BioEssays
www.bioessays-journal.com

Analyzing Horizontal Transfer of Transposable Elements on a Large Scale: Challenges and Prospects

Jean Peccoud, Richard Cordaux, and Clément Gilbert*

Whoever compares the genomes of distantly related species might find aberrantly high sequence similarity at certain loci. Such anomaly can only be explained by genetic material being transferred through other means than reproduction, that is, a horizontal transfer (HT). Between multicellular organisms, the transferred material will likely turn out to be a transposable element (TE). Because TEs can move between loci and invade chromosomes by replicating themselves, HT of TEs (HTT) profoundly impacts genome evolution. Yet, very few studies have quantified HTT at large taxonomic scales. Indeed, this task currently faces difficulties that range from the variable quality of available genome sequences to limitations of analytical procedures, some of which have been overlooked. Here we review the many challenges that an extensive analysis of HTT must overcome, we expose biases and limits of current methods, suggest solutions or workarounds, and reflect upon approaches that could be developed to better quantify this phenomenon.

1. Introduction

Horizontal transfer (HT) of genetic material is the transmission of DNA between organisms that are not necessarily closely related, through mechanisms other than reproduction.[1] While the frequency, impact, and mechanisms underlying these transfers are well understood in prokaryotes, HT in eukaryotes is less studied and remains relatively obscure.[2] One type of HT among eukaryotes is widespread though, that of transposable elements (TEs). Transposable elements are mobile DNA segments found in the genomes of virtually all organisms.[3] They can transpose from one genomic locus to another, duplicate themselves and generate large numbers of copies. One of the first cases of HT of TEs (HTT) documented in eukaryotes was that of the *P* element.[4] The *P* element was found to be patchily distributed within a subgenus of *Drosophila*, and to be almost identical in *D. melanogaster* and *D. willistoni* despite the >26 million years of divergence separating the two species. Realizing that TEs can cross species barriers has deeply transformed our understanding of both TE evolutionary dynamics and host genome evolution. HT not only explains the persistence of certain TEs over evolutionary times in the face of host defense factors, as TEs jump from immunized to naive genomes, it may also have played a substantial role in the composition and evolution of eukaryote genomes.[5–7]

Hundreds of other HTT cases have so far been uncovered in eukaryotes, mainly in plants, animals, and fungi.[8] Overall, these cases suggest that HTT preferentially involves DNA transposons (class 2 TEs) over retrotransposons (class 1 TEs)[6,9,10] and may be facilitated by host-parasite relationships.[11–14] However, because almost all HTT studies have so far focused on one or few TE types and/or on host groups that represent limited phylogenetic breadth, the trends, and factors influencing HTT are still largely unknown or unsupported statistically. The increasing availability of genome sequences from various eukaryotic organisms makes it possible to start filling this gap.

With this aim in mind, we recently conducted a systematic analysis of HTT among all 195 insect species whose genome sequences were publicly available.[15] Our analysis showed that at least 2248 HTT events occurred among these insects in the last 10 million years. Each event corresponds to the movement of at least one TE copy from one lineage to the germline of another lineage, without reproduction, followed by amplification of the TE by transposition in the receiving genome. These thousands of HTT events were shown to have generated 2% of the insect genomic content on average, demonstrating the crucial impact of HTT on genome evolution in insects, and allowed to statistically support a role of phylogenetic and geographic proximity in facilitating HTT. Therefore, available genomic resources now make it possible to analyze HTT as an evolutionary process, rather than as isolated cases, and to assess how this process is shaped by interspecific interactions.[16] The robust identification and quantification of HTT required before any high-level statistical analysis can be performed is however far from trivial.

Dr. J. Peccoud, Dr. R. Cordaux
UMR CNRS 7267
Ecologie et Biologie des Interactions
Equipe Ecologie Evolution Symbiose
Université de Poitiers
86000 Poitiers, France
E-mail: jean.peccoud@univ-poitiers.fr

Dr. Clément Gilbert
UMR CNRS 9191
UMR 247 IRD
Laboratoire Evolution, Génomes, Comportement, Écologie
Université Paris-Sud
91198 Gif-sur-Yvette, France

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/bies.201700177.

DOI: 10.1002/bies.201700177

Until recently, most HTT cases have been uncovered fortuitously while characterizing TEs in genomes. Because a TE replicates in a genome into many copies that follow separate paths of degradation, identifying its exact ends and structural components generally requires inspecting several copies and reconstructing the most complete and the least possible degraded version of its sequence, called a consensus. HTT is typically suspected during the annotation/classification of these consensus sequences, if similarity searches reveal an unexpectedly high level of DNA sequence identity with a TE previously characterized in a distantly related species. To formally test whether the sharing of similar TEs is due to HT rather than vertical inheritance from a common ancestor, three criteria are generally applied (**Figure 1**): 1) between-species nucleotide divergence of the TE must be lower than that of orthologous, vertically inherited, genes; 2) taxonomic distribution of the TE may be patchy; and 3) TE phylogeny may be incongruent with that of the species in which it was identified.[6,7,17]

While this approach can effectively reveal HTT events, steps that rely on manual annotation and visual inspection of sequence alignments or trees do not scale up well. For instance, our study on 195 insect genomes[15] yielded 53 452 consensus sequences (representing about 9.2 million complete or partial TE copies longer than 100 bp) that cannot possibly be curated manually. The sheer number of genome sequences and TE copies to process, their rapid evolutionary diversification and degradation, together with the impossibility to apply wet-lab verifications greatly constrain large-scale comprehensive and statistical analyses of HTT. In the following, we expose the difficulties that arise with such analyses and the shortcomings of certain procedures currently in use. To tackle these challenging issues, we propose solutions and workarounds, or call for new developments.

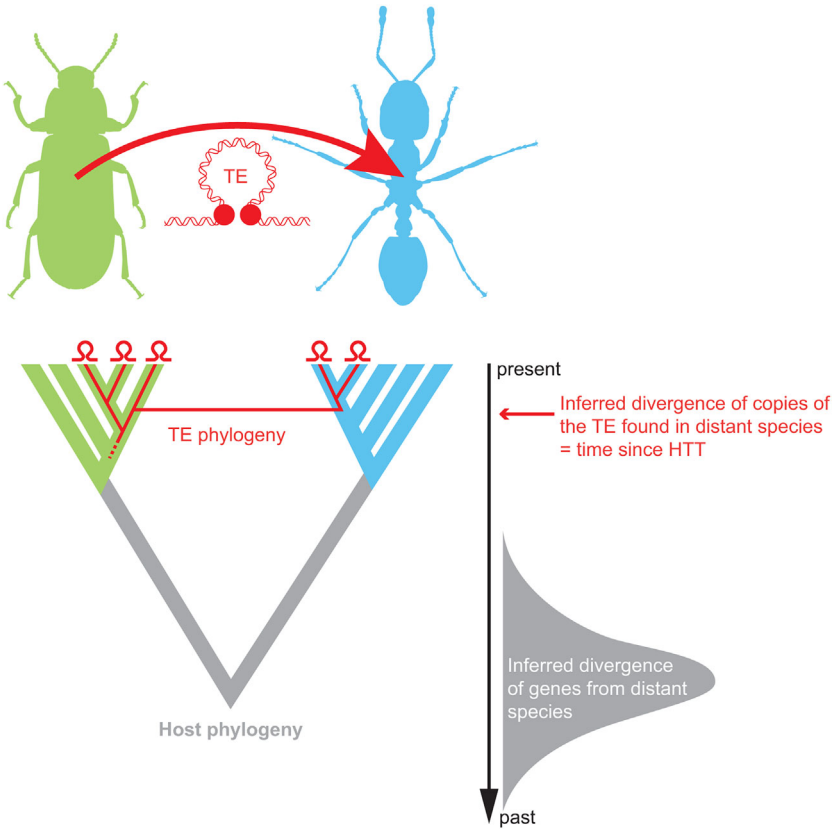

Figure 1. Phylogenetic patterns resulting from a horizontal transfer (curved red arrow) of transposable element (TE) between two distantly related lineages. This hypothetical example shows a DNA transposon linked to transposases (shown as red circles), a configuration taken during transposition. The phylogeny of the transferred TE is expected to differ from that of the host lineages (criterion 2, see text) and the TE may only be found in some of the species among the clade where the TE is present (criterion 3). DNA sequence divergence between TE copies from the two lineages should be much lower than that measured between genes inherited from the species' common ancestor (criterion 1).

2. How to Reliably Detect TEs Across Many Genomes?

2.1. Issues Related to TE Annotation and Genome Sequence Quality

The comprehensive and accurate annotation of TEs in a species currently requires manual (non-automatic) verification steps and is refined through many dedicated studies.[18–20] This task has been undertaken on several model species, but it cannot be applied to many genomes in any single analysis, considering the limited time and human resources. Automated pipelines of de novo TE detection have to be applied and may be combined for better results.[21–23] Still, the limited sensitivity and accuracy of these tools may affect large-scale HTT studies. Sensitivity is limited, in particular, by the fact that algorithms initially select candidate TEs among repeated sequences. Recently transferred TEs may thus be preferentially overlooked as they have had less time to replicate in a recipient genome. This issue appears to be difficult to solve so long as TEs are identified separately in each genome prior to being compared across species.

Conversely, automated pipelines may erroneously annotate repeated genes as TEs. It is therefore important to filter out purported TEs that present lower similarity to known TE than to non-TE proteins. Doing so in insects led us to remove ~6% of all consensus sequences initially returned by the TE annotation tool.[15] At any rate, the limitations of annotation pipelines require applying the same protocol for all genomes under investigation, and discarding prior TE annotations obtained from different methods of varying efficiency.

Independently of the method used, the ability to detect a TE present in a species may be negatively impacted by the incompleteness of genome assembly (encompassing the proportion of undetermined bases) and fragmentation, which represents the shortness and number of contigs relative to chromosomes. Both metrics greatly vary across published

genome sequences, and may cause detection biases in a study that deals with many species. Genome sequence incompleteness has probably lower impact, considering that most TE groups constitute a fair amount of copies distributed within genomes and can be detected even with low sequence coverage.[24] Fragmentation is more troubling. While most contigs are generally long enough to include full TEs, fragmentation often associates with the inability to resolve repeat sequences, hence, the collapsing of these repetitions in single sequences during the assembly process.[25] This may happen if sequencing reads are shorter than identical repeated sequences. The rate of collapsing is expected to increase with identity between repeats, and might thus be more pronounced for recently active TE families, hence, for more recent transfer events. It is not clear yet how collapsing affects TE detection and, hence, of HTT. An appropriate test of the influence of assembly quality on TE detection and HTT would compare results from searches using different genome assemblies for the same species: "draft" assemblies built only on short reads and more "finished" less fragmented assemblies. If poor genome sequence quality turns out to be problematic, improving genome assemblies with technologies generating long sequencing reads may be required. Long reads have indeed proven very useful in assembling sequences of genome regions populated with TEs or other repeated elements (e.g., Faino et al.[26] Krsticevic et al.[27]). Unfortunately, many published genome sequences of non-model species may remain as "draft" versions of potentially mediocre quality when it comes to the resolution of repeated elements.

2.2. DNA Contamination Between Species

A TE that is annotated in a genome sequence may actually be absent from the source species and instead result from contamination by DNA from other organisms before or during genome sequencing (e.g., Koutsovoulos et al.[28]). Two types of contaminations may cause spurious HTT signals: "direct" contamination between species investigated for HTT, or separate contaminations of at least two studied species by DNA from a pool of closely related organisms not included in the study. The former may happen if several species are sequenced by the same laboratory or subcontractor,[29] while the latter can be caused by symbionts in the broad sense (e.g., bacteria, fungi, and other types of parasites), whose DNA sequence may partly end up in the genome sequence of the organism that carried them, or reciprocally. As wet-lab confirmation (by PCR) of TE presence cannot be undertaken for every candidate HTT if there are hundreds of them, in silico verifications should be applied at least to identify and exclude the most suspicious cases.

Contaminant TEs from distantly related organisms can be easy to spot. For instance, bacterial TEs form well-characterized groups that are naturally very rare in eukaryote genomes.[30] These can be more specifically investigated for removal. Contamination between eukaryotes may not safely be removed solely based on TE classification. Obviously, the occurrence of these issues can be minimized by proper decontamination of genome sequences before their publication, using protocols that can take advantage of the information carried by raw sequencing reads,[31] in particular the sequencing coverage of contigs.

These procedures cannot detect or remove contamination between genome sequences from closely related species. However, this type of contamination is more likely to be of the "direct" type, as hosts and symbionts are typically distantly related. We thus expect 100% sequence similarity at contaminating TEs, if the contaminant species is also analyzed in the HTT study. Candidate HTT events involving 100% identical TEs may be investigated specifically. If the HTT causing 100% sequence identity really occurred, this event had to be so recent that identical sequences should still correspond to functional TEs. It appears less likely that a degraded TE copy could effectively transfer and transpose in the recipient genome.

The inability of contaminant TEs to transpose in a recipient genome may be used to inspect candidate HTT events in another way. A naturally transferred TE is typically represented by a set of similar copies in the donor and recipient genome, due to transposition after and before the transfer. This may not be the case for a very recent transfer, but if a degraded TE copy is found in a genome, denoting some older transfer, other copies should be present. One may thus impose a minimal number of copies in each genome for each suspected HTT event. The maximal number of copies of a TE that contamination can bring is however hard to predict, as it depends on the degree of contamination and TE composition of the source genome.

Finally, sequence identity between orthologous genes that should not be horizontally transferred, if it reaches suspiciously high levels, may indicate the presence of contamination. However, absence of apparent contaminant genes does not necessarily exclude contaminant TEs, as TEs are much more prevalent than genes in genomes.

As it appears, in silico controls can remove false positives but have limited efficiency. More robust methods should be able to identify spurious HTT signals as statistical outliers, based on comparisons to real transfer events with respect to several features of involved TEs, namely diversity within host lineages, similarity between lineages and genomic context in contigs.

3. What TE Sequences to Retain and to Compare?

Traditional searches for HTT involve comparing consensus sequences of TEs, each of which is obtained from a group of similar copies within a genome. Such a group of copies corresponds to a TE "family." In a TE family, copies must present some degree of DNA sequence similarity over a certain portion of their length (both set at 80%[32]). TEs from the same family are assumed to be derived from an ancestral TE that invaded a genome. One HTT event per family at most is eventually inferred from comparisons between consensus sequences (e.g., Pace et al.[33]) or between selected TE copies (e.g., Wallau et al.[34]). This approach provides an easy way to count HTT events and reduces the number of sequences to compare between species at the same time. However, although some families are well known and characterized in model species, most are only defined by the arbitrary 80% grouping criterion.[32] More importantly, statistical analyses of the evolution of certain TE types[35] have shown that a defined TE family may correspond to multiple separate acquisition events. The delineation of HTT events from family grouping is therefore not optimal.

To forgo the definition of TE families, the sequence of TE copies, rather than family consensus sequences, can be directly compared between species to detect horizontally transferred sequences prior to delineating HTT events.[15,36,37] This approach must however carefully consider the selection of TE copies to compare. Indeed, some TE copies identified across analyzed genomes may simply be too short and degraded for meaningful homology searches or prevent the generation of multiple sequence alignments that are necessary to build TE phylogenies (discussed below). One may also have to ensure that degraded copies representing non-overlapping parts of the same TE are not considered as unrelated TEs, which could inflate the number of inferred HTTs.[15] While it seems necessary to select TE copies of a minimal length, restriction to "full-size" or autonomous TE copies may lead to many HTT events being overlooked. For instance, 1197 of the 2248 HTT events we uncovered in insects involve TE copies that do not exceed 90% of the length of their respective family consensus (details not shown).[15] As it appears, there is currently no ideal solution to the issues related to the nature and selection of TE sequences to compare between species.

4. How to Reliably Infer Horizontal Transfer?

4.1. Excessive Sequence Similarity Between TEs from Distinct Species

HTT is suspected when some TEs appear so similar that they must have diverged more recently than their host species. To test this hypothesis for a given species pair, one evaluates if TEs show higher between-species sequence similarity than that of vertically-inherited genes (criterion 1). These "control" genes vary between studies and their sequences are generally obtained from prior dedicated genome annotations in the case of large-scale HTT studies.[15,36,37]

For a test based on criterion 1 to apply, divergence of DNA sequences must correlate with divergence time without being affected by natural selection. The condition imposes that sequence divergence be measured at synonymous sites of protein-coding regions. Unfortunately, divergence at synonymous sites (dS) cannot be estimated on TEs lacking distinguishable coding sequences, greatly limiting the number of copies that can be used in the context of a comprehensive study of HTT. This issue is mitigated by grouping TEs into families, which allows restricting comparisons to the most complete copies per family, but it is not solvable for TE types that simply lack proteins (e.g., SINEs).

As an alternative to the dS of TEs, some estimates of overall nucleotide divergence can be compared to the dS of genes[15] under the commonly accepted assumption that TEs, which are not useful to their hosts, evolve neutrally within lineages (i.e., are not subject to natural selection).[38] It is however unclear how much vertical evolution of TEs deviates from this assumption. Better estimates of global versus synonymous DNA sequence evolution of TEs may help evaluate this hypothesis. Without this knowledge and the possibility to compute the dS of TEs, it may be advised to impose a large difference between the interspecific similarity of control genes and that of TEs that would be considered as horizontally transferred.[15]

This difference itself has to be carefully evaluated. Typically, HTT is inferred if the (synonymous) between-species identity of TEs is higher than that of most (95% or more) genes inherited by the species sharing these TEs, with possible consideration of codon usage bias (the fact that certain codons are favored over synonymous counterparts in genes) to minimize the influence of natural selection on estimated divergence.[34] Such a threshold comes with a risk of false positive, as one cannot absolutely exclude that a particular TE was vertically inherited if few genes showing higher between-species identity were. In closely related species, HTT is supported by the fact that the fraction of TE pairs with low dS values can be much higher than that of genes of similar dS (**Figure 2**A). One may thus attribute these TE pairs as resulting from HTT using statistical approaches.[39]

However, dS distributions of genes and TEs cannot easily be compared in more divergent host lineages. This is because DNA sequence homology cannot be detected beyond a certain degree of divergence, which is 30–40% for most algorithms. Hence, most TEs that were vertically inherited from an ancient common ancestor, and which are highly divergent, will simply not be included in the dS distribution. By contrast, anciently diverged gene orthologs can be aligned using their conserved protein sequences and can thus take dS values that cannot possibly be measured between TEs (up to 100% dS and more, thanks to models of nucleotide evolution taking into account recurrent mutations). As a result, the distribution of dS of TEs may show a much lower mode than that of control genes simply because it is truncated at the maximal value of detectable DNA sequence homology (Figure 2B). Even in the absence of HTT, the number of low-dS TE pairs may still far exceed that of orthologous control genes of similar dS values, as TEs are much more numerous than genes in most species and, due to amplification, constitute a much greater pool of potentially similar pairs (if all individual TE copies are compared). It is therefore important to check if the distances measured between TEs are constrained (truncated) by the sensitivity of homology searches. If TE distances are well below the sensitivity of the homology search algorithm (~50% divergence), and below that of reference genes, then there is little evidence for truncation and for the presence of vertically inherited elements in homologous TEs. On the opposite, a truncated dS distribution in TEs will make it difficult to select a dS threshold below which vertical inheritance cannot reasonably explain interspecific similarities. A method to estimate this threshold should not only consider the synonymous divergence of vertically inherited genes, but also the TE composition of both species as it conditions the number of possible homologous TE pairs. Such a method would help the analysis of HTT between species with intermediate degrees of divergence, that is, where some fraction of orthologs show dS values that exceed the maximal level of detectable DNA sequence identity (~50%). In highly divergent species pairs where all gene dS far exceed this value (Figure 2C), it may safely be argued that no pairs of vertically inherited TEs should be preserved enough to show detectable homology.

4.2. Incongruent Phylogeny and Taxonomic Distribution of TEs

Two additional criteria listed in the introduction can validate or reject candidate HTT: (criterion 2) patchy TE distribution within

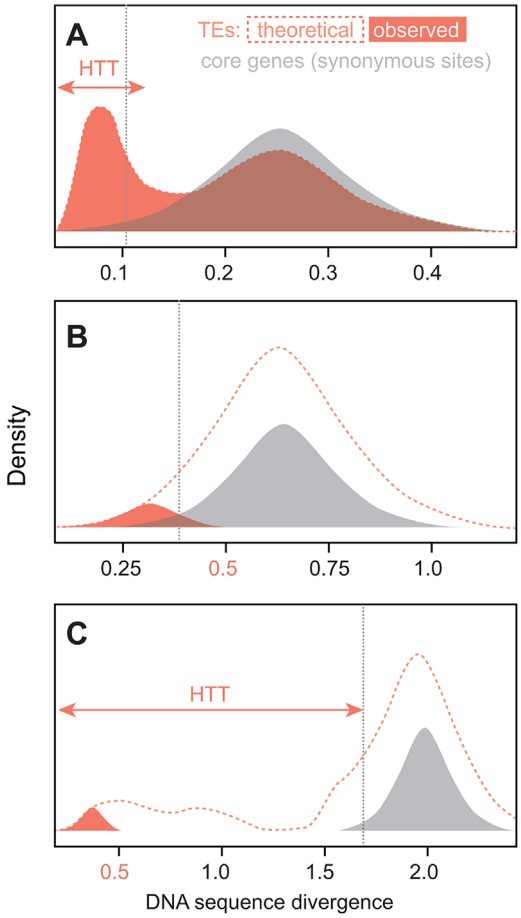

Figure 2. Comparison of nucleotide distance distributions of transposable elements (TEs) and core genes belonging to two lineages. A double arrow represents the occurrence of horizontal transfer of TEs (HTT) and the dotted segment represents the distance threshold below which similar TEs are considered as horizontally transferred. The distance value in red indicates the typical distance above which homology between DNA sequences cannot be found, which may constrain the observed TE distances (filled curves) to low values. Theoretical distance distributions of TEs (doted curves) assume that all homologies are found. The higher surface of TE curves represents the higher number of TEs compared to genes. In (A) lineages are closely related, such that the whole distance distribution of TEs is known (dotted and filled curves are identical). A clear excess of short distances between TEs in comparison to core genes indicates HTT. In (B) between more divergent lineages, distance can only be measured between the most similar TEs, so that the full distance distribution is not known. The observed distribution of TE distances suggests HTT, but it may instead merely represent the most conserved vertically inherited TEs. In (C) host lineages are highly divergent, hence, TEs inferred as homologous are very unlikely to have been vertically inherited.

the host phylogeny, and (criterion 3) incongruence between TE and host phylogenies.[17] While intuitive and visually appealing, these criteria appear to have limited usefulness in a study that aims at quantifying HTT on a large scale. Indeed, they do not consider a pair of species, but a group of species which, as a whole, can share similar TEs through both HTT and vertical inheritance. A binary "yes or no" test may indicate that a type of TEs was horizontally transferred, but cannot tell how many HTT events are needed to explain the TE phylogeny, hence, it cannot confirm a number of events suggested by criterion 1. In principle, a quantitative test based on criterion 3 could be inspired from methods developed to estimate the relative importance of co-speciation (vertical transmission) and host-switching (HT) in the co-evolutionary dynamics between symbionts and their hosts.[40] However, adaptation of these methods would have to consider that a TE diversifies and degrades within its hosts independently of speciation events, contrarily to a regular symbiont. For meaningful phylogenetic comparisons, this TE should be a single genomic locus represented by orthologous TE copies, that is, homologous TEs found in equivalent genomic locations. Only under this condition may the phylogeny of a TE be expected to conform that of its hosts. But the presence of orthologous TEs is enough to prove the vertical inheritance of these TEs,[41] voiding the need to compare phylogenies. Absence of visible orthologous TEs among genomes, however, does little to confirm HTT as this absence may also result from the excision, degradation, or ineffective detection of TE copies. Granted, blatant incongruences between TE and host trees, such as the grouping of TEs found in distinct phyla, are not reasonably explained by non-orthology. However, incongruent divergence times are more powerfully evaluated by criterion 1.

As for patchy TE distribution among hosts, this pattern may simply result from degeneration of vertically inherited elements in certain lineages, especially if these TEs are ancient (e.g., Fawcett and Innan[42]). Conversely, non-patchy presence of a TE in a host clade could result from the tendency of TEs to horizontally transfer between closely related species.[15] Criterion 2 may thus lead to both false positives and false negatives. Its ease of use can still enable a global assessment of vertical inheritance of TEs in a large-scale study. One expects the degree of patchiness to correlate with the divergence time between two lineages that vertically inherited a TE from a common ancestor (as it reflects the age, hence, the degradation of the TE). Such a correlation is not expected if the TE was horizontally transferred.[15]

5. How to Delineate Transfer Events?

HTT events between two lineages are generally counted as the numbers of pairs of TE families presenting horizontally transferred elements, with precautions avoiding considering a given TE family as contributing to more than one such event.[36,43] As previously argued, TE families may not represent HTT events well. Indeed, families are defined independently in each genome, irrespective of the phylogenetic patterns that HTT should yield. Methods that are not based on prior definition of TE families may therefore be preferred for the task of delineating HTT events.

Two clustering approaches relying on comparisons of individual TE copies have been developed. One method[44] uses single linkage clustering to connect TE copies, irrespective of their host genome, if their identity reaches a threshold set at 80%. One HTT event is counted for each cluster that comprises host lineages between which vertical transfer is excluded, mostly by criterion 1 (more than one HTT event could be inferred in a

cluster that comprises more than two such lineages). While this method does not strictly require a transfer event to correspond to a TE family within each species, it relies on the same method as used for family definition (identity-based single-linkage clustering) and may therefore yield similar results. Sensitivity, but also accuracy, of this method is certainly influenced by the identity threshold set for clustering.

We developed another clustering method[15] based on the expectation that separate HTT events between the same two host lineages should yield as many clades of TE copies grouping the lineages involved (**Figure 3**). The method considers two pairs of homologous TEs between the same two lineages as representing different HTT events if similarity of TEs between lineages is greater than similarity within lineage. This method technically clusters pairs of TEs rather than individual TEs. It has the advantage of not relying on an absolute identity threshold, but it is complex to implement as successive rounds of clustering are required to estimate a conservative number of HTT events if more than two species are involved. Separate transfer events may not be resolved as distinct clusters if the TEs involved in different transfers are similar enough to form homologous pairs used in the clustering procedure.

These clustering methods are relatively insensitive as they do not exploit the relationships between TEs beyond their presence in the same or different clusters. In particular, the "chaining phenomenon," to which single linkage clustering is prone, might group non-homologous TEs in the same cluster. Spurious grouping of non-homologous sequences may cause a single HTT event to be inferred for several unrelated TE families, and thus lead to underestimates of the number of events. This risk has not been evaluated, to our knowledge. If required, other clustering algorithms less prone to chaining may be considered, as well as methods that subdivide clusters into well-connected "communities" of sequences.[15]

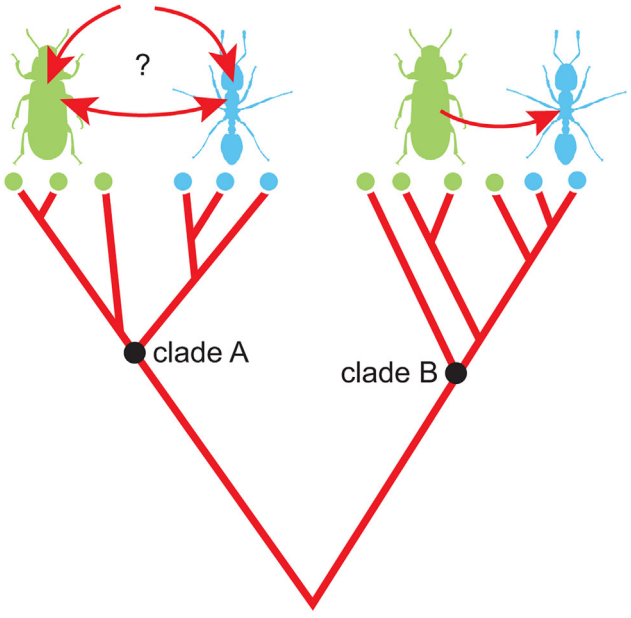

Figure 3. If a phylogeny of TE copies (shown as colored circles) presents several clades grouping two distantly related host species, at least as many HTT events must have involved these species' lineages (assuming other criteria for inferring HTT are fulfilled, see text). In clade A, TEs from each species are monophyletic, hence, the source of TEs found in either species cannot be inferred with confidence. In clade B, TEs from the beetle are paraphyletic with respect to TEs from the ant, so an HTT from beetle to ant can confidently be inferred.

Beside limitations related to sensitivity, TE clusters give little contextual information about identified transferred events, in particular the age and direction of transfers. A more explicit method should analyze phylogenetic trees of TE copies. Clades representing HTT events (Figure 3) can be counted via tests of monophyly. The direction of transfer may be inferred if the diversity of TEs from one lineage is embedded in the diversity of the other (which is then considered paraphyletic with respect to the former, clade B on Figure 3), as the donor lineage would already carry other similar TEs at the time of the transfer. Such a pattern also excludes the possibility that two lineages sharing similar TEs acquired them from a third (potentially unknown) lineage, as these events should instead lead to reciprocal monophyly of TEs from different lineages within a clade (as in clade A).

As appealing as a tree-based method may look, its effectiveness is compromised by the complexity and idiosyncrasy of TE diversification. Taking TEs that are too divergent or degraded will limit the length of the shared aligned region to build a phylogeny from, as well as the number of informative sites. On the other hand, missing or discarding certain TE copies may mislead inferences about monophyly or paraphyly. These shortcomings currently restrict the counting of HTT events to cluster-based methods. Finally, all aforementioned methods provide a minimum number of HTT events required to explain the data, not the most likely number of events. While case studies on HTT try to avoid false positives much more than they do false negatives, and rightfully so, we suggest that new quantitative approaches not aimed at outlining particular HTT cases provide the most accurate numbers of events rather than conservative estimates.

6. Conclusion

The increasing availability of genomic resources enables inference on the breadth of HTT and on the factors shaping this process. This goal cannot be achieved by traditional approaches that rely on manual curation of TEs and visual inspections of sequences, alignments, and trees. New methods must be automated to more reliably detect and annotate TEs, to remove contaminants, to assess the horizontality of transfers against alternative scenarios, and to delineate and characterize HTT events. The latter task, in particular, is not well served by the customary aggregation of TEs into families within genomes. Using individual TE copies involves other problems that relate to the selection of sequences to be compared and the number of HTT events to infer from retained homologies.

Reflecting upon these issues, it appears to us that the detection and classification of TEs should be undertaken jointly with the inference about the fraction of each type of transfer (vertical or horizontal) that best explains their diversity and distribution. This joint inference would be performed on a pangenome, that is, the aggregated sequence of all genomes

under study. Due to the complexity of the data and of the underlying evolutionary processes, statistically supported inferences may require simulations based on evolutionary models of transposition and TE diversification within lineages, considering co-divergence between hosts and TEs, and HTT. Such an integrated approach may be able to delineate trans-species TE lineages that result from vertical or horizontal transfer, give probabilities for the existence of HTT events, for their directions, their age and possibly other features. Its conception would certainly require rethinking the identification and classification of TEs, and overcoming obstacles that prevent reconstructing comprehensive phylogenies of TEs.

Whatever these developments may be, efficient large-scale studies of HTT will bring significant insights on the evolution of TEs, not only as molecular symbionts that deeply affect genomes,[45,46] but also as molecular fossils[47,48] that can reveal hitherto inaccessible information about present and past interactions between species.

Abbreviations

dS, divergence at synonymous sites; HT, horizontal transfer; HTT, horizontal transfer of transposable elements; TE, transposable element.

Acknowledgments

This work was supported by Agence Nationale de la Recherche grant to CG (ANR-15-CE32-0011-01 TransVir) and intramural funds from the Centre National de la Recherche Scientifique and the Université de Poitiers. We thank Dr. Brachhold for her editorial work and reviewers for their constructive comments.

Conflict of Interest

The authors have declared no conflict of interest.

Keywords

bioinformatics, genome evolution, phylogenetics, repeat elements, retrotransposons, transposons

Received: September 27, 2017
Revised: November 22, 2017
Published online: December 28, 2017

[1] S. M. Soucy, J. Huang, J. P. Gogarten, *Nat. Rev. Genet.* **2015**, *16*, 472.
[2] W. F. Martin, *BioEssays* **2017**, *39*, 1700115.
[3] N. L. Craig, R. Craigie, *Mobile DNA II*. American Society for Microbiology Press, Washington (DC) **2002**.
[4] S. B. Daniels, K. R. Peterson, L. D. Strausbaugh, M. G. Kidwell, A. Chovnick, *Genetics* **1990**, *124*, 339.
[5] A. M. Ivancevic, A. M. Walsh, R. D. Kortschak, D. L. Adelson, *BioEssays* **2013**, *35*, 1071.
[6] S. Schaack, C. Gilbert, C. Feschotte, *Trends Ecol. Evol.* **2010**, *25*, 537.
[7] G. L. Wallau, M. F. Ortiz, E. L. Loreto, *Genome Biol. Evol.* **2012**, *4*, 689.
[8] B. R. Dotto, E. L. Carvalho, A. F. Silva, L. F. Duarte Silva, P. M. Pinto, M. F. Ortiz, G. L. Wallau, *Bioinformatics* **2015**, *31*, 2915.
[9] H. S. Malik, W. D. Burke, T. H. Eickbush, *Mol. Biol. Evol.* **1999**, *16*, 793.
[10] J. C. Silva, E. L. Loreto, J. B. Clark, *Curr. Issues Mol. Biol.* **2004**, *6*, 57.
[11] C. Gilbert, S. Schaack, C. Feschotte, *Med. Sci. (Paris)* **2010**, *26*, 1025.
[12] C. Gilbert, S. Schaack, J. K. Pace, P. J. Brindley, C. Feschotte, *Nature* **2010**, *464*, 1347.
[13] X. Guo, J. Gao, F. Li, J. Wang, *Sci. Rep.* **2014**, *4*, 5119.
[14] S. Kuraku, H. Qiu, A. Meyer, *Genome Biol. Evol.* **2012**, *4*, 929.
[15] J. Peccoud, V. Loiseau, R. Cordaux, C. Gilbert, *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 4721.
[16] S. Venner, V. Miele, C. Terzian, C. Biémont, V. Daubin, C. Feschotte, D. Pontier, *PLoS Biol.* **2017**, *15*, e2001536.
[17] E. L. Loreto, C. M. Carareto, P. Capy, *Heredity* **2008**, *100*, 545.
[18] M. L. Carroll, A. M. Roy-Engel, S. V. Nguyen, A. H. Salem, E. Vogel, B. Vincent, J. Myers, Z. Ahmad, L. Nguyen, M. Sammarco, W. S. Watkins, J. Henke, W. Makalowski, L. B. Jorde, P. L. Deininger, M. A. Batzer, *J. Mol. Biol.* **2001**, *311*, 17.
[19] A. F. Smit, A. D. Riggs, *Proc. Natl. Acad. Sci. USA* **1996**, *93*, 1443.
[20] J. Xing, A.-H. Salem, D. J. Hedges, G. E. Kilroy, W. S. Watkins, J. E. Schienman, C.-B. Stewart, J. Jurka, L. B. Jorde, M. A. Batzer, *J. Mol. Evol.* **2003**, *1*, S76.
[21] D. R. Hoen, G. Hickey, G. Bourque, J. Casacuberta, R. Cordaux, C. Feschotte, A.-S. Fiston-Lavier, A. Hua-Van, R. Hubley, A. Kapusta, E. Lerat, F. Maumus, D. D. Pollock, H. Quesneville, A. Smit, T. J. Wheeler, T. E. Bureau, M. Blanchette, *Mobile DNA* **2015**, *6*, 13.
[22] E. Lerat, *Heredity* **2010**, *104*, 520.
[23] A. F. A. Smit, R. Hubley, P. Green, unpublished data. Current Version: open-4.0.6 (RMLib: 20160829 & Dfam: 2.0).
[24] C. Goubert, L. Modolo, C. Vieira, C. ValienteMoro, P. Mavingui, M. Boulesteix, *Genome Biol. Evol.* **2015**, *7*, 1192.
[25] T. J. Treangen, S. L. Salzberg, *Nat. Rev. Genet.* **2011**, *13*, 36.
[26] L. Faino, M. F. Seidl, E. Datema, G. C. M. van den Berg, A. Janssen, A. H. J. Wittenberg, B. Thomma, *mBio* **2015**, *6*, 11.
[27] F. J. Krsticevic, C. G. Schrago, A. B. Carvalho, *G3-Genes Genomes Genet.* **2015**, *5*, 1145.
[28] G. Koutsovoulos, S. Kumar, D. R. Laetsch, L. Stevens, J. Daub, C. Conlon, H. Maroon, F. Thomas, A. A. Aboobaker, M. Blaxter, *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 5053.
[29] M. Ballenghien, N. Faivre, N. Galtier, *BMC Biol.* **2017**, *15*, 25.
[30] C. Gilbert, R. Cordaux, *Genome Biol. Evol.* **2013**, *5*, 822.
[31] S. Kumar, M. Jones, G. Koutsovoulos, M. Clarke, M. Blaxter, *Front. Genet.* **2013**, *4*, 237.
[32] T. Wicker, F. Sabot, A. Hua-Van, J. L. Bennetzen, P. Capy, B. Chalhoub, A. Flavell, P. Leroy, M. Morgante, O. Panaud, E. Paux, P. SanMiguel, A. H. Schulman, *Nat. Rev. Genet.* **2007**, *8*, 973.
[33] J. K. Pace, C. Gilbert, M. S. Clark, C. Feschotte, *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 17023.
[34] G. L. Wallau, P. Capy, E. Loreto, A. Le Rouzic, A. Hua-Van, *Mol. Biol. Evol.* **2016**, *33*, 1094.
[35] A. C. Wacholder, C. Cox, T. J. Meyer, R. P. Ruggiero, V. Vemulapalli, A. Damert, L. Carbone, D. D. Pollock, *PLoS Genet.* **2014**, *10*, 1004482.
[36] C. Bartolome, X. Bello, X. Maside, *Genome Biol.* **2009**, *10*, R22.
[37] M. El Baidouri, M. C. Carpentier, R. Cooke, D. Gao, E. Lasserre, C. Llauro, M. Mirouze, N. Picault, S. A. Jackson, O. Panaud, *Genome Res.* **2014**, *24*, 831.
[38] D. J. Lampe, D. J. Witherspoon, F. N. Soto-Adames, H. M. Robertson, *Mol. Biol. Evol.* **2003**, *20*, 554.
[39] L. Modolo, F. Picard, E. Lerat, *Genome Biol. Evol.* **2014**, *6*, 416.
[40] D. M. de Vienne, T. Giraud, J. A. Shykoff, *J. Evol. Biol.* **2007**, *20*, 1428.
[41] A. Lee, A. Nolan, J. Watson, M. Tristem, *Philos. Trans. R. Soc. Lond. B: Biol. Sci* **2013**, *368*, 20120503.
[42] J. A. Fawcett, H. Innan, *Mol. Biol. Evol.* **2016**, *33*, 2593.

[43] C. Gilbert, S. S. Hernandez, J. Flores-Benabib, E. N. Smith, C. Feschotte, *Mol. Biol. Evol*. **2012**, *29*, 503.
[44] M. El Baidouri, O. Panaud, *Genome Biol. Evol*. **2013**, *5*, 954.
[45] A. Le Rouzic, S. Dupas, P. Capy, *Gene* **2007**, *390*, 214.
[46] S. Venner, C. Feschotte, C. Biemont, *Trends Genet*. **2009**, *25*, 317.
[47] V. V. Kapitonov, J. Jurka, *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 6569.
[48] A. Suh, C. C. Witt, J. Menger, K. R. Sadanandan, L. Podsiadlowski, M. Gerth, A. Weigert, J. A. McGuire, J. Mudge, S. V. Edwards, F. E. Rheindt, *Nat. Commun*. **2016**, *7*, 11396.