Chapter 6

Transposable Elements: Classification, Identification, and Their Use As a Tool For Comparative Genomics

Wojciech Makałowski, Valer Gotea, Amit Pande, and Izabela Makałowska

Abstract

Most genomes are populated by hundreds of thousands of sequences originated from mobile elements. On the one hand, these sequences present a real challenge in the process of genome analysis and annotation. On the other hand, they are very interesting biological subjects involved in many cellular processes. Here we present an overview of transposable elements biodiversity, and we discuss different approaches to transposable elements detection and analyses.

Key words Transposable elements, Transposons, Mobile elements, Repetitive elements, Genome analysis, Genome evolution

1 Introduction

Most eukaryotic genomes contain large numbers of repetitive sequences. This phenomenon was described by Waring and Britten a half century ago using reassociation studies [1, 2]. It turned out that most of these repetitive sequences originated in transposable elements (TEs) [3], though the repetitive fraction of a genome varies significantly between different organisms, from 12% in *Caenorhabditis elegans* [4] to 50% in mammals [3], and more than 80% in some plants [5]. With such large contributions to genome sequences, it is not surprising that TEs have a significant influence on the genome organization and evolution. Although much progress has been achieved in understanding the role TEs play in a host genome, we are still far from the comprehensive picture of the delicate evolutionary interplay between a host genome and the invaders. They also pose various challenges to the genomic community, including aspects related to their detection and classification, genome assembly and annotation, genome comparisons, and mapping of genomic variants. They also pose various challenges to the genomic community, including aspects related to their detection and classification, genome assembly and annotation, genome

Maria Anisimova (ed.), *Evolutionary Genomics: Statistical and Computational Methods*, Methods in Molecular Biology, vol. 1910, https://doi.org/10.1007/978-1-4939-9074-0_6, © The Author(s) 2019

comparisons, and mapping of genomic variants. Here we present an overview of TE diversity and discuss major techniques used in their analyses.

2 Discovery of Mobile Elements

Transposable elements were discovered by Barbara McClintock during experiments conducted in 1944 on maize. Since they appeared to influence phenotypic traits, she named them *controlling elements*. However, her discovery was met with less than enthusiastic reception by the genetic community. Her presentation at the 1951 Cold Spring Harbor Symposium was not understood and at least not very well received [6]. She had no better luck with her follow-up publications [7–9] and after several years of frustration decided not to publish on the subject for the next two decades. Not for the first time in the history of science, an unappreciated discovery was brought back to life after some other discovery has been made. In this case it was the discovery of insertion sequences (IS) in bacteria by Szybalski group in the early 1970s [10]. In the original paper they wrote: "Genetic elements were found in higher organisms which appear to be readily transposed from one to another site in the genome. Such elements, identifiable by their controlling functions, were described by McClintock in maize. It is possible that they might be somehow analogous to the presently studied IS insertions" [10]. The importance of McClintock's original work was eventually appreciated by the genetic community with numerous awards, including 14 honorary doctoral degrees and a Nobel Prize in 1983 "for her discovery of mobile genetic elements" (http://nobelprize.org/nobel_prizes/medicine/laureates/1983/).

Coincidently, at the same time as Szybalski "rediscovered" TEs, Susumu Ohno popularized the term *junk DNA* that influenced genomic field for decades [11], although the term itself was used already before [12, 13].[1] Ohno referred to the so-called noncoding sequences or, to be more precise, to any piece of DNA that do not code for a protein, which included all genomic pieces originated in transposons. The unfavorable picture of transposable and transposed elements started to change in early 1990s when some researchers noticed evolutionary value of these elements [14, 15]. With the wheel of fortune turning full circle and advances of genome sciences, TE research is again focused on the role of mobile elements played in the evolution of gene regulation [16–23].

[1] The historical background of the "junk DNA" term was recently discussed by Dan Graur in his excellent blog http://judgestarling.tumblr.com/post/64504735261/the-origin-of-the-term-junk-dna-a-historical

3 Transposons Classification

3.1 Insertion Sequences and Other Bacterial Transposons

The bacterial genome is composed of a core genomic backbone decorated with a variety of multifarious functional elements. These include mobile genetic elements (MGEs) such as *bacteriophages*, *conjugative transposons*, *integrons*, *unit transposons*, *composite transposons*, and *insertion sequences* (IS). Here we elaborate upon the last class of these elements as they are most widely found and described [24].

The ISs were identified during studies of model genetic systems by virtue of their capacity to generate mutations as a result of their translocation [10]. In-depth studies in antibiotic resistance and transmissible plasmids revealed an important role for these mobile elements in formation of resistance genes and promoting gene capture. In particular, it was observed that several different elements were often clustered in "islands" within plasmid genomes and served to promote plasmid integration and excision.

Although these elements sometimes generate beneficial mutations, they may be considered genomic parasites as ISs code only for the enzyme required for their own transposition [24]. While an IS element occupies a chromosomal location, it is inherited along with its host's native genes, so its fitness is closely tied to that of its host. Consequently, ISs causing deleterious mutations that disrupt a genomic mode or function are quickly eliminated from the population. However, intergenically placed ISs have a higher chance to be fixed in the population as they are likely neutral regarding population's fitness [25].

ISs are generally compact (Fig. 1). They usually carry no other functions than those involved in their mobility. These elements contain recombinationally active sequences which define the boundary of the element, together with Tpase, an enzyme, which processes these ends and whose gene usually encompasses the entire length of the element [26]. Majority of ISs exhibit short terminal inverted-repeat sequences (IR) of length 10–40 bp. Several notable exceptions do exist, for example, the IS91, IS110, and IS200/605 families.

The IRs contain two functional domains [27]. One is involved in Tpase binding; the other cleaves and transfers strand-specific reactions resulting in transposition. IS promoters are often positioned partially within the IR sequence upstream of the *Tpase* gene. Binding sites for host-specific proteins are often located within proximity to the terminal IRs and play a role in modulating transposition activity or Tpase expression [28]. A general pattern for the functional organization of Tpases has emerged from the limited numbers analyzed. The N-terminal region contains sequence-specific DNA binding activities of the proteins while the catalytic domain is often localized toward the C-terminal end [28].

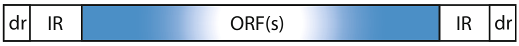


Fig. 1 Schematic representation of insertion sequences (IS). *dr* direct repeats, *IR* inverted repeats, *ORF* open reading frame

Another common feature of ISs is duplication of a target site that results in short direct repeats (DRs) flanking the IS [29]. The length of the direct repeat varies from 2 to 14 base pairs and is a hallmark of a given element. Homologous recombination between two IS elements can result in each having two different DRs [30].

ISs have been classified on the basis of (1) similarities in genetic organization (arrangement of open reading frames); (2) marked identities or similarities in their Tpases (common domains or motifs); (3) similar features of their ends (terminal IRs); and (4) fate of the nucleotide sequence of their target sites (generation of a direct target duplication of determined length). Based on the above rules, ISs are currently classified in 30 families (Table 1) [31].

3.2 Eukaryotic Transposable Elements

The first TE classification system was proposed by Finnegan in 1989 [32] and distinguished two classes of TEs characterized by their transposition intermediate: RNA (class I or retrotransposons) or DNA (class II or DNA transposons). The transposition mechanism of class I is commonly called "copy and paste" and that of class II, "cut and paste." In 2007 Wicker et al. [33] proposed hierarchical classification based on TEs structural characteristics and mode of replication (*see* Table 2 and Fig. 2). Below we present a brief overview of eukaryotic mobile elements that in general follows this classification.

3.2.1 Class I: Mobile Elements

As mentioned above, class I TEs transpose through an RNA intermediary. The RNA intermediate is transcribed from genomic DNA and then reverse-transcribed into DNA by a TE-encoded reverse transcriptase (RT), followed by reintegration into a genome. Each replication cycle produces one new copy, and as a result, class I elements are the major contributors to the repetitive fraction in large genomes. Retrotransposons are divided into five orders: LTR retrotransposons, DIRS-like elements, Penelope-like elements (PLEs), LINEs (*l*ong *in*terspersed *e*lements), and SINEs (*s*hort *in*terspersed *e*lements). This scheme is based on the mechanistic features, organization, and reverse transcriptase phylogeny of these retroelements. Accidentally, the retrotranscriptase coded by an autonomous TE can reverse-transcribe another RNA present in the cell, e.g., mRNA, and produce a retrocopy of it, which in most cases results in a pseudogene.

The LTR retrotransposons are characterized by the presence of *l*ong *t*erminal *r*epeats (LTRs) ranging from several hundred to several thousand base pairs. Both exogenous retroviruses and LTR retrotransposons contain a *gag* gene that encodes a viral

Table 1
Prokaryotic transposable elements as presented in the *IS Finder* database [31]

Family	Typical size range in bp	Direct repeat size in bp	IRs[a]	Number of ORFs
IS*1*	740–4600	0–10	Y	1 or 2
IS*110*	1200–1550	0	Y	1
IS*1182*	1330–1950	0–60	Y	1
IS*1380*	1550–2000	4–5	Y	1
IS1595	700–7900	8	Y	1
IS*1634*	1500–2000	5–6	Y	1
IS*200*/IS*605*	600–2000	0	Y/N	1 or 2
IS*21*	1750–2600	4–8	Y	2
IS*256*	1200–1500	8–9	Y	1
IS*3*	1150–1750	5	Y	2
IS*30*	1000–1700	2–3	Y	1
IS*4*	1150–5400	8–13	Y	1 or more
IS*481*	950–1300	4–15	Y	1
IS*5*	800–1500	2–9	Y	1 or 2
IS*6*	700–900	8	Y	1
IS*607*	1700–2500	0	N	2
IS*630*	1000–1400	2	Y	1 or 2
IS*66*	1350–3000	8–9	Y	1 or more
IS*701*	1400–1550	4	Y	1
IS*91*	1500–2000	0	N	1
IS*982*	1000	3–9	Y	1
IS*As1*	1200–1500	8–10	Y	1
IS*Azo13*	1250–2200	0–4	Y	1
IS*H3*	1225–1500	4–5	Y	1
IS*H6*	1450	8	Y	ISL
IS*Kra4*	1400–2900	0–9	Y	1 or more
IS*L3*	1300–2300	8	Y	ISKra4
IS*Lre2*	1500–2000	9	Y	1
Tn*3*	Over 3000	0	Y	More than 1
IS*NCY*	1300–2400	0–12	Y/N	1 or 2

[a] Presence (Y) or absence (N) of terminal inverted repeats

Table 2
Classification of eukaryotic transposable elements as proposed by Wicker et al. [33]

Class	Order	Superfamily	Phylogenetic distribution
Class I (retrotransposons)	LTR	Copia	Plants, metazoans, fungi
		Gypsy	Plants, metazoans, fungi
		Bel-Pao	Metazoans
		Retrovirus	Metazoans
		ERV	Metazoans
	DIRS	DIRS	Plants, metazoans, fungi
		Ngaro	Metazoans, fungi
		VIPER	Trypanosomes
	PLE	Penelope	Plants, metazoans, fungi
	LINE	R2	Metazoans
		RTE	Metazoans
		Jockey	Metazoans
		L1	Plants, metazoans, fungi
	SINE	tRNA	Plants, metazoans, fungi
		7SL	Plants, metazoans, fungi
		5S	Metazoans
		SVA[a]	Primates
		Retrogenes[a]	Plants, metazoans, fungi
Class II (DNA transposons) Subclass 1	TIR	Tc1-Mariner	Plants, metazoans, fungi
		hAT	Plants, metazoans, fungi
		Mutator	Plants, metazoans, fungi
		Merlin	Metazoans
		Transib	Metazoans, fungi
		P	Plants, metazoans
		PiggyBac	Metazoans
		PIF-harbinger	Plants, metazoans, fungi
		CACTA	Plants, metazoans, fungi
	Crypton	Crypton	Fungi
Class II (DNA transposons) Subclass 2	Helitron	Helitron	Plants, metazoans, fungi
	Maverick	Maverick	Metazoans, fungi

Please note that SVAs and retrogenes are not included in that classification
[a]Not included in the original Wicker classification

particle coat and a *pol* gene that encodes a reverse transcriptase, ribonuclease H, and an integrase, which provide the enzymatic machinery for reverse transcription and integration into the host genome. Reverse transcription occurs within the viral or viral-like particle (GAG) in the cytoplasm, and it is a multistep process [34]. Unlike LTR retrotransposons, exogenous retroviruses contain an *env* gene, which encodes an envelope that facilitates their migration to other cells. Some LTR retrotransposons may contain remnants of an *env* gene, but their insertion capabilities are limited to the originating genome [35]. This would rather suggest that they originated in exogenous retroviruses by losing the *env* gene. However, there is evidence that suggests the contrary, given that

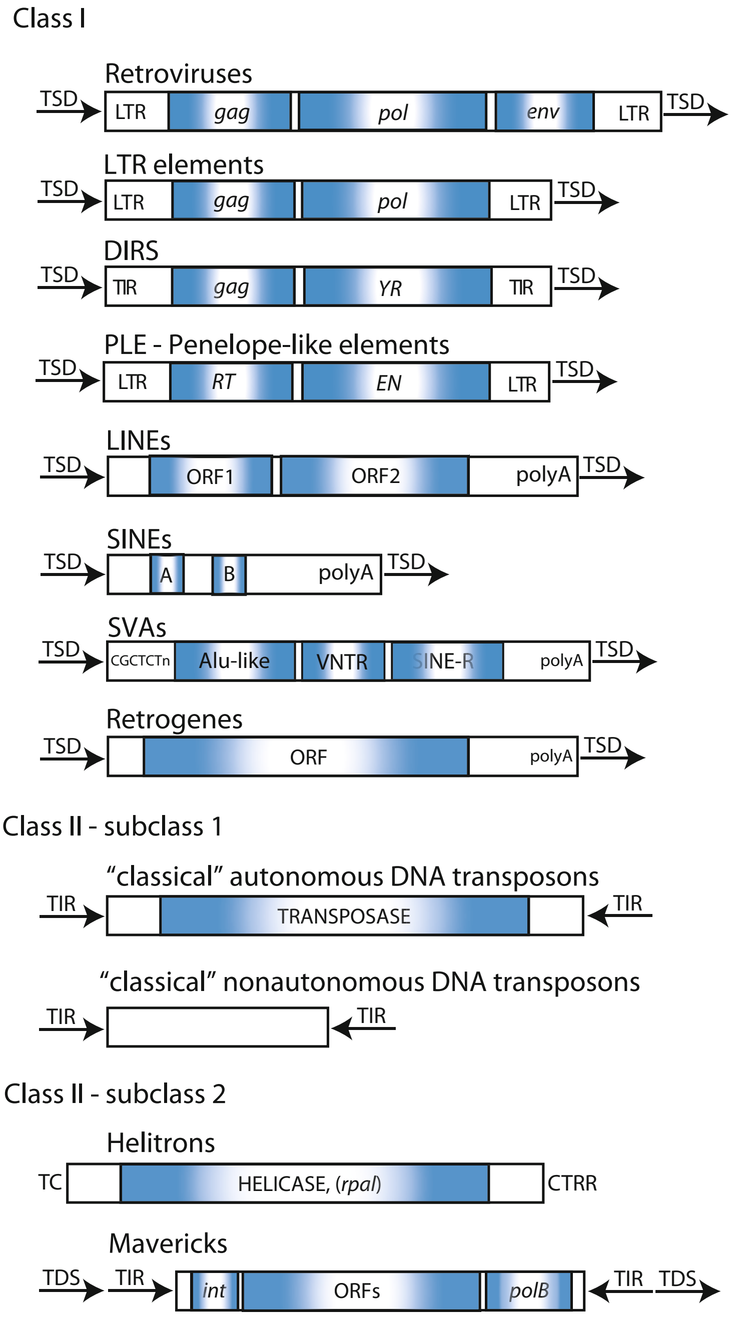


Fig. 2 Structures of eukaryotic mobile elements. See text for detailed discussion

LTR retrotransposons can acquire the *env* gene and become infectious entities [36]. Presently, most of the LTR sequences (85%) in the human genome are found only as isolated LTRs, with the internal sequence being lost most likely due to homologous recombination between flanking LTRs [37]. Interestingly, LTR retrotransposons target their reinsertion to specific genomic sites, often around genes, with putative important functional implications for a host gene [35]. Lander et al. estimated that 450,000 LTR copies make up about 8% of our genome [38]. LTR retrotransposons inhabiting large genomes, such as maize, wheat, or barley, can contain thousands of families. However, despite the diversity, very few families comprise most of the repetitive fraction in these large genomes. Notable examples are Angela (wheat) [39], BARE1 (barley) [40], Opie (maize) [41], and Retrosor6 (sorghum) [42].

The DIRS order clusters structurally diverged group of transposons that possess a tyrosine recombinase (YR) gene instead of an integrase (INT) and do not form target site duplications (TSDs). Their termini resemble either split direct repeats (SDR) or inverted repeats. Such features indicate a different integration mechanism than that of other class I mobile elements. DIRS were discovered in the slime mold (*Dictyostelium discoideum*) genome in the early 1980s [43], and they are present in all major phylogenetic lineages including vertebrates [44]. It has been showed that they are also common in hydrothermal vent organisms [45].

Another order, termed *Penelope*-like elements (PLE), has wide, though patchy distribution from amoebae and fungi to vertebrates with copy number up to thousands per genome [46]. Interestingly, no PLE sequences have been found in mammalian genomes, and apparently they were lost from the genome of *C. elegans* [47]. Although PLEs with an intact ORF have been found in several genomes, including *Ciona* and *Danio*, the only transcriptionally active representative, *Penelope*, is known from *Drosophila virilis*. It causes the hybrid dysgenesis syndrome characterized by simultaneous mobilization of several unrelated TE families in the progeny of dysgenic crosses. It seems that *Penelope* invaded *D. virilis* quite recently, and its invasive potential was demonstrated in *D. melanogaster* [46]. PLEs harbor a single ORF that codes for a protein containing reverse transcriptase (RT) and endonuclease (EN) domains. The PLE RT domain more closely resembles telomerase than the RT from LTRs or LINEs. The EN domain is related to GIY-YIG intron-encoded endonucleases. Some PLE members also have LTR-like sequences, which can be in a direct or an inverse orientation, and have a functional intron [46].

LINEs [48, 49] do not have LTRs; however, they have a poly-A tail at the 3′ end and are flanked by the TSDs. They comprise about 21% of the human genome and among them L1 with about 850,000 copies is the most abundant and best described LINE family. L1 is the only LINE retroposon still active in the human genome [50]. In the human genome, there are two other LINE-like repeats, L2 and L3, distantly related to L1. A contrasting situation has been noticed in the malaria mosquito *Anopheles gambiae*, where around 100 divergent LINE families compose only 3% of its genome [51]. LINEs in plants, e.g., Cin4 in maize and Ta11 in *Arabidopsis thaliana*, seem rare as compared with LTR retrotransposons. A full copy of mammalian L1 is about 6 kb long and contains a PolII promoter and two ORFs. The ORF1 codes for a non-sequence-specific RNA binding protein that contains zinc finger, leucine zipper, and coiled-coil motifs. The ORF1p functions as chaperone for the L1 mRNA [52, 53]. The second ORF encodes an endonuclease, which makes a single-stranded nick in the genomic DNA, and a reverse transcriptase, which uses the nicked DNA to prime reverse transcription of LINE RNA from the 3′ end. Reverse transcription is often unfinished, leaving behind fragmented copies of LINE elements; hence most of the L1-derived repeats are short, with an average size of 900 bp. LINEs are part of the CR1 clade, which has members in various metazoan species, including fruit fly, mosquito, zebrafish, pufferfish, turtle, and chicken [54]. Because they encode their own retrotransposition machinery, LINE elements are regarded as autonomous retrotransposons.

SINEs [48, 49] evolved from RNA genes, such as 7SL and tRNA genes. By definition, they are short, up to 1000 base pair long. They do not encode their own retrotranscription machinery and are considered as nonautonomous elements and in most cases are mobilized by the L1 machinery [55]. The outstanding member of this class from the human genome is the Alu repeat, which contains a cleavage site for the *Alu*I restriction enzyme that gave its name [56]. With over a million copies in the human genome, Alu is probably the most successful transposon in the history of life. Primate-specific Alu and its rodent relative B1 have limited phylogenetic distribution suggesting their relatively recent origins. The mammalian-wide interspersed repeats (MIRs), by contrast, spread before eutherian radiation, and their copies can be found in different mammalian groups including marsupials and monotremes [57]. SVA elements are unique primate elements due to their composite structure. They are named after their main components: SINE, VNTR (a variable number of tandem repeats), and Alu [58]. Usually, they contain the hallmarks of the retroposition, i.e., they are flanked by TSDs and terminated by a poly(A) tail. It seems that SVA elements are nonautonomous retrotransposons mobilized by L1 machinery, and they are thought to be transcribed by RNA polymerase II. SVAs are transpositionally active and are responsible for some human diseases [59]. They originated less than 25 million years ago, and they form the youngest retrotransposon family with about 3000 copies in the human genome [58].

Retro(pseudo)genes are a special group of retroposed sequences, which are products of reverse transcription of a spliced (mature) mRNA. Hence, their characteristic features are an absence of promoter sequence and introns, the presence of flanking direct repeats, and a 3′-end polyadenosine tract [60]. Processed pseudogenes, as sometimes retropseudogenes are called, have been generated in vitro at a low frequency in the human HeLa cells via mRNA from a reporter gene [60]. The source of the reverse transcription machinery in humans and other vertebrates seems to be active L1 elements [61]. However, not all retroposed messages have to end up as pseudogenes. About 20% of mammalian protein-encoding genes lack introns in their ORFs [62]. It is conceivable that many genes lacking introns arose by retroposition. Some genes are known to be retroposed more often than others. For instance, in the human genome there are over 2000 retropseudogenes of ribosomal proteins [63]. A genome-wide study showed that the human genome harbors about 20,000 pseudogenes, 72% of which most likely arose through retroposition [64]. Interestingly, the vast majority (92%) of them are quite recent transpositions that occurred after primate/rodent divergence [64]. Some of the retroposed genes may undergo quite complicated evolutionary paths. An example could be the RNF13B retrogene, which replaced its own parental gene in the mammalian genomes. This retrocopy was duplicated in primates, and the evolution of this primate-specific copy was accompanied by the exaptation of two TEs, Alu and L1, and intron gain via changing a part of coding sequence into an intron leading to the origin of a functional, primate-specific retrogene with two splicing variants [65].

3.2.2 Class II: Mobile Elements

Class II elements move by a conservative cut-and-paste mechanism; the excision of the donor element is followed by its reinsertion elsewhere in the genome. DNA transposons are abundant in bacteria, where they are called insertion sequences (*see* Subheading 3.1), but are present in all phyla. Wicker et al. distinguished two subclasses of DNA transposons based on the number of DNA strands that are cut during transposition [33].

Classical "cut-and-paste" transposons belong to the subclass I, and they are classified as the TIR order. They are characterized by terminal inverted repeats (TIR) and encode a transposase that binds near the inverted repeats and mediates mobility. This process is not usually a replicative one, unless the gap caused by excision is repaired using the sister chromatid. When inserted at a new location, the transposon is flanked by small gaps, which, when filled by host enzymes, cause duplication of the sequence at the target site. The length of these TSDs is characteristic for particular transposons. Nine superfamilies belong to the TIR order, including *Tc1-Mariner*, *Merlin*, *Mutator*, and *PiggyBac*. The second order Crypton consists of a single superfamily of the same name. Originally thought to be limited to fungi [66], now it is clear that they have a wide distribution, including animals and heterokonts [67]. A heterogeneous, small, nonautonomous group of elements MITEs also belong to the TIR order [68], which in some genomes amplified to thousands of copies, e.g., *Stowaway* in the rice genome [69], *Tourist* in most bamboo genomes [70], or *Galluhop* in the chicken genome [71].

Subclass II includes two orders of TEs that, just as those from subclass I, do not form RNA intermediates. However, unlike "classical" DNA transposons, they replicate without double-strand cleavage. Helitrons replicate using a rolling-circle mechanism, and their insertion does not result in the target site duplication [72]. They encode tyrosine recombinase along with some other proteins. Helitrons were first described in plants, but they are also present in other phyla, including fungi and mammals [73, 74]. Mavericks are large transposons that have been found in different eukaryotic lineages excluding plants [75]. They encode various numbers of proteins that include DNA polymerase B and an integrase. Kapitonov and Jurka suggested that their life cycle includes a single-strand excision, followed by extrachromosomal replication and reintegration to a new location [76].

4 Identification of Transposable Elements

With the ever-growing number of sequenced genomes from different branches of the tree of life, there are increasing TE research opportunities. There are several reasons why one would like to analyze TEs and their "offsprings" left in a genome. First of all, they are very interesting biological subjects to study genome structure, gene regulation, or genome evolution. In some cases, they also make genome assembly and annotation quite challenging, especially with the current NGS technology that generates reads shorter than TEs. Nevertheless, TEs should be and are worthy to study. However, it is not a simple task and requires different approaches depending on the level of analysis. We will walk through these different levels starting with raw genome sequences without any annotation and discuss different methods and software used for TE analyses. In principle, we can imagine two scenarios: in the first one, genomic or transcriptome sequences are coming from a species for which there is already some information about the transposon repertoire, for instance, a related genome has been previously characterized or TEs have been studied before. In the second scenario, we have to deal with a completely unknown genome or a genome for which little information exists with regard to TEs. In the former case, one can apply a range of techniques used in comparative genomics or try to search specific libraries of transposons using the "homology search" approach. In the latter, which is basically an approach to identify TEs de novo, first we need to find any repeats in a genome and then attempt characterization and classification of newly identified repetitive sequences. In this approach, we will find *any* repeats, not necessarily transposons. There are many algorithms, and even more software, that can be applied in both approaches.

4.1 De Novo Approaches to Finding Repetitive Elements

There are several steps involved in the de novo characterization of transposons. First, we need to find all the repeats in a genome, then build a consensus of each family of related sequences, and finally classify detected sequences. For the first step, three groups of algorithms exist: the *k*-mer approach, sequence self-comparison, and periodicity analysis.

In the *k*-mer approach, sequences are scanned for overrepresentation of strings of certain length. The idea is that repeats that belong to the same family are compositionally similar and share some oligomers. If the repeats occur many times in a genome, then those oligomers should be overrepresented. However, since repeats and transposons in particular are not perfect copies of a certain sequence, some mismatches must be allowed when oligo frequencies are calculated. The challenge is to determine optimal size of an oligo (*k*-mer) and number of mismatches allowed. Most likely, these parameters should be different for different types of transposons, i.e., low versus high copy number, old versus young transposons, and those from different classes and families. Several programs have been developed based on the *k*-mer idea using a suffix tree data structure including REPuter [77, 78], Vmatch (Kurtz, unpublished; http://www.vmatch.de/), and Repeat-match [79, 80]. Another approach is to use fixed length *k*-mers as seeds and extend those seeds to define repeat's family as it was implemented in ReAS [81], RepeatScout [82], and Tallymer [83]. Another interesting algorithm can be found in the FORRepeats software [84], which uses *factor oracle* data structure [85]. It starts with detection of exact oligomers in the analyzed sequences, followed by finding approximate repeats and their alignment.

The second group of programs developed for de novo detection of repeated sequences is using self-comparison approach. Repeat Pattern Toolkit [86], RECON [87], PILER [88, 89], and BLASTER [90] belong to this group. The idea is to use one of the fast sequence similarity tools, e.g., BLAST [91], followed by clustering search results. The programs differ in the search engine for the initial step, though most are using some of the BLAST algorithms, the clustering method, and heuristics of merging initial hits into a prototype element. For instance, RECON [87], which was developed for the repeat finding in unassembled sequence reads, starts with an all-to-all comparison using WU-BLAST engine. Then, single-linkage clustering is applied to alignment results that is followed by construction of an undirected graph with overlapping. The shortest sequence that contains connected images (aligned subsequences) creates a prototype element. However, this procedure might result in composite elements. To avoid this, all the images are aligned to the prototype element to detect potential illegitimate mergers and split those at every point with a significant number of image ends.

PILER [88, 89] is using a different approach to find initial clusters. Instead of BLAST, it uses PALS (pairwise alignment of long sequences) for the initial alignment. PALS records only hit points and uses banded search of the defined maximum distance to optimize its performance. To further improve performance of the system, PILER uses different heuristics for different types of repeats, i.e., satellites, pseudosatellites, terminal repeats, and interspersed repeats. Finally, a consensus sequence is generated from a multiple sequence alignment of the defined family members.

Dot matrix is a simple method to compare two biological sequences. The graphical output of such an analysis is called a dotplot. Dotplots can be used to detect conserved domains, sequence rearrangements, RNA secondary structure, or repeated sequences. It compares every residue in one sequence to every residue in the other sequence or to every residue of the same sequence in the self-comparison mode. In the latter case, there will be a main diagonal line representing a perfect match and a number of short diagonal lines representing similar regions (red circles in Fig. 3). Interestingly, simple repeats appear as diamond shapes on a main diagonal line or short vertical and horizontal lines outside the main diagonal line (red squares in Fig. 3). The method was introduced to biological analyses almost a half century ago [92, 93]. However, the first easy-to-use software with a graphical interface, DOTTER, was developed much later [94]. The major problem of this approach is the time required for the dotplot calculation, which is of quadratic complexity. This proved to be prohibitive for comparison of the genome-size sequences. One of the solutions to this problem is using a word index for the fast identification of substrings. Gepard implements the suffix array data structure to improve the execution time [95]. It is written in Java, which makes it platform-independent. Gepard enables analyses of sequences at the mega-base level in the matter of seconds, and it takes about an hour to analyze the whole human chromosome I [95]. The example of the dotplot produced by the Gepard is presented in Fig. 3.

4.2 Transposable Elements Determination in NGS Data

With constant improvement of sequencing technology associated with decreasing sequencing cost, the number of new sequenced genomes is exploding. As of January 2019, there are more than 7000 eukaryotic and almost 180,000 prokaryotic genomes publicly available (information retrieved on January 16, 2019, from https://www.ncbi.nlm.nih.gov/genome/browse/). However, this comes with a price; most of the recently sequenced genomes, due to the short read sequencing technology, are available at various levels of

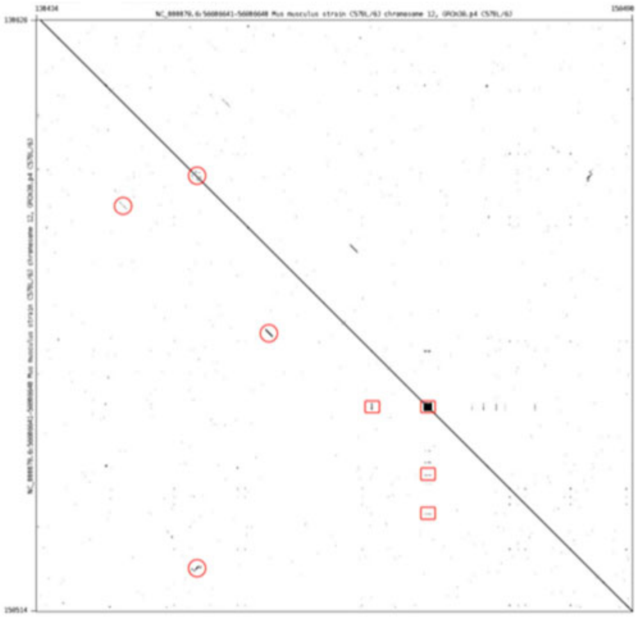


Fig. 3 Graphical output of the Gepard. A 30 kb fragment of mouse chromosome 12 was compared to itself. Similar sequences are represented by diagonal lines if both fragments are located on the same strains or by reverse diagonal lines if the fragments with significant similarity are located on opposite strands. Some of the examples are marked with the red circles. Simple repeats are represented by either diamond shapes on the main diagonal or horizontal and vertical lines. Some of the examples are marked with the red squares

"completeness" or assembly. For most non-model organisms, we are presented with draft assemblies of rather short contigs. Moreover, these genomes usually are not very well annotated, with TEs not being on the annotation priority list. Unfortunately, genome annotation pipelines do not include TE annotation, focusing on protein-coding and RNA-coding genes. To fill the gap, a number of methods have been developed to detect repeats from short reads. Two algorithms dominate in attempts to determine repeats in NGS raw reads: clustering and *k*-mer. Transposome [96] and RepeatExplorer [97] employ the former approach, while RepARK [98], REPdenovo [99], and dnaPipeTE [100] utilize the latter one. Since NGS results in the relatively short reads, assembly of selected sequences into longer contigs representing TEs is required after initial clustering of the raw reads.

4.3 Population-Level Analyses of Transposable Elements

Recent advances in sequencing technology and the sharp decrease in sequencing costs allow genomic studies at population level. Although initially focused on human populations [101–103], recent population studies of other species have been initiated as well [104, 105]. One of the common questions in such studies is

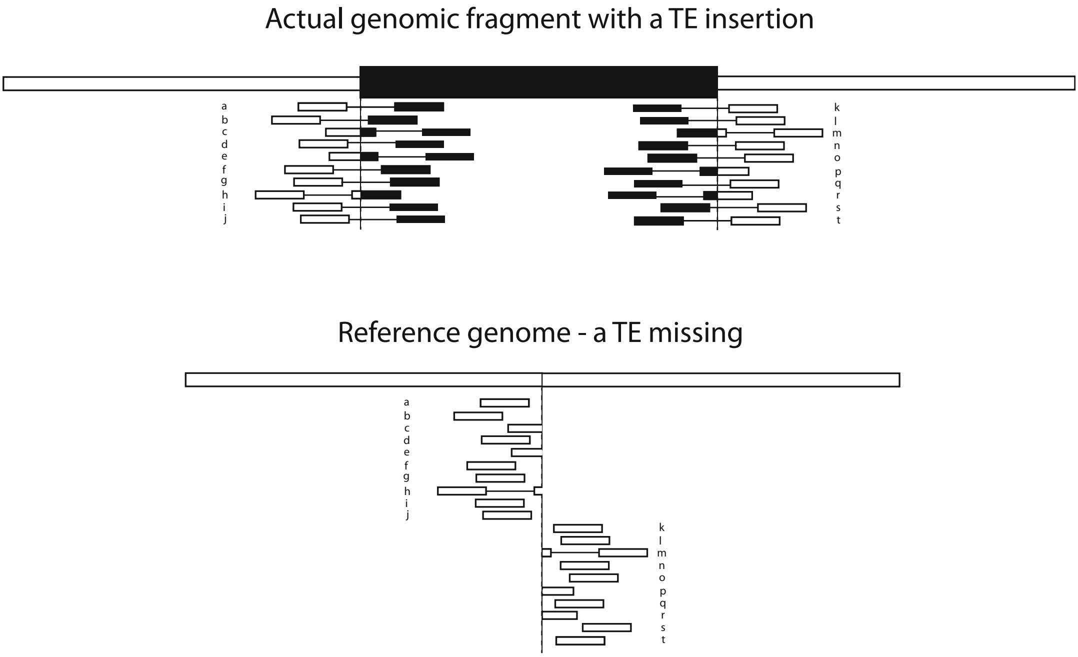

Actual genomic fragment with a TE insertion
a
b
c
d
e
f
g
h
i
j
k
l
m
n
o
p
q
r
s
t
Reference genome - a TE missing
a
b
c
d
e
f
g
h
i
j
k
l
m
n
o
p
q
r
s
t

Fig. 4 Detection of a TE insertion (polymorphic TE) from the NGS data. The upper panel shows real genomic sequence with a TE, which is not present in the reference genome (lower panel). Hypothetical discordant pair-reads (a, b, d, f, g, i, j, k, l, o, q, s, and t) have only one the pairs mapped to the reference genome, while the other would map to a consensus sequence of a TE. The hypothetical split reads (c, e, h, m, p, and r) will have part of the sequence mapped to the reference genome and the other to a TE consensus sequence

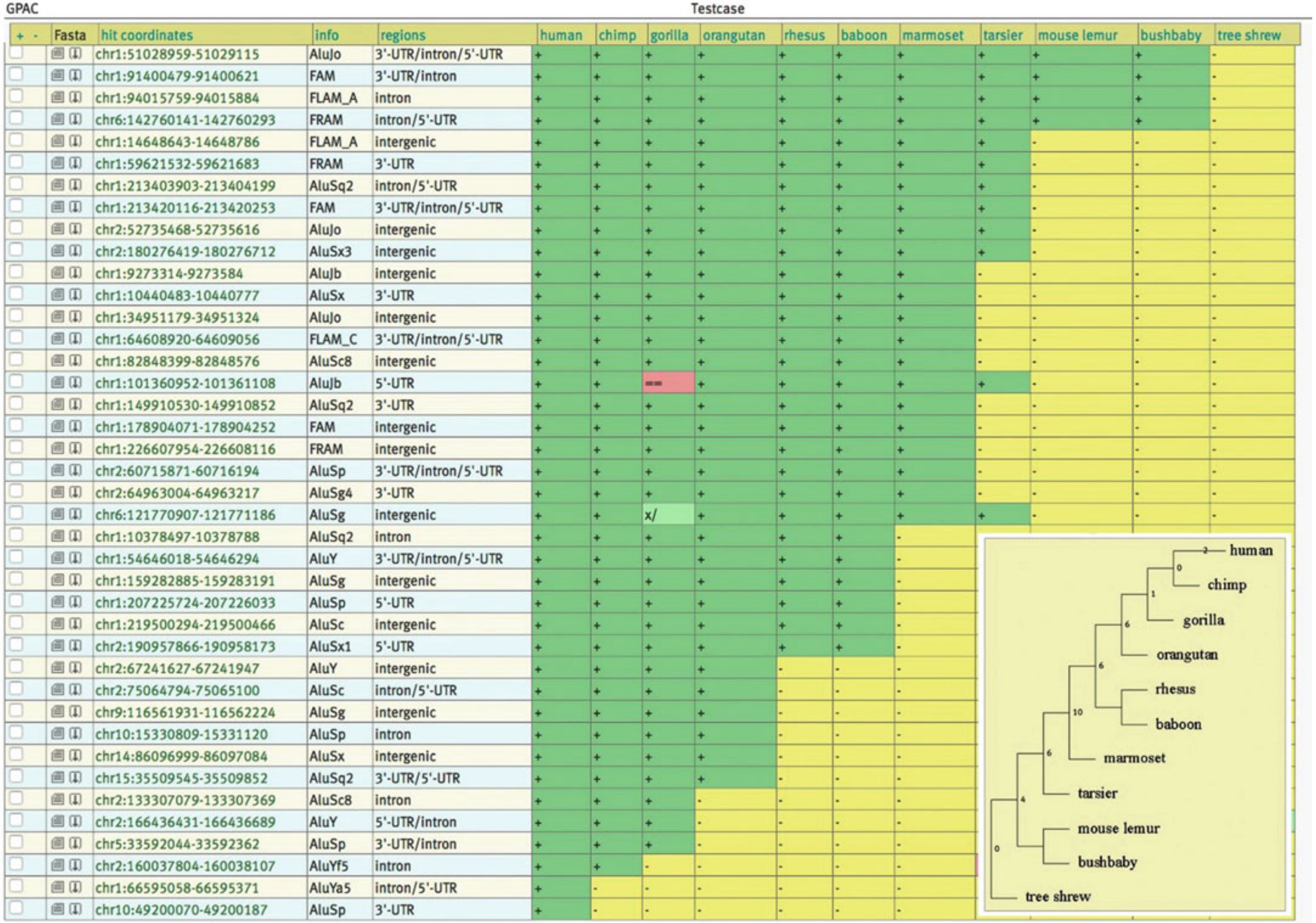
how much structural variation (SV) exists in different populations. TE insertions are responsible for about 25% of structural variants in human genomes [106]. In general, any tool designed for detection of SV should work for TE insertion analysis, but specialized software can take advantage of specific expectations related to insertions of TEs. Most of the SV-detection algorithms rely on paired-end reads and are based on discordant read pair mapping and/or split reads mapping (Fig. 4). A discordant pair read is defined as one that is inconsistent with the expected insert size in the library used for sequencing. For example, if the insert size of the library used for sequencing is 300 nt but the reads map to a reference genome within much larger distance or to two different chromosomes, such a pair is considered to be discordant. If, additionally, one of the reads maps to a TE, it might be an indication of a polymorphic TE. Usually some filtering is used to reduce a chance of false positives. These include minimum read number in the cluster mapped to a unique position, quality score of the reads, or consistency in reads orientation. However, the discordant read mapping cannot detect exact insertion position. Therefore another step is required that may include local assembly and split-read mapping.

A split read is defined as a read for which part of it maps uniquely to one position in the genome and the other part to another position. This is, for example, a very common feature of the mapping of RNA-seq data to eukaryotic genomes when reads span two exons. Split reads are being also observed if structural variants exist. In a case of a TE insertion, a part of the read will be mapped to a unique location and the rest to a TE in some other location or may not be mapped at all (Fig. 4).

Different methods for structure variant detection return different results on the same data. Recently published benchmarking demonstrates that TE detection is not an exception [107, 108]. Ewing [107] compared TranspoSeq [109] with two other tools, Tea [110] and TraFIC [111], on the same data sets. Results were not very encouraging as in both comparisons there was a high fraction of insertions detected only by a single program [107]. Similar conclusion was drawn by Rishishwar et al. [108] in a benchmark of larger number of tools including MELT [106], Mobster [112], and RetroSeq [113]. It is clear that different software have different biases, and each one can produce a high number of false positives. It is recommended then to employ several programs for high confidence results. Exhaustive tests run on real and simulated human genome data showed superior performance of MELT [106, 108]. TIPseqHunter is another tool developed to identify transposon insertion sites based on the transpose insertion profiling using next-generation sequencing [114]. It employs machine learning algorithm to ensure high precision and reliability. It is worth to note that all these tools were designed for short read sequencing methods. However, with current development of single-molecule long reads, sequencing technologies such as PacBio and Oxford Nanopore may make these methods irrelevant and obsolete. Long reads should be of superior performance and make TE insertion detection relatively easy with more traditional aligners, such as MegaBLAST [115], BLAT [116], or LAST [117].

4.4 Comparative Genomics of TE Insertions

To understand the general pattern of TE insertions in different genomes and evolutionary dynamics of TE families, a comparative approach is necessary. Although precomputed alignments of different genomes are publicly available, for example, the UCSC Genome Browser includes Multiz alignments of 100 vertebrate genomes [118], not many tools are available for such analyses. One of them is GPAC (genome presence/absence compiler) that creates a table of presence and absence of certain elements based on the precomputed multiple genomes alignment [119] (http://bioinformatics.uni-muenster.de/tools/gpac/index.hbi). The tool is quite generic, but is well suited for the TE comparative analysis (*see* Fig. 5 for an example).



GPAC

Testcase

+ -	Fasta	hit coordinates	info	regions	human	chimp	gorilla	orangutan	rhesus	baboon	marmoset	tarsier	mouse lemur	bushbaby	tree shrew
☐		chr1:51028959-51029115	AluJo	3'-UTR/intron/5'-UTR	+	+	+	+	+	+	+	+	+	+	-
☐		chr1:91400479-91400621	FAM	3'-UTR/intron	+	+	+	+	+	+	+	+	+	+	-
☐		chr1:94015759-94015884	FLAM_A	intron	+	+	+	+	+	+	+	+	+	+	-
☐		chr6:142760141-142760293	FRAM	intron/5'-UTR	+	+	+	+	+	+	+	+	+	+	-
☐		chr1:14648643-14648786	FLAM_A	intergenic	+	+	+	+	+	+	+	+	-	-	-
☐		chr1:59621532-59621683	FRAM	3'-UTR	+	+	+	+	+	+	+	+	-	-	-
☐		chr1:213403903-213404199	AluSq2	intron/5'-UTR	+	+	+	+	+	+	+	+	-	-	-
☐		chr1:213420116-213420253	FAM	3'-UTR/intron/5'-UTR	+	+	+	+	+	+	+	+	-	-	-
☐		chr2:52735468-52735616	AluJo	intergenic	+	+	+	+	+	+	+	+	-	-	-
☐		chr2:180276419-180276712	AluSx3	intergenic	+	+	+	+	+	+	+	+	-	-	-
☐		chr1:9273314-9273584	AluJb	intergenic	+	+	+	+	+	+	+	-	-	-	-
☐		chr1:10440483-10440777	AluSx	3'-UTR	+	+	+	+	+	+	+	-	-	-	-
☐		chr1:34951179-34951324	AluJo	intergenic	+	+	+	+	+	+	+	-	-	-	-
☐		chr1:64608920-64609056	FLAM_C	3'-UTR/intron/5'-UTR	+	+	+	+	+	+	+	-	-	-	-
☐		chr1:82848399-82848576	AluSc8	intergenic	+	+	+	+	+	+	+	-	-	-	-
☐		chr1:101360952-101361108	AluJb	5'-UTR	+	+	==	+	+	+	+	+	-	-	-
☐		chr1:149910530-149910852	AluSq2	3'-UTR	+	+	+	+	+	+	+	-	-	-	-
☐		chr1:178904071-178904252	FAM	intergenic	+	+	+	+	+	+	+	-	-	-	-
☐		chr1:226607954-226608116	FRAM	intergenic	+	+	+	+	+	+	+	-	-	-	-
☐		chr2:60715871-60716194	AluSp	3'-UTR/intron/5'-UTR	+	+	+	+	+	+	+	-	-	-	-
☐		chr2:64963004-64963217	AluSg4	3'-UTR	+	+	+	+	+	+	+	-	-	-	-
☐		chr6:121770907-121771186	AluSg	intergenic	+	+	x/	+	+	+	+	+	-	-	-
☐		chr1:10378497-10378788	AluSq2	intron	+	+	+	+	+	+	-				
☐		chr1:54646018-54646294	AluY	3'-UTR/intron/5'-UTR	+	+	+	+	+	+	-				
☐		chr1:159282885-159283191	AluSg	intergenic	+	+	+	+	+	+	-				
☐		chr1:207225724-207226033	AluSp	5'-UTR	+	+	+	+	+	+	-				
☐		chr1:219500294-219500466	AluSc	intergenic	+	+	+	+	+	+	-				
☐		chr2:190957866-190958173	AluSx1	5'-UTR	+	+	+	+	+	+	-				
☐		chr2:67241627-67241947	AluY	intergenic	+	+	+	+	-	-	-				
☐		chr2:75064794-75065100	AluSc	intron/5'-UTR	+	+	+	+	-	-	-				
☐		chr9:116561931-116562224	AluSg	intergenic	+	+	+	+	-	-	-				
☐		chr10:15330809-15331120	AluSp	intron	+	+	+	+	-	-	-				
☐		chr14:86096999-86097084	AluSx	intergenic	+	+	+	+	-	-	-				
☐		chr15:35509545-35509852	AluSq2	3'-UTR/5'-UTR	+	+	+	+	-	-	-				
☐		chr2:133307079-133307369	AluSc8	intron	+	+	+	-	-	-	-				
☐		chr2:166436431-166436689	AluY	5'-UTR/intron	+	+	+	-	-	-	-				
☐		chr5:33592044-33592362	AluSp	3'-UTR/intron	+	+	+	-	-	-	-				
☐		chr2:160037804-160038107	AluYf5	intron	+	+	-	-	-	-	-				
☐		chr1:66595058-66595371	AluYa5	intron/5'-UTR	+	-	-	-	-	-	-				
☐		chr10:49200070-49200187	AluSp	3'-UTR	+	-	-	-	-	-	-				

Fig. 5 The output table of the GPAC software. Several Alu elements were analyzed for presence/absence in 11 primate species. The human genome was used a reference, and "hit coordinates" refer to that genome along with the information on the annotated elements in the hit region and a type of the region. For each genome, the presence (+) or absence (−) of the hit is presented. x/ denotes that only part of the original insertion (less than 20%) is present in a given genome, and == indicates that more than 80% of the expected sequence is not alignable in a given locus. The optional phylogenetic tree constructed based on the obtained data is shown in the lower right corner

4.5 Classification of Transposable Elements

Once the consensus of a repetitive element has been constructed, it can be subjected to further analyses. There are two major categories of programs dealing with the issue of TE classification: library or similarity-based and signature-based. The latter approach is very often used in specialized software, i.e., tailored for specific type of TEs. However, some general tools also exist, e.g., TEclass [120].

The library approach is probably the most common approach for TE classification. It is also very efficient and quite reliable as long as good libraries of prototype sequences exist. In practice, it is the recommended approach when we analyze sequences from well-characterized genomes or from a genome relatively closely related to a well-studied one. For instance, since the human genome is one of the best studied, any primate sequences can be confidently analyzed using the library approach. Most likely, the first software using the similarity-based approach for repeat classification was *Censor* developed by Jerzy Jurka in the early 1990s [121]. It uses RepBase [122] as a reference collection and BLAST as a search engine [91]. However, the most popular TE detection software is RepeatMasker (RM) (http://www.repeatmasker.org). Interestingly, RM is also using RepBase as a reference collection and AB-BLAST, RM-BLAST, or cross-match as a search engine. In both cases, original search hits are processed by a series of Perl scripts to determine the structure of elements and classify them to one of known TE families. Both Censor and RM also employ user-provided libraries, including "third-party" lineage-specific libraries, e.g., TREP [123]. Over the years, RepeatMasker has become a standard tool for TE analyses, and often its output is used for more biologically oriented studies (see below). The aforementioned programs have one important drawback: since they are completely based on sequence similarity, they can detect only TEs that had been previously described. Nevertheless, similarity searches, like in many other bioinformatics tasks, should be the first approach for the analysis of repetitive elements.

Signature-based programs are searching for certain features that characterize specific TEs, for example, long terminal repeats (LTRs), target site duplications (TSDs), or primer-binding sites (PBSs). Since different types (families) of elements are structurally different, they require specific rules for their detection. Hence, many of the programs that use signature-based algorithms are specific for certain type of transposons. There are a number of programs specialized in detection of LTR transposons, which are based on a similar methodology. They take into account several structural features of LTR retroposons including size, distance between paired LTRs and their similarity, the presence of TSDs, and the presence of replication signals, i.e., the primer-binding site and the polypurine tract (PPTs). Some of the programs check also for ORFs coding for the *gag*, *pol*, and *env* proteins. LTR_STRUC [124] was one of the first programs based on this principle. It uses seed-and-extend strategy to find repeats located within user-defined distance. The candidate regions are extended based on the pairwise alignment to determine cognate LTRs' boundaries. Putative full-length elements are scored based on the presence of TSD, PBS, PPT, and reverse transcriptase ORF. However, because of the heuristics described above, LTR_STRUC is unable to find incomplete LTR transposons and in particular solo LTRs. Another limitation of this program is its Windows-only implementation that significantly prohibits automated large-scale analysis. Several other programs have been developed based on similar principles, e.g., LTR_par [125], find_LTR [126], LTR_FINDER [127], and LTRharvest [128]. Lerat tested performance of these programs [129], and although sensitivity of the methods was acceptable (between 40% and 98%), it was at the expense of specificity, which was very poor. In several cases, the number of falsely assigned transposons exceeded the number of correctly detected ones.

Another group of transposons that have a relatively conserved structure are MITEs and Helitrons. Several specialized programs were developed that take advantage of their specific structure. FINDMITE [130] and MUST [131] are tailored for MITEs, while HelitronFinder [132] and HelSearch [133] were developed for Helitron detection.

A further interesting approach to transposon classification was implemented by Abrusan et al. [120] in the software package called TEclass, which classifies unknown TE consensus sequences into four categories, according to their mechanism of transposition: DNA transposons, LTRs, LINEs, and SINEs. The classification uses support vector machines, random forests, learning vector quantization, and predicts ORFs. Two complete sets of classifiers are built using tetramers and pentamers, which are used in two separate rounds of the classification. The software assumes that the analyzed sequence represents a TE and the classification process is binary, with the following steps: forward versus reverse sequence orientation > DNA versus retrotransposon > LTRs versus nonLTRs (for retroelements) > LINEs versus SINEs (for nonLTR repeats). If the different methods of classification lead to conflicting results, TEclass reports the repeat either as unknown or as the last category where the classification methods agree (http://bioinformatics.uni-muenster.de/tools/teclass/index.hbi).

4.6 Pipelines

Recent years witnessed some attempt to create more complex, global analyses systems. One such a system is REPCLASS [134]. It consists of three classification modules: homology (HOM), structure (STR), and target site duplication (TSD). Each module can be run separately or in the pairwise manner, whereas the final step of the analysis involves integration of the results delivered by each module. There is one interesting novelty in the STR module, namely, implementation of *tRNAscan-SE* [135] to detect tRNA-like secondary structure within the query sequence, one of the signatures of many SINE families. The REPPET is another pipeline for TE sequence analyses. It uses “classical” three-step approach for de novo TE identification: self-alignment, clustering, and consensus sequences generation. However, the pipeline is using a spectrum of different methods at each step, followed by a rigorous TE classification step based on recently proposed classification of TEs [136]. Unfortunately, a complex implementation that makes installation and running the system rather difficult limits usage of the pipeline. The classification step seems to be unreliable as it may annotate lineage-specific TEs in wrong taxonomical lineages (Kouzel and Makalowski, unpublished data).

There are other attempts to create comprehensive systems for “repeatome” analysis. One of them is dnaPipeTE developed for mosquito genomes’ analyses [100]. Interestingly, dnaPipeTE works on the raw NGS data, which makes the pipeline well suited for genomes with lower sequencing depth. The raw reads are first subjected to *k*-mer count on the sampled data. The sampling of the data to size less than 0.25× of the genome is required to avoid clustering reads representing unique sequences. The determined repetitive reads are assembled into contigs using Trinity [137]. Although Trinity was originally developed for transcriptome assembly from RNA-seq data, it proves to be very useful for TEs assembly from short reads as it can efficiently determine consensus sequences of closely related transposons. In the next step, dnaPipeTE annotates repeats using RepeatMasker with either built-in or user-defined libraries. This is probably the weakest point of the pipeline as it will not annotate any novel TEs, which have no similar sequences present in the provided libraries. It would be useful to complement this step with model-based or machine learning approaches (*see* Subheading 4.5). After contigs’ annotation, copy number of the TEs are estimated using BLAST algorithm [91]. Finally, sequence identity between an individual TE and its consensus sequence is used to determine the relative age of the TEs. The pipeline produces a number of output files including several graphs, i.e., pie chart with the relative proportion of the main repeat classes and graph with the number of base pairs aligned on each TE contig and TE age distribution. Overall, the dnaPipeTE is very efficient, outperforming, according to the authors, RepeatExplorer by severalfold [100].

4.7 Meta-analyses

Most of the software developed are focused on the TE discovery and rarely offer more biological oriented analyses. Consequently, researchers interested in TE biology or using TE insertions as tools for another biological investigations need to utilize other resources. One of them is TinT (transposition in transposition), tool that applies maximum likelihood model of TE insertion probability to estimate relative age of TE families [138] (http://bioinformatics.uni-muenster.de/tools/tint/index.hbi). In the first steps, it takes RepeatMasker output to detect nested retroposons. Then, it generates a data matrix that is used by a probabilistic model to estimate chronology and activity period of analyzed families. The method was applied to resolve the evolutionary history of galliformes [139], marsupials [140], lagomorphs [141], squirrel monkey [142], or elephant shark [143].

Another interesting application that takes advantage of TEs is their use for detecting signatures of positive selection [144], a central goal in the field of evolutionary biology. A typical research scenario for this application would be investigating whether a specific TE fragment exapted into resident genomic features, such as proximal and distal enhancers or exons of spliced transcripts, has undergone accelerated evolution that could be indicative of gain of function events. In short, the test first requires the identification of all genomically interspersed TE fragments that are homolog to the TE segment of interest, which can be done through alignments with a family consensus sequence. Based on multi-species genome alignments, a second step involves identification of lineage-specific substitutions in every single homolog fragment, which are then consolidated into a distribution of lineage-specific substitutions that provides the expectation (null distribution) for a segment evolving largely without specific constraints (neutrally). A significantly higher number of lineage-specific substitutions observed in the TE fragment of interest compared to the null distribution could then be interpreted as a molecular signature of adaptive evolution. However, the possibility of confounding molecular mechanisms, such as GC-biased gene conversion [145–147], needs to be evaluated. We note that building the null distribution based only on data from intergenic regions, where transcription-coupled repair is absent, results in a more liberal estimate of the expected substitutions, which in turn leads to a more conservative estimate of the adaptive evolution. Additionally, building the null distribution requires the detection of many homolog fragments, which limits the applicability of the test to TE families with numerous members in a given genome. Prime examples would be human Alu or murine B1 SINEs. In theory, this test could also be used for detecting signatures of purifying selection by searching for fragments depleted of lineage-specific substitutions. However, the low level or complete lack of lineage-specific substitution is characteristic to many TE fragments, obscuring the effect of potential purifying forces.

5 Concluding Remarks

Annoying junk for some, hidden treasure for others, TEs can hardly be ignored [148]. With their diversity and high copy number in most of the genomes, they are not the easiest biological entities to analyze. Nevertheless, recent years witnessed increased interest in TEs. On the one hand, we observe improvement in computational tools specialized in TE analyses. Table 3 lists some of such tools and

Table 3
Selected resources for transposable elements discovery and analyses

Software	Address
AB-BLAST	http://www.advbiocomp.com/blast.html
ACLAME	http://aclame.ulb.ac.be/
BLASTER suite	http://urgi.versailles.inra.fr/index.php/urgi/Tools/BLASTER
Censor	http://www.girinst.org/censor/download.php
DOTTER	http://sonnhammer.sbc.su.se/Dotter.html
DROPOSON	ftp://biom3.univ-lyon1.fr//pub/drosoposon/
find_ltr	http://darwin.informatics.indiana.edu/cgi-bin/evolution/ltr.pl
FINDMITE	http://jaketu.biochem.vt.edu/dl_software.htm
FORRepeats	http://al.jalix.org/FORRepeats/
Gepard	http://cube.univie.ac.at/gepard
HelitronFinder	http://limei.montclair.edu/HT.html
HelSearch	http://sourceforge.net/project/showfiles.php?group_id=260708
HERVd	http://herv.img.cas.cz/
IRF	http://tandem.bu.edu/irf/irf.download.html
LTR_FINDER	http://tlife.fudan.edu.cn/ltr_finder/
LTR_MINER	http://genomebiology.com/2004/5/10/R79/suppl/s7
LTR_par	http://www.eecs.wsu.edu/~ananth/software.htm
MGEScan-LTR	http://darwin.informatics.indiana.edu/cgi-bin/evolution/daphnia_ltr.pl
MGEScan-nonLTR	http://darwin.informatics.indiana.edu/cgi-bin/evolution/nonltr/nonltr.pl
microTranspoGene	http://transpogene.tau.ac.il/microTranspoGene.html
MITE-Hunter	http://target.iplantcollaborative.org/mite_hunter.html
PILER	http://www.drive5.com/piler/
REannotate	http://www.bioinformatics.org/reannotate/index.html
ReAS	ftp://ftp.genomics.org.cn/pub/ReAS/software/
RECON	http://eddylab.org/software/recon/

(continued)

Table 3
(continued)

Software	Address
RepSeek	http://wwwabi.snv.jussieu.fr/public/RepSeek/
RepeatFinder	http://cbcb.umd.edu/software/RepeatFinder/
RepeatMasker	http://www.repeatmasker.org/
RepeatModeler	http://www.repeatmasker.org/RepeatModeler/
RepeatRunner	http://www.yandell-lab.org/software/repeatrunner.html
Repeat-match	http://mummer.sourceforge.net/
REPET	http://urgi.versailles.inra.fr/index.php/urgi/Tools/REPET
RepMiner	http://repminer.sourceforge.net/index.htm
REPuter	http://bibiserv.techfak.uni-bielefeld.de/reputer/
RetroMap	http://www.burchsite.com/bioi/RetroMapHome.html
SMaRTFinder	http://services.appliedgenomics.org/software/smartfinder/
SoyTEdb	http://www.soytedb.org
Spectral Repeat Finder	http://www.imtech.res.in/raghava/srf/
T-lex	http://petrov.stanford.edu/cgi-bin/Tlex.html
Tallymer	http://www.zbh.uni-hamburg.de/Tallymer/
TARGeT	http://target.iplantcollaborative.org/
TEclass	http://www.bioinformatics.uni-muenster.de/tools/teclass/
TE Displayer	http://labs.csb.utoronto.ca/yang/TE_Displayer/
TE nest	http://www.plantgdb.org/prj/TE_nest/TE_nest.html
TESD	http://pbil.univ-lyon1.fr/software/TESD/
TinT	http://www.bioinformatics.uni-muenster.de/tools/tint/
TIPseqHunter	https://github.com/fenyolab/TIPseqHunter
TRANSPO	http://alggen.lsi.upc.es/recerca/search/transpo/transpo.html
TranspoGene	http://transpogene.tau.ac.il/
Transposon-PSI	http://transposonpsi.sourceforge.net/
TRAP	http://www.coccidia.icb.usp.br/trap/tutorials/
TRF	http://tandem.bu.edu/trf/trf.html
TROLL	http://finder.sourceforge.net/
TSDfinder	http://www.ncbi.nlm.nih.gov/CBBresearch/Landsman/TSDfinder/
WikiPoson	http://www.bioinformatics.org/wikiposon/doku.php
VariationHunter	http://compbio.cs.sfu.ca/software-variation-hunter
Vmatch	http://www.vmatch.de/

the up-to-date list can be found at our web site: http://www.bioinformatics.uni-muenster.de/ScrapYard/. On the other hand, improved tools and new technologies enable biologists to explore new research avenues that might lead to novel, fascinating insights into the biology of mobile elements.

References

1. Waring M, Britten RJ (1966) Nucleotide sequence repetition - a rapidly reassociating fraction of mouse DNA. Science 154 (3750):791–794
2. Britten RJ, Kohne DE (1968) Repeated sequences in DNA. hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms. Science 161(841):529–540
3. Makalowski W (2001) The human genome structure and organization. Acta Biochim Pol 48(3):587–598
4. C._elegans_Sequencing_Consortium (1998) Genome sequence of the nematode C. elegans: a platform for investigating biology. Science 282(5396):2012–2018
5. SanMiguel P, Tikhonov A, Jin YK, Motchoulskaia N, Zakharov D, Melake-Berhan A, Springer PS, Edwards KJ, Lee M, Avramova Z, Bennetzen JL (1996) Nested retrotransposons in the intergenic regions of the maize genome. Science 274 (5288):765–768
6. Keller EF (1983) A feeling for the organism: the life and work of Barbara McClintock. W.H. Freeman, San Francisco
7. McClintock B (1950) The origin and behavior of mutable loci in maize. Proc Natl Acad Sci U S A 36(6):344–355
8. McClintock B (1951) Chromosome organization and genic expression. Cold Spring Harb Symp Quant Biol 16:13–47
9. McClintock B (1956) Controlling elements and the gene. Cold Spring Harb Symp Quant Biol 21:197–216
10. Malamy MH, Fiandt M, Szybalski W (1972) Electron microscopy of polar insertions in the lac operon of Escherichia coli. Mol Gen Genet 119(3):207–222
11. Ohno S (1972) So much "junk" DNA in our genome. In: Smith HH (ed) Brookhaven symposia in biology, vol 23. Gordon & Breach, New York, pp 366–370
12. Aronson AI, Bolton ET, Britten RI, Cowie DB, Duerksen JD, McCarthy BJ, McQuillen K, Roberts RB (1960) Biophysics. In: Yearbook 59, vol 59. Carnegie Institution of Washington, Washington, pp 229–279
13. Ehret CF, De Haller G (1963) Origin, development and maturation of organelles and organelle systems of the cell surface in Paramecium. J Ultrastruct Res 23(Suppl 6):1–42
14. Brosius J (1991) Retroposons--seeds of evolution. Science 251(4995):753
15. Makalowski W, Mitchell GA, Labuda D (1994) Alu sequences in the coding regions of mRNA: a source of protein variability. Trends Genet 10(6):188–193
16. Jordan IK, Rogozin IB, Glazko GV, Koonin EV (2003) Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends Genet 19(2):68–72. Pii S0168-9525(02)00006-9
17. Thornburg BG, Gotea V, Makalowski W (2006) Transposable elements as a significant source of transcription regulating signals. Gene 365:104–110. https://doi.org/10.1016/j.gene.2005.09.036. S0378-1119(05)00653-0 [pii]
18. Feschotte C (2008) Transposable elements and the evolution of regulatory networks. Nat Rev Genet 9(5):397–405. https://doi.org/10.1038/nrg2337
19. Mita P, Boeke JD (2016) How retrotransposons shape genome regulation. Curr Opin Genet Dev 37:90–100. https://doi.org/10.1016/j.gde.2016.01.001
20. Chuong EB, Elde NC, Feschotte C (2017) Regulatory activities of transposable elements: from conflicts to benefits. Nat Rev Genet 18 (2):71–86. https://doi.org/10.1038/nrg.2016.139
21. Franke V, Ganesh S, Karlic R, Malik R, Pasulka J, Horvat F, Kuzman M, Fulka H, Cernohorska M, Urbanova J, Svobodova E, Ma J, Suzuki Y, Aoki F, Schultz RM et al

(2017) Long terminal repeats power evolution of genes and gene expression programs in mammalian oocytes and zygotes. Genome Res 27(8):1384–1394. https://doi.org/10.1101/gr.216150.116

22. Wang L, Rishishwar L, Marino-Ramirez L, Jordan IK (2017) Human population-specific gene expression and transcriptional network modification with polymorphic transposable elements. Nucleic Acids Res 45 (5):2318–2328. https://doi.org/10.1093/nar/gkw1286
23. Venuto D, Bourque G (2018) Identifying co-opted transposable elements using comparative epigenomics. Develop Growth Differ 60(1):53–62. https://doi.org/10.1111/dgd.12423
24. Mahillon J, Chandler M (1998) Insertion sequences. Microbiol Mol Biol Rev 62 (3):725–774
25. Wilde C, Escartin F, Kokeguchi S, Latour-Lambert P, Lectard A, Clement JM (2003) Transposases are responsible for the target specificity of IS1397 and ISKpn1 for two different types of palindromic units (PUs), Nucleic Acid Res 31(15):4345–4353
26. Derbyshire KM, Grindley NDF (1996) Cis preference of the IS903 transposase is mediated by a combination of transposase instability and inefficient translation. Mol Microbiol 21(6):1261–1272. https://doi.org/10.1111/j.1365-2958.1996.tb02587.x
27. Ichikawa H, Ikeda K, Amemura J, Ohtsubo E (1990) Two domains in the terminal inverted-repeat sequence of transposon Tn3. Gene 86(1):11–17
28. Maekawa T, Amemura-Maekawa J, Ohtsubo E (1993) DNA binding domains in Tn3 transposase. Mol Gen Genet 236 (2–3):267–274
29. Weinert TA, Schaus NA, Grindley ND (1983) Insertion sequence duplication in transpositional recombination. Science 222 (4625):755–765
30. Turlan C, Chandler M (1995) IS1-mediated intramolecular rearrangements: formation of excised transposon circles and replicative deletions. EMBO J 14(21):5410–5421
31. Siguier P, Gourbeyre E, Varani A, Ton-Hoang B, Chandler M (2015) Everyman's guide to bacterial insertion sequences. Microbiol Spectr 3(2):MDNA3-0030-2014. https://doi.org/10.1128/microbiolspec.MDNA3-0030-2014
32. Finnegan DJ (1989) Eukaryotic transposable elements and genome evolution. Trends Genet 5(4):103–107
33. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH (2007) A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8(12):973–982. https://doi.org/10.1038/nrg2165. nrg2165 [pii]
34. Hughes SH (2015) Reverse transcription of retroviruses and LTR retrotransposons. Microbiol Spectr 3(2):MDNA3-0027-2014. https://doi.org/10.1128/microbiolspec.MDNA3-0027-2014
35. Kazazian HH Jr (2004) Mobile elements: drivers of genome evolution. Science 303 (5664):1626–1632. https://doi.org/10.1126/science.1089670. 303/5664/1626 [pii]
36. Malik HS, Henikoff S, Eickbush TH (2000) Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses. Genome Res 10(9):1307–1318
37. Leib-Mosch C, Haltmeier M, Werner T, Geigl EM, Brack-Werner R, Francke U, Erfle V, Hehlmann R (1993) Genomic distribution and transcription of solitary HERV-K LTRs. Genomics 18(2):261–269. https://doi.org/10.1006/geno.1993.1464
38. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J et al (2001) Initial sequencing and analysis of the human genome. Nature 409 (6822):860–921. https://doi.org/10.1038/35057062
39. Wicker T, Stein N, Albar L, Feuillet C, Schlagenhauf E, Keller B (2001) Analysis of a contiguous 211 kb sequence in diploid wheat (Triticum monococcum L.) reveals multiple mechanisms of genome evolution. Plant J 26(3):307–316. tpj1028 [pii]
40. Vicient CM, Kalendar R, Anamthawat-Jonsson K, Schulman AH (1999) Structure, functionality, and evolution of the BARE-1 retrotransposon of barley. Genetica 107 (1–3):53–63
41. SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL (1998) The paleontology of intergene retrotransposons of maize. Nat Genet 20(1):43–45. https://doi.org/10.1038/1695
42. Peterson DG, Schulze SR, Sciara EB, Lee SA, Bowers JE, Nagel A, Jiang N, Tibbitts DC, Wessler SR, Paterson AH (2002) Integration of Cot analysis, DNA cloning, and high-throughput sequencing facilitates genome characterization and gene discovery. Genome

Res 12(5):795–807. https://doi.org/10.1101/gr.226102. Article published online before print in April 2002

43. Zuker C, Lodish HF (1981) Repetitive DNA sequences cotranscribed with developmentally regulated Dictyostelium discoideum mRNAs. Proc Natl Acad Sci U S A 78 (9):5386–5390
44. Goodwin TJ, Poulter RT (2001) The DIRS1 group of retrotransposons. Mol Biol Evol 18 (11):2067–2082
45. Piednoel M, Bonnivard E (2009) DIRS1-like retrotransposons are widely distributed among Decapoda and are particularly present in hydrothermal vent organisms. BMC Evol Biol 9:86. https://doi.org/10.1186/1471-2148-9-86
46. Evgen'ev MB, Arkhipova IR (2005) Penelope-like elements - a new class of retroelements: distribution, function and possible evolutionary significance. Cytogenet Genome Res 110(1–4):510–521. https://doi.org/10.1159/000084984
47. Arkhipova IR (2006) Distribution and phylogeny of Penelope-like elements in eukaryotes. Syst Biol 55(6):875–885. https://doi.org/10.1080/10635150601077683
48. Singer MF (1982) Highly repeated sequences in mammalian genomes. Int Rev Cytol 76:67–112
49. Singer MF (1982) SINEs and LINEs: highly repeated short and long interspersed sequences in mammalian genomes. Cell 28 (3):433–434
50. Mills RE, Bennett EA, Iskow RC, Devine SE (2007) Which transposable elements are active in the human genome? Trends Genet 23(4):183–191. https://doi.org/10.1016/j.tig.2007.02.006
51. Biedler J, Tu Z (2003) Non-LTR retrotransposons in the African malaria mosquito, Anopheles gambiae: unprecedented diversity and evidence of recent activity. Mol Biol Evol 20(11):1811–1825. https://doi.org/10.1093/molbev/msg189. msg189 [pii]
52. Martin SL, Cruceanu M, Branciforte D, Wai-Lun Li P, Kwok SC, Hodges RS, Williams MC (2005) LINE-1 retrotransposition requires the nucleic acid chaperone activity of the ORF1 protein. J Mol Biol 348 (3):549–561. https://doi.org/10.1016/j.jmb.2005.03.003
53. Martin SL (2010) Nucleic acid chaperone properties of ORF1p from the non-LTR retrotransposon, LINE-1. RNA Biol 7 (6):706–711
54. Kapitonov VV, Jurka J (2003) Molecular paleontology of transposable elements in the Drosophila melanogaster genome. Proc Natl Acad Sci U S A 100(11):6569–6574. https://doi.org/10.1073/pnas.0732024100
55. Kajikawa M, Okada N (2002) LINEs mobilize SINEs in the eel through a shared 3′ sequence. Cell 111(3):433–444. S0092867402010413 [pii]
56. Houck CM, Rinehart FP, Schmid CW (1979) A ubiquitous family of repeated DNA sequences in the human genome. J Mol Biol 132(3):289–306
57. Jurka J, Zietkiewicz E, Labuda D (1995) Ubiquitous mammalian-wide interspersed repeats (MIRs) are molecular fossils from the mesozoic era. Nucleic Acids Res 23 (1):170–175
58. Wang H, Xing J, Grover D, Hedges DJ, Han KD, Walker JA, Batzer MA (2005) SVA elements: a hominid-specific retroposon family. J Mol Biol 354(4):994–1007. https://doi.org/10.1016/j.jmb.2005.09.085
59. Ostertag EM, Goodier JL, Zhang Y, Kazazian HH (2003) SVA elements are nonautonomous retrotransposons that cause disease in humans. Am J Hum Genet 73 (6):1444–1451. https://doi.org/10.1086/380207
60. Vanin EF (1985) Processed pseudogenes: characteristics and evolution. Annu Rev Genet 19:253–272
61. Maestre J, Tchenio T, Dhellin O, Heidmann T (1995) mRNA retroposition in human cells: processed pseudogene formation. EMBO J 14:6333–6338
62. Kabza M, Ciomborowska J, Makalowska I (2014) RetrogeneDB--a database of animal retrogenes. Mol Biol Evol 31 (7):1646–1648. https://doi.org/10.1093/molbev/msu139
63. Zhang Z, Harrison P, Gerstein M (2002) Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome. Genome Res 12:1466–1482
64. Torrents D, Suyama M, Zdobnov E, Bork P (2003) A genome-wide survey of human pseudogenes. Genome Res 13:2559–2567
65. Szcześniak MW, Ciomborowska J, Nowak W, Rogozin IB, Makałowska I (2011) Primate and rodent specific intron gains and the origin of retrogenes with splice variants. Mol Biol Evol 28:33–38

66. Goodwin TJ, Butler MI, Poulter RT (2003) Cryptons: a group of tyrosine-recombinase-encoding DNA transposons from pathogenic fungi. Microbiology 149. (Pt 11:3099–3109

67. Kojima KK, Jurka J (2011) Crypton transposons: identification of new diverse families and ancient domestication events. Mob DNA 2 (1):12. https://doi.org/10.1186/1759-8753-2-12

68. Bureau TE, Wessler SR (1994) Stowaway: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. Plant Cell 6(6):907–916. https://doi.org/10.1105/tpc.6.6.907. 6/6/907 [pii]

69. Feschotte C, Swamy L, Wessler SR (2003) Genome-wide analysis of mariner-like transposable elements in rice reveals complex relationships with stowaway miniature inverted repeat transposable elements (MITEs). Genetics 163(2):747–758

70. Zhou MB, Tao GY, Pi PY, Zhu YH, Bai YH, Meng XW (2016) Genome-wide characterization and evolution analysis of miniature inverted-repeat transposable elements (MITEs) in moso bamboo (Phyllostachys heterocycla). Planta 244(4):775–787. https://doi.org/10.1007/s00425-016-2544-0

71. Wicker T, Robertson JS, Schulze SR, Feltus FA, Magrini V, Morrison JA, Mardis ER, Wilson RK, Peterson DG, Paterson AH, Ivarie R (2005) The repetitive landscape of the chicken genome. Genome Res 15 (1):126–136. https://doi.org/10.1101/gr.2438004

72. Kapitonov VV, Jurka J (2001) Rolling-circle transposons in eukaryotes. Proc Natl Acad Sci U S A 98:8714–8719

73. Hood ME (2005) Repetitive DNA in the automictic fungus Microbotryum violaceum. Genetica 124(1):1–10

74. Pritham EJ, Feschotte C (2007) Massive amplification of rolling-circle transposons in the lineage of the bat *Myotis lucifugus*. Proc Natl Acad Sci U S A 104:1895–1900

75. Pritham EJ, Putliwala T, Feschotte C (2007) Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. Gene 390(1–2):3–17. https://doi.org/10.1016/j.gene.2006.08.008. S0378-1119(06)00537-3 [pii]

76. Kapitonov VV, Jurka J (2006) Self-synthesizing DNA transposons in eukaryotes. Proc Natl Acad Sci U S A 103:4540–4545

77. Kurtz S, Schleiermacher C (1999) REPuter: fast computation of maximal repeats in complete genomes. Bioinformatics 15 (5):426–427

78. Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R (2001) REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res 29(22):4633–4642

79. Delcher AL, Kasif S, Fleischmann RD, Peterson J, White O, Salzberg SL (1999) Alignment of whole genomes. Nucleic Acids Res 27(11):2369–2376

80. Delcher AL, Phillippy A, Carlton J, Salzberg SL (2002) Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Res 30(11):2478–2483

81. Li RQ, Ye J, Li SG, Wang J, Han YJ, Ye C, Wang J, Yang HM, Yu J, Wong GKS, Wang J (2005) ReAS: recovery of ancestral sequences for transposable elements from the unassembled reads of a whole genome shotgun. PLoS Comput Biol 1(4):313–321. https://doi.org/10.1371/Journal.Pcbi.0010043. Artn E43 [pii]

82. Price AL, Jones NC, Pevzner PA (2005) De novo identification of repeat families in large genomes. Bioinformatics 21:I351–I358. https://doi.org/10.1093/Bioinformatics/Bti1018

83. Kurtz S, Narechania A, Stein JC, Ware D (2008) A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. BMC Genomics 9:517. https://doi.org/10.1186/1471-2164-9-517

84. Lefebvre A, Lecroq T, Dauchel H, Alexandre J (2003) FORRepeats: detects repeats on entire chromosomes and between genomes. Bioinformatics 19(3):319–326. https://doi.org/10.1093/Bioinformatics/Btf843

85. Crochemore M, Ilie L, Seid-Hilmi E (2006) Factor oracles. In: Ibarra OH, Yen H-C (eds) Implementation and application of automata. Springer, Berlin, pp 78–89

86. Agrawal P, States D (1994) The Repeat Pattern Toolkit (RPT): analyzing the structure and evolution of the *C. elegans* genome. Proc Int Conf Intell Syst Mol Biol 2:9

87. Bao ZR, Eddy SR (2002) Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res 12 (8):1269–1276. https://doi.org/10.1101/Gr.88502

88. Edgar RC, Myers EW (2005) PILER: identification and classification of genomic repeats. Bioinformatics 21:I152–I158. https://doi.org/10.1093/Bioinformatics/Bti1003

89. Edgar RC (2007) PILER-CR: fast and accurate identification of CRISPR repeats. BMC Bioinform 8:18. https://doi.org/10.1186/1471-2105-8-18
90. Quesneville H, Bergman CM, Andrieu O, Autard D, Nouaud D, Ashburner M, Anxolabehere D (2005) Combined evidence annotation of transposable elements in genome sequences. PLoS Comput Biol 1 (2):166–175. Artn E22. https://doi.org/10.1371/Journal.Pcbi.0010022
91. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215(3):403–410
92. Fitch WM (1969) Locating gaps in amino acid sequences to optimize the homology between two proteins. Biochem Genet 3 (2):99–108
93. Gibbs AJ, McIntyre GA (1970) The diagram, a method for comparing sequences. Its use with amino acid and nucleotide sequences. Eur J Biochem 16(1):1–11
94. Sonnhammer EL, Durbin R (1995) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene 167(1–2):GC1–G10
95. Krumsiek J, Arnold R, Rattei T (2007) Gepard: a rapid and sensitive tool for creating dotplots on genome scale. Bioinformatics 23 (8):1026–1028. https://doi.org/10.1093/bioinformatics/btm039
96. Staton SE, Burke JM (2015) Transposome: a toolkit for annotation of transposable element families from unassembled sequence reads. Bioinformatics 31(11):1827–1829. https://doi.org/10.1093/bioinformatics/btv059
97. Novak P, Neumann P, Pech J, Steinhaisl J, Macas J (2013) RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. Bioinformatics 29(6):792–793. https://doi.org/10.1093/bioinformatics/btt054
98. Koch P, Platzer M, Downie BR (2014) RepARK--de novo creation of repeat libraries from whole-genome NGS reads. Nucleic Acids Res 42(9):e80. https://doi.org/10.1093/nar/gku210
99. Chu C, Nielsen R, Wu Y (2016) REPdenovo: inferring de novo repeat motifs from short sequence reads. PLoS One 11(3):e0150719. https://doi.org/10.1371/journal.pone.0150719
100. Goubert C, Modolo L, Vieira C, ValienteMoro C, Mavingui P, Boulesteix M (2015) 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 Biol Evol 7(4):1192–1205. https://doi.org/10.1093/gbe/evv050
101. Genome of the Netherlands Consortium (2014) Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat Genet 46 (8):818–825. https://doi.org/10.1038/ng.3021
102. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR (2015) A global reference for human genetic variation. Nature 526(7571):68–74. https://doi.org/10.1038/nature15393
103. UK10K Consortium, Walter K, Min JL, Huang J, Crooks L, Memari Y, McCarthy S, Perry JR, Xu C, Futema M, Lawson D, Iotchkova V, Schiffels S, Hendricks AE, Danecek P et al (2015) The UK10K project identifies rare variants in health and disease. Nature 526(7571):82–90. https://doi.org/10.1038/nature14962
104. Lack JB, Lange JD, Tang AD, Corbett-Detig RB, Pool JE (2016) A thousand fly genomes: an expanded Drosophila genome nexus. Mol Biol Evol 33(12):3308–3313. https://doi.org/10.1093/molbev/msw195
105. Lynch M, Gutenkunst R, Ackerman M, Spitze K, Ye Z, Maruki T, Jia Z (2017) Population genomics of Daphnia pulex. Genetics 206(1):315–332. https://doi.org/10.1534/genetics.116.190611
106. Gardner EJ, Lam VK, Harris DN, Chuang NT, Scott EC, Pittard WS, Mills RE, Genomes Project Consortium, Devine SE (2017) The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. Genome Res 27 (11):1916–1929. https://doi.org/10.1101/gr.218032.116
107. Ewing AD (2015) Transposable element detection from whole genome sequence data. Mob DNA 6:24. https://doi.org/10.1186/s13100-015-0055-3
108. Rishishwar L, Marino-Ramirez L, Jordan IK (2016) Benchmarking computational tools for polymorphic transposable element detection. Brief Bioinform. https://doi.org/10.1093/bib/bbw072
109. Helman E, Lawrence MS, Stewart C, Sougnez C, Getz G, Meyerson M (2014) Somatic retrotransposition in human cancer revealed by whole-genome and exome

sequencing. Genome Res 24(7):1053–1063. https://doi.org/10.1101/gr.163659.113

110. Lee E, Iskow R, Yang L, Gokcumen O, Haseley P, Luquette LJ III, Lohr JG, Harris CC, Ding L, Wilson RK, Wheeler DA, Gibbs RA, Kucherlapati R, Lee C, Kharchenko PV et al (2012) Landscape of somatic retrotransposition in human cancers. Science 337 (6097):967–971. https://doi.org/10.1126/science.1222077

111. Tubio JMC, Li Y, Ju YS, Martincorena I, Cooke SL, Tojo M, Gundem G, Pipinikas CP, Zamora J, Raine K, Menzies A, Roman-Garcia P, Fullam A, Gerstung M, Shlien A et al (2014) Mobile DNA in cancer. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. Science 345(6196):1251343. https://doi.org/10.1126/science.1251343

112. Thung DT, de Ligt J, Vissers LE, Steehouwer M, Kroon M, de Vries P, Slagboom EP, Ye K, Veltman JA, Hehir-Kwa JY (2014) Mobster: accurate detection of mobile element insertions in next generation sequencing data. Genome Biol 15(10):488. https://doi.org/10.1186/s13059-014-0488-x

113. Keane TM, Wong K, Adams DJ (2013) RetroSeq: transposable element discovery from next-generation sequencing data. Bioinformatics 29(3):389–390. https://doi.org/10.1093/bioinformatics/bts697

114. Tang Z, Steranka JP, Ma S, Grivainis M, Rodic N, Huang CR, Shih IM, Wang TL, Boeke JD, Fenyo D, Burns KH (2017) Human transposon insertion profiling: analysis, visualization and identification of somatic LINE-1 insertions in ovarian cancer. Proc Natl Acad Sci U S A 114(5):E733–E740. https://doi.org/10.1073/pnas.1619797114

115. Chen Y, Ye W, Zhang Y, Xu Y (2015) High speed BLASTN: an accelerated MegaBLAST search tool. Nucleic Acids Res 43 (16):7762–7768. https://doi.org/10.1093/nar/gkv784

116. Kent WJ (2002) BLAT--the BLAST-like alignment tool. Genome Res 12 (4):656–664. https://doi.org/10.1101/gr.229202. Article published online before March 2002

117. Kielbasa SM, Wan R, Sato K, Horton P, Frith MC (2011) Adaptive seeds tame genomic sequence comparison. Genome Res 21 (3):487–493. https://doi.org/10.1101/gr.113985.110

118. Casper J, Zweig AS, Villarreal C, Tyner C, Speir ML, Rosenbloom KR, Raney BJ, Lee CM, Lee BT, Karolchik D, Hinrichs AS, Haeussler M, Guruvadoo L, Navarro Gonzalez J, Gibson D et al (2018) The UCSC Genome Browser database: 2018 update. Nucleic Acids Res 46(D1): D762–D769. https://doi.org/10.1093/nar/gkx1020

119. Noll A, Grundmann N, Churakov G, Brosius J, Makalowski W, Schmitz J (2015) GPAC-genome presence/absence compiler: a web application to comparatively visualize multiple genome-level changes. Mol Biol Evol 32(1):275–286. https://doi.org/10.1093/molbev/msu276

120. Abrusan G, Grundmann N, DeMester L, Makalowski W (2009) TEclass-a tool for automated classification of unknown eukaryotic transposable elements. Bioinformatics 25 (10):1329–1330. https://doi.org/10.1093/bioinformatics/btp084

121. Jurka J, Klonowski P, Dagman V, Pelton P (1996) Censor - a program for identification and elimination of repetitive elements from DNA sequences. Comput Chem 20 (1):119–121

122. Bao W, Kojima KK, Kohany O (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA 6:11. https://doi.org/10.1186/s13100-015-0041-9

123. Wicker T, Matthews DE, Keller B (2002) TREP: a database for Triticeae repetitive elements. Trends Plant Sci 7(12):561–562. [pii] S1360-1385(02)02372-5

124. McCarthy EM, McDonald JF (2003) LTR_STRUC: a novel search and identification program for LTR retrotransposons. Bioinformatics 19(3):362–367. https://doi.org/10.1093/Bioinformatics/Btf878

125. Kalyanaraman A, Aluru S (2006) Efficient algorithms and software for detection of full-length LTR retrotransposons. J Bioinforma Comput Biol 4(2):197–216. S021972000600203X [pii]

126. Rho M, Choi JH, Kim S, Lynch M, Tang H (2007) De novo identification of LTR retrotransposons in eukaryotic genomes. BMC Genomics 8:90. https://doi.org/10.1186/1471-2164-8-90. 1471-2164-8-90 [pii]

127. Xu Z, Wang H (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res 35 (Web Server issue):W265–W268. https://doi.org/10.1093/nar/gkm286. gkm286 [pii]

128. Ellinghaus D, Kurtz S, Willhoeft U (2008) LTRharvest, an efficient and flexible software

for de novo detection of LTR retrotransposons. BMC Bioinform 9:18. https://doi.org/10.1186/1471-2105-9-18. 1471-2105-9-18 [pii]

129. Lerat E (2010) Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. Heredity 104(6):520–533. https://doi.org/10.1038/hdy.2009.165. hdy2009165 [pii]

130. Tu Z (2001) Eight novel families of miniature inverted repeat transposable elements in the African malaria mosquito, Anopheles gambiae. Proc Natl Acad Sci U S A 98 (4):1699–1704. https://doi.org/10.1073/pnas.041593198. 041593198 [pii]

131. Chen Y, Zhou F, Li G, Xu Y (2009) MUST: a system for identification of miniature inverted-repeat transposable elements and applications to Anabaena variabilis and Haloquadratum walsbyi. Gene 436(1–2):1–7. https://doi.org/10.1016/j.gene.2009.01.019. S0378-1119(09)00051-1 [pii]

132. Du C, Caronna J, He L, Dooner HK (2008) Computational prediction and molecular confirmation of Helitron transposons in the maize genome. BMC Genomics 9:51. https://doi.org/10.1186/1471-2164-9-51. 1471-2164-9-51 [pii]

133. Yang L, Bennetzen JL (2009) Structure-based discovery and description of plant and animal Helitrons. Proc Natl Acad Sci U S A 106(31):12832–12837. https://doi.org/10.1073/pnas.0905563106. 0905563106 [pii]

134. Feschotte C, Keswani U, Ranganathan N, Guibotsy ML, Levine D (2009) Exploring repetitive DNA landscapes using REPCLASS, a tool that automates the classification of transposable elements in eukaryotic genomes. Genome Biol Evol 1:205–220. https://doi.org/10.1093/Gbe/Evp023

135. Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25(5):955–964

136. Flutre T, Duprat E, Feuillet C, Quesneville H (2011) Considering transposable element diversification in de novo annotation approaches. PLoS One 6(1):e16526. https://doi.org/10.1371/journal.pone.0016526

137. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng QD, Chen ZH, Mauceli E, Hacohen N, Gnirke A, Rhind N et al (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29(7):644–U130. https://doi.org/10.1038/nbt.1883

138. Churakov G, Grundmann N, Kuritzin A, Brosius J, Makalowski W, Schmitz J (2010) A novel web-based TinT application and the chronology of the Primate Alu retroposon activity. BMC Evol Biol 10:376. https://doi.org/10.1186/1471-2148-10-376. 1471-2148-10-376 [pii]

139. Kriegs JO, Matzke A, Churakov G, Kuritzin A, Mayr G, Brosius J, Schmitz J (2007) Waves of genomic hitchhikers shed light on the evolution of gamebirds (Aves: Galliformes). BMC Evol Biol 7:190. https://doi.org/10.1186/1471-2148-7-190. 1471-2148-7-190 [pii]

140. Nilsson MA, Churakov G, Sommer M, Tran NV, Zemann A, Brosius J, Schmitz J (2010) Tracking marsupial evolution using archaic genomic retroposon insertions. PLoS Biol 8 (7):e1000436. https://doi.org/10.1371/journal.pbio.1000436

141. Kriegs JO, Zemann A, Churakov G, Matzke A, Ohme M, Zischler H, Brosius J, Kryger U, Schmitz J (2010) Retroposon insertions provide insights into deep lagomorph evolution. Mol Biol Evol 27 (12):2678–2681. https://doi.org/10.1093/molbev/msq162. msq162 [pii]

142. Baker JN, Walker JA, Vanchiere JA, Phillippe KR, St Romain CP, Gonzalez-Quiroga P, Denham MW, Mierl JR, Konkel MK, Batzer MA (2017) Evolution of Alu subfamily structure in the Saimiri lineage of new world monkeys. Genome Biol Evol 9 (9):2365–2376. https://doi.org/10.1093/gbe/evx172

143. Luchetti A, Plazzi F, Mantovani B (2017) Evolution of two short interspersed elements in Callorhinchus milii (Chondrichthyes, Holocephali) and related elements in sharks and the coelacanth. Genome Biol Evol 9(6). https://doi.org/10.1093/gbe/evx094

144. Gotea V, Petrykowska HM, Elnitski L (2013) Bidirectional promoters as important drivers for the emergence of species-specific transcripts. PLoS One 8(2):e57323. https://doi.org/10.1371/journal.pone.0057323

145. Kostka D, Hubisz MJ, Siepel A, Pollard KS (2012) The role of GC-biased gene conversion in shaping the fastest evolving regions of the human genome. Mol Biol Evol 29 (3):1047–1057. https://doi.org/10.1093/molbev/msr279

146. Capra JA, Hubisz MJ, Kostka D, Pollard KS, Siepel A (2013) A model-based analysis of GC-biased gene conversion in the human and chimpanzee genomes. PLoS Genet 9(8):

e1003684. https://doi.org/10.1371/journal.pgen.1003684

147. Gotea V, Elnitski L (2014) Ascertaining regions affected by GC-biased gene conversion through weak-to-strong mutational hotspots. Genomics 103(5–6):349–356. https://doi.org/10.1016/j.ygeno.2014.04.001

148. Makalowski W (2000) Genomic scrap yard: how genomes utilize all that junk. Gene 259 (1–2):61–67. https://doi.org/10.1016/S0378-1119(00)00436-4