The Dfam Community Resource of Transposable Element Families, Sequence Models, and Genome Annotations

Jessica Storer
Institute for Systems Biology
Robert Hubley (robert.hubley@isbscience.org)
Institute for Systems Biology https://orcid.org/0000-0001-9261-3821
Jeb Rosen
Institute for Systems Biology
Travis Wheeler
University of Montana Missoula College
Arian F.A. Smit
Institute for Systems Biology

Methodology

Keywords: Element Families, Sequence Models, Genome Annotations

Posted Date: September 17th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-76062/v1

Version of Record: A version of this preprint was published on January 12th, 2021. See the published version at https://doi.org/10.1186/s13100-020-00230-y.

Abstract

The 3.0-3.2 releases of Dfam (https://dfam.org) represent an evolution from a proof-of-principle collection of transposable element families in model organisms into a community resource for a broad range of species and for both curated and uncurated datasets. In addition, releases since Dfam 3.0 provide auxiliary consensus sequence models, transposable element protein alignments, and a formalized classification system to support the growing diversity of organisms represented in the resource. The latest release includes 266,740 new *de novo* generated transposable element families from 336 species contributed by the EBI. This expansion demonstrates the utility of many of Dfam's new features and provides insight into the long term challenges ahead for improving *de novo* generated transposable element datasets.

Introduction

Significant portions of many genomes are composed of transposable element (TE) copies. TE-derived sequence decays in the genome over time, making its discovery and characterization challenging. However, accurate annotation and description of these elements is crucial in understanding their impact on the genome in which they reside and the evolution of a species as a whole. The influence of TEs on the genome and/or species can be direct, such as insertions into coding regions, exaptation to new functions, or chromosomal rearrangements as a consequence of non-homologous recombination, or indirect, as with an "arms race" between the host and resident parasite. TE instances have long been identified in genomes through database-driven annotation: a database of sequences representing known TE families is maintained, and each sequence in that database is aligned to the genome being annotated, with the best-scoring alignment determining the label of the genomic sequence. Such sequence databases have long contained a consensus sequence representing each family. However, such searches tend to miss highly-diverged sequences, prompting us to explore the utility of profile methods [1, 2] to increase sensitivity.

In 2012, we released Dfam [3], a database of TE families from the human genome in which each family was represented by a multiple sequence alignment (MSA) and a profile hidden Markov model (HMM). Profile HMMs [4, 5] yield sensitivity gains in part by modeling the position-specific residue and indel (insertion and deletion) variability found in family MSAs. The first release of Dfam was based on the design of similar databases of protein (Pfam) and RNA (Rfam) families [6, 7]. In addition to improving annotation sensitivity through the use of profile HMMs, Dfam demonstrated decreased false discovery rates through rigorously defined thresholds [8]. An additional advantage of these databases is the preservation of a multiple sequence alignment of representative family members, the seed alignment. The seed alignment is model-agnostic, provides details on coverage and fragmentation, and supplies essential provenance for the family.

Subsequent releases of Dfam refined the prototype database and modestly expanded the curated libraries to five model organisms (4,150 families). In 2018, Dfam received funding to move from a proof-

of-concept to a production community resource by (1) scaling the system architecture, (2) supporting multiple model types (HMMs and consensus sequences) derived from the seed alignment, (3) improving annotation speed and quality, and most importantly (4) engaging the community in its further development. In this paper we will describe accomplishments represented in the latest release (Dfam 3.2) and the challenges that lay ahead.

Dfam 3.2 currently houses 273,655 families: 83,432 retrotransposons, 71,401 DNA transposons, and 118,822 other repeats which include interspersed repeats of unknown origin, satellite regions and/or other non-TE entries to avoid annotating non-coding RNA genes as TEs. As Dfam is an open collection of TEs, this database will continue to grow as the collective expertise of researchers in the TE community are able to submit new TE libraries, curate existing families, and contribute ideas and effort to this evolving resource.

Consensus Models

While the use of HMMs allows for improved detection of TE copies in genomes, most sequence analysis algorithms (Smith-Waterman, Needleman-Wunsch, Suffix Trees, Burrows Wheeler Transform, etc.) and popular sequence analysis tools (BLAST, BLAT, MAFFT etc.) act directly on string representations of sequences (e.g. consensus sequences). Likewise, programs used to define new TEs (de-novo repeat finders), to extend fragmented models, to unravel the relationship of related TEs, to classify elements or to describe biological features like exons are not typically able to generate or take advantage of HMMs directly. Each TE model in Dfam therefore needs to be accompanied by a simple sequence model, for which a consensus sequence derived from the seed alignment is the logical candidate.

The use of a consensus sequence as a first-order model for sequence families has a rich history of demonstrated utility [2, 9–11]. A consensus is typically made by considering the occupancy and composition of columns in a multiple sequence alignment of TE copies. A basic consensus caller might assign for any given column the majority nucleotide found in the column regardless of occupancy (the number of homologous nucleotides in the column vs the number of gaps). A more sophisticated caller would account for gaps and make base calls reflecting the observed rates of substitutions in the given genome.

Most TE-derived sequences are under no functional constraint and accumulate mutations in a random and neutral fashion. Given this random noise, an informed consensus for a sufficient number of properly aligned copies may be expected to reproduce the original active TE sequence. This bears out particularly well for most "class II elements" or "DNA transposons" in eukaryotes; due to the trans-activity of the transposase on the genomic copies, these do not tend to evolve during their short life in a genome and create copies that have a star-like phylogenetic relationship to the original sequence (Fig. 1) [12]. The situation is more complicated for most class I elements, which duplicate via a reverse transcription step; thanks to the cis-activity of the reverse transcriptase on its own mRNA, they may evolve, e.g. to escape the host's defense mechanisms, and propagate in a genome for hundreds of millions of years. For these

elements, careful clustering of copies into so-called subfamilies will result in a series of interrelated consensus sequences that can be interpreted as snapshots of the TE sequence during its evolution, though each may still be a composition of divergent active elements.

Figure 1 - Typical phylogenetic structure of retroposon and DNA transposon families. After multiple mutations have occurred in the evolving class I TE, the relative ordering of copies may be distinguished by these changes as they cosegregate. The presence of such clusters or "subfamilies" of TE copies is a good indication that they arose via retrotransposition.

The concept that consensus sequences approach the original TE sequence has been demonstrated by the resurrection of recently extinct TEs through modification of a dead copy to the consensus [13, 14] and by the recovery of expected sequence features for ancient elements. For example, the consensus sequences of many coding TEs that were active > 100 MYA in our genome contain full-length ORFs [15], despite the fact that their individual copies have accumulated so many mutations that they on average share less than half the original nucleotides with each other and often cannot be pairwise aligned.

While Repbase was initiated as a reference database containing prototypes of genomic interspersed repetitive DNA [16], we transformed it into a database of consensus sequences by 1994 [11]. Not only did this endeavor explain the biological origin of most repetitive DNA, the use of consensus sequences rather than genomic copies improved their detection as well. A collection of genomic copies has redundancy and contains low complexity sequences like simple repeats expanded in individual copies, both of which result in lower specificity. More importantly, individual copies of a TE are separated by, on average, twice as many mutations as they are from the original sequence approached by a consensus, improving sensitivity dramatically.

In Dfam we now provide both profile HMM and auxiliary consensus sequence models for each family in the database. Both are derived from a single seed alignment, allowing for provenance to be maintained and both models to be simultaneously improved. Most importantly, Dfam maintains a correspondence between consensus and HMM positions so that alignments produced by either may be compared directly. In Dfam, consensus sequences are generated for each family using a caller that we originally employed to build many of the consensus sequences for Repbase. It assigns the base with the highest score using a log-odds substitution matrix that reflects neutral substitution patterns in a genome, including e.g. the strong GC->AT bias in mammals. It also infers ancestral CpG sites by accounting for the frequency of TG and CA dimers in neighboring columns [17].

Classification System

Without classification, a TE library is of limited use. While entries in Dfam have always been classified, in this release we have added an interactive tool to the website, displaying our classification scheme for repetitive sequences in eukaryotic genomes in the form of an identification key.

Classification of TEs poses specific problems that may prevent a universal solution to be found [18]. A purely cladistic approach is impossible as TEs are polyphyletic (they have many independent origins) and because their relationship is reticulated (sections of TEs can have entirely different evolutionary histories, due to recombinations, gene captures and nested insertions). Classic SINEs, which have originated many times from fortuitous positioning of an internal promoter (e.g. in a small RNA gene) and the 3' fragment of an active LINE [15, 19] provide an example for both these issues. Nevertheless, most currently used classification systems for eukaryotic TEs are very similar and are based on hybrids of cladistic, mechanistic and structural approaches.

In 1989, David Finnegan introduced an early classification with just four classes [20]. His basic division between TEs that transpose via an RNA intermediate (class I) and those that “transpose directly from DNA to DNA” (class II) is still used by most. Considering the fundamental impact of the trans-activity of class II proteins on their transcripts and the cis -activity of class I proteins on their genomic copies, this division is indeed fundamental. At the time, very few types of eukaryotic TEs were known, and his further divisions of class I elements into those with and without long terminal repeats (LTRs), and divisions of class II elements into those with short and long terminal inverted repeats (TIRs) has not survived the onslaught of new data, although LTR and non-LTR (LINE) elements still form valid clades, at least from the reverse-transcriptase point of view [21].

When we introduced RepeatMasker in 1995, we needed a succinct classification to fit in the slightly modified cross_match format [22] we used to annotate genomic DNA. We chose a three-level form coded as “level1/level2-level3” (e.g. “DNA/hAT-Charlie”). We adopted Finnegan's LTR, LINE and class II (“DNA”) divisions and added SINE and a number of non-TE classes for the first divisions, the three class I elements reflecting a bias towards the frequency of elements encountered in the human and other mammalian genomes. Second and third divisions represent clades of elements based on reverse transcriptase (RT) or transposases phylogenies. Non-autonomous elements whose movement depends upon the coding capacity of autonomous elements, were grouped within the autonomous elements' classification, based on similarities of the LTRs or TIRs in the absence of any coding sequence. Entries in Repbase more or less inherited this simple classification hierarchy. In later years, attempts were made to reflect as much as possible of the classification in the name of the elements [23]. The classification system suggested in 2007 by researchers with a primarily plant genomics background [24], has the same basis in Finnegan and follows a similar logic; in order to display compact classification on an annotation line, they suggested a three-letter class-order-superfamily code to add to each "family" classification. The "subfamily" was suggested to be used in the TE's name itself.

Our classification, like those before, combines a mechanistic, cladistic and structural approach. Where possible, the relationship of the RT in class I elements and transposase, helicase, or DNA polymerase in class II elements guides the tree. While non-autonomous LTR elements tend to remain dependent on the autonomous element from which they formed and can be classified with these, LINE-dependent non-autonomous elements have a variety of origins. They are separated by those with a small RNA derived pol III internal promoter (the SINEs) and other elements. The latter category is a grab bag of sorts,

classified by the type of LINE they depend upon, and contains elements mostly consisting of LINE-material to hodgepodges like SVA [25]. The modular, classic SINEs are organized by their 5' small RNA-derived, core, and 3' LINE-derived modules. Class II elements are divided in the four fundamental mechanisms of propagation so far known in eukaryotes, "cut-and-paste" via a linear or circular dsDNA, "rolling circle", and "self-synthesizing" groups, after which the phylogenetic relationship of the transposase, recombinase, helicase, or DNA polymerase, respectively, takes over. Like non-autonomous LTR elements, most non-autonomous DNA transposons can be classified based on their TIR combined with their target site duplication (TSD) pattern. We therefore do not provide structural categories like LARD (large LTR retrotransposon derivatives) or MITE (miniature inverted–repeat transposable elements).

Figure 2 - Dfam TE Classification System. (A) A portion of the dynamic visualization of the classification system found at the Dfam website. Filled in circles represent internal nodes of the tree while hollow circles are leaf nodes in the classification tree. A classification is specified by concatenating the path through the classification tree. For example, the classification “Interspersed_Repeat;Unknown” is highlighted in the tree. (B) In addition, wherever possible a mapping is provided between classification systems. The Dfam classification for the L1 group of LINEs is shown with the equivalent classifications in several other systems.

The Dfam classification system (Fig. 2) does not display a ranked hierarchy as there will never be satisfying definitions for what a class, order, family or subfamily of TEs constitutes, while with the addition of new elements and growing knowledge of their relationship, the number of branches, and therefore subdivisions, along some parts of the tree will remain in flux. Wicker et al. proposed to define a family as a group of TEs that can be aligned over at least 80 bp and show 80%+ identity covering 80% or more of the alignment [24]. Meant as a pragmatic definition, it has been pointed out that applying it would lead "to an unpredictable mix of monophyletic, paraphyletic and polyphyletic groups" [26]. Strictly following this rule will also not be practical, as, for example, newly identified TEs intermediate between known families will force these to be merged over time and the aforementioned reticulate relationship of TEs could join radically different TEs in one family. Also, some of the ranks are already in use for other purposes: the term "family" is often used for any group of aligned TE copies for which a consensus or HMM has been derived and, in animal TE annotation, "subfamilies" either indicate subsets of class I TE copies that share multiple co-segregating differences from the rest (Fig. 1) or sets of particular internal deletion products of an autonomous class II transposon. With its lack of taxonomic ranks, our schema avoids these issues.

TEs may also be classified by their transposition mechanism and classification systems based on the mechanism of integration and chemistry of the transposition reaction have been proposed [27, 28]. These have the benefit of being able to integrate the wide variety of TEs active in prokaryotes, but are somewhat hampered by the lack of knowledge on the details of transposition by new, bioinformatically discovered TEs. Furthermore, written specifically to include prokaryotic TEs, the mechanistic classifications do not have the fundamental division in cis-active and trans-active elements, brought about by the separation of

transcription and translation in eukaryotes. While the focus of the RepeatMasker/Repbase/Wicker classification on eukaryotes and on reverse transcriptase phylogeny has been criticized [29], a unified eukaryotic/prokaryotic TE classification would be unwieldy. In the future, we will explore the use of an independent classification for prokaryotic TEs.

A TE family can be classified as belonging to any node in the classification tree by concatenating the names along the path from the root to the designated node. For example, the highlighted node in Fig. 2 is referenced with the string “Interspersed_repeat;Unknown”. This enables partial classifications to be made and node labels to be reused. All classifications are linked to the corresponding RepeatMasker, RepBase, Wicker-et-al. or Curcio-Derbyshire classification, where they are available.

While most interspersed repeats identified by *de novo* repeat finding programs are derived from TEs, alternative origins include (i) simple tandem repeats, originating independently at many sites, (ii) long tandem repeats like satellites, found at multiple (sub)telomeric and centromeric sites, (iii) segmental duplications, (iv) common coding motifs like zinc fingers, and (v) gene families. In mammals, the most common non-TE source of interspersed repeats are retro(pseudo)genes that have been accidently copied by the LINE1-mechanism; some small structural RNAs occur with over a thousand copies [30]. While our classification system includes these categories, most of these entries should not be part of Dfam. Satellites and small structural RNAs are included in Dfam, but shorter tandem repeats are better detected by specialized programs like TRF [31] and ULTRA [32] and the inclusion of segmental duplications, cellular transcripts or coding regions would lead to much false annotation.

Transposon Termini

Most methods to categorize newly identified TEs are pipelines that rely heavily on finding homology to existing classified TEs [17, 33–35]. When a curated library of a related species exists, near-full-length matches at the DNA level are often found that allow proper class assignment. A translated comparison to a TE protein database can classify many models with (remnants of) coding sequence. However, due to the recombinant and modular tendencies of TEs, sequence homology is not always sufficient evidence for classification, especially if similarities are fragmentary or weak. For example, many non-autonomous class I and II elements carry insertions or fragments of non-related TEs, while a model matching the 3' end of a LINE element may represent a SINE instead. Even functional TE coding regions can be misleading as related proteins have been repurposed in disparate TEs; for example, the transposases of the cut-and-paste Ginger transposons are closely related to the integrases of both Gypsy retrotransposons and Maverick/polinton self-synthesizing elements [36, 37]. Needless to say, TEs without (remnant) coding sequence or related entries in the database will remain unclassified using this method alone.

There are fortunately other characteristics and sequence features of TEs that can confirm the mechanism by which the element propagated, and in turn, how it should be classified. Clusters of copies with co-segregating mutations generally imply that the TE is a class I element, and, given enough copies, their

complete absence suggests a class II element. A simple-repeat tail and TSDs of variable length indicate movement via target-primed reverse transcription, which requires the protein products of LINEs. When models have TG…CA termini and carry a poly-adenylation signal, while their copies are flanked by 4–6 bp TSDs, they likely represent solitary LTRs, which, through homologous recombination, can vastly outnumber complete LTR elements in a genome.

In our experience, many of the not automatically categorized models represent non-autonomous class II elements that actually can be classified in detail based on the pattern of the terminal 20–50 bp. This is particularly true for the ubiquitous TIR transposons: their transposases specifically recognize and bind the terminal sequences, to the point that these are sufficient for an element to retain or obtain mobility [38, 39]. As a result, transposons with similar transposases (the basis of their classification) have similar termini. While these terminal homologies are too short to appear significant in a whole database search, they can be found by comparing the 30–60 bp termini of new elements to the 30–60 bp termini of all classified class II elements and filtering the output by orientation and position. Characteristic TSDs can cement classification. Over the years, we classified thousands of short sequences in Repbase this way, including ancient elements that were active in the common ancestor of all amniotes.

Figure 3 - Generation of HMMs and sequence LOGOs for DNA Transposon Termini. The first/last 60 bp of family consensi belonging to a single classification of DNA Transposons are piled up and aligned (without gaps) by hand. Profile HMMs are developed for each end and for the combination of the two to determine if a stronger signal may be obtained in that fashion. Finally, LOGOs are generated for each HMM and displayed on the Dfam website.

In Dfam, we now provide terminal sequence signatures for 64 categories of class II elements, for use in classifying new TE models. To create these signatures, we lined up the 5' and 3' 60 bp of all members of a particular type that seem to have clearly defined ends (e.g. as indicated by the presence of the expected TSDs, Fig. 3). Minor modifications were made to some of the nearly 12,000 remaining consensus sequences in order to have each start at the true beginning or end of the TE. This most commonly involved removing (partial) target site duplications or adding one or more Ns when comparison to others showed the sequence to be short. The alignments are ungapped, however, and it is possible that more signal can be obtained by allowing a few indels. We used HMMER (hmmbuild) to develop HMMs for the 5’ ends, the 3’ ends, and, in case of TIRs, the combined termini.

The LOGOs of the termini can be viewed on the “Classifications” page on the Dfam website, and are organized by class II subclasses (e.g., Crypton, Helitron, TIR, etc.). This allows for easy visualization of the base conservation at each position in the terminal sequences and comparisons between the 5’ and 3’ termini. The full set of profiles may be downloaded and will be updated as new elements are added to each class.

Dfam Growth

The development of a curated TE library for a given species is a specialized and mostly manual task that has only incrementally improved since the inception of TE databases. We recognize that as reference genome sequencing increases at a faster rate, and until automated curation methods improve, uncurated datasets (mostly *de novo* generated TE libraries) will far outpace the development of curated libraries. There are some advantages to making uncurated datasets available through a resource such as Dfam: (1) they can be used as simple genome masking libraries, and the fragmentation and redundancy that are the hallmark of datasets derived from *de novo* discovery tools are not detrimental in that context, (2) cataloging uncurated families provides a shared starting point for community curation efforts, and (3) these datasets will provide a resource for developing per-family and per-library quality metrics as well as improved automated curation tools.

Figure 4 - Dfam analysis pipeline. The full Dfam analysis pipeline consists of a set of sequential analysis steps depicted above with examples of the products produced. For uncurated families only the first portion of the pipeline (colored in blue) is initially conducted.

In the latest release of Dfam we have added support for uncurated datasets, datasets, denoting these families using the new accession prefix “DR” and limiting/altering the analysis and metadata displayed for these families (Fig. 4). For instance, a DR family has both a consensus and a profile HMM generated for it, but does not have rigorously characterized false discovery rate thresholds as for curated families [3] due to the lack of pre-calculated assembly annotations. However, uncurated families do contain provenance for the seed alignment, standard metadata (description, classification, taxa, citations etc), TE protein matches, relationships with other families and model details. By limiting the analysis for DR families to those that facilitate future curation efforts, we can scale Dfam to handle this growing data category and provide early access to newly discovered TE families.

To demonstrate the new features in the latest release of Dfam and the challenges/opportunities this type of dataset will create, we imported RepeatModeler results generated by the EBI on 336 assemblies (Fig. 5). This import of 266,740 families dwarfed the existing Dfam by 40-fold. The seed alignments typically produced by RepeatModeler were not available from these runs, therefore we used the RepeatModeler consensus libraries to generate new seed alignments from annotated instances identified by RepeatMasker. By using a library-based approach rather than individual family searches, we avoided assigning TE instances to more than one family. This step was followed by an iterative extension process (manuscript in preparation) that is based upon an approach used in RepeatScout [40] to further extend fragmented families.

Figure 5 - Phylogenetic tree of the species in the Dfam 3.2 release: Schematic of the genomes now represented in Dfam3.2. Collapsed nodes are present as triangles, with the number of species preceding a label of the species present in the triangle. A selection of species not in collapsed nodes are indicated. The dotted line indicates an axis break. Major branch points have been labeled according to the taxonomy on NCBI. Branching order within the birds follows that of Prum et al [14764687]

Due to the scale of the dataset, only minimal prefiltering of the families was performed before import. Families with > 80% tandemly repetitive sequence and less than 100 bp of contiguous non tandemly repetitive sequence (according to TRF) were removed. This filtered 894 families. TRF was run with a maximum period of 20 bp to preserve common and complex satellite sequences in the dataset.

Examples of the seed alignments generated as part of the aforementioned pipeline can be observed in Fig. 6. The seed alignment for any model provides researchers with a wealth of information including sequence coverage, length, and number of sequences contributing to the model. From a curator's perspective, information like divergence patterns and blocks of seemingly truncated sequences are also evident. For example, a block of sequences in an LTR alignment that show a different divergence pattern and/or appears to be truncated at the same position relative to the consensus sequence are likely to indicate a subfamily structure within the model (Fig. 6-a, red box). This alignment pattern reflects the biology of the TE, as LTR sequences regularly recombine their 5' and 3' ends to form new families [41]. Similarly, blocks of outwardly truncated sequences close to the 5' end in an L1 seed alignment reflect a tendency for differing patterns present in the 5' UTR regions [42](Fig. 6-d). Closer inspection of these sequence blocks within the alignment is warranted to determine if a subfamily structure is present. In addition, not only does this pipeline have the ability to generate shorter TEs like some LTRs (Fig. 6-a) and SINEs, but also longer elements like the aforementioned autonomous LINE elements (Fig. 6-d) and Helitrons (Fig. 6-c) with acceptable sequence coverage given the size of the element.

Figure 6. Seed alignment examples from raw Dfam3.2 entries. A) ERV1 (LTR) (DR0086957.1; *Eulemur macaco* (black lemur)). The red box indicates a group of sequences differing in length and divergence patterns. B) Unknown sequence (DR0087060.1; *Eulemur macaco* (black lemur)). The orange box indicates a potential segmental duplication. C) Helitron-1 (DR0096635.1; *Oreochromis niloticus* (tilapia)) D) L1 sequence (DR0215804.1; *Phyllostomus discolor* (pale spear-nosed bat)). The red lines indicate two groups of sequences that differ in their 5' alignment. Blue sequences indicate a higher match to the consensus, while red indicates a poorer match.

Not all seed alignments are clear or accurate. Figure 6-b depicts the seed alignment from an unknown TE sequence in a black lemur. This sequence has higher sequence coverage in the 5' end of the sequence, but contains a block of sequence with no coverage. In these instances, it is possible that RepeatModeler managed to stitch together a full family consensus sequence. However, when analyzed by RepeatMasker, followed by an analysis to reduce TE family redundancy, a different family outcompeted the middle section in all places in the genome, leading to a gap. Upon further analysis, it appears as though a highly repetitive satellite sequence is present within the "Unknown" seed alignment (Fig. 6-b, orange box). The segment does not match any repeats in our database.

By providing the RepeatModeler output for the additional 336 species, Dfam encourages the TE community to lend their expertise, thus improving this dataset so that it moves from a raw dataset to a highly curated one. We are working on providing tools for the community to curate and improve upon these and existing curated datasets.

Architectural And Interface Improvements

The architecture of Dfam was refactored in Dfam 3.0 to prepare the resource for housing TE families, seed alignments, and sequence models at scale. The Dfam website (https://www.dfam.org) has been updated to provide both (i) a front end intended for human interaction through a web browser (the 'portal') and (ii) an API served over HTTP to support programmatic access to Dfam data. The source code for both projects is available on GitHub (https://github.com/Dfam-consortium/) and released under the CC0 public domain dedication. The API uses computer-friendly data formats such as JSON and tab-separated values (as appropriate), which makes it more suitable as a data source for community-developed tools than the human-oriented format of the website.

The portal has gained several significant new features. To support the massive scale of family data we replaced the old one-page-per-letter approach to family organization with a "Browse" page that supports sorting and filtering on multiple criteria such as name, classification, taxon, or keywords. Some similar filters have also been added to the "Relationships" tab to restrict results to related species. The "Features" tab has been added to the per-family page to display curated features (binding sites, hand-curated coding sequences etc), as well as blastx matches against other known TE protein sequences.

The architectural changes also facilitated the merger of Dfam and Dfam_consensus [43], creating a single resource for both consensus and HMM sequence models for each family. During the merger, Dfam inherited the seed alignment visualization from Dfam_consensus. This visualization shows both coverage (how many representatives cover each portion of the alignment) and the conservation patterns within the alignment, shown as a heatmap indicating how closely each member aligns to the consensus (as seen in Fig. 6).

Dfam also now includes an authentication system and public data submission system; this is the first step in a planned development of a curation workbench that will provide users with tools to edit and curate data uploaded by themselves or (eventually) others.

Software/tool Distribution Improvements

One of the main hurdles with using bioinformatic pipelines and tools is the complexity of the installation and configuration necessary to use them. Software containers are an increasingly popular way to tackle this problem, by delivering an executable user environment that comes with pre-installed and pre-configured software packages. In order to support community curation efforts we have developed containers for Docker and Singularity, housing pre-installed versions of our latest tools (e.g. RepeatMasker, RepeatModeler, Coseg, RMBlast) and dependencies as well as external open-source tools (e.g. HMMER, mafft, cdhit, TRF). More information along with instructions for use are available at https://github.com/Dfam-consortium/TETools/. As part of our outreach efforts, we will be eliciting recommendations for including additional packages in future releases.

Notably these containers include FamDB, a new HDF5-based Dfam export format and associated query tools for offline access to Dfam. FamDB files contain family consensi/HMMs and the NCBI Taxonomy data related to these families in a format that allows for fast offline access from the command line. The current release of FamDB includes all Dfam consensus sequences, HMMs, metadata, and 61,003 taxa from NCBI's taxonomy database [44] related to these families. Lookups for information on a single taxon or family complete in about a second; extraction of consensus sequences (FASTA, EMBL) or HMMs for all TE families found in Human (including ancestral repeats) complete in about 3 to 4 seconds. Due to indexing, the run time for data queries is largely independent of the total number of TEs in the database: it takes about the same amount of time to extract the human library from a FamDB file including only the curated subset of Dfam (6,915 entries) as for the full database (273,655).

Future Challenges/directions

The curation of multi-species TE databases involves many manual tasks (e.g. defragmentation, classification validation, lineage assignment) and many subjective decisions (e.g. redundancy removal, pseudogene removal, subfamily characterization) that have not been universally standardized into protocols. With the growth of uncurated TE libraries and their inclusion in databases such as Dfam, it has become necessary to develop protocols for many of these tasks and to consider the challenges introduced by this data growth. Here, we discuss two of these topics that we will be focusing on in the near term.

Subfamilies

When a family is believed to exhibit subfamily structure there are a variety of methods available to cluster the family copies into subfamily groups. A subfamily structure is suspected when there is a wide divergence distribution of family members, or directly observing distinct groups of co-segregating subsequences in a seed alignment. Wicker used the aforementioned 80/80/80 rule to cluster sequences into subfamilies: sequences are clustered such that all members of the subfamily share at least 80% sequence identity and at least 80% sequence coverage with length greater than 80 bp. A strategy to define subfamilies was used in the analysis of bread wheat subgenomes by applying a 90/90 or 95/95 rule in which 90 or 95% sequence identity and 90 or 95% sequence coverage was used to cluster sequences into subfamilies [45, 46]. However, further testing should be completed to determine if this type of threshold accurately splits sequences into subfamilies of all TE types across a wide array of species.

The methods used to develop subfamily models vary widely, from the manual clustering of copies in a multiple sequence alignment by eye, to automated clustering algorithms [47], and network-based approaches [48]. Wildly different subfamily sets can be produced by these alternative methods, or even by using slightly different parameterizations within one method. The large number of fine-grained subfamilies produced by some of these methods is not of practical use for identifying copies of the superfamily in a genome or specifically labeling individual copies confidently with current sequence

similarity search algorithms. Still, in aggregate the tree structure and subfamily membership are valuable datasets for studying family evolution and databases can play a role in the standardisation of this data.

For future releases of Dfam we will explore ways to set a minimum sequence distance threshold for inclusion in Dfam as a subfamily. The threshold should reflect the current sensitivity and specificity of both HMM and consensus based search algorithms and act on the detailed subfamily tree to cluster closely-related subfamilies (lumping their copies together). The original detailed tree structure and individual copy membership need not be lost (Fig. 7), but stored alongside the superfamily as a combination of newick data and fine-scale seed alignment sequence subfamily labels for further study and use.

Figure 7: Database Subfamily Representation: Proposed database representation for TE subfamilies maintaining a detailed phylogenetic structure while reducing the representative models for practical genome-scale annotation. The TE seed alignment (1) from a family with evidence of subfamily structure is analysed by a clustering method to produce a detailed subfamily structure and membership (2). Sequence models are developed for subfamilies and lumped (3) if model performance isn't improved by the subdivision of two or more subfamilies. The lumped families and their corresponding seed alignments are added to the database (4) with metadata holding the detailed tree structure and seed sequence membership for each subfamily.

Redundancy/Fragmentation Detection

Ideally, a TE database should contain a single full-length entry for each transposing family. Unfortunately that overly simple definition doesn't account for the fine detail of subfamily expansions, recombinations, deletion products, and mosaicism exhibited by many TE families. These processes lead to necessary redundancy in a library. Another form of redundancy, that is not desirable, is the direct result of (1) re-detection of ancestral families in the *de novo* analysis of two or more related species, (2) the confounding effects of sequence variation on *de novo* detection methods leading to rediscovery, and (3) inadequate clustering in pipelines that run multiple discovery methods and merge the results (Fig. 8). In addition, differences in the representative set of TE copies, in the alignment parameters, and the selection of model building parameters will lead to subtle differences between models generated for the same family; making automatic redundancy detection difficult.

Through the expansion of the Dfam database via the addition of diverse sets of species and their associated TEs, it will become necessary to detect redundancy automatically. One approach would be to use a comparative genomics approach to assess TE insertions at orthologous sites to resolve interspecies redundancies while improving the taxonomic labels for each family (Fig. 8).

Figure 8

Redundancy/Fragmentation removal challenges. Both inter- and intra-library redundancy is present in *de novo* datasets and are currently resolved through manual curation. Interlibrary redundancy is often the result of unresolved subfamily structure (e.g. internal deletion products of DNA transposons) that

confounds discovery and produces both redundant and fragmented families. Intra library redundancy is an inherent aspect analyzing a single species in isolation. For each new species these ancestral families need to be resolved by comparison to existing families, and by considering presence at orthologous sites.

Fragmentation is an additional problem apparent in most *de novo* datasets. In some cases this directly relates to the structure of the observed TE copies appearing as distinct patterns within a genome (e.g. full length LTRs with internal sequence, and solo LTRs) or coverage bias in a families copies (e.g 5' end of LINE families). In either case joining these fragments into a complete family is the desired result. Fragmentation is often identified during manual curation as a family fragment is extended and subsequently matched to another fragment in the library over the extended region. Another approach would be to use genome annotations for the uncurated library to identify significant collinearity among family pairs and automatically group families together.

Conclusion

The new Dfam release has expanded the number and scope of species included the database, allowing for enhanced genome annotation while fostering the development of highly curated TE libraries for use in research. In addition, a unified eukaryotic TE classification scheme and HMMs for DNA transposon termini now on Dfam, provide additional details for researchers to utilize in their TE research. Combined with an expanded TE library, the new database architecture, improved interfaces, and simplified software distribution, Dfam offers a collaborative platform for the TE research community. Collaborative efforts and increased datasets will be necessary to tackle problems such as those mentioned above: subfamily identification, library redundancy and fragmentation. We invite the TE research community to provide feedback on the challenges discussed here and to join us in these efforts to further Dfam development.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets generated during and/or analyzed during the current study are available at the Dfam portal: https://dfam.org. Software generated for this project is located in the Dfam consortium repository: https://github.com/Dfam-consortium.

Competing interests

The authors declare that they have no competing interests.

Funding

This work is funded by NHGRI grant # U24 HG010136, and NHGRI grant # RO1 HG002939.

Authors' contributions

JR, RH, and TW developed the infrastructure and analysis pipelines for Dfam 3.x. AS developed the classification system and performed the transposon termini analysis. JS, AS curated Dfam families. JS, RH, and AS were major contributors in writing the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We would like to thank Fergal Martin and Denye Ogeh of the European Bioinformatics Institute for sharing the RepeatModeler libraries they generated. Additionally, we would like to thank David Ray (TTU), and University of Montana's Griz Shared Computing Cluster (GSCC) for providing cluster time for some of the analyses performed on the EBI dataset.

References

1. Park J, Karplus K, Barrett C, Hughey R, Haussler D, Hubbard T, et al. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. J Mol Biol [Internet]. 1998;284:1201–10. Available from: http://dx.doi.org/10.1006/jmbi.1998.2221
2. Schneider TD. Consensus sequence Zen. Appl Bioinformatics [Internet]. 2002;1:111–9. Available from: https://www.ncbi.nlm.nih.gov/pubmed/15130839
3. Wheeler TJ, Clements J, Eddy SR, Hubley R, Jones T a., Jurka J, et al. Dfam: a database of repetitive DNA based on profile hidden Markov models. Nucleic Acids Res [Internet]. 2013;41:D70–82. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3531169&tool=pmcentrez&rendertype=abstract
4. Krogh A, Brown M, Mian IS, Sjölander K, Haussler D. Hidden Markov models in computational biology. Applications to protein modeling. J Mol Biol [Internet]. 1994;235:1501–31. Available from: http://dx.doi.org/10.1006/jmbi.1994.1104
5. Wheeler TJ, Eddy SR. nhmmer: DNA Homology Search With Profile HMMs. 2013.
6. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, et al. The Pfam protein families database. Nucleic Acids Res [Internet]. 2010;38:D211–22. Available from: http://dx.doi.org/10.1093/nar/gkp985
7. Gardner PP, Daub J, Tate J, Moore BL, Osuch IH, Griffiths-Jones S, et al. Rfam: Wikipedia, clans and the "decimal" release. Nucleic Acids Res [Internet]. Oxford University Press; 2010;39:D141–5.

Available from: https://academic.oup.com/nar/article-abstract/39/suppl_1/D141/2507084

8. Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, et al. The Dfam database of repetitive DNA families. Nucleic Acids Res [Internet]. 2016;44:D81–9. Available from: http://dx.doi.org/10.1093/nar/gkv1272
9. Deininger PL, Jolly DJ, Rubin CM, Friedmann T, Schmid CW. Base Sequence Studies of 300 Nucleotide Renatured Repeated Human DNA Clones. J Mol Riol. 1981;151:17–33.
10. Sadler JR, Waterman MS, Smith TF. Regulatory pattern identification in nucleic acid sequences. Nucleic Acids Res [Internet]. 1983;11:2221–31. Available from: http://dx.doi.org/10.1093/nar/11.7.2221
11. Smit AFA. Structure and evolution of mammalian interspersed repeats. University of Southern California; 1996.
12. Smit AF, Riggs AD. Tiggers and DNA transposon fossils in the human genome. Proc Natl Acad Sci U S A [Internet]. National Acad Sciences; 1996;93:1443–8. Available from: http://dx.doi.org/10.1073/pnas.93.4.1443
13. Ivics Z, Hackett PB, Plasterk RH, Izsvák Z. Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells. Cell [Internet]. Elsevier; 1997;91:501–10. Available from: http://dx.doi.org/10.1016/s0092-8674(00)80436-5
14. Jacobs FMJ, Greenberg D, Nguyen N, Haeussler M, Ewing AD, Katzman S, et al. An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. Nature [Internet]. nature.com; 2014;516:242–5. Available from: http://dx.doi.org/10.1038/nature13760
15. Smit AFA. The origin of interspersed repeats in the human genome. Curr Opin Genet Dev [Internet]. Elsevier; 1996;6:743–8. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0959437X9680030X
16. Jurka J, Walichiewicz J, Milosavljevic A. Prototypic sequences for human repetitive DNA. J Mol Evol [Internet]. 1992;35:286–91. Available from: https://www.ncbi.nlm.nih.gov/pubmed/1404414
17. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci U S A [Internet]. 2020;117:9451–7. Available from: http://dx.doi.org/10.1073/pnas.1921046117
18. Arkhipova IR. Using bioinformatic and phylogenetic approaches to classify transposable elements and understand their complex evolutionary histories. Mob DNA [Internet]. 2017;8:19. Available from: http://dx.doi.org/10.1186/s13100-017-0103-2
19. Ohshima K, Hamada M, Terai Y, Okada N. The 3’ ends of tRNA-derived short interspersed repetitive elements are derived from the 3' ends of long interspersed repetitive elements. Mol Cell Biol [Internet]. 1996;16:3756–64. Available from: http://dx.doi.org/10.1128/mcb.16.7.3756
20. Finnegan DJ. Eukaryotic transposable elements and genome evolution. Trends Genet [Internet]. 1989;5:103–7. Available from: http://dx.doi.org/10.1016/0168-9525(89)90039-5
21. Malik HS, Eickbush TH. Phylogenetic analysis of ribonuclease H domains suggests a late, chimeric origin of LTR retrotransposable elements and retroviruses. Genome Res [Internet]. 2001;11:1187–97.

Available from: http://dx.doi.org/10.1101/gr.185101

22. Green P. Phrap/Cross_match/Swat Bioinformatics Tools [Internet]. Laboratory of Phil Green. 1998. Available from: http://phrap.org
23. Kapitonov VV, Jurka J. A universal classification of eukaryotic transposable elements implemented in Repbase. Nat Rev Genet [Internet]. 2008;9:411–2; author reply 414. Available from: http://dx.doi.org/10.1038/nrg2165-c1
24. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. Nat Rev Genet [Internet]. 2007;8:973–82. Available from: http://dx.doi.org/10.1038/nrg2165
25. Damert A. Composite non-LTR retrotransposons in hominoid primates. Mob Genet Elements [Internet]. 2015;5:67–71. Available from: http://dx.doi.org/10.1080/2159256X.2015.1068906
26. Seberg O, Petersen G. A unified classification system for eukaryotic transposable elements should reflect their phylogeny [Internet]. Nat. Rev. Genet. 2009. p. 276. Available from: http://dx.doi.org/10.1038/nrg2165-c3
27. Curcio MJ, Derbyshire KM. The outs and ins of transposition: from mu to kangaroo. Nat Rev Mol Cell Biol [Internet]. 2003;4:865–77. Available from: http://dx.doi.org/10.1038/nrm1241
28. Hickman AB, Chandler M, Dyda F. Integrating prokaryotes and eukaryotes: DNA transposases in light of structure. Crit Rev Biochem Mol Biol [Internet]. 2010;45:50–69. Available from: http://dx.doi.org/10.3109/10409230903505596
29. Piégu B, Bire S, Arensburger P, Bigot Y. A survey of transposable element classification systems–a call for a fundamental update to meet the challenge of their diversity and complexity. Mol Phylogenet Evol [Internet]. 2015;86:90–109. Available from: http://dx.doi.org/10.1016/j.ympev.2015.03.009
30. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature [Internet]. 2001;409:860–921. Available from: http://dx.doi.org/10.1038/35057062
31. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res [Internet]. 1999;27:573–80. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=148217&tool=pmcentrez&rendertype=abstract
32. Olson D, Wheeler T. ULTRA: A Model Based Tool to Detect Tandem Repeats. ACM BCB [Internet]. 2018;2018:37–46. Available from: http://dx.doi.org/10.1145/3233547.3233604
33. Feschotte C, Keswani U, Ranganathan N, Guibotsy ML, Levine D. Exploring repetitive DNA landscapes using REPCLASS, a tool that automates the classification of transposable elements in eukaryotic genomes. Genome Biol Evol [Internet]. 2009;1:205–20. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2817418&tool=pmcentrez&rendertype=abstract
34. Abrusán G, Grundmann N, DeMester L, Makalowski W. TEclass--a tool for automated classification of unknown eukaryotic transposable elements. Bioinformatics [Internet]. 2009;25:1329–30. Available

from: http://dx.doi.org/10.1093/bioinformatics/btp084

35. Hoede C, Arnoux S, Moisset M, Chaumier T, Inizan O, Jamilloux V, et al. PASTEC: an automatic transposable element classification tool. PLoS One [Internet]. 2014;9:e91929. Available from: http://dx.doi.org/10.1371/journal.pone.0091929
36. Bao W, Kapitonov VV, Jurka J. Ginger DNA transposons in eukaryotes and their evolutionary relationships with long terminal repeat retrotransposons. Mob DNA [Internet]. 2010;1:3. Available from: http://dx.doi.org/10.1186/1759-8753-1-3
37. Cerbin S, Wai CM, VanBuren R, Jiang N. GingerRoot: A Novel DNA Transposon Encoding Integrase-Related Transposase in Plants and Animals. Genome Biol Evol [Internet]. 2019;11:3181–93. Available from: http://dx.doi.org/10.1093/gbe/evz230
38. van Luenen HG, Colloms SD, Plasterk RH. The mechanism of transposition of Tc3 in C. elegans. Cell [Internet]. 1994;79:293–301. Available from: http://dx.doi.org/10.1016/0092-8674(94)90198-8
39. Vos JC, De Baere I, Plasterk RH. Transposase is the only nematode protein required for in vitro transposition of Tc1. Genes Dev [Internet]. 1996;10:755–61. Available from: http://dx.doi.org/10.1101/gad.10.6.755
40. Price AL, Jones NC, Pevzner PA. De novo identification of repeat families in large genomes. Bioinformatics [Internet]. Oxford Univ Press; 2005;21 Suppl 1:i351–8. Available from: https://www.ncbi.nlm.nih.gov/pubmed/15961478
41. Sanchez DH, Gaubert H, Drost H-G, Zabet NR, Paszkowski J. High-frequency recombination between members of an LTR retrotransposon family during transposition bursts. Nat Commun [Internet]. 2017;8:1283. Available from: http://dx.doi.org/10.1038/s41467-017-01374-x
42. Lee J, Mun S, Meyer TJ, Han K. High Levels of Sequence Diversity in the 5′ UTRs of Human-Specific L1 Elements. Comp Funct Genomics [Internet]. 2012;2012:129416. Available from: http://dx.doi.org/10.1155/2012/129416
43. Hubley R. Dfam_consensus – A new open database of transposable element consensus sequences and representative alignments. June 18-23, 2017.
44. Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database [Internet]. 2020;2020. Available from: http://dx.doi.org/10.1093/database/baaa062
45. Appels R. Wheat research and breeding in the new era of a high-quality reference genome. Frontiers of Agricultural Science and Engineering [Internet]. journal.hep.com.cn; 2019;6:225–32. Available from: http://journal.hep.com.cn/fase/EN/abstract/abstract24931.shtml
46. Wicker T, Gundlach H, Spannagl M, Uauy C, Borrill P, Ramírez-González RH, et al. Impact of transposable elements on genome structure and evolution in bread wheat. Genome Biol [Internet]. 2018;19:103. Available from: http://dx.doi.org/10.1186/s13059-018-1479-0
47. Price AL, Eskin E, Pevzner PA. Whole-genome analysis of Alu repeat elements reveals complex evolutionary history. Genome Res [Internet]. Cold Spring Harbor Laboratory Press; 2004;14:2245–52.

Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=525682&tool=pmcentrez&rendertype=abstract

48. Levy O, Binyamin †., Knisbacher A, Erez †., Levanon Y, Havlin S. Integrating networks and comparative genomics reveals retroelement proliferation dynamics in hominid genomes [Internet]. 2017. Available from: http://advances.sciencemag.org/

Figures

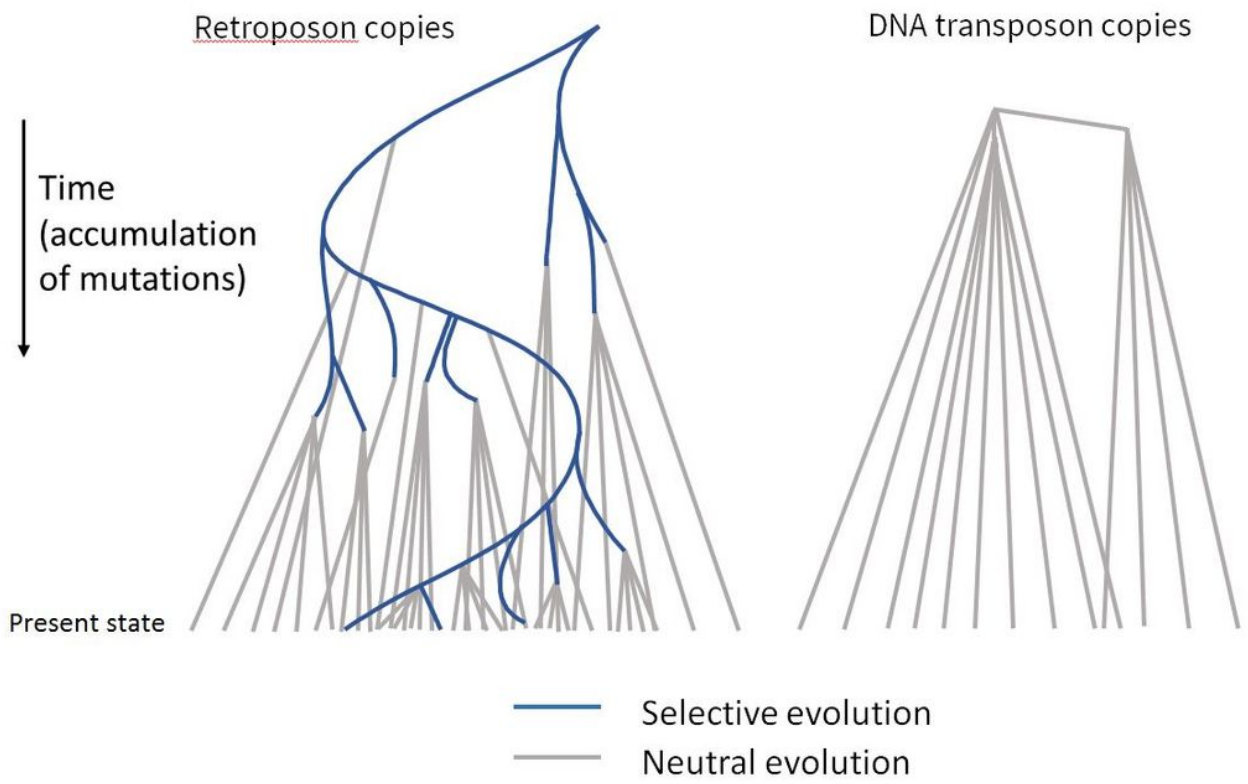

Figure 1

Typical phylogenetic structure of retroposon and DNA transposon families. After multiple mutations have occurred in the evolving class I TE, the relative ordering of copies may be distinguished by these changes as they cosegregate. The presence of such clusters or "subfamilies" of TE copies is a good indication that they arose via retrotransposition.

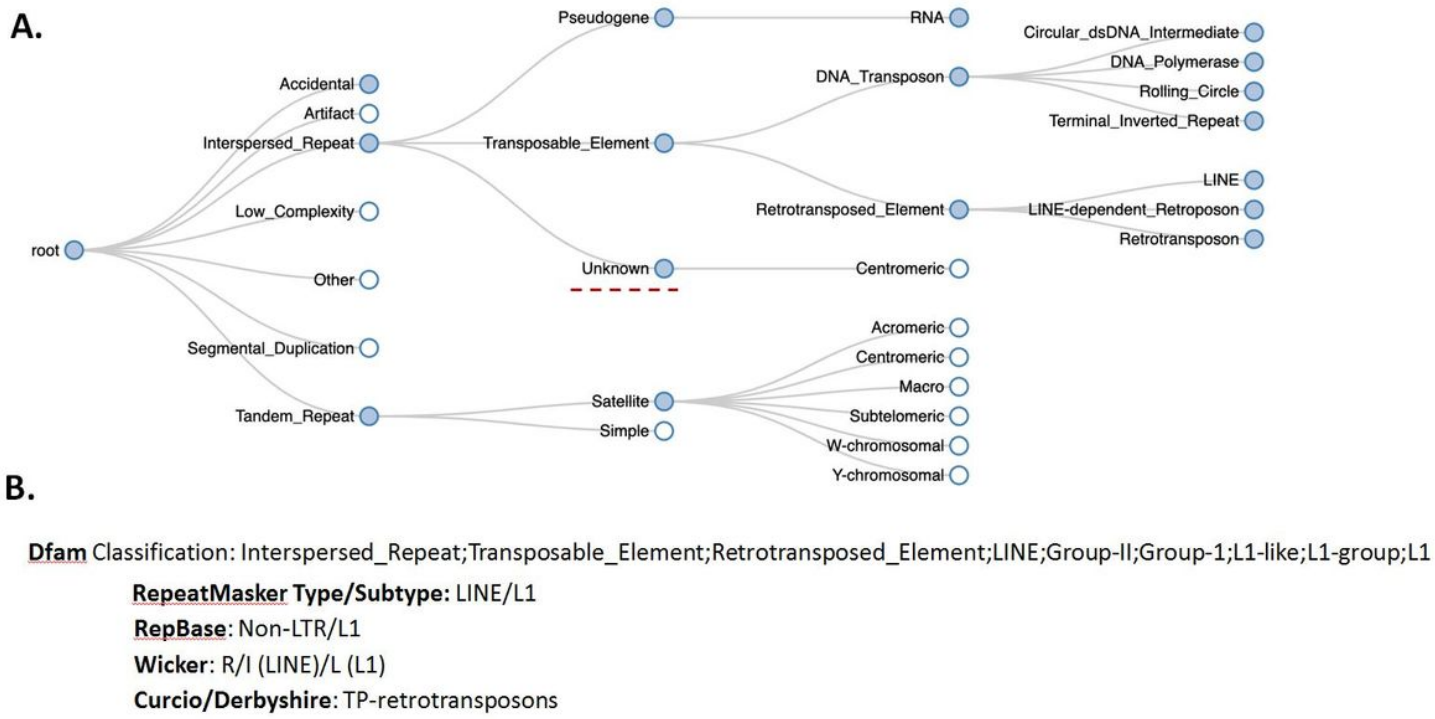

Figure 2

Dfam TE Classification System. (A) A portion of the dynamic visualization of the classification system found at the Dfam website. Filled in circles represent internal nodes of the tree while hollow circles are leaf nodes in the classification tree. A classification is specified by concatenating the path through the classification tree. For example, the classification “Interspersed_Repeat;Unknown” is highlighted in the tree. (B) In addition, wherever possible a mapping is provided between classification systems. The Dfam classification for the L1 group of LINEs is shown with the equivalent classifications in several other systems.

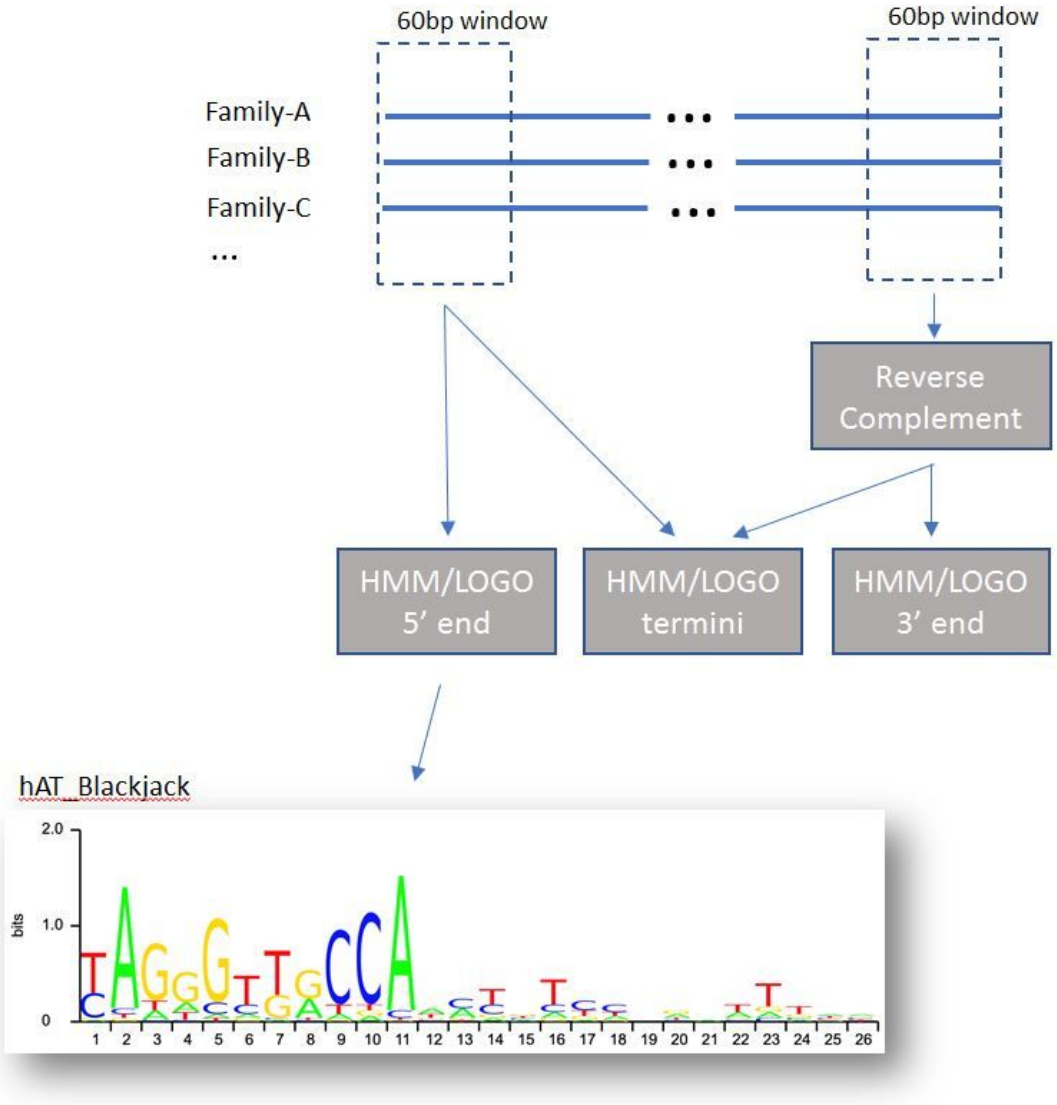

Figure 3

Generation of HMMs and sequence LOGOs for DNA Transposon Termini. The first/last 60bp of family consensi belonging to a single classification of DNA Transposons are piled up and aligned (without gaps) by hand. Profile HMMs are developed for each end and for the combination of the two to determine if a stronger signal may be obtained in that fashion. Finally, LOGOs are generated for each HMM and displayed on the Dfam website.

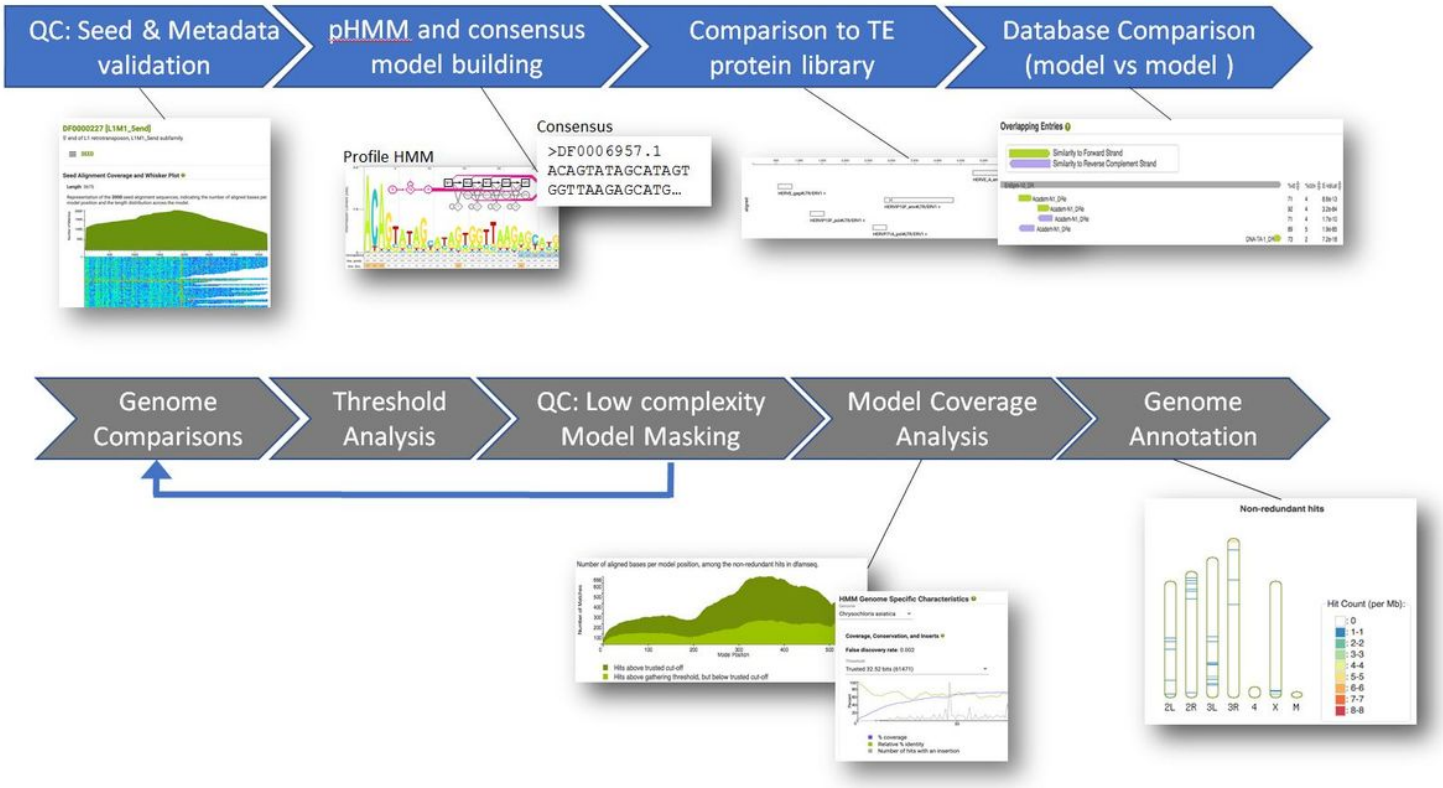

Figure 4

Dfam analysis pipeline. The full Dfam analysis pipeline consists of a set of sequential analysis steps depicted above with examples of the products produced. For uncurated families only the first portion of the pipeline (colored in blue) is initially conducted.

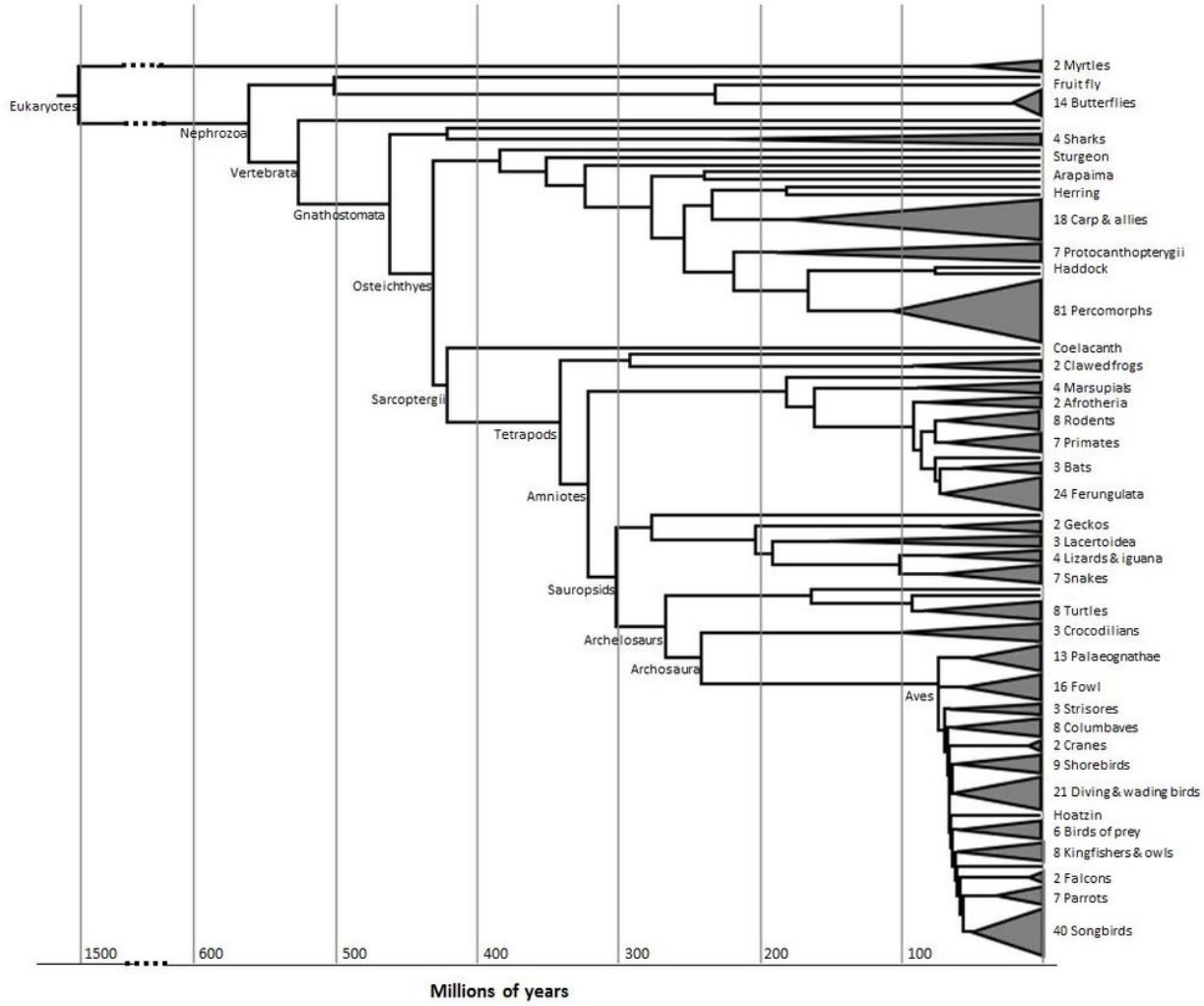

Figure 5

Phylogenetic tree of the species in the Dfam 3.2 release: Schematic of the genomes now represented in Dfam3.2. Collapsed nodes are present as triangles, with the number of species preceding a label of the species present in the triangle. A selection of species not in collapsed nodes are indicated. The dotted line indicates an axis break. Major branch points have been labeled according to the taxonomy on NCBI. Branching order within the birds follows that of Prum et al [14764687]

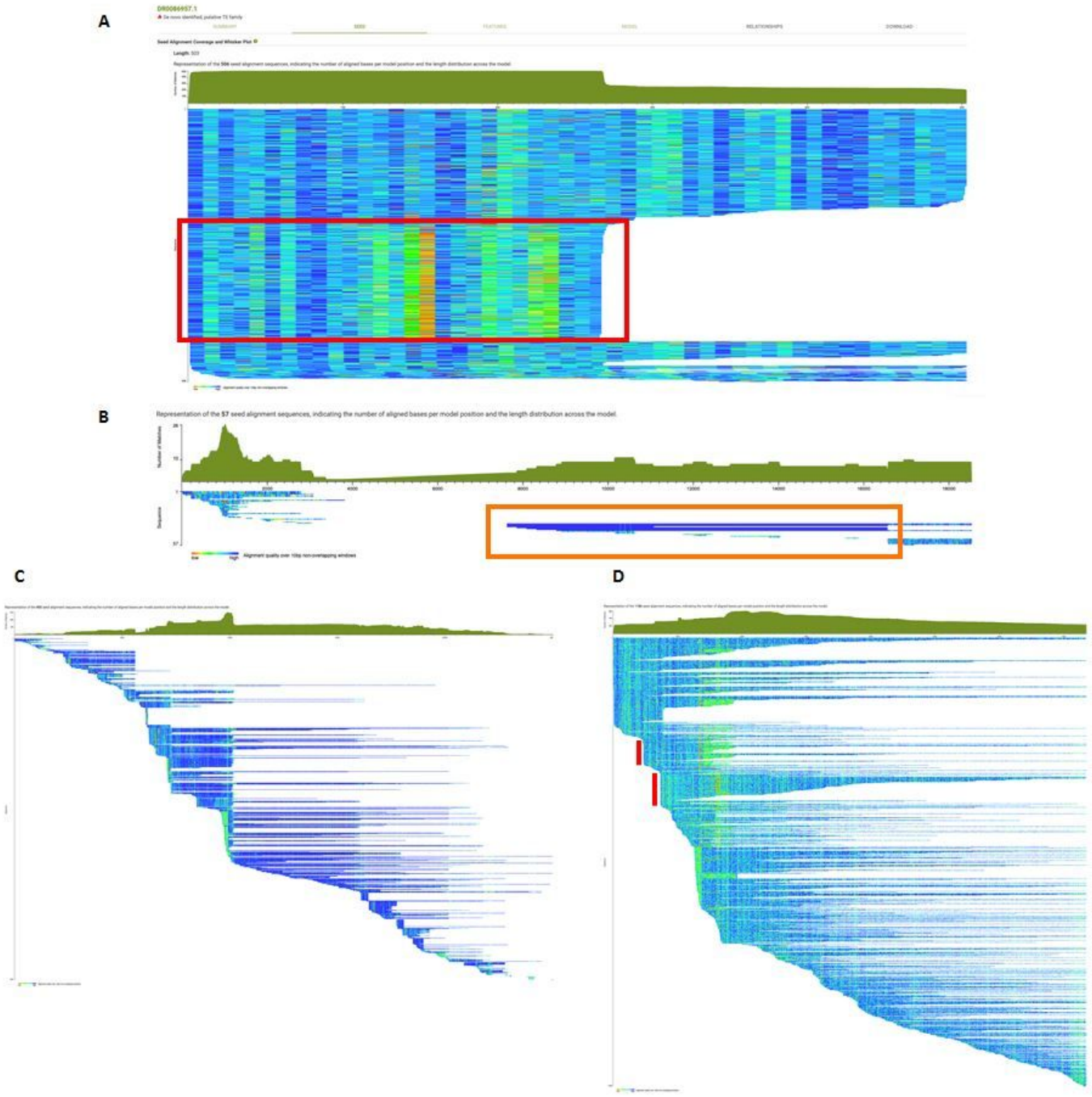

Figure 6

Seed alignment examples from raw Dfam3.2 entries. A) ERV1 (LTR) (DR0086957.1; Eulemur macaco (black lemur)). The red box indicates a group of sequences differing in length and divergence patterns. B) Unknown sequence (DR0087060.1; Eulemur macaco (black lemur)). The orange box indicates a potential segmental duplication. C) Helitron-1 (DR0096635.1; Oreochromis niloticus (tilapia)) D) L1 sequence (DR0215804.1; Phyllostomus discolor (pale spear-nosed bat)). The red lines indicate two groups of sequences that differ in their 5' alignment. Blue sequences indicate a higher match to the consensus, while red indicates a poorer match.

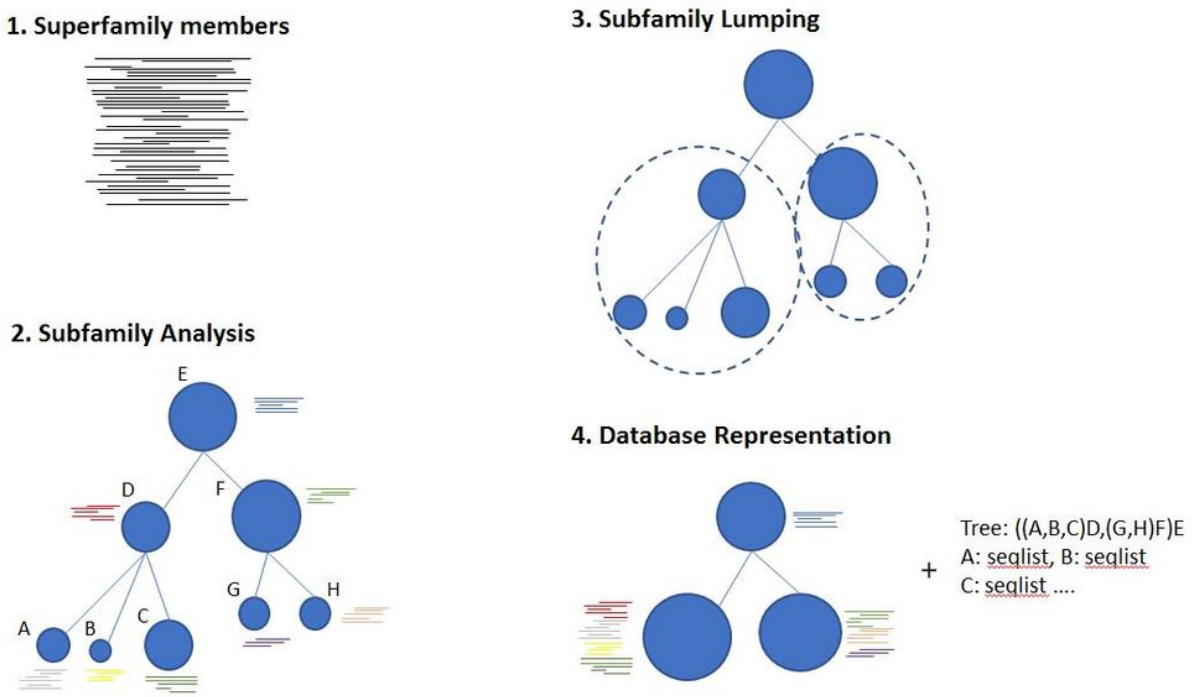

Figure 7

Database Subfamily Representation: Proposed database representation for TE subfamilies maintaining a detailed phylogenetic structure while reducing the representative models for practical genome-scale annotation. The TE seed alignment (1) from a family with evidence of subfamily structure is analysed by a clustering method to produce a detailed subfamily structure and membership (2). Sequence models are developed for subfamilies and lumped (3) if model performance isn't improved by the subdivision of two or more subfamilies. The lumped families and their corresponding seed alignments are added to the database (4) with metadata holding the detailed tree structure and seed sequence membership for each subfamily.

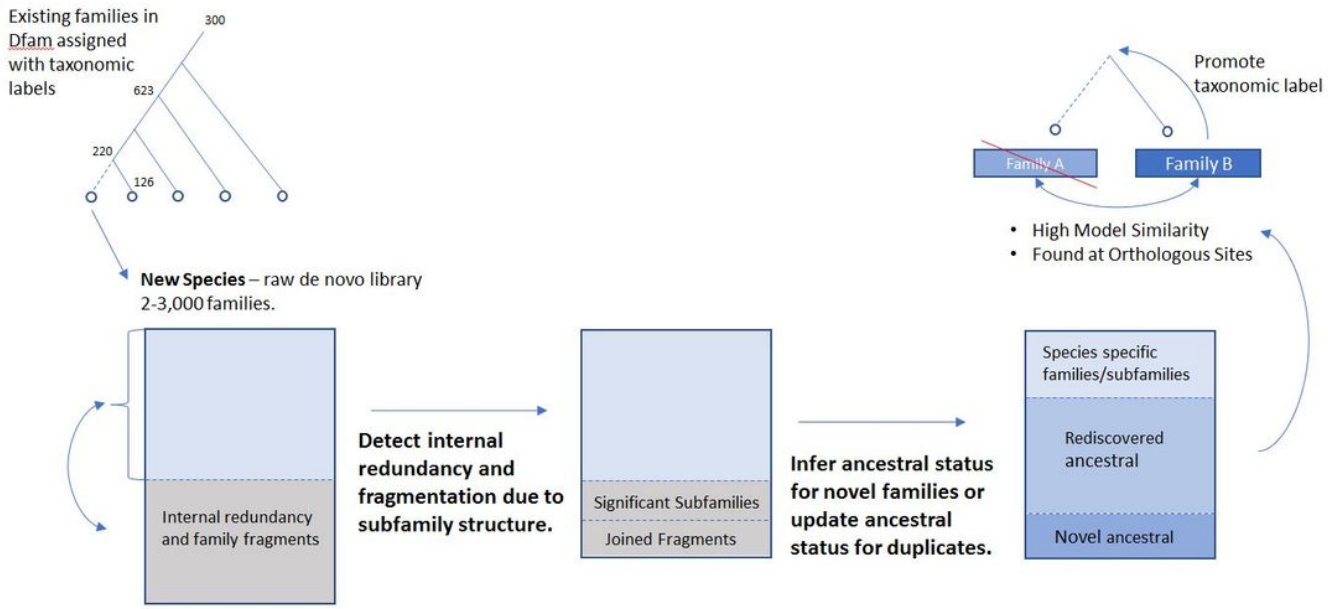

Figure 8

Redundancy/Fragmentation removal challenges. Both inter- and intra-library redundancy is present in de novo datasets and are currently resolved through manual curation. Interlibrary redundancy is often the result of unresolved subfamily structure (e.g. internal deletion products of DNA transposons) that confounds discovery and produces both redundant and fragmented families. Intra library redundancy is an inherent aspect analyzing a single species in isolation. For each new species these ancestral families need to be resolved by comparison to existing families, and by considering presence at orthologous sites.