Combined evidence annotation of transposable elements in genome sequences.

Hadi Quesneville, Casey Bergman, Olivier Andrieu, Delphine Autard, Danielle Nouaud, Michael Ashburner, Dominique Anxolabehere

To cite this version:

Hadi Quesneville, Casey Bergman, Olivier Andrieu, Delphine Autard, Danielle Nouaud, et al.. Combined evidence annotation of transposable elements in genome sequences.. PLoS Computational Biology, Public Library of Science, 2005, 1, pp.166-75. 10.1371/journal.pcbi.0010022 . inserm-00000104

HAL Id: inserm-00000104
https://www.hal.inserm.fr/inserm-00000104

Submitted on 18 Apr 2006

Combined Evidence Annotation of Transposable Elements in Genome Sequences.

Hadi Quesneville* (1), Casey M. Bergman (2), Olivier Andrieu (1), Delphine Autard (1), Danielle Nouaud (1), Michael Ashburner (2), Dominique Anxolabehere (1)

Authors' affiliations:

(1) Laboratoire Dynamique du Génome et Evolution, Institut Jacques Monod, Paris, France

(2) Department of Genetics, University of Cambridge, Downing Street, Cambridge, United Kingdom.

* Corresponding author:

Hadi Quesneville

Laboratoire Dynamique du Génome et Evolution, Institut Jacques Monod, 2, place Jussieu,

75251 Paris Cedex 05, France

tel.: (+33)-1.44.27.78.68

fax.: (+33)-1.44.27.36.60

E-mail: hq@ccr.jussieu.fr

Running head: Combined Evidence TE Annotation

Abstract

Background: Transposable elements (TEs) are mobile, repetitive sequences that comprise significant fractions of metazoan genomes. Despite their near ubiquity and importance in genome and chromosome biology, most efforts to annotate TEs in genome sequences rely on the results of a single computational method, RepeatMasker. In contrast, recent advances in gene annotation indicate that high-quality gene models can be produced from combining multiple independent sources of computational evidence.

Methodology/Principal Findings: To elevate the quality of TE annotations to a comparable status of gene models, we have developed a combined evidence-model TE annotation pipeline, analogous to systems used for gene annotation, by integrating results from multiple homology-based and *de novo* TE identification methods. As proof of principle, we have annotated "TE models" in *D. melanogaster* Release 4 genomic sequences using the combined computational evidence derived from RepeatMasker, BLASTER, TBLASTX, all-by-all BLASTN, RECON, TE-HMM and the previous Release 3.1 annotation. Our system is designed for use with the Apollo genome annotation tool, allowing automatic results to be curated manually to produce reliable annotations. The euchromatic TE fraction of *D. melanogaster* is now estimated at 5.3% (cf. 3.86% in Release 3.1) and we found a substantially higher number of TEs (n=6,013) than previously identified (n=1,572). Most of the new TEs derive from small fragments of about few hundred nucleotides long and highly abundant families not previously annotated (e.g. INE-1). We also estimated that 518 TE copies (8.6%) are inserted into at least one other TE forming a nest of elements.

Conclusions/Significance: The pipeline allows rapid and thorough annotation of even the most complex TE models, including highly deleted and/or nested elements such as those often found in heterochromatic sequences. Our pipeline can be easily adapted to other genome sequences, such as those of the *D. melanogaster* heterochromatin or other species in the genus *Drosophila*.

Introduction

Transposable elements (TEs) are mobile, repetitive DNA sequences that constitute a structurally dynamic component of genomes. The taxonomic distribution of TEs is virtually ubiquitous, they have been found in nearly all eukaryotic organisms studied with few exceptions. TEs represent quantitatively important components of genome sequences (e.g. 44.4% of the human genome, [1]), and there is no doubt that modern genomic DNA has evolved in close association with TEs. TEs show high species specificity, and the number and types of TE can differ quite dramatically between even closely related organisms. There is abundant circumstantial evidence that TEs may transfer horizontally between species by mechanisms that remain obscure. The forces controlling the dynamics of TE spread within a species are also poorly understood, as are the systemic effects of the elements on their host genomes. Insertions of individual TEs may lead to genome restructuring (e.g., the occurrence of inversions), mutations in genes or changes in gene regulation. Some TE insertions may even have become domesticated to play roles in the normal functions of the host (see [2] for review). Despite their manifold effects, abundance and ubiquity we understand very little about most aspects of TE biology.

One way of furthering our knowledge of TE biology is through the computational analysis of TEs in the growing number of complete genomic sequences. By detailed comparison of the abundance and distribution of TEs in entire genomes, we can infer the fundamental biological properties of TEs that are shared or that differ among species. However, meaningful inferences about TE biology based on computationally-derived TE annotations can only be done if we are confident about the results of these analyses. The hallmark of a strong result in computational biology should be its robustness to the particular method used. The annotation of TEs, however, typically relies on the results of a single computational method, RepeatMasker (http://www.repeatmasker.org/), which recent studies indicate may be "neither the most efficient nor the most sensitive approach" for TE annotation [3]. By contrast, recent advances in the field of gene annotation indicate that high quality gene models can be produced by combining multiple independent sources of computational evidence [4-8]. With the recent development of several new methods for TE detection [9-13], it is now possible to apply a similar "combined evidence" approach to elevate the quality of TE annotations to a comparable status as gene models.

To achieve this aim, we have developed a TE annotation pipeline that integrates results from multiple homology-based and *de novo* TE identification methods. Currently, our pipeline uses the combined computational evidence derived from RepeatMasker (http://www.repeatmasker.org/), BLASTER [12], TBLASTX, all-by-all BLASTN [14], RECON [9], TE-HMM [13], and previously published TE annotations [15]. We have designed our system to use an "evidence-model" framework and the Apollo genome annotation tool [16], allowing computational evidence to be manually curated in an efficient manner to produce reliable "TE models." The pipeline allows rapid and thorough annotation of complex TE models, providing key structural details that allow insights into the origin of highly deleted and/or nested elements. In contrast to simply masking repeats, our method provides the means to a complete and accurate annotation of TEs, supported by multiple sources of computational evidence, a goal that has important implications for experimental studies of genome and chromosome biology.

As a test case we have chosen to annotate the euchromatic genomic sequence of the fruitfly, *Drosophila melanogaster*. The 116.8 Mb Release 3 genome sequence of *D. melanogaster* is among the highest quality genome sequences and is a particularly well suited sequence for genome-wide studies of TEs, since repetitive DNA has been finished to high quality and systematically verified by restriction fingerprint analysis [17]. Moreover, the Release 3.1 annotation of *D. melanogaster* includes a manually-curated set of TE annotations [15] that can be used as a benchmark for developing and refining TE annotation methodologies. Controlled tests performed here on the Release 3 sequence show that a combined-evidence approach has superior performance over individual TE detection methods, and that a substantially larger fraction of the genome is composed of TEs than previously estimated. We have applied our pipeline to the new 118.4 Mb Release 4 sequence (http://www.fruitfly.org/annot/release4.html), which has closed several of the gaps in Release 3 and has extended the sequence of the pericentomeric regions, to produce a systematic re-annotation of TEs in the *D. melanogaster* genome. The euchromatic TE fraction is now estimated at 5.3% (cf. 3.86% in Release 3.1) and we found a substantially higher number of TEs (n=5,941) than previously identified (n=1,572). We also estimated that 518 TE copies (8.6%) are inserted into at least one other TE forming a nest of elements. Our pipeline can be easily adapted to other genome sequences, and could drastically increase the efficiency of annotating genomic regions with complex or abundant TE insertions such as heterochromatic sequences.

Results

Evaluation of methods

The first step in the development of our pipeline was to evaluate the ability of different computational tools that are available to annotate TEs in order to assess the strengths and weaknesses of each method. To do this we have re-annotated the *D. melanogaster* Release 3 sequence using different TE detection methods and compared these results to the FlyBase Release 3.1 annotation (http://www.flybase.org/annot/release3.html), which includes the results of a manually curated set of TE annotations published previously in Kaminker *et al.* (2002) [15].

Methods for TE annotation fall into two general classes: (i) those designed for the annotation of known TE families, which utilize a specific reference sequence (also called a canonical sequence); and (ii) "*ab initio*" methods designed for the annotation of anonymous TE families, for which no reference sequence has yet been identified. This distinction is necessary since it determines the relevant measures to evaluate different methods for TE detection.

Methods for the annotation of known TE families

To allow direct comparison with previous results [15], we used the Release 3 genomic sequence as a query to be scanned for similarity to reference sequences in version 7.1 of the BDGP TE data set (http://www.fruitfly.org/p_disrupt/TE.html), the same version that was used for the Release 3.1 FlyBase annotation. We initially tested three methods for TE prediction (see Materials and Methods for details): (i) BLASTER using BLASTN followed by chaining with MATCHER (called hereafter BLRn); (ii) RepeatMasker using default parameters (RM); and (iii) RM using default parameters followed by chaining with MATCHER (RMm). The last method was used to test the benefit of the “chaining algorithm” implemented in MATCHER.

We compared predictions to annotations by calculating sensitivity and specificity values from the number of nucleotides of TE sequence predicted by a method that overlap (or not) TEs in the Release 3.1 FlyBase annotation (see Material and Methods). Note that specificity is here biased as for its computation, since it assumes that all TEs in the Release 3.1 FlyBase annotation are known, which is certainly not true. Moreover, we have also compared different categories of overlap between prediction and annotation boundaries to gain deeper insight into the details of TE detection methods (see Material and Methods and Figure 3 for details). These results are summarized in Table 1.

Table 1: Results of comparisons between TE prediction methods that use reference sequences and the Release 3 FlyBase TE annotations. Relationships of predictions to annotations can be categorized as: 1-to-1, 1-to-n, n-to-1, 1-to-0, n-to-n (where n>1, see Materials and Methods for details)

	BLRn	**RM**	**RMm**	**RMBLR.opt**	**RMBLR.cons**
Sensitivity	96.9	94.3	95.8	97.8	97.8
Specificity	99.7	99.1	99.1	99.1	99.1
1-to-1					
Exact	854	664	711	717	694
Near exact	98	190	178	169	172
Equivalent	3	20	18	17	17
Near equivalent	6	37	37	28	28
One side exact	176	154	192	177	182
Similar	48	76	87	75	74
n-to-1					
Method not joined	3	110	6	4	33
Annotation over joined	14	71	28	6	8
Same TE nested	12	4	5	25	25
TE duplication	35	44	26	32	35
1-to-n					
Annotation not joined	63	45	71	61	61
1-to-0					
New TE	1515	4561	4764	4957	4996
New nest	34	24	23	30	30
Other strand	23	23	21	23	23
Different TE	31	45	45	44	44
n-to-n					
Complex structure	20	11	12	21	16

These results demonstrate that both the sensitivity and specificity to predict Release 3.1 TEs are higher for BLRn (96.9% and 99.7% respectively) than for RM (94.3% and 99.1% respectively). In addition, 28% more Release 3.1 TEs are predicted exactly by BLRn (n=854) relative to RM (n=664). BLRn also made well over an order of magnitude fewer “method not joined” errors (n=3) than RM (n=110), indicating that the BLRn strategy makes high quality automatic decisions about joining fragments of TEs. RMm has intermediate performance with respect RM and BLRn for exactly predicting Release 3.1 annotations (n=711), but, like BLRn, has few “method not joined” errors (n=6). These results may be explained partly by the fact the Release 3.1 annotation was produced using BLAST-based methods [15], and that the local alignment stop criterion significantly differs between the BLAST algorithm and the Smith and Waterman algorithm used by RM (in the final search phase). Thus the good performance of BLRn for predicting Release 3.1 TE boundaries could result from the fact that the same local alignment stop criterion has been used. However differences in local alignment matching cannot explain these results entirely, since RMm outperforms RM to recover exact matches, indicating that the chaining algorithm implemented in MATCHER is a significant improvement over raw RM results for predicting Release 3.1 TE annotations.

RM identifies approximately 3-fold more new TEs than BLRn, and thus appears to be a more sensitive method for the detection of previously unannotated TEs. But here also RMm has the better performance for detecting new TEs than RM, so the effects of chaining can also improve RM in this regard. The putative TEs predicted by RM in general are short, as can be seen by the relatively limited effect that an additional 3000+ predictions have on the genome-wide specificity of RM and RMm.

Given the different performance of these approaches, we developed and tested a fourth strategy that attempts to capitalize on the strengths of both RM and BLRn. This method, called RepeatMasker-BLASTER (RMBLR), combines hits from both BLRn and RM and gives them to MATCHER for chaining. To do this, we normalized alignment scores from BLRn and RM to be the hit length for chaining. As shown in Table 1, an optimized RMBLR (called hereafter RMBLR.opt) has higher sensitivity than RM, RMm or BLRn alone, produces the highest number of putative new TE annotations, and otherwise retains performance features similar to RMm and/or BLRn. These results show that a combined approach to TE annotation is more efficient at both recovering known and predicting new TE annotations than each method alone.

The results shown in Table 1 also suggest that there were errors in the Release 3.1 FlyBase annotation (Table 1). Among them, the tools predicted cases where two annotations could be joined automatically (category “Annotation not joined” in Table 1) and others where an annotation might be split (category “Annotation over joined” in Table 1). Using the Apollo annotation editor [16] to inspect visually these errors, we have seen that the fragmented and the nested structures of TEs often can be recovered better with these tools than in the Release 3.1 FlyBase annotation. In addition, using Apollo we have seen that the many new copies appear to be *bona fide* remnants of TEs missing from the previous annotation, however a detailed analysis of Release 4 revealed that many of these new TEs may result from spurious hits to simple repeats in the reference sequence (see below).

Methods for the annotation of anonymous TE families

We have also tested "*ab initio*" methods to predict TEs that do not use a specific reference sequence, and evaluated the ability of these methods to find TEs in the Release 3.1 *D. melanogaster* annotation. These results serve to determine the performance of each method to identify anonymous TEs, and are important for the annotation of genome sequences where a manually-curated reference set of TEs is not available. Individually, we find that these methods have lower performance than those that use specific reference sequences, but together they provide additional evidence that can be used to evaluate TE models in the final manual curation step.

TEs have been predicted anonymously using four different methods: (i) an all-by-all genome comparison with BLASTER using BLASTN followed by chaining with MATCHER and grouping with GROUPER (BLRa); (ii) RECON, using default parameters; (iii) BLASTER using TBLASTX with the entire Repbase Update as the database, followed by chaining with MATCHER (BLRtx); and (iv) a hidden Markov Model which detects TE coding sequences based on nucleotide composition (TE-HMM). Note that for BLRa, we compare coordinates of the *group* of sequences obtained by GROUPER with a coverage of 0 (*i.e.* all overlapping matches are merged; see Materials and Methods for details).

As above, sensitivity, specificity and the comparison of boundaries between predictions and annotations were used to evaluate the performance of each method. Note again that, as previously, specificity is here biased as it assumes for its computation that all TEs in the genome are known. Here, specificity may be less meaningful than above, since the ability of these methods to detect new TE is enhanced, and methods detecting many new TEs would have a correspondingly low specificity. Therefore we must be careful to interpret specificity here as the ability to detect only already known TEs.

Table 2: Results of comparisons between TE prediction methods that do not use reference sequences and the Release 3 FlyBase TE annotations. Relationships of predictions to annotations can be categorized as: 1-to-1, 1-to-n, n-to-1, 1-to-0, n-to-n (where n>1, see Materials and Methods for details)

	BLRa	RECON	BLRtx	BLRtxNoDros	TE-HMM
Sensitivity	88.3	89.8	97.2	44.2	74.3
Specificity	98.6	98.9	98	98.9	88.9
1-to-1					
Exact	5	202	126	0	0
Near exact	10	153	160	0	1
Equivalent	9	83	74	0	2
Near equivalent	14	40	75	16	23
One side exact	15	114	105	3	11
Similar	36	55	39	69	453
n-to-1					
Method not joined	9	457	1172	3587	42
Annotation over joined	209	90	44	112	283
Same TE nested	409	1	281	428	155
1-to-n					
Annotation not joined	14	53	30	11	68
1-to-0					
New TE	511	744	18260	8110	11898
n-to-n					
Complex structure	125	46	75	8	86

Table 2 shows that all *ab initio* methods have relatively high overall specificity (>88%) to detect Release 3.1 TE annotations, but that RECON gives the best performance to recover Release 3.1 TEs exactly. BLRtx has the highest overall sensitivity to detect Release 3.1 TEs (97.2%), which may be explained by the fact that this method uses Repbase Update that includes most of the *Drosophila* TEs. This can be shown by a similar analysis with *Drosophila* TEs removed from the Repbase Update (see BLRtxNoDros in Table 2), which gives lower sensitivity (44.2%), fewer new TEs (n=8110), and no "*Exact*", "*Near exact*" or "*Equivalent*" cases. BLRtx and TE-HMM detect thousands of new putative TEs relative to RECON, BLRa and the other methods detailed in Table 1, indicating that many new TE families may remain to be described in this genome [12]. These new families are probably low in copy number and represented by non-overlapping fragments, as suggested by the smaller number of new TEs found by BLRa and RECON. In fact RECON can only detect TEs that are repeated and have copies that are more or less well conserved to their extremities. BLRtx and TE-HMM would be able to detect TEs in few copies (even unique elements) that could be highly diverged and/or degenerate. It is perhaps surprising that BLRtx predicts the highest number of new TEs over TE-HMM, since TE-HMM would be able to detect copies for which no distant TE reference sequence is known. However the high number of BLRtx predictions may result from the under-joining of fragments from the same TE as suggested by the number of "*Method not joined*" cases (n=1172). Moreover its relatively high specificity indicates that most of the new TEs are rather small, as is expected if TBLASTX would mainly detect small fragments of TE coding regions. Together these results demonstrate that *ab initio* methods provide specific evidence that can be used to support TE models, however additional development is necessary to fine-tune these approaches to generate accurate TE annotations directly.

The annotation pipeline

Based on these results, we designed an integrated pipeline to compute and store evidence and TE annotations for genome sequences (Figure 1). Our annotation pipeline is composed of (i) TE detection software such as BLASTER, RepeatMasker, TE-HMM and RECON; (ii) satellite detection software such as RepeatMasker, TRF (Tandem Repeat Finder, [18]) and mreps [19] (iii) a MySQL database (http://www.mysql.com/) to manage the results of these methods and the annotations generated from them; and (iv) an OpenPBS (Open Portable Batch System, http://www.openpbs.org/) scheduling system distributing jobs on a computer cluster. The flexible architecture of this system easily allows other methods for TE detection to be added to this pipeline in the future.

To save computer time and reduce software memory requirements, we segmented the Release 4 genomic sequences into chunks of 200 kb overlapping by 10 kb. Each chunk is then independently analysed by the different analysis programs and the results are stored in the MySQL database. GAME-XML (http://www.fruitfly.org/annot/apollo/game.rng.txt) files are then generated from the results stored in the database and loaded into the Apollo genome annotation tool, allowing automatic results to be manually curated to produce a reliable annotation. For this curation we used as evidence tiers (i) the Release 3.1 FlyBase annotations with coordinates mapped to the Release 4 sequences, (ii) BLRn, RM and RMBLR results using version 9.0 of the BDGP TE reference set (iii) BLRtx using RepBase Update 8.12, and (iv) RECON, BLRa and TE-HMM (see Materials and Methods for details).

To facilitate manual curation, we automatically promoted the results or RMBLR to be the candidate annotation, which could then be validated or modified by the curator in Apollo according to the evidence available in the GAME-XML file (see Figure 2 for example). In addition, we generated a candidate list of miss-joined matches that are contiguous but not joined by MATCHER because of the size of the deletion or the insertion on the genomic sequence. This list identified potential problem cases to be considered carefully for manual joins in Apollo. Moreover, we used RMBLR with conservative settings (gap penalty of 0.05, instead of 0.04 for optimal setting), intentionally under-joining contiguous matches compared to the optimal settings (see Table 1, RMBLR.cons vs. RMBLR.opt). Hence the join decision of the most difficult cases is left to the curator. Another consequence of this conservative approach is that only a few annotations have been manually split. This happens when two small and distant fragments (generally neighboring copies of INE-1 [20]) are automatically joined, and the insert between the two fragments does not correspond to another TE (as would be the case for a nested TE). We have considered these joins excessive because of the lack of knowledge about the biology of the INE-1 TE family, for which it is difficult to find a reliable reference sequence. We initially split the 5 major chromosome arms among 5 curators for a first-pass manual curation, which was completed in less than 2 weeks. Subsequent to this, a single curator performed a second-pass manual curation in order to improve the consistency of manual edit decisions. We examined 10,348 annotations, and only 523 (5%) of them need to be edited. Finally we obtained 9,053 unique TE annotations after merging annotations in the overlaps between chunks.

During the manual curation step, we encountered an unexpectedly large number of apparently spurious hits to particular TE families resulting from similarity to simple repeats present in the reference sequence. For example, 236 of 373 predicted TEs for the *roo* family [21] are generated only by matches to the $[CA(A/G)]_n$ repeat in the *roo* reference sequence. Since the number of spurious hits resulting from simple repeats is potentially quite large, we considered several alternative strategies for their automatic removal. We rejected the possibility of masking the reference sequences and/or the genome for simple repeats, because that could have decreased dramatically the sensitivity of the detection of TEs that have many simple repeats in their reference sequence. Moreover, this strategy does not guarantee the removal of simple repeats that are too degenerate from a regular pattern to be detected, but which could still produce spurious hits because of differences in simple repeat detection versus TE detection.

Instead, we settled on a two-step post-processing of our curated predictions that first identifies all annotations that are less than a length threshold after removing regions that overlap simple repeat regions. These putative spurious hits are then used as queries in a filtered BLAST against the BDGP TE reference set to "rescue" false spurious hits (i.e. real TEs) from true spurious hits. To develop this method, we used the *roo* family as a training set, for which we could easily partition spurious from real TE annotations. We tested the ability of three methods for simple repeat detection – RepeatMasker, mreps and TRF – to discriminate real from spurious *roo* annotations as a function of length remaining after simple repeat removal. We found that using RepeatMasker with a length threshold of 170 bp allowed all 236 spurious *roo* annotations to be identified with no real annotations identified as spurious (data not shown).

Using this threshold we detected 3,058 putative spurious hits, which were then searched with BLASTN (E-value > 10e-15) using the “dust” filtering option against our reference TE sequence set. We found that only 18 of the 3,058 putative spurious hits were rescued as real annotations, indicating that our filtering thresholds have high specificity. These 3,040 putative spurious hits were removed from the final set of Release 4 TE annotations submitted to FlyBase. Finally, to understand the source of these spurious hits in the auto-promoted RMBLR.cons TE models, we analyzed the overlap of the 3040 spurious hits with Release 4 predictions generated by individually by BLRn and RM. We find that 2,898 (95%) of the spurious hits overlap a RM prediction, whereas only 1,255 (41%) of the spurious hits overlap a BLRn prediction, indicating that RM generates a greater proportion of the spurious hits than BLRn.

Discussion

We have developed and implemented a combined-evidence pipeline to annotate TEs in genome sequences and applied this novel system to the detecting TEs in the Release 4 sequence of *Drosophila melanogaster*. Our work fulfils the demand for an unified approach to TE annotation that capitalizes on the strength of multiple TE detection methods [3] and places TE annotation on common conceptual framework with gene annotation [4-8]. Compared with annotations generated for the Release 3 sequence [15], we confirmed precisely 743/1,572 TE annotations. We adjusted the boundaries of 488, joined 80, changed the strand of 66, changed the name of 14, split 16, and described 4,573 new TE annotations. (Note that the number of modifications does not total 1,572 since multiple Release 3 elements are incorporated in a single join). According to our annotation the euchromatic TE fraction is now estimated to be 5.3% (cf. 3.86% in Release 3.1) and we found a substantially higher number of TEs (n=6,013) than previously identified (n=1,572). Most of the new TEs derive from small fragments of about few hundred nucleotides long, and highly abundant families not previously annotated (e.g. INE-1). Taking into account the heterochromatic TE fraction estimated by [22] and the fraction of this compartment (1/3 of the genome), we can estimate that in *D. melanogaster* TEs represent ~20% of the whole genome (~5% of the euchromatin and ~50% of the heterochromatin). The pipeline allows rapid and thorough annotation of even the most complex TE models, including highly deleted and/or nested elements. We now estimate that 518 TE copies (8.6%) are inserted into at least one other TE forming a nest. A detailed description of abundance and distribution of TEs in Release 4 based on the result of this annotation is in preparation. The full annotation is available through FlyBase (http://www.flybase.org) and on the RepEt web site (http://dynagen.ijm.jussieu.fr/repet/)

Performance

Our studies on the Release 3 sequence provide the first detailed, genome-wide analysis of different methods for TE detection relative to a manually-curated reference set of TE annotations. These results (Tables 1 and 2) provide insight into the strengths and weaknesses of each method and therefore a deeper understanding of the consequences of algorithmic differences for TE detection. In general, our results suggest that BLRn can outperform RM with respect to the precise determination of TE boundaries, and that much of this improvement derives from the joining algorithm implemented in MATCHER. On the other hand, RM appears to be more sensitive for the detection of small and divergent TE copies. RM can detect small copies with less than 80% of identity with the reference sequence, while BLRn misses these small copies. This increase in sensitivity comes with a cost, as RM predicts many spurious hits for TE families with simple repeats in their reference sequence. Overall, we found that the differences between BLRn and RM make them very complementary for TE annotation when hits from both methods are chained with MATCHER, and that a simple-repeat-filtered version of our combined RMBLR method can be used to promote reliable TE models automatically.

There are many reasons that can explain performance differences between BLRn and RM. One obvious reason is that the initial word length used to seed the alignments is shorter for RM than for BLRn (9 for cross_match *versus* 11 for BLASTN). Another reason is that RM chose its scoring scheme (a match-mismatch matrix) according to the background GC% composition. A third explanation could come from the final Smith-Waterman alignment performed by RM, allowing it to produce longer alignment in low identity regions. Likewise, in some particularly difficult cases where a genomic TE copy has a duplicated segment, BLRn gives a better annotation because it relies only on BLASTN hits that allow a small level of overlap between adjacent hits. The final Smith-Waterman alignment performed by RM is disturbed in these cases, at best placing a gap to face the duplicated segment. The first two reasons are a matter of parameter values, and the differences may simply be due to our use of default parameters. The more sensitive parameter set of RM has a cost in term of speed, and the trade-off between speed and sensitivity between BLRn and RM is not the same (BLRn is at least 3 fold faster). Using different parameter values could improve either BLRn sensitivity and/or RM speed. It remains to be determined to what degree the sensitivity of BLRn can be improved to an equivalent level of RM just by changing the BLASTN parameters, since the use of different match-mismatch matrices (each optimal for a background GC% level) is an important difference between the two methods, and may limit BLRn sensitivity gains.

Pitfalls

From our manual edits, we are able to identify some pitfalls that could be avoided in future attempts at a fully automated TE annotation process. One of the most important problems arises from the annotation of symmetrical structures, such as Terminal Inverted Repeats (TIR) or Long Terminal Repeats (LTRs). The first concerns palindromic structures such as in the *FB* element [23]. Often the two TIRs of a genomic *FB* element are detected on different strands, i.e. the 5' TIR on the positive and the 3' TIR on the negative strand. This happens because the two TIRs are not identical in the reference sequence. Thus if the two TIRs of the genomic copy are more similar to each other than to the appropriate TIR in the reference sequence, only one TIR of the reference (the most similar one) is used to detect the two genomic TIRs, but on different strands. To avoid this type of manual edit, we suggest using a reference sequence with identical TIRs. Another similar pitfall occurs with LTR retrotransposons. If the two LTRs are not identical on the reference, a genomic copy can be detected with two 5' LTRs (or 3' LTRs) if its LTRs are more similar to each other than to the appropriate LTRs of the reference sequence. If a join is necessary because of an indel in the genomic copy, our algorithm fails since the coordinates on the reference sequence are not collinear. To avoid this, we suggest using reference sequences with identical LTRs.

Some non-LTR retrotransposons genomic copies have to be extended in 3' direction to encompass the entire the poly(A) tail. This occurs because the reference sequence has a shorter poly(A) tail than a particular genomic copy. In general, these cases are easily identified by observing an overlapping poly(A) simple repeat at the 3' end of the element. One solution to this problem is to extend the poly(A) tail of non-LTR retrotransposons in the reference set to the length of the longest observed genomic copy.

The biggest pitfall we have encountered is the problem posed by simple repeats that exist in TE reference sequences. Without a specific treatment of this problem we would have included 3,040 spurious hits – approximately one third of our original set of annotations. Filtering simple repeats on the genomic or reference sequences without affecting the sensitivity of TE detection is not obvious. We have developed an effective (but *ad-hoc*) two-step filtering strategy, but the magnitude of this problem leaves room for future improvements. Currently we employ RM to detect simple repeats, although refined parameter optimization may reveal that other more specialized simple repeat detection software such as TRF [18] or mreps [19] might be more appropriate. A careful evaluation of methods of parameters for simple repeat detection may allow decreasing our 170bp threshold and avoid the rescue step.

Regardless of the best method or criteria to detect simple repeats, the existence of simple repeats in TE reference sequences raises an important problem, since it is difficult to unambiguously determine whether a simple repeat with homology to a TE is a spurious hit or reflects a true remnant of that TE in the genome. Our methods guarantee that if we leave a spurious hit in the annotation due to homology with a simple repeat, it is more than 170bp long. Moreover, any potentially real TEs labelled as spurious hits that did not survive our rescue strategy bear no unique hallmarks of being generated by a TE. Nevertheless, the possibility of the involvement of TEs in the genesis of microsatellites [24] highlights the fundamental biological difficulty in resolving real from spurious simple repeats in a whole genome TE annotation.

Conclusions and Future Directions

We have shown in this work that a combined-evidence framework can improve the quality and confidence of TE annotations in *D. melanogaster*. Our automated pipeline allows us to annotate TEs on a genomic scale quickly and accurately, and the integration of our pipeline with the Apollo annotation tool also allows rapid evaluation and manual editing of TE annotations for even complex TE models. Based on the lessons learned in this study, we are continuing to develop and improve our pipeline. We are automating several classes of the manual edits that we have identified and expect that progressively fewer manual edits would be necessary in the future, allowing application of our pipeline to larger genome sequences such as the human. In addition, the combined-evidence framework is inherently flexible and allows inclusion of other computational methods to detect TEs as they become available in the future.

We have observed several cases in the genome annotation where one or more *ab initio* method (RECON, BLRa, BLRtx, and TE-HMM) simultaneously supports a potential sequence belonging to a new TE family. In addition, results of our analyses with tools that detect anonymous TEs (Table 2) suggests that there may be many additional families of TEs yet to be discovered in the *D. melanogaster* genome. Since the methods that support these predictions potentially suffer from high false positive rate, we have chosen not to include them in our current annotation, since more work need to be done to validate these potential new TE families. Nevertheless the combined evidence for some of these elements is compelling and such cases are available for mining in our current results.

In general, the problem of TE discovery remains a major challenge for TE annotation. A good TE annotation relies critically on an expertly assembled reference sequence set, data that currently cannot be obtained in an automatic fashion. This crucial step is now the bottleneck in any method or pipeline to annotate TEs in genome sequences (see also [3]). The task to assemble such reference set will be most difficult in genomes where only few TE families are known. In these situations, we will need good *ab initio* TE detection procedures that can only be trained and evaluated properly using high quality TE annotations in well-studied systems such as *Drosophila*. We hope that the TE annotations presented here will serve to further the development and refinement of TE discovery and annotation methods in general, as the Release 3.1 annotations have served for the development our current methods.

Finally, we are also developing our pipeline to include methods for the detailed annotation of the structural features (ORFs, LTRs, etc) in TE sequences. Development of such detailed annotation methodologies will allow a detailed evaluation of the coding and expression potential of individual TE annotations in genomic sequences. Moreover, the ability to automatically annotate structural features of TEs will facilitate the manual curation and validation of candidate TE sequences resulting from one or several different *ab initio* TE discovery methods. Continued development of this pipeline, together with other advances in the field of TE genome informatics, will lead to a robust computational framework which can shed light on the origin and impact of TEs in modern genomes.

Material and Methods

Data

The *D. melanogaster* genomic sequences and TE references sets are available at BDGP (Berkeley *Drosophila* Genome Project) web site (http://www.fruitfly.org/p_disrupt/TE.html). The Release 3.1 *D. melanogaster* genomic sequences and their TE annotations have been extracted from the GAME-XML files. The Release 4 *D. melanogaster* genomic sequences have been downloaded as fasta files. TE reference sequence sets v.7.1 (used by Kaminker *et al.* 2002 [15]) and v.9.0 have been downloaded from BDGP web site.

Sequences of the transposable elements were also obtained from the Repbase Update database release 8.12 [25], which contains all known repeated sequences including TEs (downloaded from http://www.girinst.org). We use them to detect unknown families by similarity with TEs from other species.

Sequence Analysis software

We have improved three C++ programs: BLASTER, MATCHER, and GROUPER, previously presented in Quesneville *et al.* (2003):

BLASTER

This program can compare two sets of sequences: a query databank against a subject databank. For each sequence in the query databank, BLASTER launches one of the BLAST programs (BLASTN, TBLASTN, BLASTX, TBLASTX, BLASTP, MegaBLAST) [14,26-28] to search the subject databank. Each BLAST search is launched in parallel on a computer cluster. BLASTER is not limited by the length of sequences. It cuts long sequences before launching BLAST and re-assembles the results afterwards. Hence, it can work on whole genomes, in particular to compare a genome with itself to detect repeats. The results of BLASTER can then be treated by the MATCHER and GROUPER programs described below. For the experiments conducted here, NCBI-BLAST2 (ftp://ftp.ncbi.nlm.nih.gov/blast/) programs are used with default parameters, using as a query genomic fragments of 50kb, overlapping by 100bp.

MATCHER

This program has been developed to map match results onto query sequences by first filtering overlapping hits. When two matches overlap on the genomic (query) sequence, the one with the best alignment score is kept, the other is truncated so that only non-overlapping regions remain on the match. As a result of this procedure a match is totally removed only if it is included in a longer one with a best score. All matches that have E-value > 1e-10 or length <=20 are eliminated

Long insertions (or deletions) in the query or subject could result in two matches, instead of one with a long gap. Thus the remaining matches are chained by dynamic programming. A score is calculated by summing match scores and subtracting a gap penalty (0.05 times the gap length) and a mismatch penalty (0.2 times the mismatch length region) as in [29].

The chaining algorithm [30] (pp. 325-329) is modified to produce local alignments. A match is chained with a chain of other matches, only if the resulting score is greater than the score of the match alone. Thus, the chaining is stopped if the score of the resulting chain of matches is less than if the match is not chained. The best scoring chain is kept. Then to identify other match chains, the chain previously found is removed, and we search again for the next best match chain. This is done iteratively until no chain is found. This algorithm is repeated independently for match on strand +/+, +/-, and -/+. A maximum of 20% of overlap between the matches is allowed. The chaining algorithm allows the recovery of TEs containing long insertions, and therefore can identify nested elements accurately as they appear as a long insertion inside another TE.

GROUPER

This program uses matches (or chained matches) to gather similar sequences into *groups* by simple link clustering. A match belongs to a *group* if one of the two matching sequence coordinates overlaps a sequence coordinate of this *group* by more than a given length coverage percentage threshold (a program parameter). If the two matches overlap with this constraint, their coordinates are merged taking the extremum of the both. *Groups* that share sequence locations - not previously grouped because of a too low length coverage percentage - are regrouped into what we call a *cluster*. As a result of these procedures, each *group* contains sequences that are homogeneous in length. A given region may belong to several *groups*, but all of these groups belong to the same *cluster*.

RepeatMasker

RepeatMasker (http://repeatmasker.genome.washington.edu/) screens for TEs and low complexity DNA sequences. It detects TEs in nucleic acid sequences by nucleic sequence alignment with previously characterized elements using the program cross_match (http://www.phrap.org/phredphrapconsed.html) or WU-BLAST (http://blast.wustl.edu) with the script MaskerAid [31]. Both alignment program perform their Smith-Waterman alignments by first identifying exact word matches and restricting the alignment to a band or matrix surrounding this exact match(es). According to the background GC% composition, different similarity matrices (each optimal for a background GC% level) are used. It annotates the parts of sequences that are very similar to an element from a reference set of "known elements". Low-complexity DNA regions are detected when stretches of nucleotides are GC- or AT-rich. Simple repeats are detected by searching all di- to pentameric and some hexameric repeats, possibly polymorphic.

RECON

RECON [9] is an automated *de novo* identification of new repeat sequence families in sequenced genomes. It searches genomic sequences for long repeats and clusters them in groups of similar sequences. TE copies from a given family are expected to cluster together. Its algorithm cluster repeats obtained by an all-by-all sequence comparisons (here using BLASTER with blastn) and redefined the clusters by the aggregation of endpoints in a multiple alignment of the identified regions. In that way it tends to distinguish true TE copies from copies in a segmental duplication.

TE-HMM

We have shown previously how base compositional differences can be used as a tool for detection and analysis of novel TE sequences [13]. Hidden Markov Models (HMM) are used to take into account the base composition of the sequences and the heterogeneity between coding and non-coding parts of sequences. We use three sets of sequences from *D. melanogaster* containing class I TEs, class II TEs and cellular genes. Each of these sets has a distinct, homogeneous composition, enabling us to distinguish between the two classes of TE and the genes. This approach can be used to detect and annotate TEs in genomic sequences and complements the current homology-based TE detection methods. Furthermore, the HMM method is able to identify the parts of a sequence in which the nucleotide composition resembles that of a coding region of a TE. This is useful for the detailed annotation of TE sequences, which may contain an ancient, highly diverged coding region that is no longer fully functional.

Comparison of Predictions and Annotations

We have automatically compared predictions obtained with different computational methods to the Release 3.1 TE reference annotations in two ways, each implemented in a custom python script.

The first calculates the nucleotide overlaps between the predictions and reference annotations, and computes the genome-wide sensitivity and the specificity. These values are obtained from formula (1) and (2) and the nucleotide counts of true positive (TP - correctly annotated as belonging to a TE), false positive (FP - falsely predicted as belonging to a TE), true negative (TN - correctly annotated as not belonging to a TE), and false negative (FN, falsely predicted as not belonging to a TE).

$$Sensitivity = \frac{TP}{TP + FN} \qquad (1)$$

$$Specificity = \frac{TN}{TN + FP} \qquad (2)$$

A high sensitivity indicates that a method misses few TE nucleotides (false negative). A high specificity indicates that a method finds few false positive nucleotides.

The second python script compares the boundaries of predictions to the boundaries of the reference annotations. For each prediction under test, we search the reference annotations that overlap on the same genomic region. Different cases could be distinguished according to one-to-one, one-to-many, many-to-one, or many-to-many relationships (see Figure 3 for details).

For those that have a one-to-one correspondence with the same TE family, we calculate the difference in distances between predictions and annotations for their respective 5’ and 3’ coordinates. We categorize the difference in distance into 3 classes: ≤1, ≤10 or >10 bp. We call *“Exact”* annotations which have distance at both extremities ≤1 bp, *“Near exact”* when one is ≤1 bp and the other >1 bp and ≤10 bp, *“Exact one side”* when one is ≤1 bp and the other >10 bp. Cases where both distances are >1 bp and ≤10 bp are called *“Equivalent”*, *“Near equivalent”* if one is >1 bp and ≤10 bp and the other >10 bp, and *“Similar”* if both distances are >10 bp.

We also consider many-to-one relationships. Some are method errors when a genomic copy (given by the reference annotation) has a large insertion or deletion. In this case, the two fragments (flanking the indel) are predicted as two separate copies, the fragments are not joined. We call this error class: *"Method not joined"*. We have also found cases where two predictions are falsely considered as one in the reference annotation. Here, a long region of mismatch separates two fragments and the most parsimonious explanation is the independent insertion of two copies. These are *"Annotation over-joined"* cases. We have also found cases considered as one copy by the reference annotation, but that are in fact copies with a self-duplicated region. If the duplication is nested we call it *"Same TE nested"*, or if not nested, *"TE duplication"*.

One-to-many relationships are cases where two annotations in the reference are found joined by the method. We call this *"Annotation not joined"*.

One-to-zero relationships correspond to cases where a prediction does not correspond to a reference annotation. *"New TE"* are copies identified by the method under test but not present in the reference annotation, and *"Different TE"* are those overlapping a reference annotation but with a different TE family name. A TE prediction included in a prediction of a different family already involved in a given relationship with reference annotations, is called *"New nest"* if no corresponding reference annotation can be found. Annotation correspondence of the same TE family but on different strand is called "*Other strand*" if the relationship is one-to-one, otherwise they are *"New TE"*.

Finally we have a *"Complex structure"* case when the relation is many-to-many.

The script can be also used in an anonymous mode to test boundaries of *"ab initio"* predictions that do not use a specific reference sequence. The information used for such comparisons is of poorer quality since we do not have alignment coordinates on the reference sequence (i.e. RECON, TE-HMM), which renders several categories meaningless (e.g. *"Different TE",* but also *"New nest", "Other strand",* and "*TE duplication"*).

Acknowledgments

We thank Clémentine Vitte for help testing of BLASTER; Chris Mungall for early comparative analyses on Release 3; Emmanuel Mongin for the initial Release 4 mapping of Release 3.1 TE annotations; and Dave Emmert for assistance loading our Release 4 TE annotations into FlyBase. This work was supported by the "Centre National de Recherche

Scientifique" (CNRS), the Universities P. and M. Curie and D. Diderot (Institut Jacques Monod, UMR 7592, Dynamique du Génome et Evolution) and by the "programme Bio-Informatique" (CNRS). C. Bergman was supported by a Royal Society USA Research Fellowship. Work in M. Ashburner's laboratory was supported by an MRC Programme Grant to M.A. and S. Russell.

Abbreviations

BDGP, Berkeley *Drosophila* Genome Project; HMM, Hidden Markov Model; OpenPBS, Open Portable Batch System; ORF, open reading frame; TE, Transposable Element; TRF, Tandem Repeat Finder; TIR, Terminal Inverted Repeats; LTR, Long Terminal Repeats.

References

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860-921.
2. Kidwell MG, Lisch DR (2001) Perspective: transposable elements, parasitic DNA, and genome evolution. Evolution Int J Org Evolution 55: 1-24.
3. Juretic N, Bureau TE, Bruskiewich RM (2004) Transposable element annotation of the rice genome. Bioinformatics 20: 155-160.
4. Mungall CJ, Misra S, Berman BP, Carlson J, Frise E, et al. (2002) An integrated computational pipeline and database to support whole-genome sequence annotation. Genome Biol 3: RESEARCH0081.
5. Allen JE, Pertea M, Salzberg SL (2004) Computational gene prediction using multiple sources of evidence. Genome Res 14: 142-148.
6. Ding L, Sabo A, Berkowicz N, Meyer RR, Shotland Y, et al. (2004) EAnnot: a genome annotation tool using experimental evidence. Genome Res 14: 2503-2509.
7. Ashurst JL, Chen CK, Gilbert JG, Jekosch K, Keenan S, et al. (2005) The Vertebrate Genome Annotation (Vega) database. Nucleic Acids Res 33: D459-465.
8. Haas BJ, Wortman JR, Ronning CM, Hannick LI, Smith RK, Jr., et al. (2005) Complete reannotation of the *Arabidopsis* genome: methods, tools, protocols and the final release. BMC Biol 3: 7.
9. Bao Z, Eddy SR (2002) Automated *de novo* identification of repeat sequence families in sequenced genomes. Genome Res 12: 1269-1276.
10. Biedler J, Tu Z (2003) Non-LTR retrotransposons in the African malaria mosquito, *Anopheles gambiae*: unprecedented diversity and evidence of recent activity. Mol Biol Evol 20: 1811-1825.
11. McCarthy EM, McDonald JF (2003) LTR_STRUC: a novel search and identification program for LTR retrotransposons. Bioinformatics 19: 362-367.
12. Quesneville H, Nouaud D, Anxolabehere D (2003) Detection of new transposable element families in *Drosophila melanogaster* and *Anopheles gambiae* genomes. J Mol Evol 57 Suppl 1: S50-59.
13. Andrieu O, Fiston AS, Anxolabehere D, Quesneville H (2004) Detection of transposable elements by their compositional bias. BMC Bioinformatics 5: 94.
14. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403-410.

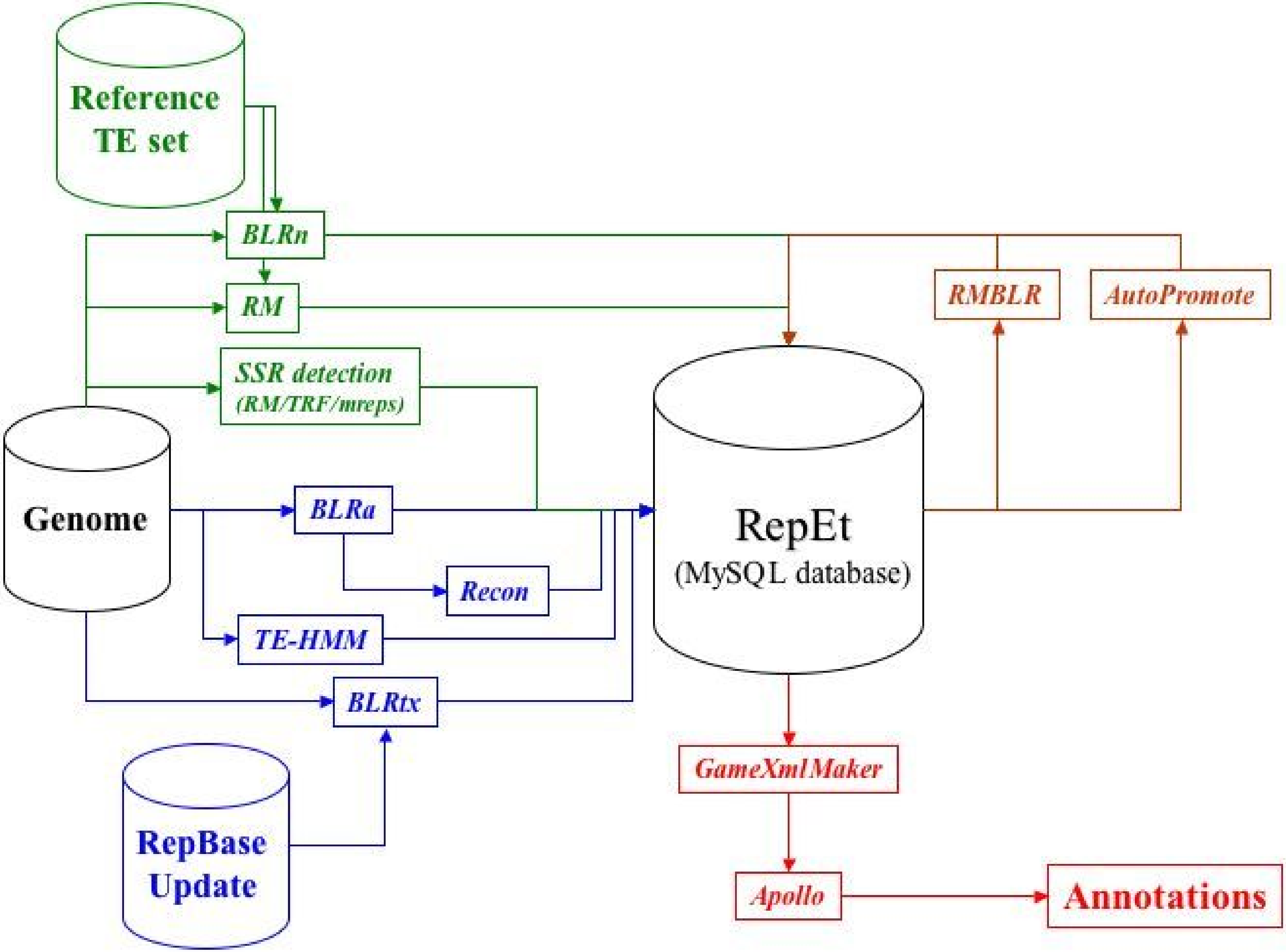
15. Kaminker JS, Bergman CM, Kronmiller B, Carlson J, Svirskas R, et al. (2002) The transposable elements of the *Drosophila melanogaster* euchromatin: a genomics perspective. Genome Biol 3: RESEARCH0084.
16. Lewis SE, Searle SM, Harris N, Gibson M, Lyer V, et al. (2002) Apollo: a sequence annotation editor. Genome Biol 3: RESEARCH0082.
17. Celniker SE, Wheeler DA, Kronmiller B, Carlson JW, Halpern A, et al. (2002) Finishing a whole genome shotgun sequence assembly: release 3 of the *Drosophila* euchromatic genome sequence. Genome Biology 3: RESEARCH0079.
18. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27: 573-580.
19. Kolpakov R, Bana G, Kucherov G (2003) mreps: Efficient and flexible detection of tandem repeats in DNA. Nucleic Acids Res 31: 3672-3678.
20. Locke J, Howard LT, Aippersbach N, Podemski L, Hodgetts RB (1999) The characterization of DINE-1, a short, interspersed repetitive element present on chromosome and in the centric heterochromatin of *Drosophila melanogaster*. Chromosoma 108: 356-366.
21. Meyerowitz EM, Hogness DS (1982) Molecular organization of a *Drosophila* puff site that responds to ecdysone. Cell 28: 165-176.
22. Hoskins RA, Smith CD, Carlson JW, Carvalho AB, Halpern A, et al. (2002) Heterochromatic sequences in a *Drosophila* whole-genome shotgun assembly. Genome Biol 3: RESEARCH0085.
23. Potter S, Truett M, Phillips M, Maher A (1980) Eucaryotic transposable genetic elements with inverted terminal repeats. Cell 20: 639-647.
24. Wilder J, Hollocher H (2001) Mobile elements and the genesis of microsatellites in dipterans. Mol Biol Evol 18: 384-392.
25. Jurka J (2000) Repbase update: a database and an electronic journal of repetitive elements. Trends Genet 16: 418-420.
26. Gish W, States DJ (1993) Identification of protein coding regions by database similarity search. Nat Genet 3: 266-272.
27. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402.
28. Zhang Z, Schwartz S, Wagner L, Miller W (2000) A greedy algorithm for aligning DNA sequences. J Comput Biol 7: 203-214.
29. Chao KM, Zhang J, Ostell J, Miller W (1995) A local alignment tool for very long DNA sequences. Comput Appl Biosci 11: 147-153.
30. Gusfield D (1997) Algorithms on strings, trees, and sequences: computer science and computational biology. New York, NY: Cambridge University Press. 534 p.
31. Bedell JA, Korf I, Gish W (2000) MaskerAid: a performance enhancement to RepeatMasker. Bioinformatics 16: 1040-1041.

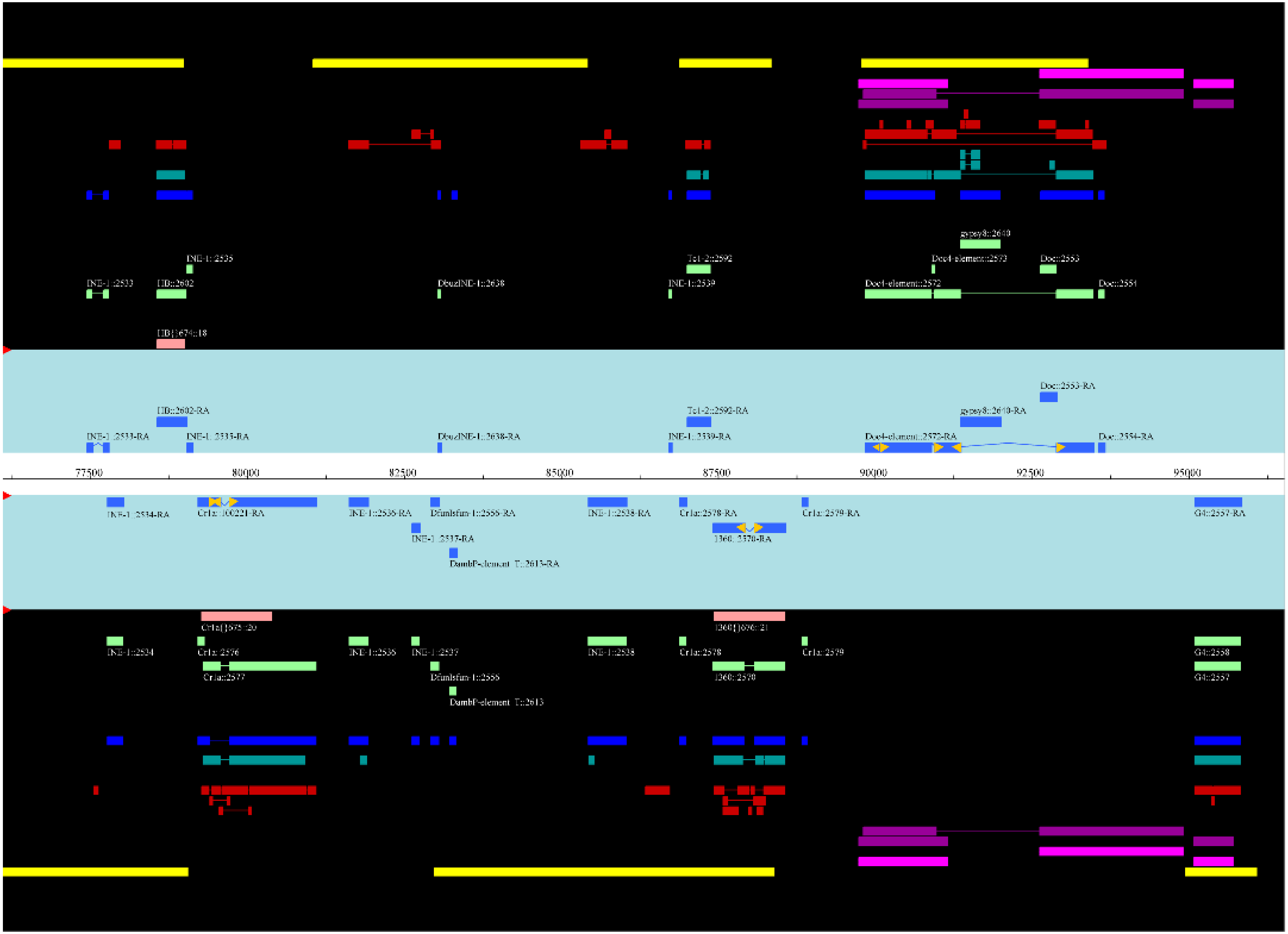
Figure Legends

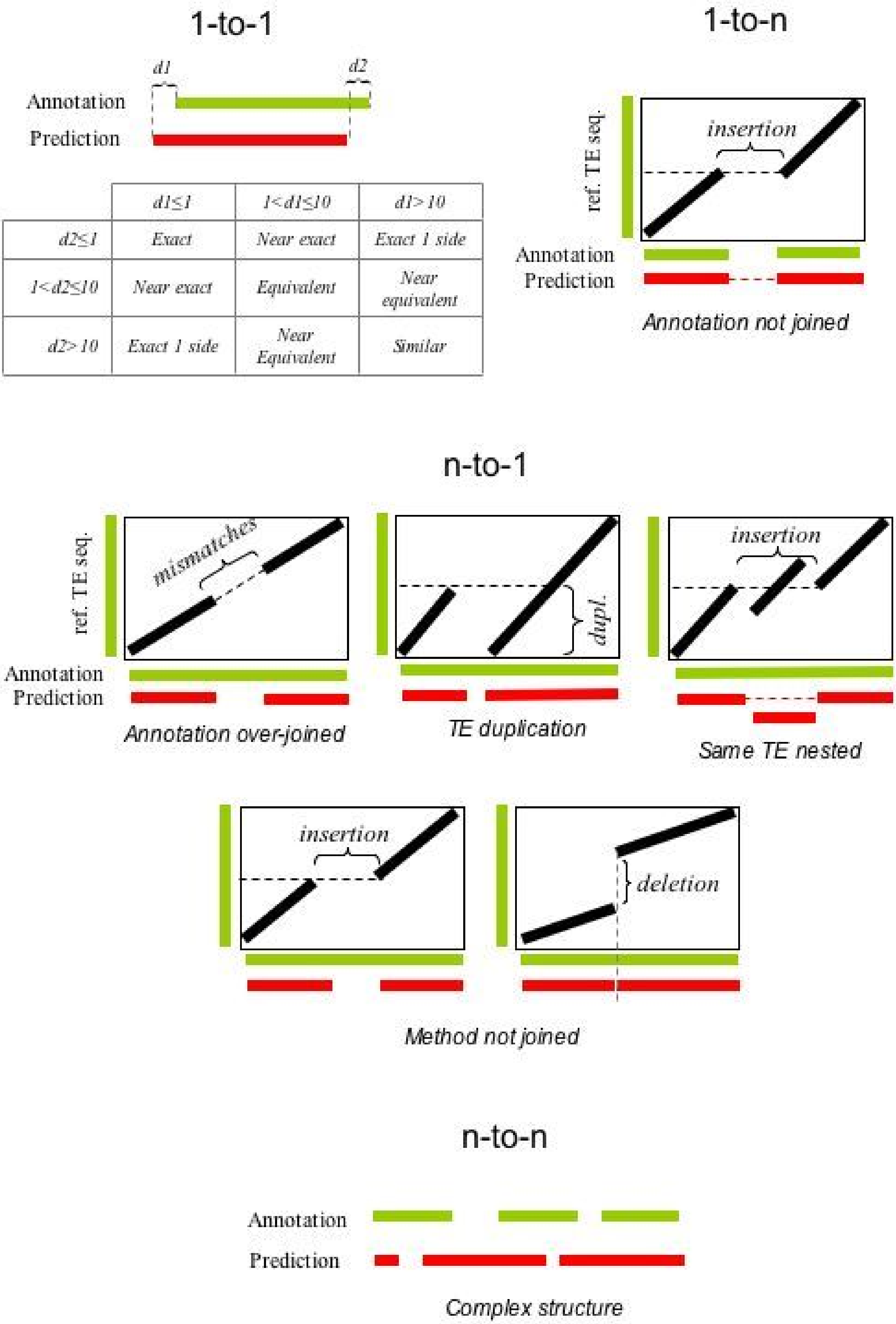
Figure 1: Schematic of our TE annotation pipeline. The pipeline is composed of (i) known TE families detection methods such as BLRn, RM, and RMBLR; (ii) satellite detection software such as RM, TRF, and mreps; (iii) anonymous TE detection methods: BLRa, TE-HMM, RECON, and BLRtx; (iv) a MySQL database called RepEt to manage the results and the annotations. GAME-XML files are then generated from the results stored in the database and loaded into the Apollo genome annotation tool, allowing automatic results to be manually curated to produce a reliable annotation. To facilitate manual curation, we automatically promoted RMBLR results to be the candidate annotation.

Figure 2: Screenshot of an Apollo view for a peri-centromeric region with extreme TE density. Curated annotations on both forward strand (top) and reverse strand (bottom) are displayed in the light blue panels. Evidence tiers are shown in the black panels: Release 3.1 FlyBase annotations (light pink), BLRn (grey-blue), RM (blue) RMBLR (light green), BLRtx (red), RECON (pink), BLRa (violet) and TE-HMM (yellow).

Figure 3: Categories of possible boundary comparisons between predictions and reference annotations. The different cases taken into account can be grouped according to a one-to-one (1-to-1), one-to-many (1-to-n), many-to-one (n-to-1), or many-to-many (n-to-n) relationships.


Reference TE set
BLRn
RM
RMBLR
AutoPromote
SSR detection (RM/TRF/mreps)
Genome
BLRa
RepEt (MySQL database)
Recon
TE-HMM
BLRtx
RepBase Update
GameXmlMaker
Apollo
Annotations


gypsy8::2640
INE-1::2535
Tc1-2::2592
Doc4-element::2573
Doc::2553
INE-1::2533
HB::2602
DbuzINE-1::2638
INE-1::2539
Doc4-element::2572
Doc::2554
HB{}674::18
Doc::2553-RA
HB::2602-RA
Tc1-2::2592-RA
gypsy8::2610-RA
INE-1::2533-RA
INE-1::2535-RA
DbuzINE-1::2638-RA
INE-1::2539-RA
Doc4-element::2572-RA
Doc::2554-RA
77500
80000
82500
85000
87500
90000
92500
95000
INE-1::2534-RA
Cr1a::100221-RA
INE-1::2536-RA
Dfunlsfun-1::2566-RA
INE-1::2538-RA
Cr1a::2578-RA
Cr1a::2579-RA
G4::2557-RA
INE-1::2537-RA
1360::2570-RA
DambP-element_T::2613-RA
Cr1a{}575::20
1360{}676::21
INE-1::2534
Cr1a::2576
INE-1::2536
INE-1::2537
INE-1::2538
Cr1a::2578
Cr1a::2579
G4::2558
Cr1a::2577
Dfunlsfun-1::2566
1360::2570
G4::2557
DambP-element_T::2613



	d1≤1	*1<d1≤10*	*d1>10*
d2≤1	*Exact*	*Near exact*	*Exact 1 side*
1<d2≤10	*Near exact*	*Equivalent*	*Near equivalent*
d2>10	*Exact 1 side*	*Near Equivalent*	*Similar*