genes

Review

Methodologies for the *De novo* Discovery of Transposable Element Families

Jessica M. Storer, Robert Hubley, Jeb Rosen and Arian F. A. Smit *

Institute for Systems Biology, Seattle, WA 98109, USA; jessica.storer@isbscience.org (J.M.S.); robert.hubley@isbscience.org (R.H.); jeb.rosen@isbscience.org (J.R.)
* Correspondence: arian.smit@isbscience.org

Abstract: The discovery and characterization of transposable element (TE) families are crucial tasks in the process of genome annotation. Careful curation of TE libraries for each organism is necessary as each has been exposed to a unique and often complex set of TE families. *De novo* methods have been developed; however, a fully automated and accurate approach to the development of complete libraries remains elusive. In this review, we cover established methods and recent developments in *de novo* TE analysis. We also present various methodologies used to assess these tools and discuss opportunities for further advancement of the field.

Keywords: repeats; transposon; transposable element; *de novo* methods; signature-based methods; genome annotation; curation

check for updates

Citation: Storer, J.M.; Hubley, R.; Rosen, J.; Smit, A.F.A. Methodologies for the *De novo* Discovery of Transposable Element Families. *Genes* **2022**, *13*, 709. https://doi.org/10.3390/genes13040709

Academic Editors: Jürgen Schmitz and Liliya Doronina

Received: 23 March 2022
Accepted: 15 April 2022
Published: 17 April 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Genomes have likely always battled with subsequences that evolved to multiply independently of genome replication. For billions of years, these transposable elements (TEs) have littered genomes with interspersed copies that are generally detrimental or useless for their hosts and thus tend to wither away over time. Depending on the relative rate of TE reproduction and genomic clean-up through random deletions, significant fractions of present genomes are ultimately derived from TEs. Recognized portions are as high as 84% in some cereals and 90% in lungfish [1–3]. Since 1980, it has been suggested that most of the 85–90% of our own genome that is not under functional constraint is TE derived [4]; and by 1996, we could confirm that for almost half the genome [5]. Because of relatively low TE activity and DNA loss, much of our and other vertebrate TE-derived DNA was introduced a long time ago and, through the accumulation of mutations, ranges from difficult to impossible to recognize as such. Over half of the human DNA recognizably derived from (~4 million) TE insertions became part of our genome over 80 million years ago, in a common ancestor of all placental mammals [6,7].

While the persistent onslaught of TEs has been a bane for genomes, as evidenced by the many and wide-ranging defense mechanisms they evolved against them, it forms a veritable boon for phylogenetic research. The advantages of TE insertions as a phylogenetic tool include their high abundance and interspersed distribution, the near-neutral nature of most insertions fixed in a population, the built-in knowledge of the ancestral (absent) state, the virtual absence of back-mutations or parallel events leading to the same sequence pattern (homoplasy), and our ability to recognize ancient events [8]. Not all TEs are equally suitable; less reliable are class II elements that excise from their locus during transposition or elements with more specific target site preferences. Most LINE elements, such as L1 in mammals and CR1 in birds, are close to ideal: random 5′ truncation of most insertions and the variable target site duplication (TSD) lengths distinguish even the rare event of same-site insertions in the same orientation in related genomes.

These qualities of TE insertions have been used with great aplomb to resolve long-standing phylogenetic problems before the availability of complete genome data [9–13].

Genome assemblies for most vertebrates and many other eukaryotes are being rapidly produced and can increasingly be studied in the context of complete and reliable multi-species genome alignments [14–17]. One could expect that the rich detail contained in whole-genome alignments would be more reliable in phylogenetic studies and that the role of TEs in phylogenetic studies will steadily decline, but they will always remain relevant in population studies and may even continue to be the best tool to resolve the trickiest phylogenies, such as species radiations. For these, individual markers show conflicting species trees because two or more speciation events took place when the loci were still polymorphic (incomplete lineage sorting) or due to interspecific gene flow (introgression), primarily via hybridization. Quartet-based summary coalescent methods have the potential to solve these knots when using TE insertions as input data [18,19]. For this to work, it is critical that the ancestral and derived loci are always correctly called, which is dependent on solid repeat annotation with full-length reconstructed TEs and knowledge of insertion behavior, and that the number of phylogenetic markers is high [20]. For phylogenetic purposes, the best product of *de novo* programs therefore is an as complete as possible library of reconstructed TEs, for anything but the youngest elements best presented by consensus sequences or profile HMMs.

TE libraries for the first sequenced genomes were years in the making and the need for automation was apparent early on, especially because the number of sequenced species was expected to grow exponentially. Indeed, in March 2021, the International Nucleotide Sequence Database Collaboration contained genome assemblies for 6480 unique species [21], including higher-quality assemblies for approximately 3300 animal and 800 plant species [22,23]. This is a small fraction of what awaits, with, for example, the Vertebrate Genomes Project aiming to generate complete reference genomes for all ~70,000 extant vertebrates [24] and the Darwin Tree of Life Project planning the same for all ~73,000 eukaryotic species in the UK [25]. These and many other such efforts are coordinated by the Earth BioGenome project to sequence all organisms in the forthcoming decade [21].

Earlier brute force TE library building efforts somewhat simplify repeat analysis in tetrapod genomes, as many ancient repeats are shared between these species and the general nature of the TE fauna is familiar. For most organisms, however, the vast majority of TE copies are lineage specific; a library has to be built from scratch and may contain heretofore unknown elements with their own idiosyncratic challenges.

Early in the 2000s, automation was addressed by a number of labs who developed programs such as RepeatFinder [26], REPuter [27], RECON [28], RepeatScout [29] and PILER [30] that are still in use today. We released an automated version of part of our own pipeline (RepeatModeler) in 2008 and new methods have been introduced steadily, to the point that prospective repeat analysts may be overwhelmed by the choices. Here, we provide an overview of the popular or promising new methods and pipelines to identify interspersed repeats *de novo*. We do not address the analysis of tandem repeats and satellites; the most recent reviews on this that we are aware of are from 2013 and 2015 [31,32] and quite a few promising newer methods have been published since [33–37]. The analysis of segmental duplications, which has seen considerable progress in recent years [38–40] also falls outside the scope of this review, though one should be aware of their existence as they can interfere with the discovery and analysis of TE families. We also do not address the genotyping of TE insertions compared to a reference genome (see reviews [41,42]) and more recently published tools [43–45].

Several reviews on the subject of *de novo* TE analysis have been written in the last dozen years [46–50]. We especially recommend the broader review by Nicolas and colleagues, originally written in 2016 and updated this year [51], which contains excellent introductions to the concepts of sequence indexing and their application to repeat detection. In addition, a comprehensive list of tools for TE analysis is currently being maintained as part of the TEHub project (http://tehub.org; accessed on 16 April 2022). Our focus will be on a comparison of the methodology of the most commonly used programs, the different ways the programs present the results, the need for a standard benchmark to meaningfully

compare results of different programs, and some open problems that none of the programs have truly solved.

2. Why Is *De novo* Repeat Analysis So Hard?

At first thought, identification of interspersed repeats and subsequent calculation of a consensus sequence approximating the original TE appear straightforward. If most instances of TE copies have decayed in a neutral fashion, the accumulation of substitutions and indels should be random and with some knowledge of neutral mutation patterns, expectations can be set for what are likely dispersed copies of the same element instead of chance similarities. With enough copies, a reconstruction of the mobile element should be straightforward, but many complications exist.

While neutral decay provides advantages for repeat detection, the lack of selective constraint means that structural signals of TEs perish as quickly as any other sequence. Thus, a translational comparison or a search for characteristic terminal sequences does not increase sensitivity (quite the opposite).

The size of the genome can interfere with the detection of older and/or lower copy number elements. If TE instances have undergone 20% substitutions since arrival (an example of these are TE copies that arrived in the mouse genome at the time of speciation from hamsters), the distance between any two instances is on average 40%. Detecting matches of such a high divergence level requires very sensitive settings in self-comparison of the genome, making the process impractically slow. To allow more sensitive self-comparison, programs could work with smaller samples of a genome, but lower copy number elements may then go unnoticed.

Extensive fragmentation creates a challenge for algorithms to find the true ends of the TE. Older TE instances tend to be highly fragmented, either through partial deletions or through interruption by insertions, usually of other TEs. In many species, TE copies mostly accumulate in defined heterochromatic, gene-poor or intergenic regions of the genome, in part because their impact is more likely to be neutral. In those regions, overall repeat density can approach 100% of densely nested TE insertions [52,53]. On top of this, some elements tend to be truncated upon insertion. This is particularly so for LINEs, a class of elements that make up the majority of repeats in many vertebrates.

To make things worse, full-length insertions are less likely to occur or persist than fragments, impeding their reconstruction. This can happen during transposition: cut and paste DNA transposons with an internal deletion appear to have an advantage over full, coding elements, perhaps because the transposase has a better chance binding to both termini. The ratio of short elements over full copies can be very high, hampering reconstruction of the long element. Often, the activity of a DNA transposon is only evidenced by the presence of tiny elements with terminal inverted repeats (TIRs) [54,55]. Autonomous elements with long terminal repeats (LTRs) may be outnumbered by elements with a reduced internal sequence [56,57] and LINE elements sometimes give free rides to internal deletion products [58]. Long insertions are also more likely to be selectively disadvantageous to the genome. Severely truncated LINE insertions are thus more likely to be fixed. For the same reason, full-length LTR elements are often reduced to solo LTRs via LTR–LTR recombination and the internal sequences of many LTR elements remain unknown.

Whereas the original sequence of most class II elements can be precisely reconstructed, class I elements evolve in a genome and the differences between instances are a mixture of neutral mutations accumulated in the fixed copies and evolved changes in the source gene(s). A consensus sequence of such families may not match any state of the evolving TE precisely. Over time, class I TEs can change to the point that homology between old and young copies or between copies of two branches is obscure. *De novo* algorithms will not consider these dissimilar copies to represent the same TE family. This is usually preferable, as a consensus or profile HMM sequence model of these aligned yet dissimilar sequences could be meaningless. Instead, several models will be built that represent often partially and sometimes wholly overlapping sets of instances of the evolving TE. Proper clustering

of the instances is complicated, even if sometimes aided by apparent bursts of activity of the TE, resulting in clear "subfamilies" of instances. None of the *de novo* programs currently attempt an automated subfamily analysis.

Regional homology can exist between otherwise unrelated TEs, further complicating defining the true edges of a TE as well as its classification. These regional similarities have multiple origins. (1) A (fragment of a) TE could insert in or be recombined into an active other element. Some TEs, such as Helitrons and non-autonomous LTR elements, are particularly impartial to foreign intrusions. The anomalous L1-dependent SVA and LAVA elements active in ape genomes harbor *Alu* fragments, a retroviral LTR and a low complexity tandem repeat; had their copies been ancient and highly mutated, they would have been painful to reconstruct and classify. (2) Different elements may use the same functional module. The best examples of this phenomenon are probably SINEs. Classical SINEs originate by the happenstance recombination of a small structural RNA, containing an internal pol III promoter, and the 3′ end of a LINE element, which lets them hitch a ride with the latter. (3) Recombination between active TEs is a common feature. This is especially true for LTR elements, where disparate RNAs can be packaged in the same viral particle and template switching of the reverse transcriptase between the two RNA genomes is required for normal replication. Through this mechanism, chimeric-looking LTRs originate, identical LTRs can flank entirely different internal sequences (and vice versa) and integrases of one class of ERVs can even be combined with reverse transcriptases of another [59].

Besides these true recombinant mobile elements, *de novo* programs are also prone to build in silico chimeras of TEs, in part because integration site preferences of prolific TEs can make them frequently appear in tandem or at the (near) same site of other interspersed repeats. For instance, in mammals, L1-dependent SINEs insert in A-rich regions, most frequently provided by the poly-A tail of other SINEs. Because such incidental pairs can become a successful TE, the dimeric primate *Alus* being a prime example, they cannot be dismissed offhand.

Perhaps in part due to the frequency of tandem copies mentioned above, TE instances appear overrepresented in satellite-like tandem repeats [60]. Co-duplication of a TE instance such as that or along with segmental duplications can obfuscate its true extend. Given enough such copies and a significant decay from the original sequence, the model for the TE may become distorted as well. On top of that, *de novo* programs often build models of higher copy number segmental duplications or large tandem repeats as putative TE families. Being "random" fragments of the genome, these models may include coding regions of cellular genes and other unintended sequences. They usually contain copies of TEs and may prevent the discovery of some of these.

Until recently, known eukaryotic TEs ranged in size from 80 bp to approximately 15 kb. Unfortunately, oversized transposable units have now come to light in non-model organisms. These include members of known classes, such as 30 kb LTR elements in planarians, as well as exotic new TEs, such as the up to 180 kb Teratorns in fish [61]. These are a problem for TE class-specific programs, which necessarily impose size limitations, and tend to be fragmented by general *de novo* programs.

Finally, given the low information density present in a four-nucleotide alphabet, low complexity is the common source of false positives in the initial search for repetitive signals, false confidence in alignments of non-homologous sequences, and false extensions of real matches. This is of course particularly problematic for genomes with extreme GC content, but low-complexity problems can arise in subtler ways. *De novo*-created libraries almost always contain models representing common (degraded) simple repeats flanked by unrelated DNA of merely similar composition. Mini-satellites with the same periodicity and just a few bases in common will align "significantly" with each other in the long run and will show up as well.

3. Approaches to TE Discovery and Annotation

The complex nature of TE sequence analysis is reflected in the often-ambiguous usage of terms such as "discovery" and "annotation". It has been useful to define TE discovery as the process of reconstructing/modeling TE families directly from sequence data to generate or augment a TE library. Similarly, TE genome annotation can be viewed as the process of identifying and characterizing all recognizable instances of a TE family or a set of TE families in a genome. The large spectrum of methodologies developed for these tasks over the past two decades has blurred the lines between strict discovery and annotation processes. For this reason, we will focus on characterizing the granularity of results; from methods producing sequence ranges labeled simply as repetitive in nature to methods that produce complete family models and genome annotations.

A further distinction is often made between methods which: (1) discover TE families based on general principles such as subsequence repetition/locality (*de novo*/ab initio methods); (2) employ domain knowledge to detect signatures of known TE class activity/composition (signature-based methods); and (3) methods that produce highly detailed genome annotation of TE instances and depend entirely on a predefined library of TE family models (library-based methods) (Figure 1). *De novo* and signature-based methods are typically employed to generate the input for library-based methods and will be the focus of this review.

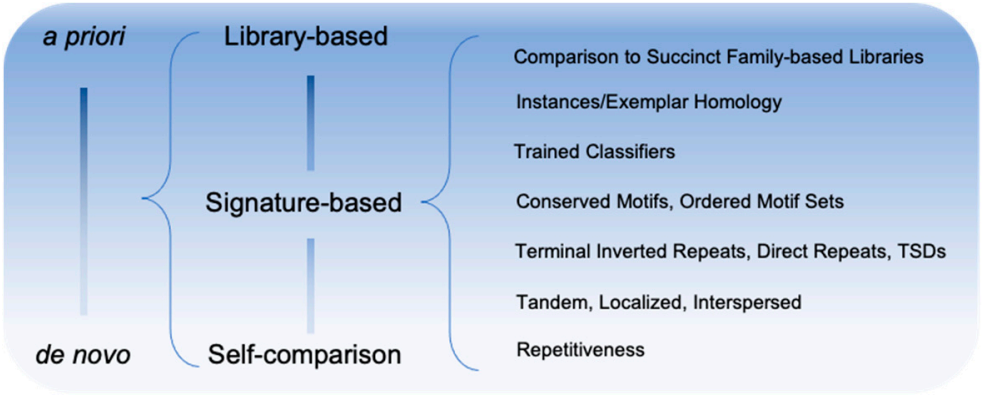

Figure 1. Spectrum of methodologies for the discovery of TE sequences.

The results of a *de novo* analysis come in a dizzying array of forms and granularity. Tools that provide genome annotations (marked with "Annotation Generation" in our software tables) may only separate a sequence into repetitive and non-repetitive subsequences, may report pairwise associations between repetitive subsequences, may group repetitive regions by broad TE classifications (e.g., LTRs and MITEs), or may rigorously report on distinct instances of clustered TE families. Tools falling into the last category typically also generate a library of sequence models for each family that has been discovered. These may take the form of consensus sequences, or representative instance(s) chosen from each family (exemplars). Furthermore, some programs provide provenance for the family definition in the form of sequence ranges for the identified family instances or a multiple sequence alignment of family instances (seed alignment).

4. *De novo* Methodologies

De novo methods have the advantage over signature-based methods that they can identify families that do not belong to a known class of TE or do not share one or more of the diagnostic features signature-based methods employ. They work by detecting exact or closely matching sequence repetitions, extending these matches, and in some cases grouping them into families of related sequences. In Table 1, *de novo* tools are characterized by the level of granularity they produce (families, instances, other), and by the types of family models produced (consensi, exemplars, or other tool-specific representations).

Table 1. *De novo* methods.

Tool	Approach	Input	Granularity	Library Generation				Annotation	Ref.
			Families, Instances or Other	Consensi	Exemplars	Other	Provenance		
CARP	Self-comparison, clustering	Assembly	Families	x			x	x	[62]
dnaPipeTE	Read sampling and assembly	NGS reads	Families	x				x	[63]
LongRepMarker	Pre-assembly, k-mer coverage	SMS/NGS reads, or assembly	Instances					x	[64]
mer-engine	K-mer	Assembly	Other: k-mer counts					x	[65]
P-Clouds	K-mer	Assembly	Other: k-mer clusters			x		x	[66]
phRAIDER	Spaced k-mer	Assembly	Families			x		x	[67]
RAP	Spaced k-mer	Assembly	Instances					x	[68]
ReAS	Read k-mer seed and extend	Reads	Families	x					[69]
RECON	Self-comparison, clustering	Assembly	Families			x		x	[28]
RED	K-mer, supervised learning	Reads, or assembly	Instances					x	[70]
RepAHR	Read filtering, and assembly	NGS reads	Families	x				x	[71]
RepARK	K-mer assembly	NGS reads	Families	x					[72]
REPdenovo	K-mer assembly, contig assembly	NGS reads	Families	x					[73]
RepeatExplorer2	Read sampling, clustering and assembly	NGS reads	Families	x			x		[74]
RepeatFinder	MEMs	Assembly	Families		x			x	[26]
RepeatScout	K-mer seeded multiple alignment	Assembly	Families	x					[29]
RepLong	Read clustering and assembly	SMS reads	Families	x					[75]
TE_finder	Self-comparison, clustering	Assembly	Families			x			[76]
Tedna	K-mer, de-bruijn graph	NGS reads	Families	x					[77]
Vmatch/REPuter	MEMs	Assembly	Other: repeat pairs					x	[27]
WindowMasker	K-mer	Assembly	Instances					x	[78]

De novo methods that process whole-genome assemblies typically employ a self-comparison or (spaced) k-mer seeding approach. K-mer approaches identify over-represented exact *k*-length subsequences (k-mers) or *k*-length subsequences with a fixed pattern of match/mismatch positions (spaced k-mers) in an input sequence (Figure 2). These k-mer counts may be simply reported at every position (mer-engine), or thresholded to identify ranges of repetitive sequences (RAP, WindowMasker). The RED tool first identifies repetitive regions using spaced k-mer abundance, trains a classifier on these regions, and finally uses the classifier to annotate the genome. phRAIDER tiles spaced k-mers into maximal approximate matches (MAMs) identifying families of approximate repeats without indels. RepeatScout identifies abundant k-mers and employs them as seeds for a multiple sequence alignment extension and consensus generation. Finally, P-Clouds takes a statistical approach to cluster the k-mers with sequence overlap into groups (clouds). Regions showing significant coverage by k-mers present in one cloud are then considered repetitive and are annotated with that cloud.

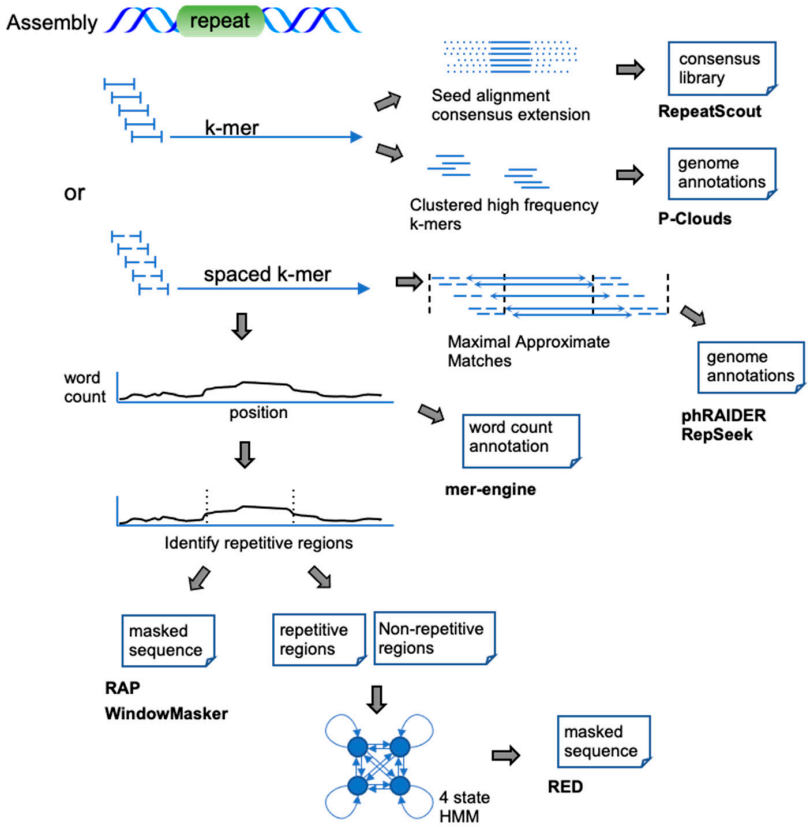

Figure 2. K-mer-based approaches on sequence assemblies. Upon characterizing the k-mer composition of the assembly, the word counts are either: simply used to annotate each base of the sequence (mer-engine), used to discriminate regions of high repetitiveness (RAP, WindowMasker, RED), clustered (P-Clouds), or used as anchors in a seed and extension process (RepeatScout, phRAIDER, RepSeek).

Self-comparison approaches identify repetitive regions using computationally intensive alignment algorithms (Figure 3) followed by clustering strategies to resolve TE families from the pairwise alignment data. Accurate clustering of these alignments is challenging due to high fragmentation and mosaicism present in TE families. Grouper approaches this problem by applying single-linkage clustering, an agglomerative clustering technique that merges two clusters based on the shortest distance between any two members. RECON first applies single-linkage clustering, and then evaluates significant groupings of sequence endpoints within these clusters to identify composite sequences which are split apart accordingly. PILER uses several independent clustering approaches to identify tandem, local, and interspersed repeats from self-comparison alignment data. The interspersed repeat PILER method is also used in the CARP tool. It identifies clusters of aligned sequences that can be considered globally alignable.

Many *de novo* methods have been developed that directly operate on next-generation sequencing (NGS) or single-molecule sequencing (SMS) reads. The primary advantage of this approach is to avoid assembler biases that often cause low-divergence repetitive sequences to be mis-assembled or left out entirely. The predominant approaches (Figure 4) attempt to treat repetitive sequence reads as a special case of sequence assembly or employ clustering methods to group reads/k-mers directly into repetitive families. In both cases, reads enriched for repetitive sequences are obtained either by downsampling the read dataset (to 0.1–0.5x coverage) or by filtering reads composed mostly of low-frequency k-mers. A few methods apply the assembly/clustering strategies directly to the k-mers rather than the reads.

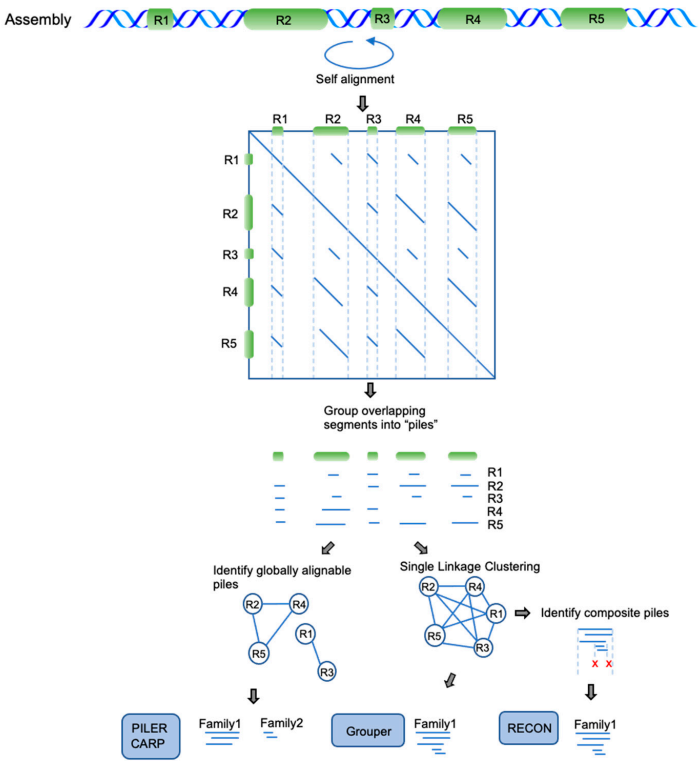

Figure 3. Self-comparison approaches. These methods attempt an all-vs.-all self-alignment using the whole assembly or a portion thereof. The self-alignments, viewed as a dot plot, will have many off-diagonal alignments representing dispersed similarities. These methods group the alignments into "piles", defined by their distinct coverage across a region of the assembly. The primary difference between methods is in how they group piles into families. PILER and CARP require that elements are globally alignable, thereby identifying R1/R3 as a distinct family rather than fragments. Grouper and RECON apply single-linkage clustering, which, in this example, groups all fragments into a single family. RECON further attempts to identify composite families by looking for overrepresented internal edges—in this example, the internal edges were not deemed significant (red x's).

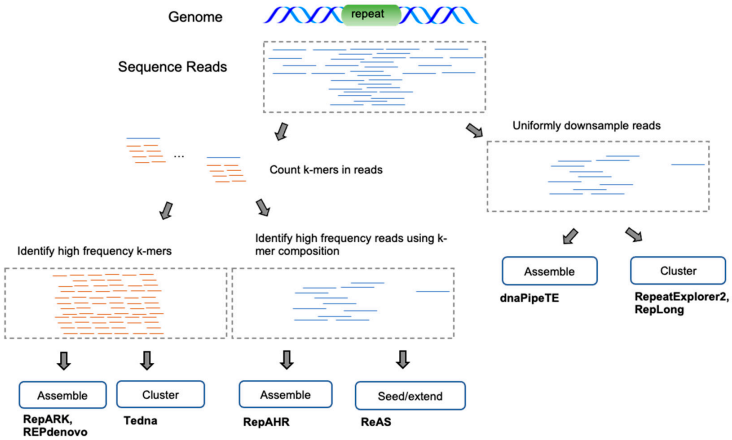

Figure 4. Read-based *de novo* methodologies. Due to the overwhelming size of read datasets, methods often start by either downsampling or filtering low-coverage regions based upon read k-mer frequencies. At this stage, either the remaining reads or the k-mers themselves are assembled into contigs or clustered into distinct groups representing repetitive families.

5. Signature-Based Methodologies

Purely *de novo* methods should be able to detect all classes of TE families. However, detecting TEs by sequence repetition alone has the potential to miss low-copy or certainly

single-copy members of any well-characterized class of TEs and leads to the inclusion of non-TE sequences, such as processed pseudogenes and high-copy gene families. Without expectations regarding the structure of mobile elements, these methods also produce many fragmented or overextended TE models. LTR elements are particularly vulnerable to this, and the output of *de novo* programs may contain (fragments of) solo LTRs, single LTRs with a fragment of an internal sequence on either or both sides, all the way up to LTR-int-LTR-int-LTR structures. Signature-based methods are less susceptible to these particular problems.

Signature-based methods (Table 2) identify TE instances (Figure 5) by recognizing features of specific classes of TEs (terminal inverted repeats, direct repeats, transcription factor binding sites, protein motifs, etc.) as well as hallmarks of TE insertions, such as target site duplications. Often, several features must be used in concert to overcome the low specificity of each; even still, signature-based methods typically suffer from high false-positive rates.

Table 2. Signature-based methods.

Tool	Repeat Types/Classes	Approach	Granularity	Library Generation				Annotation	Ref.
			Families, Instances or Other	Consensi	Exemplars	Other	Provenance		
TESeeker	Protein-coding TEs	Homology based.	Families	x					[79]
Generic Repeat Finder	TIR, LTR, Interspersed	Complex-number sequence encoding/comparison. Identifies direct/inverted repeats with low-frequency indels and mismatches.	Families	x			x	x	[80]
ReDoSt	DIRS	Homology detection to RT, MT and YR protein domains.	Instances					x	[81]
DARTS	LTR	Homology detection to TE-specific subset of the NCBI CDD database. LTR detection and clustering of families.	Families		x		x		[82]
LTR_FINDER	LTR	Suffix array seed and chain strategy for locating intact LTRs. Further validation of TG . . . CA box, PBS, TSR and RT domains.	Instances					x	[83]
LTR_par	LTR	Suffix array for finding MEMs within a constrained distance followed by flanking seq alignment. TSR and TG...CA box validation.	Instances					x	[84]
LTR_STRUC	LTR	Greedy alignment extension of inexact sequence matches.	Instances					x	[85]
LtrDetector	LTR	Nearest k-mer pair distances are chained into LTR pairs. Filters pairs that exhibit TIRs. Validates TSD, TG...CA box, and PPT.	Instances					x	[86]

Table 2. *Cont.*

Tool	Repeat Types/Classes	Approach	Granularity	Library Generation				Annotation	Ref.
			Families, Instances or Other	Consensi	Exemplars	Other	Provenance		
LTRdigest	LTR	*De novo* result polisher that identifies PPT, PBS, protein domains. Feature-based clustering of results into groups.	Families		x			x	[87]
LTRharvest	LTR	Enhanced suffix array for finding MEMs seeds, followed by alignment extension. Validated by distance constraints and by presence of TSD.	Instances					x	[88]
TE-Learner	LTR	Detection using homology to known LTR proteins followed by a machine learning approach for the classification of LTR results.	Instances					x	[89]
MGEScan	LTR/LINE	LTR identification using suffix array, followed by protein domain validation, and clustering into families. Non-LTR identification using RT/APE signal detection and 12-way classifier based on detailed protein domain structure.	Instances					x	[90]
RetroTector	LTR/ERV	Identifies candidates using a set of enriched ERV hexamers. Identifies poly-A, R-U5 and many other characterized motifs and validates their relative positions.	Instances					x	[91]
SINE_Scan	SINE	An enhanced SINE-Finder polisher that looks for matches to tRNA, 7SLRNA and 5SRNA as well as locates TSDs, poly-A regions. Filters common false positives matching LTR/TIR results and/or not repetitive in the genome.	Families	x				x	[92]
SINE-Finder	SINE	Identifies SINES by scanning for patterns matching: TSD, Box A/B motifs, poly-A and appropriate spacers.	Instances					x	[93]
EAHelitron	Helitron	Identify helitrons by looking for matches to: TC . . . CTAG termini containing a GC-rich hairpin structure	Instances					x	[94]

Table 2. *Cont.*

Tool	Repeat Types/Classes	Approach	Granularity	Library Generation				Annotation	Ref.
			Families, Instances or Other	Consensi	Exemplars	Other	Provenance		
HelitronScanner	Helitron	Identifies terminal structures using a set of local combinational variables.	Instances					x	[95]
MITE-Hunter	TIR/MITE	Identifies TIR and TSDs, clusters sequences into families through MSA, able to discover other short non-autonomous Class 2 TEs.	Families	x				x	[96]
TIR-Finder	TIR/MITE	Suffix tree approach to finding specific or arbitrary TIR/TSDs patterns allowing for mismatches.	Instances					x	[97]
MITE Digger	TIR/MITE	Identifies TIR and TSDs iteratively using self-comparison and masking of discovered families.	Families		x			x	[98]
detectMITE	TIR/MITE	Complex-number scoring and comparison of subsequences can detect TIRs with mismatches only. Lempel-Ziv filter for low complexity sequences. CD-HIT clustering for family detection.	Families		x			x	[99]
MiteFinderII	TIR/MITE	K-mer search for identical/imperfect TIRs. Merged k-mers are filtered for low complexity, absence of TSDs, and score poorly to a model based on known MITEs in RepBase.	Families		x			x	[100]
MITE Tracker	TIR/MITE	Identifies TIR and TSDs using self comparison in small batches using BLAST allowing for mismatches and gaps. Candidates are filtered based on sequence complexity, length/distance characteristics, and copy number after clustering.	Families		x			x	[101]

LTR retrotransposons and non-autonomous DNA transposons (aka MITEs when very short) are particularly suited to this approach due to the presence of long direct and inverted repeats flanking intact copies, respectively. Fast computational approaches have been developed to identify generic locally duplicated direct or inverted repeats, but this property alone is not sufficient to identify a TE instance. LTR and MITE finders use these methods to identify potential candidate matches, which are then further evaluated for the presence of target site duplications (short genomic sequences duplicated at the time of insertion), non-repetitive flanking sequences (e.g., not part of a larger repetitive element), and the presence of motifs/protein domains.

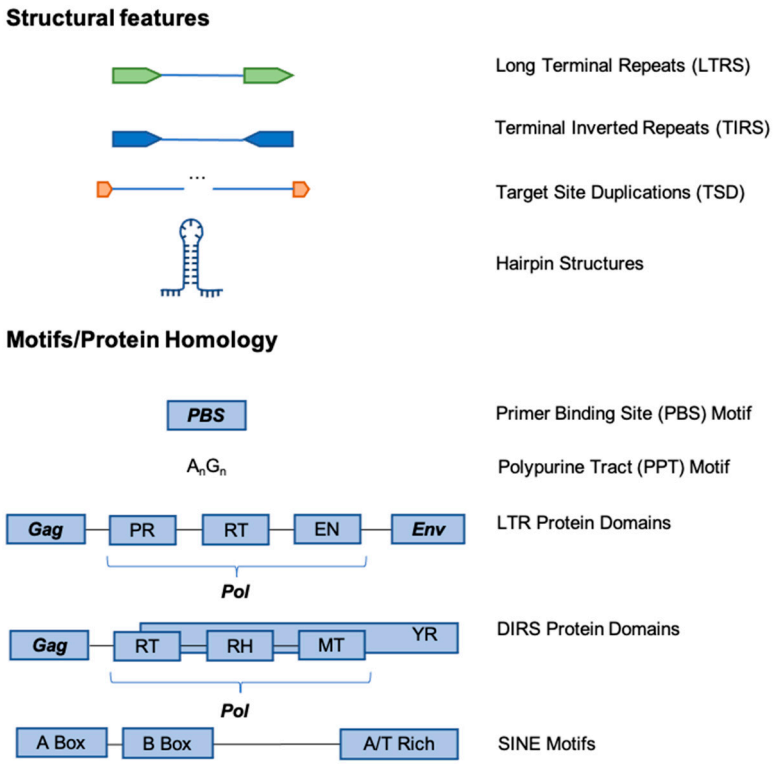

Figure 5. Examples of commonly used TE Signatures for Detection. Structural features: the identification of LTR/ERV elements, class II elements, non-LTR retrotransposable elements, and Helitrons can be achieved by searching for LTRs (~100–1000 bp direct repeats), TIRs (~10–40 bp inverted repeats), TSDs (6–21 bp on average duplications), and hairpin structures, respectively. In addition, the A and B boxes seen in RNA polymerase III promoters and 3′ terminal A/T-rich sequence can be used to identify SINE elements. Motifs/Protein Homology: the order, orientation, and similarity to protein domains is key to homology-based searches. Gag: group-specific antigen; PR: pathogenesis-related; RT: reverse transcriptase; EN: endonuclease; Env: envelope; RH: ribonuclease H; MT: methyltransferase; YR: tyrosine recombinase. Other sequence structures (not seen in the figure above) observed in LINEs are their poly-A or simple-repeat tails, and the RT and apurinic–apyrimidinic EN (APE) domains of the Pol protein.

6. TE Discovery Pipelines

A TE discovery pipeline is defined herein as a combination of previously published or discrete *de novo* algorithms to comprehensively describe all TE classes within any given genome. A pipeline represents a protocol for the orchestration of various tools and their parameters, often augmented with algorithms for clustering TE instances, defragmentation, sequence modeling, and the reduction in false positives and redundancy. The strategies of several popular pipelines are outlined in Figure 6, and further detailed in Supplementary Materials.

Integrating a variety of tools into a single pipeline is an effective way to overcome the shortcomings of any one particular approach; however, it also introduces its own set of challenges. For instance, the outputs of different tools may substantially overlap with each other requiring complex adjudication strategies to eliminate the redundancy. This process is complicated by the natural fragmentation and high sequence divergence present in many TE instances. Managing the overall false-positive rate is a further challenge when integrating several discovery approaches, as each additional tool will contribute its own distinct set of false positives. Finally, each additional method requires evaluation of a larger set of possibly dependent parameters for the overall process.

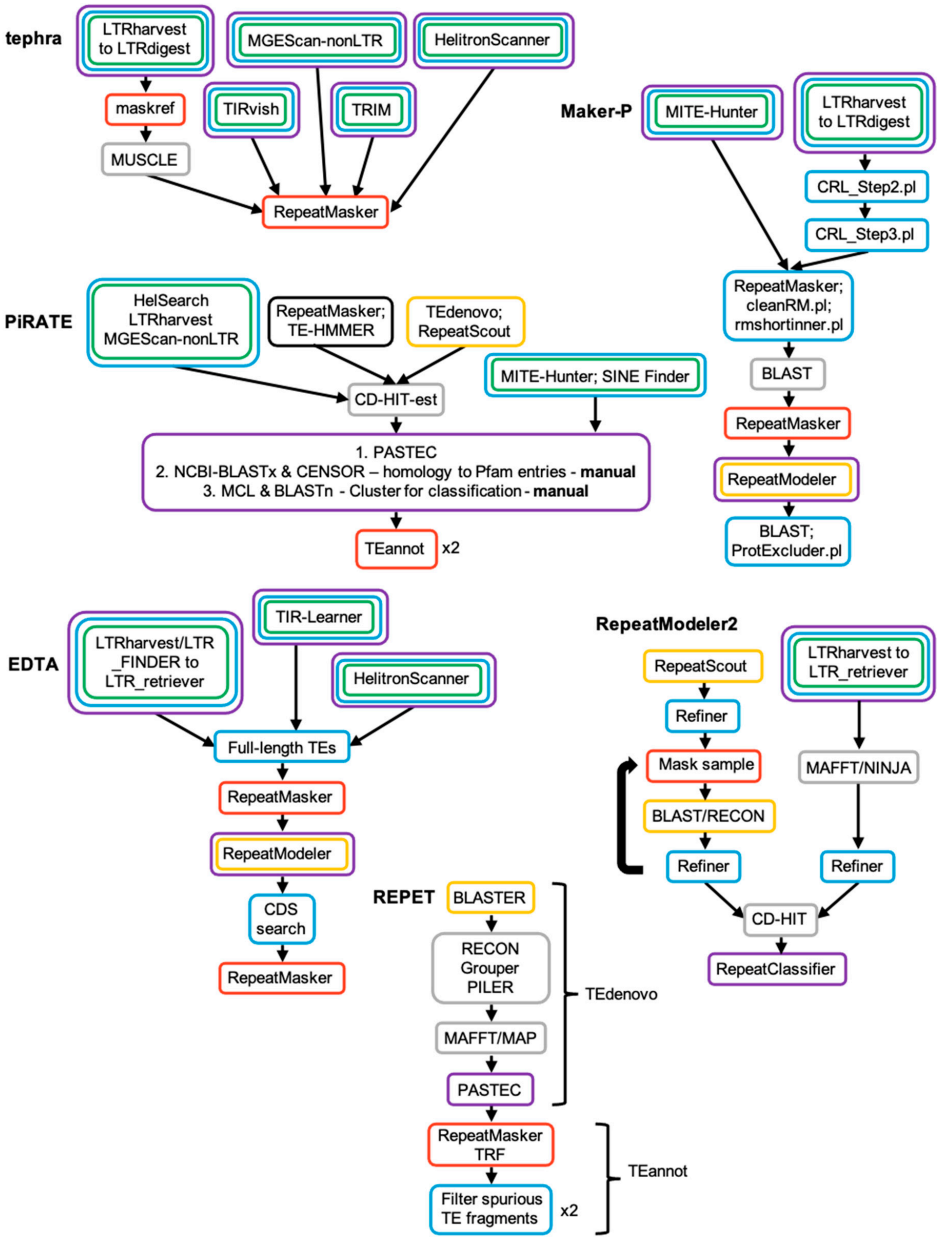

Figure 6. Workflow of select TE discovery pipelines. Each process in the pipeline has been categorized as classification (purple), signature-based TE detection (green), *de novo* TE detection (gold), homology-based detection (black), genome annotation (red), filter and/or refinement (blue) and clustering (grey). Arrows indicate the general workflow direction. NOTE: the above image is meant to describe the high-level organization of each pipeline, and does not reflect the inherent complexity contained within. Refer to Supplementary Materials for additional details.

Various strategies have been employed to produce either a library consisting of a unique set of TE family exemplars or consensi, or of distinct genomic instances. To that end, pipelines utilize clustering methods to collapse similar instances into families, or even redundant families into a single entry. One approach is to use fast sequence clustering algorithms such as CD-HIT-est for this purpose, efficiently grouping sequences with high sequence similarity (>75% sequence identity [102]) (PiRATE, RepeatModeler). A similar approach uses alignment tools (BlastN, Vmatch, etc.) to more accurately assess sequence similarity (albeit less efficiently) and cluster the pairwise sequence distances using single-linkage or complete linkage clustering techniques (EDTA, MAKER-P, tephra, REPET). In addition, pipelines that use sequential discovery and masking stages avoid the inter-

tool clustering problem altogether (RepeatModeler). This is an area that is likely to see improvement in coming years as novel sequence distance estimation [103] and clustering techniques [104] are evaluated in the context of TE families.

The *de novo* tools invoked by a pipeline may have varying levels of false positives, including matches to coincidental groupings of low-specificity sequence signatures, inclusion of sequences in segmental duplications, low-complexity/tandem sequences, or identification of gene families. Combining the results of multiple tools compounds these problems. One approach to reduce false positives has been to consider the flanking or partially flanking regions of instances, filtering those that demonstrate a level of repetitiveness either genome-wide (evidence that either edge is part of a larger repeat), or between two edges (indicating that the extents of the repeat were not fully recognized) (EDTA, MAKER-P). Filtering such instances may be effective in reducing false positives but may also catch true instances. An approach that specifically targets false positives induced by low-complexity sequences and tandem repeats is to pre-mask these regions prior to running discovery tools and only restore them if they are found to be flanked by repetitive sequences (RepeatModeler, REPET).

De novo methods often produce family definitions representing mere fragments of a full-length family. In many cases, more than one fragment is present, representing different regions of the same TE family and creating a problem when they are not recognized as such. Family fragmentation produces inaccurate estimates of families and their abundance, often hampers correct classification of the family, and produces confusing nestings of annotation. However, popular pipelines have yet to tackle this crucial problem.

Since the final output of a TE discovery pipeline may consist of instances, exemplar sequences, and/or consensus sequences from a variety of underlying algorithms, it is important to be aware of the relative limitations of the different data types and appropriate uses. Given perfectly random, neutral decay, and a consensus sequence that precisely matches the ancestral TE, the substitution level of an instance from a consensus is twice as low as that from the average other instance. While this ideal situation is not always met, alignments against exemplars will give an average higher divergence of the TE family, resulting in a higher estimate of the age of the TEs. The sensitivity of the alignments is also reduced, drastically so when the actual average substitution level is over 15–20%. Some tools may provide outputs of the intermediate results or provide the multiple sequence alignments (seed alignments) for the derived TE family consensi. The latter provides a useful definition for the family and from which a consensus may be further improved or other forms of sequence modeling, such as profile Hidden Markov Modeling (pHMMs) may be applied.

The pipelines discussed here have been evaluated to varying degrees on non-model organisms, distantly related species to those used while developing the pipeline, or on species harboring differing TE content, which may represent a wider range of sequence divergence. This is often attributed to the different compositions of TE classes within different species. Therefore, the appropriate pipeline and associated strengths and weaknesses should be considered before beginning any genome analysis. Unfortunately, it can be difficult to make a direct comparison between even *de novo* tools, and much more so for pipelines. In particular, there is not a clear standardized or widely adopted benchmarking method for comparing the relative quality of libraries or genome annotations.

7. Benchmarking

The high diversity of benchmarking approaches applied to TE discovery is a barrier to both the understanding of the true performance of a method and to the competitive evaluation of methods. While this issue has been previously identified [105], a universal approach to this problem has yet to be developed. Of the many benchmarking methods, the comparison of results to existing highly curated libraries or genome annotations (i.e., gold standard dataset) has been the most frequently employed method for assessing true positives (TP) and false negatives (FN). The gold standard used is dependent upon the

product and/or goal of the program in question. For example, if the products are consensi, these are typically compared to Repbase [106] or Dfam [107] consensi in their entirety or to a random sample. For algorithms targeting one type of TE (e.g., LTR elements), these elements are extracted from the Repbase consensi based on the criteria the authors have set and compared to the program output. Alternatively, two common options are utilized if genome annotation is the output and/or goal: (1) the generated library and the Repbase library are each used for RepeatMasker runs, and the loci compared or (2) loci obtained from previously published data are utilized for comparison. The main assumptions when using gold standards (e.g., Repbase, RepeatMasker annotations, or previously published data) are that these data are complete and accurate.

Similarly, the assessment of false positives (FP) is often accomplished by running a given method on a randomized, shuffled, masked or simulated genomic sequence in which any result is necessarily false. The simplest approaches, randomizing bases or shuffling words, do not maintain the complexity of the genome (e.g., maintaining isochores and commonly repeated k-mers) which can produce a less challenging benchmark. Masking out known elements has the advantage of preserving natural background sequences and other non-target genomic elements, but assumes that the masking process is ideal and no true copies of the target are present. Sequence simulation is an attempt to use sequence models, trained on the natural sequence, to generate sequences with similar complexity. The GARLIC [108] tool is an example of this last approach, in which a model trained on the genome is used to generate sequences with the addition of simple/tandem repeats to create a realistic background sequence. In addition, the inclusion of simulated instances generated from TE consensus sequences and fragmented/mutated to natural divergence levels allows the same benchmark to be used to comprehensively assess TP, FN and FP results. Simulation is particularly well suited to repetitiveness-based *de novo* algorithms, but may be less appropriate for programs that detect intra-TE signatures such as TSDs without extra treatment in the simulation.

In addition to assessing their newly generated algorithm, authors may compare their approach to similar tools. In these cases, the most common metrics include a copy number comparison between genome annotations, and/or comparison of the number of models generated, the length of the sequences generated as part of the program output and the N50 of the library. Such metrics promote the idea that more is better. However, this notion does not take into account the quality of the dataset produced.

8. Cleaning Up a TE Library

While *de novo* programs are sometimes used to directly annotate genomes, a careful comparison against a pipeline's complete library or the combination of the output with previously established models has the advantage that the best match can be chosen between two or more related models. Other advantages of the use of libraries are reproducibility, provenance, and the possibility of incremental improvements.

An ideal library would contain only full-length models of all significantly distinct TEs that have left copies in a genome. Full-length models would not only make annotations easier to interpret and allow reconstruction of evolutionary events, important for, e.g., phylogenetic analysis as pointed out in the introduction, but also avoid unfair competition between more or less complete models of related TEs. While such an ideal may never be reached, all automatically created TE libraries currently still need extensive editing before they can be accepted in curated databases such as Dfam or Repbase. The work involved is so intense that the great majority of Dfam submissions is currently housed in a non-curated section [109]. Libraries can always be improved, and many imperfections that seem hard to address automatically may be fixed in updated TE repositories after manual intervention.

There are some significant common library deficiencies that pipelines can ameliorate with additional filters or modules. These include extensive redundancy, and the presence of false positives, artifacts and genic DNA. Below, we also briefly discuss the generation of relevant (sub)family models, but do not address complex problems such as filtering

composite artifacts, identifying overextensions and finishing or merging fragmentary models. These are largely open problems for automated methodologies and still require extensive manual curation to identify and remedy.

While some pipelines (soft) mask simple repeats before *de novo* analysis, the output from all still includes many low complexity sequences with degenerate simple repeats at their cores. Entries that are almost entirely masked by a combination of low complexity and simple repeat finding programs could generally be dismissed automatically at the end of a pipeline. Even if some simple repeats, such as telomeres, are interesting to annotate, genome annotation programs tend to include a separate tandem repeat finding module that would identify these before comparison to the library.

The output of signature-based programs usually contains a number of false positives comprising random genomic sequence. Their uniqueness is a red flag noted by some pipelines, but should be weighed against the evidence, as low/single-copy active TEs are often of considerable interest. *De novo* programs also produce entries looking like random genomic DNA. These may be long 3′ UTRs of processed pseudogenes or, more often, fragments of segmental duplications. In seed alignments, they show a lack of defined ends and true TE instances within them show up as short, dispersed regions with a high number of seeds aligned. Lacking fully automated interpreters of seed alignments as yet, pipelines could instead mark entries as possible segmental duplications if they match multiple other library entries with various classification and if the genome annotation step, already part of most pipelines, shows them to be primarily localized in distribution.

Models that represent signature-based false positives or segmental duplications can contain coding regions of cellular genes. These are unwelcome in repeat libraries, if only for the natural tendency of most researchers to ignore DNA annotated as repetitive. All *de novo* methods also create (partial) models of highly expressed mRNAs from the interspersed processed pseudogenes in a genome, and often create models for common conserved or tandemly repeated protein domains, zinc-finger motifs being a stalwart. While some pipelines offer filtration of genic DNA, these rely on user-supplied protein libraries or repeat-free genomic DNA, the quality of which is critical and not easy to achieve. Furthermore, proper TE models may be dismissed by distant homologies to cellular proteins. Instead, pipelines could offer a competitive comparison to a curated TE protein database [110] and a domain database such as Pfam [111] from which TE protein domains have been removed. TEs may carry (fragments of) cellular genes along, so this filter should be conservative.

Each discovery program has a tendency to produce redundant but non-identical entries for some TE families, especially when they are abundant. For a primate genome, for example, dozens of generic *Alu* models are built. A major drawback of using multiple programs through a pipeline is rediscovery of the same family, resulting in a dramatic increase in redundancy. This causes confusing genome annotation, with instances of the same TE receiving different labels, and more false-positive matches. This is not a small problem; in our experience automatically produced libraries are often more than twice too large. Most pipelines include a clustering step to reduce redundancy, but this usually only involves one class of elements. Defining a redundant set and choosing the best representatives is not a straightforward task, especially since many programs do not preserve the seeds and alignments that led to each model. Simple rules, such as merging all models for which the consensus sequences are 80% similar over 90% of the length, generally do not suffice. Under those restrictions, good models representing distinct subfamilies of a class I element may be lost, while bad models of the same TE (with significant errors, many ambiguous bases and/or long false extensions) will be retained. For *Alu*, the redundant models are often quite diverged, because the many, highly mutagenic CpG sites have been called differently to TG or CA in each. As many models are incomplete, two models with long non-aligned extensions on opposite sites may very well represent a single TE family but will not be joined following such rules.

On top of the problems of recognizing genuine duplicates, trying to define what constitutes a TE family using set cutoffs is impracticable. For phylogeneticists, recognition of the youngest, potentially dimorphic subfamilies of a TE is important. However, if all instances more than 80% similar over 80% of their length should constitute a single family and model [112], young branches of a long-term resident class I TE would go unnoticed as they will be absorbed by a general model. Additionally, still relatively young elements with a > 10% substitution level, would never be built, since most or all of their copies are more than 20% diverged from each other. Generation of one model for many related TEs in general leads to an overestimate of the divergence and therefore age of the copies, which can confuse evolutionary analyses. For older TEs, it also leads to a significant loss in sensitivity during annotation.

Our hands-on strategy to reduce redundancy is to combine the seeds of all interrelated models and perform subfamily analysis using Coseg when all models involved align over much of their length, or create subfamily models with a CD-HIT based code when relationships are partial or modular [113–115]. Such a strategy could be incorporated in pipelines.

As curators of Dfam, we are also acutely aware of redundancy between libraries of related genomes. For example, most of the TEs described in the human genome have left copies that are (originally) shared between all placental mammals. The majority of a library constructed for a slow-evolving genome such as that of a rhinoceros will match those ancestral elements and would form redundant entries (as all ancestral models are or should be included in the analysis of a genome). Since very similar TEs have been active in separate lineages of mammals, sequence similarity is not the right basis for detecting this redundancy. Instead, observation of their presence at orthologous sites outside the lineage is key [116]. Fortunately, as mentioned in the introduction, in the near future, so many species will have been sequenced that most genomes can be studied in the context of multi-species genome alignments. In those settings, not only can it be quickly determined if TE copies are ancestral to two species, but one may avoid rediscovering ancestral TEs by focusing on those parts of a genome that are lineage specific. Many other advantages of such subtractive TE detection, which is as old as the publication of the first aligned mammalian genomes [117–119], suggest that this method may be the future of *de novo* TE detection.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/genes13040709/s1. Supplementary material is an Excel spreadsheet detailing the TE discovery approaches (including the algorithms used for signature-based, homology-based, *de novo*-based TE discovery, as well as filtering, clustering, classification and annotation methods, in addition to the level of expertise required, program versions, and the output of the pipeline) presented in Figure 6 in the main text.

Author Contributions: J.M.S., R.H., J.R. and A.F.A.S. contributed to all aspects of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Human Genome Research Institute grant grants # U24 HG010136 (A.F.A.S.) and # RO1 HG002939 (A.F.A.S.).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The supplementary data files are available on the online version of this paper.

Conflicts of Interest: The authors declare no conflict of interest.

Glossary

Transposable Element (TE) (n): A mobile DNA sequence evolved to replicate within and throughout genomes independently of the host cell DNA.
Tandem Repeat (n): A (possibly degraded) sequence pattern that is repeated directly adjacent to each other.
Satellite DNA (n): Tandem repeats in large arrays that can span megabases and may form essential chromosome structures such as centromeres and telomeres.
Interspersed Repeat (IR) (n): Any, sometimes highly degraded, sequence pattern of paralogous origin that is repeated a handful to millions of times at mainly non-adjacent places throughout a genome. Most IRs are copies of TEs at various levels of decay, but, e.g., processed pseudogenes and gene families can also be considered IRs. Simple tandem repeats appear all over a genome as well, but are not of paralogous origin.
Segmental Duplications (n): Continuous portions of DNA lacking defined ends that map to two dozen genomic locations. Typically ranging in size from 1 to 200 kb, they are composed of apparently normal genomic DNA, may contain TE copies and (fragments of) cellular genes, and can occur both interspersed and in tandem. Their distribution often appears localized to (former) pericentromeric and subtelomeric regions, but copies may occur anywhere in a genome.
Family/Subfamily (n): A collection of similar genomic subsequences produced by a single biological process.
Instance (n): An individual fragment or full-length copy of a family in a genome.
Exemplar (n): Exemplars (aka prototypes) are one or more genomic instances of a family that are used to represent the family in a library.
Source gene (n): TE copy that gave rise to a (group of) particular TE instances.
Consensus sequence (n): A nucleotide sequence representation of a family, often simply calculated as an average of aligned instances, but ideally approaching the original sequence of the source gene.
Profile Hidden Markov Model (pHMM) (n): A model of a repetitive sequence family which encodes the relative observed frequencies of nucleotide matches, insertions, and deletions at each position in a multiple sequence alignment.
Sequence Model (n): A summary representation of a set of nucleotide or protein sequences, typically a consensus sequence, a sequence profile, or a profile Hidden Markov Model.
Classification, classification system (n): A hierarchy or other arrangement of transposable element families, defined by phylogenetic, structural, or other criteria.
Classify (v): To determine and assign the classifications of transposable element families.
Genome annotation (n,v): A labeling of a genome with the locations, names, classifications, and other properties of genomic features such as simple repeats and transposable elements.
Clustering (v): To group sequences or sequence models together, usually based in some way on the relative similarity between sequences.
Multiple sequence alignment (MSA) (n): A collection of sequences, in which each sequence is aligned to the others and/or to a "reference" sequence based on the similarity of each sequence to the others.
Seed alignment (n): A multiple sequence alignment of representative instances of a TE family which can be used to build a model.
Pipeline (n): A comprehensive strategy with which to discover, classify, and annotate all classes of transposable elements in a given genome assembly; compilations of previously published algorithms, supplemented with algorithms for clustering, false-positive reduction, etc.
Single-linkage clustering (v): A hierarchical clustering algorithm that forms clusters on the basis of the minimum distances between pairs.
Complete-linkage clustering (v): A hierarchical clustering algorithm that forms clusters on the basis of the maximum distances between pairs.
Maximal Exact Match (MEM) (n): An exact match of two subsequences of any length that cannot be extended in either direction without introducing a mismatch.

Maximal Approximate Match (MAM) (n): An approximate match between two subsequences of any length that cannot be extended in either direction without introducing an additional mismatch that exceeds the total mismatches allowed.

References

1. Schnable, P.S.; Ware, D.; Fulton, R.S.; Stein, J.C.; Wei, F.; Pasternak, S.; Liang, C.; Zhang, J.; Fulton, L.; Graves, T.A.; et al. The B73 maize genome: Complexity, diversity, and dynamics. *Science* **2009**, *326*, 1112–1115. [CrossRef] [PubMed]
2. International Barley Genome Sequencing Consortium; Mayer, K.F.; Waugh, R.; Brown, J.W.; Schulman, A.; Langridge, P.; Platzer, M.; Fincher, G.B.; Muehlbauer, G.J.; Sato, K.; et al. A physical, genetic and functional sequence assembly of the barley genome. *Nature* **2012**, *491*, 711–716.
3. Meyer, A.; Schloissnig, S.; Franchini, P.; Du, K.; Woltering, J.M.; Irisarri, I.; Wong, W.Y.; Nowoshilow, S.; Kneitz, S.; Kawaguchi, A.; et al. Giant lungfish genome elucidates the conquest of land by vertebrates. *Nature* **2021**, *590*, 284–289. [CrossRef] [PubMed]
4. Doolittle, W.F.; Sapienza, C. Selfish genes, the phenotype paradigm and genome evolution. *Nature* **1980**, *284*, 601–603. [CrossRef] [PubMed]
5. Smit, A.F.A. The origin of interspersed repeats in the human genome. *Curr. Opin. Genet. Dev.* **1996**, *6*, 743–748. [CrossRef]
6. Lander, E.S.; Linton, L.M.; Birren, B.; Nusbaum, C.; Zody, M.C.; Baldwin, J.; Devon, K.; Dewar, K.; Doyle, M.; FitzHugh, W.; et al. Initial sequencing and analysis of the human genome. *Nature* **2001**, *409*, 860–921.
7. Smit Arian, A.F.; Hubley, R. RepeatMasker Human Genome Dataset. 2013. Available online: http://www.repeatmasker.org/species/hg.html (accessed on 16 April 2022).
8. Doronina, L.; Reising, O.; Clawson, H.; Ray, D.A.; Schmitz, J. True Homoplasy of Retrotransposon Insertions in Primates. *Syst. Biol.* **2019**, *68*, 482–493. [CrossRef]
9. Nikaido, M.; Rooney, A.P.; Okada, N. Phylogenetic relationships among cetartiodactyls based on insertions of short and long interpersed elements: Hippopotamuses are the closest extant relatives of whales. *Proc. Natl. Acad. Sci. USA.* **1999**, *96*, 10261–10266. [CrossRef]
10. Salem, A.-H.; Ray, D.A.; Xing, J.; Callinan, P.A.; Myers, J.S.; Hedges, D.J.; Garber, R.K.; Witherspoon, D.J.; Jorde, L.B.; Batzer, M.A. Alu elements and hominid phylogenetics. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 12787–12791. [CrossRef]
11. Roos, C.; Schmitz, J.; Zischler, H. Primate jumping genes elucidate strepsirrhine phylogeny. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 10650–10654. [CrossRef]
12. Nishihara, H.; Maruyama, S.; Okada, N. Retroposon analysis and recent geological data suggest near-simultaneous divergence of the three superorders of mammals. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 5235–5240. [CrossRef] [PubMed]
13. SSuh, A.; Paus, M.; Kiefmann, M.; Churakov, G.; Franke, F.A.; Brosius, J.; Kriegs, J.O.; Schmitz, J. Mesozoic retroposons reveal parrots as the closest living relatives of passerine birds. *Nat. Commun.* **2011**, *2*, 443. [CrossRef] [PubMed]
14. Armstrong, J.; Fiddes, I.T.; Diekhans, M.; Paten, B. Whole-Genome Alignment and Comparative Annotation. *Annu. Rev. Anim. Biosci.* **2019**, *7*, 41–64. [CrossRef] [PubMed]
15. Zoonomia Consortium. A comparative genomics multitool for scientific discovery and conservation. *Nature* **2020**, *587*, 240–245. [CrossRef]
16. Feng, S.; Stiller, J.; Deng, Y.; Armstrong, J.; Fang, Q.; Reeve, A.H.; Xie, D.; Chen, G.; Guo, C.; Faircloth, B.C.; et al. Dense sampling of bird diversity increases power of comparative genomics. *Nature* **2020**, *587*, 252–257. [CrossRef]
17. Gundappa, M.K.; To, T.-H.; Grønvold, L.; Martin, S.A.M.; Lien, S.; Geist, J.; Hazlerigg, D.; Sandve, S.R.; Macqueen, D.J. Genome-Wide Reconstruction of Rediploidization Following Autopolyploidization across One Hundred Million Years of Salmonid Evolution. *Mol. Biol. Evol.* **2022**, *39*, msab310. [CrossRef]
18. Springer, M.S.; Molloy, E.K.; Sloan, D.B.; Simmons, M.P.; Gatesy, J. ILS-Aware Analysis of Low-Homoplasy Retroelement Insertions: Inference of Species Trees and Introgression Using Quartets. *J. Hered.* **2020**, *111*, 147–168. [CrossRef]
19. Simmons, M.P.; Springer, M.S.; Gatesy, J. Gene-tree misrooting drives conflicts in phylogenomic coalescent analyses of palaeognath birds. *Mol. Phylogenet. Evol.* **2022**, *167*, 107344. [CrossRef]
20. Molloy, E.K.; Gatesy, J.; Springer, M.S. Theoretical and practical considerations when using retroelement insertions to estimate species trees in the anomaly zone. *Syst. Biol.* **2021**, syab086. [CrossRef]
21. Lewin, H.A.; Richards, S.; Aiden, E.L.; Allende, M.L.; Archibald, J.M.; Bálint, M.; Barker, K.B.; Baumgartner, B.; Belov, K.; Bertorelle, G.; et al. The Earth BioGenome Project 2020: Starting the clock. *Proc. Natl. Acad. Sci. USA* **2022**, *119*, e2115635118. [CrossRef]
22. Marks, R.A.; Hotaling, S.; Frandsen, P.B.; VanBuren, R. Representation and participation across 20 years of plant genome sequencing. *Nat. Plants* **2021**, *7*, 1571–1578. [CrossRef] [PubMed]
23. Hotaling, S.; Kelley, J.L.; Frandsen, P.B. Toward a genome sequence for every animal: Where are we now? *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2109019118. [CrossRef] [PubMed]
24. Rhie, A.; McCarthy, S.A.; Fedrigo, O.; Damas, J.; Formenti, G.; Koren, S.; Uliano-Silva, M.; Chow, W.; Fungtammasan, A.; Kim, J.; et al. Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **2021**, *592*, 737–746. [CrossRef] [PubMed]

25. Darwin Tree of Life Project Consortium. Sequence locally, think globally: The Darwin Tree of Life Project. *Proc. Natl. Acad. Sci. USA* **2022**, *119*. [CrossRef]
26. Volfovsky, N.; Haas, B.J.; Salzberg, S.L. A clustering method for repeat analysis in DNA sequences. *Genome. Biol.* **2001**, *2*, research0027.1. [CrossRef] [PubMed]
27. Kurtz, S.; Choudhuri, J.V.; Ohlebusch, E.; Schleiermacher, C.; Stoye, J.; Giegerich, R. REPuter: The manifold applications of repeat analysis on a genomic scale. *Nucleic. Acids. Res.* **2001**, *29*, 4633–4642. [CrossRef] [PubMed]
28. Bao, Z.; Eddy, S.R. Automated *de novo* identification of repeat sequence families in sequenced genomes. *Genome. Res.* **2002**, *12*, 1269–1276. [CrossRef]
29. Price, A.L.; Jones, N.C.; Pevzner, P.A. *De novo* identification of repeat families in large genomes. *Bioinformatics* **2005**, *21* (Suppl. 1), i351–i358. [CrossRef]
30. Edgar, R.C.; Myers, E.W. PILER: Identification and classification of genomic repeats. *Bioinformatics* **2005**, *21* (Suppl. 1), i152–i158. [CrossRef]
31. Lim, K.G.; Kwoh, C.K.; Hsu, L.Y.; Wirawan, A. Review of tandem repeat search tools: A systematic approach to evaluating algorithmic performance. *Brief. Bioinform.* **2013**, *14*, 67–81. [CrossRef]
32. Anisimova, M.; Pečerska, J.; Schaper, E. Statistical approaches to detecting and analyzing tandem repeats in genomic sequences. *Front. Bioeng. Biotechnol.* **2015**, *3*, 31. [CrossRef] [PubMed]
33. Olson, D.; Wheeler, T. ULTRA: A Model Based Tool to Detect Tandem Repeats. In Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, Washington, DC, USA, 29 August–1 September 2018; Association for Computing Machinery: New York, NY, USA, 2018; pp. 37–46.
34. Harris, R.S.; Cechova, M.; Makova, K.D. Noise-cancelling repeat finder: Uncovering tandem repeats in error-prone long-read sequencing data. *Bioinformatics* **2019**, *35*, 4809–4811. [CrossRef] [PubMed]
35. Gao, Y.; Liu, B.; Wang, Y.; Xing, Y. TideHunter: Efficient and sensitive tandem repeat detection from noisy long-reads using seed-and-chain. *Bioinformatics* **2019**, *35*, i200–i207. [CrossRef] [PubMed]
36. Velasco, A.; James, B.T.; Wells, V.D.; Girgis, H.Z. Look4TRs: A *de novo* tool for detecting simple tandem repeats using self-supervised hidden Markov models. *Bioinformatics* **2020**, *36*, 380–387. [CrossRef] [PubMed]
37. Shortt, J.A.; Ruggiero, R.P.; Cox, C.; Wacholder, A.C.; Pollock, D.D. Finding and extending ancient simple sequence repeat-derived regions in the human genome. *Mob. DNA* **2020**, *11*, 11. [CrossRef]
38. Pu, L.; Lin, Y.; Pevzner, P.A. Detection and analysis of ancient segmental duplications in mammalian genomes. *Genome. Res.* **2018**, *28*, 901–909. [CrossRef]
39. Delehelle, F.; Cussat-Blanc, S.; Alliot, J.-M.; Luga, H.; Balaresque, P. ASGART: Fast and parallel genome scale segmental duplications mapping. *Bioinformatics* **2018**, *34*, 2708–2714. [CrossRef]
40. Vollger, M.R.; Dishuck, P.C.; Sorensen, M.; Welch, A.E.; Dang, V.; Dougherty, M.L.; Graves-Lindsay, T.A.; Wilson, R.K.; Chaisson, M.J.P.; Eichler, E.E. Long-read sequence and assembly of segmental duplications. *Nat. Methods* **2019**, *16*, 88–94. [CrossRef]
41. Ewing, A.D. Transposable element detection from whole genome sequence data. *Mob. DNA* **2015**, *6*, 24. [CrossRef]
42. Chu, C.; Zhao, B.; Park, P.J.; Lee, E.A. Identification and Genotyping of Transposable Element Insertions From Genome Sequencing Data. *Curr. Protoc. Hum. Genet.* **2020**, *107*, e102. [CrossRef]
43. Jain, D.; Chu, C.; Alver, B.H.; Lee, S.; Lee, E.A.; Park, P.J. HiTea: A computational pipeline to identify non-reference transposable element insertions in Hi-C data. *Bioinformatics* **2021**, *37*, 1045–1051. [CrossRef] [PubMed]
44. Quadrana, L.; Silveira, A.B.; Caillieux, E.; Colot, V. Detection of Transposable Element Insertions in Arabidopsis Using Sequence Capture. *Methods Mol. Biol.* **2021**, *2250*, 141–155. [PubMed]
45. Chu, C.; Borges-Monroy, R.; Viswanadham, V.V.; Lee, S.; Li, H.; Lee, E.A.; Park, P.J. Comprehensive identification of transposable element insertions using multiple sequencing technologies. *Nat. Commun.* **2021**, *12*, 3836. [CrossRef] [PubMed]
46. Lerat, E. Identifying repeats and transposable elements in sequenced genomes: How to find your way through the dense forest of programs. *Heredity* **2010**, *104*, 520–533. [CrossRef]
47. Flutre, T.; Permal, E.; Quesneville, H. Transposable Element Annotation in Completely Sequenced Eukaryote Genomes. In *Plant Transposable Elements: Impact on Genome Structure and Function*; Grandbastien, M.-A., Casacuberta, J.M., Eds.; Springer: Berlin/Heidelberg, Germany, 2012; pp. 17–39.
48. Bennetzen, J.L.; Park, M. Distinguishing friends, foes, and freeloaders in giant genomes. *Curr. Opin. Genet. Dev.* **2018**, *49*, 49–55. [CrossRef]
49. Goerner-Potvin, P.; Bourque, G. Computational tools to unmask transposable elements. *Nat. Rev. Genet.* **2018**, *19*, 688–704. [CrossRef]
50. Makałowski, W.; Gotea, V.; Pande, A.; Makałowska, I. Transposable Elements: Classification, Identification, and Their Use As a Tool For Comparative Genomics. *Methods Mol. Biol.* **2019**, *1910*, 177–207.
51. Nicolas, J.; Tempel, S.; Fiston-Lavier, A.-S.; Cherif, E. Finding and Characterizing Repeats in Plant Genomes. *Methods Mol. Biol.* **2022**, *2443*, 327–385.
52. Chakraborty, M.; Chang, C.-H.; Khost, D.E.; Vedanayagam, J.; Adrion, J.R.; Liao, Y.; Montooth, K.L.; Meiklejohn, C.D.; Larracuente, A.M.; Emerson, J. Evolution of genome structure in the Drosophila simulans species complex. *Genome. Res.* **2021**, *31*, 380–396. [CrossRef]

53. SanMiguel, P.; Tikhonov, A.; Jin, Y.K.; Motchoulskaia, N.; Zakharov, D.; Melake-Berhan, A.; Springer, P.S.; Edwards, K.J.; Lee, M.; Avramova, Z.; et al. Nested retrotransposons in the intergenic regions of the maize genome. *Science* **1996**, *274*, 765–768. [CrossRef]
54. Bureau, T.E.; Ronald, P.C.; Wessler, S.R. A computer-based systematic survey reveals the predominance of small inverted-repeat elements in wild-type rice genes. *Proc. Natl. Acad. Sci. USA* **1996**, *93*, 8524–8529. [CrossRef] [PubMed]
55. Smit, A.F.A.; Riggs, A.D. Tiggers and other DNA transposon fossils in the human genome. *Proc. Natl. Acad. Sci. USA* **1996**, *93*, 1443–1448. [CrossRef] [PubMed]
56. Smit, A.F. Identification of a new, abundant superfamily of mammalian LTR-transposons. *Nucleic. Acids. Res.* **1993**, *21*, 1863–1872. [CrossRef] [PubMed]
57. Witte, C.P.; Le, Q.H.; Bureau, T.; Kumar, A. Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 13778–13783. [CrossRef]
58. Smit, A.F.A. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.* **1999**, *9*, 657–663. [CrossRef]
59. Vargiu, L.; Rodriguez-Tomé, P.; Sperber, G.O.; Cadeddu, M.; Grandi, N.; Blikstad, V.; Tramontano, E.; Blomberg, J. Classification and characterization of human endogenous retroviruses; mosaic forms are common. *Retrovirology* **2016**, *13*, 7. [CrossRef]
60. Paço, A.; Freitas, R.; Vieira-Da-Silva, A. Conversion of DNA Sequences: From a Transposable Element to a Tandem Repeat or to a Gene. *Genes* **2019**, *10*, 1014. [CrossRef]
61. Arkhipova, I.R.; Yushenova, I.A. Giant Transposons in Eukaryotes: Is Bigger Better? *Genome. Biol. Evol.* **2019**, *11*, 906–918. [CrossRef]
62. Zeng, L.; Kortschak, R.D.; Raison, J.M.; Bertozzi, T.; Adelson, D.L. Superior ab initio identification, annotation and characterisation of TEs and segmental duplications from genome assemblies. *PLoS ONE* **2018**, *13*, e0193588. [CrossRef]
63. Goubert, C.; Modolo, L.; Vieira, C.; ValienteMoro, C.; Mavingui, P.; Boulesteix, M. *De novo* assembly and annotation of the Asian tiger mosquito (*Aedes albopictus*) repeatome with dnaPipeTE from raw genomic reads and comparative analysis with the yellow fever mosquito (*Aedes aegypti*). *Genome. Biol. Evol.* **2015**, *7*, 1192–1205. [CrossRef]
64. Liao, X.; Li, M.; Hu, K.; Wu, F.-X.; Gao, X.; Wang, J. A sensitive repeat identification framework based on short and long reads. *Nucleic. Acids. Res.* **2021**, *49*, e100. [CrossRef] [PubMed]
65. Healy, J.; Thomas, E.E.; Schwartz, J.T.; Wigler, M. Annotating large genomes with exact word matches. *Genome. Res.* **2003**, *13*, 2306–2315. [CrossRef] [PubMed]
66. Gu, W.; Castoe, T.A.; Hedges, D.J.; Batzer, M.A.; Pollock, D.D. Identification of repeat structure in large genomes using repeat probability clouds. *Anal. Biochem.* **2008**, *380*, 77–83. [CrossRef] [PubMed]
67. Schaeffer, C.E.; Figueroa, N.D.; Liu, X.; Karro, J.E. phRAIDER: Pattern-Hunter based Rapid Ab Initio Detection of Elementary Repeats. *Bioinformatics* **2016**, *32*, i209–i215. [CrossRef] [PubMed]
68. Campagna, D.; Romualdi, C.; Vitulo, N.; Del Favero, M.; Lexa, M.; Cannata, N.; Valle, G. RAP: A new computer program for *de novo* identification of repeated sequences in whole genomes. *Bioinformatics* **2005**, *21*, 582–588. [CrossRef]
69. Li, R.; Ye, J.; Li, S.; Wang, J.; Han, Y.; Ye, C.; Wang, J.; Yang, H.; Yu, J.; Wong, G.K.; et al. ReAS: Recovery of ancestral sequences for transposable elements from the unassembled reads of a whole genome shotgun. *PLoS Comput. Biol.* **2005**, *1*, e43. [CrossRef]
70. Girgis, H.Z. Red: An intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. *BMC Bioinformatics* **2015**, *16*, 227. [CrossRef]
71. Liao, X.; Gao, X.; Zhang, X.; Wu, F.-X.; Wang, J. RepAHR: An improved approach for *de novo* repeat identification by assembly of the high-frequency reads. *BMC Bioinform.* **2020**, *21*, 463. [CrossRef]
72. Koch, P.; Platzer, M.; Downie, B.R. RepARK–*de novo* creation of repeat libraries from whole-genome NGS reads. *Nucleic. Acids. Res.* **2014**, *42*, e80. [CrossRef]
73. Chu, C.; Nielsen, R.; Wu, Y. REPdenovo: Inferring *De Novo* Repeat Motifs from Short Sequence Reads. *PLoS ONE* **2016**, *11*, e0150719. [CrossRef]
74. Novák, P.; Neumann, P.; Macas, J. Global analysis of repetitive DNA from unassembled sequence reads using RepeatExplorer2. *Nat. Protoc.* **2020**, *15*, 3745–3776. [CrossRef] [PubMed]
75. Guo, R.; Li, Y.-R.; He, S.; Ou-Yang, L.; Sun, Y.; Zhu, Z. RepLong: *de novo* repeat identification using long read sequencing data. *Bioinformatics* **2018**, *34*, 1099–1107. [CrossRef] [PubMed]
76. Sohrab, V.; López-Díaz, C.; Di Pietro, A.; Ma, L.-J.; Ayhan, D. TEfinder: A Bioinformatics Pipeline for Detecting New Transposable Element Insertion Events in Next-Generation Sequencing Data. *Genes* **2021**, *12*, 224. [CrossRef] [PubMed]
77. Zytnicki, M.; Akhunov, E.; Quesneville, H. Tedna: A transposable element *de novo* assembler. *Bioinformatics* **2014**, *30*, 2656–2658. [CrossRef]
78. Morgulis, A.; Gertz, E.M.; Schäffer, A.A.; Agarwala, R. WindowMasker: Window-based masker for sequenced genomes. *Bioinformatics* **2006**, *22*, 134–141. [CrossRef]
79. Kennedy, R.C.; Unger, M.F.; Christley, S.; Collins, F.H.; Madey, G.R. An automated homology-based approach for identifying transposable elements. *BMC Bioinform.* **2011**, *12*, 130. [CrossRef]
80. Shi, J.; Liang, C. Generic Repeat Finder: A High-Sensitivity Tool for Genome-Wide *De Novo* Repeat Detection. *Plant Physiol.* **2019**, *180*, 1803–1815. [CrossRef]
81. Piednoël, M.; Gonçalves, I.R.; Higuet, D.; Bonnivard, E. Eukaryote DIRS1-like retrotransposons: An overview. *BMC Genom.* **2011**, *12*, 621. [CrossRef]

82. Biryukov, M.; Ustyantsev, K. DARTS: An Algorithm for Domain-Associated Retrotransposon Search in Genome Assemblies. *Genes* **2021**, *13*, 9. [CrossRef]
83. Xu, Z.; Wang, H. LTR_FINDER: An efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic. Acids. Res.* **2007**, *35*, W265–W268. [CrossRef]
84. Kalyanaraman, A.; Aluru, S. Efficient algorithms and software for detection of full-length LTR retrotransposons. *J. Bioinform. Comput. Biol.* **2006**, *4*, 197–216. [CrossRef] [PubMed]
85. McCarthy, E.M.; McDonald, J.F. LTR_STRUC: A novel search and identification program for LTR retrotransposons. *Bioinformatics* **2003**, *19*, 362–367. [CrossRef] [PubMed]
86. Valencia, J.D.; Girgis, H.Z. LtrDetector: A tool-suite for detecting long terminal repeat retrotransposons de-novo. *BMC Genom.* **2019**, *20*, 450. [CrossRef] [PubMed]
87. Steinbiss, S.; Willhoeft, U.; Gremme, G.; Kurtz, S. Fine-grained annotation and classification of *de novo* predicted LTR retrotransposons. *Nucleic. Acids. Res.* **2009**, *37*, 7002–7013. [CrossRef] [PubMed]
88. Ellinghaus, D.; Kurtz, S.; Willhoeft, U. LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. *BMC Bioinform.* **2008**, *9*, 18. [CrossRef] [PubMed]
89. Schietgat, L.; Vens, C.; Cerri, R.; Fischer, C.N.; Costa, E.; Ramon, J.; Cararetto, C.M.A.; Blockeel, H. A machine learning based framework to identify and classify long terminal repeat retrotransposons. *PLoS Comput. Biol.* **2018**, *14*, e1006097. [CrossRef]
90. Lee, H.; Lee, M.; Ismail, W.M.; Rho, M.; Fox, G.C.; Oh, S.; Tang, H. MGEScan: A Galaxy-based system for identifying retrotransposons in genomes. *Bioinformatics* **2016**, *32*, 2502–2504. [CrossRef]
91. Sperber, G.O.; Airola, T.; Jern, P.; Blomberg, J. Automated recognition of retroviral sequences in genomic data—RetroTector. *Nucleic. Acids. Res.* **2007**, *35*, 4964–4976. [CrossRef]
92. Mao, H.; Wang, H. SINE_scan, an efficient tool to discover short interspersed nuclear elements (SINEs) in large-scale genomic datasets. *Bioinformatics* **2017**, *33*, 743–745. [CrossRef]
93. Wenke, T.; Döbel, T.; Sörensen, T.R.; Junghans, H.; Weisshaar, B.; Schmidt, T. Targeted identification of short interspersed nuclear element families shows their widespread existence and extreme heterogeneity in plant genomes. *Plant Cell* **2011**, *23*, 3117–3128. [CrossRef]
94. Hu, K.; Xu, K.; Wen, J.; Yi, B.; Shen, J.; Ma, C.; Fu, T.; Ouyang, Y.; Tu, J. Helitron distribution in Brassicaceae and whole Genome Helitron density as a character for distinguishing plant species. *BMC Bioinform.* **2019**, *20*, 354. [CrossRef] [PubMed]
95. Xiong, W.; He, L.; Lai, J.; Dooner, H.K.; Du, C. HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 10263–10268. [CrossRef] [PubMed]
96. Han, Y.; Wessler, S.R. MITE-Hunter: A program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic. Acids. Res.* **2010**, *38*, e199. [CrossRef] [PubMed]
97. Gambin, T.; Startek, M.; Walczak, K.; Paszek, J.; Grzebelus, D.; Gambin, A. TIRfinder: A Web Tool for Mining Class II Transposons Carrying Terminal Inverted Repeats. *Evol. Bioinform. Online* **2013**, *9*, EBO.S10619. [CrossRef]
98. Yang, G. MITE Digger, an efficient and accurate algorithm for genome wide discovery of miniature inverted repeat transposable elements. *BMC Bioinformatics* **2013**, *14*, 186. [CrossRef]
99. Ye, C.; Ji, G.; Liang, C. detectMITE: A novel approach to detect miniature inverted repeat transposable elements in genomes. *Sci. Rep.* **2016**, *6*, 19688. [CrossRef]
100. Hu, J.; Zheng, Y.; Shang, X. MiteFinderII: A novel tool to identify miniature inverted-repeat transposable elements hidden in eukaryotic genomes. *BMC Med. Genom.* **2018**, *11*, 101. [CrossRef]
101. Crescente, J.M.; Zavallo, D.; Helguera, M.; Vanzetti, L.S. MITE Tracker: An accurate approach to identify miniature inverted-repeat transposable elements in large genomes. *BMC Bioinform.* **2018**, *19*, 348. [CrossRef]
102. Pipes, L.; Nielsen, R. AncestralClust: Clustering of Divergent Nucleotide Sequences by Ancestral Sequence Reconstruction using Phylogenetic Trees. *Bioinformatics* **2021**, *38*, 663–670. [CrossRef]
103. Joudaki, A.; Rätsch, G.; Kahles, A. Fast Alignment-Free Similarity Estimation By Tensor Sketching. *bioRxiv* **2021**. [CrossRef]
104. Girgis, H.Z. MeShClust v3.0: High-quality clustering of DNA sequences using the mean shift algorithm and alignment-free identity scores. *bioRxiv* **2022**. [CrossRef]
105. Hoen, D.R.; Hickey, G.; Bourque, G.; Casacuberta, J.; Cordaux, R.; Feschotte, C.; Fiston-Lavier, A.-S.; Hua-Van, A.; Hubley, R.; Kapusta, A.; et al. A call for benchmarking transposable element annotation methods. *Mob. DNA* **2015**, *6*, 13. [CrossRef] [PubMed]
106. Bao, W.; Kojima, K.K.; Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **2015**, *6*, 11. [CrossRef] [PubMed]
107. Wheeler, T.J.; Clements, J.; Eddy, S.R.; Hubley, R.; Jones, T.A.; Jurka, J.; Smit, A.F.A.; Finn, R.D. Dfam: A database of repetitive DNA based on profile hidden Markov models. *Nucleic. Acids. Res.* **2013**, *41*, D70–D82. [CrossRef] [PubMed]
108. Caballero, J.; Smit, A.F.A.; Hood, L.; Glusman, G. Realistic artificial DNA sequences as negative controls for computational genomics. *Nucleic. Acids. Res.* **2014**, *42*, e99. [CrossRef] [PubMed]
109. Storer, J.; Hubley, R.; Rosen, J.; Wheeler, T.J.; Smit, A.F. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob. DNA* **2021**, *12*, 2. [CrossRef] [PubMed]
110. Smit, A.F.A.; Hubley, R.; Green, P. RepeatMasker Open-4.0. In: RepeatMasker Project [Internet]. 2008 [cited 2021]. Available online: http://www.repeatmasker.org (accessed on 16 April 2022).

111. Mistry, J.; Chuguransky, S.; Williams, L.; Qureshi, M.; Salazar, G.A.; Sonnhammer, E.L.L.; Tosatto, S.C.E.; Paladin, L.; Raj, S.; Richardson, L.J.; et al. Pfam: The protein families database in 2021. *Nucleic. Acids. Res.* **2021**, *49*, D412–D419. [CrossRef]
112. Wicker, T.; Sabot, F.; Hua-Van, A.; Bennetzen, J.L.; Capy, P.; Chalhoub, B.; Flavell, A.; Leroy, P.; Morgante, M.; Panaud, O.; et al. A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **2007**, *8*, 973–982. [CrossRef]
113. Storer, J.M.; Hubley, R.; Rosen, J.; Smit, A.F.A. Curation Guidelines for *de novo* Generated Transposable Element Families. *Curr. Protoc.* **2021**, *1*, e154. [CrossRef]
114. Price, A.L.; Eskin, E.; Pevzner, P.A. Whole-genome analysis of Alu repeat elements reveals complex evolutionary history. *Genome. Res.* **2004**, *14*, 2245–2252. [CrossRef]
115. Li, W.; Godzik, A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **2006**, *22*, 1658–1659. [CrossRef] [PubMed]
116. Churakov, G.; Zhang, F.; Grundmann, N.; Makalowski, W.; Noll, A.; Doronina, L.; Schmitz, J. The multicomparative 2-n-way genome suite. *Genome. Res.* **2020**, *30*, 1508–1516. [CrossRef] [PubMed]
117. Gibbs, R.A.; Pachter, L. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **2004**, *428*, 493–521. [PubMed]
118. Sequencing, T.C.; Consortium, A. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **2005**, *437*, 69–87.
119. Caspi, A.; Pachter, L. Identification of transposable elements using multiple alignments of related genomes. *Genome. Res.* **2006**, *16*, 260–270. [CrossRef] [PubMed]