HAL
open science

Repeat in genomes: how and why you should consider them in genome analyses

Emmanuelle Lerat

To cite this version:

Emmanuelle Lerat. Repeat in genomes: how and why you should consider them in genome analyses. Michael Gribskov; Christian Schonbach; Kenta Nakai; Shoba Ranganathan. Encyclopedia of Bioinformatics and Computational Biology, 2, Elsevier, pp.210-220, 2019, 9780128114148. 10.1016/B978-0-12-809633-8.20227-6 . hal-02989096

HAL Id: hal-02989096

https://cnrs.hal.science/hal-02989096

Submitted on 5 Nov 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Repeat in genomes: how and why you should consider them in genome analyses.

Emmanuelle LERAT

Univ Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive UMR5558, F-69622 Villeurbanne, France
emmanuelle.lerat@univ-lyon1.fr
33+ 4 72 43 29 18

Abstract
Repeats, and more particularly Transposable Elements (TEs) are major components of eukaryotes due to their effect at short and long terms on the genome evolution and species adaptation. It is thus important to be able to identify them in genome sequences to allow further analyses that can decipher their action. To perform these tasks, numerous tools have been developed to answer specific biological questions going from their annotation in assembled and non-assembled genome to polymorphic sequence variation among natural populations.

Keywords
Repeats, Transposable element, genome annotation, NGS sequencing, polymorphic insertion

1. Introduction

1.1. Repeats: a large bestiary grouping very different sequences

Although the genome size has been shown in prokaryotes to largely correlate to its complexity, i.e. the number of genes, it is not the case in eukaryotes. This observation, often referred to as the C-value paradox, can be partly explained by the fact that coding genes in eukaryotic genomes may represent only a tiny fraction of the total genome (Elliott and Gregory 2015). In fact, it has been shown that the proportion of non-coding sequences in genomes may be particularly huge. For example, the coding genes in the human genome only represent 2% (Lander et al. 2001). A large proportion of the non-genic sequences are represented by repeats. In a genome, repeats correspond to non-coding sequences present in several occurrences. Two main categories of repeats are described, the tandem repeats and the interspersed repeats (figure 1). Another type of repeats exists in a genome which correspond to segmental duplications, which are very large nearly identical sequences (from 1 to 400 kb) often referred to as “low-copy repeats” (Sharp et al. 2005).

The tandem repeats correspond to sequences having a size from few to hundred of base pairs (bp) occurring in tandem and over several hundred of kilo base pairs (kb), which characterized them as “highly repeated sequences”. We can distinguish among this category the Simple Sequence Repeats (SSR) also called Short Tandem Repeats (STR) or microsatellites, which are a short tract of adjacent DNA motifs (around 1 to 13 bp) and the minisatellites which are longer (around 14 to 500 bp), both types being repeated several times (from 5 to 50 times for example). This type of repeats usually occurs at particular regions of the chromosomes corresponding to the telomeres and centromeres, but they also can be found throughout the genome, especially in regulatory and coding regions of genes (Richard et al. 2008).

The interspersed repeats stand for sequences that are more often known under the name transposable

elements (TEs). These particular sequences were discovered during the 1950's by Barbara McClintock (McCLINTOCK 1950). TEs have the particularity to be able to move from one position to another along the chromosomes since the majority of them encode all the proteins necessary for their transposition. Various types exist, according to structural features, the transposition intermediate (RNA or DNA), and their evolutionary origin (Wicker et al. 2007; Kapitonov and Jurka 2008). Retrotransposons use an RNA intermediate and form the class I, in which are found the LTR (Long Terminal Repeat)-retrotransposons (endogenous retrovirus-like mobile elements) which possess from two to three open reading frames (gag, pol, and env) and the non-LTR retrotransposons grouping the LINE and the SINE elements (standing for Long and Short Interspersed Nuclear Elements respectively). DNA transposons use a DNA intermediate and form the class II. They possess short terminal inverted repeats (TIRs) at their extremities surrounding one open reading frame coding for a transposase. Different types of DNA transposons have been classified mainly on the basis of the presence or absence of a catalytic site in the protein responsible for their transposition. A specific category of non-autonomous elements among the DNA transposons exist, the MITEs (Miniature Inverted-repeat Transposable Elements) that may be particularly numerous in some genomes. Depending on the organism, the proportion of TEs can be highly variable and at times very large. For example, their proportion in genomes represent 3% in yeast (Kim et al. 1998), 15% in *Drosophila* (Dowsett and Young 1982), 45% in human and in the mouse (Lander et al. 2001; Waterston et al. 2002), and more than 80% in maize (Schnable et al. 2009).

1.2. “From junk to funk”: the importance of repeats in the genome functioning and its evolution

Repeats, and more particularly TEs, have long been considered as selfish and unnecessary components of the genomes. However, since the work of Britten and Davidson (Britten and Davidson 1969) where TEs were for the first time considered as potentially playing a role in the gene regulation, numerous examples have flourished allowing to show the genuine importance of such sequences in genomes (Biemont 2010). By their ability to move and because they are repeated, TEs can promote various types of mutations, which are expected to be mostly deleterious when affecting functional regions. When TE insertions occur in or near protein-coding genes, they can result in coding sequence modification or alteration of their splicing or polyadenylation patterns, therefore disrupting the protein coding capacity of the gene. Moreover, because TEs possess their own regulatory sequences, they can alter the normal expression pattern of neighboring genes while inserted in an intergenic region (Kidwell and Lisch 2000; Biémont and Vieira 2006). The possibility of homologous recombination between copies can also promote illegitimate recombinations, chromosome breakages, deletions and genome rearrangements (Kidwell and Lisch 2000; Biémont and Vieira 2006). In human, 0.3 % of TE insertions have been suggested for causing disease (Belancio et al. 2008) and approximately 96 new transposition events were directly linked to single-gene diseases (Hancks and Kazazian 2012). For example, the *Alport* syndrome has been shown to be due to a TE mediated rearrangement resulting in the partial deletion and fusion of two genes (Segal et al. 1999). More specific to cancer, the disruption of the APC gene caused by the TE insertion is involved in a colon cancer (Miki et al. 1992). Despite the deleterious effects they may have, TEs have also been associated with useful adaptation for their host genome. For example, the antigen receptor gene assembly by V(D)J recombination in vertebrates is performed by genes that originated from a DNA transposon (Agrawal et al. 1998). TE insertions near specific genes confer resistance to insecticide for some insects (Rostant et al. 2012). These examples make TEs to be now considered as major players in genome evolution due to the genetic and epigenetic diversity they can promote (Biémont and Vieira 2006). Similarly, tandem repeats have been found to play a fundamental role in the organization of the genome (Dumbovic, Sonia-V. Forcales, et al. 2017). For

example, changes in mini- and micro-satellites correlate with various diseases but are also associated with gene regulation (Dumbovic, Sonia-V Forcales, et al. 2017).

In addition to the fundamental biological role repeats have in genomes, in the current big sequencing era, they also represent a technical challenge that may complicate the task of genome assembly and sequence alignment. For example, the presence of repeats in a genome is the major source of genome mis-assemblies via rearrangement assembly errors and collapsed repeats but also in the assignment of splicing events and gene expression estimate in transcriptome analyses (Treangen and Salzberg 2012). Since simply ignoring these sequences is not an issue in genomics, it is important to be able to identify them.

2. Bioinformatic approaches to identify, annotate and analyze repeats in genomes

Since several decades, numerous bioinformatics tools have been developed to allow a better identifications of repeats in genome assemblies (Lerat 2010; Modolo and Lerat 2013; Saha et al. 2008; Bergman and Quesneville 2007). New tools continue to arise regularly to follow the progress in sequencing technology in particular, but also to response to specific biological questions. According to the type of data on which they can work and the biological question, the different tools can be separated into various categories (table 1).

2.1. Detection of repeats in assembled genomes

Diverse methods exist to allow the detection and annotation of repeats in assembled genomes. These methods depend on the type of repeats and on the knowledge we have concerning their content inside the organism under investigation. Repeat annotation is a particularly complex computing problem due to the nature of the sequences, especially in the case of the TEs. Indeed, the TEs are not always exact repeats since a large divergence among copies can occur and TEs also can be found in the genome inserted inside each others (nested insertions). The detection of TEs has led to the development of a large number of different tools that fall in different categories according to the approach they use. Two main types of approaches exist: the library- or signature-based methods and the *ab initio* methods.

The library- or signature-based methods all require a certain amount of knowledge concerning the searched TEs. Library-based methods compare the genome sequence to a set of reference sequences corresponding to TE consensus (i.e. library) to search for their occurrences in the genome. It is thus an approach by sequence homology. The most known and used program from this category is *RepeatMasker* (Smit et al. 1996-2010) whose search engines include *nhmmer*, *cross_match*, *ABBlast/WUBlas*t, *RMBlast* and *Decypher*. It relies on a library of consensus sequences of TEs called Repbase (Bao et al. 2015). It is also able to identify low complexity DNA sequences, which can correspond to tandem repeats, by using sequence homology and the *Tandem Repeat Finder* program (Benson 1999). The main outputs of the program are a global annotation of the repeats that are present in the query sequence as well as a modified version of the query sequence in which all the annotated repeats have been masked. Recently, a program has been developed to allow a more detailed analyses of one output as well as an easy way to retrieve the identified sequences (Bailly-Bechet et al. 2014). Although *RepeatMasker* is fast and quite efficient, the main drawback consists in the need to already know the sequences of the TEs that are present in the genome under investigation or at least in not too far closely related ones. It is thus not possible by this approach to detect new TE families. The signature-based methods allow to have less knowledge concerning the searched TEs since they seek the

genome sequence for nucleotidic or proteic motifs, or particular structural features from a specific class of TEs. LTR-retrotransposons can be identified based on their structure (presence of two direct repeats (LTR) at their extremities, size, presence of protein motifs inside the coding parts etc.) by different programs like *LTR_STRUC* (McCarthy and McDonald 2003), *find_LTR* (Rho et al. 2007) , *LTR-finder* (Xu and Wang 2007) and *LTRharvest* (Ellinghaus et al. 2008). Outputs from *LTRharvest* can further be analyzed by *LTRdigest* (Steinbiss et al. 2009) to annotate internal features of LTR-retrotransposons. These programs have various amount of success since they can find a lot of false positives, meaning that their output files need to be manually curated. They also are designed to find only full-length elements with two LTRs at both extremities of the element and sufficiently conserved. The identification of non-LTR retrotransposons has been the goal of some programs like *TSDfinder* (Szak et al. 2002) and *SINEDR* (Tu et al. 2004). DNA transposons can also be searched for using structural and sequence characteristics of such elements using programs like *TRANSPO* (Santiago et al. 2002), *MUST* (Chen et al. 2009), recently updated into a new version MUSTv2 (Ge et al. 2017), and *MITE-hunter* (Han and Wessler 2010). All signature-based programs are thus mainly designed to help the researcher finding very specific types of TEs and thus concerning very specific biological questions since with these approaches, a large proportion of repeats remains ignored.

Alternatively to the precedent approaches, *ab initio* methods have been developed to detect virtually all kind of repeats in a genome without any *a priori* knowledge. These approaches can be separated into two distinct categories. In the first category, the methods first use self-comparison approaches of sequences to identify repeats and then use clustering methods to group them into families, before generating a consensus sequence for each detected family (method implemented in *RECON* (Bao and Eddy 2003) or *PILER* (Edgar and Myers 2005)). In the second category are grouped programs using k-mer and spaced seed approaches , which count “words” (method implemented in *ReAS* (Li et al. 2005) or *RepeatScout* (Price et al. 2005)). In that case, a repeat is defined as a sub-sequence that appears more than once in a longer sequence. All *ab-initio* programs have very varying rates of success in the identification of repeats depending on the data (quality of the genome assembly, proportion and age of the repeats in the genome). Moreover, the results produced by these methods are generally raw implying further analyses to identify the different type of repeats. To help with this step, some classification programs have been proposed like *TEclass* (Abrusán et al. 2009), *REPCLASS* (Feschotte et al. 2010), and *PASTEC* (Hoede et al. 2014), which use structural feature and sequence similarities to determine to which type of TE a sequence can be associated. The level of precision is variable according to the tool, *PASTEC* having currently the finer level in the assignation of TE type according to the classification proposed by Wicker and colleagues (Wicker et al. 2007). Globally, no stand alone *ab initio* program can discover all repeated sequences present in a genome (Platt et al. 2016). This is why approaches using several different programs to optimize repeat finding have been developed like, for example, the pipelines *REPET* (Flutre et al. 2011) and *RepeatModeler* (Smit and Hubley; http://www.repeatmasker.org/RepeatModeler.html). They tend to give better results and have been largely used in the scientific community. For example, *RepeatModeler* has recently been used to identify TEs in the genome of a fungi (Castanera et al. 2016) while this program was used in addition to *REPET* in the genome of the Atlantic salmon (Lien et al. 2016).

A large number of tools also exist to detect specifically tandem repeats (see for a review (Lim et al. 2013)). One of the most used tools to identify these sequences is the *Tandem Repeat Finder* program, which has been developed almost 20 years ago and that is still maintain by his developer (Benson 1999). The algorithm of this program uses the approach of matching k-tuples, i.e. two windows of k consecutive characters from a nucleotide sequence that have identical content. This requires no *a priori*

knowledge concerning the pattern of the repeat, its size or the number of copies. A similar approach is used in programs like *XSTREAM* (Newman and Cooper 2007) and *MREPS* (Kolpakov et al. 2003). Other programs use improved dynamic programming algorithms like *STAR* (Delgrange and Rivals 2004) and *TRED* (Sokol et al. 2007). More recently, the program *ReD tandem* (Audemard et al. 2012) was developed using a flow based chaining algorithm. Some programs, like any type of dot-plot programs, are based solely on sequence self-alignment (SSA) algorithms, which are particularly efficient for the detection of long repeats. Their drawbacks are that these programs are relatively slow due to their time complexity and usually fail to identify short repeats.

2.2. Detection of repeats in unassembled genomes

The development of the next generation sequencing (NGS) technologies has overturned our approach to genomics with a huge amount of data being produced everyday (Margulies et al. 2005; Mardis 2017). We thus now have access to more data on very various organisms at low cost and with fewer bias than the previous sequencing technologies (Wicker et al. 2006). Currently however, the data produced by regular NGS technology like Illumina, correspond to rather short sequences due to the small size of the reads (maximum of 300 bp length) (Mardis 2017). This short size implies that the assembly of the original DNA sequences is the most challenging and time-consuming step especially when the considered organism is rich in repeats. To assemble a genome in this condition often leads to unfinished drafts of very numerous scaffolds, a large number of them corresponding to unplaced repeats on chromosomes.

By their repetitive nature, repeats represent portions of the genome with the best coverage, especially in the case of genome survey sequencing, where a sample of a complete genome is actually sequenced. Indeed, for a genomic coverage of 0.01X, each repeat having 1000 occurrences will theoretically have a coverage of 10X (Macas et al. 2007). This situation has allowed the development of new approaches to detect repeats directly from the raw data, without the need for any homology search nor assembly. The first method to have been developed based on this assumption and that is able to work with short reads is the *AAARF* algorithm (DeBarry et al. 2008). This approach uses *BLAST* (Altschul et al. 1997) to compare a read against all the others and obtain its nucleotide coverage. This value is used to determine overlapping reads that will be aligned to reconstruct a new sequence. Iteratively, the program will elongate each new sequence to assemble a set of TE contigs. Another type of approach that is also working on genomic sample is based on the construction of sequence clusters like the *SeqGraphR* program implemented in the *RepeatExplorer* pipeline (Novák et al. 2010; Novák et al. 2013). In this method, reads are clustered using a hierarchical agglomeration algorithm. Various graph metrics are computed to discriminate between different types of repeats and the assembly of TE sequences gives consensus sequences. Based on a similar approach, the *Transposome* program has been proposed more recently that also uses a graph-based analysis of similarity between reads. (Staton and Burke 2015). As an alternative to read clustering, another approach, *DNApipeTE* (Goubert et al. 2015), proposes to use the RNAseq assembler *Trinity* (Grabherr et al. 2011) to build repeat contigs from genomic samples of less than 1X of coverage. The use of *Trinity* allows to recover alternative repeat consensus of a given family by producing distinct contigs for each structural variant. Alternatively, other tools like *Tedna* (Zytnicki et al. 2014), *RepARK* (Koch et al. 2014) and *REPdenovo* (Chu et al. 2016) use directly a de Bruijn graph on the most represented k-mers to perform TE assembly.

Although the previous methods are supposed to be able to find any kind of repeats, they usually are best suited to discover TEs. They may be less powerful concerning tandem repeats. This is why some

specific tools have been designed to specifically uncover tandem repeats from raw reads. The first tool, *VNTRseek* (Gelfand et al. 2014), is a variant detection tool that compares reads in which tandem repeats have been detected using *Tandem Repeat Finder* (Benson 1999) to a set of tandem repeats present in a reference genome. By this comparison, the variable number of tandem repeats is determined. A recent pipeline called *TAREAN* (Novák et al. 2017) uses the principles of graph-based repeat clustering, as implemented in the *RepeatExplorer* pipeline, as well as tools facilitating the unsupervised identification and characterization of satellite repeats from raw reads. The reconstruction of the satellite sequences is based on k-mer decomposition and counting.

Globally, all these programs allow to determine the global proportion of repeats inside a genome, with sometime the possibility to estimate the copy number, as well as a catalog of the different types of repeats with the production of consensus sequences. However, since these programs work on raw reads, the information concerning the exact positions of these repeats is missing. Other approaches have thus been developed to specifically answer this question by comparing the copy number variation of repeats between two genomes.

2.3. Identification of the repeat copy number variation

When the first genomes were sequenced, it was considered that only one would be enough to understand the functioning and evolution of a given species. However, having only one genome is not enough to uncover the polymorphism existing among individuals. Particularly for repeats, some of which being able to move and replicate themselves in the genome, it is known that variations exist in term of copy number and insertion sites among natural populations (Petrov et al. 2011; Boulesteix et al. 2006). With the decrease in cost of sequencing, it is now possible to obtain data from several individuals of a given population and of several populations of a given species. This has open the door to perform high throughput population genomics when using pooled sequencing data. Since it is not always possible to obtain good assembly with these data, new tools have been developed to specifically search for differences when compared to a reference genome. Thus, the goal of the bioinformatic tools developed for this purpose is to determine either one or the three types of TE insertions: insertions shared between the reference and the analyzed data (fixed insertions), and polymorphic insertions corresponding to insertions either absent from the analyzed data or absent from the reference genome (new insertions) (figure 2).

Several tools have been developed to determine the structural variation due to TEs in genomes and some attempts have been made to evaluate and review them (Ewing 2015; Rishishwar et al. 2016). The various methods have all in common in their process to first map the reads on a reference genome and/or on a set of annotated TE sequences before applying various filters and metrics to retain the informative ones. Then, two approaches exist that may be combined to analyze the results and to detect the presence/absence of a TE insertion. In the first approach, the program considers discordant read pairs, which are read pairs with one member matching uniquely on the reference genome sequence and the other matching on different copies from a TE family (figure 3A). This type of approach is used in the programs *TE-locate* (Platzer et al. 2012), *TraFiC* (Tubio et al. 2014) and *TE-Tracker* (Gilly et al. 2014). The other approach consists in considering split reads, i.e. reads which overlap a junction between the genome and an inserted TE copies, with one part of the read mapping uniquely on the genome while the other part maps on TE sequences (figure 3B). The programs *RelocaTE* (Robb et al. 2013), *ngs-TE-mapper* (Linheiro and Bergman 2012), *TIDAL* (Rahman et al. 2015), *ITIS* (Jiang et al. 2015), and *T-Lex2* (Fiston-Lavier et al. 2015) use this approach. Other programs use both approaches

like *Tea* (Lee et al. 2012), *RetroSeq* (Keane et al. 2013), *TEMP* (Zhuang et al. 2014), *Mobster* (Thung et al. 2014), *Tangram* (Wu et al. 2014), *TranspoSeq* (Helman et al. 2014), *Jitterbug* (Hénaff et al. 2015), *DD_Detection* (Kroon et al. 2016), *popoolation_TE2* (Kofler et al. 2016) and *MELT* (Gardner et al. 2017). These different programs have been developed to answer specific biological questions, which make them either consider individual or population data to estimate the insertion frequency, to consider in majority polymorphic insertions (especially new insertions in the case of cancer research) but sometimes also shared insertions, and in some cases to take into account the genotype status of the detected insertions, i.e. if the insertion is present on only one (heterozygous) or two (homozygous) homologous chromosomes. For example, the program *T-Lex2* has been developed to compute the frequency insertions of TE copies that are present in the reference genome of *Drosophila melanogaster* in natural populations, whereas the program *Tea* has been developed to identify only new TE insertions in human cancers corresponding to somatic insertions.

3. Discussion

Repeats, and more particularly TEs, are important component of the eukaryote genomes that cannot be simply ignored when performing sequencing and assembly tasks. A lot of various bioinformatics tools have been developed during the last 20 years to allow a better handling of these particular sequences. Such tools need to evolve jointly with the evolution of sequencing technologies. Currently, one of the major problem with the current whole genome sequencing data produced to identify repeat insertions is the size of the available sequence reads (< 300 bp). Full-length TE insertions ranged between 500 bp to 10 kb, which implies that several reads are needed to cover an entire TE copy. This step may be particularly difficult since several insertions may present a very high degree of nucleotide identity between themselves. In that case, it is often impossible to recover some insertions and the only way to distinguish them is to take into account the genomic environment, i.e. the flanking regions around the insertion. However, sequence reads being shorter than the majority of TE insertions and reads overlapping the junction of the genomic region and the TE insertions being not numerous and difficult to map, this task is particularly challenging and lead to a loss of information. A way to tackle this difficulty is to produce longer reads. The third generation of DNA sequencing is currently under development and some already used techniques may be helpful, although still expensive or not optimized. For example, PacBio sequencing allows to produce sequences up to 20 kb but the rate of errors is still quite high (Mardis 2017). However, this technique may be used in addition to short-read sequencing techniques to enhance the quality of a genome assembly (Pendleton et al. 2015). The Illumina synthetic long-read technique has been successfully tested to perform the *de novo* assembly and resolve TE sequences in the genome of *Drosophila melanogaster* (McCoy et al. 2014). However this technique is still expensive since the coverage that is needed is very high (Mardis 2017). The MinION technology, which performs a single molecule sequencing, has been recently successfully used to detect new TE insertions in the plant genome *Arabidopsis thaliana* allowing to show that the high error rate of the technique could be compensated by the read length in this particular question (Debladis et al. 2017). Of course, as soon as these technologies will become the new standard, the current tools for TE analyses will become obsolete and new methodological procedures will need to be developed. In this way, a bioinformatic tool has recently been proposed to determine the presence/absence of TE insertions using reads produced by the PacBio technology (*LorTE*, (Disdero and Filée 2017)).

4. Future Directions

The identification of TE insertions is also becoming very important in the field of epigenetics research.

Indeed, TEs are known to be associated to particular epigenetic modifications that may impact the neighboring genes (Eichten et al. 2012; Estecio et al. 2012). Very few tools have been developed to allow the direct association between TEs and epigenetic modifications. For example, a web interface has been developed to specifically study histone modification enrichment of repeats taking advantage of their increased sequence coverage in ChIP-seq data (Day et al. 2010). An advantage of this method is that it incorporates both ambiguously and uniquely mapped reads to avoid bias due to read mapping on consensus TE sequences. However, the results are not given at the insertion level and thus the information concerning the genomic environment of each insertion is lost. Several efforts have been made to develop methods allowing the association of small RNA data to TE sequences either at a global (*piPipes*, (Han et al. 2015)) or at a individual copy level (*TEtools*, (Lerat et al. 2017)). More recently, a new program has been developed allowing both the identification of new TE insertions and their associated DNA methylation in MethylC-seq data (*EpiTEome*, (Daron and Slotkin 2017). In summary, methodological efforts are still needed to study the epigenetic modifications directly associated to TE sequences. It is particularly important to consider the global sequence diversity of a TE family that may be very variable but also the genomic environment of a given insertion that may have consequences on the associated epigenetic modifications. These issues are currently difficult to handle due to the short size of sequence reads. However, as long read technologies will continue to be developed, it should open the door to new developments in a close future.

5. Closing remarks

The domain of TE annotation in genome sequences is constantly producing new tools, which are supposed to outperform the previous ones and to offer new ways to handle the ever growing sequence data. However, the question of the impartial evaluation of these tools still remains. Indeed, the different tools have usually different competencies and may be at some point complementary. A clear problem underlying this situation is the lack of common standard data that would allow an unbiased estimation of any new tools (Hoen et al. 2015). There is still room in that domain to propose various benchmarks according to the biological questions asked behind each tool.

Nomenclature

Bp, base pairs; Kb, kilo base pairs; SSR, Simple Sequence Repeat; STR, Short Tandem Repeats; TEs, Transposable Elements; SSA, sequence self-alignment; NGS, Next Generation Sequencing; LINE, Long Interspersed Nuclear Elements; SINE, Short Interspersed Nuclear Elements; TIR, terminal inverted repeat; ChIP, Chromatin Immuno Precipitation; MITEs, Miniature Inverted-repeat Transposable Elements.

Figure legends

Figure 1: schematic representation of the different types of repeats

Figure 2: the different types of TE insertions when comparing a reference genome and a newly sequenced one.

Figure 3: schematic representation of discordant read pairs (A) and split reads (B) that are used to identify TE insertion in raw data by comparison to a reference genome. Lines in black correspond to reads mapping on unique genomic regions and lines in red correspond to reads mapping on TE sequences.

References

Abrusán, G., Grundmann, N., DeMester, L. and Makalowski, W. (2009) TEclass - A tool for automated classification of unknown eukaryotic transposable elements. *Bioinformatics*, **25**, 1329–1330.

Agrawal, A., Eastman, Q.M. and Schatz, D.G. (1998) Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. *Nature*, **394**, 744–51.

Altschul, S.F. et al., (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, **25**, 3389–402.

Audemard, E., Schiex, T. and Faraut, T. (2012) Detecting long tandem duplications in genomic sequences. *BMC Bioinformatics*, **13**, 83.

Bailly-Bechet, M., Haudry, A. and Lerat, E. (2014) "One code to find them all": a perl tool to conveniently parse RepeatMasker output files. *Mobile DNA*, **5**, 13.

Bao, W., Kojima, K.K. and Kohany, O. (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, 6, 11.

Bao, Z. and Eddy, S.R. (2003) Automated *de novo* identification of repeat sequence families in sequenced genomes. *Genome Research*, **13**, 1269–1276.

Belancio, V., Hedges, D.J. and Deininger, P. (2008) Mammalian non-LTR retrotransposons: For better or worse, in sickness and in health. *Genome Research*, **18**, 343–358.

Benson, G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research*, **27**, 573–80.

Bergman, C.M. and Quesneville, H. (2007) Discovering and detecting transposable elements in genome sequences. *Briefings in Bioinformatics*, **8**, 382–392.

Biemont, C., (2010) A brief history of the status of transposable elements: from junk DNA to major players in evolution. *Genetics*, **186**, 1085–1093.

Biémont, C. and Vieira, C. 2006. Genetics: junk DNA as an evolutionary force. *Nature*, **443**, 521–524.

Boulesteix, M., Simard, F., Antonio-Nkondjio, C., *et al*. 2006. Insertion polymorphism of transposable elements and population structure of *Anopheles gambiae* M and S molecular forms in Cameroon. *Molecular Ecology*. **16**, 441-452.

Britten, R.J. and Davidson, E.H. 1969. Gene regulation for higher cells: a theory. *Science*, **165**, 349–357.

Castanera, R., López-Varas, L., Borgognone, A., *et al.* (2016) Transposable elements versus the fungal Genome: impact on whole-genome architecture and transcriptional profiles. *PLoS Genetics*, **12**, e1006108.

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**, 1–7.

Chu, C., Nielsen, R. and Wu, Y. (2016) REPdenovo: inferring *de novo* repeat motifs from short sequence reads. *PLoS ONE*, **11**, e0150719.

Daron, J. and Slotkin, R.K. (2017) EpiTEome: Simultaneous detection of transposable element

insertion sites and their DNA methylation levels. *Genome biology*, **18**, 91.

Day, D.S., Luquette, L.J., Park, P.J. and Kharchenko, P.V. (2010) Estimating enrichment of repetitive elements from high-throughput sequence data. *Genome biology*, **11**, R69.

DeBarry, J.D., Liu, R. and Bennetzen, J.L. (2008) Discovery and assembly of repeat family pseudomolecules from sparse genomic sequence data using the Assisted Automated Assembler of Repeat Families (AAARF) algorithm. *BMC Bioinformatics*, **9**, 235.

Debladis, E., Llauro, C., Carpentier, M.C., Mirouze, M. and Panaud O. (2017) Detection of active transposable elements in *Arabidopsis thaliana* using Oxford Nanopore Sequencing technology. *BMC genomics*, **18**, 537.

Delgrange, O. and Rivals, E. 2004. STAR: an algorithm to Search for Tandem Approximate Repeats. *Bioinformatics*, **20**, 2812–2820.

Disdero, E. and Filée, J. (2017) LoRTE: Detecting transposon-induced genomic variants using low coverage PacBio long read sequences. *Mobile DNA*, **8**, 5.

Dowsett, A. and Young, M.W. 1982. Differing levels of dispersed repetitive DNA among closely related species of Drosophila. *Proceedings of the National Academy of Sciences of the United States of America*, **79**, 4570–4574.

Dumbovic, G., Forcales, S.-V. and Perucho, M. (2017) Emerging roles of macrosatellite repeats in genome organization and disease development. *Epigenetics*, **12**, 515–526.

Edgar, R.C. and Myers, E.W. 2005. PILER: identification and classification of genomic repeats. *Bioinformatics*, **21**(Suppl 1), 152–158.

Eichten, S.R., Ellis, N.A., Makarevitch, I., *et al*. (2012) Spreading of heterochromatin is limited to specific families of maize retrotransposons. *PLoS Genetics*, **8**, e1003127.

Ellinghaus, D., Kurtz, S. and Willhoeft, U. (2008) LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC bioinformatics*, **9**, 18.

Elliott, T.A. and Gregory, T.R. (2015) What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **370**, 20140331.

Estécio, M.R., Gallegos, J., Dekmezian, M., *et al*. (2012) SINE retrotransposons cause epigenetic reprogramming of adjacent gene promoters. *Molecular Cancer Research*, **10**, 1332–1342.

Ewing, A.D. (2015) Transposable element detection from whole genome sequence data. *Mobile DNA*, **6**, 24.

Feschotte, C., Keswani, U., Ranganathan, N., Guibotsy, M.L. and Levine, D (2010) Exploring repetitive DNA landscapes using REPCLASS, a tool that automates the classification of transposable elements in eukaryotic genomes. *Genome Biology and Evolution*, **1**, 205–220.

Fiston-Lavier, A.S., Barrón, M.G., Petrov, D.A. and González, J (2015) T-lex2: Genotyping, frequency estimation and re-annotation of transposable elements using single or pooled next-generation sequencing data. *Nucleic Acids Research*, **43**, e22.

Flutre, T. Duprat, E, Feuillet, C, Quesneville, H (2011) Considering transposable element diversification in *de novo* annotation approaches. *PLoS ONE*, **6,** e16526.

Gardner, E.J., Lam, V.K., Harris, D.N., *et al*. (2017) The Mobile Element Locator Tool (MELT): Population-scale mobile element discovery and biology. *Genome Research*, gr.218032.116.

Ge, R., Mai, G., Zhang, R., *et al*. (2017) MUSTv2: An improved de novo detection program for recently active Miniature Inverted repeat Transposable Elements (MITEs). *Journal of Integrative Bioinformatics*, **14**.

Gelfand, Y., Hernandez, Y.2, Loving, J., Benson, G (2014) VNTRseek—a computational tool to detect tandem repeat variants in high-throughput sequencing data. *Nucleic Acids Research*, **42**, 8884–8894.

Gilly, A., Etcheverry, M., Madoui, M.A. *et al*. (2014) TE-Tracker: systematic identification of transposition events through whole-genome resequencing. *BMC Bioinformatics*, **15**, 377.

Goubert, C., Modolo, L., Vieira, C. *et al*., (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 Biology and Evolution*, **7**, 1192–1205.

Grabherr, M.G., Haas, B.J., Yassour, M. et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology*, **29**, 644–652.

Han, B.W.,Wang, W., Zamore, P.D. and Weng, Z (2015) piPipes: a set of pipelines for piRNA and transposon analysis via small RNA-seq, RNA-seq, degradome- and CAGE-seq, ChIP-seq and genomic DNA sequencing. *Bioinformatics*, **31**, 593–595.

Han, Y. and Wessler, S.R. (2010) MITE-Hunter: A program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Research*, **38**, 1–8.

Hancks, D.C. and Kazazian, H.H. (2012) Active human retrotransposons: Variation and disease. *Current Opinion in Genetics and Development*, **22**, 191–203.

Helman, E., Lawrence M.S., Stewart, C. *et al*. (2014) Somatic retrotransposition in human cancer revealed by whole-genome and exome sequencing. *Genome Research*, **24**, 1053–1063.

Hénaff, E., Zapata, L., Casacuberta, J.M. and Ossowski, S. (2015) Jitterbug: somatic and germline transposon insertion detection at single-nucleotide resolution. *BMC Genomics*, **16**, 768.

Hoede, C. Arnoux, S., Moisset, M. *et al*. (2014) PASTEC: an automatic transposable element classification tool. *PLoS ONE*, **9**, e91929.

Hoen, D.R., Hickey, G., Bourque, G. *et al.* (2015) A call for benchmarking transposable element annotation methods. *Mobile DNA*, **6**, 13.

Jiang, C., Chen, C., Huang, Z., Liu, R. and Verdier, J (2015) ITIS, a bioinformatics tool for accurate identification of transposon insertion sites using next-generation sequencing data. *BMC Bioinformatics*, **16**, 72.

Kapitonov, V.V and Jurka, J. (2008) A universal classification of eukaryotic transposable elements implemented in Repbase. *Nature reviews Genetics*, **9**, 411–412

Keane, T.M., Wong, K. and Adams, D.J. (2013) RetroSeq: Transposable element discovery from next-generation sequencing data. *Bioinformatics*, **29**, 389–390.

Kidwell, M.G. and Lisch, D.R. (2000) Transposable elements and host genome evolution. *Trends in*

Ecology and Evolution, **15**, 95–99.

Kim, J.M., Vanguri, S., Boeke, J.D., Gabriel, A. and Voytas, D.F. (1998) Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome sequence. *Genome research*, **8**, 464–78.

Koch, P., Platzer, M. and Downie, B.R. (2014) RepARK - *de novo* creation of repeat libraries from whole-genome NGS reads. *Nucleic Acids Research*, **42**, 1–12.

Kofler, R., Gómez-Sánchez, D. and Schlötterer, C. (2016) PoPoolationTE2: comparative population genomics of transposable elements using Pool-Seq. *Molecular Biology and Evolution*, **33**, 2759–2764.

Kolpakov, R., Bana, G. and Kucherov, G. (2003) mreps: Efficient and flexible detection of tandem repeats in DNA. *Nucleic acids research*, **31**, 3672–8.

Kroon, M., Lameijer, E.W., Lakenberg, N. *et al*. (2016) Detecting dispersed duplications in high-throughput sequencing data using a database-free approach. *Bioinformatics*, **32**, 505–510.

Lander, E.S. *et al*. (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Lee, E., Iskow, R., Yang, L. *et al*. (2012) Landscape of somatic retrotransposition in human cancers. *Science*, **337**, 967–971.

Lerat, E. (2010) Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. *Heredity*, **104**, 520–533.

Lerat, E., Fablet, M., Modolo, L., Lopez-Maestre, H. and Vieira, C. (2017) TEtools facilitates big data expression analysis of transposable elements and reveals an antagonism between their activity and that of piRNA genes. *Nucleic Acids Research*, **45**,e17.

Li, R., Ye, J., Li, S. *et al*. (2005) ReAS: Recovery of ancestral sequences for transposable elements from the unassembled reads of a whole genome shotgun. *PLoS Computational Biology*, **1**, 313–321.

Lien, S., Koop, B.F., Sandve, S.R. *et al*. (2016) The Atlantic salmon genome provides insights into rediploidization. *Nature*, **533**, 200–205.

Lim, K.G., Kwoh, C.K., Hsu, L.Y. and Wirawan, A. (2013) Review of tandem repeat search tools: a systematic approach to evaluating algorithmic performance. *Briefings in Bioinformatics*, **14**, 67–81.

Linheiro, R.S. and Bergman, C.M. (2012) Whole genome resequencing reveals natural target site preferences of transposable elements in Drosophila melanogaster. *PLoS ONE*, **7**, e30008.

Macas, J., Neumann, P. and Navrátilová, A. (2007) Repetitive DNA in the pea (*Pisum sativum L.*) genome: comprehensive characterization using 454 sequencing and comparison to soybean and *Medicago truncatula*. *BMC Genomics*, **8**, 427.

Mardis, E.R. (2017) DNA sequencing technologies: 2006-2016. *Nature protocols*, **12**, 213–218.

Margulies, M., Egholm, M., Altman, W.E. *et al*. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.

McCarthy, E.M. and McDonald, J.F. (2003) LTR_STRUC: a novel search and identification program for LTR retrotransposons. *Bioinformatics*, **19**, 362–367.

McClintock, B. (1950) The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences of the United States of America*, **36**, 344–355.

McCoy, R.C., Taylor, R.W., Blauwkamp, T.A. et al. (2014) Illumina TruSeq Synthetic Long-Reads empower de novo assembly and resolve complex, highly-repetitive transposable elements. *PLoS ONE*, **9**, e106689.

Miki, Y., Nishisho, I., Horii, A. et al. (1992) Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer. *Cancer Research*, **52**, 643–645.

Modolo, L. and Lerat, E. (2013) Identification and analysis of transposable elements in genomic sequences. In Poptsova, M. (ed) *Genome analysis: current procedures and applications*. pp165-181. Place: Caister academic press.

Newman, A.M. and Cooper, J.B. (2007) XSTREAM: A practical algorithm for identification and architecture modeling of tandem repeats in protein sequences. *BMC Bioinformatics*, **8**, 382.

Novák, P., Neumann, P., Pech, J., Steinhais, J. and Macas, J. (2013) RepeatExplorer: A Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics*, **29**, 792–793.

Novák, P., Ávila Robledillo, L., Koblížková, A. *et al*. (2017) TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads. *Nucleic acids research*, **45**, e111.

Novák, P., Neumann, P. and Macas, J. (2010) Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC bioinformatics*, **11**, 378.

Pendleton, M., Sebra, R., Pang, A.W. *et al*. (2015) Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nature Methods*, **12**, 780–786.

Petrov, D.A., Fiston-Lavier, A.S., Lipatov, M., Lenkov, K., González, J. (2011) Population Genomics of Transposable Elements in *Drosophila melanogaster*. *Molecular biology and evolution*, **28**, 1633–1644.

Platt, R.N., Blanco-Berdugo, L. and Ray, D.A. (2016) Accurate transposable element annotation is vital when analyzing new genome assemblies. *Genome Biology and Evolution*, **8**, 403–410.

Platzer, A., Nizhynska, V. and Long, Q. (2012) TE-Locate: A tool to locate and group transposable element occurrences using paired-end next-generation sequencing data. *Biology*, **1**, 395–410.

Price, A.L., Jones, N.C. and Pevzner, A. (2005) *De novo* identification of repeat families in large genomes. *Bioinformatics*, **21**(SUPPL. 1), 351–358.

Rahman, R., Chirn, G.W., Kanodia A. et al. (2015) Unique transposon landscapes are pervasive across *Drosophila melanogaster* genomes. *Nucleic acids research*, **43**, 10655–10672.

Rho, M., Choi, J.H., Kim, S., Lynch, M. and Tang, H. (2007) *De novo* identification of LTR retrotransposons in eukaryotic genomes. *BMC genomics*, **8**, 90.

Richard, G.-F., Kerrest, A. and Dujon, B. (2008) Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. *Microbiology and molecular biology reviews*, **72**, 686–727.

Rishishwar, L., Mariño-Ramírez, L. and Jordan, I.K. (2016) Benchmarking computational tools for polymorphic transposable element detection. *Briefings in Bioinformatics*, bbw072.

Robb, S.M.C., Lu, L., Valencia, E. et al. (2013) The use of RelocaTE and unassembled short reads to produce high-resolution snapshots of transposable element generated diversity in rice. *G3*, **3**, 949–957.

Rostant, W.G., Wedell, N. and Hosken, D.J. (2012) Transposable elements and insecticide resistance. *Advances in genetics*, **78**, 169–201.

Saha, S., Bridges, S., Magbanua, Z.V. and Peterson, D.G. (2008) Computational approaches and tools used in identification of dispersed repetitive DNA sequences. *Tropical Plant Biology*, **1**, 85–96.

Santiago, N., Herráiz, C., Goñi, J.R., Messeguer, X. and Casacuberta, J.M. (2002) Genome-wide analysis of the Emigrant family of MITEs of *Arabidopsis thaliana*. *Molecular biology and evolution*, **19**, 2285–2293.

Schnable, S., Ware, D., Fulton, R.S., *et al*. (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science*, **326**, 1112–1115.

Segal, Y., Peissel, B., Renieri, A. *et al*. (1999) LINE-1 elements at the sites of molecular rearrangements in Alport syndrome-diffuse leiomyomatosis. *American journal of human genetics*, **64**, 62–69.

Sharp, A.J., Locke, D.P., McGrath, S.D. *et al*. (2005) Segmental duplications and copy-number variation in the human genome. *American Journal of Human Genetics*. **77**:78-88.

Sokol, D., Benson, G. and Tojeira, J. (2007) Tandem repeats over the edit distance. *Bioinformatics*, **23**, e30–e35.

Staton, S.E. and Burke, J.M. (2015) Transposome: a toolkit for annotation of transposable element families from unassembled sequence reads. *Bioinformatics*, **31**, 1827-1829.

Steinbiss, S., Willhoeft, U., Gremme, G. and Kurtz, S. (2009) Fine-grained annotation and classification of *de novo* predicted LTR retrotransposons. *Nucleic Acids Research*, **37**, 7002–7013.

Szak, S.T., Pickeral, O.K., Makalowski, W. *et al*. (2002) Molecular archeology of L1 insertions in the human genome. *Genome biology*, **3**, research0052.

Thung, D.T., de Ligt, J., Vissers, L.E. *et al*. (2014) Mobster: accurate detection of mobile element insertions in next generation sequencing data. *Genome Biology*, **15**, 488.

Treangen, T.J. and Salzberg, S.L. (2011) Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, **13**, 36-46

Tu, Z., Li, S. and Mao, C. (2004) The changing tails of a novel short interspersed element in *Aedes aegypti*: Genomic evidence for slippage retrotransposition and the relationship between 3' tandem repeats and the poly(dA) tail. *Genetics*, **168**, 2037–2047.

Tubio, J.M.C., Li, Y., Ju, Y.S. *et al*. (2014) Mobile DNA in cancer. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. *Science*, **345**, 1251343.

Waterston, R.H., Lindblad-Toh, K., Birney, E. *et al*. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–62.

Wicker, T., Schlagenhauf, E., Graner, A. *et al.* (2006) 454 sequencing put to the test using the complex genome of barley. *BMC genomics*, **7**, 275.

Wicker, T., Sabot, F., Hua-Van, A. *et al.* (2007) A unified classification system for eukaryotic

transposable elements. *Nature reviews Genetics*, **8**, 973–982.

Wu, J., Lee, W.P., Ward, A. *et al*. (2014) Tangram: a comprehensive toolbox for mobile element insertion detection. *BMC Genomics*, **15**, 795.

Xu, Z. and Wang, H. (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research*, **35**(Web Server), W265–W268.

Zhuang, J., Wang, J., Theurkauf, W. and Weng, Z. (2014) TEMP: A computational method for analyzing transposable element polymorphism in populations. *Nucleic Acids Research*, **42**, 6826–6838.

Zytnicki, M., Akhunov, E. and Quesneville, H. (2014) Tedna: a transposable element *de novo* assembler. *Bioinformatics*, **30**:2656-2658.

Author Biography and Photograph

Emmanuelle Lerat is a CNRS researcher since 2005 working in the Laboratory “Biométrie et Biologie Evolutive” at the University Lyon 1, France. Her major research interest concerns the evolution of genomes in the light of their repeat content and more particularly the transposable elements (TEs). She has a strong background in molecular biology, in evolution, and in bioinformatics, with a long history in the annotation and analysis of TEs in various eukaryotic genomes, especially in Drosophila. She currently coordinates different subjects allowing to link informatic to biological questions concerning the functional impact of TEs on genome and gene evolution. She is a member of the Editorial Board of the international SMBE journal “Genome Biology and Evolution” since 2008.

Table 1: non-exhaustive list of tools to identify and analyze repeats

Program Name	Type of repeats	Input data	Approach	References	Web site
Tandem Repeat Finder	Tandem Repeat	Assembled genome	Identification	Benson 1999	https://tandem.bu.edu/trf/trf.html
XSTREAM	Tandem Repeat	Assembled genome	Identification	Newman and Cooper 2007	http://jimcooperlab.mcdb.ucsb.edu/xstream/
MREPS	Tandem Repeat	Assembled genome	Identification	Kolpakov et al. 2003	http://mreps.univ-mlv.fr/
STAR	Tandem Repeat	Assembled genome	Identification	Delgrange and Rivals 2004	http://www.atgc-montpellier.fr/star/
TRED	Tandem Repeat	Assembled genome	Identification	Sokol et al. 2007	http://tandem.sci.brooklyn.cuny.edu/tandem/
ReD Tandem	Tandem Repeat	Assembled genome	Identification	Audemard et al. 2012	NA
RepeatMasker	TEs, tandem repeat	Assembled genome	Identification	Smit et al.	http://www.repeatmasker.org/
LTR_STRUC	LTR-retrotransposons	Assembled genome	Identification	McCarthy and McDonald 2003	http://www.mcdonaldlab.biology.gatech.edu/ltr_struc.htm
LTR_FINDER	LTR-retrotransposons	Assembled genome	Identification	Xu and Wang 2007	http://tlife.fudan.edu.cn/ltr_finder/
find_LTR	LTR-retrotransposons	Assembled genome	Identification	Rho et al. 2007	http://darwin.informatics.indiana.edu/cgi-bin/evolution/ltr.pl
LTR_harvest	LTR-retrotransposons	Assembled genome	Identification	Ellinghaus et al. 2008	http://www.zbh.uni-hamburg.de/forschung/arbeitsgruppe-genominformatik/software/ltrharvest.html
TSDfinder	Non-LTR retrotransposons	Assembled genome	Identification	Szak et al. 2002	https://www.ncbi.nlm.nih.gov/CBBresearch/Landsman/TSDfinder/
SINEDR	Non-LTR retrotransposons	Assembled genome	Identification	Tu et al. 2004	NA
TRANSPO	MITEs	Assembled genome	Identification	Santiago et al. 2002	NA
MUST / MUSTv2	MITEs	Assembled genome	Identification	Chen et al. 2009; Ge et al. 2017	http://www.healthinformaticslab.org/supp/resources.php
MITE-hunter	MITEs	Assembled genome	Identification	Han and Wessler 2010	http://target.iplantcollaborative.org/
One code to find them all	TEs, tandem repeat	Output files from RepeatMasker	Analysis	Bailly-Béchet et al. 2014	http://doua.prabi.fr/software/one-code-to-find-them-all

LTRdigest	LTR-retrotransposons	GFF3 format	Analysis	Steinbiss et al. 2009	http://www.zbh.uni-hamburg.de/forschung/arbeitsgruppe-genominformatik/software/ltrdigest.html
RECON	TEs, tandem repeat	Assembled genome	Identification	Bao and Eddy 2003	http://eddylab.org/software/recon/
PILER	TEs, tandem repeat	Assembled genome	Identification	Edgar and Myers 2005	https://www.drive5.com/piler/
ReAS	TEs, tandem repeat	Assembled genome	Identification	Li et al. 2005	Freely available via ReAS@genomics.org.cn
RepeatScout	TEs, tandem repeat	Assembled genome	Identification	Price et al. 2005	https://bix.ucsd.edu/repeatscout/
TEclass	TEs	Individual sequences	Classification	Abrusan et al. 2009	http://www.mybiosoftware.com/teclass-2-1-classification-te-consensus-sequences.html
REPCLASS	TEs	Individual sequences	Classification	Feschotte et al. 2010	https://github.com/feschottelab/REPCLASS
PASTEC	TEs	Individual sequences	Classification	Hoede et al. 2014	https://urgi.versailles.inra.fr/Tools/PASTEClassifier
REPET	TEs, tandem repeat	Assembled genome	Pipeline (Identification and classification)	Flutre et al. 2011	https://urgi.versailles.inra.fr/Tools/REPET
RepeatModeler	TEs, tandem repeat	Assembled genome	Pipeline (Identification and classification)	Smit and Hubley, unpublished	http://www.repeatmasker.org/RepeatModeler/
AAARF	TEs, tandem repeat	Short reads	Identification	DeBarry et al. 2008	https://sourceforge.net/projects/aaarf
Tedna	TEs, tandem repeat	Short reads	Identification	Zytnicki et al. 2014	https://urgi.versailles.inra.fr/Tools/Tedna
RepeatExplorer (SeqGraphR)	TEs, tandem repeat	Short reads	Identification	Novak et al. 2010; Novak et al. 2013	http://www.repeatexplorer.org/ http://w3lamc.umbr.cas.cz/lamc/?page_id=301
DNApipeTE	TEs, tandem repeat	Short reads	Identification	Goubert et al. 2015	https://lbbe.univ-lyon1.fr/-dnaPipeTE-.html
RepARK	TEs, tandem repeat	Short reads	Identification	Koch et al. 2014	https://github.com/PhKoch/RepARK
REPdenovo	TEs, tandem repeat	Short reads	Identification	Chu et al. 2016	https://github.com/Reedwarbler/REPdenovo
Transposome	TEs, tandem repeat	Short reads	Identification	Staton and Burke 2015	https://github.com/sestaton/Transposome
VNTRseek	Tandem repeat	Short reads	Identification	Gelfand et al. 2014	https://github.com/yzhernand/VNTRseek
TAREAN	Tandem repeat	Short reads	Identification	Novak et al. 2017	http://w3lamc.umbr.cas.cz/lamc/?page_id=312

TE-locate	TEs	Short reads	TE insertion variants	Platzer et al. 2012	https://sourceforge.net/projects/te-locate/
TraFiC	TEs	Short reads	TE insertion variants	Tubio et al. 2014	https://gitlab.com/mobilegenomes/TraFiC
TE-Tracker	TEs	Short reads	TE insertion variants	Gilly et al. 2014	http://www.genoscope.cns.fr/externe/tetracker/
RelocaTE	TEs	Short reads	TE insertion variants	Robb et al. 2013	https://github.com/srobb1/RelocaTE
Ngs-TE-mapper	TEs	Short reads	TE insertion variants	Linheiro and Bergman 2012	https://github.com/bergmanlab/ngs_te_mapper
TIDAL	TEs	Short reads	TE insertion variants	Rahman et al. 2015	https://github.com/laulabbrandeis/TIDAL
ITIS	TEs	Short reads	TE insertion variants	Jiang et al. 2015	http://bioinformatics.psc.ac.cn/software/ITIS/
T-Lex2	TEs	Short reads	TE insertion variants	Fiston-Lavier et al. 2015	http://petrov.stanford.edu/cgi-bin/Tlex.html
Tea	TEs	Short reads	TE insertion variants	Lee et al. 2012	http://compbio.med.harvard.edu/Tea/
RetroSeq	TEs	Short reads	TE insertion variants	Keane et al. 2013	https://github.com/wtsi-svi/RetroSeq
TEMP	TEs	Short reads	TE insertion variants	Zhuang et al. 2014	https://github.com/JialiUMassWengLab/TEMP
Mobster	TEs	Short reads	TE insertion variants	Thung and et al. 2014	https://sourceforge.net/projects/mobster/
Tangram	TEs	Short reads	TE insertion variants	Wu et al. 2014	https://github.com/jiantao/Tangram/issues
TranspoSeq	TEs	Short reads	TE insertion variants	Helman et al. 2014	http://archive.broadinstitute.org/cancer/cga/transposeq
Jitterbug	TEs	Short reads	TE insertion variants	Hénaff et al. 2015	https://github.com/elzbth/jitterbug
DD-Detection	TEs	Short reads	TE insertion variants	Kroon et al. 2016	https://bitbucket.org/mkroon/dd_detection
popoolation_TE2	TEs	Short reads	TE insertion variants	Kofler et al. 2016	https://sourceforge.net/p/popoolation-te2/wiki/Home/
MELT	TEs	Short reads	TE insertion variants	Gardner et al. 2017	http://melt.igs.umaryland.edu/manual.php
LorTE	TEs	Long reads	TE insertion variants	Disdero and Filée 2017	http://www.egce.cnrs-gif.fr/?p=6422
Repeat Histone enrichment	TEs	ChipSeq data	Histone enrichment	Day et al. 2010	http://compbio.med.harvard.edu/repeats/
piPipes	TEs	RNAseq data	Differential expression analyses	Han et al. 2015	https://github.com/bowhan/piPipes
TEtools	TEs	RNAseq data	Differential expression	Lerat et al. 2017	https://github.com/l-modolo/TEtools

			analyses		
EpiTEome	TEs	Short reads	TE insertion variants and DNA methylation analysis	Daron and Slotkin 2017	https://github.com/jdaron/epiTEome