

# Chapter 2

## Transposable Element Annotation in Completely Sequenced Eukaryote Genomes

Timothée Flutre, Emmanuelle Permal, and Hadi Quesneville

**Abstract** With the development of new sequencing techniques, the number of sequenced plant genomes is increasing. However, accurate annotation of these sequences remains a major challenge, in particular with regard to transposable elements (TEs). The aim of this chapter is to provide a roadmap for researchers involved in genome projects to address this issue. We list several widely used tools for each step of the TE annotation process, from the identification of TE families to the annotation of TE copies. We assess the complementarities of these tools and suggest that combined approaches, using both *de novo* and knowledge-based TE detection methods, are likely to produce reasonably comprehensive and sensitive results. Nevertheless, existing approaches still need to be supplemented by expert manual curation. Hence, we describe good practice required for manual curation of TE consensus sequences.

**Keywords** Annotation • Bioinformatics • Classification • Curation • Identification • Pipeline

### 2.1 Introduction

Transposable elements (TEs) are mobile genetic elements that shape the eukaryotic genomes in which they are present. They are virtually ubiquitous and make up, for instance, 20% of a typical *D. melanogaster* genome (Bergman et al. 2006), 50% of a *H. sapiens* genome (Lander et al. 2001), and 85% of a *Z. mays* genome (Schnable et al. 2009). They are classified into two classes depending on their transposition mode: via RNA for class I retrotransposons and via DNA for class II transposons

---

T. Flutre • E. Permal • H. Quesneville (✉)  
INRA, UR 1164, URGI, Unité de Recherche en Génomique-Info,  
78026 Versailles cedex, France  
e-mail: [hadi.quesneville@versailles.inra.fr](mailto:hadi.quesneville@versailles.inra.fr)

(Finnegan 1989). Each class is also subdivided into several orders, superfamilies, and families (Wicker et al. 2007). Due to their unique ability to transpose and because they frequently amplify, TEs are major determinants of genome size (Petrov 2001; Piegu et al. 2006) and cause genome rearrangements (Gray 2000; Fiston-Lavier et al. 2007). Once described as the “ultimate parasites” (Orgel and Crick 1980), TEs are commonly found to regulate the expression of neighboring genes (Feschotte 2008; Bourque 2009) or even to have been domesticated so as to provide a specific host function (Zhou et al. 2004; Bundock and Hooykaas 2005; Santangelo et al. 2007; Kapitonov and Jurka 2005).

As a consequence of the development of new rapid sequencing techniques, the number of available sequenced eukaryotic genomes is constantly increasing. However, the first step of the analysis, i.e., accurate annotation, remains a major challenge, particularly concerning TEs. Correct genome annotation of genes and TEs is an indispensable part of thorough genome-wide studies. Consequently, efficient computational methods have been proposed for TE annotation (Bergman and Quesneville 2007; Lerat 2010; Janicki et al. 2011). Given that the pace at which genomes are sequenced is unlikely to decrease in the coming years; the process of TE annotation needs to be made widely accessible.

This chapter lays down a clear road map detailing the order in which computational tools (or combinations of such tools) should be used to annotate TEs in a whole genome. We distinguish three steps (1) identifying TEs by searching for reference sequences (e.g., full-length TE sequences) and building consensus from similar sequences, (2) manual curation to define and classify TE families, and (3) annotation of every TE copy. We also provide some hints on manual curation, a step that is still necessary.

## 2.2 *De Novo* Detection of Transposable Elements

Various efficient computational methods are available to identify unknown TEs in genomic sequences. Each method is based on specific assumptions that have to be understood to optimize selection and combination of the methods to ensure they are appropriate for any particular analytic goal.

### 2.2.1 *Computing Highly-Repeated Words*

TEs, due to their capacity to transpose, are often present in a large number of copies within the same genome. Although TE sequences degenerate with time, words (i.e., short subsequences of few nucleotides) that compose them are consequently repeated throughout the genome. Software, such as the TALLYMER (Kurtz et al. 2008) and P-CLOUDS (Gu et al. 2008), has been designed to find repeats rapidly in genome sequences by counting highly frequent words of a given length  $k$ , called  $k$ -mers. These programs are very useful for quickly providing a view of the repeated

fraction in a given set of genomic sequences, including especially unassembled sequences. However, they do not provide much detail about the TEs present in these sequences. Their output only identifies highly repeated regions without indicating precise TE fragment boundaries or TE family assignments. These methods are quick and simple to use but allow only limited biological interpretations and no real TE annotation.

Other methods also start by counting frequent  $k$ -mers but then go on to try to define consensus. ReAS (Li et al. 2005) applies this approach directly to shotgun reads. For each frequent  $k$ -mer, a multiple alignment of all short reads containing it is built and then extended iteratively. REPEATSCOUT (Price et al. 2005) has a similar approach but works on assembled sequences. These tools return a library of consensus sequences. Although their results are more biologically relevant than those of previous methods, the consensus are usually too short and correspond to truncated versions of ancestral TEs (Flutre et al. 2011). Substantial manual inspection and editing is therefore needed to obtain a meaningful list of consensus sequences.

### 2.2.2 *All-by-All Alignment and Clustering of Interspersed Repeats*

Repeats can also be identified by self-alignment of genomic sequences, starting with an all-by-all alignment of the assembled sequences.

Several tools can be used for this. Some, such as BLAST (Altschul et al. 1997) and BLAST-like algorithms, use heuristics. For instance, BLASTER (Quesneville et al. 2003) performs this search by launching BLAST repeatedly over the genome sequences. Others are exact algorithms. Hence, PALS uses “q-Gram filters” that unlike a heuristic (e.g., BLAST), it rapidly and stringently eliminates a large part of the search space from consideration before the alignment search but nevertheless guarantees not to eliminate a region containing a match (Rasmussen et al. 2005). As the amount of input data is usually large, the computations are intensive. Consequently, stringent parameters are applied: good results are obtained with BLAST-like tools when matches shorter than 100 bp or with identity below 90% or with an E-value above  $1e-300$  are dismissed (Flutre et al. 2011). As most TEs are shorter than 25 kb, segmental duplications can also be filtered out by removing longer matches. To speed up the computations, such alignment tools can be launched in parallel on a computer cluster.

With these parameters, only closely related TE copies will be found. Note that the aim of this step is not to recover all TE copies of a family but to use those that are well conserved to build a robust consensus (see below). Stringent alignment parameters are crucial for successful reconstruction of a valid consensus. Interestingly, even with these stringent criteria, this approach is still more sensitive than other methods for identifying repeats. However, it is also the most computer intensive. It also misses single-copy TE families because at least two copies are required for detection by self-alignment.

Once the matches corresponding to repeats have been obtained, they need to be clustered into groups of similar sequences. The aim is for each cluster to correspond to copies of a single TE family. However, TEs may include divergent interspersed repeats, often nested within each other, making the task difficult. Algorithms have been designed to cluster identified sequences appropriately, limiting the artifacts induced by nested and deleted TE copies and non-TE repeats such as segmental duplications. The various tools that are available are based on different assumptions about (1) the sequence diversity within a TE family, (2) the evolutionary dynamics of TE sequences, (3) nested patterns, and (4) repeat numbers.

GROUPE (Quesneville et al. 2003; Flutre et al. 2011) starts by connecting fragments belonging to the same copy by dynamic programming, and then applies a single link clustering algorithm with (1) a 95% coverage constraint between copies of the same cluster and (2) cluster selection based on the number of copies not included in larger copies of other clusters. The rationale here is to detect copies that have the same length as they most probably correspond to mobile entities. Indeed, copies can diverge rapidly by accumulating deletions leading to copies with different sizes. Copies that are almost intact can transpose conserving their original, presumably functional, size. RECON (Bao and Eddy 2002) also starts with a single link-clustering step. If a cluster includes nested repeats and is thus chimerical, it can be subdivided according to the distribution of its all-by-all genome alignment ends. Indeed, nested repeats exhibit a specific pattern in alignments of sequences obtained in an all-by-all genome comparison: the alignment ends of any one inner repeat are all in the relative same position.

PILER-DF (Edgar and Myers 2005) identifies lists of matches covering a maximal contiguous region, defines them as piles, and then builds clusters of globally alignable piles. The rationale here is identical to that used by GROUPE where copies of identical length are sought; however, PILER-DF has no specific attitude to indels.

The three clustering programs behave differently according to the sequence diversity of TE families. For instance, GROUPE better distinguishes groups of mobile elements differing by their sizes inside a TE family. It also better recovers fragmented copies due to its dynamic programming joining algorithm. But, it produces more redundant results and only correctly recovers TE families if there are at least three complete copies. RECON is better for TE families with fewer than three complete copies, being able to reconstruct the complete TE from fragments. PILER is fast and very specific. It is a useful option for large genomes when time is an issue, or if a non-exhaustive search is sufficient.

Once clusters are defined, a filter is usually applied to retain only those with at least three members, thereby eliminating the vast majority of segmental duplications. Finally, for each remaining cluster, a multiple alignment is built from which a consensus sequence is derived. Numerous algorithms are available for this but only those complying with the following criteria should be used (1) speed, because the number of clusters is usually very large and (2) ability to handle appropriately sequences of different lengths, which is the case for the clusters generated by RECON. MAP (Huang 1994) and MAFFT (Katoh et al. 2002) comply

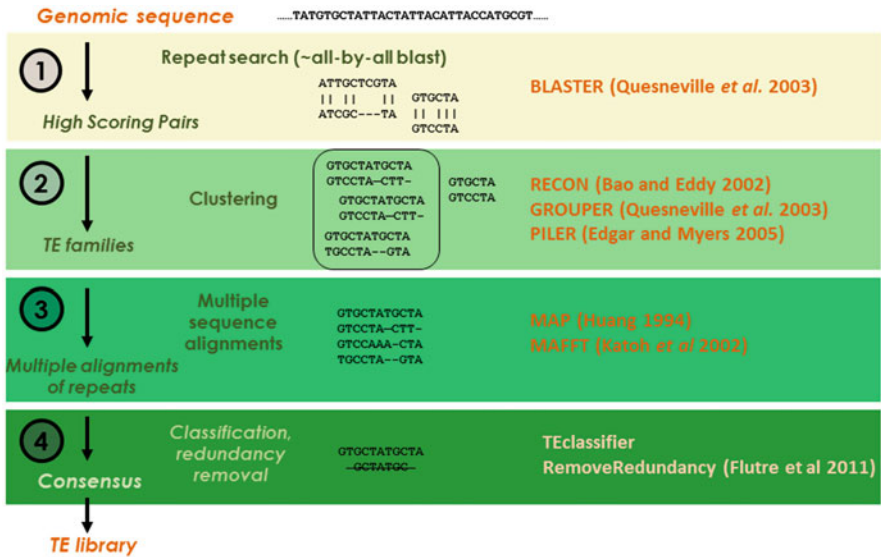


Fig. 2.1 Workflow of the 4-step *de novo* TE detection pipeline (Flutre et al. 2011)

with these criteria and give good results (Flutre et al. 2011). Taking the 20 longest sequences is generally sufficient to build the consensus. The set of consensus sequences obtained represents a condensed view of all TE families present in the genome being studied.

For easy identification of TE families, i.e., those for which there are full-length copies that are very similar to each other, all clustering methods will find roughly the same consensus. However, for other families, which may be numerous, different methods generate different clusters, because they rely on different assumptions. Therefore, manual curation is required to identify an appropriate set of representative sequences (see below).

This all-by-all genome comparison strategy has been implemented in a pipeline called TEdenovo (Fig. 2.1). The TEdenovo pipeline is part of the REPET package (Flutre et al. 2011) and was designed to be used on a computer cluster for fast calculations. It allows the use of different software at each step to exploit the best strategy according to the genome size and the TE identification goal.

### 2.2.3 Features-Based Methods

Alternatively, TEs can be detected using prior knowledge about TE features. For example, class I LTR retrotransposons characteristically have LTR at both ends of the element, and this can be used for their detection. Numerous class II TEs encompass TIR structures that can be used as markers. Many TE families generate

a double-strand break when they insert into the DNA sequence. The break is caused by the enzymatic machinery of the TE that generally cuts the DNA with a shift between the two DNA strands. After the insertion, DNA repair processes generate a short repeat of few nucleotides (up to 11) at each end; these repeats are called Target Site Duplications (TSDs) and are characteristic of particular TE families.

There are many different types of TEs and several tools to detect them are available (Table 2.1). Most of these tools have been described in detail in various reviews (Bergman and Quesneville 2007; Lerat 2010; Janicki et al. 2011). Here, we will address the general principles behind their design.

As class I LTR retrotransposons are easily characterized on the basis of their LTRs and are abundant in genomes, there have been substantial efforts to design bioinformatics tools for their detection. Some of these tools also use the characteristics of some of the substructures of the LTR retrotransposons. The programs available are: LTR\_STRUC (McCarthy and McDonald 2003), LTR\_MINER (Pereira 2004), SmaRTFinder (Morgante et al. 2005b), LTR\_FINDER (Xu and Wang 2007), LTR\_par (Kalyanaraman and Aluru 2006), find\_LTR (Rho et al. 2007), which is now called MGEScanLTR, LTRharvest (Ellinghaus et al. 2008), and LTRdigest (Steinbiss et al. 2009) that also identifies protein-coding regions within the LTR element. The algorithms of these tools are generally divided into two parts: they first build a data structure to speed up searches for repeats, and then use this structure to search for repeats in the genomic sequences. For example, LTRharvest builds suffix-array using the “suffixerator” tool from GenomeTools package (Lee and Chen 2002). Some of these tools add a third step to refine the search by looking for additional substructures, such as Primer Binding Sites (PBS) and Poly-Purine Tracks (PPT) that are important signals for LTR retrotransposon transposition. These programs also allow searching for TSD and coding regions, including those encoding protein domains, specific to these TEs.

There are also tools aimed at detecting class I non-LTR retrotransposons, e.g., *Long Interspersed Nuclear Elements (LINE)* and *Short Interspersed Nuclear Elements (SINE)*. TSDfinder (Szak et al. 2002) is based on the L1 TE insertion signature which is constituted in part by two Target Site Duplications (TSDs) and a polyA tail. RTAnalyzer (Lucier et al. 2007) is a Web server that follows the same approach as TSDfinder. SINEDR (Tu et al. 2004) is designed to look for SINE elements, a group of non-LTR retrotransposons, in sequence databases. MGEScan-non-LTR (Rho and Tang 2009) identifies and classifies non-LTR TEs in genomic sequences using probabilistic models. It is based on the structure of the 12 TE clades that are non-LTR TEs. It uses two separate Hidden Markov Model (HMM) profiles, one for the Reverse Transcriptase (RT) gene and one for the endonuclease (APE) gene, both of which are well conserved among non-LTR TEs.

Class II TEs, but not Helitrons and Cryptons, are structurally characterized by TIRs. Some class II-specific bioinformatics tools, for example, FindMite (Tu 2001), Transpo (Santiago et al. 2002), and MAK (Yang and Hall 2003), search for defined TIR features in sequences. Must (Chen et al. 2009) is designed to search for TEs containing two TIRs and two direct repeats (i.e., TSD) to identify MITE candidates. Two new tools were published recently: MITE-Hunter (Han and Wessler 2010)

**Table 2.1** Availability of feature-based detection programs for TE *de novo* identification

TEclass	Program	URL	Web server or program
I LTR	LTR_STRUC	<a href="http://www.mcdonaldlab.biology.gatech.edu/finalLTR.htm">http://www.mcdonaldlab.biology.gatech.edu/finalLTR.htm</a>	Software available upon request
	LTR_MINER	<a href="http://genomebiology.com/2004/5/10/R79/additional">http://genomebiology.com/2004/5/10/R79/additional</a>	Script available as additional file
	SmartFinder	<a href="http://services.appliedgenomics.org/software/smartfinder/">http://services.appliedgenomics.org/software/smartfinder/</a>	Downloadable software
	LTR_FINDER	<a href="http://tlife.fudan.edu.cn/ltr_finder/">http://tlife.fudan.edu.cn/ltr_finder/</a>	Web server and software available upon request
	LTR_par	<a href="http://www.eecs.wsu.edu/~ananth/software.htm">http://www.eecs.wsu.edu/~ananth/software.htm</a>	Software available upon request
	Find_LTR (MGEScanLTR)	<a href="http://darwin.informatics.indiana.edu/cgi-bin/evolution/ltr.pl">http://darwin.informatics.indiana.edu/cgi-bin/evolution/ltr.pl</a>	Downloadable software
	LTRharvest	<a href="http://www.zbh.uni-hamburg.de/forschung/genomformatik/software/ltrharvest.html">http://www.zbh.uni-hamburg.de/forschung/genomformatik/software/ltrharvest.html</a>	Downloadable as part of the genomeTool package ( <a href="http://genometools.org/pub/">http://genometools.org/pub/</a> )
	LTRdigest	<a href="http://www.zbh.uni-hamburg.de/forschung/genomformatik/software/ltrdigest.html">http://www.zbh.uni-hamburg.de/forschung/genomformatik/software/ltrdigest.html</a>	Downloadable as part of the genomeTool package ( <a href="http://genometools.org/pub/">http://genometools.org/pub/</a> )
I Non-LTR	TSDfinder	<a href="http://www.ncbi.nlm.nih.gov/CBBresearch/Landsman/TSDfinder/">http://www.ncbi.nlm.nih.gov/CBBresearch/Landsman/TSDfinder/</a>	Script available as additional file
	RTAanalyzer	<a href="http://www.riboclub.org/cgi-bin/RTAnalyzer/index.pl?page = rt_find">http://www.riboclub.org/cgi-bin/RTAnalyzer/index.pl?page = rt_find</a>	Web server
	SINEDR	Not available	Software available upon request
	MGEScan-non-LTR	<a href="http://darwin.informatics.indiana.edu/cgi-bin/evolution/nonltr/nonltr.pl">http://darwin.informatics.indiana.edu/cgi-bin/evolution/nonltr/nonltr.pl</a>	Downloadable software
II TIR	Transpo	<a href="http://algggen.lsi.upc.es/recerca/search/transpo/transpo.html">http://algggen.lsi.upc.es/recerca/search/transpo/transpo.html</a>	Web server and downloadable software
	FindMITE	No longer available	No longer available
	MAK	No longer available	No longer available
	MUST	<a href="http://csbl1.bmb.uga.edu/ffzhou/MUST/">http://csbl1.bmb.uga.edu/ffzhou/MUST/</a>	Web server
	MITE-Hunter	<a href="http://target.iplantcollaborative.org/mite_hunter.html">http://target.iplantcollaborative.org/mite_hunter.html</a>	Downloadable software
	TS clustering	Not available	Software available upon request
II Helitron	HelSearch	<a href="http://sourceforge.net/projects/helsearch/files/">http://sourceforge.net/projects/helsearch/files/</a>	Downloadable software
	HelitronFinder	<a href="http://limei.montclair.edu/HF.html">http://limei.montclair.edu/HF.html</a>	Web server and software available upon request

Feature-based *de novo* TE identification is generally fast and efficient. Unfortunately, only well-described TEs that also have a strong signature can be found. Some TEs do not have such characteristics and thus cannot be found by this type of approach. Consequently, feature-based *de novo* TE identification cannot be used alone to provide an exhaustive inventory of TEs in a genome. Nevertheless, this approach can be used to supplement the findings of all-by-all genome comparison TE searches, in particular for low copy TE families that are otherwise difficult to detect. Surprisingly, these feature-based tools also suffer from high false-positive detection rates such that careful curation is required (data not shown)

which is a five-step pipeline, with the first step involving a TIR-like structure search and TS clustering (Hikosaka and Kawahara 2010), which is dedicated to finding T2-MITEs.

Despite there being no TIR structures in Helitrons, programs have also been designed for their detection: HelitronFinder (Du et al. 2008) is based on known consensus sequences and HelSearch (Yang and Bennetzen 2009) looks for a Helend structure constituted by a six base-pair hairpin and CTRR nucleotide motif.

### 2.2.4 Evidence for TE Mobility

The identification of a long indel by sequence alignments between two closely related species is suggestive of the presence of a TE. The rest of the genome can then be searched for this sequence to assess its repetitive nature. This approach has been used (Caspi and Pachter 2006) and appears to work well for recent TE insertions: indeed, it will only detect insertions that occurred after speciation. Using several alignments with species diverging at different times may lead to more TEs being identified (Caspi and Pachter 2006), as each alignment allows detection of TEs inserted at different times. However, one limitation is the difficulty of correctly aligning long genomic sequences from increasingly divergent species.

This idea could be also used within a genomic sequence by considering segmental duplications. A long indel apparent in sequence alignments of genomic duplications may similarly be an indication of the presence of a TE (Le et al. 2000). Various controls are needed, however, to confirm the TE status of the sequence. For example, TE features such as terminal repeats (e.g., LTR, TIR) or similarity to other TE sequences could be used. This approach only detects TE insertions that occur after the duplication event and may thus be limited to rare events.

TSDs are hallmarks of a transposition event, but they can be difficult to find in old insertions because they are short, and they can be altered by mutations or deletions. In addition, the size of the TSD depends on the family and not all TEs generate a TSD upon insertion.

## 2.3 Classification and Curation of Transposable Element Sequences

When they amplify, TE copies may nest within each other in complex patterns (Bergman et al. 2006), thereby fragmenting the elements. With time, the sequences accumulate (1) point substitutions, (2) deletions that truncate copies, and (3) insertions that interrupt their sequences (Blumenstiel et al. 2002). These events generate complex remnants of TEs. Various *de novo* tools use these remnants to try to infer the ancestral sequence that actually transposed.



When starting with a self-alignment (i.e., all-by-all genome comparison) of genomic sequences, the optimal strategy is to use several tools and even combine them. However, all the relevant tools and every *de novo* approach can encounter difficulties when trying to distinguish true TEs from segmental duplications, multi-member gene families, tandem repeats, and satellites. It is, therefore, strongly recommended to confirm that the predicted sequences can be classified as being TEs. Computerized analysis therefore still needs to be complemented by manual curation.

### 2.3.1 Classification

Sequences believed to correspond to TEs can be classified according to their similarity to known TEs, for example, those recorded in databases like Repbase Update (Jurka et al. 2005). A tool called TEclass (Abrusan et al. 2009) implements a support vector machine, using oligomer frequencies, to classify TE candidates.

However, for most previously unknown TE sequences obtained via *de novo* approaches from nonmodel organisms, classification requires the specific identification of several TE features [see (Wicker et al. 2007) for complete description]. By searching for structural features, such as terminal repeats, features characteristic of various TE types can be identified: long terminal repeats specific to class I LTR retrotransposons, terminal inverted repeats specific to the class II DNA transposons, and poly-A or SSR-like tails specific to class I non-LTR retrotransposons. In addition, using BLASTN, BLASTX, and TBLASTX to compare TE candidates with a reference data bank, can provide hints for classification, as long as the reference data bank contains elements similar to the TE candidate. Therefore, it is also recommended to search for matches for sequences encoding TE-specific protein profiles in TE sequences. For example, the presence of a transposase gene is strongly indicative of a class II DNA transposon. Such protein profiles can be obtained from the Pfam database which includes protein families represented by multiple sequence alignments and hidden Markov models (HMM) (Finn et al. 2010). These profiles can be used by programs such as HMMER to find matches within the candidate TE sequences.

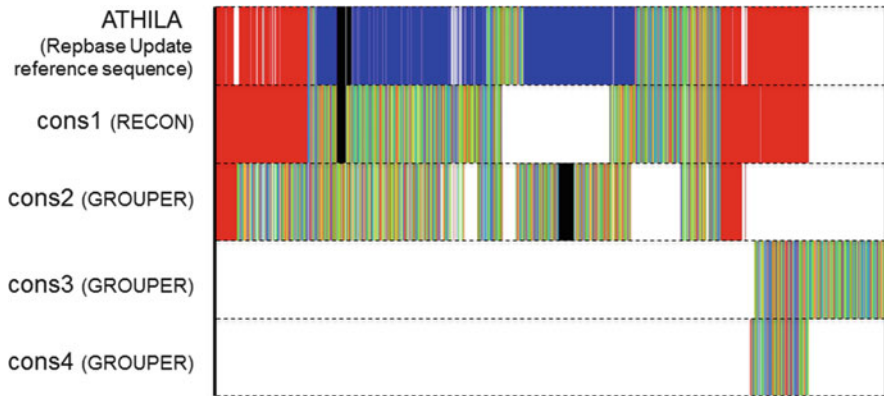
Some tools classify TE sequences according to their features, usually via a decision tree. The TEclassifier in the REPET package (Flutre et al. 2011) and REPCLASS (Feschotte et al. 2009) searches for all the features listed above. In addition, REPCLASS allows TE candidates to be filtered on the basis of the number of copies they have in the genome. TEclassifier interestingly allows the removal of redundancy from among potential TE sequences. It uses the classification to eliminate redundant copies (a sequence contained within a longer one) and retains well-classified TE candidate sequences preferentially over less well-classified TE candidate sequences. This tool is particularly useful for reconciling different TE reference libraries obtained independently, as it guarantees to retain well-classified TE candidate sequences.

### 2.3.2 *Identification of Families*

Once the newly identified TE sequences have been classified, manual curation is required as some consensus sequences may not have been classified previously and there may still be some redundant consensus sequences. Manual curation is crucial because the annotation of TE copies, as described in the next section, depends on the quality of the TE library. One way to curate a library of TE consensus sequences is to gather these sequences into clusters that may constitute TE families. A tool like BLASTCLUST in the NCBI-BLAST suite can quickly build such clusters via simple link clustering based on sequence alignment coverage and identity. Eighty percent identity and coverage, as proposed by (Wicker et al. 2007), gives good results. Typical clusters will contain well-classified consensus (e.g., class I—LTR—Gypsy element) as well as unclassified consensus (without structural features and little sequence similarity either with known TEs or any TE domain).

Then, computing a multiple sequence alignment (MSA) for each cluster gives a useful view of the relationships between the consensus sequences such that it is possible to assess whether they belong to the same TE family. One of the programs detailed above, MAP or MAFFT, can be used. It can also be informative to build a MSA with the consensus and with the genomic sequences from which these consensus were derived and/or the genomic copies that each of these consensus can detect. In such cases, we advise first building a single MSA for each consensus with the genomic sequences it detects, and then building a global MSA by aligning these multiple alignments together, for example, using the “profile” option of the MUSCLE program (Edgar 2004). Finally, after a visual check of the MSA with the evidence used to assign a classification to the consensus, it is then possible to tag all consensus sequences in the same cluster with the most frequent TE class, order, superfamily, and family, if one has been assigned (Fig. 2.2). The MSA can be also edited by splitting it or deleting sequences to obtain a MSA corresponding to a single TE family. Indeed, in some cases, consensus are only similar along a small segment or display substantial sequence divergence. In these cases, the MSA can be split into as many MSA as there are candidate TE families. In other cases, an insertion appears to be specific to one consensus sequences and may sometimes show evidence (e.g., BLAST hits) for a different TE order. This may indicate a chimeric consensus that can be either removed from the library, if artifactual according to the sequences used to build the consensus (also visible in the MSA), or used to build a new TE family (if several copies support it). In all these cases, finding a genomic copy that aligns along almost all the length of a consensus (e.g., 95% coverage) appears to be a reasonable criterion for retaining the consensus. Those that fail generally appear to be artifacts or at least could be considered to be of no value.

Phylogenies of TE family copies and/or consensus sequences provide another view of the members in a TE family. This can serve as an aid to curation if the cluster has many members or if two or more subfamilies are present. In such cases, sub-families can be hard to detect by examination of the MSA alone, but may



**Fig. 2.2** Alignment (Jalview (Clamp et al. 2004) screenshot) of *de novo* TE consensus sequences with Athila, the best-matching known TEs in the Rebase Update. They are represented with some of the features shown: LTRs (*red zones*), ORFs (*blue zone*), and matches with HMM profiles (*black*). The differences between the consensus sequences obtained by different methods, here RECON (cons1) and GROUPER (cons2, cons3, cons4), are indicated. Manual curation would remove cons3 as it corresponds to a single LTR with short sequences not present in the Athila family and cons4 as it corresponds to a LTR probably formed from the Athila solo-LTRs of the genome. A good consensus for the family would be a combination of cons1 and cons2

become evident in a phylogeny if distinct sub-trees emerge. Such phylogenies can be constructed from the MSA with currently available software, including the PhyML program (Guindon and Gascuel 2003). Note, however, as most phylogeny programs do not consider gaps, branch length may be biased when sequences are of very different lengths. Divergence between the sequences can also be a criterion. Some authors (Wicker et al. 2007) have suggested a 80–80–80 rule: two sequences can be considered to belong to same TE family if they can be aligned along more than 80 bp, over more than 80% of their length, with more than 80% of identity. This rule is empirical but appears to be useful for classifying TE sequences into families that are consistent for the following annotation step, the annotation of their copies. These authors also suggest a nomenclature system for naming new TEs.

## 2.4 Annotation of Transposable Element Copies

This third phase annotates all TE copies in the genome, resolving the most complex degenerate or nested structures. This requires a library of reference sequences representing the TE families. In the best case, the library is both exhaustive and non-redundant, i.e., each ancestral TE, autonomous or not, is represented by a single consensus sequence. We usually use the manually curated library built as described in the previous section, as well as known TE sequences present in the public data banks. Note that some TE families, particularly those including

structural variants with independent amplification histories, are best represented by several consensus. In such cases, manual curation would retain several consensus for a family, considered here as nonredundant.

### ***2.4.1 Detecting TE Fragments***

The first step mines the genomic sequences with the TE library via local pairwise alignments. Several tools were designed specifically for this purpose, such as REPEATMASKER (Smit et al. 1996–2004), CENSOR (Jurka et al. 1996; Kohany et al. 2006), and BLASTER (Quesneville et al. 2003). Some of these tools incorporate scoring matrices to be used with particular GC percentages, as is the case for isochores in the human genome. All these tools propose a small set of parameter combinations depending on the level of sensitivity required by the user.

Although similar, these tools are complementary. We have shown previously that combining these three programs is the best strategy (Quesneville et al. 2005). The MATCHER program (Quesneville et al. 2003) can then be used to assess the multiple results and keep only the best for each location.

Whatever parameters are used for the pairwise alignments, some of the matches will be false positives, i.e., a TE reference sequence will match a locus although no TE is present. For protein-coding genes, full-length cDNAs can be used for confirmation; unfortunately, there is no equivalent way of checking for TE annotation. An empirical statistical filter, such as implemented in the TEannot pipeline (REPET package) (Flutre et al. 2011), can be used to assess the false positive risk. The genomic sequences are shuffled and screened with the TE library. The alignments obtained on a shuffled sequence can be considered as false positives, then the 95-percentile alignment score is used to filter out spurious alignments obtained with the true genome. Only the matches with the true genomic sequences having a higher score are kept. This procedure guarantees that no observed match scores used for the annotation can be obtained for random sequences with a probably greater than 5%.

### ***2.4.2 Filtering Satellites***

Short simple repeats (SSRs) are short motifs repeated in tandem. Many TE sequences contain SSRs but SSRs are also present in the genome independently. It is therefore necessary to filter out TE matches if they are restricted to SSR that the TE consensus may contain. This can be done by annotating SSRs and then removing TE matches included in SSR annotations. Several efficient programs, for example, TRF (Benson 1999), MREPS (Kolpakov et al. 2003), and REPEATMASKER, are available for SSR annotation. In TEannot from the REPET package, these three programs are launched in parallel, and their results are subsequently combined to be used to eliminate hits due to only SSRs in TE consensus.

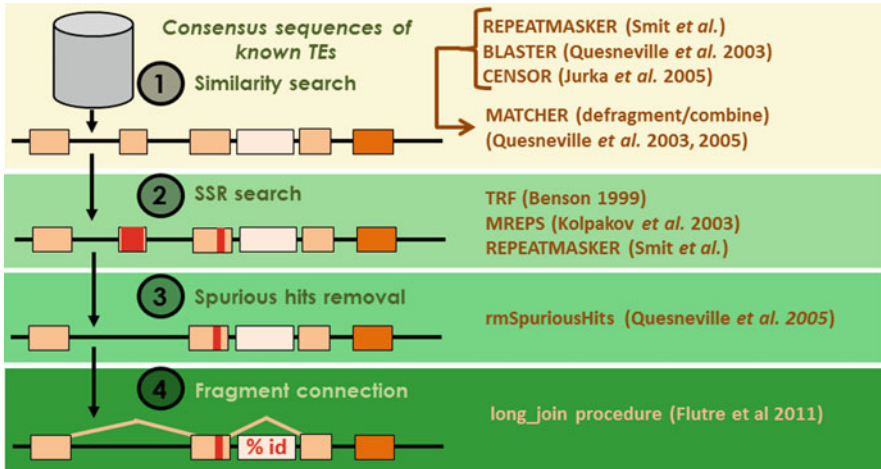


Fig. 2.3 The four steps of the TEannot pipeline (Quesneville et al. 2005)

Satellites are longer motifs, around 100 bp long, also repeated in tandem. Although they are not TEs, they are sometimes difficult to distinguish because they may contain parts of TEs. PILER-TA (Edgar and Myers 2005) detects pyramids in a self-alignment of the genomic sequences. These pyramids can be used to make a consensus of the satellite unit motif. These consensus can then be aligned on the whole genome to find all their occurrences and to distinguish them from TEs.

### 2.4.3 Connecting TE Fragments to Recover TE Copies

Even when TE fragments have been mapped in the genome, the work is only half-finished. Indeed, TE copies can be disrupted into several fragments. A complete TE annotation requires retrieving all copies and thus linking fragments belonging to the same copy when it has transposed.

The first, historical method was manual curation using dot plots. However, this is laborious and curator dependent, and is impractical for large genomes. It requires the curator having detailed knowledge of transposable elements. Moreover, it ignores the age of nested fragments, potentially leading to incongruities. Therefore several computational approaches have been proposed. Many of them are reviewed in the article by Pereira (Pereira 2008).

Joining TE fragments to reconstruct a TE copy is known as a “chain problem” as it corresponds to finding the best chain of local pairwise alignments. The optimal solution is found via dynamic programming as implemented in MATCHER. Subsequently, an additional procedure implemented in the TEannot pipeline (Fig. 2.3) called “long join,” can be used to take into account additional considerations related to TE biology. Two TE fragments distant from each other but mostly separated by

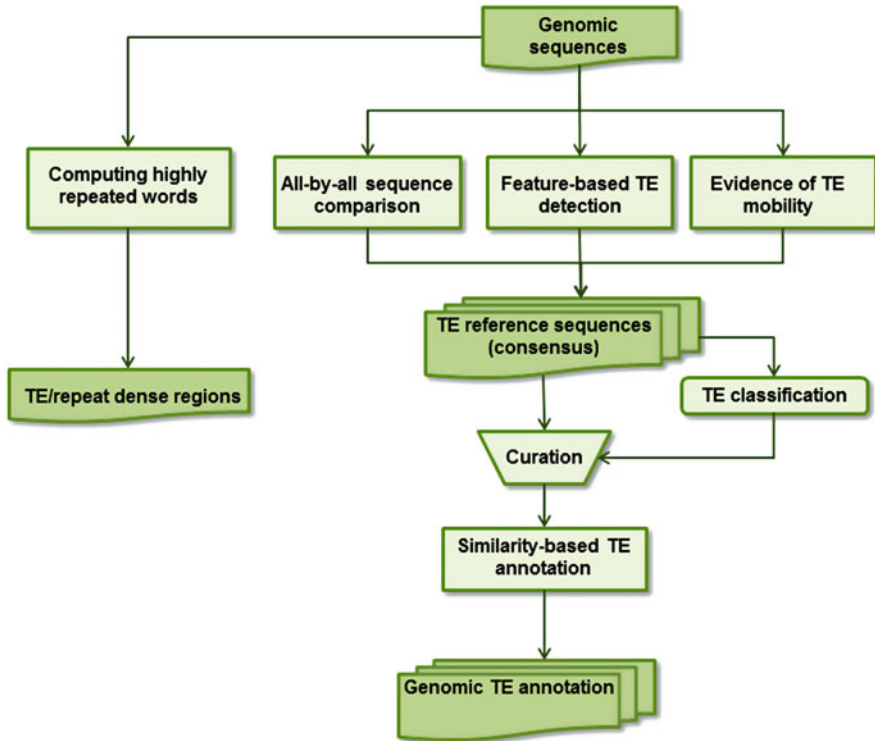
other TE fragments (e.g., at least 95% as in heterochromatin) can be joined as long as the TE fragments between them are younger. The age can be approximated using the percent identity of the matches between the TE reference sequences and the fragments.

## 2.5 Discussion

The contribution of TEs to genome structure and evolution, and their impact on genome assembly has generated an increasing interest in the development of improved methods for their computational analysis. The most common strategy is to detect pairs of similar sequences at different locations in an all-by-all genome comparison, and then cluster these pairs to obtain families of repeats. These methods are not specific to TEs and, therefore, find repeats generated by many different processes, including tandem repeats, segmental duplications, and satellites. Moreover, TE copies can be highly degenerated, deleted, or nested. So repeat detection methods can make errors in the detection of individual TE copies and consequently in defining TE families. We believe that existing automatic approaches still need to be supplemented by expert manual curation. At this step, careful examination is required because some identified families that may appear to be artifactual can in fact be unusual TE families. Indeed, well documented cases illustrate how TE families can appear confusing as they may (1) include cellular genes or parts of genes [e.g., pack-MULEs (Jiang et al. 2004) or *Helitrons* (Morgante et al. 2005a)], (2) be restricted to rDNA genes [e.g., the R2 Non-LTR retroelement superfamily (Eickbush et al. 1997)], or (3) form telomeres [in *Drosophila* (Clark et al. 2007)]. Close examination of noncanonical cases may also reveal new and interesting TE families or particular transposition events [e.g., macrotranspositions (Gray 2000)].

Knowledge-based TE detection methods (i.e., based on structure or similarity to distant TEs) have distinct advantages over *de novo* repeat discovery methods. They capitalize on prior knowledge established from the large number of previously reported TE sequences. Thus, they are more likely to detect *bona fide* TEs, including even those present as only a single copy in the genome. However, these methods are not well suited to the discovery of new TEs (especially of new types). Moreover, these methods have intrinsic ascertainment biases. For example, miniature inverted repeat transposable elements (MITEs) and short interspersed nuclear elements (SINEs) will be under-identified if only similarity-based methods are used because these TEs are composed entirely of noncoding sequences.

For some species, only parts of the genomic sequences are available as BAC sequences assembly. Working on a genome subset could be difficult for all-by-all genome comparison approaches as a TE might appear not repeated if other copies are not yet sequenced. Detection sensitivity of such approaches increase on both the sequenced fraction of the genome and its repeat density. Consequently, according to the sequence size and the repeat density, all-by-all genome comparison approaches may be used with more or less success. Interestingly, detection



**Fig. 2.4** Workflow for annotating TEs in genomic sequences

sensitivity of knowledge-based approaches (i.e., based on structure or similarity to distant TEs) is independent of the sequenced fraction, making them highly recommended here.

Through our experience with many genome projects (Cock et al. 2010; Abad et al. 2008; Amselem et al. 2011; Cuomo et al. 2007; Duplessis et al. 2011; Martin et al. 2008, 2010; Nene et al. 2007; Quesneville et al. 2003, 2005; Rouxel et al. 2011; Spanu et al. 2010), we have assessed the relative benefits of using different programs for TE detection, clustering, and multiple alignments. Our investigations suggest that only combined approaches, using both *de novo* and knowledge-based TE detection methods, are likely to produce reasonably comprehensive and sensitive results. Figure 2.4 shows the general workflow to follow for annotating TEs. In view of this, the REPET package (Flutre et al. 2011) has been developed. It is composed of two pipelines, TEdenovo and TEannot. These pipelines launch several different prediction programs in parallel and then combine their results to optimize the accuracy and exhaustiveness of TE detection. Even with this sophisticated pipeline, manual curation is still needed. Hence, in addition to the automation of all the steps required for the TE annotation, it computes data that are useful for the manual curation, including TE sequence multiple alignments, TE sequence phylogenies,

and TE evidence. Sequencing costs have dropped dramatically and sequences have thus become easier to obtain. However, sequence analysis remains a major bottleneck. Efficient analysis pipelines are required. They need to be quick and robust to accelerate the pace of data production; they should also exploit the knowledge of the few specialists able to perform genome analysis on a large scale so that TE annotations are made available to the wider community of scientists.

**Acknowledgments** This work was supported in part by grants from the Agence Nationale de la Recherche (Holocentrism project, to HQ [grant number ANR-07-BLAN-0057]) and the Centre National de la Recherche Scientifique—Groupement de Recherche “Elements Transposables.” TF was supported by a PhD studentship from the Institut National de la Recherche Agronomique. EP was supported by a postdoctoral fellowship from the Agence Nationale de la Recherche.

## References

- Abad P, Gouzy J, Aury JM, Castagnone-Sereno P, Danchin EG, Deleury E, Perfus-Barbeoch L, Anthouard V, Artiguenave F, Blok VC, Caillaud MC, Coutinho PM, Dasilva C, De Luca F, Deau F, Esquibet M, Flutre T, Goldstone JV, Hamamouch N, Hewezi T, Jaillon O, Jubin C, Leonetti P, Magliano M, Maier TR, Markov GV, McVeigh P, Pesole G, Poulain J, Robinson-Rechavi M, Sallet E, Segurens B, Steinbach D, Tytgat T, Ugarte E, van Ghelder C, Veronico P, Baum TJ, Blaxter M, Blevé-Zacheo T, Davis EL, Ewbank JJ, Favery B, Grenier E, Henriessat B, Jones JT, Laudet V, Maule AG, Quesneville H, Rosso MN, Schiex T, Smant G, Weissenbach J, Wincker P (2008) Genome sequence of the metazoan plant-parasitic nematode *Meloidogyne incognita*. *Nat Biotechnol* 26:909–915
- Abrusan G, Grundmann N, DeMester L, Makalowski W (2009) TEclass—a tool for automated classification of unknown eukaryotic transposable elements. *Bioinformatics* 25:1329–1330
- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402
- Amselem J, Cuomo CA, van Kan JA, Viaud M, Benito EP, Couloux A, Coutinho PM, de Vries RP, Dyer PS, Fillinger S, Fournier E, Gout L, Hahn M, Kohn L, Lapalu N, Plummer KM, Pradier JM, Quevillon E, Sharon A, Simon A, ten Have A, Tudzynski B, Tudzynski P, Wincker P, Andrew M, Anthouard V, Beever RE, Beffa R, Benoit I, Bouzid O, Brault B, Chen Z, Choquer M, Collemare J, Cotton P, Danchin EG, Da Silva C, Gautier A, Giraud C, Giraud T, Gonzalez C, Grossetete S, Guldener U, Henriessat B, Howlett BJ, Kodira C, Kretschmer M, Lappartient A, Leroch M, Levis C, Mauceli E, Neuveglise C, Oeser B, Pearson M, Poulain J, Poussereau N, Quesneville H, Rasclé C, Schumacher J, Segurens B, Sexton A, Silva E, Sirven S, Soanes DM, Talbot NJ, Templeton M, Yandava C, Yarden O, Zeng Q, Rollins JA, Lebrun MH, Dickman M (2011) Genomic analysis of the necrotrophic fungal pathogens *Sclerotinia sclerotiorum* and *Botrytis cinerea*. *PLoS Genet* 7:e1002230
- Bao Z, Eddy SR (2002) Automated *de novo* identification of repeat sequence families in sequenced genomes. *Genome Res* 12:1269–1276
- Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* 27:573–580
- Bergman CM, Quesneville H (2007) Discovering and detecting transposable elements in genome sequences. *Brief Bioinform* 8:382–392
- Bergman CM, Quesneville H, Anxolabehere D, Ashburner M (2006) Recurrent insertion and duplication generate networks of transposable element sequences in the *Drosophila melanogaster* genome. *Genome Biol* 7:R112



- Blumenstiel JP, Hartl DL, Lozovsky ER (2002) Patterns of insertion and deletion in contrasting chromatin domains. *Mol Biol Evol* 19:2211–2225
- Bourque G (2009) Transposable elements in gene regulation and in the evolution of vertebrate genomes. *Curr Opin Genet Dev* 19:607–612
- Bundock P, Hooykaas P (2005) An Arabidopsis hAT-like transposase is essential for plant development. *Nature* 436:282–284
- Caspi A, Pachter L (2006) Identification of transposable elements using multiple alignments of related genomes. *Genome Res* 16:260–270
- 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
- Clamp M, Cuff J, Searle SM, Barton GJ (2004) The Jalview Java alignment editor. *Bioinformatics* 20:426–427
- Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, Kaufman TC, Kellis M, Gelbart W, Iyer VN, Pollard DA, Sackton TB, Larracuente AM, Singh ND, Abad JP, Abt DN, Adryan B, Aguade M, Akashi H, Anderson WW, Aquadro CF, Ardell DH, Arguello R, Artieri CG, Barbash DA, Barker D, Barsanti P, Batterham P, Batzoglou S, Begun D, Bhutkar A, Blanco E, Bosak SA, Bradley RK, Brand AD, Brent MR, Brooks AN, Brown RH, Butlin RK, Caggese C, Calvi BR, Bernardo de Carvalho A, Caspi A, Castrezana S, Celniker SE, Chang JL, Chapple C, Chatterji S, Chinwalla A, Civetta A, Clifton SW, Comeron JM, Costello JC, Coyne JA, Daub J, David RG, Delcher AL, Delehaunty K, Do CB, Ebling H, Edwards K, Eickbush T, Evans JD, Filipinski A, Findeiss S, Freyhult E, Fulton L, Fulton R, Garcia AC, Gardiner A, Garfield DA, Garvin BE, Gibson G, Gilbert D, Gnerre S, Godfrey J, Good R, Gotea V, Gravely B, Greenberg AJ, Griffiths-Jones S, Gross S, Guigo R, Gustafson EA, Haerty W, Hahn MW, Halligan DL, Halpern AL, Halter GM, Han MV, Heger A, Hillier L, Hinrichs AS, Holmes I, Hoskins RA, Hubisz MJ, Hultmark D, Huntley MA, Jaffe DB, Jagadeeshan S, Jeck WR, Johnson J, Jones CD, Jordan WC, Karpen GH, Kataoka E, Keightley PD, Kheradpour P, Kirkness EF, Koerich LB, Kristiansen K, Kudrna D, Kulathinal RJ, Kumar S, Kwok R, Lander E, Langley CH, Lapoint R, Lazzaro BP, Lee SJ, Levesque L, Li R, Lin CF, Lin MF, Lindblad-Toh K, Llopart A, Long M, Low L, Lozovsky E, Lu J, Luo M, Machado CA, Makalowski W, Marzo M, Matsuda M, Matzkin L, McAllister B, McBride CS, McKernan B, McKernan K, Mendez Lago M, Minx P, Mollenhauer MU, Montooth K, Mount SM, Mu X, Myers E, Negre B, Newfield S, Nielsen R, Noor MA, O'Grady P, Pachter L, Papaceit M, Parisi MJ, Parisi M, Parts L, Pedersen JS, Pesole G, Phillippy AM, Ponting CP, Pop M, Porcelli D, Powell JR, Prohaska S, Pruitt K, Puig M, Quesneville H, Ram KR, Rand D, Rasmussen MD, Reed LK, Reenan R, Reily A, Remington KA, Rieger TT, Ritchie MG, Robin C, Rogers YH, Rohde C, Rozas J, Rubenfield MJ, Ruiz A, Russo S, Salzberg SL, Sanchez-Gracia A, Saranga DJ, Sato H, Schaeffer SW, Schatz MC, Schlenke T, Schwartz R, Segarra C, Singh RS, Sirot L, Sirota M, Sisneros NB, Smith CD, Smith TF, Spieth J, Stage DE, Stark A, Stephan W, Strausberg RL, Stempel S, Sturgill D, Sutton G, Sutton GG, Tao W, Teichmann S, Tobar YN, Tomimura Y, Tsolas JM, Valente VL, Venter E, Venter JC, Vicario S, Vieira FG, Vilella AJ, Villasante A, Walenz B, Wang J, Wasserman M, Watts T, Wilson D, Wilson RK, Wing RA, Wolfner MF, Wong A, Wong GK, Wu CI, Wu G, Yamamoto D, Yang HP, Yang SP, Yorke JA, Yoshida K, Zdobnov E, Zhang P, Zhang Y, Zimin AV, Baldwin J, Abdouelleil A, Abdulkadir J, Abebe A, Abera B, Abreu J, Acer SC, Aftuck L, Alexander A, An P, Anderson E, Anderson S, Arachi H, Azer M, Bachantsang P, Barry A, Bayul T, Berlin A, Bessette D, Bloom T, Blye J, Boguslavskiy L, Bonnet C, Boukhgalter B, Bourzgui I, Brown A, Cahill P, Channer S, Cheshatsang Y, Chuda L, Citroen M, Collymore A, Cooke P, Costello M, D'Aco K, Daza R, De Haan G, DeGray S, DeMaso C, Dhargay N, Dooley K, Dooley E, Doricent M, Dorje P, Dorjee K, Dupes A, Elong R, Falk J, Farina A, Faro S, Ferguson D, Fisher S, Foley CD, Franke A, Friedrich D, Gadbois L, Gearin G, Gearin CR, Giannoukos G, Goode T, Graham J, Grandbois E, Grewal S, Gyaltsen K, Hafez N, Hagos B, Hall J, Henson C, Hollinger A, Honan T, Huard MD,

- Hughes L, Hurlhala B, Husby ME, Kamat A, Kanga B, Kashin S, Khazanovich D, Kisner P, Lance K, Lara M, Lee W, Lennon N, Letendre F, LeVine R, Lipovsky A, Liu X, Liu J, Liu S, Lokyitsang T, Lokyitsang Y, Lubonja R, Lui A, MacDonald P, Magnalis V, Maru K, Matthews C, McCusker W, McDonough S, Mehta T, Meldrim J, Meneus L, Mihai O, Mihalev A, Mihova T, Mittelman R, Mlenga V, Montmayeur A, Mulrain L, Navidi A, Naylor J, Negash T, Nguyen T, Nguyen N, Nicol R, Norbu C, Norbu N, Novod N, O'Neill B, Osman S, Markiewicz E, Oyono OL, Patti C, Phunkhang P, Pierre F, Priest M, Raghuraman S, Rege F, Reyes R, Rise C, Rogov P, Ross K, Ryan E, Settupalli S, Shea T, Sherpa N, Shi L, Shih D, Sparrow T, Spaulding J, Stalker J, Stange-Thomann N, Stavropoulos S, Stone C, Strader C, Tesfaye S, Thomson T, Thoulutsang Y, Thoulutsang D, Topham K, Topping I, Tsamla T, Vassiliev H, Vo A, Wangchuk T, Wangdi T, Weiland M, Wilkinson J, Wilson A, Yadav S, Young G, Yu Q, Zembek L, Zhong D, Zimmer A, Zwirko Z, Jaffe DB, Alvarez P, Brockman W, Butler J, Chin C, Gnerre S, Grabherr M, Kleber M, Mauceli E, MacCallum I (2007) Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* 450:203–218
- Cock JM, Sterck L, Rouze P, Scornet D, Allen AE, Amoutzias G, Anthouard V, Artiguenave F, Aury JM, Badger JH, Beszteri B, Billiau K, Bonnet E, Bothwell JH, Bowler C, Boyen C, Brownlee C, Carrano CJ, Charrier B, Cho GY, Coelho SM, Collen J, Corre E, Da Silva C, Delage L, Delaroque N, Dittami SM, Doulbeau S, Elias M, Farnham G, Gachon CM, Gschloessl B, Heesch S, Jabbari K, Jubin C, Kawai H, Kimura K, Kloareg B, Kupper FC, Lang D, Le Bail A, Leblanc C, Lerouge P, Lohr M, Lopez PJ, Martens C, Maumus F, Michel G, Miranda-Saavedra D, Morales J, Moreau H, Motomura T, Nagasato C, Napoli CA, Nelson DR, Nyvall-Collen P, Peters AF, Pommier C, Potin P, Poulain J, Quesneville H, Read B, Rensing SA, Ritter A, Rousvoal S, Samanta M, Samson G, Schroeder DC, Segurens B, Strittmatter M, Tonon T, Tregear JW, Valentin K, von Dassow P, Yamagishi T, Van de Peer Y, Wincker P (2010) The *Ectocarpus* genome and the independent evolution of multicellularity in brown algae. *Nature* 465:617–621
- Cuomo CA, Guldener U, Xu JR, Trail F, Turgeon BG, Di Pietro A, Walton JD, Ma LJ, Baker SE, Rep M, Adam G, Antoniw J, Baldwin T, Calvo S, Chang YL, Decaprio D, Gale LR, Gnerre S, Goswami RS, Hammond-Kosack K, Harris LJ, Hilburn K, Kennell JC, Kroken S, Magnuson JK, Mannhaupt G, Mauceli E, Mewes HW, Mitterbauer R, Muehlbauer G, Munsterkötter M, Nelson D, O'Donnell K, Ouellet T, Qi W, Quesneville H, Roncero MI, Seong KY, Tetko IV, Urban M, Waalwijk C, Ward TJ, Yao J, Birren BW, Kistler HC (2007) The *Fusarium graminearum* genome reveals a link between localized polymorphism and pathogen specialization. *Science* 317:1400–1402
- 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
- Duplessis S, Cuomo CA, Lin YC, Aerts A, Tisserant E, Veneault-Fourrey C, Joly DL, Hacquard S, Amselem J, Cantarel BL, Chiu R, Coutinho PM, Feau N, Field M, Frey P, Gellhaye E, Goldberg J, Grabherr MG, Kodira CD, Kohler A, Kues U, Lindquist EA, Lucas SM, Mago R, Mauceli E, Morin E, Murat C, Pangilinan JL, Park R, Pearson M, Quesneville H, Rouhier N, Sakthikumar S, Salamov AA, Schmutz J, Selles B, Shapiro H, Tanguay P, Tuskan GA, Henrissat B, Van de Peer Y, Rouze P, Ellis JG, Dodds PN, Schein JE, Zhong S, Hamelin RC, Grigoriev IV, Szabo LJ, Martin F (2011) Obligate biotrophy features unraveled by the genomic analysis of rust fungi. *Proc Natl Acad Sci USA* 108:9166–9171
- Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5:113
- Edgar RC, Myers EW (2005) PILER: identification and classification of genomic repeats. *Bioinformatics* 21(Suppl 1):i152–i158
- Eickbush TH, Burke WD, Eickbush DG, Lathe WC 3rd (1997) Evolution of R1 and R2 in the rDNA units of the genus *Drosophila*. *Genetica* 100:49–61
- Ellinghaus D, Kurtz S, Willhoeft U (2008) LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* 9:18

- Feschotte C (2008) Transposable elements and the evolution of regulatory networks. *Nat Rev Genet* 9:397–405
- Feschotte C, Keswani U, Ranganathan N, Guibotsy ML, Levine D (2009) Exploring repetitive DNA landscapes using REPCCLASS, a tool that automates the classification of transposable elements in eukaryotic genomes. *Genome Biol Evol* 2009:205–220
- Finn RD, Mistry J, Tate J, Coghill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer EL, Eddy SR, Bateman A (2010) The Pfam protein families database. *Nucleic Acids Res* 38:D211–D222
- Finnegan DJ (1989) Eukaryotic transposable elements and genome evolution. *Trends Genet* 5:103–107
- Fiston-Lavier AS, Anxolabehere D, Quesneville H (2007) A model of segmental duplication formation in *Drosophila melanogaster*. *Genome Res* 17:1458–1470
- Flutre T, Duprat E, Feuillet C, Quesneville H (2011) Considering transposable element diversification in *de novo* annotation approaches. *PLoS One* 6:e16526
- Gray YH (2000) It takes two transposons to tango: transposable-element-mediated chromosomal rearrangements. *Trends Genet* 16:461–468
- Gu W, Castoe TA, Hedges DJ, Batzer MA, Pollock DD (2008) Identification of repeat structure in large genomes using repeat probability clouds. *Anal Biochem* 380(1):77–83
- Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52:696–704
- Han Y, Wessler SR (2010) MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res* 38:e199
- Hikosaka A, Kawahara A (2010) A systematic search and classification of T2 family miniature inverted-repeat transposable elements (MITEs) in *Xenopus tropicalis* suggests the existence of recently active MITE subfamilies. *Mol Genet Genomics* 283:49–62
- Huang X (1994) On global sequence alignment. *Comput Appl Biosci* 10:227–235
- Janicki M, Rooke R, Yang G (2011) Bioinformatics and genomic analysis of transposable elements in eukaryotic genomes. *Chromosome Res* 19:787–808
- Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR (2004) Pack-MULE transposable elements mediate gene evolution in plants. *Nature* 431:569–573
- Jurka J, Klonowski P, Dagman V, Pelton P (1996) CENSOR—a program for identification and elimination of repetitive elements from DNA sequences. *Comput Chem* 20:119–121
- Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* 110:462–467
- Kalyanaraman A, Aluru S (2006) Efficient algorithms and software for detection of full-length LTR retrotransposons. *J Bioinform Comput Biol* 4:197–216
- Kapitonov VV, Jurka J (2005) RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. *PLoS Biol* 3:e181
- Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 30:3059–3066
- Kohany O, Gentles AJ, Hankus L, Jurka J (2006) Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics* 7:474
- Kolpakov R, Bana G, Kucherov G (2003) mreps: efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Res* 31:3672–3678
- 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
- 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, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A,

- Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramsler J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglu S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kasprzyk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrino A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ (2001) Initial sequencing and analysis of the human genome. *Nature* 409:860–921
- Le QH, Wright S, Yu Z, Bureau T (2000) Transposon diversity in *Arabidopsis thaliana*. *Proc Natl Acad Sci USA* 97:7376–7381
- Lee W, Chen SL (2002) Genome-tools: a flexible package for genome sequence analysis. *Biotechniques* 33:1334–1341
- 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
- Li R, Ye J, Li S, Wang J, Han Y, Ye C, Wang J, Yang H, Yu J, Wong GK, Wang J (2005) ReAS: Recovery of ancestral sequences for transposable elements from the unassembled reads of a whole genome shotgun. *PLoS Comput Biol* 1:e43
- Lucier JF, Perreault J, Noel JF, Boire G, Perreault JP (2007) RTAnalyzer: a web application for finding new retrotransposons and detecting L1 retrotransposition signatures. *Nucleic Acids Res* 35:W269–W274
- Martin F, Aerts A, Ahrn D, Brun A, Danchin EG, Duchaussoy F, Gibon J, Kohler A, Lindquist E, Pereda V, Salamov A, Shapiro HJ, Wuyts J, Blaudez D, Buee JM, Brokstein P, Canback B, Cohen D, Courty PE, Coutinho PM, Delaruelle C, Detter JC, Deveau A, DiFazio S, Duplessis S, Fraissinet-Tachet L, Lucie E, Frey-Klett P, Fourrey C, Feussner I, Gay G, Grimwood J, Hoegger PJ, Jain P, Kilaru S, Labbe J, Lin YC, Legue V, Le Tacon F, Marmeisse R, Melayah D, Montanini B, Muratet M, Nehls U, Niculita-Hirzel H, Oudot-Le Secq MP, Peter M, Quesneville H, Rajashekar B, Reich M, Rouhier N, Schmutz J, Yin T, Chalot M, Henrissat B, Kues U, Lucas S, Van de Peer Y, Podila GK, Polle A, Pukkila PJ, Richardson PM, Rouze P, Sanders IR, Stajich JE, Tunlid A, Tuskan G, Grigoriev IV (2008) The genome of *Laccaria bicolor* provides insights into mycorrhizal symbiosis. *Nature* 452:88–92
- Martin F, Kohler A, Murat C, Balestrini R, Coutinho PM, Jaillon O, Montanini B, Morin E, Noel B, Percudani R, Porcel B, Rubini A, Amicucci A, Amselem J, Anthouard V, Arcioni S,

- Artiguenave F, Aury JM, Ballario P, Bolchi A, Brenna A, Brun A, Buee M, Cantarel B, Chevalier G, Couloux A, Da Silva C, Denoeud F, Duplessis S, Ghignone S, Hilselberger B, Iotti M, Marçais B, Mello A, Miranda M, Pacioni G, Quesneville H, Riccioni C, Ruotolo R, Splivallo R, Stocchi V, Tisserant E, Viscomi AR, Zambonelli A, Zampieri E, Henrissat B, Lebrun MH, Paolocci F, Bonfante P, Ottonello S, Wincker P (2010) Perigord black truffle genome uncovers evolutionary origins and mechanisms of symbiosis. *Nature* 464:1033–1038
- McCarthy EM, McDonald JF (2003) LTR\_STRUC: a novel search and identification program for LTR retrotransposons. *Bioinformatics* 19:362–367
- Morgante M, Brunner S, Pea G, Fengler K, Zuccolo A, Rafalski A (2005a) Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nat Genet* 37:997–1002
- Morgante M, Policriti A, Vitacolonna N, Zuccolo A (2005b) Structured motifs search. *J Comput Biol* 12:1065–1082
- Nene V, Wortman JR, Lawson D, Haas B, Kodira C, Tu ZJ, Loftus B, Xi Z, Megy K, Grabherr M, Ren Q, Zdobnov EM, Lobo NF, Campbell KS, Brown SE, Bonaldo MF, Zhu J, Sinkins SP, Hogenkamp DG, Amedeo P, Arensburger P, Atkinson PW, Bidwell S, Biedler J, Birney E, Bruggner RV, Costas J, Coy MR, Crabtree J, Crawford M, Debruyen B, Decaprio D, Eiglmeier K, Eisenstadt E, El-Dorry H, Gelbart WM, Gomes SL, Hammond M, Hannick LI, Hogan JR, Holmes MH, Jaffe D, Johnston JS, Kennedy RC, Koo H, Kravitz S, Kriventseva EV, Kulp D, Labutti K, Lee E, Li S, Lovin DD, Mao C, Mauceli E, Menck CF, Miller JR, Montgomery P, Mori A, Nascimento AL, Naveira HF, Nusbaum C, O’Leary S, Orvis J, Perteu M, Quesneville H, Reidenbach KR, Rogers YH, Roth CW, Schneider JR, Schatz M, Shumway M, Stanke M, Stinson EO, Tubio JM, Vanzee JP, Verjovski-Almeida S, Werner D, White O, Wyder S, Zeng Q, Zhao Q, Zhao Y, Hill CA, Raikhel AS, Soares MB, Knudson DL, Lee NH, Galagan J, Salzberg SL, Paulsen IT, Dimopoulos G, Collins FH, Birren B, Fraser-Liggett CM, Severson DW (2007) Genome sequence of *Aedes aegypti*, a major arbovirus vector. *Science* 316:1718–1723
- Orgel LE, Crick FH (1980) Selfish DNA: the ultimate parasite. *Nature* 284:604–607
- Pereira V (2004) Insertion bias and purifying selection of retrotransposons in the *Arabidopsis thaliana* genome. *Genome Biol* 5:R79
- Pereira V (2008) Automated paleontology of repetitive DNA with REANNOTATE. *BMC Genomics* 9:614
- Petrov DA (2001) Evolution of genome size: new approaches to an old problem. *Trends Genet* 17:23–28
- Piegu B, Guyot R, Picault N, Roulin A, Saniyal A, Kim H, Collura K, Brar DS, Jackson S, Wing RA, Panaud O (2006) Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. *Genome Res* 16:1262–1269
- Price AL, Jones NC, Pevzner PA (2005) *De novo* identification of repeat families in large genomes. *Bioinformatics* 21(Suppl 1):i351–i358
- Quesneville H, Nouaud D, Anxolabehere D (2003) Detection of new transposable element families in *Drosophila melanogaster* and *Anopheles gambiae* genomes. *J Mol Evol* 57(Suppl 1): S50–S59
- 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* 1:166–175
- Rasmussen K, Stoye J, Myers EW (2005) Efficient q-gram filters for finding all e-matches over a given length. In: Heidelberg SB (ed) RECOMB, pp 189–203
- Rho M, Tang H (2009) MGEscan-non-LTR: computational identification and classification of autonomous non-LTR retrotransposons in eukaryotic genomes. *Nucleic Acids Res* 37:e143
- Rho M, Choi JH, Kim S, Lynch M, Tang H (2007) *De novo* identification of LTR retrotransposons in eukaryotic genomes. *BMC Genomics* 8:90

- Rouxel T, Grandaubert J, Hane JK, Hoede C, van de Wouw AP, Couloux A, Dominguez V, Anthouard V, Bally P, Bourras S, Cozijsen AJ, Ciuffetti LM, Degrave A, Dilmaghani A, Duret L, Fudal I, Goodwin SB, Gout L, Glaser N, Linglin J, Kema GH, Lapalu N, Lawrence CB, May K, Meyer M, Ollivier B, Poulain J, Schoch CL, Simon A, Spatafora JW, Stachowiak A, Turgeon BG, Tyler BM, Vincent D, Weissenbach J, Anselme J, Quesneville H, Oliver RP, Wincker P, Balesdent MH, Howlett BJ (2011) Effector diversification within compartments of the *Leptosphaeria maculans* genome affected by repeat-induced point mutations. *Nat Commun* 2:202
- Santangelo AM, de Souza FS, Franchini LF, Bumashny VF, Low MJ, Rubinstein M (2007) Ancient exaptation of a CORE-SINE retroposon into a highly conserved mammalian neuronal enhancer of the proopiomelanocortin gene. *PLoS Genet* 3:1813–1826
- Santiago N, Herraiz C, Goni JR, Messegueur X, Casacuberta JM (2002) Genome-wide analysis of the Emigrant family of MITEs of *Arabidopsis thaliana*. *Mol Biol Evol* 19:2285–2293
- Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlinson C, Strong C, Delehaunty K, Fronick C, Courtney B, Rock SM, Belter E, Du F, Kim K, Abbott RM, Cotton M, Levy A, Marchetto P, Ochoa K, Jackson SM, Gillam B, Chen W, Yan L, Higginbotham J, Cardenas M, Waligorski J, Applebaum E, Phelps L, Falcone J, Kanchi K, Thane T, Scimone A, Thane N, Henke J, Wang T, Ruppert J, Shah N, Rotter K, Hodges J, Ingenthron E, Cordes M, Kohlberg S, Sgro J, Delgado B, Mead K, Chinwalla A, Leonard S, Crouse K, Collura K, Kudrna D, Currie J, He R, Angelova A, Rajasekar S, Mueller T, Lomeli R, Scara G, Ko A, Delaney K, Wissotski M, Lopez G, Campos D, Braidotti M, Ashley E, Golser W, Kim H, Lee S, Lin J, Dujmic Z, Kim W, Talag J, Zuccolo A, Fan C, Sebastian A, Kramer M, Spiegel L, Nascimento L, Zutavern T, Miller B, Ambrose C, Muller S, Spooner W, Narechania A, Ren L, Wei S, Kumari S, Faga B, Levy MJ, McMahan L, Van Buren P, Vaughn MW, Ying K, Yeh CT, Emrich SJ, Jia Y, Kalyanaraman A, Hsia AP, Barbazuk WB, Baucom RS, Brutnell TP, Carpita NC, Chaparro C, Chia JM, Deragon JM, Estill JC, Fu Y, Jeddelloh JA, Han Y, Lee H, Li P, Lisch DR, Liu S, Liu Z, Nagel DH, McCann MC, SanMiguel P, Myers AM, Nettleton D, Nguyen J, Penning BW, Ponnala L, Schneider KL, Schwartz DC, Sharma A, Soderlund C, Springer NM, Sun Q, Wang H, Waterman M, Westerman R, Wolfgruber TK, Yang L, Yu Y, Zhang L, Zhou S, Zhu Q, Bennetzen JL, Dawe RK, Jiang J, Jiang N, Presting GG, Wessler SR, Aluru S, Martienssen RA, Clifton SW, McCombie WR, Wing RA, Wilson RK (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science* 326:1112–1115
- Smit AFA, Hubley R, Green P (1996–2004) RepeatMasker open-3.0. Institute for Systems Biology
- Spanu PD, Abbott JC, Amselem J, Burgis TA, Soanes DM, Stuber K, Loren V, van Themaat E, Brown JK, Butcher SA, Gurr SJ, Lebrun MH, Ridout CJ, Schulze-Lefert P, Talbot NJ, Ahmadinejad N, Ametz C, Barton GR, Benjdia M, Bidzinski P, Bindschedler LV, Both M, Brewer MT, Cadle-Davidson L, Cadle-Davidson MM, Collemare J, Cramer R, Frenkel O, Godfrey D, Harriman J, Hoede C, King BC, Klages S, Kleemann J, Knoll D, Koti PS, Kreplak J, Lopez-Ruiz FJ, Lu X, Maekawa T, Mahanil S, Micali C, Milgroom MG, Montana G, Noir S, O'Connell RJ, Oberhaensli S, Parlange F, Pedersen C, Quesneville H, Reinhardt R, Rott M, Sacristan S, Schmidt SM, Schon M, Skamnioti P, Sommer H, Stephens A, Takahara H, Thordal-Christensen H, Vigouroux M, Wessling R, Wicker T, Panstruga R (2010) Genome expansion and gene loss in powdery mildew fungi reveal tradeoffs in extreme parasitism. *Science* 330:1543–1546
- Steinbiss S, Willhoeft U, Gremme G, Kurtz S (2009) Fine-grained annotation and classification of *de novo* predicted LTR retrotransposons. *Nucleic Acids Res* 37:7002–7013
- Szak ST, Pickeral OK, Makalowski W, Boguski MS, Landsman D, Boeke JD (2002) Molecular archeology of L1 insertions in the human genome. *Genome Biol* 3:research0052
- Tu Z (2001) Eight novel families of miniature inverted repeat transposable elements in the African malaria mosquito, *Anopheles gambiae*. *Proc Natl Acad Sci USA* 98:1699–1704

- Tu Z, Li S, 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
- 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* 8:973–982
- 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
- Yang L, Bennetzen JL (2009) Structure-based discovery and description of plant and animal Helitrons. *Proc Natl Acad Sci USA* 106:12832–12837
- Yang G, Hall TC (2003) MAK, a computational tool kit for automated MITE analysis. *Nucleic Acids Res* 31:3659–3665
- Zhou L, Mitra R, Atkinson PW, Hickman AB, Dyda F, Craig NL (2004) Transposition of hAT elements links transposable elements and V(D)J recombination. *Nature* 432:995–1001