BRNO UNIVERSITY OF TECHNOLOGY

Faculty of Information Technology

PHD THESIS

Brno, 2021

Ing. Janka Puterová

BRNO UNIVERSITY OF TECHNOLOGY
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FACULTY OF INFORMATION TECHNOLOGY
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

DEPARTMENT OF INFORMATION SYSTEMS
ÚSTAV INFORMAČNÍCH SYSTÉMŮ

DETECTION OF REPETITIVE SEQUENCES IN GENOMES

DETEKCE REPETITIVNÍCH SEKVENCÍ V GENOMECH

PHD THESIS
DISERTAČNÍ PRÁCE

AUTHOR **Ing. JANKA PUTEROVÁ**
AUTOR PRÁCE

SUPERVISOR **doc. Ing. JAROSLAV ZENDULKA, CSc.**
ŠKOLITEL

BRNO 2021

Abstract

Repetitive sequences can make up a significant part of the genome, in some cases more than 80%, but scientists have often overlooked them. Today we know that repeats have various functions in the genomes and are divided into two main groups: interspersed and tandem repeats. This work aimed to develop bioinformatics tools to detect repetitive sequences, either directly from sequencing data generated by sequencers or assembled genomes. In the introductory part, the work provides an insight into the issue and an overview of the repeat types occurring in genomes. Furthermore, the work deals with existing approaches and tools with an aim to detect repeats directly from the assembled sequences. The main contribution to this area was developing the digIS tool, which aims to detect insertion sequences that represent the most abundant interspersed repeats in prokaryotes. digIS is based on the principle of profile hidden Markov models constructed for the catalytic domains of transposases, representing the most conserved part of the insertion sequences and retaining a secondary structure within the family. Subsequently, the work provides an overview of sequencing technologies and discusses existing methods for detecting repeats directly from sequencing data without the need for prior genome assembly. A novel approach for a detailed analysis of tandem repeats is presented. This approach extends the primary analysis of RepeatExplorer, which detects and characterizes repeats directly from sequencing data. The work further discusses the applications of repeat detection in biological research, especially from the point of view of comparative repeatome studies and the evolution of sex chromosomes. Finally, the work summarizes the research results in the form of four articles published in international journals, the full text of which is available in the appendices, and provides a general summary of the work together with possibilities for future research.

Keywords

transposons, transposable elements, tandem repeats, satellite DNA, repetitive elements, repeatome, repeat detection, profile hidden Markov models, comparative analysis, sex chromosomes, genome evolution

Reference

PUTEROVÁ, Janka. *Detection of repetitive sequences in genomes.* Brno, 2021. PhD thesis. Brno University of Technology, Faculty of Information Technology. Supervisor doc. Ing. Jaroslav Zendulka CSc.

Abstrakt

Repetitivní sekvence mohou tvořit významnou část genomu, v některých případech více než 80 %, která však bývala vědci často přehlížena. Dnes je známo, že repetice mají v genomu různé funkce a rozdělují se na dvě hlavní skupiny: rozptýlené a tandemové repetice. Cílem této práce bylo vytvoření bioinformatických nástrojů pro detekci repetic, ať už přímo ze sekvenačních dat generovaných sekvenátory, nebo ze sestavených genomů. V úvodní části práce poskytuje náhled do problematiky a přehled typů repetic vyskytujících se v genomech. Dále se práce zabývá stávajícími přístupy a nástroji zaměřenými na identifikaci repetic přímo ze sestavených sekvencí. Hlavním přínosem do této oblasti bylo vytvoření nástroje digIS, který se zaměřuje na detekci inserčních sekvencí, které přestavují nejhojněji se vyskytující rozptýlené repetice u prokaryot. digIS je založen na principu profilových skrytých Markovových modelů zkonstruovaných pro katalytické domény transpozáz, které představují nejkonzervativnější část inserčních sekvencí a zachovávají si sekundární strukturu v rámci rodiny. Následně práce poskytuje přehled sekvenačních technologií a rozebírá stávající metody pro detekci repetic přímo ze sekvenačních dat, bez nutnosti procházejícího sestavení genomu. Je představen nový přístup pro detailní analýzu tandemových repetic. Tento přístup rozšiřuje základní analýzu nástroje RepeatExplorer, který detekuje a charakterizuje repetice přímo ze sekvenačních dat. Práce dále diskutuje aplikace detekce repetic v biologickém výzkumu zejména z pohledu srovnávacích studií repeatomu a evoluce pohlavních chromozomů. V závěrečné části práce poskytuje souhrn dosažených výsledků výzkumu v podobě čtyř článků publikovaných v mezinárodních časopisech, jejichž plné znění je dostupné v přílohách, a celkové shrnutí práce a možnosti budoucího výzkumu.

Klíčová slova

transpozony, transpozibilné elementy, tandemové repetice, satelitní DNA, repetitivní elementy, repeatom, detekce repetic, profilové skryté Markovovy modely, komparativní analýza, pohlavní chromozomy, evoluce genomu

Citace

PUTEROVÁ, Janka. *Detekce repetitivních sekvencí v genomech.* Brno, 2021. Disertační práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Školitel doc. Ing. Jaroslav Zendulka CSc.

Detection of repetitive sequences in genomes

Declaration

Hereby I declare that this Thesis was prepared as an original author's work under the supervision of doc. Ing. Jaroslav Zendulka, CSc. and Ing. Tomáš Martínek, Ph.D. The results section is based on four papers that I have written together with my supervisors and colleagues. My contribution to these papers is described in detail in the Chapter Research results summary. All the relevant information sources used during the preparation of this thesis are appropriately cited and included in the list of references.

....
Janka Puterová
September 8, 2021

Acknowledgements

I received a lot of support and assistance during my doctoral studies from many people. First and foremost, I am grateful to my supervisors, doc. Ing. Jaroslav Zendulka CSc. and Ing. Tomáš Martínek Ph.D., for their expertise, feedback, continuous support, guidance, and patience with me. I want to acknowledge colleagues from the Institute of Biophysics of the Czech Academy of Sciences, Department of Plant Developmental Genetics, especially Eduard Kejnovský and Zdeněk Kubát, for sharing their knowledge, guidance, and help.

Thanks to my friends, Marta Jaroš, Gabriela Nečasová, and Dominika Regéciová, for their mental support during the days I wanted to give up. I would also like to thank my coworkers from DNAnexus, who always listened to me and supported me during the finalization of this work.

Finally, the greatest thanks belong to my fiancé for his support, understanding, and patience during my doctoral studies.

This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic from the National Programme of Sustainability II, IT4Innovations excellence in science, grant number LQ1602; the GAČR grant GA15-02891S; the BUT grant FIT-S-14-2299; and the BUT grant FIT-S-17-3964.

Contents

List of Figures

List of Abbreviations

A	Adenine
bp	Base pair
C	Cytosine
DNA	Deoxyribonucleic acid
DIRS	Dictyostelium intermediate repeat sequences
DR	Direct repeat
EN	Endonuclease
eORF	Extra open reading frame
FP	False positive
G	Guanine
Gb	Gigabase
GS	Genome size
HEL	Helicase domain
HGT	Horizonatal gene transfer
HMM	Hidden Markov model
INT	Integrase
IS	Insertion sequence
ISE	Insertion sequence element
IR	Inverted repeat
kb	Kilobase
LINE	Long interspersed nuclear element
LTR	Long terminal repeat
Mb	Megabase
MGE	Mobile genetic element
MITE	Miniature inverted-repeat transposable element
MSA	Multiple sequence alignment
mya	Million years ago

NGS	Next-generation sequencing
ORF	Open reading frame
ONT	Oxford Nanopore Technologies
PCR	Polymerase chain reaction
pHMM	Profile hidden Markov model
PLE	Penelope-like elements
PacBio	Pacific Biosciences
PR	Protease
R	Purine
REP	Replication initiator motif
RNA	Ribonucleic acid
RT	Reverse transcriptase
SARS-CoV-2	Severe acute respiratory syndrome coronavirus 2
SMRT	Single-molecule real-time
T	Thymine
Tb	Terabase
TE	Transposable element
TIR	Terminal inverted repeat
TPase	Transposase
TR	Tandem repeat
TSD	Target side duplication
UPGMA	Unweighted pair group method with arithmetic mean
UTR	Untranslated region
satDNA	Satellite DNA
SINE	Short interspersed nuclear element
YR	Tyrosine recombinase

Chapter 1

Introduction

1.1 Motivation

Repetitive sequences are motifs of Deoxyribonucleic acid (DNA) that can occur thousands of times across a genome. They are abundant in a wide range of species, from simple prokaryotic organisms such as bacteria to complex eukaryotic organisms represented by plants, fungi, and mammals, including humans. For example, two-thirds of the human genome consists of repetitive sequences, and in plant species, the proportion of repeats in the genome can reach more than 80 %, like in maize [123].

Repeats can be divided into two main categories: tandem repeats and interspersed repeats. Tandem repeats (TRs) are composed of large arrays of tandemly repeating patterns, and they occur directly next to each other. On the other hand, interspersed repeats, also known as transposable elements (TE), are dispersed in the genome at various locations. TEs can be presented in many copies, can move within the host genome, and even create new copies of themselves.

At present, we know that repetitive DNA plays various roles in the genome. It can be beneficial when it has a specific cellular function, e.g., it serves as telomeric DNA. Another role of repetitive sequences is in genome organization, as it can significantly influence genome size due to the ability to amplify itself. Repetitive DNA has been associated with large chromosomal rearrangements such as deletions, duplications, or inversions. These rearrangements can affect the host's fitness in both positive or negative way. As TEs can move within the genome, they represent potent mutagenic agents. Their new copies can integrate directly into the gene and disrupt its function, which may result in disease. Examples of diseases caused by TEs include Hemophilia A and B (blood disease) caused by insertion of LINE1 TE [59] or Porphyria (a liver disease) induced by insertion of Alu element [94]. In other cases, they may have a regulatory function and influence gene expression by their activity. A recent study also showed that DNA copies of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) sequences can be integrated into the host genome by a retroposition mechanism mediated by LINE1 TE [158]. This integration of viral sequences can lead to PCR-positive tests even after the patient's recovery from COVID-19.

Sequencing technologies are moving forward in terms of availability, accuracy, and speed. However, the very process of obtaining data, sequencing and subsequent assembly, is complicated by the presence of repeats which is especially evident in eukaryotes, where the repeat content in the genome can reach up to 80%. Genome assembly is a computationally exhaustive and expensive process, mainly for eukaryotic species with large genomes and highly repetitive content. Thus, only a limited number of fully assembled eukaryotic genomes is available, particularly for model organisms such as human or mouse.

In the case of studying and analyzing new genomes, scientists often use low-pass sequencing, when not entire, but only a small portion representing several percent of the genomic content is sequenced. By applying this approach, we can examine the repeats' composition in the genomes without assembling and annotating the entire genome as it can capture highly and moderately abundant repeats. Although low-pass sequencing has its limitations, for example, it cannot capture low-copy repeats, it represents a good compromise. This approach can provide insight into the composition and evolution of studied genomes and perform large-scale comparative repeatome studies. For example, such studies of closely related species can shed more light on their evolutionary dynamics and determine which repetitive elements played a role in genome size evolution.

For the prokaryotic genomes, the situation is better as repeats are not represented at such a high rate as in eukaryotic genomes. Due to the smaller size of prokaryotic genomes, the process of data acquisition and analysis is simpler. Therefore, assembled genomes of prokaryotic organisms are highly available in public databases.

Despite the increasing availability of data, we are still far away from a comprehensive understanding of the function and behavior of repetitive sequences. To fully uncover the various roles of repetitive DNA in genomes, we need efficient bioinformatics tools for their detection and analysis.

The main aim of this Thesis was to develop or improve bioinformatics approaches and tools for the detection of repetitive sequences. Although plenty of computational methods and software tools focusing on this challenging task have been developed, there is still room for improvement.

1.2 Objectives of the Thesis

Due to the interdisciplinary character of this work, it will be necessary to understand the topic of this dissertation from several angles. The objectives of the presented dissertation can be summarized as follows:

- Study of repetitive sequences (biological background), their structure, understanding their complexity; study of existing approaches and methods for detecting repetitive sequences and problems associated with repeat detection from an algorithmic and computational point of view.

- Improvement of existing methods focused on detecting repetitive sequences in eukaryotic genomes and their application in biological research.

- Addressing identified shortcomings associated with detection of repetitive sequences in prokaryotic genomes.

- Design, development, and evaluation of an original and efficient method for detecting repetitive sequences in prokaryotic genomes.

1.3 Organization of the Thesis

Considering the interdisciplinary nature of this Thesis, Chapter 2 provides the reader with an overview and biological background of the central object of this work, repetitive elements, together with a summary of available repositories collecting known repetitive sequences. Chapter 3 is focused on the assembly-based computational approaches for the detection of repetitive sequences which require an already assembled reference genome. Chapter 4 provides an overview of sequencing technologies and discusses how different types of sequencing data affect repeat detection when assembly-free methods are used. The main focus is on the principles of the assembly-free methods, their advantages, and weak points. The practical use of repeat detection and its application in biological research is discussed in Chapter 5. In Chapter 6, achieved publication results are summarized, which are represented by four original publications. Full texts of these publications are available in Appendices A-D of this Thesis. Lastly, Chapter 7 is devoted to the concluding remarks and discusses possible future research.

Chapter 2

Overview of repetitive sequences

The genetic information of each living organism is encoded within a genome, and it contains a set of genetic instructions needed for the development and functioning of that organism. The genome is stored in long molecules of DNA, a double helix molecule, which is composed of four nucleotides: adenine (A), cytosine (C), guanine (G), and thymine (T) and is packaged into thread-like structures called chromosomes. Genome content can be divided into two parts: coding and non-coding DNA. Coding DNA represents regions that code for a protein product, protein-coding genes. The rest of the genome comprises non-coding DNA, which includes functional non-coding RNA (transfer RNA, ribosomal RNA, and regulatory RNAs), sequences controlling transcription and translation of protein-coding or repetitive sequences.

Cellular life forms are classified into three domains: Archaea, Bacteria, and Eukaryota. Bacteria and Archaea domains represent prokaryotic, unicellular organisms that lack a membrane-bound nucleus, and their genetic information is stored in the so-called nucleoid. These organisms possess small, compact genomes that are densely packed with protein-coding regions. The repeat content can occupy up to a quarter of the genome [125].

Eukaryotes are organisms whose cells have a nucleus within a membrane. Their genetic material is stored within this nucleus and is divided into multiple chromosomes. Organisms in this domain include animals, plants, fungi (mainly multicellular), and other groups of organisms unitedly classified as protists (many of which are single-cell organisms).

Coding DNA forms only around one percent of the human genome. The rest of the genome is formed by non-coding DNA, and studies have shown that more than two-thirds of the human genome can be made up of different types of repetitive sequences [63]. In the case of plant species, the proportion of repeats in their genome can be even higher. For example, repeats can take up more than 80% of maize genome [123].

Repetitive sequences were initially considered junk DNA, and scientists have been convinced that these sequences have no function in the genome. In the last decade, high throughput sequencing helped to revealed many new repetitive sequences. Nowadays, we know that repetitive DNA plays an essential role in genome evolution [6, 47], chromosomal rearrangements [72], gene formation and regulation [22, 85], increases genetic variation, influences genome size [90], or is involved in processes of plant sex chromosome evolution [17].

In the following sections, repetitive sequences present in both prokaryotic and eukaryotic genomes are described concerning their structural characteristics.

2.1 Repetitive sequences in prokaryotic genomes

Genomes of prokaryotes are greatly compact, with high gene density approaching 85% [92]. Prokaryotic genome sizes may vary from small and simple as the 112 kilobase (kb) genome of *Nasuia deltocephalinicola* (symbiotic bacterium) [5] to large and complex genomes as the 13.03 megabase (Mb) genome of *Sorangium cellulosum* So ce56 (soil-dwelling myxobacterium) [124]. In general, the entire prokaryotic genome is contained in a double-stranded DNA molecule organized as a single circular chromosome. The genome may also include independent smaller, circular or linear DNA molecules called plasmids, which can carry additional genes, for example, genes for antibiotic resistance [70]. As prokaryotes show variation in genome organization, some bacterial species have linear [14] or multiple chromosomes [136].

The abundance of repetitive DNA in prokaryotes is highly variable [138], and it can constitute a significant fraction of prokaryotic genomes reaching up to 25% in the genome of *Enterococcus faecalis* [125]. Prokaryotic repetitive DNA is represented by two main groups of TEs: insertion sequence elements (ISEs) and transposons, which will be described in greater detail in the following sections. TRs have been reported to be present in prokaryotes occasionally. As the structure of TRs in prokaryotes and eukaryotes is the same, they will be described in detail together with other repeats occurring in eukaryotic genomes later in this chapter.

2.1.1 Insertion sequence elements

ISEs are short fragments of the DNA sequence representing the simplest mobile genetic elements (MGEs). They can move independently within the genome and act as bacterial mutagenic agents. ISEs have a considerable impact on prokaryotic genome plasticity and adaptability, and help the host genome to adapt to new environmental challenges. They are involved in antibiotic/xenobiotic resistance, modulate metabolic activities, or virulence [142].

ISEs vary in size from 700 base pairs (bp) up to 5 kb. Their body is typically surrounded by short inverted repeats (IRs) and possesses one or two open reading frames (ORFs) coding for a protein involved in the transposition process [83, 128], the transposase (TPase), forming most of their body. Some ISEs generate direct repeats (DRs) on insertion. TPases are composed of three functional domains: the N-terminal site-specific DNA binding domain, the catalytic core, and the C-terminal protein-protein interaction domain [115]. However, only the catalytic core domain is conserved.

TPases form five main groups based on the type of chemistry they catalyze and are named after amino acid residues located in their conserved catalytic core, which include: DDE, DEDD, HUH, Tyrosine (Y), and Serine (S). DDE TPases are the most common TPases in ISEs, and their catalytic core has a typical secondary structure, a mixed alpha-beta fold, $\beta1 - \beta2 - \beta3 - \alpha1 - \beta4 - \alpha2/3 - \beta5 - \alpha4 - \alpha5/6$, also referred to as „RNase H-like fold“ [49]. The general structure of ISEs is depicted in Figure 2.1.

ISEs are classified based on a variety of characteristics, including sequence similarity of the TPases, the length and the sequences of short imperfect terminal IRs enveloping the body of ISE, the length and the sequence of the short flanking DRs, the organization of ORFs or the target region into which they insert [83]. Currently, there are 29 families of ISEs reported in the ISfinder database [129].

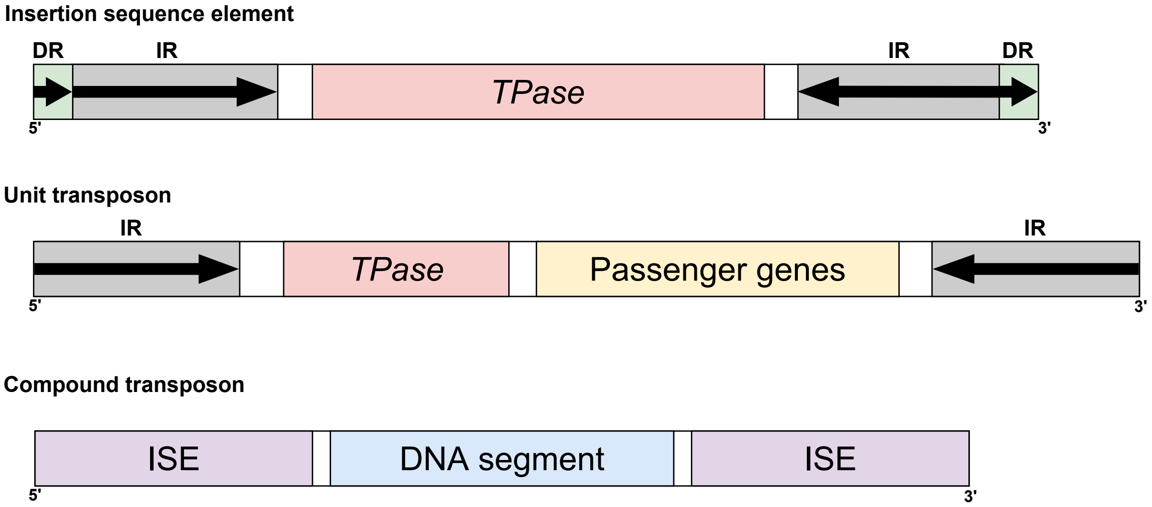


Figure 2.1: Structural features of insertion sequence elements, unit and compound transposons. DR - direct repeat, ISE - insertion sequence element, IR - inverted repeat, TPase - transposase.

2.1.2 Transposons

Transposons are more complex than ISEs. Besides the genes for transposition, they also encode additional genes. Prokaryotic transposons involve *unit transposons* and *compound transposons* and their structure is shown in Figure 2.1.

Members of the first group, unit transposons, encode additional genes, such as drug resistance genes, which are an inherent part of the transposon structure. Examples of this group are Tn3 transposons, which often carry passenger genes, particularly for mercury resistance, catabolism of xenobiotics, or individual genes engaged in antibiotic resistance.

The second group, compound transposons, represents MGEs consisting of a pair of ISEs of the same type flanking a DNA segment. Surrounding ISEs can be in either direct or inverted orientation. The central region encompasses additional genes coding for antibiotic resistance, xenobiotic catabolism, or symbiosis [15].

2.2 Repetitive sequences in eukaryotic genomes

The genome size (GS) of eukaryotic organisms is markedly more variable in comparison to prokaryotes. While the largest known prokaryotic genome reaches just several Mb, the largest eukaryotic genome belongs to a rare Japanese flower *Paris japonica* whose genome size reaches more than 149 billion bp [105] and is roughly 50 times bigger than the human genome. One of the key drivers of genome size variation in eukaryotes is repetitive sequences. They represent a heterogeneous set including dozens of families, which vary in motif length, copy number, or overall structure. There are multiple types of repetitive sequences present in eukaryotic genomes represented by two main groups: TRs and TEs.

2.2.1 Tandem repeats

TRs are highly abundant within complex eukaryotic genomes. They consist of tandemly organized repeat units called monomers. The monomer length can range in size from 2 up to several hundred nucleotides. Monomers typically form long arrays containing thousands of copies and can occupy up to 25% of plant nuclear DNA [110], or even higher proportions in some insect genomes [108]. In the human genome, they compose around 8% [31]. Copies of monomers are not entirely identical and show sequence polymorphism. Based on the monomer length and array size, TRs are categorized into three groups: *i)* microsatellites with monomer length < 9 nucleotides and array size of < 1 kb, *ii)* minisatellites with monomer length between 10 and 100 bp and *iii)* satellite DNA (satDNA) having monomers longer than 100 bp and often forms arrays longer than 100 Mb [78, 89]. Schematic representation of TR is depicted in Figure 2.2.

Different TR families may be present in a species. For example, there are 12 satDNA families in *Hippophae rhamnoides* [110], 62 families in *Locusta migratoria* [120], or 9 families within the human genome [91]. TRs can have various localization patterns and be localized on multiple chromosomes, accumulated on sex chromosomes, or appear only on the Y chromosome [110]. Multiple studies have reported accumulation of satDNA on sex chromosomes across several species [52, 84, 110].

Although TRs were initially considered to be non-functional DNA, at present, we know they have many functions in the genome. TRs are involved in chromosome organization, the control of telomere elongation, in transcriptional response during stress, or the modulation of gene expression [106, 107]. They could influence the adaptability of a host genome and influence sex chromosomes evolution [51].

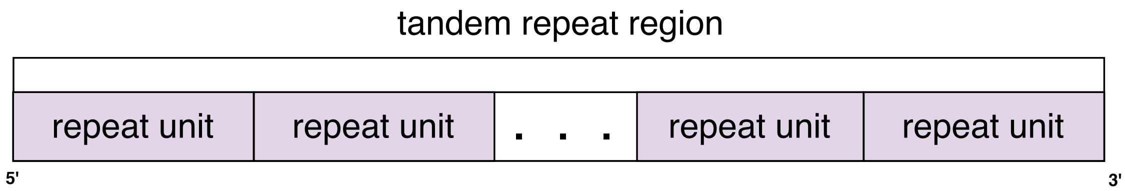


Figure 2.2: Schematic representation of tandem repeat. Repeat unit represents tandemly repeating DNA sequence – monomer.

2.2.2 Transposable elements

TEs, also called jumping genes, were firstly discovered in the 1940s by geneticist Barbara McClintock [87, 88]. TEs are dispersed in the genome at various locations. They have the ability to move or even copy themselves from one genomic location to another, which can result in their rapid amplification in the genome. TEs can occur in hundreds or even thousands of copies. Due to their repetitive nature and variability, they are challenging to analyze and remain a major challenge in the bioinformatics field.

In the beginning, they were considered junk DNA without any function and were overlooked by many researchers. Nowadays, it is known that TEs have many roles: they affect genome size [21], play an essential role in chromosomal rearrangements [41], or have been crucial players in genome evolution [10]. TEs compose a significant part of eukaryotic genomes and have been found in almost every organism studied so far. For example, they occupy 37% of the mouse genome [146], about 50% of the human genome [65] and around 80% of the maize genome [123]. Many different types of TEs have been found since McClintock's discovery. TEs are divided into two major classes according to whether they transpose via an RNA intermediate, *Class I* - retrotransposons, or a DNA intermediate, *Class II* - DNA transposons. These classes are further subdivided into several subclasses with respect to their chromosomal integration mechanism. Another classification of both *Class I* and *Class II* elements is based on whether TEs encode all domains needed for their transposition or not and divides them into autonomous and non-autonomous elements, respectively. Non-autonomous ones may arise in various ways, e.g., deriving from autonomous copies which gathered mutations; thus, they do not encode necessary domains for transposition anymore.

In the following sections, structure and characteristic features of *Class I* and *Class II* TEs will be described based on [35, 93, 149].

2.2.3 *Class I* – Retrotransposons

Elements included in this class transpose via an RNA intermediate when the RNA intermediate is transcribed from a genomic copy followed by reverse transcription into DNA by reverse transcriptase (RT) encoded in TE. This mechanism is generally called copy-and-paste because each complete replication cycle generates a new copy of TE. The classification given by Wicker et al. [149] divides retrotransposons according to their mechanistic features, organization, and reverse transcriptase phylogeny into five groups: LTR retrotransposons, *DIRS*-like elements, *Penelope*-like elements, LINEs, and SINEs. The structure of elements belonging to these groups is depicted in Figure 2.3.

LTR retrotransposons are composed of long terminal repeats (LTRs), which enclose the retrotransposon's body coding protein domains essential for the transposition process. LTRs length ranges from a few hundred bp up to 6 kb, and they begin with 5'-TA-3' and end with 5'-CA-3' pattern. LTR retrotransposons usually contain two protein-coding ORFs, *gag* and *pol*, but as an exception, additional ORFs of unknown function may be present [60, 64]. *pol* encodes several protein domains (reverse transcriptase – RT, protease – PR, RNase H – RH, and integrase – INT), which carry out reverse transcription and integration into a new location in the genome. After integration, they generate a target site duplication (TSD) of length 4-6 bp. The overall length of these elements can reach surprising 25 kb [81].

Dictyostelium Intermediate Repeat Sequences (DIRSs) encode a tyrosine recombinase (YR) domain instead of the INT and therefore do not produce TSDs. Elements in this group possess terminal sequences which resemble either IRs or split DRs.

***Penelope*-like elements** (PLEs) encode only two protein domains, RT and endonuclease (EN). Their enclosing repeats can be in direct or inverse orientation.

Long Interspersed Nuclear Elements (LINEs) can be several kb long, absent LTRs, and contain *pol* ORF encoding at least the RT and a nuclease (EN or an apuric or apyrimidic EN). A *gag*-like ORF of unknown function is sometimes found 5' to *pol*. The coding region can be flanked by untranslated regions (UTRs) from each side of the element [122]. LINEs produce TSDs, but they are difficult to find because of truncated 5' ends. Their 3' end may contain a poly(A) tail, TR, or A-rich region.

Short Interspersed Nuclear Elements (SINEs) are non-autonomous and originate from accidental retrotransposition of various polymerases III transcripts. They depend on partner LINEs because they use the RT domain from LINEs for their reverse transcription. These elements are relatively short, ranging between 80 and 500 bp, and generate TSDs from 5 to 15 bp. SINEs are terminated by A- or AT-rich region or by poly(T) tail.

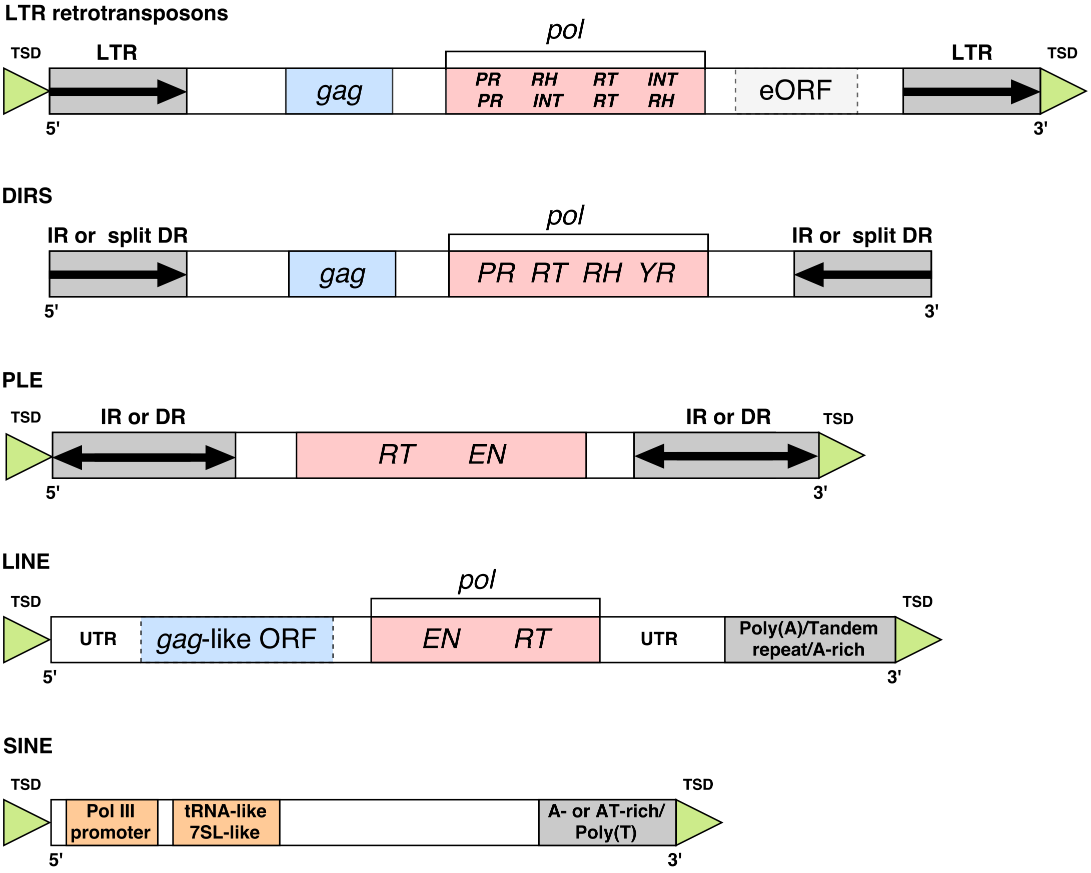


Figure 2.3: Structural features of retrotransposons. DR - direct repeat, EN - endonuclease, eORF - extra open reading frame, *gag* - *gag* gene, INT - integrase, IR - inverted repeat, LTR - long terminal repeat, ORF - open reading frame, *pol* - *pol* gene, PR - protease, RH - RNase H, RT - reverse transcriptase, UTR - untranslated region, TSD - target site duplication, YR - tyrosine recombinase.

2.2.4 *Class II* – DNA transposons

DNA transposons usually transpose in the genome by a cut-and-paste mechanism using a DNA intermediate, but there are some exceptions. DNA transposon is cut out from the current chromosomal location and reinserted into a new one during this process. Because most DNA transposons move through a non-replicative mechanism (do not generate copies of themselves), they usually occur in low copy numbers. Most eukaryotic DNA transposons have relatives among the prokaryotic ISEs [49, 50]. According to the classification proposed by Wicker et al. [149], DNA transposons include two subclasses based on the number of DNA strands that are cut during transposition: *Subclass I* and *Subclass II*.

Subclass I involves two orders of elements, TIR and *Crypton*, which transpose by the cut-and-paste mechanism and their structure is showed in Figure 2.4.

- TIR DNA transposons are characterized by terminal inverted repeats (TIRs) varying in length and the TPase surrounded by these TIRs. In the transposition process, elements are cut out from a current location in the genome and reintegrated into a new chromosomal location as double-stranded DNA. This process is mediated by the TPase encoded in the element. TIR DNA transposons are capable of increasing their copy numbers by moving during chromosome replication when they transpose from a position that has been already replicated to another position that the replication fork has not yet passed [149]. They are distinguished into nine superfamilies, Tc1/*mariner*, PIF/*Harbinger*, *h*AT, Mutator, Merlin, Transib, P, *piggyBac* and CACTA, according to the TIR sequences and TSD size. TIR DNA transposons vary in length ranging between 2 kb and 15 kb, and some superfamilies possess the second ORF of unknown function.

- *Crypton* DNA transposons were found only in fungi and are poorly known. They are composed of the YR, lack TIRs, and seem to generate TSDs. Their transposition also requires cutting both DNA strands.

Subclass II consists of DNA transposons called *Helitrons* and *Mavericks* that use transposition process requiring replication without cleavage of both DNA strands and transpose by the copy-and-paste mechanism. Their structure is illustrated in Figure 2.4.

- *Helitrons* replicate through a rolling-circle mechanism when only one DNA strand is cut. They encode Y2-type tyrosine recombinase with a helicase domain (HEL) and replication initiator motif (REP). *Helitrons* do not generate TSDs, and their ends can be determined by TC or CTRR motifs (R is purine) and short hairpin structure before 3' end.

- *Mavericks* reach from 10 to 20 kb in length. They are surrounded by long TIRs and can encode up to 11 proteins. *Mavericks* encode a DNA polymerase B, INT, and do not contain RT. This suggests that replicative transposition without RNA intermediate is used.

DNA transposons also include non-autonomous elements known as Miniature Inverted-repeat Transposable Elements (MITEs), which do not encode proteins and have no coding potential. Hence, their transposition is presumably dependent on autonomous transposons. They are widely distributed in eukaryotes [45, 140]. MITEs are typically short, and their length varies between 50-800 bp. They contain short conserved TIRs flanked by TSDs, which are common features of DNA transposons [93]. MITEs are classified into superfamilies based on the composition of their TIRs and the length of TSDs. They are usually located in gene-rich regions and presented in high copy numbers [11, 12]. Several studies suggested that MITEs can affect the expression of nearby genes [95, 153], play an important role in genome size evolution as the number of MITE sequences significantly correlates with genome size [19] and can even have a role in phenotypic diversity [20].

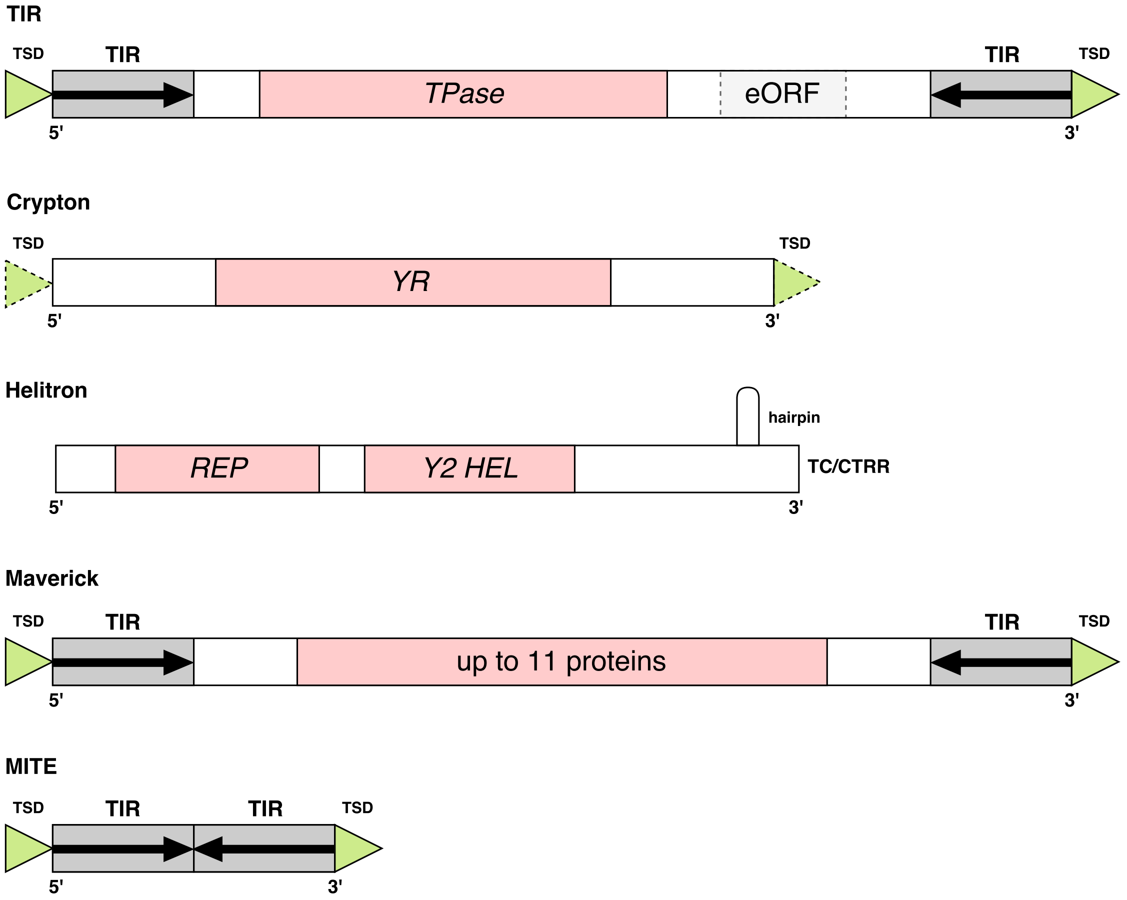


Figure 2.4: Structural features of DNA transposons. eORF - extra open reading frame, HEL - helicase domain, REP - replication initiator motif, TIR - terminal inverted repeat, TPase - transposase gene, TSD - target site duplication, YR - tyrosine recombinase, Y2 - Y2-type tyrosine recombinase.

2.3 Databases and repositories of repetitive sequences

In addition to the annotation of gene space, detailed annotation of repeated sequences is critical for understanding the structure of genomes and an important tool for improving the quality of genome assemblies. Since more and more genomes are being studied each year, numerous repositories focused on gathering repeat sequences started to emerge as the amount of data is growing. They are used to classify and annotate repeats in genomes and are utilized by many bioinformatics tools that require a reference library of repetitive sequences. These repositories can be divided into two types: *i)* repeat-centric, and *ii)* genome-centric. This division is not unambiguous as one database can focus on a certain type of repeats in a particular group of organisms, such as plants. The following sections are based on the recent reviews [38, 101].

2.3.1 Repeat-centric repositories

The most frequently used repository is Repbase [4], which contains the most extensive collection of eukaryotic TEs and other repetitive sequences in the form of consensus sequences. As one repeat family occurs in multiple copies within the genome, which are not identical due to mutations gathered over time, the consensus sequence is calculated from several copies by performing a multiple sequence alignment (MSA), and the most frequent residue is used at each position. One major drawback of consensus sequences is that we lose information about the variability of a given repeat. To capture the divergence of a given repeat family, it is appropriate to use profile hidden Markov models (pHMMs) that capture position-specific information about how conserved each nucleotide or amino acid is together with a degree to which gaps and insertions have occurred.

Another database of repeat families, Dfam [54, 135, 148], contains MSA and pHMM built from that alignment instead of the consensus sequence for each repeat family. As pHMMs improve detection of remote homologs of known repeat families by increased sensitivity, they are able to find degenerated copies of repeats that accumulated mutations over time and are difficult to detect by using other approaches otherwise.

Gypsy Database (GyDB [77]) is focused on LTR retrotransposons, especially on those from Gypsy and Copia superfamilies along with Bel/Pao LTR retrotransposons, Retroviridae-like elements, and the Caulimoviridae pararetroviruses of plants. SINEBase [144] is a dedicated resource of SINEs found in eukaryotic genomes. Plant MITE (P-MITE [19]) database stores MITE families identified in plant species. TRs were not forgotten by researchers either, and several repositories that collect them are available, such as PlantSat [80], Tandem Repeat Database (TRDB) [37], or MicroSatellite Database (MSDB) [3].

All the databases mentioned above are focused on repeat families in eukaryotic genomes. On the other hand, ISfinder database [129] collects ISEs isolated from prokaryotic genomes and stores them in the form of individual sequences. The Transposon Registry (TTR) [137] aims to transposons in the bacterial and archaeal genomes and provides a searchable repository for all transposons. Recently, a database of prokaryotic transposons, TnCentral [117] was posted.

2.3.2 Genome-centric repositories

These repositories collect all known repetitive sequences from a single species or group of closely related species. These comprehensive databases are built to better understand the role of TEs and other repetitive sequences, their structural, functional, and evolutionary dynamics. The single-species databases include:

- *BmTEdb:* a database collecting TEs of silkworm (*Bombyx mori*) [151].
- *GrTEdb:* the first web-based database of TEs in cotton (*Gossypium raimondii*) [152].
- *MnTEdb:* a resource of TEs in mulberry (*Morus notabilis*) [79].
- *RepPop:* a database focusing on repeats in cottonwood (*Populus trichocarpa*) [159].
- *RetrOryza:* a reference database of LTR retrotransposons in rice (*Oryza sativa*) [16].
- *SoyTEdb:* a database of TEs in the soybean (*Glycine max*) genome [30].

Databases collecting repeat families from groups of organisms are represented by RiTE database [26], DPTEdb [73], FishTEdb [126], ConTEdb [155], SPTEdb [154], or TIGR Plant Repeat Database [102][1]. Plant Genome and System Biology Repeat Database (PGSB-REdat) [133] stores individual sequences of repetitive elements found in plants and offers various browsing methods. REXdb [97] is a comprehensive database of retrotransposons protein domains sampled from 80 species representing major groups of green plants. It provides a reference for efficient and unified annotation of LTR retrotransposons in plant genomes. MITEs from 98 insect genomes are available in iMITEdb [46].

[1] TIGR Plant Repeat Database was discontinued on February 8, 2017

Chapter 3

Assembly-based approaches for detection of repetitive sequences

Repeat detection or identification is a process by which we search for genomic regions representing repetitive sequences. Many computational tools have been developed to detect repetitive sequences in both prokaryotic and eukaryotic assembled genomes.

They employ several approaches, often utilizing available knowledge about the repeat families, which is considered during the detection process and include *i)* library-based, *ii)* signature-based, and *iii) de novo* approach. These tools may also include an annotation step during which they assign the repeat family, subfamily, or even label the structural features such as LTRs or ORFs to detected repeat.

In this chapter, we will focus on detecting TEs and other repeat types in already assembled sequences, providing an overview of existing approaches and tools available, followed by a summary of each approach's characteristics and discussing their limitations. Finally, we will focus on the main contributions to assembly-based approaches for identifying repetitive elements developed as a part of this Thesis. The following overview is based on the review of tools for repeat detection by Lerat from 2010 [68] and has been updated with current approaches and tools for repeat detection.

3.1 Library-based approach

This approach is based on the similarity searches of input sequences, e.g., reference genomes, against a library of known repeats collected in a database and its main principle is depicted in Figure 3.1. Library-based tools (also known as homology-based or repository-based) are dependent on a source of known repeats. Their performance is affected mainly by the quality of the used reference library. The reference library can be either created and customized by the user or a comprehensive and curated repeat library such as Repbase [4].

Based on the type of provided reference library of repeats, the library-based method can be divided further into *i)* sequence-based and *ii)* profile-based methods.

3.1.1 Sequence-based method

Tools that utilize the sequence-based method use either sequences of individual copies of repeats or consensus sequences of repeat families as a reference library. For performing the sequence-similarity search between the reference database and input sequences, a software tool for local pairwise alignment such as BLAST [2] is often used. In short, it finds regions with local similarity and calculates the statistical significance of individual matches. It allows for mismatches and shorter gaps (insertions and deletions) to be introduced in the alignment. Tools employing this method can detect closely related sequences only and struggle with longer gaps; thus, their ability to detect remotely homologous members of known repeats families or new repeats is considerably limited.

3.1.2 Profile-based method

This method takes advantage of pHMMs known to improve detection sensitivity over the conventional sequence-similarity search method. pHMMs are statistical models used for modeling a particular sequence family of interest. They convert MSA into a position-specific scoring system that reflects variation levels within the sequence family [32]. pHMMs are capable of dealing with a higher level of divergence within repeat families than the sequence-based method and detect remote homologs. Concerning repetitive sequences, pHMMs are usually constructed for protein domains contained in them, such as TPase or EN. However, they can also be built for terminal regions of repeats such as LTRs. pHMMs are used in combination with HMMER software package [33], which detects homologous sequences by searching the input sequences using the constructed pHMMs.

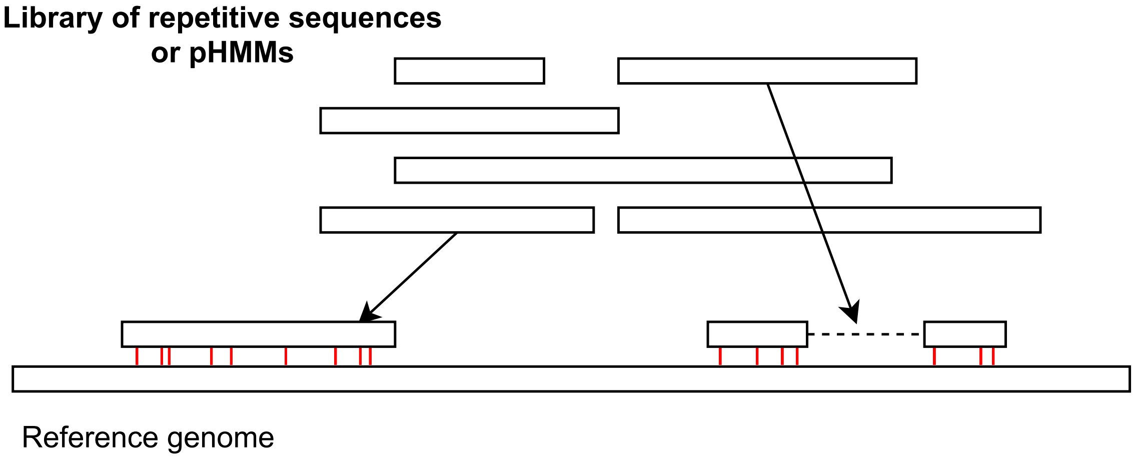


Figure 3.1: Principle of the library-based approach. The reference genome is searched for similarities with a library of repetitive sequences or pHMMs. Here, two mathces with the library were found. Mismatches are depicted in red vertical line, gaps are depicted in dashed horizontal line.

3.2 Signature-based approach

Tools based on this approach take advantage of prior knowledge about repeats. As described in Chapter 2, each repeat type has a set of unique structural features, for example, TSDs, a poly-A tail, TIRs, LTRs, hairpin loops, or conserved motifs on 3' or 5' ends. Tools utilizing this approach search for the occurrences of these structures and conserved motifs that are characteristic for a given repeat type by employing various algorithms such as suffix-arrays, and considering other characteristics such as size range of the repeat family, the maximum distance between LTRs or TIRs of the element, or the percentage of identity between LTRs.

Besides, they have the potential to find new repeat elements but not a new repeat type, which may have completely different structural characteristics compared to hitherto known repeat types as our current knowledge about repetitive elements limits them. The principle of the signature-based approach is shown in Figure 3.2.

Tools adopting the signature-based approach do not rely on repeat databases and are ideal for detecting repeats in newly assembled genomes for which a library of known repeats is not available in sufficient quality. They can detect repeat elements with low sequence similarity to known sequences of repeats or families with an atypical structure, such as non-autonomous elements missing ORFs. However, they suffer from a high number of false positive (FP) results [68] and also struggle to detect fragmented elements whose structure was disrupted by larger insertion or deletion.

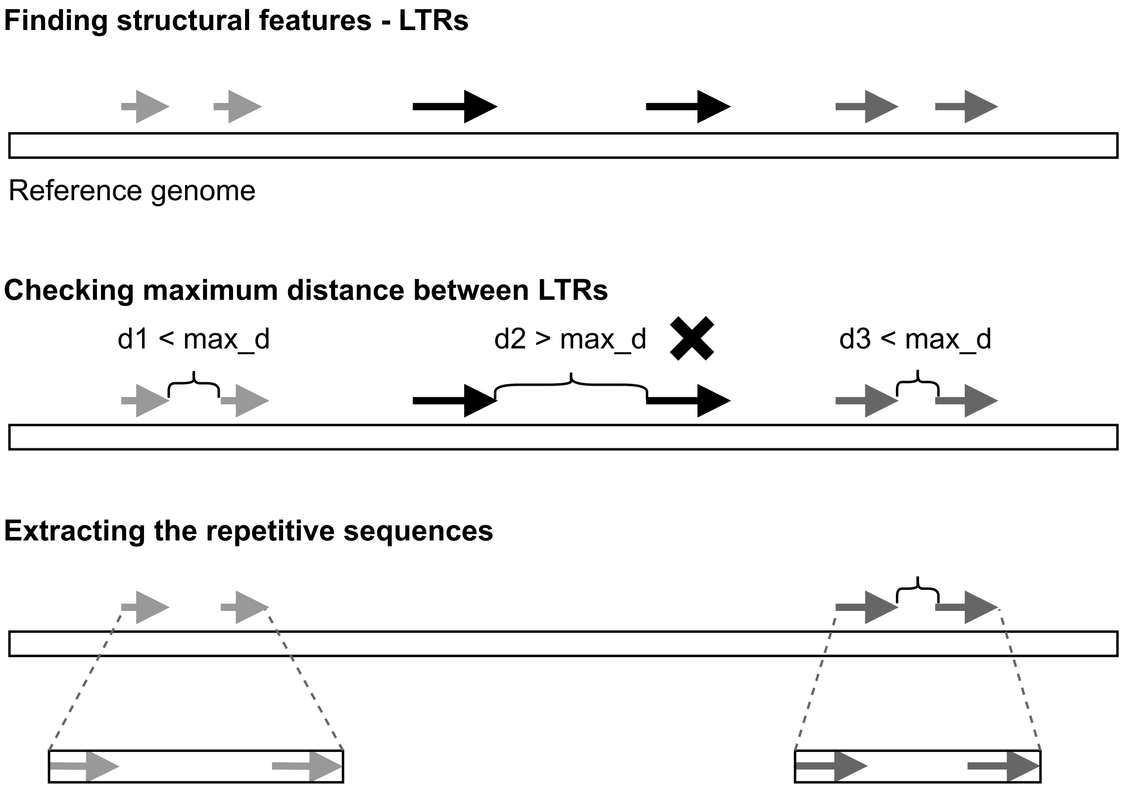

Finding structural features - LTRs
Reference genome
Checking maximum distance between LTRs
d1 < max_d
d2 > max_d
d3 < max_d
Extracting the repetitive sequences

Figure 3.2: Principle of the signature-based approach. The reference genome is searched for structural features, for example, LTRs, depicted as arrows. Other structural features and repeat type characteristics are further verified to eliminate FPs. In this case, we look at the maximum distance between LTRs. If the distance is smaller than the set threshold *max_d*, the repeat is extracted and reported.

3.3 *De novo* repeat detection

De novo (meaning „from the new“) approach does not rely on any prior knowledge about repeats, such as structural features or known sequences. Before high-throughput sequencing technologies were invented, *de novo* methods were applied to the entire assembled genomes. There are two main approaches which are used for *de novo* repeat detection in both assembled and raw sequences: *i)* self-comparison, and *ii)* k-mer counting approach. Tools employing the *de novo* approach do not focus on a particular repeat type and can detect all repeats in general. This section will focus on *de novo* methods and their use to detect repetitive sequences in assembled sequences. Their application to detect repetitions in raw sequencing data will be discussed in detail in Chapter 4.

3.3.1 Self-comparison approach

The principle of the self-comparison approach is that it uses sequence similarity detection algorithms such as BLAST [2] to perform pairwise alignment between sequences of interest or with the sequence itself to detect repeated regions. These regions can be extracted from the assembled input sequences and subsequently analyzed, for example, clustered based on their sequence similarity into groups representing repeat families.

The main advantage of this method is that it can deal with the diversity of repeats. On the other hand, tools using this approach are dependent on the availability and quality of the assembled genomes, do not scale well, and are not suitable for the analysis of large genomes sequenced nowadays.

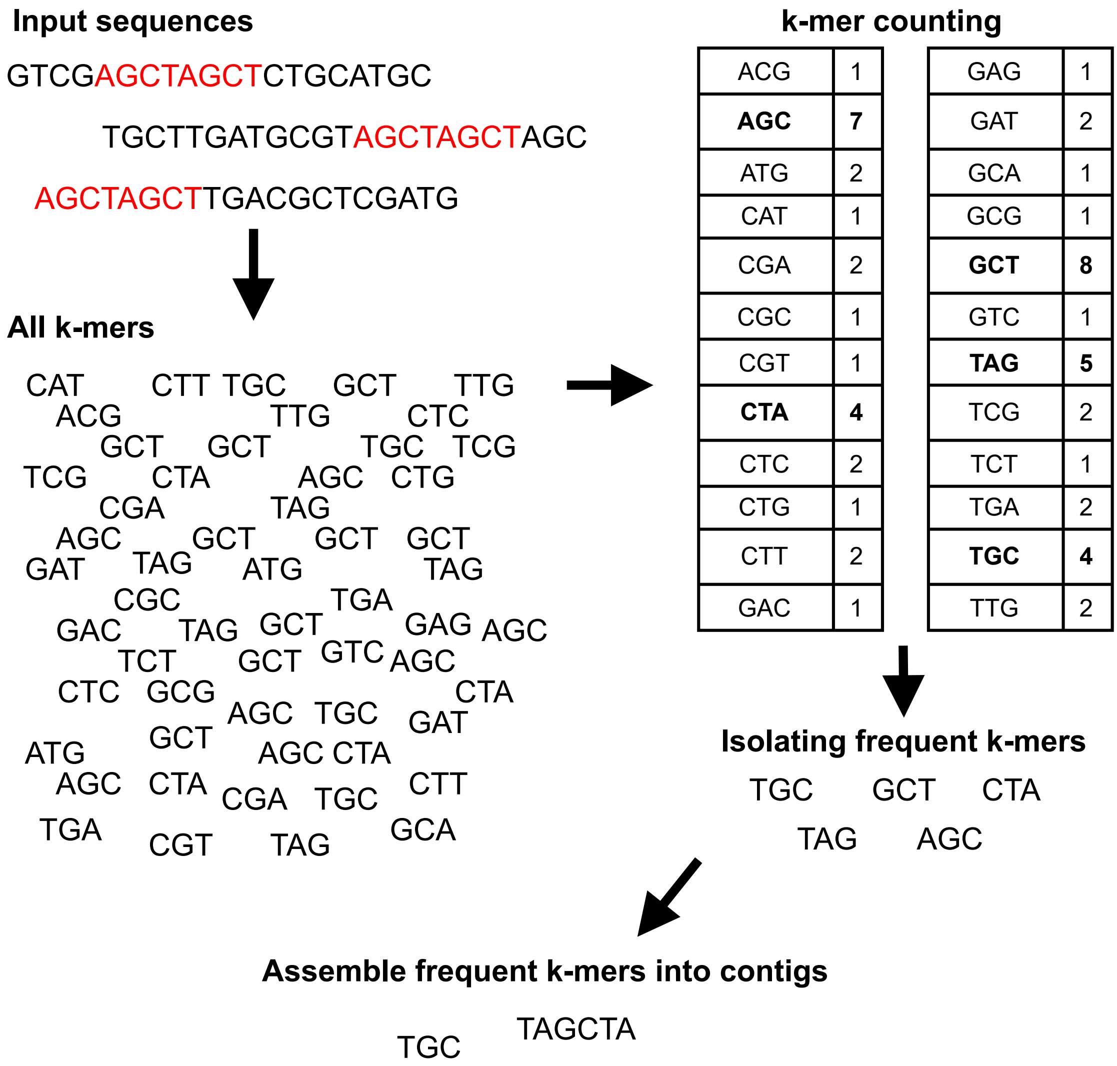
3.3.2 k-mer counting approach

The fundamental principle of the k-mer counting approach, illustrated in Figure 3.3, is that it views repeats as substrings of length k called k-mers. Thus, a genomic region containing frequent k-mers is highly likely to be a repeat. Input sequences are split into overlapping fragments - k-mers. The occurrence of each exact k-mer is counted, and efficient data structures for their storage and fast lookup are utilized, such as suffix trees, suffix arrays, or hash tables. The highly frequent k-mers can be further assembled into contigs. By using this approach, only conserved repeats can be detected.

The number of possible k-mers, which can appear in the genomic sequence, grows exponentially with k-mer length what makes this approach computationally challenging. Choosing the right length of k-mers is a difficult task. Short k-mers lead to low specificity and high sensitivity as they are found too often within the genome. On the other hand, longer k-mers result in high specificity and low sensitivity, but they are not suitable for the detection of degenerated repeats.

As repeats accumulate mutations over time, new approaches have emerged from k-mer approach to detect divergent repeats, such as *seed extension* or *spaced seeds* approaches. The seed extension approach uses frequent k-mers as seeds which are subsequently extended into a longer consensus sequence. Spaced seeds approach brings a higher level of variability tolerance. Instead of looking for exact k-mers, it allows for variation in the length or sequence identity of the seed. All these approaches help provide a quick initial overview of repeat content in newly assembled genomes, for which a library of known repetitive sequences is not available. However, more and more organisms have been sequenced with much higher coverage. To perform genome assembly of these organisms is not only computationally consuming but also financially demanding. The k-mer counting and seed approaches are nowadays used to detect repeats directly from NGS data to avoid the expensive assembly step. If researchers still decide to assemble the genome, these approaches can be used for filtering purposes to remove repeats before performing the assembly. Whereas researchers are usually only interested in coding regions, removing repeats can significantly reduce assembly cost and improve its quality as a large number of repeats in the genome can lead to fragmented and low-quality assembly.

k-mer and seed approaches struggle with the detection of low-copy repeats, especially if they are highly divergent, and do not provide many details about individual repeats like precise boundaries.



ACG	1	GAG	1
AGC	**7**	GAT	2
ATG	2	GCA	1
CAT	1	GCG	1
CGA	2	**GCT**	**8**
CGC	1	GTC	1
CGT	1	**TAG**	**5**
CTA	**4**	TCG	2
CTC	2	TCT	1
CTG	1	TGA	2
CTT	2	**TGC**	**4**
GAC	1	TTG	2

Figure 3.3: Principle of the *k*-mer counting approach. Input sequences are split into *k*-mers, frequency of each *k*-mer is counted. *k*-mers with high number of occurrences are assembled into contigs. Repetitive sequence presented in the input sequences is labeled in red color.

3.4 Hybrid approaches for repeat detection

Choosing the right method for repeat detection depends on various factors. The first factor is the availability of data. When the existing repeat library is not of sufficient quality, or we are studying new genome, which has not been sequenced yet and may contain new repeats, choosing a suitable method that can detect repeats without any prior knowledge or repeats database is necessary. Such methods include *de novo* and signature-based methods.

Another factor that needs to be considered is the repeat's structural characteristics. For detecting repeat types having structural features which have a strong signal, e.g., LTRs are long enough and are present in the repeat family, signature-based tools are preferred. As discussed in Section 3.2, signature-based methods tend to report many FPs. To tackle this issue, the signature-based approach is often combined with the library-based approach, especially in a situation when the repeat type contains conserved protein domains, to eliminate the FPs.

On the other hand, for repeat types without structural features or with highly variable characteristics across a given kind of repeat, the signature-based method will not be applicable. For such repeats, it is advisable to use the library-based approach and focus mainly on detecting conserved parts of these repeats, such as protein domains.

3.5 Novel approach for detection of ISEs and the digIS tool

As discussed earlier in Section 2.1.1, ISEs are small segments of DNA occurring in prokaryotic genomes which can move within it. ISEs may play various roles in the genome. For example, they can be involved in antibiotic resistance or enabling the host genome to adapt to a new environment. Their body typically encodes only for a protein that catalyzes the transposition, TPase, and is flanked by short IRs and DRs. TPase consists of three functional domains from which the catalytic core domain is conserved.

ISEs currently include 29 subfamilies, which vary in length, number of ORFs, or differ in the presence of IRs and DRs. Since the structural features are not shared across all ISEs subfamilies, tools designed for their detection focus on finding their conserved part, TPase, first, and the structural features are searched afterward.

Several tools have been developed over the last 15 years to detect ISEs. The first tools used a sequence-based approach and include IScan [145], ISsaga [143], and OASIS [116]. Successors of these tools use a profile-based approach and are represented by TnpPred [114], pipeline proposed by Kamoun et al. [58], and ISEScan [150].

However, only OASIS and ISEScan are currently publicly available, can be installed locally, and run on a computing cluster. Moreover, OASIS requires an already annotated genome what makes it unsuitable for a large-scale analysis of newly assembled prokaryotic genomes. Although ISEScan uses pHMMs, these models can be too specific as they were constructed for whole TPases, leading to high levels of FPs and ignoring distant members of known insertion sequence (IS) families and novel ISEs.

In response to the shortcomings of the tools designed to detect ISEs, we proposed and implemented new approach for detecting novel ISEs and distant members of already known IS families. This approach is implemented in a software tool digIS [?]. The fundamental principle of this tool lies in the detection of the most conserved part of TPases – their catalytic domain – instead of entire TPases.

digIS utilizes manually curated pHMMs constructed for catalytic domain of TPases. By applying this approach, it is able to search for these catalytic domains in prokaryotic genomes. The resulting hits serve as *seeds* and are further filtered based on various thresholds (e.g., domain bit score, e-value) to eliminate FPs. Overlapping consecutive seeds or seeds within a certain distance are merged. Their boundaries are further extended by matching against the database of known ISEs - ISfinder [129]. Subsequently, additional filtration criteria are applied (noise cutoff score, length, duplicates removal) to reduce FPs. Finally, the remaining extended seeds are classified based on GenBank annotation (if provided) and sequence similarity.

digIS is implemented as a command-line tool developed in Python3 and incorporates external bioinformatics tools and libraries (BLAST [2], HMMER [33], Biopython [25]). As an input, it takes assembled genomic sequences in FASTA format and GenBank annotation as an optional input, which is used to improve the classification of detected ISEs.

digIS was evaluated against OASIS, ISsaga, ISEScan, and ISEScan in configuration to search for fragments on two benchmark datasets (*E.coli* [48] and ISbrowser [61]) and two large datasets (NCBI Archaea (347 genomes) and NCBI Bacteria (2 500 genomes) [121]). During the evaluation, we also pointed out deficiencies in evaluating these tools using benchmark datasets, for example, reporting different types of outputs – full-length ISEs, fragments of ISEs, and ORFs.

The evaluation results showed that digIS could find already known as well as putative novel ISEs while maintaining a moderate level of FPs and high sensitivity, which was demonstrated by providing examples of putative novel ISEs found exclusively by digIS. The full text of the published article is available in Appendix D.

Chapter 4

Assembly-free approaches for detection of repetitive sequences

The vast majority of organisms do not have an available assembled reference genome, their reference genome is of low quality, or only little information about repeat content is available. With the rise of high-throughput sequencing technologies in the last two decades, there has been a demand for tools that would identify repetitive sequences directly from raw, unassembled sequencing reads, without the need to assemble the entire genome since it is an expensive and computationally intensive process. In this scenario, when dealing with an utterly unknown genome for which only sequencing reads are available, assembly-free approaches are used to detect repetitive elements. Multiple algorithms have been developed to address assembly-free repeat detection and based on their core idea they can be divided into three categories: *i) k*-mer-based, *ii) de Bruijn* graph-based, and *iii)* graph-based clustering algorithms.

At the beginning of this chapter, we will describe existing sequencing technologies and focus on the characteristics of the data they produce. Afterward, we will provide an overview of assembly-free methods used for detecting repetitive sequences and discuss their strengths and weaknesses in Sections 4.2 and 4.3. The last part of this chapter is devoted to our original contribution to assembly-free methods for detecting repeats.

4.1 Introduction to sequencing technologies

Sequencing technologies have made massive progress in the last 20 years, and new methods are continually emerging and commercialized. Nevertheless, we are still unable to obtain the DNA sequence of organisms as a continuous string, but only in the form of relatively shorter or longer DNA pieces called sequencing reads produced by sequencing machines. The following sections are devoted to individual sequencing technologies.

4.1.1 Sanger sequencing

In 1977, Frederick Sanger and his colleagues developed DNA sequencing technology, known as Sanger sequencing. This technology was adopted as a primary sequencing method for the next three decades. It produces sequencing reads of length up to 1000 bp with high per-base accuracy of 99.999%. On the other hand, it is limited in throughput as it can perform only hundreds of sequencing reactions at once, producing only several Mb of data per day [127].

4.1.2 Next-generation sequencing technologies

The goal to reduce the cost of human genome sequencing to $1000 stimulated the development of so-called next-generation sequencing (NGS) technologies. Compared to Sanger sequencing, NGS technologies are able to perform several thousand to millions of sequencing reactions in a parallel way. They revolutionized the genomics field by reducing the per-base cost of sequencing, providing high throughput and their ability to sequence entire genomes in a matter of days or even hours.

Their main drawback is that they produce relatively short reads. On the other hand, the loss of information about sequence continuity is compensated by high coverage depth. Different types of NGS technologies will be described in the following sections.

454 sequencing

The first NGS technology, called 454 sequencing, was introduced in 2005 by 454 Life Sciences. It is based on the pyrosequencing method, and in its beginnings, it was able to generate 110 bp long reads with throughput 20 Mb per run [141]. The latest sequencing machines could reach a read length of about 700 bp with an accuracy of 99.9%. The limitation of this method is a high error rate in poly-bases longer than 6 bp. Data output can reach 14 gigabases (Gb) per run depending on the used machine [76]. At present, this technology is not supported and was discontinued because other NGS technologies overran it.

Illumina sequencing

A year later, in 2006, Illumina released a sequencing method using the sequencing by synthesis approach. First Illumina instruments generated an output of 1 Gb per run and produced very short reads, only 35 bp [141]. Since then, Illumina has improved sequencing technology gradually and currently provides various types of sequencing machines. They range from small, benchtop sequencers, e.g., iSeq 100, whose maximum output is only 1.2 Gb per run. They are relatively affordable and suitable for research laboratories.

On the other hand, the production-scale sequencers, such as NovaSeq 6000, can generate the maximum output of 6000 Gb within 1 billion reads at 2 × 150 bp read length in less than two days. Production-scale sequencers are costly as their price can reach almost one million dollars. Illumina supports multiple sequencing protocols, including genomic, exome, targeted sequencing, metagenomic, or RNA sequencing[1]. Illumina sequencing has a higher error rate than Sanger sequencing reaching 0.1% [113]. The most common error is the base substitution [29].

Ion Torrent

Another sequencing technology, Ion Torrent, was firstly introduced in 2010. It uses semiconductor sequencing technology [119] and provides various sequencing machines with different throughput. For example, Ion S5 XL System can generate 15 Gb of data representing 60-80 million reads within less than one day. The length of reads generated by Ion Torrent machines started at 100 bp and now can reach 600 bp[2]. This technology

[1] www.illumina.com

[2] www.thermofisher.com

has high error rates in homopolymer repeats (e.g., TTTTTT), same as 454 sequencing technology [141]. The overall error rate is 1% [113].

4.1.3 Third-generation sequencing technologies

In recent years, new sequencing technologies producing much longer reads compared to NGS technologies started to emerge. These sequencing technologies are referred to as „third-generation“ and offer various advantages over NGS. Currently, two technologies lead in the long-read sequencing field: single-molecule real-time (SMRT) sequencing from Pacific Biosciences (PacBio) and nanopore sequencing from Oxford Nanopore Technologies (ONT).

SMRT sequencing

PacBio developed the SMRT sequencing method, which was commercially released in 2011. The SMRT sequencing generated longer reads with a maximum read length over 60 kb in the early stages. The biggest weakness of PacBio reads is a high error rate (∼11%) of single-pass reads with insertions or deletions as prevalent error [112]. These errors are distributed randomly; thus, multiple passes over the same DNA template and calling the consensus of these reads allow for accurate reads, so-called circular consensus sequences. By applying this approach, SMRT sequencing can generate long high-fidelity (HiFi) reads with accuracy over 99.9% while maintaining the average read length of 13.5 kb [147]. The throughput of the newest sequencers from a Sequel II System series is averaging 160 Gb per SMRT Cell within one day[3].

Nanopore sequencing

Nanopore sequencing from ONT is based on the idea that individual nucleotides induce different ionic current changes. The current is measured as the DNA strand is passed through a tiny channel, a biological nanopore. The first commercially available sequencer using this method, MinION, was released in 2014. MinION is a tiny portable sequencer. Its low device price (only $1000 for basic starter pack) makes it stand out from other existing sequencing technologies and is affordable for small laboratories [103, 113]. Its compactness and portability allowed, for example, to sequence DNA in microgravity on the International Space Station. The first reports claimed the error rate of reads more than 38% [67]. Since then, the chemistry and base calling software have been improved significantly. The other two sequencers offered by ONT, GridION and PromethION, are benchtop size sequencers. The latter has the capacity to generate up to 8 terabases (Tb) of data, making it suitable for projects focused on larger genomes or population-scale sequencing[4]. The longest read ever sequenced by this technology has an exceptional 2.3 Mb [104]. However, the current error rate is still relatively high, ranging in 5-20% [111]. The errors include both insertions and deletions, which seem to be a systematic error, unlike SMRT sequencing, and additional NGS data are typically required for error correction [55].

[3] www.pacb.com

[4] www.nanoporetech.com

4.1.4 Effects of sequencing data on repeat detection

Nowadays, NGS data dominate over third-generation sequencing data. However, the situation can rapidly change in the next few years. NGS sequencers generate data characterized by short read length and high accuracy, and they are currently used in most existing tools designed for the detection or reconstruction of repeats. These tools tend to identify only fragments of repeats, and often subsequent assembly step is required to recover long repeats.

Data generated by third-generation sequencers is characterized by a higher level of errors and greater length compared to NGS data. Due to their length, they can span entire repeated elements and help resolve gaps, ambiguity and reconstruct full-length repeats [57]. With the advent of third-generation sequencing technologies, approaches for repeat detection utilizing both NGS and third-generation sequencing reads have started to emerge. These approaches combine the advantages of both types of data and use them in the following manners:

- Accurate NGS data are used to correct high error rates of third-generation data, and then long reads are used for assembly.
- Short NGS reads are used for assembly and resulting contigs are joined based on the long reads.

At present, the accuracy of third-generation sequencing reads is improving rapidly. Thus error correction by NGS reads will no longer be needed. Tools for repeat detection using solely long reads have already started to be developed.

4.2 *k*-mer-based approach

The main principle of the k-mer based approach was described in Section 3.3.2 as it has initially been used for the detection of repetitive sequences in assembled genomes. With the expansion of sequencing technologies, this approach was quickly adopted to detect repeats in NGS data.

4.2.1 Overview of tools for repeat detection from sequencing reads employing *k*-mer-based approach

Several computational tools for *de novo* repeat detection from sequence reads have employed k-mer frequency counting approach. One of the first tools that have been developed utilizing this idea was ReAS [71]. This tool recovers ancestral sequences of TEs from unassembled reads. The ancestral sequence is the TE when it was inserted into the genome for the first time. The algorithm works for TEs that satisfy two conditions: they occur in high copy numbers over the genome, and they are not too old (they did not accumulate too many mutations). Therefore, they are still identifiable compared to the ancestral sequences. In order to output meaningful results, it requires high genome coverage (2-fold coverage or higher). By the term coverage, we mean the average number of nucleotide bases that align and thus cover the reference sequences. For example, 2-fold genome coverage means that each base in the reference genome is spanned by data twice on average. The ReAS algorithm starts by selecting a high-depth k-mer. This k-mer has to meet several conditions: it can not be a simple repeat, k-mer depth must meet a specified threshold D, and it should not be in previously reconstructed TE. Afterward, it finds all reads containing this k-mer, and they are trimmed to 100 bp fragments with k-mer in the center. These fragments are aligned to each other, and the algorithm searches for groups containing at least D fragments with mutual identity 95% and assembles them into an initial consensus sequence. Next, the consensus sequence is extended based on the k-mers at the ends. This process is repeated until no further extensions are possible. In the end, misassemblies and ambiguities are resolved by using information from paired-end data. ReAS reports consensus sequences that are ancestral sequences of TEs.

Another tool utilizing k-mer-based approach is RepARK [62] which was designed to construct *de novo* repeat libraries from NGS data. In the beginning, the frequency of k-mers in input sequencing reads is counted, and only k-mers exceeding a given threshold are classified as abundant and isolated. These k-mers are further assembled by a genome assembly program (either CLC Assembly Cell [1] or Velvet [157]) into repeat consensus sequences. RepARK does not use information from paired-end reads to post-process output raw contigs, so they often represent only fragments of full repeats. To address this drawback, Chu et al. developed the tool REPdenovo [23], which provides *de novo* estimation of low-divergent and highly abundant repeats from sequence reads. It improves the RepARK assembly phase and uses a frequency-based assembly approach, in which all frequent k-mers are divided into groups with similarly binned frequencies and assembled separately into contigs. These contigs, called raw contigs, represent the final output of RepARK. Because they are usually only fragments of repeats, REPdenovo tries to merge them into longer contigs. Finally, the assembled contigs are filtered and verified by aligning reads back to the constructed repeat sequences. Wrongly assembled repeats that have no or low coverage are removed. If the coverage is uneven, low coverage regions are filtered out. The REPdenovo algorithm was improved, and the authors implemented two approaches to deal with highly divergent and low-copy-number repeats [24]. The two main improvements compared to original REPdenovo algorithm are *i)* finding more repeat-related k-mers, and *ii)* randomized algorithm is used to generate consensus k-mers, which are more reliable for assembly of highly divergent repeats.

In recent years, tools focusing on the detection of repeats from long reads started to emerge. Despite many advantages over NGS data, long reads produced by third-generation sequencing technologies suffer from high error rates, which can significantly impact repeat detection. *De novo* approach DLR (Detection of Long Repeats) [75] based on the long PacBio reads overcomes this problem by introducing a long read error correction step on the raw reads. These corrected reads are converted into unique k-mers. Those with a frequency above a specified threshold are aligned back to the corrected long reads by Bowtie2 [66] with a large proportion of error tolerance set. Regions that are covered by high-frequency k-mers are recorded. Repetitive regions with duplicate or inclusion relationship are merged, and final sequences of repeats are collected.

4.3 *de Bruijn* graph-based approach

Tools belonging to this category take advantage of *de Bruijn* graphs, a data structure frequently used in *de novo* sequence assemblers such as Trinity [40], ABySS [130], or Velvet [157]. In general, the *de Bruijn* graph is a directed multigraph that consists of a set of vertices and a multiset of directed edges. It is constructed using the unique k-mers that occur in the input data - sequencing reads. k-1-mers are added to the graph as nodes and k-mers as edges. During the graph construction, two branching structures are formed:

tips and bubbles. Tips often called „dead-ends“ are mostly caused by sequencing errors [130]. Bubbles represent false branches in the graph, which are a result of divergence due to single-point mutations. These branching structures can make the constructed graph very large, leading to the high memory usage of *de Bruijn* graph-based assemblers. To find the assembly, every edge in the graph is visited exactly once, representing an Eulerian path problem. This task is easy to solve even for large graphs containing millions of vertices since algorithms can resolve it in a linear time.

4.3.1 Overview of tools for repeat detection from sequencing reads employing *de Bruijn* graph-based approach

To the best of our knowledge, there is only tool explicitly designed for the assembly of TEs – Tedna [160]. It utilizes the idea of the most frequent k-mers combined with the *de Bruijn* graph assembly approach. Tedna uses Illumina paired-end reads as an input and outputs a list of consensus sequences of repeated elements. Each connected component in the graph, representing putative TE, is handled independently, and the frequency of each path in the graph is estimated by linear programming. Because the most frequent path typically corresponds to incomplete TE, the longest path is chosen to recover the full-length element. Tedna also eliminates duplicates and searches for big loops indicating LTRs in the *de Bruijn* graph, opens them, and adds LTRs at both ends of the transposon. Subsequently, Tedna tries to merge the sequences' ends into longer ones using k-mer comparison and Needleman-Wunsch algorithm [96]. The resulting merges generate a graph where nodes correspond to sequences, and edges indicate that sequences can be joined and the longest sequences are formed. The last step of Tedna's approach uses paired-end read information to scaffold the TEs.

dnaPipeTE [39] is a pipeline designed to assemble, annotate, and quantify repetitive sequences directly from NGS data. As an input, it requires at least three samples of the original genomic dataset. Two are used in the iterative runs of the Trinity assembler, and the third is used in the quantification step. dnaPipeTE performs uniform sampling of the input data to create low coverage data. The primary assumption is that only repetitive sequences will have sufficient representation to be assembled in such datasets. For assembly purposes, dnaPipeTE uses Trinity, a *de Bruijn* graph-based assembler, originally designed for transcriptome assembly, which can recover complete alternative consensus sequences of repeats. The resulting contigs are annotated by using RepeatMasker [131] and Repbase database and quantified.

A hybrid approach that combines short, high-quality reads with long reads for *de novo* detection of TRs was proposed MixTaR [34]. A set of short reads is used to build a *de Bruijn* graph. The graph is searched for cycles representing potential TR patterns. These patterns are validated using long reads. TRs are constructed from overlapping short reads by using local greedy assemblies.

4.4 Graph-based clustering approach

The first step in graph-based clustering is finding the sequence similarities between the input reads utilizing all-against-all pairwise alignment. This step can be run in a highly parallel manner to speed up the comparison of hundreds of thousands of sequencing reads. Further, only read pairs that meet specified similarity and overlap length thresholds are kept for further analysis. The second step consists of clustering, which utilizes information about sequencing reads similarities and constructs a graph in which the vertices represent sequence reads, overlapping reads are connected by edges, and an edge weight reflects the similarity between connected reads. The graph is scanned for connected components, groups of mutually connected vertices, which represent repetitive sequences.

Graph-based clustering approach can deal better with the variability of repetitive sequences compared to k-mer or *de Bruijn*-graph-based approaches as the desired similarity and overlap length thresholds can be set. Another significant advantage is that the input reads do not have to be split into k-mers; thus, the their continuity is not lost.

The main drawbacks of the graph-based clustering approach are *i*) chimeric clusters and *ii*) splitting one repeat family into multiple clusters. Both of these happen due to divergence of repeat and different conservation levels within the repeat families. If the similarity threshold is too benevolent, multiple repeats can be presented in one cluster, leading to the formation of a chimeric cluster. If the similarity threshold is too strict, then even a highly conserved repeat family will be divided into several clusters.

4.4.1 Overview of tools for repeat detection from sequencing reads employing graph-based clustering approach

One of the most commonly used tools for genome-wide characterization of repetitive elements utilizing the graph-based clustering approach, RepeatExplorer [99, 100], detects and quantifies the proportion of repeated sequences directly from NGS data.

Depending on the similarity threshold setting, resulting clusters may include various repetitive elements due to partial sequence similarity across them. Additional analysis of the graph is employed to detect communities [9], groups of nodes in the graph that are densely connected compared with the rest of the graph. Layouts of resulting clusters are calculated and characterized based on similarity hits to known repetitive sequences and protein domains. Graph layouts provide helpful information about the structural variants of the repeat, but some can shade in the graph. Finally, the sequence assembly of reads is performed for each cluster independently by CAP3 assembler [53].

RepeatExplorer provides a fundamental analysis of found repeats, and manual inspection of the output is often required. To extend this primary analysis, we developed a novel bioinformatics approach to analyze satDNA [110]. It includes identification of monomers and their length based on the distance between the same k-mers followed by clustering of the identified monomers to detect satDNA families. A more detailed description of the approach is available in Section 4.5.

It is necessary to point out that this work was published before TAREAN [98], a computational pipeline utilizing the graph-based clustering approach based on RepeatExplorer, focusing on the detection of satDNA from unassembled NGS data. TAREAN requires paired-end short-read data as an input. Graph-based clustering is followed by detecting circular structures in clustered graphs, a typical graph structure for TRs. The putative TR clusters are identified based on the proportion of broken paired-end reads, which should be low. Finally, the satellite monomers' consensus sequence is constructed from the most frequent k-mers in individual TR clusters.

Another tool utilizing graph-based clustering is Transposome [134]. It annotates TE families from unassembled sequence reads and estimates the genomic abundance of TE families. Transposome borrows the main idea from RepeatExplorer and tries to eliminate a few of its shortcomings, such as computational inefficiency and lack of modularity. The all-against-all sequence comparison is highly parallelized, and Transposome provides a programmatic interface for manipulating NGS data, allowing users to construct custom analysis pipelines. Unlike RepeatExplorer, this tool does not report consensus sequences of repeats.

Detection of repeats directly from long PacBio reads employing graph-based clustering approach was proposed in RepLong [44]. To identify mutual similarities and overlaps between every read pair, it uses MHAP [7], an efficient algorithm for similarity estimation between long reads. Reads with overlap longer than 100 bp are used for the graph construction, followed by community extraction using modularity optimization. Representative reads are extracted from each community to build a repeat library. RepLong showed that it could find longer repeats than tools using short reads and comparable results to genome-based tool RepeatModeler in shorter runtime [44].

4.5 Novel approach for detailed analysis of satDNA

The approach proposed by Puterova et al. was published in [110]. It extends the basic analysis of the RepeatExplorer pipeline and post-processes its results. The approach is composed of three steps:

1. *Detection of satellite monomers.* Contigs of selected clusters are extracted from RepeatExplorer output. For each contig, the monomer length is estimated from distances between the same *k*-mers in the contig. The monomer sequence is extracted from the most covered region of the contig.

2. *Estimation of satellite families composition and their annotation.* To estimate the composition of satellite families, the sequence similarity between monomers is determined by performing a semi-global alignment to compute a distance matrix, which is further processed by unweighted pair group method with arithmetic mean (UPGMA) clustering method [132]. The resulting dendrogram is cut to define the individual satellite families and visualized using igraph library[5]. Annotation of monomers is performed by querying them against nucleotide collection nt/nr and PlantSat database [80] using blastn [2]. To estimate the diversity within each satellite family, reads belonging to the family are mapped to the representative monomer using BWA-MEM aligner [69] and sequence logo is generated by WebLogo tool [28].

3. *Visualization of satellite families homogeneity.* Reads of each satellite family are merged and sampled randomly to decrease the computational demands for highly abundant families. Sequence similarity of these reads is estimated by all-against-all alignment performed by megablast [13]. Only pairs of reads that meet the set thresholds (70% sequence identity over at least 55% of sequence length) are used for graph construction and visualization. A relative abundance of male and female reads in each family is estimated. Such information can be useful to determine the chromosomal location, whether the satellite family is present on sex chromosomes or autosomes.

[5] https://igraph.org/r/

By applying the approach described above, we were able to identify 12 satellite families in the seabuckthorn (*Hippophae rhamnoides*) genome, including Y-specific, X-accumulated, and sex-chromosome-accumulated satellite families. The discovery of the Y-specific satellite helped to show that seabuckthorn has small Y and large X chromosomes since it was previously thought to be exactly the opposite [139]. For more details, the full text of the published article is available in Appendix A.

Chapter 5

Application of repeat detection in biological research

Genomes are continuously changing their structure and size over time. This process is called genome evolution, and several mechanisms contribute to it. In prokaryotes, two main mechanisms are shaping the genome: *i*) mutations (including transposition of ISEs), and *ii*) horizontal gene transfer (HGT) – a movement of genetic material between organisms (other than from parent to offspring). For example, antibiotic resistance genes are spread among bacteria by HGT. Various mechanisms contribute to the genome evolution of eukaryotes, which include: gene and genome duplications, mutations, accumulation/elimination of repetitive elements, exon shuffling, gene loss, or sexual reproduction.

Eukaryotic genomes are, in general, significantly bigger compared to prokaryotic genomes. For example, the genome of *E.coli* (prokaryote) is only ca. 4.6 Mb long [8] in comparison with the human genome (eukaryote), which is around 3.2 Gb in length. However, eukaryotic organisms display incredible variation in their GS. The plant genome sizes are ranging from as small as 60 Mb/1C[1] of carnivorous plant *Genlisea margaretae* [36] to the astonishingly large genome of *Paris japonica*, possessing the largest known genome so far, having roughly 150 Gb/1C [105] and being almost 50x larger than the human genome. In the animal kingdom, a similar phenomenon was observed where GS ranges from 20 Mb in the plant-pathogenic nematode *Pratylenchus coffeae* to almost 130 Gb in the marbled lungfish *Protopterus aethiopicus* [42]. The smallest eukaryotic genome ever reported belongs to the fungus *Encephalitozoon intestinalis*, whose genome is only around 2.3 Mb long [27].

From the information on genome sizes provided above, it is noticeable that genome size does not correlate with an organism's complexity as plant species can have a 50 times larger genome than humans. This discrepancy is known as the „C-value enigma“ [43] and refers to the observed variable proportion of non-coding DNA, including repetitive sequences, identified within the genomes of eukaryotic organisms while the number of genes remains quite stable [6]. However, there is relatively little knowledge about the molecular and evolutionary mechanisms contributing to genome size diversification.

Repetitive elements may comprise a considerable fraction of the genome and play a significant role in genome evolution (for more details, see Chapter 2). Therefore, a comparative analysis of the repeat content, nowadays known under the term

[1] The amount of DNA contained within a haploid nucleus.

„repeatome“ [86], of individuals from the same species, closely related species, or species with large genomes may elucidate the C-value enigma and contribute to a better understanding of the dynamics of genome size.

In the following sections, we will discuss the comparative repeatome analysis and its application in biological research, including our contribution in this field.

5.1 Comparative repeatome studies

The comparative analysis of repeats is still one of the greatest bioinformatics challenges. These studies can be performed, for example, on closely related species or within species to study intraspecific variation caused by repetitive elements. A suitable and the most desirable approach to perform comparative studies would be on assembled and annotated genomes. Such genomes are available mainly for the most important species, such as model organisms (human, mouse) or crop plants (e.g., maize, wheat, pea), which are essential from the agricultural point of view.

As we already know, genomes of eukaryotic organisms can reach the size of several dozens of Gb. However, performing the genome assembly and annotating such large genomes is computationally extensive and requires expensive analytical approaches. Furthermore, repetitive sequences are still a big challenge for *de novo* assemblers. Repeats can cause misarrangements or gaps in the assembly resulting in highly fragmented and low-quality assembly [74]. Another issue that can arise during assembly is that several copies of the same repeat can be collapsed into a single contig resulting in the distortion of the genome's repetitive content. Due to the reasons mentioned above, this approach is currently not applicable in practice. That leaves the researchers with the last and only option: to perform repeatome comparative analyses directly on NGS data without performing the whole genome's assembly.

As repeats may appear in thousands of copies within the genome, it is unnecessary to sequence the whole genome but only a small proportion of it, and we will still be able to capture highly and moderately abundant repeats. At the same time, it is unlikely that single-copy sequences will be present several times in the data. When the average sequencing depth of the genome is less than 1× coverage, this approach is called low-pass or low-coverage sequencing. In combination with the NGS data, which is becoming more and more available nowadays, it provides a remarkable opportunity to perform repeatome's comparative studies of dozens of either closely or distantly related species and study the genome's evolution in a reasonable time for an affordable price.

Several comparative repeatome studies combining low-pass genome sequencing with novel bioinformatics approaches focused on non-model species have been conducted in recent years. One of the first studies of this nature was performed on 23 plant species from *Fabaea* tribe to study different types of repeats and how they contribute to genome size evolution within a phylogenetic context [82]. This study provided a proof of concept that it is possible to use low-coverage sequencing representing a cost-effective approach with a combination of bioinformatics tools to perform comparative repeatome analysis in a large number of non-model species without prior knowledge about the studied genomes and the need for assembled reference genomes. Another study of repeat content of 52 fish species showed an association of specific repeat families with fish habitat suggesting the potential role of repetitive elements in fish adaptation to their living environments [156].

5.2 Sex chromosomes evolution

A special part of the genome are sex chromosomes, which undergo different evolutionary processes compared to the rest of the genome resulting in the accumulation/elimination of repetitive sequences and deviating from each other and the rest of the genome.

Humans, many animal and dozens of plant species possess sex chromosomes. These chromosomes determine the sex of an organism. Their origin and evolution have been a subject of interest of evolutionary biologists for a long time. There are multiple sex-determination systems; for example, humans have an XY sex-determination system: females have two copies of the same sex chromosome (XX), males have two distinct sex chromosomes (XY). Other sex-determination systems are the X0 system (females have two copies of the same sex chromosome, males have only one), ZW system, which is an opposite to the XY sex-determination system (females have two different sex chromosomes and males have two chromosomes of the same kind), or Z0 system (males have two copies of the same sex chromosome, whereas females have only one).

Recombination, a process during which genetic material is exchanged between multiple different chromosomes or between different regions within the same chromosome, is suppressed in sex chromosomes. Thus, other evolutionary forces influence their development compared to the rest of the genome. Sex chromosome evolution is rather cyclic than linear process and has various phases. In general, the main stages of the sex chromosomes evolution are [18][2]:

1. *Establishment of the sex-determining region.* Gain of a sex-determining gene or genes on a chromosome that was not formerly a sex chromosome.

2. *Evolution of suppressed recombination between sex chromosomes.* The sex-linked regions do not undergo chromosomal crossover, a process during which genetic material is exchanged between two homologous chromosomes. This non-recombining region may even extend beyond the region containing the sex-determining genes. Often, several small regions of sex chromosomes called pseudoautosomal regions still undergo recombination.

3. *Divergence between Y-linked and X-linked homologs.* In this stage, repetitive sequences accumulate, and differences between X and Y chromosomes develop, such as heteromorphism or low gene density.

4. *Degeneration of Y-linked genes.* Function of Y-linked genes may degenerate as the accumulation of transposable elements in their proximity may affect the expression of these genes.

5. *Evolution of dosage compensation process.* As the homogametic sex will have unequal gene expression compared to the heterogametic sex due to two copies of X-linked genes, a compensation process for equalizing the gene expression is established.

6. *Shrinkage phase.* Y-linked regions decay and lose functional genes due to deleterious mutations, which may lead to an overall loss of the sex chromosome.

[2] For simplicity, we use the XY sex-determination system for explanation purposes as the reader is the most familiar with it as this sex-determination system is present in humans.

The human sex chromosomes evolved roughly more than 200 million years ago (mya) and are considered old. During their evolution, the recombination restriction, gene loss, and deleterious mutations have resulted in morphological differences - the X chromosome is large (∼155 Mb) [118] and Y chromosome is three times smaller. In contrast to evolutionary old sex chromosomes in humans, most dioecious plants have evolutionary young sex chromosomes. For example, sex chromosomes of *Silene latifolia* (white campion), a well-established model organism to study the evolution of sex chromosomes, evolved only ca. 6 mya [64] and possess a large, evolutionary young Y chromosome.

To better understand the sex chromosomes evolution, mainly events in its early stages, their structure, dynamics and to bring new insights into previously unstudied species possessing sex chromosomes, several studies were conducted in recent years [56, 109, 110].

5.3 Our contribution to the field of comparative repeatome studies

Only around twenty plant species are dioecious and possess sex chromosomes. Most of the plant species carry evolutionary young sex chromosomes - large Y and small X chromosomes.

One of the little-studied plants having sex chromosomes is seabuckthorn (*Hippophae rhamnoides*). The first comprehensive study of its genome content was presented in Puterova et al. [110]. The study revealed several interesting findings by conducting a comparative repeatome analysis of the male and female genome. The seabuckthorn's genome contains a huge number of satellite repeats compared to most other plant genomes. Secondly, satDNA accumulated on the X chromosome (HRTR8), Y chromosome-specific (HRTR12), and both sex chromosomes-accumulated (HRTR2) satellites were identified. Identifying the Y-specific satellite repeat enabled to demonstrate that seabuckthorn possesses heteromorphic sex chromosomes with large X and small Y chromosomes. However, until then, it was thought that seabuckthorn has a large Y and small X chromosome as it is common in dioecious plant species. As a part of this study, a novel bioinformatics approach for comprehensive analysis of satDNA was developed and was described in Section 4.5. More details can be found in the published article available in Appendix A.

In another study, we analyzed the repeatome of several European ecotypes of *S.latifolia*, a well-established model for studying sex chromosomes evolution, focusing on a comparison of repeats composition and differences in genome dynamics among these ecotypes. The study showed that despite an intraspecific genome size variation, the Y chromosome has retained its size. This finding indicates that the expansion of the evolutionary young Y chromosome in *S.latifolia* has already reached its peak [109]. For more details, see Appendix B.

In the last study, we performed a sex chromosome-specific repeatome characterization for common sorrel (*Rumex acetosa*), a dioecious plant possessing XY_1Y_2 sex determination system [56]. In this study, the sex chromosomes were flow-sorted (separated from the rest of the genome), which allowed to sequence them in greater coverage and perform their repeatome characterization at a finer level. The analysis revealed several novel satellite repeats that contribute to the expansion of the sorrel's Y chromosomes. Although the repeat fraction was similar for the X and Y chromosomes, composition of repeat families varied considerably. Further details can be found in the full text of this study available in Appendix C.

Chapter 6

Research results summary

This chapter will summarize the research results that arose in connection with this dissertation and its central topic. The research results are composed of four original peer-reviewed publications published in impacted and international journals. We referenced these publications in previous chapters. For each publication, abstract, author's contribution, and other relevant information are included. The full texts of these publications are available in the appendices of this Thesis. Lastly, a list of other author's publications is provided at the end of this chapter.

6.1 Publication 1 - Novel approach of satDNA analysis

PUTEROVA, J., RAZUMOVA, O., MARTINEK, T., ALEXANDROV, O., DIVASHUK, M., KUBAT, Z., HOBZA, R., KARLOV, G., KEJNOVSKY, E. (2017). Satellite DNA and Transposable Elements in Seabuckthorn (*Hippophae rhamnoides*), a Dioecious Plant with Small Y and Large X Chromosomes. *Genome Biology and Evolution*, 9(1). 197-212. doi:10.1093/gbe/evw303

- Author's participation: 50%
- Journal impact factor (2020): 3.416 (Q2 - Genetics & Heredity, Evolutionary Biology)
- Number of citations as of 18.7.2021: 15 (WoS without self-citations)

Author's contribution

Designing and performing bioinformatics data analysis, data visualization, designing and implementing a novel bioinformatics approach for detailed analysis of satellite DNA, writing the part of the manuscript describing methods and the novel bioinformatics approach.

Abstract

Background: Seabuckthorn (*Hippophae rhamnoides*) is a dioecious shrub commonly used in the pharmaceutical, cosmetic, and environmental industry as a source of oil, minerals, and vitamins. The size of the seabuckthorn's genome is 2.55 Gbp/2C, but there is a dearth of information on its composition. In this study, we analyzed the transposable elements and satellites in its genome.

Methods: We carried out Illumina DNA sequencing and reconstructed the main repetitive DNA sequences. For repeat detection and characterization, we used RepeatExplorer. For a detailed analysis of tandem repeats, we developed a new bioinformatics approach that extends the basic analysis of the RepeatExplorer pipeline (described in section 4.5. For the determination of the chromosomal localization of transposons and satellites, fluorescence *in situ* hybridization (FISH) was used.

Results: The data showed that about 25% of the genome consists of satellite DNA and about 24% is formed of transposons, dominated by Ty3/*Gypsy* and Ty1/*Copia* LTR retrotransposons. FISH mapping revealed X chromosome-accumulated, Y chromosome-specific, or both sex chromosomes-accumulated satellites, but most satellites were found on autosomes. Transposons were located mostly in the subtelomeres of all chromosomes. The 5S rDNA and 45S rDNA were localized on one autosomal locus each.

Conclusion: We presented the first comprehensive analysis of the seabuckthorn (H. rhamnoides) genome. Although we demonstrated the small size of the Y chromosome of the seabuckthorn and accumulated satellite DNA there, we were unable to estimate the age and extent of the Y chromosome degeneration. Analysis of dioecious relatives such as Shepherdia would shed more light on the evolution of these sex chromosomes. The manuscript is available in Appendix A.

6.2 Publication 2 - Biological research

PUTEROVA, J.[*], KUBAT, Z.*, KEJNOVSKY, E., JESIONEK, W., CIZKOVA, J., VYSKOT, B., HOBZA, R. (2018). The slowdown of Y chromosome expansion in dioecious *Silene latifolia* due to DNA loss and male-specific silencing of retrotransposons. *BMC Genomics*, 19, 153. doi:10.1186/s12864-018-4547-7

- Author's participation: 40%
- Journal impact factor (2020): 3.969 (Q2 - Genetics & Heredity)
- Number of citations as of 18.7.2021: 8 (WoS without self-citations)

Author's contribution

Designing and performing bioinformatics data analysis, data visualization, writing the part of the manuscript describing methods and results.

Abstract

Background: The rise and fall of the Y chromosome was demonstrated in animals but plants often possess the large evolutionarily young Y chromosome that is thought has expanded recently. Break-even points dividing expansion and shrinkage phase of plant Y chromosome evolution are still to be determined. To assess the size dynamics of the Y chromosome, we studied intraspecific genome size variation and genome composition of male and female individuals in a dioecious plant *Silene latifolia*, a well-established model for sex-chromosomes evolution.

Results: Our genome size data are the first to demonstrate that regardless of intraspecific genome size variation, Y chromosome has retained its size in *S. latifolia*. Bioinformatics

[*] These authors contributed equally to this work.

study of genome composition showed that constancy of Y chromosome size was caused by Y chromosome DNA loss and the female-specific proliferation of recently active dominant retrotransposons. We show that several families of retrotransposons have contributed to genome size variation but not to Y chromosome size change.

Conclusions: Our results suggest that the large Y chromosome of *S. latifolia* has slowed down or stopped its expansion. Female-specific proliferation of retrotransposons, enlarging the genome with exception of the Y chromosome, was probably caused by silencing of highly active retrotransposons in males and represents an adaptive mechanism to suppress degenerative processes in the haploid stage. Sex specific silencing of transposons might be widespread in plants but hidden in traditional hermaphroditic model plants. The manuscript is available in Appendix B.

6.3 Publication 3 - Biological research

JESIONEK, W., BODLÁKOVÁ, M., KUBÁT, Z., ČEGAN, R., VYSKOT, B., VRÁNA, J., ŠAFÁŘ, J., **PUTEROVÁ, J.** , HOBZA, R. (2021). Fundamentally different repetitive element composition of sex chromosomes in *Rumex acetosa*. *Annals of Botany*, 127(1), 33–47. doi:10.1093/aob/mcaa160

- Author's participation: 10%
- Journal impact factor (2020): 4.357 (Q1 - Plant Sciences)
- Number of citations as of 18.7.2021: 2 (WoS without self-citations)

Author's contribution

Bioinformatics data analysis, writing part of the manuscript describing the methods.

Abstract

Background and aims: Dioecious species with well-established sex chromosomes are rare in the plant kingdom. Most sex chromosomes increase in size but no comprehensive analysis of the kind of sequences that drive this expansion has been presented. Here we analyse sex chromosome structure in common sorrel (Rumex acetosa), a dioecious plant with XY1Y2 sex determination, and we provide the first chromosome-specific repeatome analysis for a plant species possessing sex chromosomes.

Methods: We flow-sorted and separately sequenced sex chromosomes and autosomes in R. acetosa using the two-dimensional fluorescence in situ hybridization in suspension (FISHIS) method and Illumina sequencing. We identified and quantified individual repeats using RepeatExplorer, Tandem Repeat Finder and the Tandem Repeats Analysis Program. We employed fluorescence in situ hybridization (FISH) to analyse the chromosomal localization of satellites and transposons.

Key results: We identified a number of novel satellites, which have, in a fashion similar to previously known satellites, significantly expanded on the Y chromosome but not as much on the X or on autosomes. Additionally, the size increase of Y chromosomes is caused by non-long terminal repeat (LTR) and LTR retrotransposons, while only the latter contribute to the enlargement of the X chromosome. However, the X chromosome is populated by different LTR retrotransposon lineages than those on Y chromosomes.

Conclusions: The X and Y chromosomes have significantly diverged in terms of repeat composition. The lack of recombination probably contributed to the expansion of diverse satellites and microsatellites and faster fixation of newly inserted transposable elements (TEs) on the Y chromosomes. In addition, the X and Y chromosomes, despite similar total counts of TEs, differ significantly in the representation of individual TE lineages, which indicates that transposons proliferate preferentially in either the paternal or the maternal lineage. The manuscript is available in Appendix C.

6.4 Publication 4 - Novel approach for detection of ISEs and the digIS tool

PUTEROVÁ, J. and MARTÍNEK, T. (2021) digIS: towards detecting distant and putative novel insertion sequence elements in prokaryotic genomes *BMC Bioinformatics* 22, 258. doi:10.1186/s12859-021-04177-6

- Author's participation: 75%
- Journal impact factor (2020): 3.169 (Q2 - Mathematical & Computational Biology)
- Number of citations as of 18.7.2021: 0 (WoS without self-citations)

Author's contribution

Designing and implementing the software, performing evaluation experiments, collecting evaluation datasets, writing the manuscript.

Abstract

Background: The insertion sequence elements (IS elements) represent the smallest and the most abundant mobile elements in prokaryotic genomes. It has been shown that they play a significant role in genome organization and evolution. To better understand their function in the host genome, it is desirable to have an effective detection and annotation tool. This need becomes even more crucial when considering rapid-growing genomic and metagenomic data. The existing tools for IS elements detection and annotation are usually based on comparing sequence similarity with a database of known IS families. Thus, they have limited ability to discover distant and putative novel IS elements.

Results: In this paper, we present digIS, a software tool based on profile hidden Markov models assembled from catalytic domains of transposases. It shows a very good performance in detecting known IS elements when tested on datasets with manually curated annotation. The main contribution of digIS is in its ability to detect distant and putative novel IS elements while maintaining a moderate level of false positives. In this category it outperforms existing tools, especially when tested on large datasets of archaeal and bacterial genomes.

Conclusion: We provide digIS, a software tool using a novel approach based on manually curated profile hidden Markov models, which is able to detect distant and putative novel IS elements. Although digIS can find known IS elements as well, we expect it to be used primarily by scientists interested in finding novel IS elements. The tool is available at https://github.com/janka2012/digIS. The manuscript is available in Appendix D.

6.5 Other original research publications

TOKAN, V., **PUTEROVA, J.**, LEXA, M., KEJNOVSKY, E. (2018) Quadruplex DNA in long terminal repeats in maize LTR retrotransposons inhibits the expression of a reporter gene in yeast. *BMC Genomics*, 19, 184. doi:10.1186/s12864-018-4563-7

- Author's participation: 20% (data analysis, writing manuscript)
- Journal impact factor (2020): 3.969 (Q2 - Genetics & Heredity)
- Number of citations as of 18.7.2021: 6 (WoS without self-citations)

CHYRÁ, Z., ŠEVČÍKOVÁ, T., VOJTA, P., **PUTEROVÁ, J.**, BROŽOVÁ, L., GROWKOVÁ, K., FILIPOVÁ, J., ZÁTOPKOVÁ, M., GROSICKI, S., BARCHANICKA, A., JĘDRZEJCZAK, Wiesław W., WASZCZUK-GAJDA, A., JUNGOVÁ, A., MIKULÁŠOVÁ, A., HAJDÚCH, M., MOKREJŠ, M., POU,R L., ŠTORK, M., HARVANOVÁ, Ľ., MISTRÍK, M., MIKALA, G., ROBAK, P., CZYŻ, A., DĘBSKI, J., USNARSKA-ZUBKIEWICZ, L., JURCZYSZYN, A., STEJSKA,L L., MORGAN, G., KRYUKOV, F., BUDINSKÁ, E., ŠIMÍČEK, M., JELÍNEK, T., HRDINKA, M. HÁJEK, R. (2021) Heterogenous mutation spectrum and deregulated cellular pathways in aberrant plasma cells underline molecular pathology of light-chain amyloidosis. *Haematologica*, 106(2), 601-604. doi:10.3324/haematol.2019.239756

- Author's participation: 7.5% (data analysis, writing manuscript)
- Journal impact factor (2020): 9.941 (Q1 - Hematology)
- Number of citations as of 18.7.2021: 0 (WoS without self-citations)

Chapter 7

Conclusion

In summary, this Thesis dealt with the development and improvement of bioinformatics approaches and tools designed to detect repetitive sequences.

Repetitive sequences can compose a considerable portion of both eukaryotic and prokaryotic genomes. As discussed in Chapter 2, repetitive sequences have many roles in the genome. They can be responsible for structural rearrangements, gene formation, and regulation, or even might be involved in the evolution of genome or sex chromosomes. However, we are still far from fully understanding their function and behavior. Therefore, research in this area from both biological and computational point of view is necessary.

There are several main principles for the analysis and detection of repetitive elements depending on the availability of the input data on which we have focused in Chapter 3 and Chapter 4. The ever-increasing development of sequencing technologies improves the quality and availability of the data, which follows hand in hand in advancing computational tools aimed for repeat detection.

In this Thesis, we presented several research results. The first result, a novel approach for detecting ISEs implemented as the digIS tool, helped to extend the field of assembly-based methods for repeat detection. It utilizes pHMMs built for the most conserved region of ISEs – catalytic domain – together with additional filtering settings to eliminate FP hits. The approach was implemented in the form of the digIS tool and was evaluated against other tools aiming at ISEs detection.

The second presented result, a novel bioinformatics approach for the analysis of satDNA, contributed to the field of assembly-free methods for repeat detection. This approach extends the primary analysis of the RepeatExplorer pipeline. The extension consists of identifying monomers – their sequence and length – followed by their clustering resulting in the identification of satDNA families.

Finally, we presented how the detection of repetitive sequences can be applied in biological research through three studies focused on the detection and characterization of repetitive sequences in plant genomes. These studies dealt with repeats at various levels - chromosome, genome, and population level. In the first study, a comprehensive analysis of the entire seabuckthorn's genome utilizing a novel bioinformatics approach described in Section 4.5 was performed.

Repetitive content of dioecious plant *S.latifolia* was analyzed at the population level in the second study. Looking at the repeat content of individuals from multiple geographic locations suggests that the expansion of evolutionary young Y chromosome in *S.latifolia* has already reached the top.

The last study focused on analyzing and quantifying individual repeats in common sorrel on the autosomes and sex chromosomes and uncovered novel satellite repeats responsible for enlarging common sorrel's Y chromosome.

7.1 Future work

Research is a never-ending task. Many different ideas and extensions have not been incorporated or implemented due to the lack of time and complexity of topic of this Thesis. Therefore, future work may go in several directions. Here, we present several ideas for future work that were identified, mainly related to the novel approach for detection of ISEs:

- *Automatic build of pHMMs for the catalytic domain of TPase.* pHMMs used in digIS were built in a semi-automatic way and needed manual curation to identify the catalytic domain based on its secondary structure. With a potential improvement in detecting protein secondary structures, it might be possible to automate this task.
- *Detailed analysis of putative novel ISEs.* Comprehensive analysis of putative novel ISEs reported by digIS would be a great benefit, especially for biologists. We proposed a procedure for a thorough inspection of putative novel ISEs within Additional file 9 of digIS's manuscript. However, this procedure still requires manual inspection of MSAs or dot plots, and its full automation is welcomed.
- *Detection of other repeat types occurring in prokaryotic genomes.* digIS focuses mainly on the detection of ISEs. However, other types of TEs are present in prokaryotic genomes as well. Recently, a new database focusing on TEs in prokaryotic genomes, TnCentral [117], was posted. Sequences stored in this database can be used to build additional pHMMs that can be subsequently integrated into the digIS tool.

Bibliography

[1] CLC Genomics Workbench 20.0.
Retrieved from: https://digitalinsights.qiagen.com

[2] Altschul, S. F.; Gish, W.; Miller, W.; et al.: Basic local alignment search tool. *Journal of Molecular Biology*. vol. 215, no. 3. 1990: pp. 403–410. ISSN 00222836. doi:10.1016/S0022-2836(05)80360-2.

[3] Avvaru, A. K.; Sharma, D.; Verma, A.; et al.: MSDB: A comprehensive, annotated database of microsatellites. *Nucleic Acids Research*. vol. 48, no. D1. 2020: pp. D155–D159. ISSN 1362-4962. doi:10.1093/nar/gkz886.

[4] Bao, W.; Kojima, K. K.; Kohany, O.: Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*. vol. 6, no. 1. 2015: page 11. ISSN 1759-8753. doi:10.1186/s13100-015-0041-9.

[5] Bennett, G. M.; Moran, N. A.: Small, smaller, smallest: The origins and evolution of ancient dual symbioses in a phloem-feeding insect. *Genome Biology and Evolution*. vol. 5, no. 9. 2013: pp. 1675–1688. ISSN 17596653. doi:10.1093/gbe/evt118.

[6] Bennetzen, J. L.; Wang, H.: The Contributions of Transposable Elements to the Structure, Function, and Evolution of Plant Genomes. *Annual Review of Plant Biology*. vol. 65, no. 1. 2014: pp. 505–530. ISSN 1543-5008. doi:10.1146/annurev-arplant-050213-035811.

[7] Berlin, K.; Koren, S.; Chin, C.-S.; et al.: Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nature Biotechnology*. vol. 33, no. 6. 2015: pp. 623–630. ISSN 1546-1696. doi:10.1038/nbt.3238.

[8] Blattner, F. R.; Plunkett, G.; Bloch, C. A.; et al.: The complete genome sequence of *Escherichia coli* K-12. *Science*. vol. 277, no. 5331. 1997: pp. 1453–1462. ISSN 00368075. doi:10.1126/science.277.5331.1453.

[9] Blondel, V. D.; Guillaume, J.-L.; Lambiotte, R.; et al.: Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*. vol. 2008, no. 10. 2008: page P10008.

[10] Bowen, N. J.; Jordan, I. K.: Transposable elements and the evolution of eukaryotic complexity. *Current issues in molecular biology*. vol. 4, no. 3. 2002: pp. 65–76. ISSN 1467-3037.

[11] Bureau, T. E.; Wessler, S. R.: Tourist: a large family of small inverted repeat elements frequently associated with maize genes. *The Plant Cell*. vol. 4, no. 10. 1992: pp. 1283–1294. ISSN 1040-4651. doi:10.1105/tpc.4.10.1283.

[12] Bureau, T. E.; Wessler, S. R.: Stowaway: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. *The Plant Cell*. vol. 6, no. 6. 1994: pp. 907–916. ISSN 1040-4651. doi:10.1105/tpc.6.6.907.

[13] Camacho, C.; Coulouris, G.; Avagyan, V.; et al.: BLAST+: architecture and applications. *BMC Bioinformatics*. vol. 10, no. 1. 2009: page 421. ISSN 1471-2105. doi:10.1186/1471-2105-10-421.

[14] Chaconas, G.; Chen, C. W.: *Replication of Linear Bacterial Chromosomes: No Longer Going Around in Circles*. chapter 29. John Wiley & Sons, Ltd. 2004. ISBN 9781683672043. pp. 525–539. doi:10.1128/9781555817640.ch29.

[15] Chandler, M.: *Transposons: Prokaryotic*. 2016. ISBN 9780470015902. pp. 1–9. doi:10.1002/9780470015902.a0000591.pub2.

[16] Chaparro, C.; Guyot, R.; Zuccolo, A.; et al.: RetrOryza: A database of the rice LTR-retrotransposons. *Nucleic Acids Research*. vol. 35, no. SUPPL. 1. 2007: pp. 66–70. ISSN 03051048. doi:10.1093/nar/gkl780.

[17] Charlesworth, D.: Plant sex chromosome evolution. *Journal of Experimental Botany*. vol. 64, no. 2. 2012: pp. 405–420. ISSN 0022-0957. doi:10.1093/jxb/ers322.

[18] Charlesworth, D.: Young sex chromosomes in plants and animals. *New Phytologist*. vol. 224, no. 3. 2019: pp. 1095–1107. doi:10.1111/nph.16002.

[19] Chen, J.; Hu, Q.; Zhang, Y.; et al.: P-MITE: a database for plant miniature inverted-repeat transposable elements. *Nucleic Acids Research*. vol. 42, no. D1. 10 2013: pp. D1176–D1181. ISSN 0305-1048. doi:10.1093/nar/gkt1000.

[20] Chen, J.; Lu, C.; Zhang, Y.; et al.: Miniature inverted-repeat transposable elements (MITEs) in rice were originated and amplified predominantly after the divergence of Oryza and Brachypodium and contributed considerable diversity to the species. *Mobile Genetic Elements*. vol. 2, no. 3. 2012: pp. 127–132. ISSN 2159-256X. doi:10.4161/mge.20773.

[21] Chénais, B.; Caruso, A.; Hiard, S.; et al.: The impact of transposable elements on eukaryotic genomes: From genome size increase to genetic adaptation to stressful environments. *Gene*. vol. 509, no. 1. 2012: pp. 7–15. ISSN 0378-1119. doi:10.1016/j.gene.2012.07.042.

[22] Cho, J.; Paszkowski, J.: Regulation of rice root development by a retrotransposon acting as a microRNA sponge. *eLife*. vol. 6. 2017: page e30038. ISSN 2050-084X. doi:10.7554/eLife.30038.

[23] Chu, C.; Nielsen, R.; Wu, Y.: REPdenovo: Inferring *De Novo* repeat motifs from short sequence reads. *PLoS ONE*. vol. 11, no. 3. 2016: pp. 1–17. ISSN 19326203. doi:10.1371/journal.pone.0150719.

[24] Chu, C.; Pei, J.; Wu, Y.: An improved approach for reconstructing consensus repeats from short sequence reads. *BMC Genomics*. vol. 19. 2018. ISSN 1471-2164. doi:10.1186/s12864-018-4920-6.

[25] Cock, P. J. A.; Antao, T.; Chang, J. T.; et al.: Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics.* vol. 25, no. 11. 2009: pp. 1422–1423. ISSN 1460-2059. doi:10.1093/bioinformatics/btp163.

[26] Copetti, D.; Zhang, J.; Baidouri, M. E.; et al.: RiTE database: a resource database for genus-wide rice genomics and evolutionary biology. *BMC Genomics.* vol. 16, no. 1. 2015: page 538. ISSN 1471-2164. doi:10.1186/s12864-015-1762-3.

[27] Corradi, N.; Pombert, J.-F.; Farinelli, L.; et al.: The complete sequence of the smallest known nuclear genome from the microsporidian *Encephalitozoon intestinalis*. *Nature Communications.* vol. 1, no. 1. 2010: pp. 1–7. ISSN 2041-1723. doi:10.1038/ncomms1082.

[28] Crooks, G. E.; Hon, G.; Chandonia, J.-M.; et al.: WebLogo: a sequence logo generator. *Genome Research.* vol. 14, no. 6. 2004: pp. 1188–1190. ISSN 1088-9051. doi:10.1101/gr.849004.

[29] Dohm, J. C.; Lottaz, C.; Borodina, T.; et al.: Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research.* vol. 36, no. 16. 2008: page e105. ISSN 03051048. doi:10.1093/nar/gkn425.

[30] Du, J.; Grant, D.; Tian, Z.; et al.: SoyTEdb: a comprehensive database of transposable elements in the soybean genome. *BMC Genomics.* vol. 11:113. 2010. ISSN 1471-2164. doi:10.1186/1471-2164-11-113.

[31] Duitama, J.; Zablotskaya, A.; Gemayel, R.; et al.: Large-scale analysis of tandem repeat variability in the human genome. *Nucleic Acids Research.* vol. 42, no. 9. 2014: pp. 5728–5741. ISSN 1362-4962. doi:10.1093/nar/gku212.

[32] Eddy, S. R.: Profile hidden Markov models. *Bioinformatics (Oxford, England).* vol. 14, no. 9. 1998: pp. 755–763. ISSN 1367-4803. doi:10.1093/bioinformatics/14.9.755.

[33] Eddy, S. R.: Accelerated Profile HMM Searches. *PLoS Computational Biology.* vol. 7, no. 10. 2011: page e1002195. ISSN 1553-7358. doi:10.1371/journal.pcbi.1002195.

[34] Fertin, G.; Jean, G.; Radulescu, A.; et al.: Hybrid de novo tandem repeat detection using short and long reads. *BMC Medical Genomics.* vol. 8, no. 3. 2015: pp. 1–19. ISSN 17558794. doi:10.1186/1755-8794-8-S3-S5.

[35] Feschotte, C.; Pritham, E. J.: DNA transposons and the evolution of eukaryotic genomes. *Annual Review of Genetics.* vol. 41, no. 1. 2007: pp. 331–368. ISSN 0066-4197. doi:10.1146/annurev.genet.40.110405.090448.

[36] Fleischmann, A.; Michael, T. P.; Rivadavia, F.; et al.: Evolution of genome size and chromosome number in the carnivorous plant genus Genlisea (Lentibulariaceae), with a new estimate of the minimum genome size in angiosperms. *Annals of botany.* vol. 114, no. 8. 2014: pp. 1651–1663. ISSN 1095-8290. doi:10.1093/aob/mcu189.

[37] Gelfand, Y.; Rodriguez, A.; Benson, G.: TRDB—the tandem repeats database. *Nucleic acids research.* vol. 35, no. suppl_1. 2007: pp. D80–D87. ISSN 03051048. doi:10.1093/nar/gkl1013.

[38] Goerner-Potvin, P.; Bourque, G.: Computational tools to unmask transposable elements. *Nature Reviews Genetics.* vol. 19, no. 11. 2018: pp. 688–704. ISSN 14710064. doi:10.1038/s41576-018-0050-x.

[39] Goubert, C.; Modolo, L.; Vieira, C.; et al.: De Novo Assembly and Annotation of the Asian Tiger Mosquito (*Aedes albopictus*) Repeatome with dnaPipeTE from Raw Genomic Reads and Comparative Analysis with the Yellow Fever Mosquito (*Aedes aegypti*). *Genome Biology and Evolution.* vol. 7, no. 4. 2015: pp. 1192–1205. ISSN 1759-6653. doi:10.1093/gbe/evv050.

[40] Grabherr, M. G.; Haas, B. J.; Yassour, M.; et al.: Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology.* vol. 29, no. 7. 2011: pp. 644–652. ISSN 1546-1696. doi:10.1038/nbt.1883.

[41] Gray, Y. H.: It takes two transposons to tango:transposable-element-mediated chromosomal rearrangements. *Trends in Genetics.* vol. 16, no. 10. 2000: pp. 461–468. ISSN 0168-9525. doi:10.1016/S0168-9525(00)02104-1.

[42] Gregory, T.: Animal Genome Size Database. 2021. accessed: 2021-05-16. Retrieved from: http://www.genomesize.com/

[43] Gregory, T. R.: Coincidence, coevolution, or causation? DNA content, cellsize, and the C-value enigma. *Biological Reviews.* vol. 76, no. 1. 2001: pp. 65–101. ISSN 1469185X. doi:10.1111/j.1469-185X.2000.tb00059.x.

[44] Guo, R.; Li, Y. R.; He, S.; et al.: RepLong: *de novo* repeat identification using long read sequencing data. *Bioinformatics.* vol. 34, no. 7. 2018: pp. 1099–1107. ISSN 1460-2059. doi:10.1093/bioinformatics/btx717.

[45] Han, M. J.; Shen, Y. H.; Gao, Y. H.; et al.: Burst expansion, distribution and diversification of MITEs in the silkworm genome. *BMC Genomics.* vol. 11, no. 1. 2010: page 520. ISSN 1471-2164. doi:10.1186/1471-2164-11-520.

[46] Han, M. J.; Zhou, Q. Z.; Zhang, H. H.; et al.: IMITEdb: The genome-wide landscape of miniature inverted-repeat transposable elements in insects. *Database.* vol. 2016. 2016: page baw148. ISSN 1758-0463. doi:10.1093/database/baw148.

[47] Harkess, A.; Mercati, F.; Abbate, L.; et al.: Retrotransposon Proliferation Coincident with the Evolution of Dioecy in Asparagus. *G3: Genes,Genomes,Genetics.* vol. 6, no. 9. 2016: pp. 2679–2685. ISSN 2160-1836. doi:10.1534/g3.116.030239.

[48] Hayashi, K.; Morooka, N.; Yamamoto, Y.; et al.: Highly accurate genome sequences of *Escherichia coli* K-12 strains MG1655 and W3110. *Molecular Systems Biology.* vol. 2. 2006: page 2006.0007. ISSN 1744-4292. doi:10.1038/msb4100049.

[49] Hickman, A. B.; Chandler, M.; Dyda, F.: Integrating prokaryotes and eukaryotes: DNA transposases in light of structure. *Critical Reviews in Biochemistry and*

Molecular Biology. vol. 45, no. 1. 2010: pp. 50–69. ISSN 1040-9238. doi:10.3109/10409230903505596.

[50] Hickman, A. B.; James, J. A.; Barabas, O.; et al.: DNA recognition and the precleavage state during single-stranded DNA transposition in *D. radiodurans*. *The EMBO Journal*. vol. 29, no. 22. 2010: pp. 3840–3852. ISSN 0261-4189. doi:10.1038/emboj.2010.241.

[51] Hobza, R.; Cegan, R.; Jesionek, W.; et al.: Impact of Repetitive Elements on the Y Chromosome Formation in Plants. *Genes*. vol. 8, no. 11. 2017: page 302. ISSN 2073-4425. doi:10.3390/genes8110302.

[52] Hobza, R.; Lengerova, M.; Svoboda, J.; et al.: An accumulation of tandem DNA repeats on the Y chromosome in *Silene latifolia* during early stages of sex chromosome evolution. *Chromosoma*. vol. 115, no. 5. 2006: pp. 376–382. ISSN 0009-5915. doi:10.1007/s00412-006-0065-5.

[53] Huang, X.; Madan, A.: CAP3: A DNA sequence assembly program. *Genome Research*. vol. 9, no. 9. 1999: pp. 868–877. ISSN 1088-9051. doi:10.1101/gr.9.9.868.

[54] Hubley, R.; Finn, R. D.; Clements, J.; et al.: The Dfam database of repetitive DNA families. *Nucleic Acids Research*. vol. 44, no. D1. 11 2015: pp. D81–D89. ISSN 0305-1048. doi:10.1093/nar/gkv1272.

[55] Jain, M.; Koren, S.; Miga, K. H.; et al.: Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology*. vol. 36, no. 4. 2018: pp. 338–345. ISSN 1546-1696. doi:10.1038/nbt.4060.

[56] Jesionek, W.; Bodláková, M.; Kubát, Z.; et al.: Fundamentally different repetitive element composition of sex chromosomes in *Rumex acetosa*. *Annals of Botany*. vol. 127, no. 1. 2021: pp. 33–47. ISSN 0305-7364. doi:10.1093/aob/mcaa160.

[57] Kamath, G. M.; Shomorony, I.; Xia, F.; et al.: HINGE: Long-read assembly achieves optimal repeat resolution. *Genome Research*. vol. 27, no. 5. 2017: pp. 747–756. ISSN 15495469. doi:10.1101/gr.216465.116.

[58] Kamoun, C.; Payen, T.; Hua-Van, A.; et al.: Improving prokaryotic transposable elements identification using a combination of *de novo* and profile HMM methods. *BMC Genomics*. vol. 14, no. 1. 2013: page 700. ISSN 1471-2164. doi:10.1186/1471-2164-14-700.

[59] Kazazian, H. H.; Wong, C.; Youssoufian, H.; et al.: Haemophilia A resulting from *de novo* insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature*. vol. 332, no. 6160. 1988: pp. 164–166. ISSN 1476-4687. doi:10.1038/332164a0.

[60] Kejnovsky, E.; Kubat, Z.; Macas, J.; et al.: *Retand*: a novel family of gypsy-like retrotransposons harboring an amplified tandem repeat. *Molecular Genetics and Genomics*. vol. 276, no. 3. 2006: pp. 254–263. doi:10.1007/s00438-006-0140-x.

[61] Kichenaradja, P.; Siguier, P.; Pérochon, J.; et al.: ISbrowser: an extension of ISfinder for visualizing insertion sequences in prokaryotic genomes. *Nucleic Acids*

Research. vol. 38, no. suppl_1. 2010: pp. D62–D68. ISSN 0305-1048. doi:10.1093/nar/gkp947.

[62] Koch, P.; Platzer, M.; Downie, B. R.: RepARK—*de novo* creation of repeat libraries from whole-genome NGS reads. *Nucleic Acids Research*. vol. 42, no. 9. 2014: page e80. ISSN 0305-1048. doi:10.1093/nar/gku210.

[63] de Koning, A. P.; Gu, W.; Castoe, T. A.; et al.: Repetitive elements may comprise over Two-Thirds of the human genome. *PLoS Genetics*. vol. 7, no. 12. 2011: page e1002384. ISSN 15537390. doi:10.1371/journal.pgen.1002384.

[64] Kubat, Z.; Zluvova, J.; Vogel, I.; et al.: Possible mechanisms responsible for absence of a retrotransposon family on a plant Y chromosome. *New Phytologist*. vol. 202, no. 2. 2014: pp. 662–678. doi:10.1111/nph.12669.

[65] Lander, E. S.; Linton, L. M.; Birren, B.; et al.: Initial sequencing and analysis of the human genome. *Nature*. vol. 409, no. 6822. 2001: pp. 860–921. ISSN 0028-0836. doi:10.1038/35057062. 11237011.

[66] Langmead, B.; Salzberg, S. L.: Fast gapped-read alignment with Bowtie 2. *Nature Methods*. vol. 9, no. 4. 2012: pp. 357–359. ISSN 1548-7091. doi:10.1038/nmeth.1923.

[67] Laver, T.; Harrison, J.; O'Neill, P. A.; et al.: Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomolecular Detection and Quantification*. vol. 3. 2015: pp. 1–8. ISSN 2214-7535. doi:10.1016/j.bdq.2015.02.001.

[68] Lerat, E.: Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. *Heredity*. vol. 104, no. 6. 2010: pp. 520–33. ISSN 1365-2540. doi:10.1038/hdy.2009.165.

[69] Li, H.: Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*. 2013.

[70] Li, Q.; Chang, W.; Zhang, H.; et al.: The Role of Plasmids in the Multiple Antibiotic Resistance Transfer in ESBLs-Producing Escherichia coli Isolated From Wastewater Treatment Plants. *Frontiers in Microbiology*. vol. 10. 2019: page 633. ISSN 1664-302X. doi:10.3389/fmicb.2019.00633.

[71] Li, R.; Ye, J.; Li, S.; et al.: ReAS: Recovery of ancestral sequences for transposable elements from the unassembled reads of a whole genome shotgun. *PLoS Computational Biology*. vol. 1, no. 4. 2005: page e43. ISSN 1553-7358. doi:10.1371/journal.pcbi.0010043.

[72] Li, S. F.; Su, T.; Cheng, G. Q.; et al.: Chromosome evolution in connection with repetitive sequences and epigenetics in plants. *Genes*. vol. 8, no. 10. 2017: page 290. ISSN 2073-4425. doi:10.3390/genes8100290.

[73] Li, S.-F.; Zhang, G.-J.; Zhang, X.-J.; et al.: DPTEdb, an integrative database of transposable elements in dioecious plants. *Database*. vol. 2016. 2016: page baw078. ISSN 1758-0463. doi:10.1093/database/baw078.

[74] Liao, X.; Li, M.; Zou, Y.; et al.: Current challenges and solutions of *de novo* assembly. *Quantitative Biology*. vol. 7, no. 2. 2019: pp. 90–109. ISSN 20954697. doi:10.1007/s40484-019-0166-9.

[75] Liao, X.; Zhang, X.; Wu, F. X.; et al.: De novo repeat detection based on the third generation sequencing reads. *Proceedings - 2019 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2019*. 2019: pp. 431–436. doi:10.1109/BIBM47256.2019.8982959.

[76] Liu, L.; Li, Y.; Li, S.; et al.: Comparison of next-generation sequencing systems. *Journal of Biomedicine and Biotechnology*. vol. 2012. 2012: page 251364. ISSN 1110-7251. doi:10.1155/2012/251364.

[77] Llorens, C.; Futami, R.; Covelli, L.; et al.: The Gypsy Database (GyDB) of Mobile Genetic Elements: Release 2.0. *Nucleic Acids Research*. vol. 39, no. suppl_1. 2011: pp. D70–D74. ISSN 0305-1048. doi:10.1093/nar/gkq1061.

[78] López-Flores, I.; Garrido-Ramos, M. A.: The repetitive DNA content of eukaryotic genomes. *Repetitive DNA*. vol. 7. 2012: pp. 1–28. ISSN 1660-9263. doi:10.1159/000337118.

[79] Ma, B.; Li, T.; Xiang, Z.; et al.: MnTEdb, a collective resource for mulberry transposable elements. *Database*. vol. 2015. 2015. ISSN 1758-0463. doi:10.1093/database/bav004.

[80] Macas, J.; Mészáros, T.; Nouzová, M.: PlantSat: A specialized database for plant satellite repeats. *Bioinformatics*. vol. 18, no. 1. 2002: pp. 28–35. ISSN 1460-2059. doi:10.1093/bioinformatics/18.1.28.

[81] Macas, J.; Neumann, P.: Ogre elements — A distinct group of plant Ty3/gypsy-like retrotransposons. *Gene*. vol. 390, no. 1-2. 2007: pp. 108–116. ISSN 0378-1119. doi:10.1016/J.GENE.2006.08.007.

[82] Macas, J.; Novák, P.; Pellicer, J.; et al.: In Depth Characterization of Repetitive DNA in 23 Plant Genomes Reveals Sources of Genome Size Variation in the Legume Tribe Fabeae. *PLoS ONE*. vol. 10, no. 11. 2015: page e0143424. ISSN 1932-6203. doi:10.1371/journal.pone.0143424.

[83] Mahillon, J.; Chandler, M.: Insertion sequences. *Microbiology and Molecular Biology Reviews*. vol. 62, no. 3. 1998: pp. 725–774.

[84] Mariotti, B.; Manzano, S.; Kejnovský, E.; et al.: Accumulation of Y-specific satellite DNAs during the evolution of Rumex acetosa sex chromosomes. *Molecular Genetics and Genomics*. vol. 281, no. 3. 2009: pp. 249–259. ISSN 1617-4615. doi:10.1007/s00438-008-0405-7.

[85] Martin, A.; Troadec, C.; Boualem, A.; et al.: A transposon-induced epigenetic change leads to sex determination in melon. *Nature*. vol. 461, no. 7267. 2009: pp. 1135–1138. ISSN 1476-4687. doi:10.1038/nature08498.

[86] Maumus, F.; Quesneville, H.: Deep Investigation of Arabidopsis thaliana Junk DNA Reveals a Continuum between Repetitive Elements and Genomic Dark Matter.

PLoS ONE. vol. 9, no. 4. 2014: page e94101. ISSN 1932-6203. doi:10.1371/journal.pone.0094101.

[87] McClintock, B.: Mutable loci in maize. *Carnegie Institution of Washington Year Book*. vol. 47. 1947.

[88] McClintock, B.: The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences of the United States of America*. vol. 36, no. 6. 1950: pp. 344–355. ISSN 0027-8424. doi:10.1073/pnas.36.6.344.

[89] Mehrotra, S.; Goyal, V.: Repetitive Sequences in Plant Nuclear DNA: Types, Distribution, Evolution and Function. *Genomics, Proteomics and Bioinformatics*. vol. 12, no. 4. 2014: pp. 164–171. ISSN 1672-0229. doi:10.1016/j.gpb.2014.07.003.

[90] Michael, T. P.: Plant genome size variation: Bloating and purging DNA. *Briefings in Functional Genomics and Proteomics*. vol. 13, no. 4. 2014: pp. 308–317. ISSN 1477-4062. doi:10.1093/bfgp/elu005.

[91] Miga, K. H.: Completing the human genome: the progress and challenge of satellite DNA assembly. *Chromosome Research*. vol. 23, no. 3. 2015: pp. 421–426. ISSN 1573-6849. doi:10.1007/s10577-015-9488-2.

[92] Mira, A.; Ochman, H.; Moran, N. A.: Deletional bias and the evolution of bacterial genomes. *Trends in Genetics*. vol. 17, no. 10. 2001: pp. 589–596. ISSN 0168-9525. doi:10.1016/S0168-9525(01)02447-7.

[93] Munoz-Lopez, M.; Garcia-Perez, J. L.: DNA Transposons: Nature and Applications in Genomics. *Current Genomics*. vol. 11, no. 2. 2010: pp. 115–128. doi:10.2174/138920210790886871.

[94] Mustajoki, S.; Ahola, H.; Mustajoki, P.; et al.: Insertion of Alu element responsible for acute intermittent porphyria. *Human mutation*. vol. 13, no. 6. 1999: pp. 431–438. ISSN 1059-7794. doi:10.1002/(SICI)1098-1004(1999)13:6<431::AID-HUMU2>3.0.CO;2-Y.

[95] Naito, K.; Zhang, F.; Tsukiyama, T.; et al.: Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. *Nature*. vol. 461, no. 7267. 2009: pp. 1130–1134. ISSN 0028-0836. doi:10.1038/nature08479.

[96] Needleman, S. B.; Wunsch, C. D.: A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*. vol. 48, no. 3. 1970: pp. 443–453. ISSN 0022-2836. doi:10.1016/0022-2836(70)90057-4.

[97] Neumann, P.; Novák, P.; Hoštáková, N.; et al.: Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. *Mobile DNA*. vol. 10, no. 1. 2019: pp. 1–17. ISSN 1759-8753. doi:10.1186/s13100-018-0144-1.

[98] Novák, P.; Ávila Robledillo, L.; Koblížková, A.; et al.: TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads. *Nucleic Acids Research*. vol. 45, no. 12. 2017: page e111. ISSN 0305-1048. doi:10.1093/nar/gkx257.

[99] Novák, P.; Neumann, P.; Macas, J.: Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinformatics*. vol. 11. 2010: page 378. ISSN 1471-2105. doi:10.1186/1471-2105-11-378.

[100] Novák, P.; Neumann, P.; Pech, J.; et al.: RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics*. vol. 29, no. 6. 2013: pp. 792–793. ISSN 1460-2059. doi:10.1093/bioinformatics/btt054.

[101] Orozco-Arias, S.; Isaza, G.; Guyot, R.: Retrotransposons in Plant Genomes: Structure, Identification, and Classification through Bioinformatics and Machine Learning. *International Journal of Molecular Sciences*. vol. 20, no. 15. 2019: page 3837. ISSN 1422-0067. doi:10.3390/ijms20153837.

[102] Ouyang, S.; Buell, C. R.: The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants. *Nucleic Acids Research*. vol. 32, no. suppl_1. 2004: pp. D360–D363. ISSN 1362-4962. doi:10.1093/nar/gkh099.

[103] Oxford Nanopore Technologies webpage. https://nanoporetech.com. 2021. accessed: 18.05.2021.

[104] Payne, A.; Holmes, N.; Rakyan, V.; et al.: BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files. *Bioinformatics*. vol. 35, no. 13. 2019: pp. 2193–2198. doi:10.1093/bioinformatics/bty841.

[105] Pellicer, J.; Fay, M. F.; Leitch, I. J.: The largest eukaryotic genome of them all? *Botanical Journal of the Linnean Society*. vol. 164, no. 1. 2010: pp. 10–15. ISSN 0024-4074. doi:10.1111/j.1095-8339.2010.01072.x.

[106] Pezer, Ž.; Brajković, J.; Feliciello, I.; et al.: Satellite DNA-Mediated Effects on Genome Regulation. In *Genome Dynamics*, vol. 7. Karger. 2012. ISBN 1660-9263. pp. 153–169. doi:10.1159/000337116.

[107] Plohl, M.; Meštrović, N.; Mravinac, B.: Satellite DNA Evolution. In *Genome Dynamics*, vol. 7. Karger. 2012. ISBN 1660-9263. pp. 126–152. doi:10.1159/000337122.

[108] Pons, J.; Bruvo, B.; Juan, C.; et al.: Conservation of satellite DNA in species of the genus *Pimelia* (Tenebrionidae, Coleoptera). *Gene*. vol. 205, no. 1-2. 1997: pp. 183–190. ISSN 0378-1119. doi:10.1016/S0378-1119(97)00402-2.

[109] Puterova, J.; Kubat, Z.; Kejnovsky, E.; et al.: The slowdown of Y chromosome expansion in dioecious *Silene latifolia* due to DNA loss and male-specific silencing of retrotransposons. *BMC Genomics*. vol. 19, no. 1. 2018. ISSN 1471-2164. doi:10.1186/s12864-018-4547-7.

[110] Puterova, J.; Razumova, O.; Martinek, T.; et al.: Satellite DNA and Transposable Elements in Seabuckthorn (*Hippophae rhamnoides*), a Dioecious Plant with Small Y and Large X Chromosomes. *Genome Biology and Evolution*. vol. 9, no. 1. 2017: page evw303. ISSN 1759-6653. doi:10.1093/GBE/EVW303.

[111] Rang, F. J.; Kloosterman, W. P.; de Ridder, J.: From squiggle to basepair: Computational approaches for improving nanopore sequencing read accuracy. *Genome Biology*. vol. 19, no. 1. 2018: pp. 1–11. ISSN 1474-760X. doi:10.1186/s13059-018-1462-9.

[112] Reuter, J. A.; Spacek, D. V.; Snyder, M. P.: High-Throughput Sequencing Technologies. *Molecular Cell*. vol. 58, no. 4. 2015: pp. 586–597. ISSN 1097-4164. doi:10.1016/j.molcel.2015.05.004.

[113] Rhoads, A.; Au, K. F.: PacBio Sequencing and Its Applications. *Genomics, Proteomics and Bioinformatics*. vol. 13, no. 5. 2015: pp. 278–289. ISSN 2210-3244. doi:10.1016/j.gpb.2015.08.002.

[114] Riadi, G.; Medina-Moenne, C.; Holmes, D. S.: TnpPred: A Web Service for the Robust Prediction of Prokaryotic Transposases. *Comparative and Functional Genomics*. vol. 2012. 2012: page 678761. ISSN 1532-6268. doi:10.1155/2012/678761.

[115] Rice, P. A.; Baker, T. A.: Comparative architecture of transposase and integrase complexes. *Nature Structural Biology*. vol. 8, no. 4. 2001: pp. 302–307. ISSN 1072-8368. doi:10.1038/86166.

[116] Robinson, D. G.; Lee, M.-C.; Marx, C. J.: OASIS: an automated program for global investigation of bacterial and archaeal insertion sequences. *Nucleic Acids Research*. vol. 40, no. 22. 2012: page e174. ISSN 1362-4962. doi:10.1093/nar/gks778.

[117] Ross, K.; Varani, A. M.; Snesrud, E.; et al.: TnCentral: A Prokaryotic Transposable Element Database and Web Portal for Transposon Analysis. *bioRxiv*. 2021. doi:10.1101/2021.05.26.445724.
Retrieved from:
https://www.biorxiv.org/content/early/2021/05/26/2021.05.26.445724

[118] Ross, M. T.; Grafham, D. V.; Coffey, A. J.; et al.: The DNA sequence of the human X chromosome. *Nature*. vol. 434, no. 7031. 2005: pp. 325–337. ISSN 0028-0836. doi:10.1038/nature03440.

[119] Rothberg, J. M.; Hinz, W.; Rearick, T. M.; et al.: An integrated semiconductor device enabling non-optical genome sequencing. *Nature*. vol. 475, no. 7356. 2011: pp. 348–352. ISSN 0028-0836. doi:10.1038/nature10242.

[120] Ruiz-Ruano, F. J.; López-León, M. D.; Cabrero, J.; et al.: High-throughput analysis of the satellitome illuminates satellite DNA evolution. *Scientific Reports*. vol. 6. 2016: page 28333. ISSN 2045-2322. doi:10.1038/srep28333.

[121] Sayers, E. W.; Cavanaugh, M.; Clark, K.; et al.: GenBank. *Nucleic Acids Research*. vol. 47, no. D1. 2019: pp. D94–D99. ISSN 1362-4962. doi:10.1093/nar/gky989.

[122] Schmidt, T.: LINEs, SINEs and repetitive DNA: non-LTR retrotransposons in plant genomes. *Plant Molecular Biology*. vol. 40, no. 6. 1999: pp. 903–910. ISSN 0167-4412. doi:10.1023/A:1006212929794.

[123] Schnable, P. S.; Ware, D.; Fulton, R. S.; et al.: The B73 maize genome: complexity, diversity, and dynamics. *Science*. vol. 326, no. 5956. 2009: pp. 1112–1115. ISSN 1095-9203. doi:10.1126/science.1178534.

[124] Schneiker, S.; Perlova, O.; Kaiser, O.; et al.: Complete genome sequence of the myxobacterium *Sorangium cellulosum*. *Nature Biotechnology*. vol. 25, no. 11. 2007: pp. 1281–1289. ISSN 1087-0156. doi:10.1038/nbt1354.

[125] Sebaihia, M.; Wren, B. W.; Mullany, P.; et al.: The multidrug-resistant human pathogen Clostridium difficile has a highly mobile, mosaic genome. *Nature Genetics*. vol. 38, no. 7. 2006: pp. 779–786. ISSN 1061-4036. doi:10.1038/ng1830.

[126] Shao, F.; Wang, J.; Xu, H.; et al.: FishTEDB: A collective database of transposable elements identified in the complete genomes of fish. *Database*. vol. 2018, no. 2018. 2018: page bax106. ISSN 1758-0463. doi:10.1093/database/bax106.

[127] Shendure, J.; Ji, H.: Next-generation DNA sequencing. *Nature Biotechnology*. vol. 26, no. 10. 2008: pp. 1135–1145. ISSN 1546-1696. doi:10.1038/nbt1486.

[128] Siguier, P.; Gourbeyre, E.; Chandler, M.: Bacterial insertion sequences: Their genomic impact and diversity. *FEMS Microbiology Reviews*. vol. 38, no. 5. 2014: pp. 865–891. ISSN 1574-6976. doi:10.1111/1574-6976.12067.

[129] Siguier, P.; Perochon, J.; Lestrade, L.; et al.: ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Research*. vol. 34, no. suppl_1. 2006: pp. D32–D36. ISSN 0305-1048. doi:10.1093/nar/gkj014.

[130] Simpson, J. T.; Wong, K.; Jackman, S. D.; et al.: ABySS: a parallel assembler for short read sequence data. *Genome Research*. vol. 19, no. 6. 2009: pp. 1117–1123. ISSN 1088-9051. doi:10.1101/gr.089532.108.

[131] Smit, A.; Hubley, R.; Green, P.: RepeatMasker Open-4.0.
Retrieved from: http://www.repeatmasker.org

[132] Sokal, R.; Michener, C. D.: A statistical method for evaluating systematic relationships. *University of Kansas science bulletin*. vol. 38. 1958: pp. 1409–1438.

[133] Spannagl, M.; Nussbaumer, T.; Bader, K. C.; et al.: PGSB plantsDB: Updates to the database framework for comparative plant genome research. *Nucleic Acids Research*. vol. 44, no. D1. 2016: pp. D1141–D1147. ISSN 13624962. doi:10.1093/nar/gkv1130.

[134] Staton, S. E.; Burke, J. M.: Transposome: a toolkit for annotation of transposable element families from unassembled sequence reads. *Bioinformatics*. vol. 31, no. 11. 2015: pp. 1827–1829. ISSN 1460-2059. doi:10.1093/bioinformatics/btv059.

[135] Storer, J.; Hubley, R.; Rosen, J.; et al.: The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mobile DNA*. vol. 12, no. 1. 2021: pp. 1–14. ISSN 17598753. doi:10.1186/s13100-020-00230-y.

[136] Suwanto, A.; Kaplan, S.: Physical and genetic mapping of the Rhodobacter sphaeroides 2.4.1 genome: presence of two unique circular chromosomes. *Journal of Bacteriology*. vol. 171, no. 11. 1989: pp. 5850–5859. ISSN 0021-9193. doi:10.1128/jb.171.11.5850-5859.1989.

[137] Tansirichaiya, S.; Rahman, M. A.; Roberts, A. P.: The Transposon Registry. *Mobile DNA*. vol. 10, no. 1. 2019: page 40. ISSN 1759-8753. doi:10.1186/s13100-019-0182-3.

[138] Touchon, M.; Rocha, E. P. C.: Causes of Insertion Sequences Abundance in Prokaryotic Genomes. *Molecular Biology and Evolution*. vol. 24, no. 4. 2007: pp. 969–981. ISSN 1537-1719. doi:10.1093/molbev/msm014.

[139] Truta, E.; Capraru, G.; Rosu, C. M.; et al.: Morphometric pattern of somatic chromosomes in three Romanian seabuckthorn genotypes. *Caryologia*. vol. 64, no. 2. 2011: pp. 189–196. ISSN 0008-7114. doi:10.1080/00087114.2002.10589783.

[140] Tu, Z.: Eight novel families of miniature inverted repeat transposable elements in the African malaria mosquito, Anopheles gambiae. *Proceedings of the National Academy of Sciences*. vol. 98, no. 4. 2001: pp. 1699–1704. ISSN 00278424. doi:10.1073/pnas.041593198.

[141] Van Dijk, E. L.; Auger, H.; Jaszczyszyn, Y.; et al.: Ten years of next-generation sequencing technology. *Trends in Genetics*. vol. 30, no. 9. 2014: pp. 418–426. ISSN 0168-9525. doi:10.1016/j.tig.2014.07.001.

[142] Vandecraen, J.; Chandler, M.; Aertsen, A.; et al.: The impact of insertion sequences on bacterial genome plasticity and adaptability. *Critical Reviews in Microbiology*. vol. 43, no. 6. 2017: pp. 709–730. ISSN 1040-841X. doi:10.1080/1040841X.2017.1303661.

[143] Varani, A. M.; Siguier, P.; Gourbeyre, E.; et al.: ISsaga is an ensemble of web-based methods for high throughput identification and semi-automatic annotation of insertion sequences in prokaryotic genomes. *Genome Biology*. vol. 12, no. 3. 2011: page R30. ISSN 1465-6906. doi:10.1186/gb-2011-12-3-r30.

[144] Vassetzky, N. S.; Kramerov, D. A.: SINEBase: a database and tool for SINE analysis. *Nucleic Acids Research*. vol. 41, no. D1. 2013: pp. D83–D89. ISSN 0305-1048. doi:10.1093/nar/gks1263.

[145] Wagner, A.; Lewis, C.; Bichsel, M.: A survey of bacterial insertion sequences using IScan. *Nucleic Acids Research*. vol. 35, no. 16. 2007: pp. 5284–5293. ISSN 1362-4962. doi:10.1093/nar/gkm597.

[146] Waterston, R. H.; Lindblad-Toh, K.; Birney, E.; et al.: Initial sequencing and comparative analysis of the mouse genome. *Nature*. vol. 420, no. 6915. 2002: pp. 520–562. ISSN 0028-0836. doi:10.1038/nature01262.

[147] Wenger, A. M.; Peluso, P.; Rowell, W. J.; et al.: Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*. vol. 37, no. 10. 2019: pp. 1155–1162. ISSN 1546-1696. doi:10.1038/s41587-019-0217-9.

[148] Wheeler, T. J.; Clements, J.; Eddy, S. R.; et al.: Dfam : a database of repetitive DNA based on profile hidden Markov models. *Nucleic Acids Research*. vol. 41, no. D1. 2013: pp. D70–D82. ISSN 0305-1048. doi:10.1093/nar/gks1265.

[149] Wicker, T.; Sabot, F.; Hua-Van, A.; et al.: A unified classification system for eukaryotic transposable elements. *Nature reviews. Genetics*. vol. 8, no. 12. 2007: pp. 973–982. ISSN 1471-0064. doi:10.1038/nrg2165.

[150] Xie, Z.; Tang, H.: ISEScan: automated identification of insertion sequence elements in prokaryotic genomes. *Bioinformatics*. vol. 33, no. 21. 2017: pp. 3340–3347. ISSN 1460-2059. doi:10.1093/bioinformatics/btx433.

[151] Xu, H. E.; Zhang, H. H.; Xia, T.; et al.: BmTEdb: A collective database of transposable elements in the silkworm genome. *Database*. vol. 2013. 2013: page bat055. ISSN 1758-0463. doi:10.1093/database/bat055.

[152] Xu, Z.; Liu, J.; Ni, W.; et al.: GrTEdb: The first web-based database of transposable elements in cotton (Gossypium raimondii). *Database*. vol. 2017, no. 1. 2017: page bax013. ISSN 1758-0463. doi:10.1093/database/bax013.

[153] Yang, G.; Lee, Y. H.; Jiang, Y.; et al.: A two-edged role for the transposable element *Kiddo* in the *rice ubiquitin2* promoter. *The Plant Cell*. vol. 17, no. 5. 2005: pp. 1559–1568. ISSN 1040-4651. doi:10.1105/tpc.104.030528.

[154] Yi, F.; Jia, Z.; Xiao, Y.; et al.: SPTEdb: a database for transposable elements in salicaceous plants. *Database*. vol. 2018. 2018: page bay024. ISSN 1758-0463. doi:10.1093/database/bay024.

[155] Yi, F.; Ling, J.; Xiao, Y.; et al.: ConTEdb: a comprehensive database of transposable elements in conifers. *Database*. vol. 2018. 2018: page bay131. ISSN 1758-0463. doi:10.1093/database/bay131.

[156] Yuan, Z.; Liu, S.; Zhou, T.; et al.: Comparative genome analysis of 52 fish species suggests differential associations of repetitive elements with their living aquatic environments. *BMC Genomics*. vol. 19, no. 1. 2018: page 141. ISSN 1471-2164. doi:10.1186/s12864-018-4516-1.

[157] Zerbino, D. R.: Using the Velvet de novo Assembler for Short-Read Sequencing Technologies. *Current Protocols in Bioinformatics*. vol. 31, no. 1. 2010: pp. 11.5.1–11.5.12. ISSN 1934-3396. doi:10.1002/0471250953.bi1105s31.

[158] Zhang, L.; Richards, A.; Inmaculada Barrasa, M.; et al.: Reverse-transcribed SARS-CoV-2 RNA can integrate into the genome of cultured human cells and can be expressed in patient-derived tissues. *Proceedings of the National Academy of Sciences of the United States of America*. vol. 118, no. 21. 2021: page e2105968118. ISSN 10916490. doi:10.1073/pnas.2105968118.

[159] Zhou, F.; Xu, Y.: *RepPop*: A database for repetitive elements in *Populus trichocarpa*. *BMC Genomics*. vol. 10. 2009: pp. 1–9. ISSN 1471-2164. doi:10.1186/1471-2164-10-14.

[160] Zytnicki, M.; Akhunov, E.; Quesneville, H.: Tedna: A transposable element *de novo* assembler. *Bioinformatics*. vol. 30, no. 18. 2014: pp. 2656–2658. ISSN 1460-2059. doi:10.1093/bioinformatics/btu365.

Appendices

Appendix A

Satellite DNA and Transposable Elements in Seabuckthorn (*Hippophae rhamnoides*), a Dioecious Plant with Small Y and Large X Chromosomes

Janka Puterova, Olga Razumova, Tomas Martinek, Oleg Alexandrov, Mikhail Divashuk, Zdenek Kubat, Roman Hobza, Gennady Karlov, and Eduard Kejnovsky

Genome Biology and Evolution. 2017, 9(1), **DOI:** 10.1093/gbe/evw303

Satellite DNA and Transposable Elements in Seabuckthorn (*Hippophae rhamnoides*), a Dioecious Plant with Small Y and Large X Chromosomes

Janka Puterova[1,2], Olga Razumova[3], Tomas Martinek[2], Oleg Alexandrov[3], Mikhail Divashuk[3], Zdenek Kubat[1], Roman Hobza[1,4], Gennady Karlov[3,5], and Eduard Kejnovsky[1,*]

[1]Department of Plant Developmental Genetics, Institute of Biophysics, Academy of Sciences of the Czech Republic, Brno, Czech Republic

[2]Department of Information Systems, Faculty of Information Technology, Brno University of Technology, Brno, Czech Republic

[3]Centre for Molecular Biotechnology, Russian State Agrarian University – Moscow Timiryazev Agricultural Academy, Moscow, Russia

[4]Institute of Experimental Botany, Center of the Region Haná for Biotechnological and Agricultural Research, Olomouc, Czech Republic

[5]All-Russia Research Institute of Agricultural Biotechnology, Moscow, Russia

*Corresponding author: E-mail: kejnovsk@ibp.cz.

Accepted: January 3, 2017

Data deposition: This project has been deposited in European Nucleotide Archive (ENA) under study accession PRJEB14480. http://www.ebi.ac.uk/ena/data/view/PRJEB14480.

Abstract

Seabuckthorn (*Hippophae rhamnoides*) is a dioecious shrub commonly used in the pharmaceutical, cosmetic, and environmental industry as a source of oil, minerals and vitamins. In this study, we analyzed the transposable elements and satellites in its genome. We carried out Illumina DNA sequencing and reconstructed the main repetitive DNA sequences. For data analysis, we developed a new bioinformatics approach for advanced satellite DNA analysis and showed that about 25% of the genome consists of satellite DNA and about 24% is formed of transposable elements, dominated by Ty3/*Gypsy* and Ty1/*Copia* LTR retrotransposons. FISH mapping revealed X chromosome-accumulated, Y chromosome-specific or both sex chromosomes-accumulated satellites but most satellites were found on autosomes. Transposable elements were located mostly in the subtelomeres of all chromosomes. The 5S rDNA and 45S rDNA were localized on one autosomal locus each. Although we demonstrated the small size of the Y chromosome of the seabuckthorn and accumulated satellite DNA there, we were unable to estimate the age and extent of the Y chromosome degeneration. Analysis of dioecious relatives such as Shepherdia would shed more light on the evolution of these sex chromosomes.

Key words: sex chromosomes, genome composition, chromosomal localization, repetitive DNA.

Introduction

Seabuckthorn (*Hippophae rhamnoides*) is a hardy, deciduous dioecious shrub belonging to the Elaeagnaceae family with a natural habitat extending widely across Europe and Asia. It is used in traditional Chinese, Tibetan and Siberian medicine and has special characteristics exploitable in biotechnology, pharmaceutical and cosmetic sciences, as a source of oil, minerals and vitamins. The size of seabuckthorn genome is ~2.55 Gbp/2C (Zhou et al. 2010) but there is a dearth of information on its composition. The ribosomal DNA ITS regions were compared among *H. rhamnoides* ssp chinensis from different geographical areas of China and showed distinct genetic variation (Chen et al 2010). RAPD markers (Sharma et al. 2010) were identified with the aim of determining the sex of individuals. Cytogenetic analysis is represented only by the older works of Shchapov (1979) and Rousi and Arohonka (1980) who both determined the diploid chromosome number $2n=24$. Shchapov (1979) revealed the small Y and large X chromosomes. Seabuckthorn transcriptome has been analyzed recently providing a resource for gene discovery and development of molecular markers (Ghangal et al. 2013).

Sex chromosomes have evolved repeatedly and independently in the plant kingdom with different age and degree of degeneration shown in various dioecious species (Ming et al.

2011; Hobza and Vyskot 2015; Charlesworth 2016). The evolution of the Y chromosomes is characterized by gene erosion/loss and accumulation of repetitive DNA (Kejnovsky et al. 2009). The most studied dioecious model species with heteromorphic sex chromosomes are white campion (*Silene latifolia*, Kejnovsky and Vyskot 2010), sorrel (*Rumex acetosa*, Steflova et al. 2013; *R. hastatulus*, Hough et al. 2014), ivy gourd (*Coccinia grandis*, Sousa et al. 2013), and members of the Cannabaceae family (*Humulus lupulus*, Divashuk et al. 2011; *H. japonicus*, Alexandrov et al. 2012; *Cannabis sativa*, Divashuk et al. 2014).

The majority of large plant genomes are formed of repetitive DNA, mostly by transposable elements and tandem repeats (satellite DNA). The processes of repetitive DNA amplification and elimination are only partially understood. Turnover of repeats is high and corresponds only to million of years (Lim et al. 2007). The localization of repetitive DNA on sex chromosomes is different from that of autosomes, reflecting different repeat dynamics, especially on the nonrecombining regions of the Y chromosomes (Kejnovsky et al. 2009). Satellite DNA has mostly discrete localization in the genome and some satellites are thus Y chromosome-specific (Mariotti et al. 2009). In contrast, transposable elements have more homogenous distribution and are only slightly enriched on the Y chromosome (Charlesworth 1991; Cermak et al. 2008) or alternatively absent on the Y chromosome as shown in *Silene latifolia* (Cermak et al. 2008; Kubat et al. 2014) and *Rumex acetosa* (Steflova et al. 2013) despite their presence in the rest of genome. The striking example is the large Y chromosome of the dioecious plant *Coccinia grandis* showing accumulation of transposable elements, satellites, and organellar DNA (Souza et al. 2016). One review published recently discusses the role of repetitive DNA in the evolution of sex chromosomes and includes a database of transposable elements of dioecious plants (Li et al. 2016a, 2016b).

In this study, we analyzed the transposable elements and satellites in the seabuckthorn genome and determined the chromosomal localization of these repeats. We showed that seabuckthorn has an XY system with large X and small Y chromosomes.

Materials and Methods

Illumina Sequencing

DNA isolation from male (Pollinator 1) and female (cv "Botanicheskaya lyubitelskaya") plants was carried out according to Doyle and Doyle (1990). One Illumina MiSeq sequencing run was performed for each male and female genomic DNA. The voucher specimen of the plants used in the study was kept for record in the herbarium (AT) of Department of Botany and Breeding of Horticultural Crops of the Russian State Agrarian University – MTAA (Voucher No.5470). Sequencing reads were analyzed by quality control tool FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/; last accessed January 4, 2017) followed by quality filtering based on the sequence quality score, adaptors trimming, filtering out short or unpaired sequences and trimming all reads to lengths of 230 nucleotides using the Trimmomatic tool (Bolger et al. 2014), leading to 1,848,543 male and 1,863,670 female paired-end reads. Quality-filtered reads were randomly sampled to 415,650 paired-end reads for both male and female individuals and the reads were merged together (totally 1,662,600 reads). As the nuclear DNA content of *H. rhamnoides* reported in Zhou et al. (2010) was determined to be ~2.61/2C pg (without detailed specification of male or female) we converted it to genome size (in bp) using following formula (Doležel et al. 2003): $g = \text{DNA content (pg)} \times (0.978 \times 10^9)$, resulting into ~2.55 Gbp/2C, our samples represent ~30% of haploid genome. Genome coverage was calculated as follow: $cov = (r \times l)/g$, where *r* corresponds to number of reads used in our analysis, *l* to read length and *g* to haploid genome size of *H. rhamnoides*.

Repeat Identification and Annotation

In order to identify repetitive sequences in the *H. rhamnoides* genome we employed comparative graph-based clustering analysis of sequenced reads by RepeatExplorer pipeline (Novak et al. 2013). Only clusters containing at least 0.01% of all clustered reads were considered and they corresponded to 58.5% of the genome. These were further manually characterized based on the similarity search results from RepeatMasker (http://www.repeatmasker.org; last accessed January 4, 2017) against Viridiplantae database and blastn and blastx (Altschul et al. 1990) against GenBank nr (Benson et al. 2009), which are part of the RepeatExplorer output. Cluster shapes were also used for repeat identification as tandem repeats with monomer longer than read length have typical donut-shaped clusters (Novak et al. 2010). Additionally, advanced analysis of satellite sequences, described in the section Satellite DNA sequences analysis, was used in the manual annotation of clusters.

Structural Annotation of LTR Retrotransposons

We reconstructed several Ty3/*Gypsy* and Ty1/*Copia* retrotransposons. The reconstruction comprised several steps. First, clusters belonging to particular element were visualized in SeqGrapheR (https://cran.r-project.org/web/packages/SeqGrapheR/index.html; last accessed January 4, 2017) program and contigs which together covered the whole elements were selected. These contigs were searched for occurrences of protein domains (GAG, RT, RH, AP, INT) by querying them to CDD (Marchler-Bauer et al. 2015). We then did multiple sequence alignment to create a consensus sequence of these contigs using progressive pairwise alignment implemented in Geneious 8.1.7 (http://www.geneious.com; last accessed January 4, 2017, Kearse et al. 2012). If necessary, resulting alignments were manually modified with respect to the order of domains for particular type of transposable element. The consensus sequence of reconstructed elements was then searched for the structural motif characteristics (ORFs and LTRs). Possible ORFs were detected by ORF Finder (https://www.ncbi.nlm.nih.gov/orffinder/; last accessed January 4, 2017). LTRs were determined on the basis of shape of a cluster and the element's coverage. Male and female coverage of reconstructed elements was determined by mapping reads which formed a current element to its consensus sequence using BowTie2 tool (Langmead and Salzberg 2012). Structural features and male and female coverage of reconstructed elements were visualized by custom R script and graph layouts of reconstructed elements were depicted by SeqGrapheR.

Phylogeny and Classification

Firstly, we created custom databases of plant LTR retrotransposon RT domains from sequences available in TREP (Wicker et al. 2002) and GyDB (Llorens et al. 2011) databases, independently for Ty3/*Gypsy* and Ty1/*Copia* retrotransposons. Contigs corresponding to retrotransposons were examined for the presence of a reverse transcriptase domain and Ty3/*Gypsy* and Ty1/*Copia* cores of RT domains were trimmed from these contigs based on the exact localization designated by CDD (Marchler-Bauer et al. 2015). Cores of RT domains were aligned by MUSCLE algorithm (Edgar 2004) together with our custom-made database of RT domains, and the resulting multiple sequence alignment was used as an input to create Neighbor-Joining tree (Saitou and Nei 1987) with Jukes-Cantor distance model using Geneious 8.1.7 (http://www.geneious.com; last accessed January 4, 2017, Kearse et al. 2012).

Preparation of Chromosomes and Probes and Fluorescence *In Situ* Hybridization

For chromosome preparations vegetatively propagated for commercial use, male ("Pollinator 1" and "Pollinator 3") and female (cv "Lomonosovskaya" and cv "Botanicheskaya ljubitelskaya") plants were used. Plant material was kindly provided by Dr G. Boyko, Lomonosov Moscow State University. The root tips were harvested separately from the individual male and female plants grown in pots. The harvested root tips were immediately pre-treated with a 2 mM aqueous solution of 8-hydroxyquinoline for 6 h at 20 °C. A 3:1 ethanol/glacial acetic acid (v/v) mix was used for fixation. Meristems 2 mm long were cut from the fixed root tips and digested in 10 µl enzyme solution [0.5% cellulase Onozuka R-10 (Serva, Germany) and 0.5% pectolyase Y-23 (Seishin Corp., Japan)] in 10 mM citrate buffer (pH = 4.9) for 2.5 h at 37 °C. The suspended cells were used for chromosome preparation as described by Kirov et al. (2014). The quality of spreads was assessed microscopically using phase-contrast and only preparations with at least 20 well-spread metaphases were used.

Probes for fluorescence *in situ* hybridization were generated using PCR-DIG Labeling Mix PLUS (Roche Diagnostics Gmbh) or by Biotin-11-dUTP 1/3 PCR labeling Mix (ZAO Sileks, Moscow). Primers for RT domain of selected transposable elements and determined monomer sequence of satellites were designed by Primer3 tool (Untergasser et al. 2012), were synthesized by ZAO "Syntol" (Moscow). These are available in supplementary table S1, Supplementary Material online. The pTa71 (45S rDNA) and pCT4.2 (5S rDNA) clones labeled by DIG-Nick translation kit were also used (Gerlach and Bedbrook 1979; Campell et al. 1992).

FISH experiments were performed as described in Alexandrov and Karlov (2016). For digoxigenin and biotin detection, slides were incubated with anti-DIG-FITC conjugate (Roche) and/or streptavidin-Cy3 conjugate (Sigma). The chromosomes were counterstained with DAPI (2 µg/ml) and mounted in Vectashield (Vector). An AxioImager M1 fluorescent microscope (Zeiss) was used to observe metaphase plates with fluorescent signals that were photographed with a monochrome AxioCam MRm CCD camera and visualized using Axiovision software (Zeiss).

Satellite DNA Sequences Analysis

As the seabuckthorn genome is abundant in satellite DNA and manual inspection would be exhaustive, we developed a custom bioinformatics approach which extended the basic analysis of RepeatExplorer tool. As an input the satellite clusters identified by RepeatExplorer are required. It is highly recommended to do manual inspection of these clusters and verify their structure and interaction with other clusters based on similarities among other clusters and pair-end reads connections. Our approach consisted of three basic steps.

(i) *Detection of satellite monomers*: First, assembled contigs of selected clusters were extracted from RepeatExplorer output and for each contig the monomer length was estimated from the distribution of distances between the same k-mers. The resulting monomer sequence was then extracted from the most covered part of the contig of previously determined length. Only the monomers with clearly distinguishable length, longer than 100 bp and reaching average coverage 50x and more were taken into account.

(ii) *Estimation of satellite families composition in genome and their annotation*: First, all to all monomer similarity was calculated. In order to do alignment of tandemly repeated monomers correctly (offsets between monomers are not known) we used one monomer as a subject and two copies in a row of the second monomer as a query. The similarity between monomers was then determined based on semiglobal alignment. To estimate the composition of satellite families in the genome, we clustered the monomer's similarity matrix using UPGMA method. The resulting dendrogram was then inspected by the user and cut off at the level that best discriminated the individual families (usually 70-85% of monomer identity). Identified families were visualized by the algorithm described by Fruchterman and Reingold (1991) implemented in igraph library and only connections that exceeded specified cut-off were considered and depicted. Secondly, to annotate identified families, all monomers were searched for similarity hits with sequences in the public nucleotide database and PlantSat database (Macas et al. 2002) using blastn (Altschul et al. 1990) with word size set to 11. Only results with an e-value lower than 10^{-20} were considered as significant. Finally, to depict satellite diversity inside the family, we chose the most covered monomer as a reference and mapped all reads belonging to the family onto its reference using BWA-MEM mapping tool (Li 2013). Conservation of different parts of the monomer was depicted using sequence logo created by WebLogo (Crooks et al. 2004) tool.

(iii) *Visualization of satellite families homogeneity:* First, the relative abundance of male and female reads was calculated in each tandem repeat family. This enabled us to predict their presence in sex chromosomes. We visualized the satellite homogeneity using the following procedure: reads from each identified family were merged together and sampled randomly to limit the maximum number of reads to speed up the following analysis. Similarity of sampled reads from all families was calculated using the megablast tool (Camacho et al. 2009) that performed all against all sequence comparison. Pairs of reads that met specific similarity threshold (70% sequence identity over at least 55% of sequence length) were further used for graph construction and visualization. Male and female reads were distinguished by color (male—blue, female—red), tandem repeat families were highlighted by different colors and the algorithm by Fruchterman and Reingold (1991) was used to depict the results. Additionally, graphs for selected families were refined with similarity thresholds ranging from 70% to 95% sequence identity to show satellite composition more clearly. Each satellite falling within individual satellite family was marked by a different color.

Results

Genomic Composition

We performed one Illumina MiSeq platform sequencing run for each male and female genomic DNA followed by graph-based clustering of reads and characterization of repetitive sequences by RepeatExplorer (Novak et al. 2013). All 223 clusters (with more than 167 reads) contained 973,049 reads corresponding to 58.5% of genome (fig. 1) and their identification showed that dominant (first) clusters corresponded to satellite DNA followed by Ty3/*Gypsy* and Ty1/*Copia* LTR retrotransposons. One cluster (CL97) corresponded to 5S rDNA, two clusters (CL40, CL71) to 45S rDNA and 15 clusters to chloroplast DNA (cpDNA). Although the majority of chloroplast DNA reads probably originated from contaminating cpDNA, some proportion could come from nuclear cpDNA insertions (NUPTs).

We identified main types of repetitive DNA and their genome proportions in male and female individuals (table 1). All transposable elements represented together 24% of male and 23% of female genome. Ty1/*Copia* retrotransposons formed 12%, Ty3/*Gypsy* retrotransposons 11% and DNA transposons 1.5% of male genome. The most abundant among Ty1/*Copia* retrotransposons were Angela/Tork and Ale/Retrofit, among Ty3/*Gypsy* retrotransposons Athila and chromoviruses dominated. No LINE elements were found in the whole seabuckthorn genome. Satellites together comprised about 27% of male and 24% of female genomes. The 45S rDNA formed 0.7% of both male and female genomes and 5S rDNA represented 0.2% of both male and female genomes.

Transposable Elements

To determine the phylogenetic relationships of Ty1/*Copia* and Ty3/*Gypsy* retrotransposons, we aligned their reverse transcriptase (RT) domains from individual clusters and constructed the phylogenetic trees. Both Ty3/*Gypsy* (fig. 2*A*) and Ty1/*Copia* (fig. 2*B*) trees contained families identified in our clusters (in red) mixed with representatives of known subfamilies of Ty1/*Copia* or Ty3/*Gypsy* from other plant species (in black). Among Ty3/*Gypsy* retrotransposons, we identified five clusters containing Athila subfamilies, one CRM subfamily, one Galadriel, one Reina and one Tat/Ogre subfamily (fig. 2*A*). Among Ty1/*Copia* retrotransposons, we found four subfamilies of Ale/Retrofit, four Angela/Tork subfamilies, one Maximus/SIRE subfamily, two TAR subfamilies and two Ivana/Oryco subfamilies (fig. 2*B*). The Angela/Tork and Ale/Retrofit subfamilies showed higher variability while Athila subfamilies were homogenous. Highest homogeneity were shown by chromoviruses where all reads were assembled into a single cluster for CRM, Galadriel and Reina families (fig. 2*A*).

We reconstructed the structure of the main Ty3/*Gypsy* and Ty1/*Copia* subfamilies (fig. 3) and identified all main features such as *gag* and *pol* genes (with all domains) and long terminal repeats (LTRs). In some retrotransposons (CL6, CL16) LTR regions were assembled into one long terminal repeat while in other clusters (CL7, CL27) right and left LTR were distinguished. This may be a consequence of lower or higher mutual diversity of LTRs in one element, and could correspond to age differences of elements. Graph layouts (right part of fig. 3) show the variability of specific parts of elements as well as alternative variants of elements, e.g., potential spliced variant (Novak et al. 2010). The similar coverage of elements by male and female reads indicates that elements are present on all chromosomes without accumulation/absence on

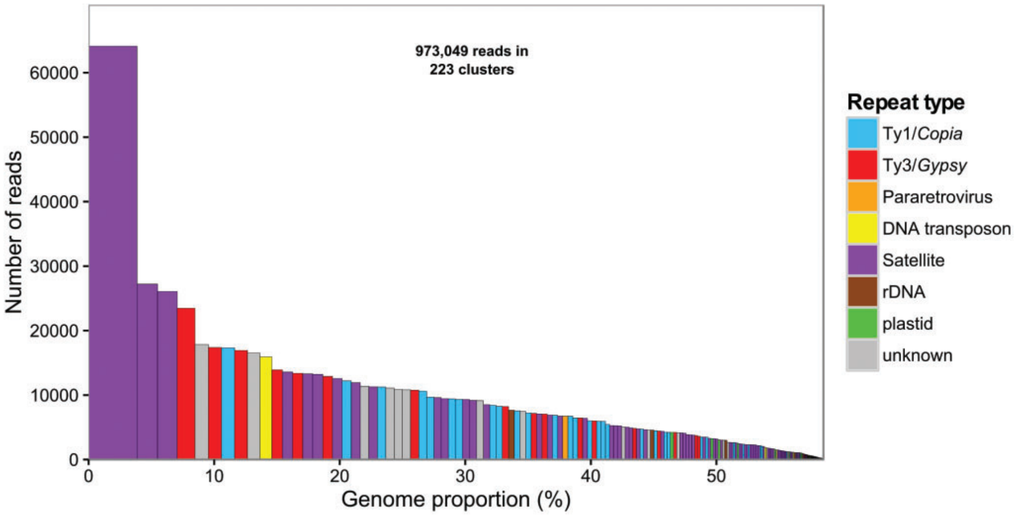


Fig. 1.—Repeat composition of clusters and their genomic proportions. Each column corresponds to one cluster and repeat types are distinguished by colors. The height of columns represents number of reads in each cluster, the width of column indicate genomic proportion of cluster.

Table 1
Repeat Composition in *Hippophae rhamnoides* Genome

Classification			Genome Proportion (%)	
Repeat Type	**Super Family**	**Family**	**Male**	**Female**
LTR retroelements	Ty1/Copia	Angela/Tork	4.83	4.90
		Ale/Retrofit	4.93	4.38
		TAR	1.34	1.06
		Maximus/SIRE	0.44	0.57
		Ivana/Oryco	0.25	0.23
		Total Ty1/*Copia*	11.79	11.15
	Ty3/Gypsy	Athila	6.39	5.36
		Chromovirus—CRM	2.98	3.58
		Chromovirus—Galadriel	1.28	0.80
		Chromovirus—others	0.27	0.31
		Chromovirus—Reina	0.06	0.04
		Tat/Ogre	0.05	0.05
		Total Ty3/*Gypsy*	11.04	10.15
DNA transposons			1.52	1.46
Total transposable elements			24.35	22.76
Pararetrovirus			0.48	0.59
rDNA	45S		0.77	0.69
	5S		0.20	0.16
Satellites			26.92	23.74
All repetitive elements			52.72	47.94
Unclassified			6.96	11.39
Low/single copy			38.96	39.50
Plastids			1.36	1.17

Note.—Types of repetitive DNA and their genome proportions.

the X or Y chromosome. Some elements had uninterrupted ORF corresponding to *gag* and *pol* (CL7, CL27, and CL43) and hence they can be active. Interruption of ORFs in other elements may have been caused by assembling errors during reconstruction (CL6, CL16, and CL37).

Satellite DNA

We developed a new bioinformatics approach for detailed analysis of satellite DNA in genomes. This method includes: (i) identification of satellite monomers based on distribution of distances of k-mers in assembled contigs, (ii) clustering of monomers allowing identification and annotation of satellite families in genome, and (iii) visualization of satellites homogeneity and male/female composition allowing better prediction of their localization with respect to sex chromosomes. Detailed description of the whole procedure is available in the section Materials and Methods and in supplementary figure S4, Supplementary Material online.

We utilized this approach for analysis of the seabuckthorn genome, but it is generally applicable in genomic studies of other species as well. As an input we used the 38 largest manually inspected satellite clusters from RepeatExplorer output extended by five smaller clusters with potentially interesting chromosomal localization (X, Y chromosomes). All clusters were grouped into 12 main superclusters that correspond to the 12 main families of satellite DNA in the seabuckthorn genome. Satellites were named HRTR1-HRTR12 (supplementary fig. S1, Supplementary Material online and table 2). Copy number of individual satellite families was determined based on following formula: *cn* = *[(s x l)/m]/cov*, where *s* represents number of reads of individual satellite family, *l* corresponds to

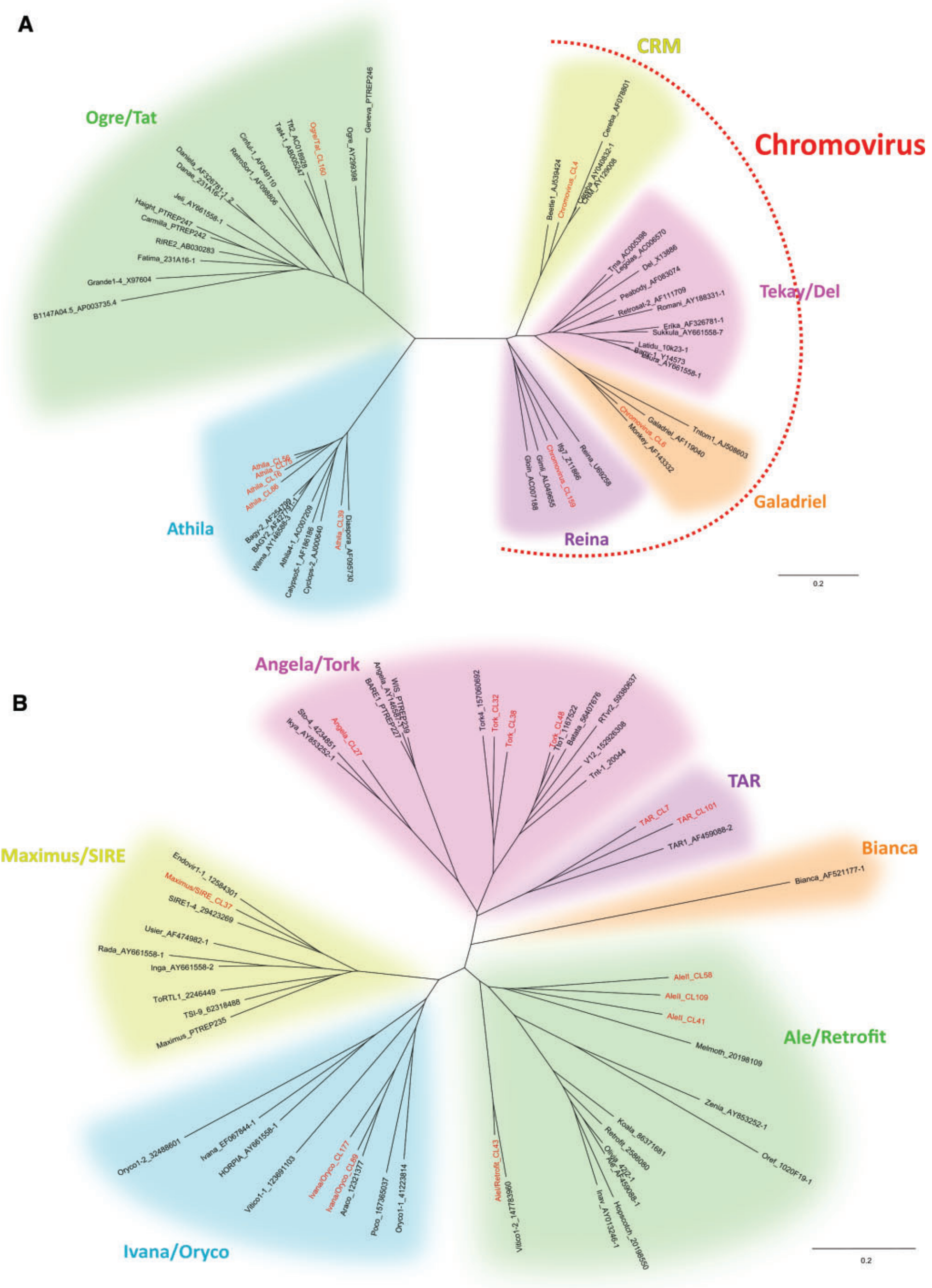


FIG. 2.—Phylogenetic trees of *Hippophae rhamnoides* Ty3/*Gypsy* (*A*) and Ty1/*Copia* (*B*) retrotransposons based on reverse transcriptase sequences. RT domains of retrotransposons reconstructed from Illumina reads in this study are in red, representative RT domains of retrotransposons from other plant species (from TREP and GyDB) are in black. Individual families are highlighted by different colors.

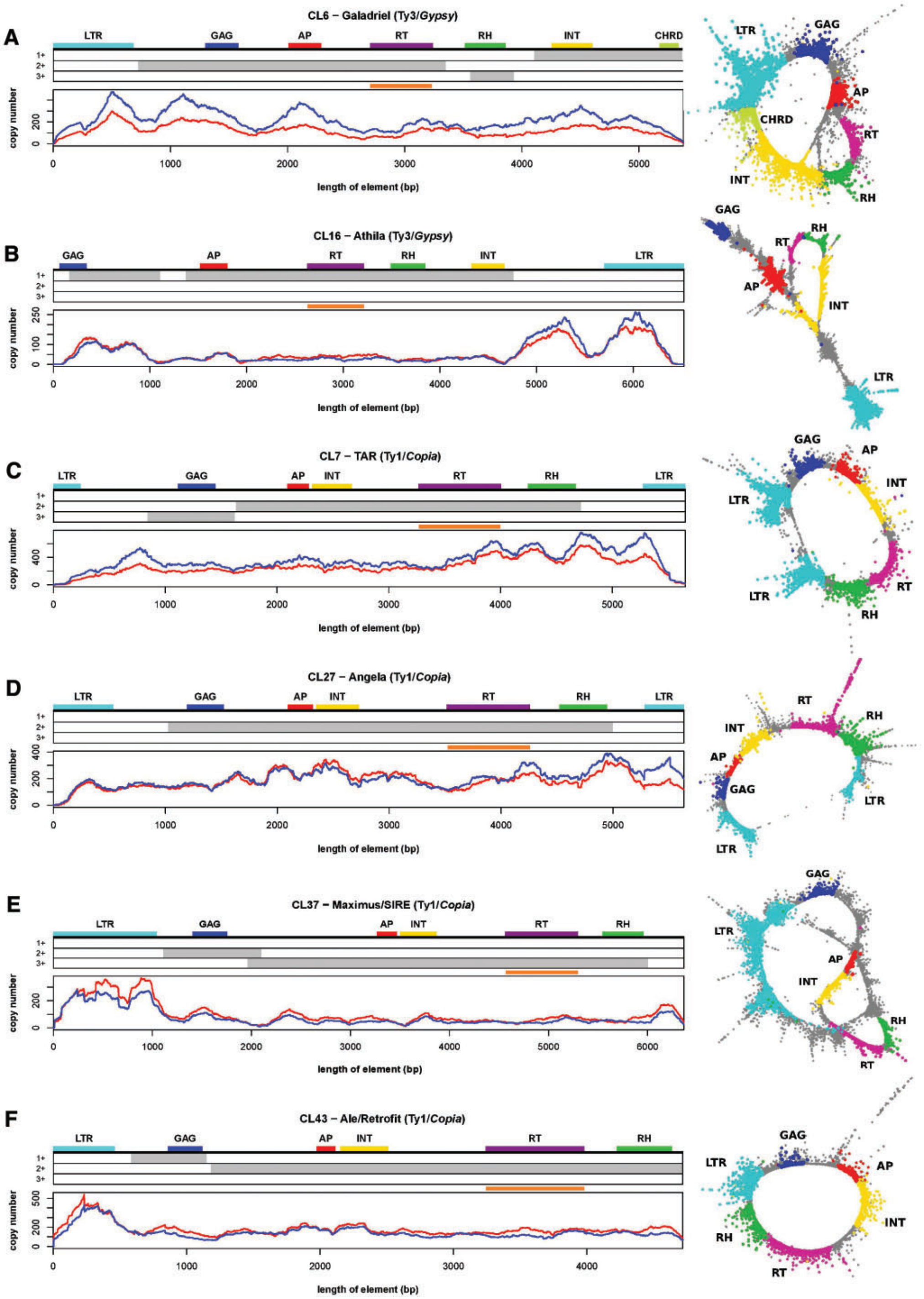


Fig. 3.—Comparison of structure of selected retrotransposon families in *Hippophae rhamnoides*. Graphs of coverage by male (in blue) and female (in red) genomic reads are showed under the structure of Ty3/*Gypsy* (*A*, *B*) and Ty1/*Copia* (*C*–*F*) elements shown in phylogenetic tree (fig. 2). Graph layouts on the right are visualized by SeqGrapheR program (http://cran.rproject.org/web/packages/SeqGrapheR/index.html). Protein domains and possible LTRs are distinguished by colors, found possible different three ORFs are marked by grey rectangles and orange line represents sequence for probes used for FISH.

Table 2
Main Satellite Families in *Hippophae rhamnoides* Genome

Name	Number of Reads	Localization	Monomer Length	M (%)	F (%)	Copy Number
HRTR1	129843	Strong signal on six pairs of small autosomes and weak signal on one pair of small autosomes	363	59.90	40.10	82270
HRTR2	60455	X and Y chromosome and weak signal on one pair of large and one pair of small autosomes	541	43.03	56.97	25702
HRTR3	46881	Dispersed signal on two large autosomal pairs	656	49.60	50.40	16437
HRTR4	27219	One pair of large and one pair of small autosomes	720	51.30	48.70	8695
HRTR5	23060	One pair of small autosomes	819	57.61	42.39	6476
HRTR6	19415	Three pairs of small autosomes	198[a]	53.67	46.33	5784[b]
HRTR7	14861	One pair of large autosomes and one pair of small autosomes	493[a]	68.38	31.62	4828[b]
HRTR8	12570	X chromosome and weak signal on one pair of small autosomes	826	35.06	64.94	3500
HRTR9	11155	One pair of small autosomes	354	69.52	30.48	7248
HRTR10	7476	Centromere of one pair of small autosomes	940	49.80	50.20	1829
HRTR11	4088	One pair of small autosomes	643	66.78	33.22	1462
HRTR12	1718	Y chromosome	257	100.00	0.00	1538

Note.—Names, monomer lengths, copy numbers, chromosomal localizations, and genome proportions.
[a]Shared length of the monomer in the family.
[b]Estimated based on average monomer length. 772 bp for HRTR6 and 708 bp for HRTR7.

read length, *m* represents estimated monomer length for satellite family and *cov* is genome coverage. Sequence logos show the monomer sequences of the main satellites and the sequence variability (supplementary fig. S2*A*–*L*, Supplementary Material online). Only HRTR1 and HRTR12 showed significant similarity hits with blast nucleotide (nr/nt) database (to previously deposited microsatellite markers of *H. rhamnoides*). There were no significant hits with PlantSat database for all satellite groups.

Based on our detailed analysis of HRTR6 and HRTR7, sharing small part of monomers (supplementary fig. S3*C*, Supplementary Material online), we decided to retain them as two separate tandem repeat families instead of one. These two families were very divergent and each showed variability in monomer' length (HRTR6: 730–810 bp, HRTR7: 475–830 bp). Monomers in each family had a common sequence (HRTR6: 198 bp, HRTR7: 493 bp) while other parts of monomers were significantly different from each other. For this reason, we only created sequence logos for the shared part of monomers for each family (supplementary fig. S2*F* and *G*, Supplementary Material online).

Male versus Female Comparison

To compare male and female genomes and to predict which repetitive DNA is specific for or accumulated on the X and Y chromosomes, we plotted the numbers of male versus female reads corresponding to individual clusters (fig. 4). This analysis involved all 223 clusters. The majority of clusters was located on the diagonal and these corresponded to transposable elements, rDNA and some satellites. However, some clusters containing satellites were enriched or even specific for males and represented potential Y-specific repeats. Other repeats, mostly satellites, were more abundant in females which could reflect their enrichment or specific localization on the X chromosome.

The greatest differences in composition of male and female reads were observed in satellites (five clusters located in the left; fig. 4). Detailed analysis showed that one of these (CL123—HRTR12) formed an isolated family composed of male reads only which suggests its localization only on the Y chromosome (fig. 5). The other four male biased satellites represented either a variant of a specific widespread cluster with Y chromosome presence (CL99 and CL144—HRTR2) or a satellite with a minor presence on the Y chromosome (CL150—HRTR1 and CL132—HRTR3). Eight satellites contained more female than male reads (2:1) indicating its localization on the X chromosome (female has two X chromosomes, male only one). HRTR2 satellite also contained more female than male reads but the ratio was 1.3 to 1 which could be explained by the localization on both sex chromosomes with greater abundance on the X than on the Y chromosome (fig. 5). Most other satellites had similar abundance of male and female reads, suggesting their localization (at least mostly) on autosomes.

Chromosomal Localization of Transposable Elements and Satellites

For determination of the chromosomal localization of transposable elements and satellites in seabuckthorn, we prepared probes representing reverse transcriptase region of individual TE families or part of a satellite monomer (supplementary fig.

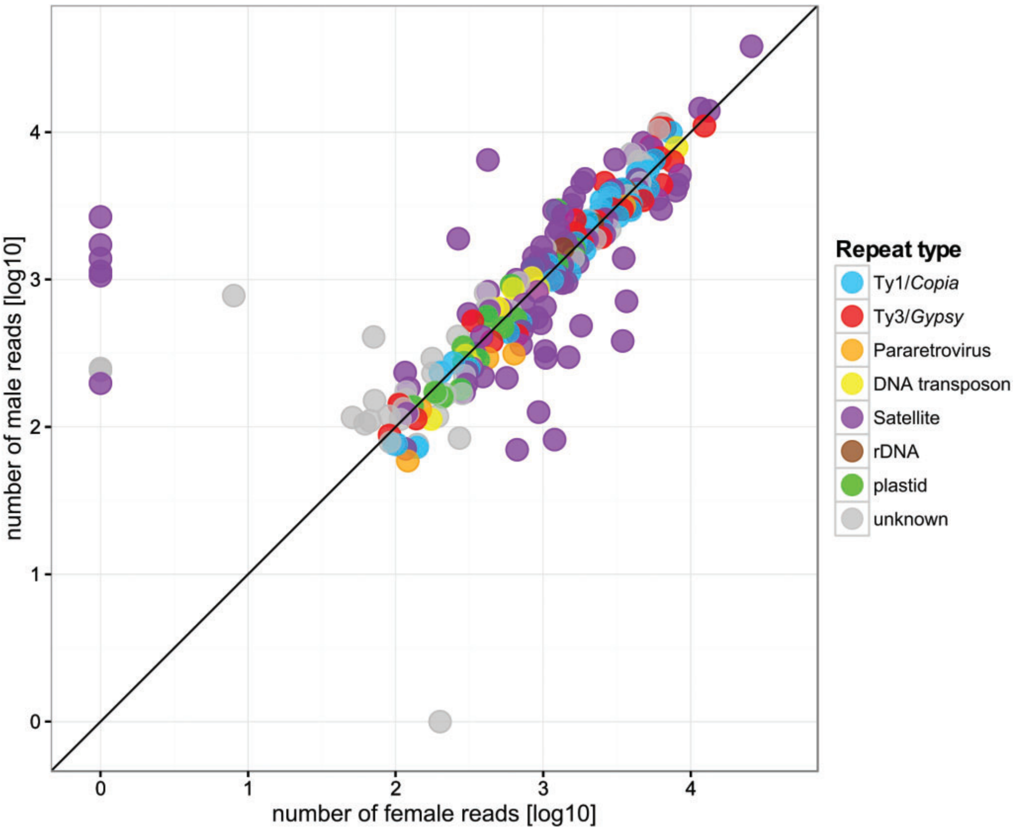


Fig. 4.—Comparison of repeats in male and female of *Hippophae rhamnoides*. Number of male versus female reads corresponding to individual clusters. Each circle in plot represents one cluster. Repeat types are marked by different color. Clusters in left upper part of graph are enriched (or specific) for males and thus potentially located on the Y chromosome while clusters in the right bottom part are enriched in female and thus potentially located on the X chromosome.

S1, Supplementary Material online) and used them for fluorescence *in situ* hybridization (FISH). In all FISH experiments we used both male (Pollinator 1, Leningradskaya region) and female (cv "Botanicheskaya lyubitelskaya") metaphases from plants that was used for sequencing. FISH experiments were also expanded to male ("Pollinator 3" Kaliningrad region) and female (cv "Lomonosovskaya"). In all ecotypes, we got the same results with X and Y.

FISH with satellite DNA showed various localization patterns on metaphase chromosomes of *H. rhamnoides* (fig. 6). The HRTR2, HRTR8 and HRTR12 show the sex specific or accumulation pattern of hybridization, while for HRTR3, HRTR4, HRTR5, HRTR6, HRTR7, HRTR9, HRTR10, and HRTR11 the hybridization patterns was the same for male as well as for female. The HRTR1 satellite hybridized mainly to heterochromatic arms of six pairs of small autosomes and weakly on one more pair of small autosomes (fig. 6*A* and *B*). In addition, a weak signal was detected distal to centromere on one arm of one large chromosome (chromosome X) in male (fig. 6*A*) and two large chromosomes in female (fig. 6*B*). The HRTR2 satellite gave a strong FISH signal on one large chromosome (chromosome X) and on one small chromosome (chromosome Y) in male (fig. 6*C*) and a strong FISH signal on two large chromosomes (chromosome X) in female (fig. 6*D*). Also a weak signal on the centromeric region of a pair of large and a pair of small autosomes was detected in both sexes. The HRTR3 satellite was localized on two large autosomal pairs with the FISH signal dispersed along these chromosomes (fig. 6*E*). The HRTR4 localized on one pair of large and on one pair of small autosomes (fig. 6*F*). The HRTR5 signal was detected on one pair of small autosomes only (fig. 6*G*). HRTR6 gave a strong signal on one autosomal pair and a weaker signals on two autosomal pairs (fig. 6*H*). The HRTR7 showed two sites of hybridization on one arm of a pair of large autosomes and on the centromeric region of a pair of small autosomes (fig. 6*I*). The HRTR8 hybridized mainly to the one large chromosome

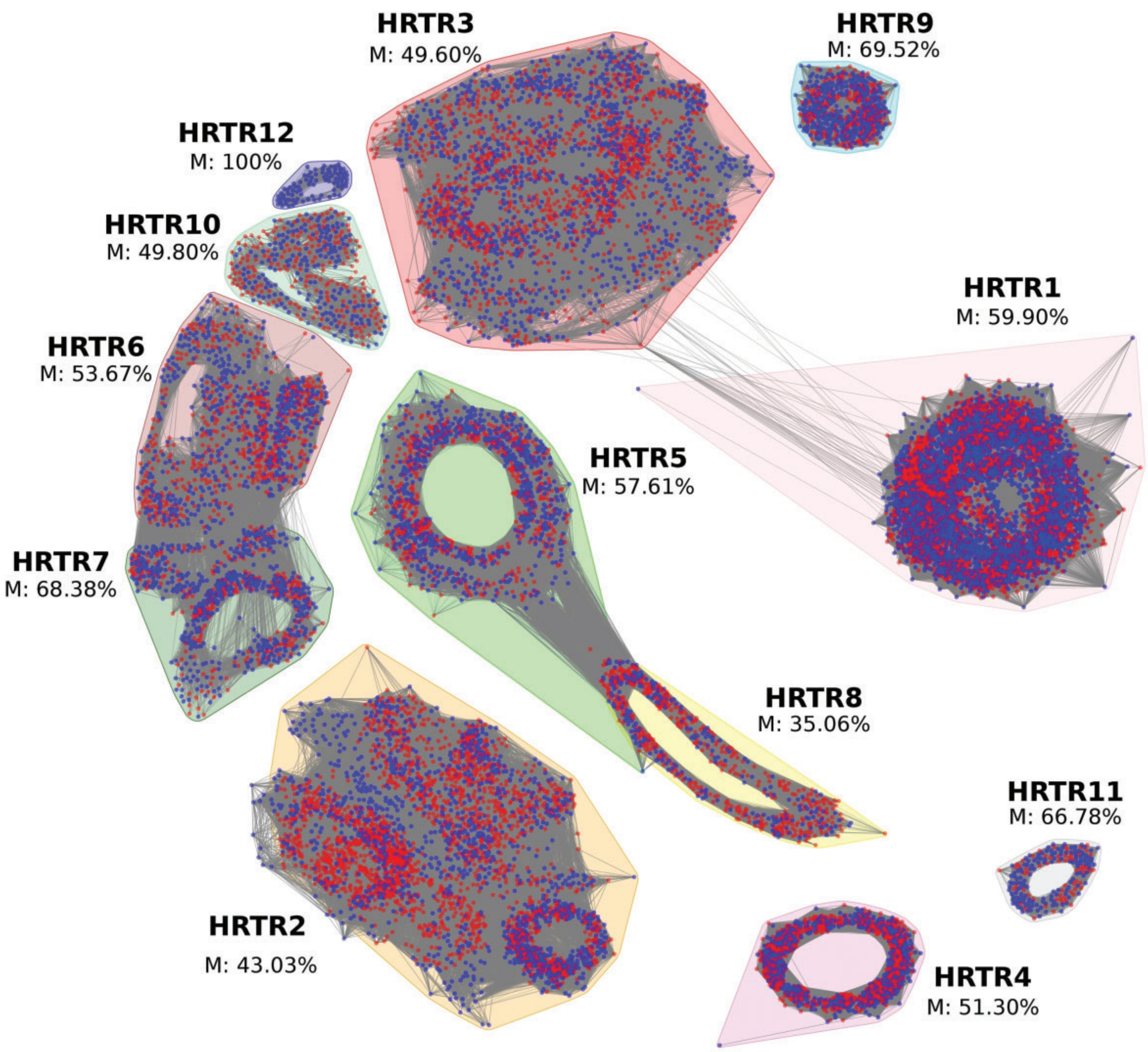


Fig. 5.—Visualization of male/female reads homogeneity in satellite families. Graph nodes correspond to sequenced reads and edges connect overlapping reads with more than 70% of sequence identity over at least 55% read length. Distances between reads are inversely proportional to their sequence similarity. Male reads are labeled by blue and female reads by red color. Individual families are highlighted by different colors. Please note HRTR12 family that is composed of male reads only assuming to be Y-specific.

(chromosome X) in male (Fig. 6*J*) and to the two large chromosomes (chromosomes X) in female (fig. 6*K*). A weak signal was also detected on one pair of small autosomes. The HRTR9, HRTR10, and HRTR11 were localized on one pair of small autosomes each (fig. 6*L–N*). The HRTR12 hybridized specifically to the small chromosome (Y chromosome) (fig. 6*O*) in male and no signal was detected in female (fig. 6*D*). The FISH signal intensity from HRTRs on X chromosomes varied depending on genotype.

Localization of the HRTR1 and the Y-specific (HRTR12), X-accumulated (HRTR8) and X and Y-accumulated (HRTR2) satellites on sex chromosomes was demonstrated by bicolor FISH using combinations of these probes and is summarized in a scheme (fig. 7). This together with specific or enriched representation of clusters in male and female (figs. 4 and 5), clearly demonstrates that *H. rhamnoides* has heteromorphic sex chromosomes (XY system) with large X and the small Y chromosomes.

We also mapped ribosomal genes. 45S rDNA was localized on one pair of small autosomes (fig. 8*A*) and 5S rDNA was localized on another pair of autosomes (fig. 8*B*). FISH with probes derived from transposable elements showed that

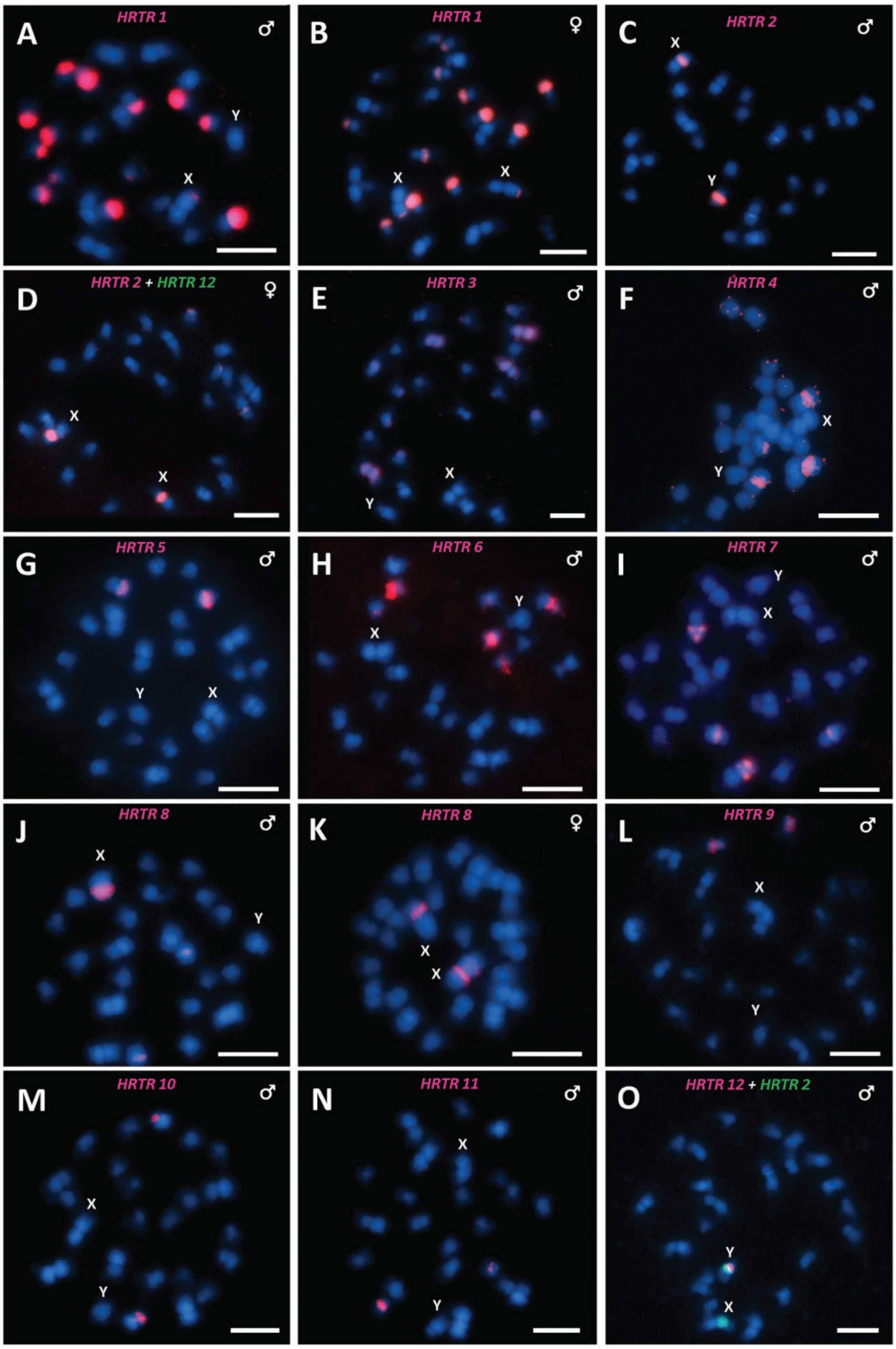


Fig. 6.—Localization of main satellite families on metaphase chromosomes of *Hippophae rhamnoides* using fluorescence *in situ* hybridization. The name of satellite family and sex of individual are indicated inside each figure. Blue are DAPI stained chromosomes, red and green signals show chromosomal localization of satellite families. Bar indicates 5 µm.

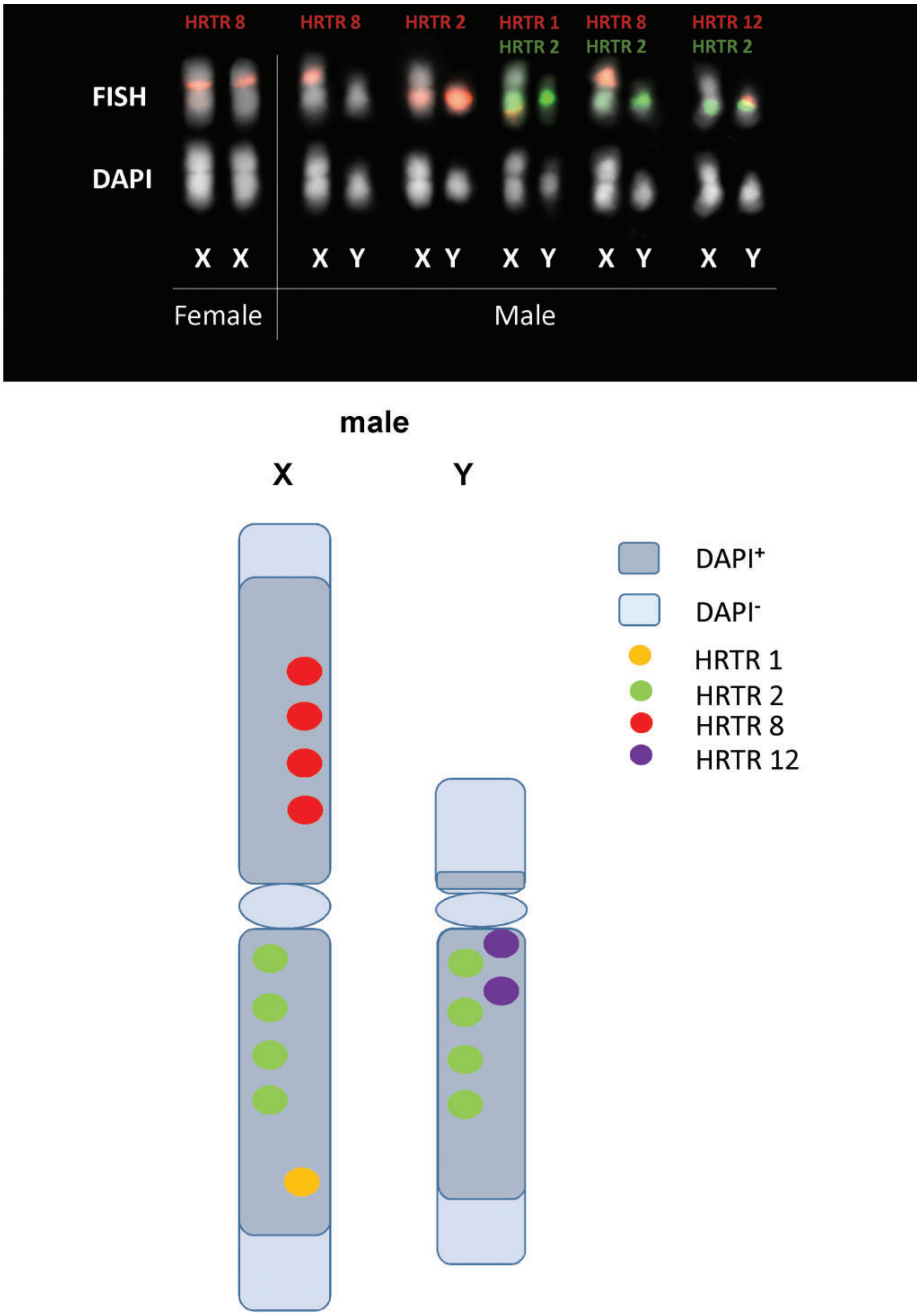


Fig. 7.—FISH and scheme of four satellites on sex chromosomes. The HRTR1, Y-specific HRTR12, X-accumulated HRTR8, sex chromosome-accumulated HRTR2.

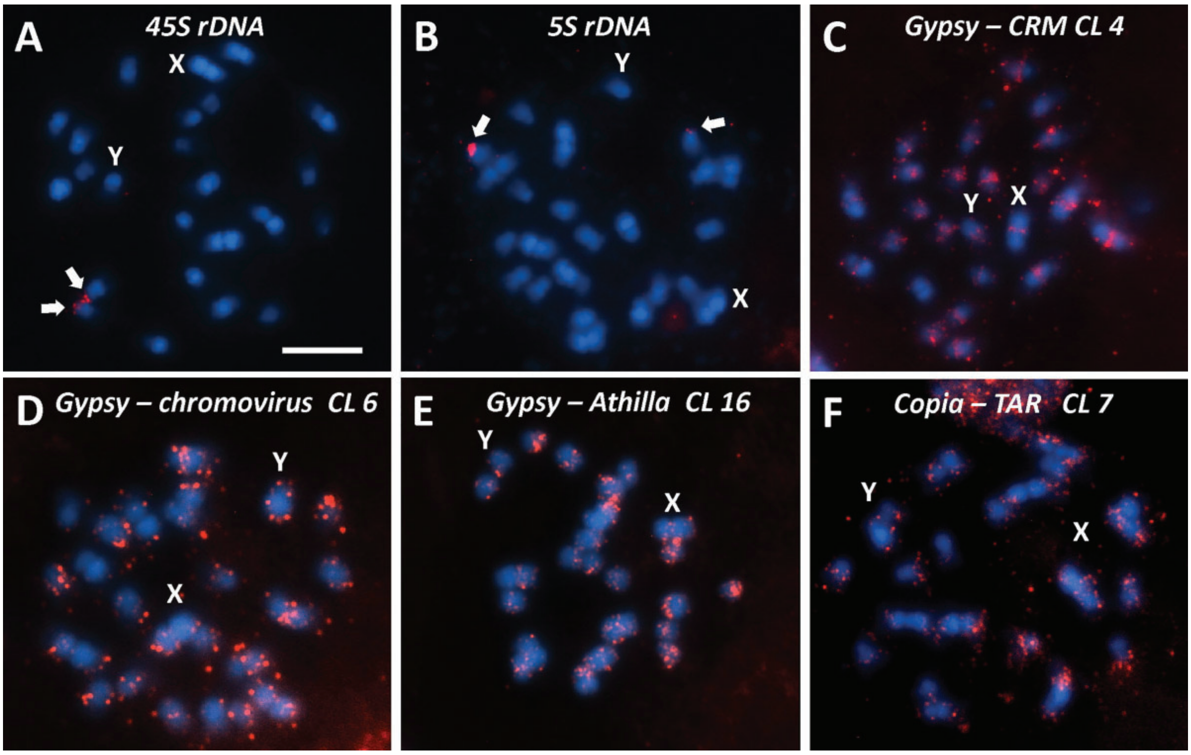


Fig. 8.—Localization of transposable elements and rDNA on metaphase chromosomes of *Hippophae rhamnoides* using fluorescence *in situ* hybridization. The name of transposable element family (together with the number of corresponding cluster) or type of rDNA cluster is inside each figure. Blue are DAPI stained chromosomes, red signal shows chromosomal localization of selected transposable elements and 45S and 5S rDNA. Bar indicates 5 μm.

three of four studied groups of TEs are present mainly in subtelomeres of all chromosomes (fig. 8*D–F*) and only the CRM retroelements (CL4) that was localized in the centromeric region of all chromosomes (fig. 8*C*).

Discussion

We present the first comprehensive analysis of seabuckthorn (*H. rhamnoides*) genome. We found that about one quarter of the genome is composed of TEs and another quarter of satellite DNA which is comparable to other plant genomes. Nevertheless, the seabuckthorn genome contains an unusually large number of different satellites (table 2, 12 main tandem repeats) compared with most other plant genomes (Mehrotra and Goyal 2014). Moreover, some satellites evolve rapidly into new variants. In particular, HRTR2 and HRTR3 satellite superclusters are comprised of a number of smaller clusters where each cluster represents an individual satellite (supplementary fig. S3, Supplementary Material online). Thus, the number of different satellites may be even higher if more strict criteria were used for tandem repeat classification. Transposable elements are represented by all main families of both Ty3/*Gypsy* and Ty1/*Copia* retrotransposons (fig. 2) with chromoviruses (CRM and Galadriel) and TAR families dominating (table 1). Most transposable element families are represented by only one or two clusters indicating their long term presence without changes in sequence or structure. Only Athila, Angela, Tork and Ale/Retrofit retrotransposons are found in multiple clusters (data not shown) suggesting higher divergence. Well preserved long ORFs in some TEs indicate the recent amplification/younger age and low level of degeneration of these elements. All in all, high variability of some satellites and TE families indicate high tempo of their diversification in the seabuckthorn genome, while other repeats remain relatively conserved. Nevertheless, this conclusion should be verified by comparative analysis of at least two closely related species. Recent analysis by Macas et al. (2015) showed that it is not transposable elements but satellites that are the most variable repeats among closely related species of Fabae genus.

Comparison of numbers of male and female reads constituting satellite superclusters, enabled us to predict satellites localized on the Y chromosome, X chromosome, on both sex chromosomes or on autosomes as each specific ratio of abundance of male and female reads in a cluster corresponded to specific chromosomal distribution. Our FISH results showed that this prediction works well in most cases as verified by satellites accumulated on the X chromosome (HRTR8) and both X and Y chromosomes, and specific for the Y chromosome (HRTR12) and for autosomes (HRTR1, 3, 4, 5, 6, and 10). It is a question whether or not the higher number of different satellites in the seabuckthorn genome than in the majority of plant genomes (Mehrotra and Goyal 2014) somehow correlates with the presence of sex chromosomes representing a specific genomic context, each shaped by different evolutionary forces.

The localization of satellites is remarkable and shows that satellites are gathered not only on the nonrecombining region of the Y chromosome but some are specific for the X chromosome or for both sex chromosomes. They are gathered in heterochromatic parts of sex chromosomes what can reflect possible role of satellites in heterochromatinization. The list of chromosomal localization of satellites and TEs in dioecious plants was recently presented by Li et al. (2016a). Although Y chromosome divergence and specific repeat composition is a generally accepted feature, an accumulation of X-specific repeats during plant sex chromosome evolution has been suggested only by limited number of studies (Hobza et al. 2004). As satellites localized on either X or Y chromosomes are mutually different, we prefer the explanation that these satellites originated and expanded on the sex chromosomes long after the X–Y divergence. Therefore, it would be interesting to compare X and Y-linked variants of HRTR2 satellite and, if present, to assess the extent of X- and Y-linked satellite divergence.

The localization of transposable elements mainly in subtelomeres is a feature characteristic of the seabuckthorn genome. However, transposable elements are accumulated in subtelomeres in other plant species too (Zhang and Wessler 2004), and, among dioecious plants, subtelomeric localization was shown in *Retand* retrotransposon in *Silene latifolia* (Kejnovsky et al. 2006). Retrotransposons are found in or around centromeres as well (Miller et al. 1998; Neumann et al. 2011).

Our results clearly confirm the existence of the XY system in seabuckthorn found by Shchapov (1979) and they show that the Y chromosome is small and the X chromosome large. We mention in passing the work of Truta et al. (2011) who initially found a large Y chromosomes and small X chromosome in three Romanian seabuckthorn genotypes that later investigation of Romanian genotypes failed to confirm (Dr. Elena Truta, Institute of Biological Research Iasi, Romania, personal communication, June 15, 2016). Another cytogenetic study on seabuckthorn using C-banding that unfortunately showed only female karyotype without marking sex chromosomes (Rousi and Arohonka 1980).

Estimation of the age of sex chromosomes is not yet possible in this species because no X- and Y-linked genes are known. It remains a question whether the large size difference between X and Y chromosomes, the small size of the Y chromosome and accumulation of different satellites on both sex chromosomes indicates greater age of these sex chromosomes or not. It is remarkable that another genus of the Elaeagnaceae family—Shepherdia (Elaeagnaceae contains three genera—Elaeagnus, Hippophae, and Shepherdia) contains only three species that are all dioecious (Veldkamp 1986). Moreover, the Elaeagnaceae family belongs to the order of Rosales containing other plants with heteromorphic sex chromosomes like Humulus and Cannabis. Although karyotypes were described in Elaeagnus ($2n = 28$ in *E. angustifolia*) and Shepherdia ($2n = 26$ in *S. argentea* and $2n = 22$ in *S. canadensis*), the sex chromosomes were not revealed (Rousi and Arohonka 1980). Therefore, it is not possible to draw conclusions about the formation or age of sex chromosomes during phylogeny.

The small Y chromosome containing several satellite DNA and a large X chromosome revealed in seabuckthorn resemble the mammalian sex chromosomal system. To the best of our knowledge, such a system is very rare among plants. Sex chromosomes in plants are mostly evolutionarily young—e.g., *Silene latifolia* (6 Ma, Kubat et al. 2014), *Rumex acetosa* (12–13 Ma, Navajas-Perez et al. 2005), or *Coccinia grandis* (3 Ma, Sousa et al. 2013)—and only sex chromosomes of *Marchantia polymorpha* are thought to be older (Yamato et al. 2007). A small Y chromosome and the large X chromosome were revealed in *Humulus lupulus* (Shephard et al. 2000; Karlov et al. 2003) and also in gymnosperm species *Cycas revoluta* (Segawa et al. 1971). The small size of the seabuckthorn Y chromosome may be caused by the loss of DNA which indicates that the Y chromosome could be in a shrinkage phase of evolution [reviewed in Hobza et al. (2015)] and thus could represent a rare example of an evolutionarily old plant sex chromosome. This assumption is supported by the FISH results which indicate that the large part of the Y chromosome arm that is homologous to the arm of the X chromosome, carrying HRTR8, was lost (fig. 7).

In this study, we developed and used a new bioinformatics approach for analysis of satellite DNA allowing prediction of satellite monomers, their grouping into clusters corresponding to main satellite families in the genome and visualization of their male/female homogeneity. This enabled prediction of satellite localization with respect to the sex determination system in species studied.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

This work was supported by the Czech Science Foundation (grant P501/12/G090). Access to computing and storage

Appendix B

The slowdown of Y chromosome expansion in dioecious *Silene latifolia* due to DNA loss and male-specific silencing of retrotransposons.

Janka Puterova*, Zdenek Kubat*, Eduard Kejnovsky, Wojciech Jesionek, Jana Cizkova, Boris Vyskot and Roman Hobza

BMC Genomics. 2018, 19:153, **DOI:** 10.1186/s12864-018-4547-7

*These authors contributed equally to this work.

Puterova *et al. BMC Genomics* (2018) 19:153
https://doi.org/10.1186/s12864-018-4547-7

BMC Genomics

RESEARCH ARTICLE Open Access

The slowdown of Y chromosome expansion in dioecious *Silene latifolia* due to DNA loss and male-specific silencing of retrotransposons

Janka Puterova[1,2†], Zdenek Kubat[1*†], Eduard Kejnovsky[1], Wojciech Jesionek[1], Jana Cizkova[3], Boris Vyskot[1] and Roman Hobza[1,3*]

Abstract

Background: The rise and fall of the Y chromosome was demonstrated in animals but plants often possess the large evolutionarily young Y chromosome that is thought has expanded recently. Break-even points dividing expansion and shrinkage phase of plant Y chromosome evolution are still to be determined. To assess the size dynamics of the Y chromosome, we studied intraspecific genome size variation and genome composition of male and female individuals in a dioecious plant *Silene latifolia*, a well-established model for sex-chromosomes evolution.

Results: Our genome size data are the first to demonstrate that regardless of intraspecific genome size variation, Y chromosome has retained its size in *S. latifolia*. Bioinformatics study of genome composition showed that constancy of Y chromosome size was caused by Y chromosome DNA loss and the female-specific proliferation of recently active dominant retrotransposons. We show that several families of retrotransposons have contributed to genome size variation but not to Y chromosome size change.

Conclusions: Our results suggest that the large Y chromosome of *S. latifolia* has slowed down or stopped its expansion. Female-specific proliferation of retrotransposons, enlarging the genome with exception of the Y chromosome, was probably caused by silencing of highly active retrotransposons in males and represents an adaptive mechanism to suppress degenerative processes in the haploid stage. Sex specific silencing of transposons might be widespread in plants but hidden in traditional hermaphroditic model plants.

Keywords: Epigenetics, Genome size, *Silene latifolia*, Transposable elements, Y chromosome

Background

Sex chromosomes evolved independently in plants and animals from a pair of ordinary autosomes. Contrary to animals, only 19 plant species possess well-established sex chromosomes. Most of these species bear large Y chromosomes, suggesting an early expanding stage of sex chromosome evolution [1]. Expansion of mainly non-recombining parts of sex chromosomes is frequently accompanied by accumulation of repetitive sequences. This often results in significant genome size variation among closely related dioecious and non-dioecious (gynodioecious, hermaphroditic) species as was shown in *Silene* [2] and *Asparagus* [3]. Out of all repeats, major contributors to genome size variation present transposable elements (TEs). TEs have been reported as players in sex chromosome size dynamics not only in species with established heteromorphic sex chromosomes such as *Silene latifolia* [4], *Rumex acetosa* [5] and *Coccinia grandis* [6] but also participate in the evolution of the young homomorphic sex chromosome system in *Carica papaya* [7].

S. latifolia (white campion) possesses a well-established sex determination system with the dominant Y chromosome in males. Contrary to the evolutionary old sex chromosomes in humans, *S. latifolia* sex chromosomes

* Correspondence: kubat@ibp.cz; hobza@ibp.cz
†Equal contributors
[1] Department of Plant Developmental Genetics, Institute of Biophysics, Czech Academy of Sciences, Kralovopolska 135, 612 00 Brno, Czech Republic
Full list of author information is available at the end of the article

evolved relatively recently, ca. 6 mya [8]. The nuclear genome of *S. latifolia* is arranged in 11 autosomal pairs and one pair of sex chromosomes. The Y chromosome in *S. latifolia* is the largest chromosome in the entire genome, approximately 1.4 times larger than the X chromosome [9]. Although the *S. latifolia* Y chromosome is not heterochromatinised; it has accumulated a significant number of DNA repeats. It was shown that chloroplast and mitochondrial DNA sequences have been transferred on sex chromosomes in *S. latifolia* [10]. Moreover, some microsatellites [11] and satellites [12, 13] are specifically distributed or accumulated on the Y chromosome in this species. A global survey of all the major types of repeats shows that two antagonistic processes - repeat accumulation and repeat spread suppression - form the Y chromosome in *S. latifola* [8].

Here we compare the global genome composition of several *S. latifolia* ecotypes. We focus on differences in genome size dynamics among the ecotypes at the autosomal and sex chromosome level. We address the following questions: How much the Y chromosome varies among *S. latifolia* populations? Does this variation correlate with genome size? Is the Y chromosome still expanding in *S. latifolia*? Which repetitive elements dominantly contribute to Y chromosome expansion in *S. latifolia*? Are these repetitive elements also the main contributors to genome size expansion?

Methods

Biological material and genome size estimation

S. latifolia seeds of each sex were collected from wild populations across Europe at seven geographical locations (Additional file 1, Additional file 2: Table S1). *S. latifolia* is not protected or endangered species in European countries. Collection of *S. latifolia* seeds comply with national and international guidelines and no permissions were needed. Seeds for all investigated plants were archived and are available upon request at the Institute of Biophysics, Department of Plant Developmental Genetics, Brno, Czech Republic. Plants were grown under greenhouse conditions. Three male and three female individuals were analyzed for each *S. latifolia* accession, and each individual was measured three times on three different days. Nuclear genome size was estimated using flow cytometry according to [14]. Genome size (2C value) was determined considering 1 pg DNA is equal to 0.978×10^9 bp [15] and average genome size of samples from distinct populations is available in Additional file 2: Table S2.

Processing of whole genome sequencing data

The *S. latifolia* genomes were sequenced by Illumina Nextera MiSeq platform using paired-end protocol. For detailed information about sequencing libraries of individual samples see Additional file 2: Table S3. Raw reads were examined and filtered by quality using FastQC [16] and Trimmomatic tool [17]. All 14 datasets were randomly sampled to represent approximately 0.015×/1C (the exact number of reads is shown in Additional file 2: Table S4) and 3,479,090 reads were analyzed altogether. RepeatExplorer pipeline [18, 19] was used for de novo repeat identification. Resulting clusters were characterized based on similarity searches against RepeatMasker libraries, user custom libraries, in blastn and blastx [20]. Reference sequences of main LTR retrotransposon subfamilies presenting in *S. latifolia* genome were collected using assembled contigs published in [21]. Contigs of these LTR retrotransposons were used as queries for megablast [22] searches against nr/nt database with default settings. For significant hits with GenBank database see Additional file 3. In case of significant hits with unannotated GenBank sequences or no hits, contigs were further searched for the presence of protein domains using CD-Search [23] with default settings. Annotated contigs were used as queries to search for similarities against assembled *S. latifolia* bacterial artificial chromosome (BAC) clones using Geneious 8.1.7 software (http://www.geneious.com, [24]), with similarity threshold set to 80%. Full length genomic copies from BACs were manually annotated in Geneious 8.1.7 and aligned using MAFFT v7.017 [25].

TE abundance and copy number estimation

To estimate approximate abundance and copy number of main LTR retrotransposon subfamilies in *S. latifolia*, genomic reads were uniquely mapped onto reference sequences of individual subfamilies using Bowtie 2 v2.3.0 [26]. Coverage of subfamilies was obtained by samtools tool [27] using bedcov utility and copy number for the whole genome was calculated using a formula: *(subfamily coverage [bp]/subfamily_length [bp])*(100/0.75*), where 0.75 represents 0.75% 1C coverage. Density of OgreCL5 subfamily in X chromosomes in comparison to autosomes was estimated according to formula *((F-M)/F)*2/0.15*, where *F* is a copy number of OgreCL5 subfamily in female (2n), *M* is a copy number of OgreCL5 subfamily in male (2n) and 0.15 accounts for genome length of X chromosome [9]. To display changes in copy number of individual LTR retrotransposons subfamilies in ecotypes, a difference between male and female copy number was calculated and illustrated using heatmap (see Additional file 4).

Fluorescence in situ hybridization

Fluorescence in situ hybridization experiments were performed according to [9] with slight modifications. Primers for probe preparation were designed on LTR and GAG or ORF region of selected LTR retrotransposons using Primer3 [28] and are available in Additional file 5. To distinguish Y chromosome arms, X43.1. tandem repeat hybridizing only on the q arm of the Y chromosome has been used [29]. All the above-mentioned procedures and methods were conducted as thoroughly described in Additional file 6.

Results

Genome size varies more than Y chromosome size in *S. latifolia* ecotypes

In order to assess possible intraspecific genome and Y chromosome size variation in *S. latifolia*, male and female genome size in seven distinct ecotypes from central and southern Europe was measured using flow cytometry. Map with the locations of sample collection is depicted in Additional file 1. As shown in Fig. 1a, genome size varies substantially among ecotypes and is always larger in males than females. Male genome sizes vary between 5.90 ± 0.01 pg/2C and 6.31 ± 0.02 pg/2C while female genomes are in the range 5.69 ± 0.02 pg/2C and 6.09 ± 0.01 pg/2C representing 1.07-fold variation in genome size. The excessiveness of male genomes over female genomes (Fig. 1a) reflects the enormous size of the Y chromosome, which is approximately 1.4 times larger than the X [9]. Nevertheless, the proportion of the Y chromosome tends to be in negative correlation with whole genome size (Fig. 1b) which indicates that genome size variation among *S. latifolia* ecotypes is caused predominantly by processes taking place on autosomes and X chromosomes.

Genome composition

To decipher how individual repeat types contribute to genome size, whole genome shotgun sequencing was performed on males and females of seven ecotypes using Illumina MiSeq platform generating raw 300 bp long paired-end reads. The reads were analyzed by RepeatExplorer [18, 19] as specified in Materials and Methods. The global repeat composition is summarized in Table 1. LTR (Long Terminal Repeat) retrotransposons represented the major fraction of all analyzed genomes, comprising of up to 70% of nuclear DNA. They were mostly represented by *Ty3/Gypsy*-like elements (~ 50%), while *Ty1/Copia*-like elements represented roughly 20% in all genomes. Non-LTR retrotransposons and DNA transposons were much less abundant and occupied ~ 0.3 and ~ 3.3% of genomes, respectively. Tandem repeats formed clusters with a small number of reads in our analysis, and thus they might not present a significant portion of studied genomes.

Correlation between repeat abundance and genome size increase uncovered active repeats contributing to recent genome size variation

To identify recently active repeats, a correlation between repeat amount (obtained using RepeatExplorer tool) and genome size of both sexes was assessed across ecotypes. Figure 1c shows that most repeat types are positively correlated with genome size, but only some could be considered as statistically significant (marked with asterisks). This might reflect either different behavior of repeats in distinct ecotypes or conflicting effects of divergent lineages within respective repeat families. Therefore, the effect of particular LTR retrotransposon subfamilies was also assessed (Fig. 1d). The nine largest LTR retrotransposon subfamilies, previously classified in [21] were analyzed in detail. It was found that each subfamily has a specific behavioral pattern not necessarily identical to the whole family (Fig. 1c). Out of three Ogre subfamilies, OgreCL5 was found to be positively correlated while OgreCL11 was negatively correlated with the genome size (Fig. 1d). Overall, correlation analysis disclosed repeats influencing genome size variability across all ecotypes in a positive manner (AngelaCL1, AthilaCL3, OgreCL5, Caulimoviridae, and Helitrons) as well as in a negative manner (TekayCL4, OgreCL11). These repeats represent transpositionally active and silent TEs, respectively. Nevertheless, other TEs might also contribute to genome size variation but their activity differs in individual ecotypes. Another noteworthy finding is that correlation is not always similar for males and females as exemplified by AthilaCL3, OgreCL5, Chromoviruses and TAR elements showing positive correlation in females but lower or even negative correlation in males (Fig. 1c and d). This indicates higher insertional activity of mentioned TEs in the female genome (autosomes and X chromosomes), i.e. low insertional activity into Y chromosome. In contrast, only AngelaCL7 and minor TE families, LINE and Caulimoviridae, have higher insertional activity on the Y chromosome.

Most of the retrotransposons are depleted on the Y chromosome

To assess the potential impact of individual LTR retrotransposon subfamilies on genome size, their copy number was estimated in all samples (Fig. 1e). The copy numbers were plotted against genome size to assess two key behavioral features of studied LTR retrotransposons; change of an LTR retrotransposon copy number towards bigger genomes (Fig. 1e, dashed lines), and relative abundance of a retrotransposon in males in comparison to females (Fig. 1e, solid colored lines). Due to a negligible genomic proportion of endogenous retroviruses and DNA transposons, only LTR retrotransposons were examined. Figure 1e shows scenarios of TEs behavior. Steeply increasing copy numbers of AngelaCL1, OgreCL5 and AthilaCL10 suggest that these LTR retrotransposons are main genome size drivers in most ecotypes (dashed lines). In contrast, TekayCL4, OgreCL6, and OgreCL11 show low or no insertional activity as implied from decreasing quantity of their genomic copies. However, most of the LTR retrotransposons show to some extent variable transposition in individual ecotypes.

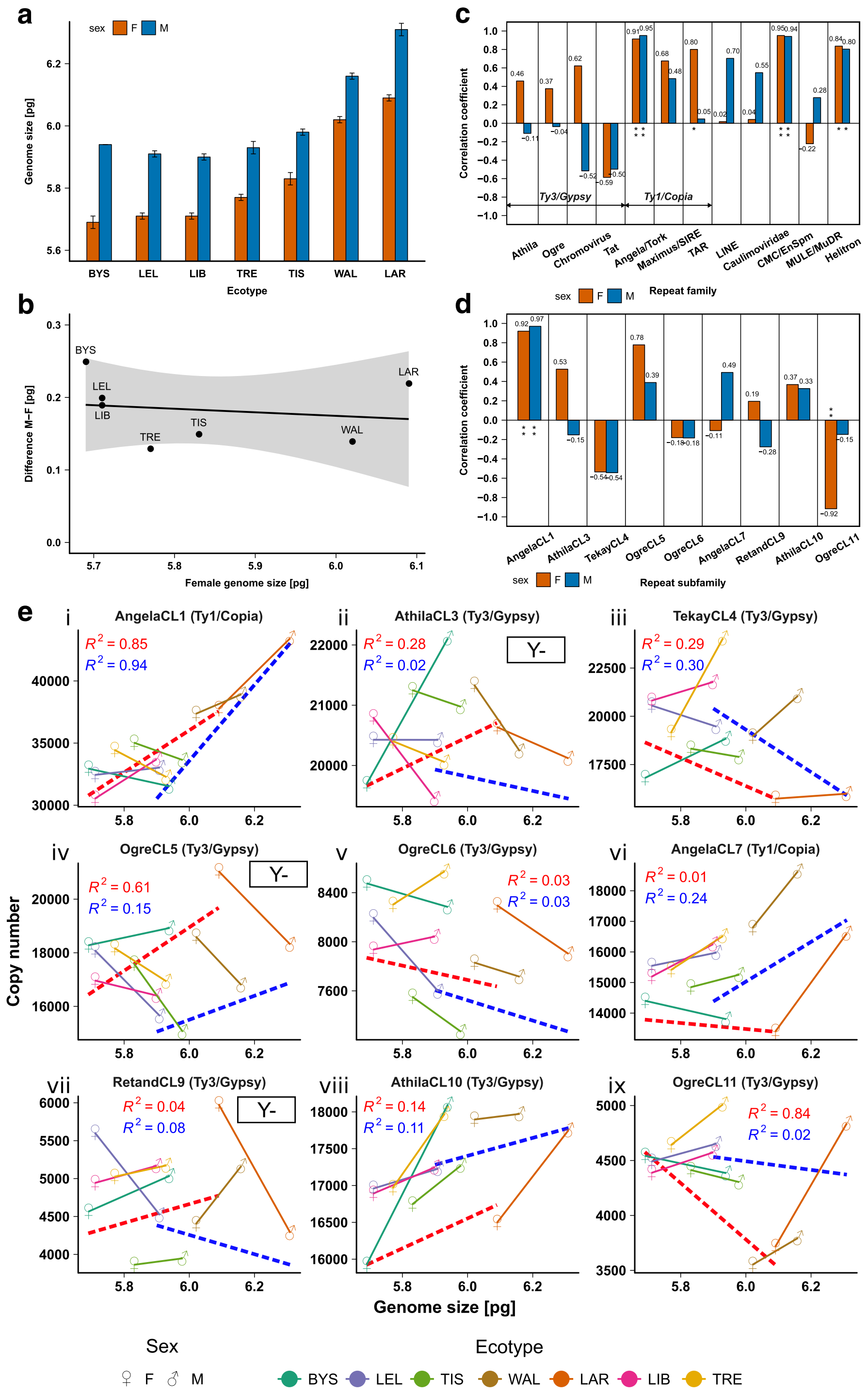


Fig. 1 (See legend on next page.)

(See figure on previous page.)

Fig. 1 Genome size and composition of *Silene latifolia* ecotypes. **a** Genome sizes of *S. latifolia* male and female genome from eight distinct ecotypes measured by flow-cytometry. Genome size varies from 5.90 pg (LIB) to 6.31 pg (LAR) in males and 5.69 pg (BYS) to 6.09 pg (LAR) in females. Error bars represent SEM. **b** Difference in genome size between sexes caused by Y chromosome. Difference was calculated using a formula: *(M-F)/F*, where *M* corresponds to male genome size and *F* to female genome size. It varies between 2.24% (WAL) and 4.32% (BYS). Black line represents linear regression line of plotted data. Grey area displays 95% confidence interval. **c** Correlation between abundance of repeat families and genome size of both sexes in *S. latifolia*. Correlation coefficient represents Pearson correlation coefficient, n (number of samples) = 7, degrees of freedom = 5. **d** Correlation between abundance of main LTR retrotransposon subfamilies and genome size of both sexes in *S. latifolia*. Correlation coefficient represents Pearson correlation coefficient, n (number of samples) = 7, degrees of freedom = 5. **e** Detailed contribution (copy number vs. genome size) of main LTR retrotransposons to genome size in both sexes. Dashed lines correspond to linear regression between female genome size and element's copy number (red), and male genome size and element's copy number (blue). R^2 represents coefficient of determination (square of the Pearson correlation coefficient), n (number of samples) = 7, degrees of freedom = 5

Remarkably, most of the TEs differ in their abundance in male and female genomes (Fig. 1e, solid colored lines). Based on the fact that male genomes are ~ 4% larger than female genomes, slightly more TE copies are expected in males. However, most retrotransposons show even larger deviation from this expectation towards both directions. While some TEs are significantly more abundant in males (AngelaCL7, AthilaCL10), other TEs are significantly less abundant in male than female genome (AthilaCL3, OgreCL5). The former case indicates accumulation of TEs on the Y chromosome due to either reduced loss of DNA on the Y chromosome or higher activity of TEs in males. The latter case suggests the exact opposite; lower density of retrotransposon insertions on the Y chromosome than in the rest of the genome, which might be a consequence of either accelerated loss of DNA on the non-recombining Y chromosome or lower activity of retrotransposons in males. Unequal distribution of TEs on sex chromosomes assessed by a bioinformatics approach is in concordance with fluorescence in situ hybridization (FISH) experiments summarized in Table 2. For TEs with no published cytogenetic data available, we performed FISH on meiotic chromosomes of TIS ecotype (Fig. 2). Nevertheless, in specific cases, LTR retrotransposons differ in their behavior among ecotypes, as exemplified by AngelaCL1 which is underrepresented on Y chromosomes of all ecotypes except WAL and LAR (Fig. 1e (i)).

To decipher the likely role of low Y diversity [30] in Y chromosome size constancy we constructed a copy number variability graph in male and female genomes (Additional file 4). The copy number values are adopted from Fig. 1e. The graph displays higher variability of TE copy numbers in males for the most abundant TE families. This additional copy number variability is driven by Y-linked TE copies and indicates that Y chromosome of each ecotype has unique repeat composition.

The most active LTR retrotransposons preferentially proliferate in females

The conspicuous case among all repeats is LTR retrotransposon subfamily OgreCL5 which is virtually absent on the Y chromosome [8]. OgreCL5 is still an active element in all ecotypes as suggested by Fig. 1e (iv) and may be one of the dominant players in genome size variation among all *S. latifolia* ecotypes studied. An earlier publication proposed that OgreCL5 proliferates transgenerationally only in the female lineage [8]. This hypothesis was tested by estimating the density of OgreCL5 elements in X chromosomes in comparison with autosomes according to the formula $((F\text{-}M)/F) \times 2/\ 0.15$ where F is a TE copy number in female (2C), M is a TE copy number in male (2C), and X chromosome accounts for 15% of genome length [9]. Since X chromosomes spend $^2/_3$ of their lifetime in females, while autosomes only $^1/_2$, the probability of insertion into the X chromosome for TE proliferating in females only is 1.33 times higher than into an autosome. In ecotypes LEL, TIS, WAL and LAR, X chromosome contains roughly 20–30% of all genomic OgreCL5 copies, 1.3–2 times more than an average autosome supporting the idea that OgreCL5 spreads preferentially in females and not in males. The computation is approximate due to the presence of a low but unknown number of OgreCL5 copies on the Y chromosome (mainly in pseudoautosomal region), thus it is slightly different from a theoretical value of 1.33. Because other retrotransposons with similar chromosomal pattern have even more Y-linked copies according to FISH experiments, the computation cannot be used for their copy number estimation – resulting copy number of X-linked TE copies would be undervalued in that case. Figure 1e and results of previous publications [4, 31, 32] examining the chromosomal localization of repeats (Table 2) suggest that at least *Ty3/Gypsy* LTR retrotransposons AthilaCL3, OgreCL6, and RetandCL9 also spread predominantly through female lineage but their recent retrotransposition activity is rather low in most ecotypes.

Discussion

We have shown here that regardless of intraspecific genome size variation, the Y chromosome size is similar in European *S. latifolia* populations. Since *S. latifolia* is thought to have found refuge in North Africa during the

Table 1 Transposable element composition of *Silene latifolia* genome

Classification			Genome proportion [%]													
Repeat Type	Superfamily	Family	BYS-fe	BYS-ma	LEL-fe	LEL-ma	LIB-fe	LIB-ma	TRE-fe	TRE-ma	TIS-fe	TIS-ma	WAL-fe	WAL-ma	LAR-fe	LAR-ma
LTR retrotransposons	Gypsy	Athila	17.54	18.79	18.18	18.53	18.37	17.90	18.08	18.61	18.00	18.31	18.37	17.88	16.90	17.09
		Ogre	17.10	16.61	17.22	15.53	16.28	16.11	16.96	16.80	15.93	14.54	15.87	15.00	16.67	15.21
		Chromovirus	11.28	12.86	12.63	12.45	12.75	12.90	12.17	13.62	12.19	11.94	12.44	12.54	12.13	10.94
		Tat	3.25	3.12	3.78	3.38	3.52	3.67	3.44	3.86	2.98	3.00	3.07	3.43	2.93	2.67
		sum	49.17	51.38	51.82	49.89	50.92	50.57	50.66	52.89	49.11	47.79	49.75	48.85	48.62	45.90
	Copia	Angela/Tork	16.51	16.09	16.18	16.92	16.36	16.45	16.92	15.88	18.08	17.91	17.97	18.61	17.65	19.49
		Maximus/SIRE	2.56	2.02	2.31	2.55	2.45	2.39	2.55	2.26	2.49	2.65	2.50	2.38	2.45	2.50
		TAR	0.22	0.32	0.20	0.23	0.24	0.22	0.19	0.22	0.21	0.23	0.24	0.25	0.29	0.22
		sum	19.29	18.42	18.69	19.71	19.05	19.07	19.66	18.36	20.78	20.80	20.72	21.24	20.39	22.22
Non-LTR retrotransposons	LINE		0.27	0.24	0.22	0.21	0.20	0.25	0.22	0.23	0.22	0.23	0.19	0.22	0.24	0.27
	Caulimoviridae		0.07	0.08	0.06	0.08	0.08	0.08	0.09	0.05	0.06	0.07	0.05	0.06	0.09	0.11
DNA transposons	CMC-EnSpm		1.70	1.63	1.66	1.67	1.67	1.70	1.74	1.63	1.69	1.69	1.80	1.75	1.76	1.76
	MULE-MuDR		1.65	1.33	1.45	1.56	1.48	1.50	1.52	1.51	1.44	1.55	1.48	1.54	1.19	1.42
	Helitron		0.18	0.20	0.16	0.19	0.16	0.18	0.16	0.16	0.18	0.19	0.18	0.19	0.25	0.23
		sum	3.23	3.16	3.27	3.42	3.31	3.38	3.42	3.30	3.30	3.44	3.46	3.48	3.41	3.41
Total TEs			71.69	72.96	73.78	73.02	73.28	73.02	73.74	74.54	73.19	72.03	73.92	73.57	72.21	71.53

Genome proportions of transposable element superfamilies and families in percentage for male and female from seven ecotypes

Table 2 Chromosomal distribution of retrotransposons with special emphasis on sex chromosomes revealed by fluorescence in situ hybridization (FISH) experiments

Subfamily	FISH	Citation
Ty1/Copia/AngelaCL1	Y-, X+	Fig. 2
Ty3/Gypsy/AthilaCL3	Y-, X+	Kralova et al., 2014
Ty3/Gypsy/TekayCL4	homogeneous	Fig. 2
Ty3/Gypsy/OgreCL5	Y-, X+	Kubat et al., 2014
Ty3/Gypsy/OgreCL6	Y- (slightly), X+ (slightly)	Kubat et al., 2014
Ty1/Copia/AngelaCL7	Y+, X-	Fig. 2
Ty3/Gypsy/RetandCL9	Y- (slightly), X+ (slightly)	Kejnovsky et al., 2006
Ty3/Gypsy/AthilaCL10	homogeneous	Kralova et al., 2014
Ty3/Gypsy/OgreCL11	homogeneous	Kubat et al., 2014

X+, Y+, the retrotransposon shows stronger hybridizing signal on the X and Y chromosome than on autosomes, respectively; X-, Y-, the retrotransposon shows weaker hybridizing signal on the X and Y chromosome in comparison to autosomes, respectively

last glaciations and to colonize its current range with the spread of agriculture [33, 34], the diversification of genome size is probably of recent origin. Unanswered questions remain: what is the ancestral state and what this variability of genomic sizes represents; are we observing rather expansion or reduction of genomes, or a combination of both phenomena here? If there is selective pressure to reduce the genome, there is no reason why X chromosome and autosomes should lose DNA faster than the largely heterochromatic (unpublished data) and genetically degrading non-recombining Y chromosome [35–38], which has lost 30% Y-linked genes [39, 40] and its diversity is reduced most likely due to strong selection against deleterious mutations [30]. Moreover, the genome of closely related *S. vulgaris* without sex-chromosomes is 2.7-fold smaller (see Plant DNA C-value Database, http://data.kew.org/cvalues/) indicating relatively recent genome expansion in *S. latifolia*. Thus, *S. latifolia* genome enlargement most probably continues as previously proven by [2] and also observed in other dioecious species [41], but at a various tempo in distinct populations. 1.07-fold variation in female genome size (Fig. 1a) indicates rapid genome size changes. And, importantly, the Y chromosome most likely contributes to genome size increase less than the rest of chromosomes.

This is in contradiction with existing assumptions that the evolutionarily recent Y chromosome (about 6 million years, [8]) is still in the expansion phase of evolution [1]. Extreme Y chromosome size [6, 42], gene degeneration [36, 43] and high content of repetitive sequences such as microsatellites [44], mobile elements and tandem repeats [4, 21, 45] and recent insertions of chloroplast DNA [46] as well as increased fixation of transposons on the Y chromosome in comparison to X and autosomes [47] illustrate the low efficiency of repair mechanisms requiring recombination.

The first possible explanation of almost constant Y chromosome size arises from low Y diversity [30, 35, 48, 49] caused most likely by selection against Y chromosomes with damaged essential genes [50] and by a selective sweep. Background selection and within-population hitch-hiking processes may lead to fixation of Y chromosomes with lower TE content that are now present across all populations. This

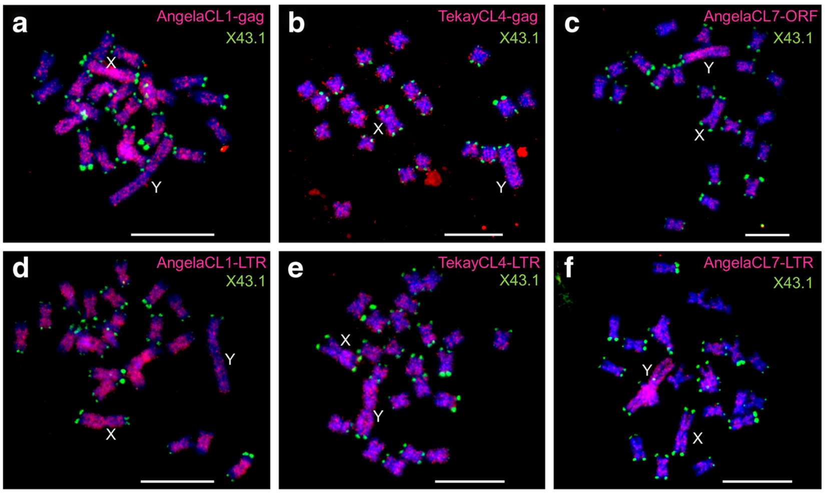


Fig. 2 Localization of LTR retrotransposons on mitotic metaphase chromosomes of male *Silene latifolia* (Tišnov population) using fluorescence in situ hybridization (FISH). **a** AngelaCL1 gag and (**d**) LTR probe, (**b**) TekayCL4 gag and (**e**) LTR probe, (**c**) AngelaCL7 ORF and (**f**) LTR probe. Chromosomes were counterstained with DAPI (blue), LTR retrotransposon probes are represented by red signals, the tandem repeat X43.1 (green) labels most chromosomal subtelomeres, but only q-arm of the Y chromosome. Bars indicate 10 μm

is consistent with fixation of MITE copies on the Y chromosome of many European populations [47] and also with the fact that the Y chromosome effective population size is much smaller than that of X and autosomes [51, 52]. In this scenario, all Y chromosomes have to be homomorphic across populations not only on genic level but also in other sites as are in TE insertions. The latter condition is not met in case of *S. latifolia*. We constructed a copy number variability graph for TE families in male and female genomes (Additional file 4). The graph shows higher copy number variability of some TE families in male than female genomes across populations. The additional variability in male TE copy numbers is caused by TEs present on the Y chromosomes. This suggests that the Y chromosomes are polymorphic in TE composition, at least in case of the most abundant TE families. The genetic uniformity and reduced effective population size (at genic level) would be remnants of the last common ancestor, but in terms of TE content the Y chromosomes evolve independently since the subdivision of studied populations after the last glaciation.

The second hypothesis says that the slowdown of Y expansion is due to the increasing prevalence of deletion loss of non-recombining parts of the Y chromosome over the accumulation of repeats. This is consistent with massive loss of genes on the Y chromosome [39, 40]. Although this hypothesis seems to be likely, our data also favor an additional explanation that retrotransposons tend to spread more in the maternal line than in the paternal, resulting in a low frequency of insertions into the Y chromosome and its lack of growth over the rest of the genome. This phenomenon was initially observed by cytogenetic analyses when it was found that several LTR retrotransposons show a lower hybridization signal on the Y chromosome of *S. latifolia* [4, 8, 32, 53] and *R. acetosa* [5].

Whether the loss of DNA on the Y or male-specific silencing of TEs dominates is difficult to determine without comparisons of high quality reference genomes. Nevertheless, previous works confirmed that there is a number of active TEs in *Silene*, some of them with sex-specific mode of spread. For example, all Ogre elements, OgreCL5 absent on the Y chromosome as well as OgreCL6 and OgreCL11 present on the Y chromosome, peaked their retrotransposition activity after Y chromosome formation [8, 53]. This indicates rather male specific silencing of OgreCL5 than selective removal of this retrotransposon family from the Y. Several tens of thousands to 1 million years old TE insertions were also documented in X- and Y-linked BACs [45]. Moreover, some retrotransposons, especially *Ty1/Copia* group (AngelaCL7), recently accumulated on the Y chromosome (Fig. 1d, e (vi); Fig. 2c, f; [4]). Altogether, these facts suggest simultaneous activity of both TE types: dominating LTR retrotransposons that do not insert into the Y chromosome as well as LTR retrotransposons that contribute to Y chromosome enlargement, but not sufficiently to keep pace with the rest of the genome. Thus, the restricted expansion of the Y chromosome is likely caused by combination of both factors: (i) insertion of active LTR retrotransposons apart from the Y chromosome and (ii) deletion loss of DNA that to some extent compensates for the activity of transposons incorporating to the Y chromosome.

As noted above, high-quality *S. latifolia* reference genome sequence should enable us to obtain more rigorous evidence for TE activity within certain chromosomal regions, such as TE insertions age, location, and copy number. Unfortunately, only not-enough representative partial sequencing data (e. g. BAC clones or partially reconstructed genic sequences) are available so far. Moreover, only very complete reference genome sequence with high-quality assembly of TE islands can address all questions regarding TE age distribution and copy number. Thus, we believe that our approach based on a combination of FISH and TE copy number estimation from whole genome sequencing datasets obtained from several populations is sufficient for the conclusions.

Our bioinformatics and FISH analyses show that LTR retrotransposons follow one of three behavioral patterns: (i) LTR retrotransposons of the first group spread equally in all chromosomes and are represented by TekayCL4. (ii) The second group spreads preferentially in a female genome, which is manifested by their lower proportion on the Y chromosome and higher proportion on the X chromosome compared to autosomes (as a consequence of X chromosome spending $^2/_3$ of its existence in females, but only $^1/_3$ in males). This group exhibits a large variability. There are elements almost totally missing on the Y chromosome as well as elements only slightly underrepresented on the Y chromosome. The group is represented mostly by *Ty3/Gypsy* LTR retrotransposons, for instance, AthilaCL3, OgreCL5, and RetandCL9. (iii) LTR retrotransposons of the third group accumulate on the Y chromosome and have a lower copy number on the X chromosome than on autosomes, they spread predominantly in males and are represented by two smaller LTR retrotransposon families, AngelaCL7 and AthilaCL10. A unique case is AngelaCL1, which is accumulated on X chromosomes of most ecotypes but reveals Y chromosome accumulation in the southern European Larzac ecotype. This indicates not negligible degree of freedom in how a TE behaves in certain genetic background. All three behavioral patterns are also observable in *R. acetosa* [5].

A major question is whether the sex-dependent retrotransposition is specific for dioecious plants, or it is a common feature of retrotransposons in angiosperms? The second closely related question that resonates is how can retrotransposons be active preferentially in either male or female genome? To our knowledge, only a few cases of sex-specific retrotransposition have been documented in model plants, so far. Activated LTR retrotransposons EVADE (EVD) expand only if transmitted through the paternal germline but are epigenetically suppressed in female flowers of *Arabidopsis thaliana* [54]. Such retrotransposon regulation would result in accumulation on the Y chromosome in the dioecious system with XY sex-chromosomes. In contrast, OgreCL5 LTR retrotransposons absent on the Y chromosome of dioecious *S. latifolia* were shown to be most probably silenced during pollen grain development also by the epigenetic mechanism [8]. It has been suggested that TEs take advantage of temporal lack of epigenetic silencing during plant gametogenesis for their transposition [55, 56] but plants possess defensive mechanisms based on siRNA production in companion cells of plant gametes [57–60]. Nevertheless, epigenetic regulation is in current view a complex array of mutually interconnected pathways sharing signal molecules (siRNAs, lncRNAs) as well as proteins and enzymes (reviewed in [61, 62]). Thus, the way of certain TE silencing might be strongly individualized, which results in diverse chromosomal distribution of TEs in dioecious plants.

Another extremely important factor influencing TE silencing and activity is its position in the genome: near a gene, within a gene, in a TE island or at the centromere core (reviewed in [63]). In maize, TEs located near genes are subject of intensive RNA directed *de-novo* DNA methylation (RdDM), while TEs in intergenic regions remain densely condensed and heterochromatinized and show very low transcriptional activity, siRNA production and association with RdDM [64–66]. Unlike *Arabidopsis*, in large plant genomes, the near-gene RdDM activity may be critical for creating a boundary that prevents the spread of open, active chromatin to adjacent transposons [67]. Thus, proximity to genes is a major factor inducing RdDM, regardless of transposon sequence or identity, and is more associated with DNA transposons that tend to insert near genes and with short low-copy number retrotransposons than with long high-copy number LTR retrotransposons [64–66]. Therefore, long high-copy number LTR retrotransposons, that play a dominant role in genome expansion, are not likely target of RdDM but rather post-transcriptionally silenced by other small RNA based mechanisms. Several recent publications suggest that male reproductive organs adopted unique epigenetic pathways that utilize micro RNAs and tRNAs for efficient post-transcriptional silencing of TEs in pollen grains [60, 68]. Particularly tRNAs derived small RNAs were proved to target mainly *Ty3/Gypsy* LTR retrotransposons, which are dominant TEs in dioecious plants. Thus, the male germline might possess a reinforced epigenetic barrier against TE transposition compared to egg cell. The male-specific silencing of highly active retrotransposons might be an adaptive mechanism to retain genes essential for haploid pollen tube growth. In dioecious species, it would slow down genetic degeneration of Y-linked genes in addition to haploid purifying selection previously confirmed in *S. latifolia* [50]. A growing body of evidence indicates that male and female gamete formation is accompanied with differently efficient TE silencing mechanisms, what leads to diversity of TE ability to proliferate preferentially through either male or female lineage and subsequently to sex-chromosome specific distribution of TEs.

Conclusions

Taken together, based on a combination of genome size estimation, repetitive DNA assembly, and analysis at the population level, we show that Y chromosome expansion has already peaked in *S. latifolia*. Our data suggest that first stage of sex chromosome evolution accompanied with Y chromosome expansion might present a relatively short period in raise and fall of sex chromosomes, since *S. latifolia* Y chromosome, in contrast to the human Y chromosome, is only partially degenerated. For a more complex view, genetic and genomic analysis should be combined in future experiments.

Additional files

Additional file 1: Map with highlighted geographical locations where samples of wild *S. latifolia* plants were collected. Google is acknowledged for providing the map under fair use principles. (JPEG 291 kb)

Additional file 2: Title: Information about analyzed data. **Table S1** Geographical locations of wild *S. latifolia* populations used in this study. **Table S2** Genome size of individual samples estimated by flow cytometry. **Table S3** Detailed information about sequencing libraries of individual samples. **Table S4** Number of preprocessed reads used in analyses. (XLSX 13 kb)

Additional file 3: Information about studied LTR retrotransposons. (XLSX 8 kb)

Additional file 4: Plot displaying copy number variability of individual LTR retrotransposons between male and female genome in studied ecotypes. Values are adopted from the Fig. 1e. If Y-linked TE copy number is fixed, the copy number variability has to be lower in males than females. Equal or higher variability in males is clear sign of TE copy number variability on Y chromosomes. The figure suggests that Y chromosomes from distinct populations are highly polymorphic in TE content. (PDF 42 kb)

Additional file 5: Primers used for fluorescent in situ hybridization (FISH). (XLSX 8 kb)

Additional file 6: Detailed description of methods. (DOCX 41 kb)

Abbreviations

BAC: Bacterial artificial chromosome; CD-Search: Conserved domain search; DNA: Deoxyribonucleic acid; FISH: Fluorescence in situ hybridization; lncRNA: Long non-coding RNA; LTR: Long terminal repeat; ORF: Open reading frame; RdDM: RNA-directed DNA methylation; siRNA: Small interfering RNA; TE: Transposable element; tRNA: Transfer ribonucleic acid

Acknowledgements

Access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum, provided under the programme "Projects of Large Infrastructure for Research,

Appendix C

Fundamentally different repetitive element composition of sex chromosomes in *Rumex acetosa.*

Wojciech Jesionek, Markéta Bodláková, Zdeněk Kubát, Radim Čegan, Boris Vyskot, Jan Vrána, Jan Šafář, **Janka Puterová**, Roman Hobza

Annals of Botany. 2021, 127(1), **DOI:** 10.1093/aob/mcaa160

Annals of Botany **127**: 33–47, 2021
doi: 10.1093/aob/mcaa160, available online at www.academic.oup.com/aob

ANNALS OF BOTANY Founded 1887

PART OF A HIGHLIGHT ON GENOMIC EVOLUTION

Fundamentally different repetitive element composition of sex chromosomes in *Rumex acetosa*

Wojciech Jesionek[1,4,*], Markéta Bodláková[1], Zdeněk Kubát[1], Radim Čegan[1], Boris Vyskot[1], Jan Vrána[2], Jan Šafář[2], Janka Puterova[1,3] and Roman Hobza[1,*]

[1]Department of Plant Developmental Genetics, The Czech Academy of Sciences, Institute of Biophysics, Královopolská 135, 61200 Brno, Czech Republic, [2]Institute of Experimental Botany, Centre of the Region Haná for Biotechnological and Agricultural Research, Šlechtitelů 31, 78371 Olomouc-Holice, Czech Republic, [3]Brno University of Technology, Faculty of Information Technology, Centre of Excellence IT4Innovations, Bozetechova 2, 61266 Brno, Czech Republic and [4]Department of Experimental Biology, Faculty of Science, Masaryk University, Kamenice 5, 62500 Brno, Czech Republic
**For correspondence. E-mail: hobza@ibp.cz or wjesionek@ibp.cz*

Received: 29 August 2020 Returned for revision: 28 July 2020 Editorial decision: 27 August 2020 Accepted: 31 August 2020
Electronically published: 9 September 2020

- **Background and Aims:** Dioecious species with well-established sex chromosomes are rare in the plant kingdom. Most sex chromosomes increase in size but no comprehensive analysis of the kind of sequences that drive this expansion has been presented. Here we analyse sex chromosome structure in common sorrel (*Rumex acetosa*), a dioecious plant with XY_1Y_2 sex determination, and we provide the first chromosome-specific repeatome analysis for a plant species possessing sex chromosomes.
- **Methods:** We flow-sorted and separately sequenced sex chromosomes and autosomes in *R. acetosa* using the two-dimensional fluorescence *in situ* hybridization in suspension (FISHIS) method and Illumina sequencing. We identified and quantified individual repeats using RepeatExplorer, Tandem Repeat Finder and the Tandem Repeats Analysis Program. We employed fluorescence *in situ* hybridization (FISH) to analyse the chromosomal localization of satellites and transposons.
- **Key Results:** We identified a number of novel satellites, which have, in a fashion similar to previously known satellites, significantly expanded on the Y chromosome but not as much on the X or on autosomes. Additionally, the size increase of Y chromosomes is caused by non-long terminal repeat (LTR) and LTR retrotransposons, while only the latter contribute to the enlargement of the X chromosome. However, the X chromosome is populated by different LTR retrotransposon lineages than those on Y chromosomes.
- **Conclusions:** The X and Y chromosomes have significantly diverged in terms of repeat composition. The lack of recombination probably contributed to the expansion of diverse satellites and microsatellites and faster fixation of newly inserted transposable elements (TEs) on the Y chromosomes. In addition, the X and Y chromosomes, despite similar total counts of TEs, differ significantly in the representation of individual TE lineages, which indicates that transposons proliferate preferentially in either the paternal or the maternal lineage.

Key words: *Rumex acetosa*, sex chromosomes, genome dynamics, transposable elements, satellites.

INTRODUCTION

The formation of sex chromosomes from a pair of ordinary autosomes is repeatedly associated with recombination restriction and the subsequent expansion of a non-recombining region in the vicinity of the sex-determining gene(s) (Vyskot and Hobza, 2004; Ming *et al.*, 2011). In some cases, the non-recombining region extended along most of the sex chromosome with the exception of a small pseudoautosomal region (PAR). Why suppressed recombination of sex chromosomes evolved is the subject of numerous theoretical studies, but experimental findings remain ambiguous and point to a role of species-specific features and changeable ecological conditions, e.g. mating system, dissimilarity of sexual roles, fluctuating selection regimes and population sizes (Ponnikas *et al.*, 2018; Charlesworth, 2019). The best-substantiated explanation is that recombination cessation evolved because selection favours linkage between sex-determining and sexually antagonistic genes (B. Charlesworth and D. Charlesworth, 1978; D. Charlesworth and B. Charlesworth, 1980; Rice, 1984, 1987). Other proposed hypotheses consider meiotic drive (Jaenike, 2001; Kozielska *et al.*, 2010; Ubeda *et al.*, 2015), heterozygote advantage (de Waal Malefijt and Charlesworth, 1979; Charlesworth and Wall, 1999) and genetic drift (Lande, 1979, 1985; Charlesworth *et al.*, 1987), reviewed in Ponnikas *et al.* (2018). Nevertheless, the evolution of a large non-recombining region is not a rule. In most amphibians and some other poikilothermic vertebrates, the sex-determining gene is not conserved and can be rapidly replaced by another gene on a different chromosome in a process called turnover of sex-determining genes and sex chromosomes (Schmid *et al.*, 1991; Eggert, 2004; Schartl, 2004; Miura, 2017). This can prevent the formation of non-recombining regions. Similarly in plants, out of the 5 % of flowering species that contain individuals with separate sexes (D. Charlesworth, 2016), morphologically distinguishable heteromorphic sex chromosomes were reported in <20 species (Ming *et al.*, 2011; Renner, 2014). Because the sex-determining genes are mostly unknown, the reason why so few plants carry heteromorphic sex chromosomes remains unclear (Hobza *et al.*, 2018).

When recombination restriction is established, sex chromosomes start to diverge from the autosome pair they evolved from. Characteristic features of non-recombining sex chromosomes are genetic degeneration, gene loss, change of epigenetic landscape and gene transcription, accumulation of repetitive elements, chromosome rearrangements and change of chromosome size (Ming *et al.*, 2011). Here we focus on the most noticeable change, which is size variation between pairs of sex chromosomes caused by different rates of expansion or contraction (Parker, 1990; Ainsworth, 2000). It is assumed that young sex chromosomes are homomorphic, and as they age they become heteromorphic and larger than most autosomes, and the oldest sex chromosomes contract due to the loss of genes except those for sex determination (Vyskot and Hobza, 2004). Thus, size diversification is thought to be a feature of evolutionarily old sex chromosomes, while young sex chromosomes appear homomorphic (e.g. *Carica papaya*; Liu *et al.*, 2004), despite having a relatively large non-recombining region in some species, e.g. *Mercurialis annua* (Veltsos *et al.*, 2018, 2019), *Rumex acetosella* and *Rumex suffruticosus* (Cuñado *et al.*, 2007). Heteromorphic sex chromosomes result in a substantial difference in DNA content between males and females, reaching 7 % of the total DNA content, with males having a larger genome due to the expansion of the Y chromosome (Costich *et al.*, 1991; Veuskens *et al.*, 1992; Matsunaga *et al.*, 1994; Vagera *et al.*, 1994; Doležel and Göhde, 1995; Grabowska-Joachimiak and Joachimiak, 2002; Grabowska-Joachimiak *et al.*, 2005; Błocka-Wandas *et al.*, 2007; Puterova *et al.*, 2018). The Y chromosome is the largest in most of the known plants carrying clearly heteromorphic sex chromosomes, e.g. *Cannabis sativa* (hemp) (Sakamoto *et al.*, 2000; Divashuk *et al.*, 2014), *Hippophae rhamnoides* (sea buckthorn) (Truţă *et al.*, 2010; Puterova *et al.*, 2017), *Coccinia grandis* (ivy gourd) (Hossain *et al.*, 2016; Sousa *et al.*, 2016) and *Silene latifolia* (white campion) (Vyskot and Hobza, 2004; Puterova *et al.*, 2018). The evolutionarily older Y chromosome eventually starts to contract due to the loss of DNA, as seen in mammals (Ming *et al.*, 2011). The size increase often also occurs in the X chromosome. For example, in *S. latifolia*, with an XY system, the Y is the largest and X by far the second largest chromosome. In *Rumex* species with an XY_1Y_2 system, the X is the largest and the Y chromosomes are the second largest chromosomes (Navajas-Pérez *et al.*, 2009, Hough *et al.*, 2014; Kasjaniuk *et al.*, 2019). In contrast to the X, reasons for Y chromosome size increase are well rationalized by means of recombination restriction, which enables amplification of satellites, accumulation of chloroplast and mitochondrial DNA and transposable elements (TEs) (Navajas-Pérez *et al.*, 2005*a*, 2006; Mariotti *et al.*, 2006, 2009; Kubat *et al.*, 2008, 2014; Kejnovsky *et al.*, 2013; Steflova *et al.*, 2014; Hobza *et al.*, 2015, 2017, 2018). Why the plant X chromosome becomes larger is less understood due to limited knowledge of the specificities of X chromosome structure. It is assumed that less frequent X recombination taking place only in females might cause effects similar to those seen in completely non-recombining Y chromosomes, i.e. accumulation of diverse spectra of repetitive elements. However, the evolutionarily young X chromosome of the papaya accumulated solely insertions of long terminal repeat (LTR) retrotransposons. Accumulation of other repetitive sequences such as satellites and organellar DNA in comparison with the corresponding region of an autosome from a closely related monoecious species has not been found in papaya (Gschwend *et al.*, 2012; Na *et al.*, 2014). This emphasizes the potential role of other mechanisms in the X size increase. For example, a number of X-accumulated LTR retrotransposons suggest female-specific activity of some mobile elements in *S. latifolia* and *Rumex acetosa* (Cermak *et al.*, 2008; Steflova *et al.*, 2013; Kralova *et al.*, 2014; Kubat *et al.*, 2014). Therefore, the precise structures and compositions of X and Y chromosomes and autosomes at different evolutionary stages in a larger number of species are needed to elucidate potential reasons for X and Y chromosome size expansion.

We chose *R. acetosa* (common garden sorrel), a dioecious plant with XY_1Y_2 males and XX females (Kihara and Ono, 1923) for our study of the potential causal agents of sex chromosome size diversification. *Rumex acetosa*'s two Y chromosomes may have originated from a Y chromosome that underwent centromere fission (Lengerova and Vyskot, 2001); however, it is also possible that one of the Y chromosomes could be a neo-Y chromosome arising from the fusion of the X chromosome with an autosome, as in *Rumex hastatulus* (Smith, 1964; Grabowska-Joachimiak *et al.*, 2015; Kasjaniuk *et al.*, 2019). The sex chromosomes of *R. acetosa* form a Y_1-X-Y_2 trivalent during the zygotene phase of male meiosis (Parker and Clark, 1991). The Y chromosomes pair with the telomeric regions of opposite arms of the X. Ring-shaped trivalents were also observed. During anaphase I and metaphase II chromosomes segregate in a ratio of 8:7. This results in one cell having 6A + X and the second having 6A + Y_1Y_2 chromosomes (Farooq *et al.*, 2014). The Y chromosomes of *R. acetosa* lost their sex-determining gene and sex determination changed from having a dominant Y to the ratio of the number of X chromosomes to the number of autosomes (X:A ratio) (Ainsworth *et al.*, 1998). The sum of Y-chromosome lengths is larger than the length of the X chromosome, but the X as such is by far the largest chromosome, indicating that both have acquired huge amounts of DNA. Cytological and bioinformatic experiments show that Y chromosomes are heterochromatic and full of repetitive sequences with huge arrays of satellites not present on other chromosomes (Shibata *et al.*, 1999, 2000; Navajas-Perez *et al.*, 2005*b*; Mariotti *et al.*, 2009; Steflova *et al.*, 2013). Whilst these studies have shed light on the content of sex chromosomes in *R. acetosa*, they do not fully describe the repetitive fraction of the sex chromosomes and therefore their informational value with regard to size diversification is limited.

Here we used a unique and advanced approach based on the direct sequencing and subsequent bioinformatics analysis of separated X and Y chromosomes and autosomes. We employed the fluorescence *in situ* hybridization in suspension method (FISHIS) to sort X and Y chromosomes and autosomes. Subsequent whole-chromosome sequencing and bioinformatics analysis of repetitive fractions were employed to uncover compositional and quantitative differences between the sex chromosomes and autosomes in *R. acetosa* and to answer the following crucial questions: (1) how do the X and Y chromosomes differ compositionally from each other and from the rest of the genome? (2) which sequences contributed the most to size diversification of the sex chromosomes? (3) does a potentially reduced rate of concerted evolution in non-recombining Y chromosomes lead to the diversification of repeats? and (4) can the repetitive fraction shed light on the origin of sex chromosomes in *R. acetosa*?

MATERIALS AND METHODS

Chromosome sorting using FISHIS

Chromosomes for flow cytometric experiments were prepared from *Rumex acetosa* root tips according to Vrána *et al.* (2016). Seeds of *R. acetosa* were germinated in a Petri dish, immersed in water at 25 °C for 2 d until the optimal length of roots was achieved (~1 cm). The root cells were synchronized by treatment with 2 mM hydroxyurea at 25 °C for 18 h. Accumulation of metaphases was achieved using 10 µM oryzalin solution at 25 °C for 2 h. Approximately 200 root tips were required to prepare 1 mL of sample. The chromosomes were obtained by mechanical homogenization using a Polytron PT1200 homogenizer (Kinematica, Littau, Switzerland) at 18 000 r.p.m. for 13 s and the crude suspension was then filtered. For better differentiation of Y chromosomes, we performed FISHIS with chromosome flow sorting (Giorgi *et al.*, 2013) using 1 mL of crude suspension. NaOH (10 M) was added to produce pH 12.8–13.3. The suspension was incubated for 15 min on ice, then the pH was adjusted to the range of 8.5–9.1 using Tris-Cl. A probe solution of 5′-FITC-$(CAA)_{10}$ (1 ng µL^{-1}) was added to the final concentration (180 ng mL^{-1}) and the suspension was incubated for 1 h in the dark at room temperature and kept on ice until flow cytometric analysis. The samples were counterstained with DAPI (2 µg mL^{-1} final concentration). All flow cytometric experiments were performed on a FACSAria II SORP flow cytometer (BD Biosciences, San José, CA, USA). Chromosomes were sorted by relative DNA content (DAPI signal) and $(CAA)_{10}$ microsatellite abundance (FITC signal), which had the strongest signal of accumulation on the Y chromosome and can therefore be used to accurately distinguish Y chromosomes from other chromosomes (Kejnovský *et al.*, 2013). We obtained six chromosomal fractions: X, Y_1Y_2 and four autosomal fractions. For each sample the quality was checked by microscopy. Purity was estimated at 95 %. We used ~1 million chromosomes (100 ng of DNA), which were purified according to Šimkova *et al.* (2008). The amplification of purified chromosomal DNA was performed using a GenomiPhi DNA Amplification Kit (GE Healthcare, Chalfont St Giles, UK) according to the manufacturer's instructions.

Illumina sequencing

We performed one run of paired-end Illumina MiSeq sequencing, generating 301 bp reads for autosomes and two runs of 251 bp reads for X and Y chromosomes separately (accession number PRJEB23612). We obtained 25 672 002 raw paired-end reads from autosomes, 4 591 591 raw paired-end reads from the X chromosome and 2 731 018 raw paired-end reads from the Y. Sequencing reads were checked for quality using the FastQC tool (available at http://www.bioinformatics.babraham.ac.uk/projects/fastqc). Reads were pre-processed based on quality with subsequent adaptor trimming, filtering out short or unpaired sequences and cutting back all reads to a uniform length of 235 nucleotides using Trimmomatic tools (Bolger *et al.*, 2014) with the Galaxy platform (Afgan *et al.*, 2016).

We estimated the coverage of the male genome using the chromosome length as described in Lengerova and Vyskot (2001). The genome size of *R. acetosa* was previously reported to be 7.0 pg for the female and 7.5 pg for the male genome (2C) (Blocka-Wandas *et al.*, 2007).

Identification of repetitive sequences

We randomly sampled the sequencing data proportionally to reflect the male genome, giving 1 702 340 reads from autosomes, 287 234 from the X chromosome and 376 276 from the Y, which is equivalent to ~×0.074 coverage of the male genome. Such coverage is sufficient for the assembly of highly and moderately repetitive sequences (Macas *et al.*, 2015). To identify repetitive DNA in the X and Y chromosomes and autosomes of *R. acetosa* we carried out comparative analysis using the RepeatExplorer tool (Novák *et al.*, 2010, 2013). This tool performs graph-based clustering of sequences based on their similarity. Clusters were annotated manually using Geneious software version 7.1.9 (Kearse *et al.*, 2012) and automatically using RepeatExplorer output. We screened the clustering results to find sequences that had been reported previously. Clusters containing unknown sequences were investigated for typical transposon protein domains using the Conserved Domain Database (CDD) (Marchler-Bauer *et al.*, 2017). Monomers of satellite DNA were detected by Tandem Repeat Finder (TRF 4.09) (Benson, 1999). Finally, we manually created a library of repeats using the sequences derived from the clusters.

Identification of microsatellites

To identify microsatellites on X and Y chromosomes and autosomes we used Tandem Repeat Finder (TRF 4.09) (Benson, 1999) and the Tandem Repeats Analysis Program (TRAP) (Sobreira *et al.*, 2006) with the following parameters: 2 7 7 80 10 50 1000. The results obtained served as a template to calculate the abundance of microsatellites.

Rank abundance curves

To test the hypothesis that tandem repeats originate mainly on sex chromosomes we compared the diversity of microsatellites in the three chromosome libraries by constructing rank abundance curves. Rank abundance curves are often used in ecological studies to simultaneously visualize both species richness and species evenness. To ensure equal sampling we randomly selected 1 000 000 reads from each library and analysed them in TRF 4.09 (Benson, 1999). The abundance of each unique tandem repeat was calculated as annotated nucleotides/total nucleotides used (= 3.02 × 10^8). Unique tandem repeats were ranked consecutively within each chromosome sample from most to least abundant.

Relative abundances of annotated clusters in the genome

Technical 95 % confidence intervals for repeat relative abundances on X and Y chromosomes and autosomes were constructed assuming binomial (multinomial) distribution of reads into clusters. The relative abundance of a cluster in the whole male genome was calculated as (10.8 × $A_{portion}$ + 1.8 × $X_{portion}$ + 2.35 × $Y_{portion}$)/14.95. Statistical analysis and figures were created in statistical software R (version 1.2.5019) (RStudio Team, 2020).

FISH analysis

Specific primers were designed for contigs from selected clusters (Supplementary Data Table S1). For the transposons, primers were made for the LTRs and/or the transposon domains (for instance, gag). Monomers of the satellite DNA were chosen for primer design. In the first step, template DNA was amplified using PCR with a mix containing 1× complete PCR buffer (Novazym VivaTaq DNA Polymerase buffer ×10), 0.1 mM dNTPs, 0.1 mM primers, 0.5 U Taq polymerase (Top Bio) and 10–15 ng of template DNA. Reaction conditions were as follows: 95 °C for 4 min, 34× (95 °C for 50 s + 55 °C for 50 s + 72 °C for 1 min) + 72 °C for 10 min. PCR products were checked by gel electrophoresis, cleaned using the Qiagen PCR Purification Kit, cloned into a pDrive vector (Qiagen) and transformed into *Escherichia coli*. Clones were sequenced to verify the presence of a specific product. Selected clones were then used for probe preparation for FISH by PCR and labelled using a Nick Translation Kit (Roche).

FISH was performed on mitotic metaphase chromosomes prepared from root tip cells. The hybridization mix contained 50 % formamide, 2× SSC and 10 % dextran sulphate. The labelled DNA (1–5 ng μL^{-1}) was denatured, added to a slide and hybridized at 37 °C for 18 h. Slides were then washed with medium stringency (250 s in 2× SSC at 42 °C, 250 s in 0.1 SSC at 42 °C, 250 s in 2× SSC at 42 °C, 50 s in 2× SSC at room temperature, 70 s in 4× SSC + 1 % Tween) and finally washed in 1× PBS. The chromosomes were counterstained with DAPI and mounted in Vectashield, examined under an Olympus AX70 fluorescent microscope, scanned with a CCD camera and analysed using ISIS software.

BAC library construction and screening

A BAC library was constructed from *R. acetosa* male high molecular weight genomic DNA. Briefly, DNA was digested with the HindIII enzyme and inserted into a pIndigoBAC-5 vector. Clones were then gridded in duplicate on Hybond N+ (Amersham Biosciences) nitrocellulose membrane filters in a 4 × 4 pattern that allowed us to identify the well positions and plate numbers of each clone, and incubated and processed as described in Bouzidi *et al.* (2006). The *R. acetosa* BAC library (72 000 colonies) was arrayed on six nylon filters with 18 432 colonies each and an additional one containing 9216 clones. The average insert size of the library was 128 kb. Based on nuclear size data, we estimated that coverage of the *R. acetosa* BAC library is 2.84 complements of the male haploid genome. Screening was performed by radioactive hybridization with α32P using a Prime-It II Random Primer Labelling Kit (Stratagene) according to the manufacturer's protocol. Probes were prepared by PCR amplification of the different sequences derived from the contigs. We selected clones showing strong hybridization with the probe, and only those that were confirmed by PCR with probe-derived primers were used in further analyses. Clones were sequenced using Illumina MiSeq 300 nt paired-end sequencing. Raw data processing, sequence assembly, alignment and annotation were done with Geneious software (Kearse *et al.*, 2012) and Edena v3 assembler (Hernandez *et al.*, 2008).

RESULTS

Repeat assembly, annotation and quantification

We identified the main groups of repetitive DNA in the *R. acetosa* genome using a RepeatExplorer pipeline. We estimated the proportion of the main repeat families in *R. acetosa* for X and Y chromosomes and autosomes. For the further analyses, we used 319 out of 387 reconstructed clusters. All the unused clusters were small and without any similarity to known sequences. Three hundred and nineteen used clusters formed at least 0.01 % of the genome and they comprised 57.62 % autosome, 68.07 % X and 73.75 % Y chromosome reads together (Supplementary Data Table S2A). We measured proportions and described the main types of repetitive DNA. Thirty-nine out of the 319 studied clusters were annotated as satellites and 123 clusters as transposons. It is important to note that a single repeat type can be found fragmented in several clusters. For this reason, we manually inspected all clusters and classified some of them as a single repeat type. Two clusters corresponded to 5S rDNA (CL285) and three to 45S rDNA (CL165). Since we used flow-sorted chromosomes, none of the analysed contigs contained chloroplast DNA (cpDNA), although cpDNA was found in smaller clusters, probably because of nuclear cpDNA insertions (Steflova *et al.*, 2014). Four clusters (CL54, CL66, CL77, CL115) were omitted as bacterial contamination.

Chromosome-specific comparative analysis revealed new satellites

For each identified satellite from Supplementary Data Table S2A, we reconstructed a monomer and described its size (Table 1, Supplementary Data Table S3). Known *R. acetosa* satellite DNA sequences were identified against the NCBI database: RAYSI, RAYSII, RAYSIII, RAE180, RAE730, RA160 and RA690. Newly discovered satellites were named according to the genome of origin (RAE) and monomer size, or, in the case of Y-specific satellites (based on FISH results),

TABLE 1. *Comprehensive table of* R. acetosa *satellites with estimation of distribution and abundance on X and Y chromosomes and autosomes. Newly described repeats are indicated. Estimation was based on the RepeatExplorer comparative analysis results*

Satellite sequences

Repeat name	FISH location	Reference	Proportion on chromosomes (%)		
			A	X	Y
RAE180	Mostly on Y	Shibata *et al.* (2000)	1.68	0.23	2.94
RAYSI	Y specific	Shibata *et al.* (1999)	0.08	0.05	1.39
RAYSII	Y specific	Mariotti *et al.* (2009)	0.01	0.01	0.03
RAYSIII	Y specific	Mariotti *et al.* (2009)	0.07	0.01	0.20
RA160	Y, X and 2 A	Steflova *et al.* (2013)	0.11	0.00	0.01
RA690	Y, X and 2 A	Steflova *et al.* (2013)	0.23	0.14	1.24
RAE730	Y and 1 A	Shibata *et al.* (2000)	0.08	0.02	0.46

Novel satellite sequences

Repeat name	FISH location	Putative monomer length (bp)	Proportion on chromosomes (%)		
			A	X	Y
RAE173	Mostly on Y	173	0.09	0.17	4.51
RAE244	Mostly on Y	244	0.01	0.00	0.09
RAYSIV	Y-specific	175	0.01	0.01	0.26
RAYSV	Y_1-specific	468	0.01	0.00	0.03
RAYSVI	Y-specific	445	0.02	0.01	0.22
RAYSVII	Y_1-specific	164	0.00	0.00	0.21

we continued naming repeats with the RAYS prefix (***R****umex* ***a****cetosa* **Y s**pecific) as in Shibata *et al.* (1999). The chromosomal distribution of newly described repeats was determined by FISH with RAYSI satellite used as a Y chromosome marker (Fig. 1). FISH shows that all of the known and newly described satellites occur mostly on the Y chromosomes.

The RAYS satellites are called Y-specific because FISH images show signals on Y chromosomes only (Fig. 1B–E). However, our bioinformatics analysis using the RepeatExplorer pipeline revealed that to some extent they are present also on the X chromosome and/or autosomes with the exception of clearly Y-specific RAYSVII (Table 1). To explain the discrepancy between the sequencing data and the FISH observations, we screened the *R. acetosa* BAC library with a RAYSV-derived probe. Six BACs with the strongest signal were sequenced and assembled. Sequencing data revealed that the RAYSV sequence is highly variable and individual monomers differ significantly from each other (data not shown). Similar intra-specific variability was previously recorded for RAYSI as well as RAE180 and RAE730 (Navajas-Pérez *et al.*, 2005*b*). In other words, although FISH analysis revealed distinct and specific signals of RAYS satellites on Y_1 and/or Y_2 chromosomes, sequencing data suggest that slightly different variants of these satellites are also present on autosomes and/or X chromosomes but their distribution is more dispersed, i.e. they do not form large repetitive blocks. The chromosomal distribution of other satellites, RA and RAE, is generally very similar to that of RAYS satellites, i.e. several strong signals on Y chromosomes and a few less intense signals on the X and autosomes (Table 1, Fig. 1A, F, G). Thus, we can conclude that short satellite arrays are ubiquitous in the *R. acetosa* genome, but expansion of satellites takes place mainly on the Y chromosomes, contributing to Y chromosome size increase.

Some satellites originated from LTR retrotransposons

We were interested in whether the investigated satellites had similarities with other types of repetitive DNA. By analysing clustering data, we detected two satellites associated with LTR retrotransposons. The RAE93 satellite shows a similarity to the 3′-end UTR of the RA Ogre/Tat LTR retrotransposon, which was confirmed by sequencing of Ogre-containing BACs. RAE93 forms short tandem arrays (five-monomer array) downstream from the *gag-pol* gene of RA Ogre/Tat elements (Supplementary Data Fig. S1). Further analysis using FISH revealed that while the RA Ogre/Tat probe derived from the gag protein-coding sequence paints the entire Y chromosome, with minor additional signals dispersed throughout the rest of the genome (Fig. 1H), the RAE93 satellite is concentrated into a lower number of discrete strong spots mainly on the X and Y chromosomes and minor additional signals resembling the *gag*-derived probe (Fig. 1G). From this it can be inferred that the RA Ogre/Tat element contains short tandem arrays of RAE93 and disperses them in the genome along with the element amplification. RAE93 eventually expands into long repetitive arrays in parts of the genome possessing conditions suitable for satellite expansion. Such a scenario was previously confirmed for several satellites in *Lathyrus sativus* (Vondrak *et al.*, 2020).

The Ty1/Copia RA AleII LTR retrotransposon-derived satellite has a completely different nature from any other known satellite originating from a TE. The RA AleII satellite monomer contains a full-length non-autonomous copy of the AleII retrotransposon consisting of a gag protein-like domain, DNAJ protein domain, polypurine tract (PPT), primer-binding site (PBS), both a 3′ and a 5′ end, and LTRs. The tandem nature of this satellite was confirmed by BAC sequencing (data not shown). FISH imaging shows a single discrete signal at the distal part of the shorter arm of the Y_1 and on the X chromosome (Fig. 1I) and clustering analysis revealed that RA AleII makes up 0.066 % of autosomes, 0.337 % of the X and 0.031 % of the Y chromosomes. These data together suggest that the mildly transpositionally active non-autonomous RA AleII retrotransposon gave rise to a single satellite locus only once. This locus is present in a putative pseudoautosomal region mediating recombination between the X and Y_1 chromosomes.

Analysis of micro- and minisatellite diversity

It has been hypothesized that suppressed recombination on Y chromosomes reduces the rate of concerted evolution and leads to the diversification of satellites (Navajas-Pérez *et al.*, 2006). In theory, some novel mutated satellites should be better predisposed to multiplication, and therefore satellite expansion on Y chromosomes can be a result of increased satellite diversity on the Y. Another hypothesis assumes that satellite expansion is caused by a lack of recombination repair. Since our short-read data are not suitable for the analysis of relatively long satellite monomers, we investigated the

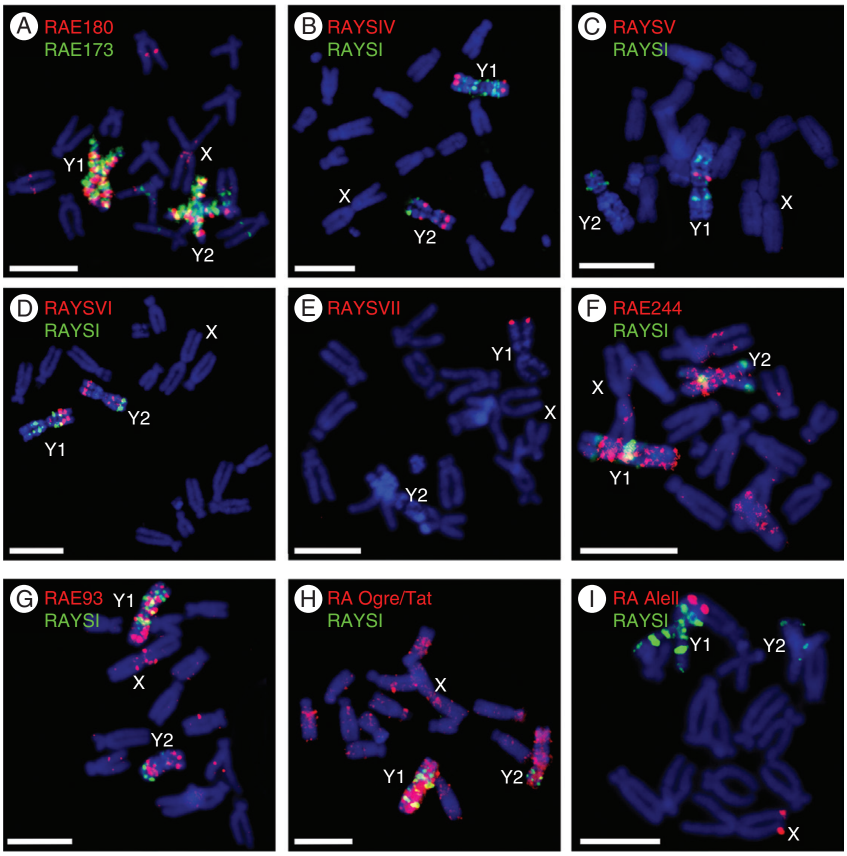


Fig. 1. Localization of satellite DNA and transposable elements on metaphase chromosomes of *R. acetosa* using FISH. Scale bars = 10 µm. (A) RAE180 (red signal) and RAE173 satellite (green signal) paint almost the entire Y chromosomes. (B) RAYSIV satellite (red signal) is present on both Y chromosomes in two (Y_1) and three (Y_2) loci. (C) RAYSV satellite (red signal) is at one locus on the Y_1 chromosome in the subcentromeric region. (D) RAYSVI satellite (red signal) gives a signal at several discrete loci on both Y chromosomes. (E) RAYSVII satellite (red signal) is present in the distal part of the Y_1 chromosome. (F) RAE244 satellite (red signal) is accumulated on Y chromosomes and a few dispersed signals are observed in the remainder of the genome. (G) RAE93 (red signal) covers both Y chromosomes and a few loci in the remainder of the genome. (H) RA Ogre/TAT retrotransposon (red signal) is accumulated mostly on both Y chromosomes. (I) RA AleII retrotransposon is located on the distal parts of the Y_1 and X chromosomes. RAYSI satellite (green signal) was used as a Y-chromosome marker; the signal is localized in four spots on each arm of the Y_1 chromosome and in two spots on each arm of the Y_2 chromosome.

Table 2. *Comprehensive table of the most abundant micro- and minisatellites in the* R. acetosa *genome*

Abundance on autosomes (%)	Abundance on X (%)	Abundance on Y (%)	Monomer size (bp)	Monomer sequence
0.0338	0.0188	0.3397	3	AAC
0.0000	0.0000	0.2988	9	AACACACCC
0.0087	0.0054	0.0049	3	AAG
0.0048	0.0033	0.0105	6	AACCCT
0.0046	0.0035	0.0087	9	AACAACAAG
0.0053	0.0040	0.0033	2	AG
0.0041	0.0030	0.0075	11	AAAAACGAGCG
0.0044	0.0023	0.0028	61	AAAAAATCGTCATCGAGCTCAAAAACGTGTTTGATGACATTATTTCGAGCTTGATGACGTT
0.0000	0.0000	0.0232	10	AAACACACCC

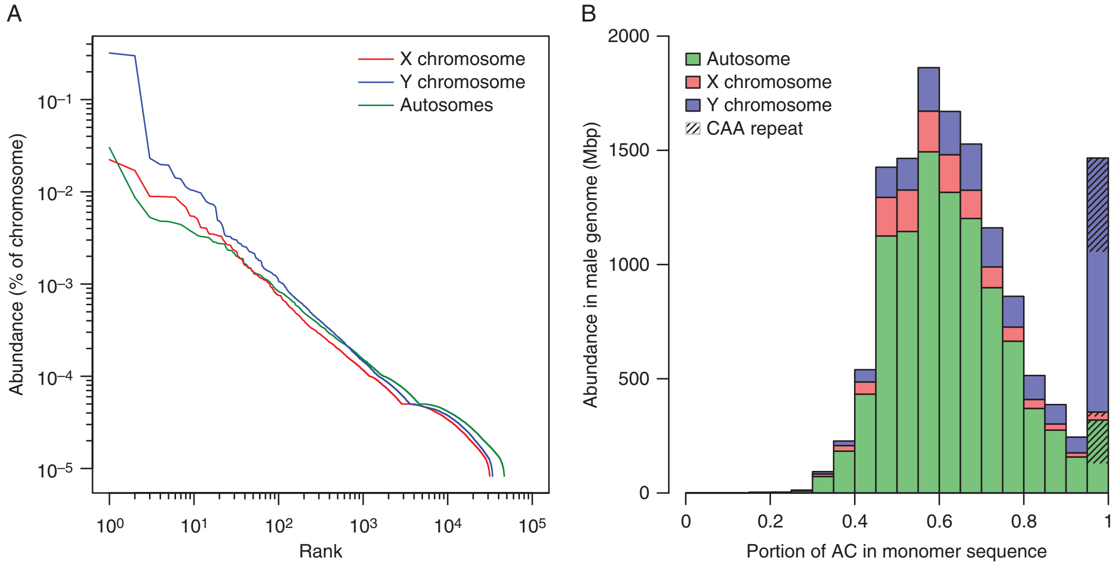


FIG. 2. Micro- and minisatellite diversity in the *R. acetosa* genome. (A) Rank abundance distributions of short tandem repeats for autosomes and sex chromosomes. Abundance of each unique tandem repeat calculated as percentage of total nucleotides from 1 million reads is on the *y*-axis. Unique tandem repeats are ranked consecutively within each chromosome library on the *x*-axis. Curves are displayed on a log–log scale for clarity. (B) Distribution of tandem repeats with respect to their adenine plus cytosine content. Contributions of each chromosome class are stacked in a histogram. Y chromosomes contain notable portions of tandem repeats consisting of pure C and A combinations, which are not true CAA repeats but could hypothetically be derived from them. No other base combination showed such deviation from normality.

chromosome-specific variability of micro- and minisatellites. Micro- and minisatellites form a rather minor genome fraction, occupying 1.82 % of autosomes, 1.34 % of X and 2.27 % of Y chromosomes (Table 2, Supplementary Data Table S4). We were particularly interested in whether micro- and minisatellites show higher diversity on the non-recombining Y chromosomes than on the X and autosomes. Our analyses considered all permutations in both complementary strands as a single satellite type. We constructed a graph with individual satellites ranked consecutively based on their abundance in equally sized sets of chromosome-specific reads (Fig. 2A). The blue curve, representing micro- and minisatellites on the Y chromosomes, is positioned higher than the red (satellites on X) and green curves (satellites on autosomes) in the graph. Thus, satellites expand with higher probability on non-recombining Y chromosomes. In addition, the green curve is less steep and extends further to the right, indicating a higher number of unique satellites on autosomes. This is consistent with the idea that random sampling of microsatellites from autosomes representing most of the genome gives a higher diversity of repeats than the relatively shorter sex chromosomes. However, the percentage of mismatches within microsatellite arrays (calculated by Tandem Repeat Finder) is higher for autosomes (weighed mean of all arrays, 14.67 %) than the X and Y chromosomes (13.60 and 12.13 %, respectively). This suggests a higher natural diversity of autosomal micro- and minisatellites. There are two possible reasons: (1) slower amplification or (2) a lower level of concerted evolution in comparison with sex chromosomal counterparts. Nevertheless, apart from the different abundance of satellites, the X and Y curves are similar in shape and gradient and suggest that X and Y chromosomes (and autosomes with high probability as well) differ in number but not diversity of micro- and minisatellites within the same-sized DNA region.

Upon closer inspection of the most prolific micro- and minisatellites (Supplementary Data Table S4), we noticed that a group of satellites that accumulated strongly on Y chromosomes had a quite high sequence similarity and contained almost exclusively A and C bases (permutations and 5′→3′ and 3′→5′ reads were merged). The sho rtlist of the most abundant CA-rich satellites is depicted in Fig. 3. Microsatellite AAC is ubiquitous in the genome but extremely propagated on Y chromosomes (Supplementary Data Fig. S2A). In addition, minisatellites potentially derived from AAC or AACACACCC are absent everywhere but Y chromosomes (Supplementary Data Fig. S2B–F). To investigate the connection between monomer expansion and base composition we inspected all identified mini- and microsatellites and constructed a graph with a histogram of the distribution of all repeats with respect to their AC content (Fig. 2B). Surprisingly, mini- and microsatellites show extreme deviation from normal distribution with respect to AC content, which suggests that AC-containing satellites are predisposed to expansion.

TE classification

Using the RepeatExplorer pipeline we classified the majority of the TEs and calculated their abundance in the *R. acetosa* genome (Supplementary Data Table S2A). A repeat content summary for the whole male genome is presented in Supplementary Data Table S5 and shows that the *R. acetosa*

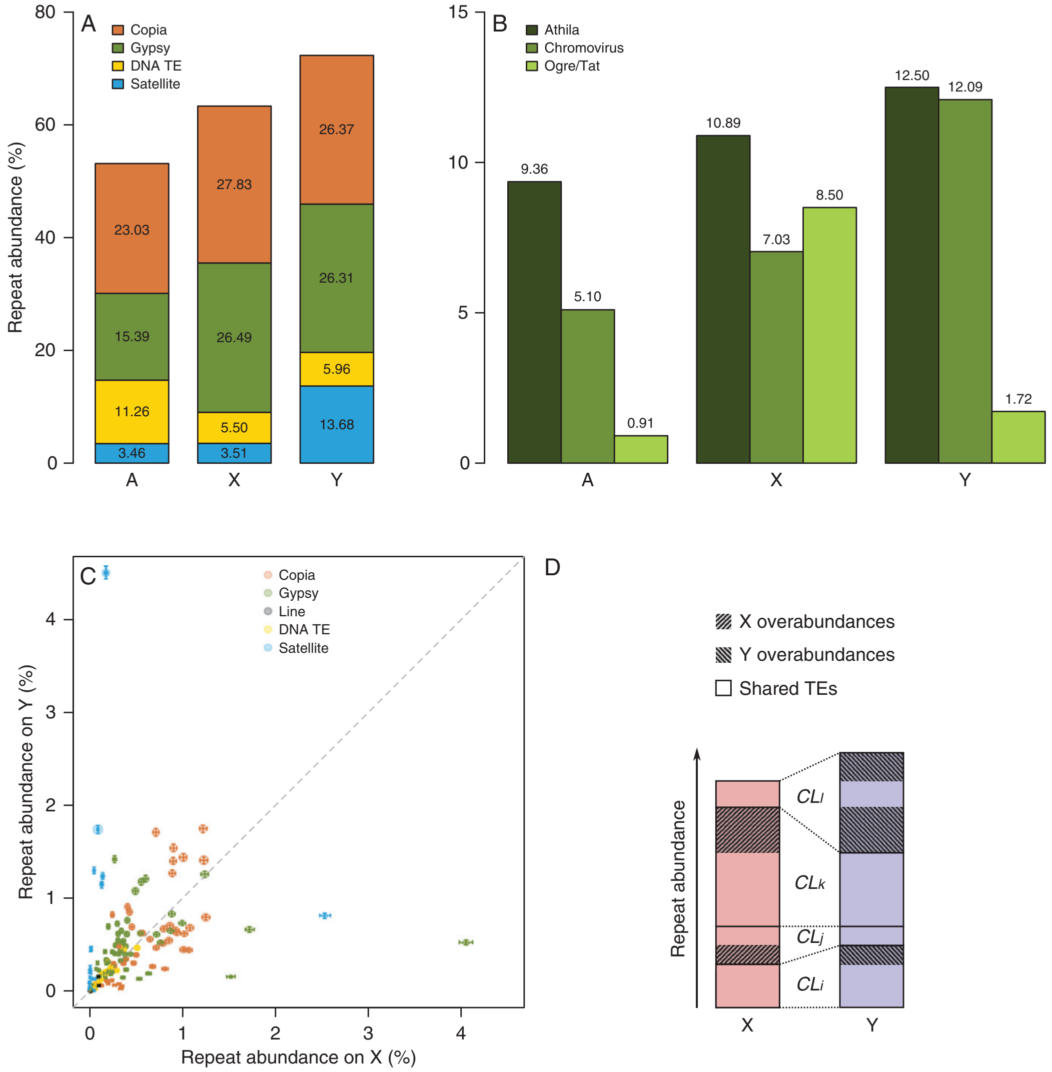


Fig. 3. Analysis of repeat composition of *R. acetosa* genome and sex chromosomes. (A) Composition of repeats on X and Y chromosomes and autosomes of *R. acetosa* estimated from Illumina sequencing data. (B) Abundance of subfamilies of Gypsy-like transposable elements on X and Y chromosomes and autosomes. (C) Relative abundances of annotated clusters on X versus Y chromosome. Error bars represent technical 95 % confidence intervals assuming binomial distribution of reads into clusters. Area of each circle is proportional to given cluster portion in male genome. Dashed line indicates a theoretical situation where the amplification rate of a repeat family (cluster) is equal on X and Y chromosomes. (D) Graphical representation of how the overabundances in Table 3 are calculated. Numbers i, j, etc. are integers. Each cluster (CLi, CLj, ...) is either more abundant on X or on Y. If the difference is only due to technical error the excess should be generally small compared with total cluster abundance. We added the excess for all clusters from a given transposon group separately for each chromosome and named these sums X and Y overabundances. They can be compared with total transposon group abundances in Table 3.

genome contains all the main types of TEs. Class I LTR and non-LTR retrotransposons are the dominant TE type and account for around 40 % of the genome. Analysis of the retrotransposon domains revealed that most Ty1/Copia-like LTR retrotransposons belong to the Maximus/SIRE family, while Ty3/gypsy elements are mostly represented by the following three families: Athila, Ogre/Tat and Chromovirus. Much less abundant class II elements (around 4.5 %) are predominantly represented by MuDR_Mutator DNA transposons.

Next, we were interested in the scale of diversity that occurs in TEs of an individual family. We manually inspected the clustering data and BAC sequences. Athila, Chromovirus and Maximus/SIRE clusters and BAC sequences evinced high fragmentation and a frequent lack of protein domains and features as functional LTRs. These findings indicate a long history of proliferation in the *R. acetosa* genome, the presence of multiple independent lineages and genetically degenerated copies, with one exception for Maximus/SIRE elements, which show higher sequence similarity among element copies and thus a comparatively lower number of independent lineages. High sequence conservation of Ty1/Copia elements has been reported in other species and it has been suggested by Macas *et al.* (2015) that it might be a general feature. In contrast, most Ogre/Tat elements are fully featured but still present in several independently spreading lineages. Based on the prevalence of full-length element copies, we conclude that Ogre/Tat retrotransposons are evolutionarily young and recently underwent an explosive proliferation. Coincidentally, Ogre elements are also the main drivers of recent genome size expansion in dioecious *S. latifolia* (Cegan *et al.*, 2012).

TEs show an inverse distribution pattern on sex chromosomes

The estimation of TE abundance on separated chromosomes revealed an interesting pattern of distribution, where both Ty1/Copia and Ty3/gypsy-type LTR retrotransposons occupy a significantly higher percentage of DNA on the sex chromosomes (X and Y chromosomes, 54.32 and 52.68 %, respectively) than on autosomes (38.42 %). Contrastingly, DNA transposons are much less abundant on sex chromosomes (5.96 %) compared with autosomes (11.26 %) (Fig. 3A). Such patterns of distribution can be explained by different speeds of amplification between class I and class II TEs. In this scenario, rapidly spreading and mutating LTR retrotransposons (Preston and Dougherty, 1996) overshadowed slowly amplifying DNA transposons in evolving sex chromosomes undergoing recent size increase. To conclude, LTR retrotransposons represent the second major cause of sex-chromosome size diversification, besides satellites.

Next, we focused on whether particular TE families contribute proportionally to sex-chromosome size increase. Surprisingly, the chromosomal abundance of Ty3/gypsy families differs (Fig. 3B). While Athila and Chromovirus LTR retrotransposons have highest abundance on the Y chromosomes and slightly less abundance on the X chromosome, Ogre/Tat elements are relatively rare on the Y and extremely abundant on the X chromosome. However, FISH revealed an opposing distribution of Ogre elements, a strong presence on the Y and a weak representation on the X chromosome (Fig. 1H). This discrepancy can be explained by clustering analysis indicating the existence of multiple independent lineages within each LTR retrotransposon family. In the case of Ogre/Tat, we can conclude that there are several Ogre/Tat lineages with contrasting chromosomal distributions in the genome. Since all the other TE families comprise multiple lineages, we were curious whether their chromosomal distribution resembles the situation within the Ogre/Tat family

We assumed that each cluster (Supplementary Data Table S2A) represents either a partial sequence of the identical TE element lineage or a different TE element lineage with potentially unique chromosomal distribution. We plotted the X-chromosome proportion against the Y-chromosome proportion of each cluster separately (Fig. 3C). The plot shows an extreme enrichment of satellites on Y chromosomes and a roughly equal abundance of DNA transposons on the X and Y chromosomes, which is in concordance with Fig. 3A. On the other hand, most LTR retrotransposon clusters are more abundant either on the X or Y chromosomes. Thus, each lineage of Maximus/SIRE, Athila, Ogre/Tat and Chromovirus LTR retrotransposons accumulates preferentially either on the X or on the Y chromosomes.

Thereafter we determined the level to which each TE lineage is enriched on the X or Y chromosomes. The TE abundance data were purged of satellites, which affects the percentage values of other repeats due to the satellite's expansion on the Y but not X chromosome (Supplementary Data Table S2B). The sum of proportions and the ratio of sex chromosome-specifically enriched elements from individual TE families is shown in Table 3 and explanatory Fig. 3D. Obviously, 47.69 % of sex chromosome DNA comprises the same shared TEs but another 14.33 % of the X chromosome and 20.18 % of the Y chromosome are made up of unique TEs, i.e. TEs enriched (over-abundant) on the X and Y chromosomes, respectively. The percentage of shared TE copies is 76.90 % and 70.26 % of all TEs on the X and Y chromosomes, respectively, indicating that individual TE lineages more probably accumulate on the Y chromosomes due to either preferential activity in males or a higher fixation rate on non-recombining Y chromosomes, or both. These summarizing data somewhat obscure the behaviour of individual TE lineages. Thus, for example, one of the Chromovirus lineages (cluster 72, Supplementary Data Table S2B) occupies 0.28 % and 1.65 % of X and Y chromosomes, respectively, which implies that 17 % of Y-TE copies are shared with the X chromosome. In other words, there are over 5 times fewer copies on the X than on the Y chromosomes. Another example is Ogre/Tat lineage (cluster 93, Supplementary Data Table S2B), occupying 1.58 % and 0.18 % of X and Y chromosomes, respectively. Accumulation on either of the chromosomes is visible for all TE types, with Ty3/gypsy LTR retrotransposons being most distinctive.

TABLE 3. *Sum of proportions and ratio of sex chromosome-specifically enriched transposon lineages from different families. Graphical representation and explanation of how overabundances are calculated is in Fig. 3D*

	Percentage of chromosome DNA				Percentage of TE copies		
Repeat type	X sum	X overabundances	Y sum	Y overabundances	Shared TEs	X TEs shared with Y	Y TEs shared with X
LTR retrotransposon							
Ty1/Copia	28.75	6.15	30.46	7.86	22.60	78.61	74.19
Ty3/Gypsy	27.26	7.82	30.12	10.68	19.44	71.33	64.55
Non-LTR							
LINE	0.31	0.06	0.38	0.13	0.25	82.08	66.06
DNA TE	5.70	0.31	6.90	1.51	5.39	94.62	78.12
Sum	62.01	14.33	67.87	20.18	47.69	76.90	70.26

Relative gain of repeats on sex chromosomes in comparison with putative ancestral autosomes

We investigated how much individual repeats changed their copy numbers along with the evolution of sex chromosomes from an ancestral autosome. We worked with the assumptions that (1) the non-repetitive fraction has not changed between ancestral autosomes and current sex chromosomes in size, (2) the ancestral autosome pair from which the current sex chromosomes originated had a repeat composition similar to that of the current autosomes, and that (3) even if ancestral autosomes had a lower repeat content, relative repeat gains along with the evolution of dioecy were uniform across chromosomes. We estimated the number of base pairs of each repeat type on the putative ancestral autosome and current sex chromosomes. Table 4 shows the relative gains of individual repeat types on sex chromosomes. Obviously, Y chromosomes acquired more repeats than X chromosomes and simultaneously lost more of some slowly proliferating TEs (DNA transposons). The latter can be explained by the accelerated genetic degeneration of old DNA transposon copies due to the raised insertion frequency of other TEs on both X and Y chromosomes, and recombination restriction on the Y chromosomes.

All in all, we can assume that the X chromosome expands almost exclusively due to an accumulation of TEs that prefer the X chromosome for insertion rather than the Y. In comparison, the expansion of Y chromosomes is caused by a combination of three factors: (1) accumulation of TEs favouring Y chromosomes; (2) accumulation of satellites; and (3) most likely increased fixation rate of repetitive elements of all types due to recombination restriction.

DISCUSSION

Non-recombining sex chromosomes frequently incorporate various types of repetitive DNA sequences. Consequently, sex chromosomes quickly diverge from each other and from the rest of the genome. Those processes can be monitored either by cytogenetic methods (e.g. visualization of heterochromatic regions and/or FISH experiments with selected probes) or by whole-genome sequence analysis. Previous studies in *R. acetosa* either provided a description of the differences between male and female genomes (Steflova *et al.*, 2013) or focused only on narrow aspects of sex chromosome divergence (Shibata *et al.*, 1999, 2000; Navajas-Perez *et al.*, 2005a, b; Mariotti *et al.*, 2009; Steflova *et al.*, 2013).

Table 4. *Relative repeat gain of current sex chromosomes compared with putative ancestral autosome(s). Size and composition of putative ancestral autosome(s) were calculated assuming (1) the composition of the ancestral autosome was similar to that of the current autosome library, and (2) the absolute amount of the non-repetitive portion of the sex chromosomes did not change drastically during their evolution. Indicated errors account for differences when Y chromosomes and X non-repetitive portion were used for calculation of ancestral autosome(s) size*

Repeat type		X relative gain (%)	Y relative gain (%)
LTR retrotransposon	Ty1/Copia	+55 ± 5	+92 ± 6
	Ty3/Gypsy	+121 ± 7	+186 ± 10
Non-LTR	LINE	−15 ± 3	+23 ± 4
DNA transposon		−37 ± 2	−11 ± 3
Satellites		+30 ± 4	+561 ± 22
Not annotated		+36 ± 5	−47 ± 2

This study represents a direct approach to the analysis and quantification of individual repetitive elements on the sex chromosomes and autosomes of common sorrel (*R. acetosa*). Using sorting and sequencing of individual chromosomes, we highlight the differences between X and Y chromosomes and autosomes of this species. We present the first quantitative analysis of repetitive sequences in plant sex chromosomes.

Satellite sequences: the key players of Y-chromosome expansion?

Although it has already been shown that the Y chromosome of *R. acetosa* possesses a greater percentage of satellite sequences than the X chromosome and autosomes, our chromosome-based approach has extended and improved the genome description at the repeatome level and has enabled the identification of six major novel satellites that make up >5 % of Y chromosomes (Table 1). Along with the seven previously published tandem repeats (RAYSI, RAYSII, RAYSIII, RAE180, RAE730, RA160, RA690) (Shibata *et al.*, 1999, 2000; Navajas-Perez *et al.*, 2005*a*, *b*; Mariotti *et al.*, 2009; Steflova *et al.*, 2013), 13 major satellites represent 13.68 % of Y chromosomes (Fig. 4). Two in particular (RAE180 and RAE173; Table 1) make up half of this number. In addition, we have identified about two dozen minor satellites, giving a total number of different satellites of around 40 in the *R. acetosa* genome. Such an elevated number of different satellite families resembles the satellite diversity present in the dioecious plant sea buckthorn (Puterova *et al.*, 2017).

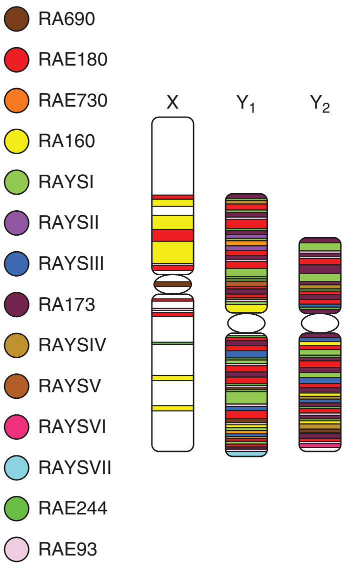


Fig. 4. Schematic map of satellite localization on sex chromosomes in *R. acetosa*.

Satellites are thought to accumulate in genomic regions with less recombination, e.g. the Y chromosome or the X chromosome, which has its recombination partner only in females. In *R. acetosa*, arrays of satellites form almost 14 % of the Y chromosomes, but their proportion on both the X chromosome and the autosomes is only 3.5 % (Fig. 3A). This information suggests that the recombination level is similar on the X chromosome and autosomes and possibly not sufficiently reduced to enable a high expansion of satellites. Moreover, it has been hypothesized that suppressed recombination on the Y chromosome reduces the rate of concerted evolution and leads to the diversification of satellites (Navajas-Pérez *et al.*, 2006). In contrast to this, we found a strong expansion of satellites on the Y chromosome, but could not confirm the increased diversity of Y satellites compared with the X chromosome and autosomes.

The previous analysis of the common sorrel genome revealed an unprecedented expansion of AC-containing microsatellites in the male genome (Kejnovsky *et al.*, 2013). Here, we performed an even more extended chromosome-specific analysis of micro- and minisatellites and conclude that AC-rich microsatellites are the prevalent type, which is derived from shorter AAC-containing motives by consecutive cycles of duplication and divergence. Exceptional richness of Y chromosomes with AAC-derived microsatellites can then influence the destiny of sex chromosomes, due to the microsatellite arrays serving as a target for TE insertions (Kejnovsky *et al.*, 2013), which is discussed below.

X- and Y-chromosome size increase is driven by different TE lineages

Although accumulation of TEs on sex chromosomes has been commonly assumed to be a natural consequence of recombination restriction and has been repeatedly confirmed in species as diverse as *Marchantia polymorpha* (Okada *et al.*, 2001), *Cannabis sativa* (Sakamoto *et al.*, 2000, 2005), *Bryonia dioica* (Oyama *et al.*, 2010), *Humulus lupulus* (Divashuk *et al.*, 2011) *C. papaya* (Yu *et al.*, 2007; Gschwend *et al.*, 2012; Na *et al.*, 2014), *Asparagus officinalis* (Li *et al.*, 2014), *S. latifolia* (Cermak *et al.*, 2008; Filatov *et al.*, 2009; Kralova *et al.*, 2014, Kubat *et al.*, 2014; Puterova *et al.*, 2018) and *R. acetosa* (Steflova *et al.*, 2013), the last two species provide a new and complex view on why TEs accumulate on sex chromosomes. The most striking feature of the TEs of white campion and sorrel is their irregular distribution along the X and Y sex chromosomes, when it appears that most TEs have a preference for either the X or Y chromosome for insertion. Insertional targeting into specific chromosomal regions such as microsatellite arrays (Akagi *et al.*, 2001; Kejnovsky *et al.*, 2013), other transposons (Jiang and Wessler, 2001) and gene promoters (Naito *et al.*, 2014) has been seen previously in a number of TEs and might be consistent with TE accumulation in the largely heterochromatic Y chromosomes of *R. acetosa* (Shibata *et al.*, 2000). However, the satellite-less, euchromatic, gene-rich X chromosome seems to have a chromatin structure comparable to that of autosomes. Why then should so many TEs be enriched on the X? We advocate that the culprit can be found among the cellular mechanisms for genome defence against deleterious activity of TEs. We have previously shown that recently spreading Ogre LTR retrotransposon elements (Cegan *et al.*, 2012), which are enriched on the X and almost absent on the Y chromosome of *S. latifolia*, might be differentially regulated by sRNA molecules involved in epigenetic regulation of TEs (Kubat *et al.*, 2014). Moreover, recent progress in the field of epigenetic regulation of TEs revealed that the most crucial time for effective TE silencing within plant life is during the formation of gametes and early embryogenesis, due to the TEs being almost inactive due to heterochromatinization in the somatic tissues (Gehring and Henikoff, 2007). Plants do not set aside germ lines early in embryogenesis and so plant gametes differentiate from the meristematic tissues of the flower. To restore the totipotent state in the zygote, epigenetic marks specific for the meristem have to be removed (Hsieh *et al.*, 2009; Calarco *et al.*, 2012) and restored during embryogenesis (Slotkin *et al.*, 2009; Ibarra *et al.*, 2012; Martínez *et al.*, 2016). TEs make use of this temporary deficiency of epigenetic control for transposition that can result in sex-specific chromosomal distribution if a TE is differently regulated between the male and female germ lines. While no female germ line-specific factors influencing the activity of TEs have been found yet, in the male germ line TE transposition can be prevented by pollen-specific TE silencing mechanisms based on small RNAs (Creasey *et al.*, 2014; Martínez *et al.*, 2017). TEs that are suppressed more efficiently in male gametes can then be found enriched on the X chromosome and depleted on the Y chromosome, exactly as is the case for many TEs in *S. latifolia* and *R. acetosa*. We have previously discussed this topic and propounded a model of sex-specific TE proliferation and its consequences in terms of chromosomal distribution of TEs that can be tested by cytological and bioinformatics approaches (Hobza *et al.*, 2017).

One exception to the rule might be represented by LINE elements: a non-LTR superfamily from the class I group of TEs. LINEs are accumulating on the Y chromosomes of *R. acetosa* (Table 4) and do not seem to involve many lineages preferring insertion into the X chromosome (Table 3, Fig. 3C). Kejnovsky *et al.* (2013) demonstrated that enrichment of AAC-containing microsatellites in the vicinity of LINE elements is 3.7 times higher than would be expected for randomly chosen chromosomal loci in *R. acetosa*. Moreover, he argued that TEs prefer DNA conformations adopted by microsatellite arrays. Therefore, the contribution of LINE elements to size increase in the Y chromosome is likely to be the result of insertional preference into AAC-containing satellites which, are exceptionally amplified on Y chromosomes (Fig. 2B, Supplementary Data Fig. S2). Nevertheless, targeting into micro- and minisatellite arrays may be a secondary factor responsible for accumulation on the Y chromosome in the case of most TE types in *R. acetosa*, because all TEs have a somewhat raised likelihood of insertion near satellites (Ramsay *et al.*, 1999; Kejnovsky *et al.*, 2013).

Localization of pseudoautosomal region

In contrast to the euchromatic X chromosome, both Y_1 and Y_2 have a heterochromatic nature (Lengerova and Vyskot, 2001).

On the other hand, a recent study showed the presence of functional genes on Y chromosomes in *R. acetosa* (Michalovova *et al.*, 2015). Little is known about the localization of potential gene regions on Y chromosomes and the pseudoautosomal region (PAR). Such information could help answer questions regarding the origin of the Y_1 and Y_2 chromosomes. Farooq *et al.* (2014) reported that during meiosis sex chromosomes of *R. acetosa* form a chain- or ring-shape trivalent (Y_1-X-Y_2). Our data support this observation since the RAAleII retrotransposon (Fig. 1I) sequence (highly conserved) occurs uniquely at the ends of the X and Y_1 chromosomes. In contrast with the RAAleII retrotransposon, the TatCL11 element is spread through all autosomes, X chromosomes and the terminal regions of the Y chromosomes (Steflova *et al.*, 2013). These results suggest that the PARs are localized in the distal parts of these sex chromosomes. So far this is the first report of a shared part of Y and X chromosomes in this species.

On the origin of the Y2 chromosome

The puzzling origin of the two Y chromosomes led to the formulation of two hypotheses. Firstly, that one Y chromosome was split into two Y chromosomes (Lengerova and Vyskot, 2001), and secondly that one of the Y chromosomes is a neo-Y chromosome, arising from the fusion of the X chromosome with an autosome. The latter scenario has already been confirmed in *R. hastatulus* (Smith, 1964; Grabowska-Joachimiak *et al.*, 2015; Kasjaniuk *et al.*, 2019) and we argue that it is most likely in *R. acetosa* as well, for the following reasons. All autosomes are submetacentric to acrocentric while sex chromosomes are clearly meta- or submetacentric, which is indicative of chromosome fusions. However, the unprecedented chromatid expansion that equalled chromatid size cannot be excluded. Also, species from the section *Acetosa* of the *Rumex* genus contain seven pairs of autosomes plus sex chromosomes, but most species of the other sections from the genus *Rumex* have nine or ten chromosomal pairs (Navajas-Pérez, 2005*a*, 2009). Additionally, we discovered that Y_2, the shorter of the two Y chromosomes, has fewer tandem repeats compared with the Y_1 chromosome. This unexpected observation was the result of the extensive analysis of satellites using FISH here (Fig. 1) and in our previous publication (Steflova *et al.*, 2013) and indicates that Y_2 had less time to accumulate satellites. Thus, Y_2 might be a neo-Y chromosome that has arisen on the base of section *Acetosa*. Nevertheless, the smaller size and relative satellite depletion of the Y_2 chromosome may reflect that it is in the shrinkage phase of its evolution (reviewed by Hobza *et al.*, 2015) and is actually the older Y chromosome. To evaluate these scenarios, future studies need to precisely assess the quantity of repeats on the Y_1 and Y_2 chromosomes of several section *Acetosa* species, such as *R. papillaris* and *R. thyrsiflorus*. Additional data can be obtained by looking also at sex-linked genes, their presence, genetic degeneration and transcript level, as shown by Hough *et al.* (2014). Unfortunately, such analyses are limited by the lack of a method to reliably map sex-linked genes to either the Y_1 or the Y_2 chromosome.

Sex-chromosome formation: a combination of a variety of effects

The widely accepted hypothesis predicts the accumulation of repeats in the non-recombining region of the Y chromosome (Charlesworth, 1991), but many repeats tend to have the opposite pattern of distribution. Cytogenetic as well as bioinformatic studies have proved not only that TEs are often absent on Y chromosomes but, even more interestingly, that many TE lineages have spread either on the X or the Y chromosome of dioecious species such as *S. latifolia* and *R. acetosa* (Cermak *et al.*, 2008; Filatov *et al.*, 2009; Steflova *et al.*, 2013; Kralova *et al.*, 2014; Puterova *et al.*, 2018). Here we demonstrated that this ‘sex-specific’ behaviour applies to most TE families and causes a substantial difference in sex chromosome TE composition, reaching 30 % of chromosome length (Table 3). This is probably the consequence of diverse and individualized mechanisms of TE regulation taking place during male and female gamete formation (Kubat *et al.*, 2014; Hobza *et al.*, 2017, 2018). Thus, besides reduced recombination levels and selective pressures, the evolution of sex chromosomes, in particular TE composition, is influenced by cellular processes that are primarily aimed at genome defence against deleterious activity of TEs in haploid phases, i.e. embryo sac and pollen grain development.

In *R. acetosa*, the X and Y chromosomes and autosomes represent three distinct genomic regions with unique repeat composition. The X and Y chromosomes both increase their size at a greater pace than the autosomes, but due to different reasons. The Y chromosomes undergo (1) expansion of satellites due to limited recombination and (2) male-preferentially active TEs. On the X chromosome, expansion of satellites is not elevated despite theoretically lower recombination in comparison with autosomes. On the other hand, the X chromosome is populated by female-preferentially active TEs. In contrast, accumulation of repeats is lowest on autosomes due to (1) recombination preventing the expansion of satellites and (2) exposure to the activity of mainly sex-specifically active TEs at a lower frequency than sex chromosomes as only half of the autosomes are present in the opposite sex (reviewed in Hobza *et al.*, 2017).

SUPPLEMENTARY DATA

Supplementary data are available online at https://academic.oup.com/aob and consist of the following. Figure S1: dot plot of BAC sequence with partially reconstructed Ogre/Tat LTR retrotransposon carrying tandem repeat RAE93. Figure S2: length distribution of selected micro- and minisatellite arrays. Script S1: script in R language used to analyse output of Tandem Repeat Finder and production of respective figures. Table S1: primers used for amplification of repetitive DNA and GenBank accession numbers of the sequences. Table S2: (A) comprehensive table of 319 clusters representing at least 0.01 % of the genome with estimation of abundance on autosomes, X and Y chromosomes of *R. acetosa*; (B) table of all manually annotated clusters. Table S3: table of all satellites identified in the *R. acetosa* genome using the RepeatExplorer pipeline. Table S4: comprehensive table of all micro- and minisatellites identified with Tandem Repeat Finder. Table S5: repeat annotation summary of *R. acetosa* male genome obtained from RepeatExplorer analysis of Illumina sequencing data.

Appendix D

digIS: towards detecting distant and putative novel insertion sequence elements in prokaryotic genomes

Janka Puterová and Tomáš Martínek

BMC Bioinformatics. 2021, 22:258, **DOI:** 10.1186/s12859-021-04177-6

Puterová and Martínek *BMC Bioinformatics* (2021) 22:258
https://doi.org/10.1186/s12859-021-04177-6

BMC Bioinformatics

SOFTWARE Open Access

digIS: towards detecting distant and putative novel insertion sequence elements in prokaryotic genomes

Janka Puterová and Tomáš Martínek*

*Correspondence:
martinto@fit.vutbr.cz
IT4Innovations Centre of Excellence, Faculty of Information Technology, Brno University of Technology, Bozetechova 2, 612 66 Brno, Czechia

Abstract

Background: The insertion sequence elements (IS elements) represent the smallest and the most abundant mobile elements in prokaryotic genomes. It has been shown that they play a significant role in genome organization and evolution. To better understand their function in the host genome, it is desirable to have an effective detection and annotation tool. This need becomes even more crucial when considering rapid-growing genomic and metagenomic data. The existing tools for IS elements detection and annotation are usually based on comparing sequence similarity with a database of known IS families. Thus, they have limited ability to discover distant and putative novel IS elements.

Results: In this paper, we present *digIS*, a software tool based on profile hidden Markov models assembled from catalytic domains of transposases. It shows a very good performance in detecting known IS elements when tested on datasets with manually curated annotation. The main contribution of *digIS* is in its ability to detect distant and putative novel IS elements while maintaining a moderate level of false positives. In this category it outperforms existing tools, especially when tested on large datasets of archaeal and bacterial genomes.

Conclusion: We provide *digIS*, a software tool using a novel approach based on manually curated profile hidden Markov models, which is able to detect distant and putative novel IS elements. Although *digIS* can find known IS elements as well, we expect it to be used primarily by scientists interested in finding novel IS elements. The tool is available at https://github.com/janka2012/digIS.

Keywords: IS elements, Mobile element, Profile HMM, Prokaryotic genomes, Genome annotation

Background

Insertion sequence elements (IS elements) are the smallest and most abundant autonomous transposable elements in prokaryotic genomes, usually ranging from 700 bp to 3 kbp. However, there are exceptions, and some IS families (Tn3) can contain elements having a length greater than 5 kbp. ISs are widespread in prokaryotic genomes and may occur in high copy numbers. They play an essential role in genome evolution, structure, and host-genome adaptability. Due to their movement ability, IS elements represent mutagenic agents and can: cause modulation of expression of neighboring genes, affect virulence, change xenobiotic or antimicrobial resistance, or modulate metabolic activities. Detailed information on IS element function in host genomes can be found in recent reviews [1, 2].

Typically, IS elements consist of one or two open reading frames (ORFs) encoding a transposase (Tpase), a product necessary for transposition within a particular genome or horizontally between genomes (in plasmids). They are flanked by short terminal inverted repeats (IRs) and direct repeats (DRs). Transposases occurring in IS elements include five groups named after amino acid residues located at their conserved catalytic domain that catalyzes the transposition: DDE, DEDD, HUH, Tyrosine (Y), and Serine (S). IS elements with DDE transposase are the most abundant, and their conserved catalytic domain has a typical secondary structure $\beta1 - \beta2 - \beta3 - \alpha1 - \beta4 - \alpha2/3 - \beta5 - \alpha4 - \alpha5/6$. Classification of IS elements into families is based mainly on Tpase structure, but other features such as IRs and DRs are also considered. Up to now, 29 IS families have been identified [1].

ISfinder [3] is a human-curated database and the most comprehensive source of known IS elements at present. Currently, the database contains more than 5000 entries and is updated regularly. As an extension of the ISfinder database, the authors implemented an ISbrowser interface [4] for visualization of IS elements inside genomes, and they prepared a benchmark dataset, consisting of 118 manually annotated prokaryotic genomes (as of November 2017), that is often used for assessment of IS detection tools performance. Another data source focused on mobile genetic elements, including manually annotated insertion sequences, is ACLAME database [5]. Unfortunately, this database has not been updated since 2009.

Even though the databases of known IS elements are growing, we are probably far from having a complete knowledge of all IS families and their structures. Therefore, for a better understanding of the IS elements function and their role in genome evolution, it is desirable to have an effective tool capable of not only annotating known families but also detecting new ones. This need becomes even more crucial when considering rapid-growing genomic and metagenomic data.

At present, there are several tools available for the detection of IS elements in prokaryotic genomes. Some of them are designed for searching in raw sequenced data (ISQuest [6], ISMapper [7], ISseeker [8], panISa [9]), and the others require assembled sequences (IScan [10], ISsaga [11], OASIS [12], ISEScan [13], TnpPred [14]). Almost all tools utilize a homology-based approach and are dependent on a source of known IS elements (they use a reference database either for verifying their results or for building searching profiles). Only the panISa tool detects IS elements solely based on structural features, such as an alignment of DR regions, and does not require a reference database.

Homology-based methods can be further divided into two main categories: (1) sequence-based and (2) profile-based methods. The first category is represented by tools IScan, OASIS, ISQuest, and ISseeker, which utilize the ISfinder database as a reference library in combination with BLAST software [15] to find close homologs. These tools are often used in annotation pipelines, where outputs with a high level of confidence are required.

The latter category includes ISsaga, TnpPred, and ISEScan. They take advantage of interpolated Markov models or profile hidden Markov models (pHMMs), which provide a more sensitive search, and detect remote homology sequences. ISsaga utilizes GLIMMER [16] and detects ORFs of IS elements or their fragments using an optimized interpolated Markov model built from the ISfinder database. TnpPred is focused on transposases detection (not full-length IS elements) and provides pHMMs for 19 of 29 IS families only. ISEScan uses 621 pHMMs built automatically from Tpases in the ACLAME database, but 355 of them are made up of one sequence only. Based on the configuration, ISEScan searches for whole Tpases or allow the presence of fragments.

Both sequence-based and profile-based tools can find new members of existing IS families, as they usually share significant sequence similarity either at the DNA or Tpase/ORF level. Profile-based methods are able to find remote members with lower similarity, which can represent hitherto undiscovered families—distant putative novel IS elements. However, the reliable identification of new IS families and their members is still challenging even for existing profile-based tools. It is mainly due to the Tpase structure, which comprises of several, often variable, domains. A search for the whole Tpase (ISEScan) is quite specific and unable to uncover novel IS elements with a distinct Tpase structure. On the other hand, allowing for fragments (ISEScan, ISsaga, and TnpPred) may result in many hits having significant similarity to a specific part of a completely different protein (i.e., false positives in terms of tool evaluation).

In this paper, we address the aforementioned challenge using a novel approach to detecting distant members of known IS families and putative novel IS elements. The fundamental idea is to search for the most conserved part of Tpase—the catalytic domain. The search is based on manually curated pHMMs with noise cutoff thresholds. Utilizing this approach, we can detect both known and putative novel IS elements with a moderate level of false positives while maintaining high sensitivity. The proposed method is implemented as *digIS* software and released as open-source at https://github.com/janka2012/digIS. The installed tool, including all dependencies, is also available as a docker image at https://hub.docker.com/r/janka2012/digis.

Implementation

digIS is a command-line tool developed in Python. It utilizes several external tools such as BLAST [15], HMMER [17], and Biopython library [18]. As an input, *digIS* accepts contigs in FASTA format. Optionally, the user can provide a GenBank annotation file for a given input sequence(s). This annotation is later used to improve the classification of identified IS elements (see “Output classification” section).

Firstly, we built a library of manually curated pHMMs, corresponding to Tpase catalytic domains of individual IS families. As a source of sequences, we used the ISfinder database, and for each pHMM, we identified the noise cutoff threshold.

Then, the *digIS* search pipeline operates in the following way:

1 The whole input nucleic acid sequence is translated into amino acid sequences (all six frames).
2 The translated sequences are searched using manually curated pHMMs.

3 Found hits, referred to as *seeds*, are filtered by domain bit score and e-value. Those that overlap or follow one another within a certain distance are merged.
4 Each seed is matched against the database of known IS elements (ISfinder) and its genomic positions are extended according to the best hit.
5 Extended seeds are filtered by noise cutoff score and length. Duplicates, corresponding to the same IS element, are removed.
6 Remaining extended seeds are classified based on sequence similarity and GenBank annotation (if available) to assess their quality.
7 Finally, the classified outputs are reported in the CSV and GFF3 format.

The overall *digIS* workflow is depicted in Fig. 1, and the individual steps are described in detail in the following sections.

Building profile hidden Markov models for the transposase catalytic domain of individual IS families

Tpase sequences were obtained from the ISfinder database. For each IS family, the pHMM was created as follows: (1) the longest ORF sequence, representing Tpase and its catalytic domain, was chosen for each IS element[1], (2) a multiple sequence alignment (MSA) for a set of Tpases belonging to the same family was created by Clustal Omega [19] and visualized using Jalview [20], (3) for each MSA, a protein secondary structure of the transposase was predicted using JPred4 [21] and used to determine the boundaries of the conserved catalytic core; the MSA was refined based on the positions of the catalytic residues (usually DDE), and the catalytic domain was manually cut using these determined boundaries, (4) such a manually modified MSA was used to construct resultant pHMM using *hmmbuild* from the HMMER package.

Since IS3, IS4, and IS5 families contain multiple subfamilies, a separate model was constructed for each of them. Moreover, IS5/IS5 and IS5/None subfamilies showed various sequence patterns (e.g., long insertions, deletions), and therefore several models were built for them concerning these patterns. MSAs with highlighted sequence groups used to construct these models are available in Additional files 1 and 2. For the ISNCY family, models were built for IS1202 and ISDol1 subfamilies only, since other subfamilies did not contain a sufficient amount of sequences. We required the models to be assembled from at least ten sequences to have a generalizing ability to find distant Tpases. Altogether, 50 pHMMs were constructed.

The remaining sequences of IS5 and ISNCY subfamilies representing outliers/distant sequences were cut with regard to the catalytic residues and secondary structure. They were used later as *individual* protein sequences in *phmmer* search. Overall, 70 outlier sequences were collected.

To eliminate false-positive hits reported by HMMER using pHMMs and still have the ability to detect distant and novel IS elements, a domain noise cutoff threshold—which represents a bit score of the highest-scoring known false positive—was

[1] Various IS families carry Tpase consisting of multiple ORFs. These ORFs are present in the ISfinder database in both individual and fusion forms. As duplicated sequences may lead to a bias in pHMMs, only the longest ORF sequence was used.

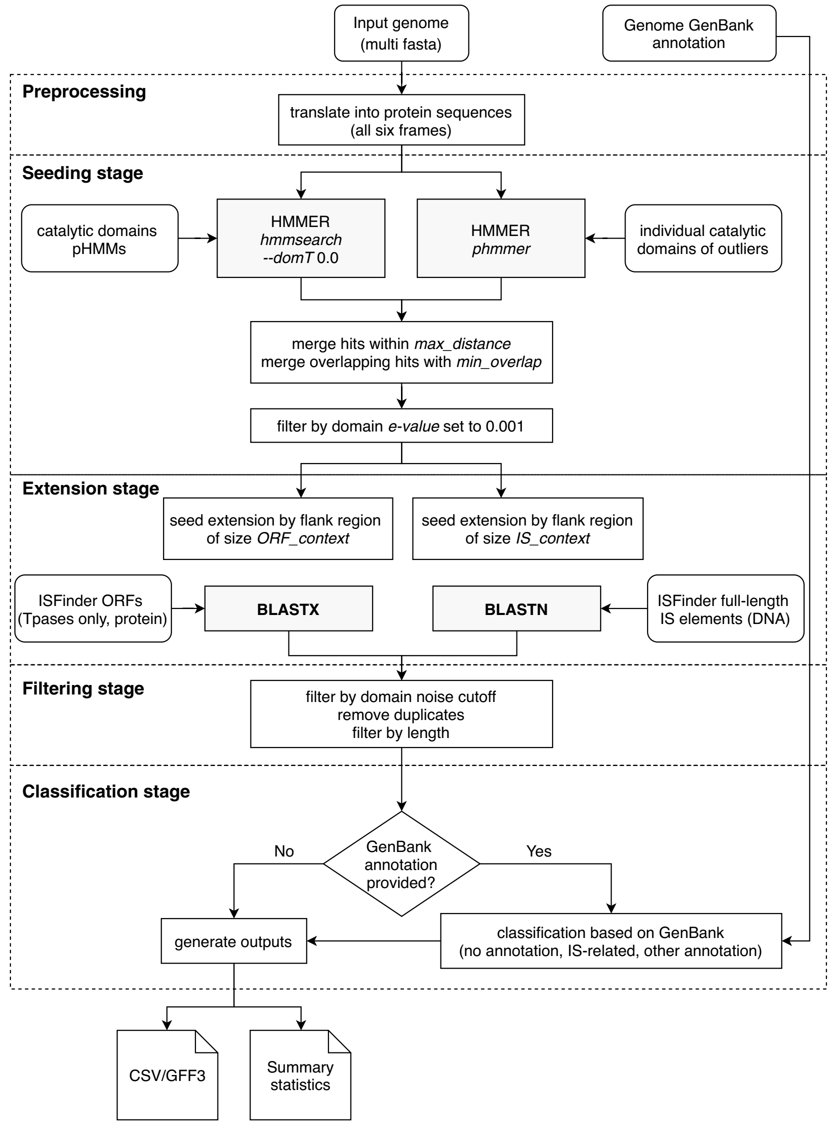

Input genome
(multi fasta)
Genome GenBank annotation
Preprocessing
translate into protein sequences (all six frames)
Seeding stage
catalytic domains pHMMs
HMMER hmmsearch --domT 0.0
HMMER phmmer
individual catalytic domains of outliers
merge hits within max_distance
merge overlapping hits with min_overlap
filter by domain e-value set to 0.001
Extension stage
seed extension by flank region of size ORF_context
seed extension by flank region of size IS_context
ISFinder ORFs (Tpases only, protein)
BLASTX
BLASTN
ISFinder full-length IS elements (DNA)
Filtering stage
filter by domain noise cutoff
remove duplicates
filter by length
Classification stage
No
GenBank annotation provided?
Yes
generate outputs
classification based on GenBank (no annotation, IS-related, other annotation)
CSV/GFF3
Summary statistics

Fig. 1 Workflow of *digIS*. digIS components and workflow, grey rectangles represent external tools, rounded rectangles represent input data, white rectangles represent digIS components

determined for each pHMM as follow: First, a database of manually curated protein sequences from Archaea and Bacteria kingdoms was collected from SwissProt [22] and RefSeq [23] databases (records labeled as 'REVIEWED'), resulting in 353051 and 232157 records (accessed on 11 March 2019), respectively. Setting this threshold is a common practice and is used, for example, in models stored in Pfam [24] database. Then, each pHMM was queried against this reference protein database employing *hmmsearch* with default settings. Finally, reported hits were sorted in a descending order based on the reported per-domain bit score and evaluated manually to estimate the bit score from which false positive hits were prevalent.

Searching for IS elements in the input sequence

In the beginning, the whole input nucleic acid sequence is translated into amino acids (all six frames). Then, the search process operates in two steps:

1. *Seeding*: The input genome is scanned using pHMMs and *individual* sequences representing Tpase catalytic domains. Each occurrence with a satisfactory score is labeled as a seed.
2. *Extension*: The genomic position of seeds identified in the previous step are extended based on the similarity boundaries with Tpases and IS elements from the ISfinder database.

In the *Seeding* stage, *digIS* utilizes *hmmsearch* from the HMMER3 package to query pHMMs against the translated sequences with an enabled domain threshold (*–domT* argument) set to 0.0 to report domain hits with a non-negative bit score only. Afterwards, *digIS* employs *phmmer* to query *individual* protein sequences against the translated sequences. The resulting hits are post-processed and filtered by a domain conditional e-value set to 0.001. Next, neighboring records, detected by the same model within a certain distance (700 bp[2]) on the same strand, are merged. This approach allows insertions or variable segments inside catalytic domains that are typical for some Tpases [25]. Next, overlapping records found by different models are merged, since there exists a sequence similarity in the catalytic domain among different Tpases, or a putative novel catalytic domain might be composed of different parts of known domains.

Please note that *digIS* scans the whole input sequence, instead of just open reading frames (ORFs), to not omit some coding regions.

During the next stage (*Extension*), the genomic position of each seed is identified in the original nucleic acid sequence and extended with *context_orf* and *context_dna* (upstream and downstream flank regions of a length 1600 bp[3], and 14000 bp[4], respectively), see Fig. 2. Next, the extended seed is matched against sequences of known Tpases (ORF level) and IS elements (DNA level), extracted from the ISfinder database, using the BLASTX and BLASTN tools. Finally, the seed's original position is adjusted (extended) according to the best BLAST hits' positions.

As the output of the *Extension* stage, the *digIS* tool reports: (1) position at DNA level if the similarity with a known IS element was found using the BLASTN tool; or (2) position at the ORF level if the similarity with a known Tpase was found using the BLASTX tool; or (3) position of the catalytic domain otherwise found during the *Seeding* stage.

[2] Merge distance 700 bp was identified based on the longest gaps within the models (see Additional file 3 for more details).

[3] ORF context size 1600 bp was identified based on the length of the longest transposase ORF in the ISfinder database divided by 2, multiplied by 3 (conversion from amino acids to nucleotides) and rounded up to the nearest hundredth

[4] DNA context size 14000 bp was identified based on the length of the longest IS element in the ISfinder database divided by 2 and rounded up to the nearest hundredth

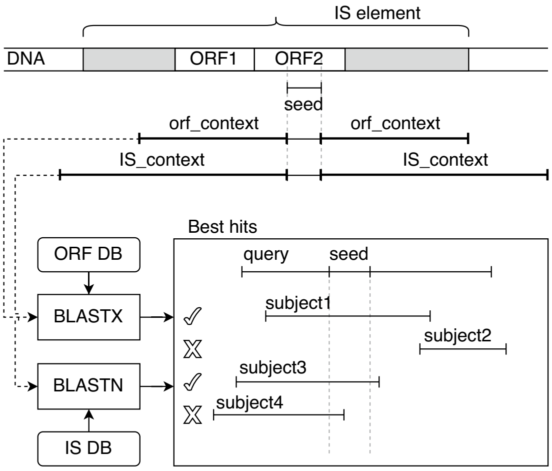


Fig. 2 Seed extension process. The seed matches the catalytic domain of the putative IS element in the input genomic sequence. This seed is extended with upstream and downstream flank regions of orf_context and IS_context size, respectively, and is searched against the database of known Tpase/ORFs and IS elements, respectively. Only the best hits, including the whole original seed, are considered for extension. Position of the seed is changed (extended) according to the best hit

Output filtering

To eliminate the number of reported false positives, *digIS* filters the hits with a score below the previously estimated noise cutoff threshold, and it removes duplicate records covering the same genomic region. Lastly, hits having less than 150 bp (50 aa) in length are filtered out.

Output classification

To help the user assess the quality of found IS elements, each output hit is supplemented by information about sequence similarity with known IS elements and Tpases extracted from the ISfinder database. The similarity is calculated as a percentage of identity between the extended seed and a known IS element or Tpase sequence, measured according to the database item's length.

In case the GenBank annotation is provided as an optional input[5], the classification process is further extended, and each *digIS* hit is classified based on the overlap with GenBank annotation records into the three categories using following rules applied in the subsequent order:

- *IS-related*—hit overlaps with a GenBank record of type: (1) mobile element or mobile element type, (2) repeat region, coding sequence (CDS), gene, or miscellaneous feature annotated as transposase, resolvase, recombinase, recombination/resolution, insertion element, mobile element, transposon, transposable element, DDE, or the

[5] GenBank annotation is a result of a complex process [26] that utilizes sources of manually curated data and automatically predicted ones with a high level of confidence.

annotation contains a name of known IS family or subfamily [27, 28]. A hit classified into this category has high confidence to be a true IS element.

- *no annotation*—hit does not overlap with any GenBank record or overlaps with a record annotated as a hypothetical protein, predicted protein, unknown, or domain of the unknown function (DUF). The hit in this category can be seen as an unknown protein or protein, where the annotation pipeline did not achieve a sufficient level of confidence. Typically, distant or putative novel IS family members may belong to this category.
- *other annotation*—otherwise. The hit in this category is probably not an IS element, because it overlaps and shares significant similarity with a different protein.

Since the previous analysis of GenBank annotation revealed that some IS element transposases were misannotated as integrases [6, 12], we classify all hits annotated as integrases and at the same time having significant identity to a known IS element in the ISfinder database (at ORF or DNA level), as *IS-related* as well.

The latest version of the GenBank annotation was newly expanded to include fragments of IS elements marked as 'pseudo' with the notation 'incomplete' [26]. To preserve a conservative approach and high confidence, these records are ignored when classifying hits.

digIS output files

The *digIS* tool generates the following output files: (1) a CSV and GFF3 file containing all found IS elements and their attributes such as sequence ID, genomic location, strand, accuracy, score, sequence similarities with known IS elements (at ORF and DNA level), and classification according to GenBank annotation (if provided); (2) a summary file containing numbers of IS elements per individual families, overall numbers of base pairs and a percentage of an input sequence occupied by IS elements. FASTA sequences of found IS elements can be extracted using the GFF3 file and BEDTools [29] (see instructions on the GitHub repository).

Results

The performance of the *digIS* tool was evaluated on different datasets and compared with related tools. Specifically, we chose ISEScan (version 1.6), OASIS (version released 18th September 2012), and ISsaga (version with the last update on 20th January 2020). Other state-of-the-art tools were excluded for various reasons. ISMapper, ISseeker, ISQuest, and panISa are designed for IS elements detection in raw sequence reads. TnpPred is available online only, and it is limited to protein sequences with a maximum length of 5000 amino acids. Even though the TnpPred pHMMs are available for download, it is unclear what kind of parameters or filtration mechanisms should be used during the search. Finally, we excluded IScan, because we were not able to install it, including all necessary dependencies.

All tools were run with default or recommended settings. Additionally, ISEScan was executed with two settings: (1) default configuration with the *removeShortIS* option enabled, when IS elements shorter than 400 bp or single copy IS elements without perfect

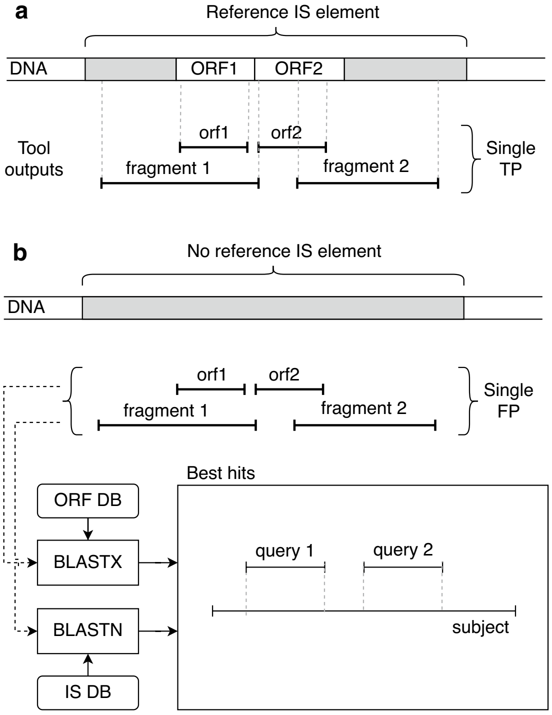


Fig. 3 Definition of true positive and false positive with respect to the tools reporting IS element's fragments or ORFs. **a** Definition of true positive: A hit is considered as a true positive (TP) if it overlaps with any item in the reference dataset. If multiple hits overlap with the same IS element in the reference dataset, then all of these hits count as one hit only. **b** Definition of false positive: Each reported hit of the tool is matched against the database of known IS elements (ISfinder), and if several adjacent hits map to the same IS element at DNA or ORF level, then these hits are counted as only one merged FP

IRs are filtered out; and (2) with *removeShortIS* turned off when all hits are reported (hereinafter referred to as ISEScan–fragments).

We faced several issues when evaluating the tools. At first, the definition of a true positive hit was ambiguous as different tools reported different types of outputs. Some tools reported entire IS elements at the DNA level (ISEScan and OASIS) or their fragments (ISEScan–fragments). Other tools reported individual ORFs or fragments thereof (ISsaga), while the proposed *digIS* tool reported outputs at one of three levels (catalytic domain, ORF, or DNA). Moreover, for tools reporting ORFs or fragments, it is common that several hits correspond to the same IS element from the reference dataset.

Considering these facts and in an effort to evaluate the tools fairly, reported hits were classified as follows: A hit is considered as a true positive (TP) if it overlaps with any item in the reference dataset, and the length of the overlapping region is $\geq$ 100 bp[6]. If multiple hits overlap with the same IS element in the reference dataset, then all of these

[6] Usually, an overlap based on a percentage of the reference IS element length is used in other studies, but when allowing for fragments, this criterion is not applicable. The requirement for at least 100 bp overlap seems to work well, even when two neighboring IS elements overlap.

hits count as one hit only (as shown in Fig. 3a). A false negative (FN) is defined as a reference dataset element without sufficient overlap with at least one reported hit. A false positive (FP) represents a reported hit without sufficient overlap with at least one item from the reference dataset.

It turns out that some reference datasets may not be complete. For example, if a new IS element is discovered, it is not included in a previously published dataset. A hit matching this new IS element is considered as an FP, even if it was identified correctly by the tool (see "Evaluation on the benchmark ISbrowser and E. coli datasets" section). The number of FPs is then even higher for tools reporting ORFs or fragments of the same IS element. To minimize this side effect, each FP was compared with a database of known IS elements (ISfinder). If several adjacent FPs mapped to the same IS element at DNA or ORF level (as shown in Fig. 3b), they were counted as one merged FP (mFP).

Evaluation on the benchmark ISbrowser and *E.coli* datasets

The first evaluation of the selected tools was performed on two benchmark datasets (1) a human-curated dataset from ISbrowser, and (2) the IS element annotation of *Escherichia coli* strain K-12 substr. MG 1655 genome [30]. The ISbrowser dataset comprises an annotation of 118 prokaryotic genomes (as of November 2017); 58 of them contain full-length IS elements, including 36 prokaryotic genomes and 22 plasmids. *E.coli* strain K-12 is one of the most well-understood model organisms [31] and is frequently used in microbial studies. The dataset of annotated IS elements for *E.coli* was obtained from the ISEScan publication (Supplementary Materials, Table 5) since EcoGene 3.0 [31], a source devoted to the structural and functional annotation of *E.coli* strain K-12, was unavailable at the time of manuscript preparation. This dataset consists of 49 IS elements of which 40 are full-length.

Results for the ISbrowser and *E.coli* datasets are shown in Table 2. Surprisingly, all tools showed a relatively high number of FPs and corresponding FDR (in the range from 8 to 24%). Therefore, we analyzed the FPs in more detail as follows: First, FPs representing fragments/ORFs of the same IS element were merged as described at the beginning of this section. Then, for each merged FP (mFP), the similarity with known IS elements in the ISfinder database was measured and by using the GenBank annotation it was classified into *IS-related*, *no annotation*, or *other annotation* category as described in "Output classification" section. Based on these results, a histogram was plotted depicting the number of mFPs as a function of similarity at both ORF and DNA levels. Finally, each bar in the histogram was divided according to the classification based on the GenBank annotation. These histograms represent an effective way to visualize the outputs of individual tools, including the identification of areas in which the tool makes errors. Please, see Additional file 4: "ISbrowser dataset" section.

In summary, many mFPs correspond to the hits that are highly likely to represent true IS elements that are not yet included in manually curated datasets. This behavior can be caused by the fact that the human-curated, whole-genome annotation might not be updated as often as databases of known IS elements. The exact numbers of true IS elements are unknown even in human-curated datasets and may evolve over time. Therefore, the common performance metrics, such as the confusion matrix, can not evaluate the tool quality fairly.

Table 1 Thresholds for classification based on the sequence similarity

Level/interpretation	Improbable member	Inter-family member	Intra-family member
IS element	SeqID < 50%	50% < SeqID ≤ 70%	70% < SeqID
Tpase/ORF	SeqID < 25%	25% < SeqID ≤ 45%	45% < SeqID

To address this issue, we decided to classify mFP hits further to distinguish between those representing IS elements with a high level of evidence and improbable/not IS elements. For these purposes, we used the GenBank annotation, which resulted from a conservative approach combining manually curated data and automatically predicted ones with a high level of confidence. Each mFP hit was classified according to the rules described in the "Output classification" section. Therefore, mFPs classified as *IS-related* can be highly likely considered as IS elements or their parts. Similarly, mFPs classified as *other annotation* can be regarded as improbable or not IS elements since they include parts that have been conservatively identified as other protein products.

The remaining hits classified as *no annotation* can be seen as unknown IS elements or those where the GenBank annotation pipeline has not achieved a sufficient level of confidence. To evaluate these outputs, additional information about sequence similarity with the database of known sequences (ISfinder) was used. Since the IS elements are divided into several independent families, it is difficult to find the exact boundary between IS and non-IS elements for mFPs. It is more appropriate to divide them into three categories:

- *Intra-family member*—a hit having similarity to the extent that is typical for members belonging to the same family.
- *Inter-family member*—a hit having similarity that is common among members of different families.
- *Improbable member*—a hit having similarity lower than usual among family members.

Although there may be several ways to categorize mFPs into these groups, we have chosen a more straightforward approach by defining two similarity thresholds (at the ORF and DNA level) that divide hits into these three categories. To determine the thresholds, a database of known IS elements (ISfinder) was used, the sequence similarities common within existing families and among them were measured, and these values were averaged. The resulting thresholds and their interpretations are given in Table 1. A detailed description of the procedure and the measured data is available in Additional file 5.

In summary, using the GenBank annotation and sequence similarity, the mFPs were classified into three categories according to the following rules:

- *IS element with a high level of evidence (eIS)*—a hit classified as *IS-related* based on the GenBank annotation, or a hit classified as *no annotation* based on the GenBank and *Intra-family member* based on the sequence similarity.
- *Distant or putative novel IS element (pNov)*—a hit classified as *no annotation* based on the GenBank and *Inter-family member* based on the sequence similarity.

- *Improbable or not an IS element (nIS)*—a hit classified as *other annotation* based on the GenBank annotation, or a hit classified as *no annotation* based on the GenBank and *Improbable member* based on the sequence similarity.

Distribution of mFP entries into these three categories is presented in Table 3, columns labeled as *Detailed classification of mFPs*. It can be seen that a large part of the hits initially classified as mFPs falls into the category *IS element with a high level of evidence*. Together with previously identified TPs, they represent the total number of IS elements with a high level of evidence (teIS). Consequently, only the hits in the nIS category are considered to be incorrectly identified by the tool (i.e. false positives). Based on these new metrics, the putative novel discovery rate (pNovDR), and nIS discovery rate (nISDR) were calculated representing the proportion of putative novel and improbable/not IS elements in reported outputs, respectively. Finally, the pNov/nIS ratio was calculated to express how many putative novel elements are found per single incorrectly identified hit.

We presume that these modified metrics reflect the tools' performance better since they address the issue of incomplete reference datasets. Concurrently, they are based on sequence similarity information with known IS elements (ISfinder) and state-of-the-art annotations with high confidence (GenBank). We are aware of possible discussions and alternatives towards defined classification rules and similarity thresholds. However, if they are applied to all tools equally, they can bring a more reliable image of their performance.

The results in Table 3 related to the ISbrowser dataset show that:

- The tools that detect both full elements and fragments (ISsaga and ISEScan–fragments) can find the highest number of teISs. On the other hand, the reported hits include the highest number of nISs. The overall nISDR is around 9%, and the ratio between pNovs and nISs is low (0.15 and 0.22).
- OASIS found the lowest number of teISs and nISs (nISDR is 1.15%), making it the most conservative tool of all. OASIS found only the hits with a high level of confidence. The output primarily includes records of known IS elements, whereas putative novel elements are rare (0.69%).
- ISEScan is the second most conservative tool in terms of the number of teISs and nISs. Surprisingly, it found even less pNovs compared to the OASIS tool.
- With respect to the number of teISs and nISs, *digIS* falls in the middle between conservative (OASIS and ISEScan) and fragment-reporting tools (ISEScan–fragments and ISsaga) representing a tool with good sensitivity (0.82) and low nISDR (3.58%). Moreover, the number of pNovs is even higher than for ISEScan–fragments. Although ISsaga found one-third more pNovs than *digIS*, it was at the cost of three times more nISs.

The tools show a similar performance on the *E.coli* dataset. However, some characteristics are violated; for instance, none of the tools found any putative novel element, and nISDR is more than double for most tools. These discrepancies are primarily caused by a too small *E.coli* dataset (a single genome with less than 50 IS elements), where some of the metrics are calculated from fewer than ten items. Similar distortion can also be seen in the ISbrowser dataset, where the numbers of pNovs and nISs are too small for the OASIS tool. It results in a disproportionately high pNov/nIS ratio.

Evaluation on the NCBI Archaea and Bacteria datasets

In the next step, tools were evaluated on much larger datasets to verify the characteristics observed in Table 3 and to specify those affected by the small number of samples. We prepared two additional datasets containing complete archaeal and bacterial genomes from the NCBI GenBank database [32]. In the case of Archaea, all 341 genomes available in the database were used (accessed on 15th June 2019). In the case of Bacteria, 2500 from 14418 available genomes were randomly selected (see Additional file 6 for detailed information about these datasets). Since OASIS could not process 25 bacterial genomes, these were excluded. Altogether, 2475 bacterial genomes were evaluated.

Unlike the ISbrowser and *E.coli* datasets, the manually curated positions of IS elements are not available. Therefore, all hits reported by the tools were considered as FPs and the detailed classification process of FPs described in "Evaluation on the benchmark ISbrowser and E. coli datasets" section was applied. To verify the accuracy of this evaluation method, it was applied to the ISbrowser dataset first. Table 4 shows the number of hits found by the tool (N), the number of merged FPs (mFPs), the output of the classification process (number of eISs, pNovs, and nISs), and an assessment in terms of pNovDR, nISDR, and pNov/nIS ratio. As the number of TPs is not available, the teIS is reduced to eIS.

By comparing the evaluation results for the ISbrowser dataset with and without human-curated annotation (Tables 3, 4), certain differences can be seen. Detailed analysis revealed that these changes arose primarily because the ISbrowser reference dataset contains not only full-length elements, but also annotated fragments of various lengths (a total of 127 fragments). If a tool finds some of these fragments, they are distributed among the categories eIS, pNov, and nIS based on the GenBank annotations and similarities with the ISfinder database. This behavior causes the number of pNovs and nISs to increase at the expense of the total number of eIS. As a side effect, the pNovDR, nISDR, and pNov/nIS ratio are slightly higher. The small changes can also be observed in the histograms (see Additional file 4: "ISbrowser dataset without reference" section), but their overall character remains the same. Considering these subtle differences, it is possible to conclude that the above-described classification allows us an assessment of the tool performance, even when the manually curated annotation is not available.

The results on large NCBI GenBank Archaea and Bacteria datasets in Table 4 confirmed the tools' characteristics seen on the ISbrowser dataset. Only the following differences were observed:

- The proportion of nISs in the outputs is higher compared to the ISbrowser dataset. For ISEScan and *digIS*, the nISDR is approximately twice as large on the Archaea dataset. ISsaga achieved the highest nISDR (around 20%) for both Archaea and Bacteria datasets. A detailed analysis of the hits revealed that this is primarily due to the higher number of items classified as *other annotation*. A list of the most common GenBank record products that overlapped with these hits is given in Additional file 7.

- Larger NCBI datasets enabled to assess the ratio between pNov and nIS for OASIS more accurately, as it was affected by a small number of items in the *E.coli* and ISbrowser datasets before. This ratio decreased significantly to 0.27 and 0.21. Also, the number of pNovs found by OASIS is no longer higher than those found by the ISEScan tool.
- The histograms depicting the similarity of the outputs with the ISfinder database and their classification according to the GenBank annotation show the same characteristics as for the ISbrowser dataset, except for minor deviations (see Additional file 4: "NCBI Archaea and Bacteria datasets without reference" section).

In summary, tools that also detect fragments (ISsaga and ISEScan–fragments) can identify the most eISs, but at the cost of a large number of nISs. On the other side of the spectrum are conservative tools (OASIS and ISEScan), which show the lowest numbers of nISs, but also eISs. The performance of the proposed *digIS* tool in terms of eISs is closer to fragment-reporting tools, and at the same time, it achieves the number of nISs closer to conservative tools. Moreover, *digIS* is dominant in finding distant/putative novel IS elements with respect to the numbers of nISs (pNov/nIS ratio). This feature is significant, especially on large datasets (NCBI GenBank Archaea/Bacteria), where the *digIS* tool shows the best performance. Please note that *digIS* found even more putative novel elements than the ISEScan–fragments in these datasets.

Discussion

In this work, we focused on the detection of putative novel IS elements and aimed to find the sequence and structural features common to more IS families. The Tpases are generally considered as the most conserved parts of IS elements. Their structural variability is used as a major feature for their classification into the families [1]. On the other hand, the Tpase catalytic domain and its secondary structure are often preserved among the families [25]. Unfortunately, the accuracy of state-of-the-art tools for secondary structure prediction is not sufficient when applied to a single sequence and MSA is usually required for a more accurate prediction [33].

For this reason, we decided to make a compromise between detecting the general structure and sequence features. We built the library of manually curated pHMMs of a catalytic domain only (not whole transposase). The results of comparing *digIS* with other tools confirmed that the search based on the catalytic domain is sufficiently specific for the area of IS elements. The number of IS elements with a high level of evidence is comparable to fragment-reporting tools, while many improbable/not IS elements are filtered out. To better understand the effectiveness of the catalytic-domain-search technique compared to using the pHMM of the whole Tpase sequence, we performed a detailed analysis of individual tools' outputs. We focused on hits classified as "other annotation" according to the GenBank annotation, i.e., the records erroneously identified by the tool as IS elements or their parts. We analyzed overlapping GenBank records for these hits and created a histogram showing the number of occurrences for each type of protein or product (see Additional file 7).

Table 2 Performance of OASIS, ISEScan, ISEScan-fragments, ISsaga, and digIS on manually curated datasets

Tool	TP	FN	FP	Se	FDR (%)
Dataset ISbrowser (N = 1192)					
OASIS	791	401	77	0.66	8.87
ISEScan	925	267	94	0.78	9.22
ISEScan-fragments	1077	115	248	0.90	18.71
ISsaga	1135	57	363	0.95	24.23
digIS	979	213	194	0.82	16.54
Dataset *E. coli* (N = 49)					
OASIS	26	23	4	0.53	13.33
ISEScan	43	6	8	0.88	15.69
ISEScan-fragments	45	4	18	0.92	28.57
ISsaga	48	1	29	0.98	37.66
digIS	43	6	11	0.88	20.37

TP, FN, and FP represent the number of True Positives, False Negatives, and False Positives, respectively; Se is sensitivity; FDR is False Discovery Rate.

Table 3 Detailed analysis of false positives of digIS, ISEScan, OASIS, and ISsaga on manually curated

Tool	Common metrics		Detailed classification of mFPs				Modified metrics			
	TP	FP	mFP	eIS	pNov	nIS	teIS	pNovDR (%)	nISDR (%)	pNov/nIS
Dataset ISbrowser (N = 1192)										
OASIS	791	77	75	59	6	10	850	0.69	1.15	0.60
ISEScan	925	94	94	69	3	22	993	0.29	2.16	0.14
ISEScan-fragments	1077	248	239	103	18	118	1179	1.37	8.97	0.15
ISsaga	1135	363	323	148	31	144	1282	2.13	9.88	0.22
digIS	979	194	194	130	22	42	1108	1.88	3.58	0.52
Dataset *E. coli* (N = 49)										
OASIS	26	4	4	4	0	0	30	0.00	0.00	0.00
ISEScan	43	8	7	3	0	4	46	0.00	8.00	0.00
ISEScan-fragments	45	18	17	6	0	11	51	0.00	17.74	0.00
ISsaga	48	29	28	10	0	18	58	0.00	23.68	0.00
digIS	43	11	11	6	0	5	49	0.00	9.26	0.00

N represents the number of outputs found by the tool; mFP represents the number of False Positives after merging fragments or ORFs referencing the same IS element; eIS, pNov, and nIS represent the number of mFPs classified into categories IS element with a high level of evidence, Distant or putative novel IS element, and Improbable or not an IS element, respectively; pNovDR is putative Novel Discovery Rate; nISDR is Improbable or not an IS element Discovery Rate, and pNov/nIS shows the ratio between the number of putative novel IS elements and improbable or not an IS elements.

From the generated histograms, it can be observed that *digIS* generally reports a small number of records classified as "other annotation", which is comparable to conservative tools such as OASIS or ISEScan (see Additional file 7; Tables 1, 2, 3). On the other hand, tools that also report fragments (ISsaga and ISEScan–fragments) show a large number of these hits. If we focus on the annotations of these records, it can be seen that they usually represent products functionally related to transposases or parts thereof, such as *DNA-binding protein*, *ATP-binding protein*, *transcriptional regulator*, or *helix-turn-helix domain-containing protein*. In addition, both fragment-reporting tools (ISsaga and ISEScan–fragments) cover a large number of products that were not observed by other tools,

Table 4 Performance of digIS against ISEScan, OASIS, and ISsaga on NCBI GenBank datasets

Tool	**N**	**Detailed classification of mFPs**				**Modified metrics**		
		mFP	**eIS**	**pNov**	**nIS**	**pNovDR (%)**	**nISDR (%)**	**pNov/nIS**
Dataset ISbrowser (N = 1192)								
OASIS	895	852	828	10	14	1.17	1.64	0.71
ISEScan	1006	993	954	9	30	0.91	3.02	0.30
ISEScan-fragments	1326	1283	1089	41	153	3.20	11.93	0.27
ISsaga	1786	1459	1188	75	196	5.14	13.43	0.38
digIS	1170	1157	1051	50	56	4.32	4.84	0.89
Dataset NCBI Archaea (341 genomes)								
OASIS	5885	5789	5382	100	307	1.73	5.30	0.33
ISEScan	8404	8266	7532	207	527	2.50	6.38	0.39
ISEScan-fragments	12,016	11,550	9622	472	1456	4.09	12.61	0.32
ISsaga	17,698	14,788	10,946	822	3020	5.56	20.42	0.27
digIS	10,607	10,548	8640	728	1180	6.90	11.19	0.62
Dataset NCBI Bacteria (random selection of 2475 genomes)								
OASIS	88,552	87,428	83,992	1176	2260	1.35	2.58	0.52
ISEScan	111,974	110,357	102,266	3274	4817	2.97	4.36	0.58
ISEScan-fragments	151,540	145,248	119,392	6096	19760	4.20	13.60	0.31
ISsaga	217,345	181,880	136,903	8479	36,498	4.66	20.07	0.23
digIS	134,851	132,877	118,805	6722	7350	5.06	5.53	0.91

N represents the number of outputs found by the tool; mFP represents the number of False Positives after merging fragments or ORFs referencing the same IS element; eIS, pNov, and nIS represent the number of mFPs classified into categories IS element with a high level of evidence, Distant or putative novel IS element, and Improbable or not an IS element, respectively; pNovDR is putative Novel Discovery Rate; nISDR is Improbable or not an IS element Discovery Rate, and pNov/nIS shows the ratio between the number of putative novel IS elements and improbable or not an IS elements.

including *digIS*, such as *chromosomal replication initiator protein DnaA*, *DNA replication protein DnaC*, or *primosomal protein DnaI*. Detailed analysis revealed that portions of these proteins have significant sequence similarity to the coding segments of IS elements of the IS21 family (see Additional file 7). These examples show that searching for any fragments of IS elements can lead to a large number of false hits, which the application user must manually check. On the other hand, focusing the search on the catalytic domain can effectively filter these hits and, unlike conservative methods reporting full-length elements only, it provides a space for searching for putative novel IS elements.

When comparing the tools without a manually curated reference dataset or an incomplete one, the histogram—showing the number of outputs depending on the similarity to the database of known elements (ISfinder) and GenBank annotation—is a useful indicator of the tool's quality. It offers an independent view of the characteristics of the outputs and clearly shows, for example, the degree of tool conservation or tendency to detect other genes, that is typical for fragment-reporting tools (ISsaga and ISEScan–fragments). It also allows the identification of various anomalies in the GenBank annotation itself (see Additional file 4).

Despite the histogram's benefits, it does not allow us to easily quantify and compare the performance of the tools. The comparison is possible only if the outputs are classified into distinct categories such as TPs, TNs, FPs, FNs using manually curated benchmark datasets. In this paper, we were the first to point out the drawbacks of this approach when applied to existing tools for IS elements detection. We addressed the issue of different outputs of individual tools (full-length elements vs. fragments/ORFs). Based on a detailed analysis (see Additional file 4), we have shown that the benchmark datasets themselves are not complete, and therefore their use may skew the evaluation results.

To overcome these issues, we have chosen an alternative classification of the tools' outputs that relies on GenBank annotation and sequence similarity with the database of known elements (ISfinder). This approach allowed us to identify a group of IS elements with a high level of evidence (eIS) and a group of Improbable or not IS elements (nIS) in the category of presumed false positives. Also, since the boundary between these two groups is not strictly defined, there is a space for the putative novel IS elements group (pNov), which is the main interest of this article. We are aware that the definition of these categories is unambiguous and should be replaced by a high-quality and consistently maintained benchmark dataset in the future. On the other hand, the boundary between the groups of pNovs and nISs will probably be the subject of debate for a long time, as its precise definition would require a knowledge of all non-IS elements.

We experimented, for example, with a different definition of pNov and its effect on tools performance. Currently, pNov is defined as a sequence without a sufficiently specific GenBank annotation, having the sequence similarity that is common among members of different IS families. Without further restrictions, this category may include, for example, the found accessory genes or some of the transposase's variable domains. To make sure that the found hit is highly likely functional from a transposition point of view, it would be appropriate to require the presence of Tpase and its catalytic domain. Therefore, an analysis of the pNov hits was performed and those that overlap with the catalytic domain of any known IS element were identified (see Additional file 8). This analysis showed that many hits fall outside the catalytic domain, especially for fragment-reporting tools (ISsaga and ISEScan–fragments). If the tools were evaluated according to this stricter definition, then the proposed *digIS* would achieve the best results in the detection of pNovs on an absolute scale.

We analyzed the coverage of pNovs by individual tools to identify which of them are reported by several tools simultaneously or, conversely, exclusively by a specific tool. We also measured pNovs regarding their proximity to existing families of IS elements to reveal a possible preference of the tool to search for pNovs in a certain part of the sequence space (see Additional file 8). It turned out that various tools have a preference to search pNov elements close to various IS families. For example, *digIS* found the most pNovs close to the ISH3 family while ISsaga found the most pNovs close to the IS5 family. In summary, it can be concluded that no tool would include all pNov outputs of other tools.

Finally, we performed an analysis of the found pNovs to verify that they met the common characteristics of IS elements, such as multiple occurrences in the genome, or the presence of IR and DR regions. Using clustering, we found groups of similar hits, then performed their multiple sequence alignment, and identified IR and DR regions. Based on a manual inspection of selected clusters, we have identified four novel IS elements, of which the first two can be found by competing tools and the other two represent new ones found exclusively by the *digIS* tool (see Additional file 9).

Conclusions

In this paper we present a novel approach for IS elements detection, that is implemented in the form of *digIS* tool. It combines searching for the catalytic domains of transposases and additional filtering mechanisms that allows to detect not only known IS elements, but also distant putative novel IS elements. Simultaneously, it eliminates a large number of false hits that are typical for fragment–reporting tools.

Comparison with other state-of-the-art tools, such as ISsaga, OASIS, and ISEScan, on different datasets (*E.coli*, ISbrowser, NCBI GenBank Archaea/Bacteria) confirmed that *digIS* can find the majority of known ISs and shows the best ratio between putative novel elements and improbable/not IS elements. This makes it the right choice for scientists who are interested in finding new IS elements.

Finally, we would also highlight the technical aspects of the developed software. *digIS* is one of the few tools that still works and is ready for future use in the form of a Docker image. Simultaneously, it does not limit the user in the number of sequences to be analyzed or other search parameters, as is the case of web-based tools. *digIS* is ready to run in a grid-computing and cloud environment, which is very important for scalability. The transparency and credibility of the tool are further supported by the open-source code on GitHub (Table 4).

Abbreviations

bp:: Base pair; CDS:: Coding sequence; DR:: Direct repeat; DUF:: Domain of unknown function; eIS:: IS element with a high level of evidence; FN:: False negative; FP:: False positive; IR:: Inverted repeat; IS:: Insertion sequence; kbp:: Kilobase pair; MSA:: Multiple sequence alignment; mFP:: Merged false positive; nIS:: Improbable or not an IS element; nISDR:: Improbable/not an IS element discovery rate; ORF:: Open reading frame; pNov:: Putative novel; pNovDR:: Putative novel discovery rate; pHMM:: Profile hidden Markov model; TP:: True positive; teIS:: Total number of IS element with a high level of evidence.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-021-04177-6.

Additional file 1. Multiple sequence alignment of IS5/IS5 subfamily.

Additional file 2. Multiple sequence alignment of IS5/None subfamily.

Additional file 3. Multiple sequence alignment of ISL3 family.

Additional file 4. Analysis of merged FPs.

Additional file 5. Calculation of the similarity at IS and ORF level.

Additional file 6. Detailed information about NCBI GenBank archaeal and bacterial genomes used in the evaluation.

Additional file 7. Analysis of hits classified as *other annotation*.

Additional file 8. Analysis of putative novel elements.

Additional file 9. Putative novel IS elements detected by *digIS*.

Acknowledgements

Computational resources were supplied by the project "e-Infrastruktura CZ" (e-INFRA LM2018140) provided within the program Projects of Large Research, Development and Innovations Infrastructures. We would like to thank Alessandro Varani (ISsaga developer) for his help, Andrea Bisconti for English proofreading and correction, and two anonymous reviewers for their valuable feedback which significantly significantly improved the quality of this work.

Authors' contributions

JP and TM designed the research. The code was written and developed by JP and TM. Data was obtained by JP and TM. The evaluation was performed by JP and TM. All authors have read and approved the final manuscript.

Funding

This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic from the National Programme of Sustainability II, IT4Innovations excellence in science [grant number LQ1602]. All data was publicly sourced