Metadata of the chapter that will be visualized online

Chapter Title	Identification and Analysis of Ingi-Related Retroposons in the Trypanosomatid Genomes	
Copyright Year	2014	
Copyright Holder	Springer Science+Business Media New York	
Corresponding Author	Family Name	**Bringaud**
	Particle	
	Given Name	**Frédéric**
	Suffix	
	Division	Centre de Résonance Magnétique des Systèmes Biologiques (RMSB), UMR 5536 CNRS
	Organization	Université Bordeaux Segalen
	Address	146, rue Léo Saignat, 33076, Bordeaux, France
	Email	bringaud@rmsb.u-bordeaux2.fr
Author	Family Name	**Rogers**
	Particle	
	Given Name	**Matthew**
	Suffix	
	Division	Centre for Vaccine Research
	Organization	University of Pittsburgh
	Address	Biomedical Science Tower 3, Pittsburgh, PA, USA
Author	Family Name	**Ghedin**
	Particle	
	Given Name	**Elodie**
	Suffix	
	Division	Centre for Vaccine Research
	Organization	University of Pittsburgh
	Address	Biomedical Science Tower 3, Pittsburgh, PA, USA
Abstract	Transposable elements (TE), defined as discrete pieces of DNA that can move from one site to another site in genomes, represent significant components of eukaryotic genomes, including trypanosomatids. Up to 5 % of the trypanosomatid genome content is composed of retroposons of the ingi clade, further divided into subclades and subfamilies ranging from short extinct truncated elements (SIDER) to long active elements (ingi). Important differences in ingi-related retroposon content have been reported between trypanosomatid species. For instance, *Leishmania*	

spp. have expanded and recycled a whole SIDER family to fulfill an important biological pathway, i.e., regulation of gene expression, while trypanosome genomes are primarily composed of active elements. Here, we present an overview of the computational methods used to identify, annotate, and analyze ingi-related retroposons for providing a comprehensive picture of all these TE families in newly available trypanosomatid genome sequences.

Keywords (separated by "-")

Transposable element - Ingi-related retroposon - Trypanosomatid - Computational methods - Identification - Annotation - Classification - Evolution - Consensus sequence - (Sub)family

Chapter 6

Identification and Analysis of Ingi-Related Retroposons in the Trypanosomatid Genomes

Frédéric Bringaud, Matthew Rogers, and Elodie Ghedin

Abstract

Transposable elements (TE), defined as discrete pieces of DNA that can move from one site to another site in genomes, represent significant components of eukaryotic genomes, including trypanosomatids. Up to 5 % of the trypanosomatid genome content is composed of retroposons of the ingi clade, further divided into subclades and subfamilies ranging from short extinct truncated elements (SIDER) to long active elements (ingi). Important differences in ingi-related retroposon content have been reported between trypanosomatid species. For instance, *Leishmania* spp. have expanded and recycled a whole SIDER family to fulfill an important biological pathway, i.e., regulation of gene expression, while trypanosome genomes are primarily composed of active elements. Here, we present an overview of the computational methods used to identify, annotate, and analyze ingi-related retroposons for providing a comprehensive picture of all these TE families in newly available trypanosomatid genome sequences.

Key words Transposable element, Ingi-related retroposon, Trypanosomatid, Computational methods, Identification, Annotation, Classification, Evolution, Consensus sequence, (Sub)family

1 Introduction

Mobile genetic elements, also called transposable elements (TEs), can be defined as DNA fragments that can move into new locations in the host genome by excision or replication of an existing copy. Despite their abundance in most genomes (over 40 % of the human genome [1]), TEs are often called "junk," "selfish," or "parasitic" DNA, because they appear as functionless DNA sequences replicating themselves into as many copies as possible. This view began to change in the early 1990s and now tends to be replaced by a "functionalist" view of TE biology. This is supported by a rapidly increasing number of reports describing domestication or exaptation of TE to play a role in cellular function, such as transcriptional regulation, and contribution to protein-coding regions (for a recent review *see* [2]). Since most of these are degenerate and extinct elements that are barely detectable as TEs [3–5],

Christopher Peacock (ed.), *Parasite Genomics Protocols*, Methods in Molecular Biology, vol. 1201,
DOI 10.1007/978-1-4939-1438-8_6,

developing bioinformatics tools and approaches to identify highly degenerate TEs is a major challenge. The aim of this chapter is to provide a bioinformatics approach (including a workflow) to identify and annotate active, as well as extinct, degenerate TEs using as example the trypanosomatid ingi retroposon family, which contains highly degenerate sequences involved in the regulation of gene expression [5].

The trypanosomatid family includes some of the most important protist parasites of humans in the genera *Leishmania* and *Trypanosoma*, as well as other species parasitic in a wide variety of vertebrates, invertebrates, ciliates, and plants [6]. The genomes of nine trypanosomatids have been sequenced and published to date, i.e., *T. brucei* [7], *T. b. gambiense* [8], *T. cruzi* [9], *L. major* [10], *L. braziliensis* [11], *L. infantum* [11], *L. mexicana* [12], *L. donovani* [13], and *L. tarentolae* [14]. In addition, a number of ongoing trypanosomatid genome projects have data available online (http://tritrypdb.org/tritrypdb/showXmlDataContent.do?name=XmlQuestions.DataSources), such as for *Crithidia fasciculata,* and two other African trypanosomes, *T. congolense* and *T. vivax*. All these genomes contain active and/or traces of inactive TEs (for reviews *see*: [15, 16]).

All TEs described to date in trypanosomatids belong to four groups: the VIPER LTR retrotransposons, the site-specific SLACS/CZAR retroposons (also named non-LTR retrotransposons), the ingi/L1Tc retroposons, and TATE, for which the TE class is unknown (for a recent review *see* [15, 16]). Retroposons of the ingi clade, which will be further considered herein, include two categories of active elements (*see* Fig. 1): (1) the long (4,736–5,419 bp) and autonomous elements originally characterized in *T. brucei* (Tbingi) [17, 18] and *T. cruzi* (L1Tc) [19], and subsequently described in *T. vivax* (Tvingi) [20] and *T. congolense* (Tcoingi and L1Tco) [20], and (2) the short (260–1,030 bp) and non-autonomous elements identified in *T. brucei* (TbRIME) [21], *T. cruzi* (NARTc) [22], and *T. vivax* (TvRIME) [20]. The short elements (TbRIME, TvRIME, and NARTc) are truncated versions that are mobilized by the retrotransposition machinery of the corresponding long elements (Tbingi, Tvingi, and L1Tc, respectively) [20, 23, 24]. Consequently, the Tbingi/TbRIME, Tvingi/TvRIME, and L1Tc/NARTc associations are considered as pairs of retroposons akin to the human LINE1/Alu pairs [25, 26].

Trypanosome and *Leishmania* genomes also contain highly degenerate elements related to retroposons of the ingi clade, named DIREs for "degenerate ingi-related elements" [27, 28]. Tbingi/TbRIME, L1Tc/NARTc, and DIREs share the first 76–79 residues, which constitute the hallmark of trypanosomatid retroposons ("76–79 bp signature"). Recently, small degenerate retroposons (~0.55 kb) named LmSIDERs (for "short interspersed degenerate retroposons") containing this motif have been identi-

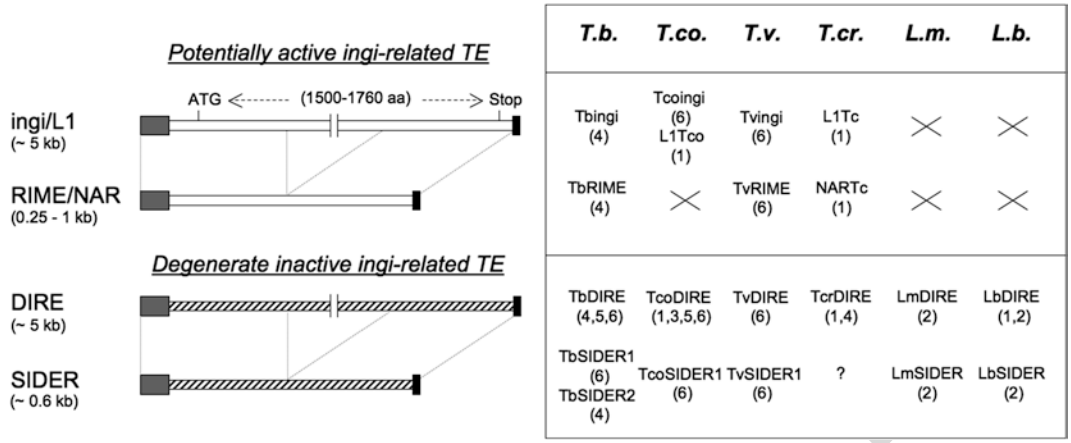

Fig. 1 Schematic representation of ingi-related retroposons identified in the trypanosomatid genomes. The *left panel* represents potentially active (ingi/L1 and RIME/NAR) and degenerate (DIRE and SIDER) retroposons. The conserved "76–79 bp signatures" and the poly(dA) tails are showed by *grey* and *black boxes*, respectively. The approximate size of these mobile elements is indicated in *brackets*. The long and autonomous ingi/L1 elements are the only retroposons coding for a protein (from 1,500 to 1,760 aa long) responsible for retrotransposition of themselves or other ingi-related retroposons. The *right panel* shows the name of each ingi-related family in the genome of *T. brucei* (*T.b.*), *T. congolense* (*T.co.*), *T. vivax* (*T.vi.*), *L. major* (*L.m.*), and *L. braziliensis* (*L.b.*), as well as the ingi subclade(s) they belong to (from 1 to 6 in *brackets*). A *cross* (x) and an *interrogation mark* (?) mean that no such TE have been identified or their presence has not been investigated, respectively

fied in the genomes of *L. major* [5], *L. infantum*, and *L. braziliensis* [5, 29]. LmSIDER constitutes the largest retroposon family described so far in trypanosomatids; members are located in the 3′-UTR of genes, where they play a role in the regulation of gene expression [5, 30, 31].

2 Materials

Computational TE analyses can be performed on a local desktop machine with Internet access. However, large-scale studies require a local software installation, typically in a UNIX environment (*see* **Note 1**), a working knowledge of the UNIX language, and the ability to install applications in a Linux environment. Also essential is a basic knowledge of genome annotation and sequence viewers (while there are many out there, the authors of this chapter are partial to Artemis). Basic knowledge of PERL is necessary for some of the pipelines proposed. For the identification and analysis of mobile elements, the following software materials will be necessary.

1. *Rapid Annotation Transfer Tool (RATT)*: freely available from its own sourceforge site (http://ratt.sourceforge.net/), or as part of the Pagit package (http://www.sanger.ac.uk/resources/software/pagit/).

2. *Exonerate*: for alignment of full-length ingi peptides against reference genomes (www.ebi.ac.uk/~guy/exonerate/).
3. *Tabix*: used for indexing and retrieving rows from a tabular file format, such as a General Feature Format File (GFF); it is freely available as part of the Samtools package (http://sourceforge.net/projects/samtools/files/).
4. *Water*: used for performing Smith-Waterman local alignments; freely available as part of the EMBOSS package (http://emboss.sourceforge.net/download/).
5. *Artemis*: or a similar sequence viewer program able to handle GFF format files (http://www.sanger.ac.uk/resources/software/artemis/).
6. *Blastall*: or a similar multifunctional Blast packaged installed locally; freely available from NCBI (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/).
7. A web browser with access to major online genome databases (*see* http://tritrypdb.org/tritrypdb/).
8. Other recommended software for DNA and protein analysis (ClustalX, DNA strider, etc.).

3 Methods

3.1 Identification of Ingi-Related Retroposons

Identification of ingi-related sequences in all the investigated trypanosomatid genomes was primarily based on Blast searches using tBlastN and BlastN approaches, followed by extensive manual annotation and curation. Here we propose a workflow that will help jumpstart curation of the elements by replacing a significant portion of what was originally done as manual work.

Ingi sequences are identifiable by BlastN searches within species (e.g., Tbingi), but are not conserved at the nucleotide sequence level across species (e.g., Tbingi versus L1Tc) and are poorly conserved at the amino acid level (Tbingi and L1Tc peptides are only 23.8 % identical). Further complicating prediction of ingi elements across trypanosomatids is the phylogenetic diversity of ingi elements, with identification of six different subclades [20], for example ingi6 (Tcoingi) and ingi1 (L1Tco) in *T. congolense*. For this reason, BlastN searches alone will not suffice in making predictions in distantly related trypanosomatid genomes. We propose a pipeline relying on peptide to DNA alignments using representative members of each ingi subclade as a query. It is important to mention at this stage that the proposed electronic annotation pipeline is useful to identify and perform a preannotation of transposable elements in a given genome; manual detailed curation, annotation, and analyses will however be required to perform in-depth analyses.

At the core of our ingi prediction pipeline is the European Bioinformatic Institute's (EBI) sequence alignment program Exonerate [32]. We use Exonerate to query representative full-length peptides from the Tbingi, Tcoingi, Tvingi, L1Tc, and L1Tco families (accession numbers: JQ917146, JQ917147, JQ917148, JQ917144, and JQ917145, respectively—*see* **Note 2**) against a reference genome. A custom Perl scripts are then employed to select the highest scoring alignment from regions where multiple ingi family members have been aligned (Fig. 2), and to generate CDS features for each predicted element. We have designed a wrapper for exonerate called Ingihelper.pl (https://ghedinlab.csb.pitt.edu/GhedinLab/Applications), which accepts as input a fasta file of peptide sequences (preferably one for each family of ingi) as query, and a whole-genome sequence as subject. This script calls exonerate to perform a protein2DNA alignment, returning a GFF file of target alignments, and then searches across the GFF file for overlapping alignments, selecting the highest scoring alignment when this occurs. This frequently occurs due to the similarity of all ingi family members at the peptide level. These results are outputted in a modified GFF format where the gene model and corresponding coding sequence features can be viewed for each ingi prediction, allowing a quick assessment of the degeneracy in each element (Fig. 2). Our goal with this method is not to provide a fully automated prediction pipeline for ingi-related elements in trypanosomatids, but to rapidly provide a mostly (>90 %) complete set of ingi predictions in a format that can then be easily viewed and combined with other predictions (e.g., Blast) by annotators familiar with the structure of these elements. In most cases, a human eye will be required to determine the absolute 5′ and 3′ ends of these elements and assess whether neighboring ingi alignments belong to the same element. A major advantage of Exonerate is the multiple output options available. Ingi predictions can be outputted not just in GFF format for viewing and further manual curation, but can also be outputted in Fasta format for multiple alignments, allowing for identification of ingi subclade using phylogenetic methods. We recommend using Artemis to view Exonerate and Blast outputs, and to annotate ingi-related retroposons. Although there are other freely available sequence viewers, the authors of this chapter favor the use of Artemis for its versatility in both sequence annotation and analysis, its ability to read and output sequence data in a variety of formats (EMBL, GenBank, and GFF), and its legacy as a tool in trypanosomatid genome annotation.

We have tested this method against manual predictions that have been performed on the *T. vivax* and *T. congolense* genomes. The *T. congolense* genome is a challenge for ingi prediction as it contains ingi elements of the ingi1 (L1Tco) and ingi6 (Tcoingi) subclades, which are distantly related at both extremities of the ingi

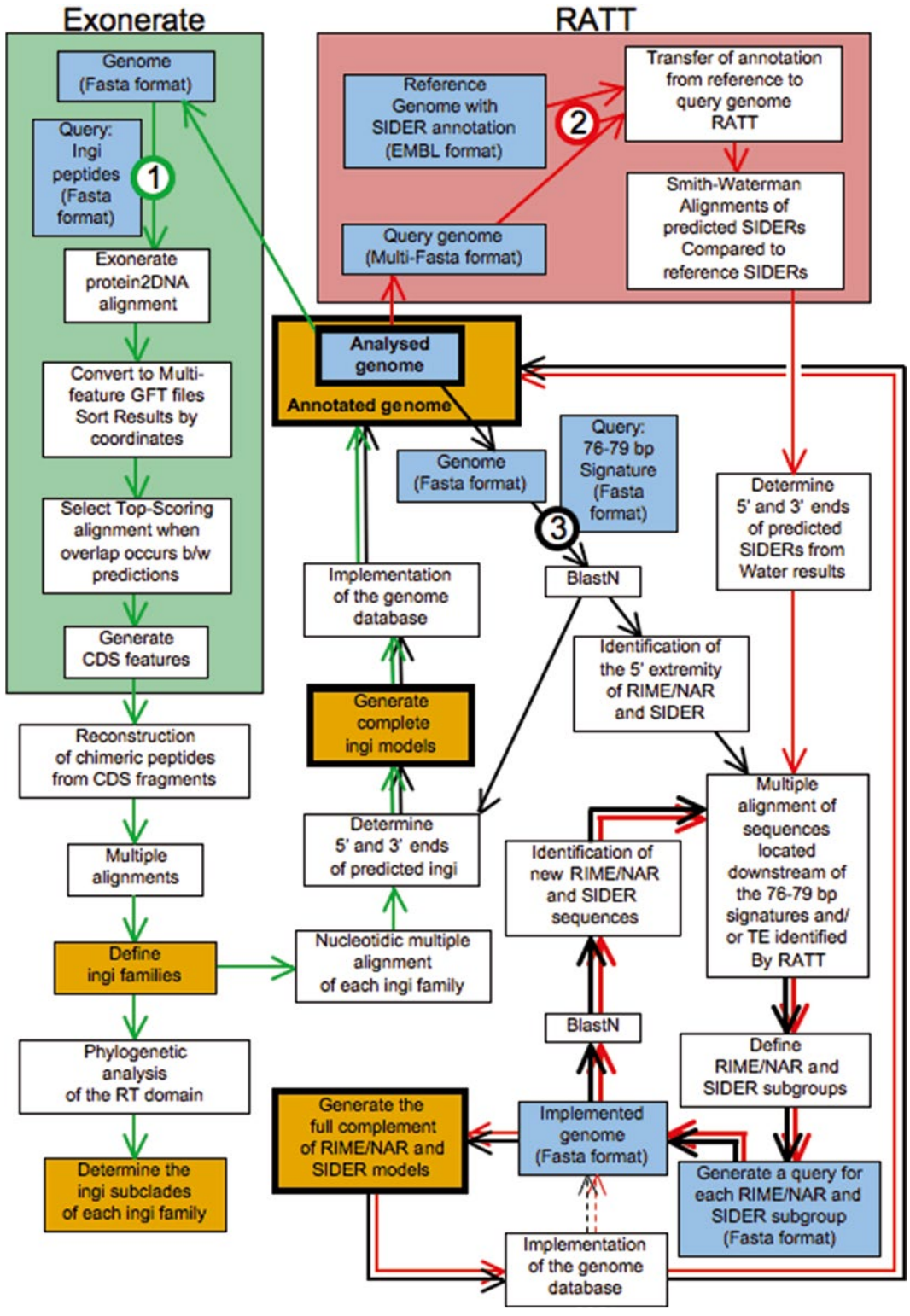

Fig. 2 Flow diagram of the annotation of ingi-related retroposons in trypanosomatid genomes. The Exonerate pipeline and the Rapid Annotation Transfer Tool (RATT) approach are shown on *green* and *red* backgrounds, respectively. *Blue* and *orange boxes* show input and output files, respectively. Identification and annotation of ingi/L1Tc-coding sequences are performed first (①), followed by identification and annotation of short non-autonomous ingi-related sequences (RIME/NAR and SIDER) by RATT (②) and BlastN approaches (③)

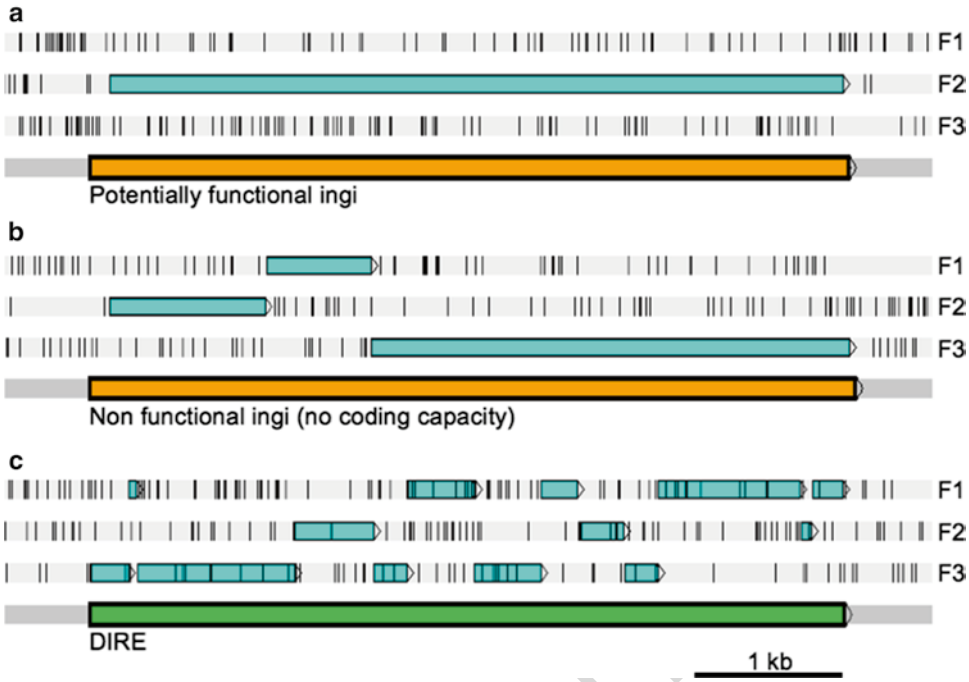

Fig. 3 Annotation of potentially functional ingi (**a**), nonfunctional ingi (**b**), and DIRE (**c**) using Exonerate. Partial (**b** and **c**) or complete (**a**) gene models generated by Exonerate are shown in *blue* as they appear in Artemis. *Vertical bars* represent stop codons found in each of the three frames (F1–F3). Potentially functional ingi contain a single long ORF (**a**). Only few frameshifts and/or stop codons are present in the coding sequence of nonfunctional ingi (**b**), while numerous frameshifts or stop codons characterize DIRE sequences (**c**). Manual curation is required to generate a complete gene model for nonfunctional ingi and DIRE, as well as to determine the 5′- and 3′-extremities of all these ingi-related TE

phylogenetic tree [20]. The *T. vivax* genome is a challenge due to the sheer number of ingi elements present, and these are frequently found on contiguated unordered contigs where TE elements may be truncated at contig boundaries. We have used earlier versions of both of these genomes in order to compare previous manual annotations [20] to Exonerate's automated predictions.

The resulting Exonerate predictions are summarized in Fig. 3. Using the full-length peptide sequences for Tbingi, Tcoingi, Tvingi, L1Tc, and L1Tco (accession numbers: JQ917146, JQ917147, JQ917148, JQ917144, and JQ917145, respectively—*see* **Note 2**) as queries, Exonerate predicts 98.7 % of Tcoingi elements and 100 % of L1Tco elements (Table 1). Of the predicted Tcoingi elements, 13 % are split into 2 or more alignments for Tcoingi, and 8.3 % for L1Tco. In most cases, these occur where the coding sequence is interrupted by a large gap. A smaller proportion (85.3 %) of TcoDIRE elements are predicted, and a larger proportion of them (19.3 %) are split into multiple alignments because of their high degree of degeneracy. The Exonerate method

[AU1]

t1.1 **Table 1**

t1.2

t1.3 t1.4 Genome	TE	Manual annotation (ref: [20])	Prediction overlap	% Correct	% Split
t1.5 *T. congolense*	Tcoingi	78	77	98.7	12.0
t1.6 *T. congolense*	L1Tco	12	12	100.0	8.3
t1.7 *T. congolense*	TcoDIRE	170	145	85.3	19.3
t1.8 *T. vivax*	Tvingi	741	701	94.6	3.7
t1.9 *T. vivax*	TvDIRE	107	97	90.6	11.3
t1.10 *T. vivax*	TvRIME	58	51	87.9	76.5

performs similarly well against the *T. vivax* genome (Table 1). Here, 94.6 % of Tvingi elements are predicted, and only 3.7 % of these are split into multiple alignments. TvDIREs and TvRIMEs are also predicted in the *T. vivax* genome (90.6 % and 87.9 %, respectively); however, none of the TvSIDER1 elements are detected. Analysis of these two trypanosome genomes confirms that this approach can detect short truncated ingi-related elements, such as TvRIME (1,030 bp), as long as they contain fragment(s) of ingi-coding sequences [20]. Identifying degenerate short ingi-related families that do not have significant matches to ingi peptides, such as TcoSIDER1 and TvSIDER1 [33], is not possible with this approach. In addition to predictions that overlap with manual annotations, Exonerate makes 43 more predictions for *T. congolense* and 27 for *T. vivax*. Most of these are, however, short alignments (median length of 384/298 bp) with low scores.

Although very powerful, Exonerate cannot predict with accuracy start and end coordinates using ingi peptide as queries, since all ingi families identified so far contain 5′-untranslated sequences ranging from 9 to 193 bp (Tbingi and L1Tco) and 3′-untranslated sequences ranging from 15 to 271 bp (L1Tc and Tbingi). Because of their lack of site specificity for insertion, ingi are flanked by non-conserved sequences. Consequently, a simple multiple alignment of all nucleotide sequences flanking the single long ingi ORF of the same family will allow to determine both 5′- and 3′-extremities. To validate the exact ingi boundaries, it is essential to remember that all ingi families start with the conserved 76–79 bp ingi signature and end with the poly(A) stretch (retroposon hallmark). Blast searches with full-length nucleotide sequences of newly identified ingi families will be necessary to complete gene model determination and curation. Taken together, Exonerate performs well in detecting ingi elements and produces a rapid survey of ingi elements in a new trypanosomatid genome. However, an exhaustive annotation of all ingi elements will require combining Exonerate

gene models with other search results (e.g., tBlastN, BlastN). Furthermore, a degree of manual annotation is necessary to correct models that are split into multiple alignments and to determine the exact ingi boundaries. Despite its shortcomings, the use of this method in generating models of ingi elements should greatly increase the speed at which these elements are called in trypanosomatid genomes.

3.2 Phylogenetic Analyses of Ingi Families and DIREs

The method described above also attempts to assign the family/subclade of ingi from the query that produces the highest scoring alignment. To conclusively assign an ingi element to a family or a subclade, a phylogenetic analysis should also be performed. Among the three ingi-coding domains, i.e., apurinic/apyrimidinic endonuclease, RNAseH, and reverse transcriptase (RT), the latter is the most conserved and ideal for phylogenetic tree reconstruction [20, 28]. It corresponds to the amino acid positions 492–770 of Tbingi and 527–822 of L1Tc. These phylogenetic analyses can include potentially active ingi families, including Tbingi, Tvingi, Tcoingi, L1Tc, and L1Tco, as well as DIRE, the degenerate ingi-related sequences. Since members of potentially active ingi families are highly conserved at the nucleotide level, it is recommended to generate for each family a consensus sequence from multiple alignments of full-length elements or of only the RT domain. Online tools available from EBI (http://srs.ebi.ac.uk/) can be used to generate the consensus (CosN). In contrast, DIREs are typically unique ingi degenerate sequences containing a number of frameshifts and stop codons in their nonfunctional coding sequences. Since phylogenetic analyses require amino acid sequences, frameshifts were removed manually from the DNA sequences using the Exonerate output to tentatively reconstitute proteins from the analyzed DIREs. This approach generated a pseudogene for each DIRE element, encoding a single ingi-like sequence, which in most cases contained numerous stop codons. Alternatively, the Blast-Extend-Repraze (BER) algorithm developed at the J. Craig Venter Institute (formerly the Institute for Genomic Research) can be used to localize frameshifts (*see* [28]). So far, RT-based phylogenetic analyses identified six ingi subclades, three of them containing potentially active ingi families [20].

3.3 Identification and Analysis of Short Ingi-Related Sequences (RIME/NAR and SIDER)

3.3.1 Identification by BlastN Searches

As mentioned above, certain short truncated ingi-related families cannot be detected by tBlastN using ingi peptides as query, implying that a BlastN-based approach is more appropriate. The "76–79 bp signature" shared by all retroposons of the ingi clade is the only query sequence that can be used to detect such elements in new trypanosomatid genomes. Several rounds of multiple alignments and BlastN searches with conserved sequences downstream of identified "76–79 bp signatures" are in fact required to annotate the full complement of the analyzed ingi-related family. For

example, BlastN with the "76–79 bp signature" detected 108 significant matches in the *L. major* genome. After 8 consecutive comparisons/BlastN cycles, 1,858 related, but highly divergent, LmSIDER sequences were identified in this genome [5]. In contrast, only a single comparison/BlastN cycle was necessary to identify and annotate all of the 70 TcoSIDER1 sequences contained in the *T. congolense* genome after identification of their "76–79 bp signature" [33].

This approach leads to the generation of a list of SIDERs that are then aligned. Because of the degenerate nature of these elements, a manual alignment is first required. A profile alignment of new SIDER elements can then be done using the initial manual alignment, leading to the identification of more SIDERs. In [29], HMM profiles of SIDER sequences were generated by optimizing parameters and maximum likelihood estimators allowing a high-quality alignment of LmSIDER1 comparable to that published for LmSIDER2. But this approach is difficult and requires genome-specific parameter optimization. For example, we used HMMER 2.32 with default parameters to identify SIDER elements in the *L. mexicana* genome using a profile generated from a manual alignment of *L. major* SIDER1 elements. Only 562 SIDERs were predicted in the *L. mexicana* genome, which we assume is on the low side considering that the *L. major* genome has 1,858 SIDERS identified.

Once the complete, or near-complete, set of SIDERs is identified, 5′- and 3′-extremities can be determined by multiple alignments of all the members of a given family, as described above for ingi/L1Tc elements.

3.3.2 Identification by Transfer of Annotation

A second proposed method for predicting SIDERs in *Leishmania* genomes relies on the transfer of annotation between genomes (Fig. 2). *Leishmania* genomes are highly syntenic with few to no breaks in gene order among the old world *Leishmania* species. Even the divergent *Leishmania Viannia* clade (represented solely by the *L. braziliensis* genome) shares obvious blocks of synteny compared to members of the *Leishmania leishmania* clade. Recent annotation of both the *L. donovani* and *L. mexicana* genomes has relied heavily on automatic predictions of gene models using synteny (in the form of Nucmer matches between genomes) and the Rapid Annotation Transfer Tool (RATT) [34]. The majority (98.3 %) of predicted *L. infantum* genes could be transferred to *L. donovani* [13] in this manner, and slightly fewer (93 %) gene models were transferred from *L. major* to *L. mexicana* [12].

As RATT will indiscriminately transfer any sequence feature, we have attempted to transfer manual predictions of SIDER elements in *L. major* [5] to *L. mexicana* for which no SIDER predictions have been made to date. This method works moderately well with 1,223 of the 1,885 manual predictions (66 %) being transferred to *L. mexicana* from *L. major*. We generated

Smith-Waterman alignments for each of these and, compared to its *L. major* orthologue, revealed that 800 of these share the same 5′ end. The benefit of this method is that it is easy to use and relatively rapid, although moderately sensitive. Drawbacks of this method are that it will only predict SIDER elements known in another species, and requires somewhat closely related *Leishmania* genomes (ideally > 95 % identity). RATT also assumes orthology and will not transfer the same model more than once, thus failing in cases where segmental duplications have resulted in multiple SIDERs of recent descent. Absence of synteny between SIDERs is probably the main reason why approximately one-third of *L. mexicana* SIDERs failed to be identified by this method. This is consistent with the identification of slightly more than 100 SIDERs in the *L. infantum* and *L. braziliensis* genomes, when more than 1,900 could be found in each genome by HMM [29].

3.3.3 Evolutionary Analyses of Ingi-Related TE Families

Once a consistent alignment of a given TE family is obtained, statistical analyses can be performed to gain insight into its evolutionary dynamics. This analysis is based on the determination of the degree of divergence within members of an ingi family. To study the extent of this divergence, the percentage of divergence between the consensus sequence deduced from the alignment and each TE copy aligned is determined. Since the consensus sequence is assumed to approximate the element's original sequence at the time of insertion, the percentage of substitutions from the consensus sequence is correlated to the age of a given element (the age corresponds to the time of retrotransposition). In other words, the younger the family, the more conserved the sequences.

Online tools available from EBI are very useful to generate the consensus sequence (CosN) of a given alignment and to calculate the percentage of divergence from the consensus sequences (InfoalignN). The values obtained can be expressed as the number of elements as a function of their divergence from their consensus sequence, to calculate the median divergence value. The higher the value, the older and degenerate the family analyzed, such as TbSIDER2, TbSIDER1, TcoSIDER, and LmSIDER2 (11, 16, 16, and >20, respectively), while low median values reflect youth and possible functionality of a TE family, as observed for TvRIME, NARTc, and TbRIME (1, 2, and 4, respectively).

3.4 Conclusion

The whole strategy developed to identify and annotate the full complement of ingi-related sequences in trypanosomatid genomes is presented in Fig. 2. This approach developed to identify protein-coding and noncoding ingi-related retroposons can be adapted to transposable element family from other organisms, with as ultimate goal to identify highly degenerate family members potentially recycled by the host genome to fulfill housekeeping cellular functions.

4 Notes

1. While UNIX is typically stated as a requirement, many of the tools commonly used also work under the UNIX-based Macintosh OS X operating system, and also under Microsoft Windows with environments like Cygwin or MSYS.
2. These entries correspond to consensus nucleotide sequences of 63 Tbingi (JQ917146), 27 Tcoingi (JQ917147), and 48 L1Tc (JQ917144) full-length elements, 46 Tvingi copies larger than 3.5 kb (JQ917148) or 8 L1Tco copies larger than 300 bp (JQ917145) [19]. Each of the five consensus sequences corresponds to a potentially functional retroposon encoding a full-length protein.

Acknowledgement

FB is supported by the Centre National de la Recherche Scientifique (CNRS), the Université Bordeaux Segalen, and the Laboratoire d'Excellence (LabEx) ParaFrap ANR-11-LABX-0024.

Ingihelper.pl is available at https://ghedinlab.csb.pitt.edu/GhedinLab/Applications.

References

1. International-Human-Genome-Sequencing-Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409: 860–921
2. Biemont C, Vieira C (2006) Genetics: junk DNA as an evolutionary force. Nature 443: 521–524
3. Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, Salama SR, Rubin EM, Kent WJ, Haussler D (2006) A distal enhancer and an ultraconserved exon are derived from a novel retroposon. Nature 441:87–90
4. Santangelo AM, de Souza FS, Franchini LF, Bumaschny VF, Low MJ, Rubinstein M (2007) Ancient exaptation of a CORE-SINE retroposon into a highly conserved mammalian neuronal enhancer of the proopiomelanocortin gene. PLoS Genet 3:1813–1826
5. Bringaud F, Muller M, Cerqueira GC, Smith M, Rochette A, El-Sayed NM, Papadopoulou B, Ghedin E (2007) Members of a large retroposon family are determinants of post-transcriptional gene expression in *leishmania*. PLoS Pathog 3:e136
6. Stevens JR, Noyes HA, Schofield CJ, Gibson W (2001) The molecular evolution of Trypanosomatidae. Adv Parasitol 48:1–56
7. Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, Bartholomeu DC, Lennard NJ, Caler E, Hamlin NE, Haas B et al (2005) The genome of the African trypanosome *Trypanosoma brucei*. Science 309: 416–422
8. Jackson AP, Sanders M, Berry A, McQuillan J, Aslett MA, Quail MA, Chukualim B, Capewell P, MacLeod A, Melville SE et al (2010) The genome sequence of *Trypanosoma brucei gambiense*, causative agent of chronic human African trypanosomiasis. PLoS Negl Trop Dis 4:e658
9. El-Sayed NM, Myler PJ, Bartholomeu DC, Nilsson D, Aggarwal G, Tran AN, Ghedin E, Worthey EA, Delcher AL, Blandin G et al (2005) The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. Science 309:409–415
10. Ivens AC, Peacock CS, Worthey EA, Murphy L, Aggarwal G, Berriman M, Sisk E, Rajandream MA, Adlem E, Aert R et al (2005)

The genome of the kinetoplastid parasite, *Leishmania major*. Science 309:436–442

11. Peacock CS, Seeger K, Harris D, Murphy L, Ruiz JC, Quail MA, Peters N, Adlem E, Tivey A, Aslett M et al (2007) Comparative genomic analysis of three *Leishmania* species that cause diverse human disease. Nat Genet 39:839–847
12. Rogers MB, Hilley JD, Dickens NJ, Wilkes J, Bates PA, Depledge DP, Harris D, Her Y, Herzyk P, Imamura H et al (2011) Chromosome and gene copy number variation allow major structural change between species and strains of *Leishmania*. Genome Res 21:2129–2142
13. Downing T, Imamura H, Decuypere S, Clark TG, Coombs GH, Cotton JA, Hilley JD, de Doncker S, Maes I, Mottram JC et al (2011) Whole genome sequencing of multiple *Leishmania donovani* clinical isolates provides insights into population structure and mechanisms of drug resistance. Genome Res 21:2143–2156
14. Raymond F, Boisvert S, Roy G, Ritt JF, Legare D, Isnard A, Stanke M, Olivier M, Tremblay MJ, Papadopoulou B et al (2012) Genome sequencing of the lizard parasite *Leishmania tarentolae* reveals loss of genes associated to the intracellular stage of human pathogenic species. Nucleic Acids Res 40:1131–1147
15. Bringaud F, Ghedin E, El-Sayed NM, Papadopoulou B (2008) Role of transposable elements in trypanosomatids. Microbes Infect 10:575–581
16. Thomas MC, Macias F, Alonso C, Lopez MC (2010) The biology and evolution of transposable elements in parasites. Trends Parasitol 26:350–362
17. Kimmel BE, Ole-MoiYoi OK, Young JR (1987) *Ingi*, a 5.2-kb dispersed sequence element from *Trypanosoma brucei* that carries half of a smaller mobile element at either end and has homology with mammalian LINEs. Mol Cell Biol 7:1465–1475
18. Murphy NB, Pays A, Tebabi P, Coquelet H, Guyaux M, Steinert M, Pays E (1987) *Trypanosoma brucei* repeated element with unusual structural and transcriptional properties. J Mol Biol 195:855–871
19. Martin F, Maranon C, Olivares M, Alonso C, Lopez MC (1995) Characterization of a non-long terminal repeat retrotransposon cDNA (L1Tc) from *Trypanosoma cruzi*: homology of the first ORF with the ape family of DNA repair enzymes. J Mol Biol 247:49–59
20. Bringaud F, Berriman M, Hertz-Fowler C (2009) Trypanosomatid genomes contain several families of ingi-related retroposons. Eukaryotic Cell 8:1532–1542
21. Hasan G, Turner MJ, Cordingley JS (1984) Complete nucleotide sequence of an unusual mobile element from *Trypanosoma brucei*. Cell 37:333–341
22. Bringaud F, Garcia-Perez JL, Heras SR, Ghedin E, El-Sayed NM, Andersson B, Baltz T, Lopez MC (2002) Identification of non-autonomous non-LTR retrotransposons in the genome of *Trypanosoma cruzi*. Mol Biochem Parasitol 124:73–78
23. Bringaud F, Biteau N, Zuiderwijk E, Berriman M, El-Sayed NM, Ghedin E, Melville SE, Hall N, Baltz T (2004) The *ingi* and RIME non-LTR retrotransposons are not randomly distributed in the genome of *Trypanosoma brucei*. Mol Biol Evol 21:520–528
24. Bringaud F, Bartholomeu DC, Blandin G, Delcher A, Baltz T, El-Sayed NM, Ghedin E (2006) The *Trypanosoma cruzi* L1Tc and NARTc non-LTR retrotransposons show relative site-specificity for insertion. Mol Biol Evol 23:411–420
25. Jurka J (1997) Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. Proc Natl Acad Sci U S A 94:1872–1877
26. Dewannieux M, Esnault C, Heidmann T (2003) LINE-mediated retrotransposition of marked *Alu* sequences. Nat Genet 35:41–48
27. Ghedin E, Bringaud F, Peterson J, Myler P, Berriman M, Ivens A, Andersson B, Bontempi E, Eisen J, Angiuoli S et al (2004) Gene synteny and evolution of genome architecture in trypanosomatids. Mol Biochem Parasitol 134:183–191
28. Bringaud F, Ghedin E, Blandin G, Bartholomeu DC, Caler E, Levin MJ, Baltz T, El-Sayed NM (2006) Evolution of non-LTR retrotransposons in the trypanosomatid genomes: *Leishmania major* has lost the active elements. Mol Biochem Parasitol 145:158–170
29. Smith M, Bringaud F, Papadopoulou B (2009) Organization and evolution of two SIDER retroposon subfamilies and their impact on the *Leishmania* genome. BMC Genomics 10:240
30. Boucher N, Wu Y, Dumas C, Dube M, Sereno D, Breton M, Papadopoulou B (2002) A common mechanism of stage-regulated gene expression in *Leishmania* mediated by a conserved 3′-untranslated region element. J Biol Chem 277:19511–19520
31. McNicoll F, Muller M, Cloutier S, Boilard N, Rochette A, Dube M, Papadopoulou B (2005)

Distinct 3′-untranslated region elements regulate stage-specific mRNA accumulation and translation in *Leishmania*. J Biol Chem 280:35238–35246

32. Slater GS, Birney E (2005) Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6:31

33. Bringaud F, Berriman M, Hertz-Fowler C (2011) TSIDER1, a short and non-autonomous Salivarian trypanosome-specific retroposon related to the ingi6 subclade. Mol Biochem Parasitol 179:30–36

34. Otto TD, Dillon GP, Degrave WS, Berriman M (2011) RATT: Rapid Annotation Transfer Tool. Nucleic Acids Res 39:e57

Author Query

Chapter No.: 6 0002174834

Query	Details Required	Author's Response
AU1	Please provide caption for Table 1.	