Transposable Elements and Factors Influencing their Success in Eukaryotes

ELLEN J. PRITHAM

Department of Biology, University of Texas, Arlington, 501 S. Nedderman Drive, Arlington, TX 76019.

Address correspondence to Ellen J. Pritham at the address above, or e-mail: pritham@uta.edu.

Abstract

Recent advances in genome sequencing have led to a vast accumulation of transposable element data. Consideration of the genome sequencing projects in a phylogenetic context reveals that despite the hundreds of eukaryotic genomes that have been sequenced, a strong bias in sampling exists. There is a general under-representation of unicellular eukaryotes and a dearth of genome projects in many branches of the eukaryotic phylogeny. Among sequenced genomes, great variation in genome size exists, however, little difference in the total number of cellular genes is observed. For many eukaryotes, the remaining genomic space is extremely dynamic and predominantly composed of a menagerie of populations of transposable elements. Given the dynamic nature of the genomic niche filled by transposable elements, it is evident that these elements have played an important role in genome evolution. The contribution of transposable elements to genome architecture and to the advent of genetic novelty is likely to be dependent, at least in part, on the transposition mechanism, diversity, number, and rate of turnover of transposable elements in the genome at any given time. The focus of this review is the discussion of some of the forces that act to shape transposable element diversity within and between genomes.

Key words: gene mapping, genome evolution, transposable element

Despite great variation in genome size, little difference in the total number of cellular genes is observed among eukaryotes (Gregory 2005; Feschotte and Pritham 2007b). In many instances, the cellular genes may represent just a small sliver of the total genomic space. These genic regions are the most stable part of an organism’s genome, largely due to purifying selection acting to preserve gene function. For many eukaryotes, the remaining genomic space is extremely dynamic and predominantly composed of a menagerie of populations of transposable elements (TEs) (for review Kidwell and Lisch 2001; Craig et al. 2002). The TEs can be viewed as genomic squatters, shackling up in the genome without providing any direct benefit to the host. TE persistence is then a reflection of both the ability to replicate faster than the host cell and of the balance between TE reinfection through outcrossing, horizontal transfer and TE loss through excision, sequence erosion, selection and drift (Craig et al. 2002). Given the dynamic nature of the genomic niche filled by TEs, it is evident that these elements have played an important role in genome evolution (Kidwell and Lisch 2002; Wessler 2006; Feschotte and Pritham 2007a). The contribution of TEs to genome architecture and to the advent of genetic novelty is likely to be dependent, at least in part, on the transposition mechanism, diversity, number and rate of turnover of TEs in the genome at any given time. The focus of this review is the discussion of some of the forces that act to shape TE diversity within and between genomes.

Transposition Mechanisms and TE Classification

TEs are classified based on the nature of the transposition intermediate (RNA or single or double-stranded DNA [ssDNA or dsDNA]), their structural features, and by homology to known elements (Feschotte and Pritham 2007a; Wicker et al. 2007). Broadly, TEs are divided into 2 classes based on whether the transposition intermediate is RNA (class 1) or DNA (class 2) (Figure 1). Class 1 TEs or retrotransposons all require reverse transcriptase to copy their RNA into DNA and can be subdivided into 3 groups based on whether the transposition intermediate is RNA (class 1) or DNA (class 2) (Figure 1) (for review Eickbush and Malik 2002). Class 1 TEs or retrotransposons all require reverse transcriptase to copy their RNA into DNA and can be subdivided into 3 groups based upon mechanism of integration (Figure 1) (for review Eickbush and Malik 2002). The reverse transcriptase encoded by each of these groups shares 7 blocks of conserved sequences suggesting that they are related by descent and share a common ancestor, although in the very distant past (Xiong and Eickbush 1990). The diversity of structures and integration mechanisms of retrotransposons is a testament to their ability to adapt and change. The long terminal repeat (LTR) and the tyrosine recombinase (YR) retrotransposons are both flanked by LTRs but they differ...
in the mechanism of integration (for review Cappello et al. 1984; Eickbush and Malik 2002; Goodwin and Poulter 2004; Poulter and Goodwin 2005). The LTR retrotransposons utilize an integrase, which is evolutionarily related to the transposase encoded by cut-and-paste DNA transposons, whereas the YR retrotransposons utilize a tyrosine recombinase. The non-LTR and probably also Penelope-like retrotransposons transpose via a process termed target-primed reverse transcription (TPRT) and integration is mediated by either an apurinic/apyrimidinic or a restriction-like endonuclease (Eickbush and Malik 2002; Evgen’ev and Arkhipova 2005).

Class 2 or DNA transposons transpose via a DNA intermediate. “Classic” DNA transposons are excised as dsDNA intermediate and reintegrated elsewhere in the genome (for review Feschotte and Pritham 2007a). These “cut-and-paste” transposons are exemplified by a relatively simple structure typically consisting of a single ORF, encoding a transposase, flanked by terminal inverted repeats (TIRs) and they are usually less than 5 kb in size (Figure 1). Helitrons, which represent a second major subclass of DNA transposons, are most likely mobilized as ssDNA intermediates through a replicative, rolling-circle-like mechanism. They encode a putative protein with a central domain homologous to the rolling-circle replication proteins encoded by rolling-circle genetic elements (e.g. plasmids, phages) and a C-terminal domain related to the PIF1 group of DNA helicases. Plant Helitrons also encode 1–3 additional putative proteins homologous to ssDNA binding proteins (for review Kapitonov and Jurka 2006, 2007; Feschotte and Pritham 2007a). Finally, Mavericks also encode a retroviral-like integrase and therefore their transposition cycle involves integration of a dsDNA intermediate. TEs that utilize unknown transposition mechanisms are being discovered and described at an unprecedented rate due to the tremendous abundance of genome data now available for scrutiny (Kapitonov and Jurka 2001; Goodwin et al. 2003; Feschotte 2004; Feschotte and Pritham 2005; Feschotte and Pritham 2007a; Pritham et al. 2007).

The Distribution of TEs in the Tree of Life

The genomes of hundreds of eukaryotes have been or are in the process of being sequenced, providing the opportunity to analyze a vast quantity of TE data. However, it should be noted that there is a strong bias in this data set toward animals and fungi (e.g., opisthokonts) and to a lesser degree apicomplexan (which include malaria parasites) and plant genomes. This bias is evident when the number and incidence of projects are considered in a phylogenetic context (see Figure 2). Surprisingly, many branches of the eukaryotic phylogeny have yet to be sampled. This bias in sampling makes difficult, and should probably preclude, any broad generalizations about TE diversity and distribution in eukaryotes. Surveys of the TE populations within the sequenced genomes have made it clear that TEs are indeed widespread and persistent entities in metazoans, fungi, and plants. In addition, a positive correlation is often seen between genome size and TE content in these genomes (Feschotte and Pritham 2007b). However, the distribution and abundance of TEs in unicellular eukaryotic organisms is far less understood,
hampered by the paucity of sequencing projects and therefore the scarcity of data, as well as a less systematic and careful scrutiny of the sequenced genomes.

Genome Sequence Comparisons Reveal Patterns in TE Diversity

Examination of complete genome sequence data allows the analysis of an organism’s genome at a single point in time. Dramatic differences between the success of retrotransposons and DNA transposons are revealed when surveys of genome sequence data are undertaken (Figure 3). Variation occurs in terms of total number, composition, and location of TEs within and between genomes. For example, both the human and mouse genomes are dominated by retrotransposons (Lander et al. 2001; Waterston et al. 2002), whereas DNA transposons have been relatively more successful in the genome of the nematode Caenorhabditis elegans (Consortium 1998). In the budding and fission yeast genomes, only a few hundreds LTR retrotransposons are found, despite a 1 billion years of divergence (Hedges 2002). Are these patterns purely stochastic or are they the result of evolutionary forces acting to influence TE success and shape genome architecture? If TE diversity and success was attributed solely to random gain and loss, than given the rapid turnover of TEs, no trends should be apparent. Given the rapid turnover of TEs and a constant rate of TE introduction and amplification, if stochasticity was the major determinant in shaping these patterns than no trends should be apparent. However, some trends in TE composition do appear to be conserved. For example, the human, macaque, mouse, rat, and dog genomes share a strikingly similar pattern of TE composition, despite continuous and extensive lineage-specific TE activity (Pace and Feschotte 2007). Similarly, analysis of the sequenced genomes of 12 Drosophila species separated by up to 40 million years of evolution shows that retrotransposons predominate in all these species, whereas DNA transposons consistently represent less than 20% of the total TE content (Clark et al. 2007). What does the conservation of TE composition between species tell us, and can inferences be made about the history of a species when a difference in pattern is observed?
Figure 3. Variation of TE composition across genomes. For each species, the relative proportion of RNA and DNA TEs was calculated. The data were compiled either from the corresponding papers reporting draft genome sequences or from the following sources: nematode C. Feschotte, personal communication; rice (Jiang et al. 2004) and N. Jiang, personal communication; Entamoeba (Pritham et al. 2005); Giardia lamblia (Arkhipova and Morrison 2001); Trichomonas vaginalis (Pritham et al. 2007); and Ellen Pritham (unpublished data). The species are Hs = Homo sapiens, Mm = Mus musculus, the nematode, Caenorhabditis elegans, Dm = Drosophila melanogaster, Dc = Drosophila erecta, Ag = Anopheles gambiæ, Aa = Aedes aegypti, Ed = Entamoeba dispar, Eh = E. histolytica, Ei = E. invadens, Em = E. moshkovskii, Sc = Saccharomyces cerevisia, Sp = Schizosaccharomyces pombe, At = Arabidopsis thaliana, Os = Oryza sativa japonica, Gi = Giardia lamblia, Tv = Trichomonas vaginalis.

Forces that Influence TE Diversity can be Subdivided into Three Groups based on the Scale at which They Act: Molecular, Genetic, and Environment

Molecular Properties of the TE

Molecular properties of the TE can function to influence TE activity and therefore accumulation in genomes, as well as the propensity of TEs to be vertically or horizontally transmitted. Some examples of molecular factors include transposition mechanism (Eickbush and Malik 2002), the pattern and timing of transposition (which can be influenced by tissue and temporal specific promoter regions or alternative splicing) (Rio 2002), TE autoregulation (Rio 2002), targeting (Devine and Boeke 1996; Eickbush and Malik 2002); (for review Lesage and Todeschini 2005) and infectivity (or the ability to move cell to cell or between organisms) (Malik et al. 2000). The non-LTR retrotransposons provide an excellent example of how the mechanism of transposition can influence both the ability of TEs to colonize specific genomic niches and their ability to propagate horizontally, as well as limit the selective impact of new insertions. For non-LTR retrotransposons, integration is coupled to reverse transcription of the mRNA in a process termed TPRT (Christensen and Eickbush 2005). Due to the inherent instability of RNA it is probable that the DNA form of the retrotransposon would be more likely to move horizontally than the RNA itself. Because the DNA is integrated directly into the nuclear genome, as it is reverse transcribed from RNA, the window of opportunity for horizontal movement appears to be temporally narrow for these elements (for review Eickbush and Malik 2002). In addition, for most non-LTR elements the TPRT process results in many copies that are truncated in the 5’-end, with the promoters being lost and further propagation inhibited. These “dead on arrival” copies are typically the most abundant product of TPRT. Thus, for a successful horizontal transfer event to occur, a rare complete non-LTR element would have to be transferred. Also, many non-LTR retrotransposons have target-site specificity, which minimizes the genomic space where integration can occur. For example, R2 and R4 elements target the 28S genes and CRE and NeSL are targeted to splice-leader sequences (Burke et al. 1987; Teng et al. 1995; Malik and Eickbush 2000) (and for review Eickbush and Malik 2002). Both TPRT and site-specific targeting are expected to reduce the potential of non-LTR retrotransposons to be moved between species by horizontal transfer. On the other hand, targeted integration into other highly repeated sequences minimizes the deleterious effects of these elements, which facilitates their vertical persistence (Malik et al. 1999). Indeed, in some cases non-LTR retrotransposons have even gone extinct in a number of mammals (Casavant et al. 2000; Grahn et al. 2005; Rinehart et al. 2005; Cantrell et al. 2008).

In contrast, horizontal movement of both DNA transposons and LTR retrotransposons appears to be more frequent. Recurrent waves of horizontal transfer of DNA transposons are thought to explain the diversity of recently active DNA transposons in the genome of the little brown bat (Myotis lucifugus) (Ray et al. 2008). The diversity of TEs in M. lucifugus deviates dramatically from the pattern observed in the genomes of other well-studied mammals where no recent DNA transposon activity has been reported and retrotransposons predominate (Pritham and Feschotte 2007; Ray et al. 2007, 2008).

The compact eukaryotic genomes, like those of the yeasts Saccharomyces cerevisiae and Schizosaccharomyces pombe, provide insight into how the ability to target specific genomic regions has allowed LTR retrotransposons to persist in these highly compact, gene-rich genomes and highlights the complex interplay between molecular properties of the TE and genetic properties of the genome. Both genomes are small (~12.5 Mb) and have limited intergenic space, proving a hazardous terrain for TE insertion. TEs that target specific sites in the genome, in effect narrow the scope of their mutagenic impact to these specific genomic locales. The Ty1 and Ty3 retrotransposons of S. cerevisiae target the upstream of genes transcribed by RNA polymerase III (Kim et al. 1998). For example, 90% of Ty1 elements are within the 750 bp flanking a tRNA gene. Ty3 retrotransposons, which naturally inhabit the genome of
S. paradoxus, actively target the silent chromatin located at the mating loci and near the telomeres (Zou et al. 1996). In S. pombe, the TFI–2 elements have evolved a different targeting strategy whereby they integrate upstream of RNA polymerase II transcripts (Leem et al. 2008). Remarkably, the targeting strategy of all these retrotransposons has evolved independently and is mediated by direct interactions between host factors and the integrase proteins encoded by the elements (Gao et al. 2008). This illustrates how TEs can establish intimate and long-term relationships with their hosts. Targeting potential does not seem to be a universal feature of LTR retrotransposons and the location and mechanism for targeting differs among elements suggesting that this intrinsic property has evolved repeatedly at different evolutionary times.

Ecological Influences on Host TE Populations

The ecology of the host population and environment are also likely to play an indirect but important role in the success of TEs within genomes. Examples of these forces include parasite load, environmental quality, resource availability, and competition. These ecological components can affect the breeding system, effective population size of the host, and exposure to vectors of horizontal transfer which in turn influence the ability of TEs to proliferate and fix in a genome (Wright and Schoen 1999; Arkhipova and Meselson 2000; Lynch and Conery 2003). Of course, these forces are not mutually exclusive and they are often interconnected. For example, effective population size will affect TE success directly by determining the likelihood of new insertions to reach fixation within a population and also indirectly by influencing other features of genome organization and gene structure which determines whether an insertion is likely to be detrimental or largely neutral for the fitness of the host (Lynch 2007). In addition, none of these factors are static with respect to time. Sex and recombination have been postulated to be of major import in the success of TEs in populations, and their influence has been investigated in a number of different systems (Arkhipova and Meselson 2000; Wright et al. 2003; Valizadeh and Crease 2008). Because TE insertions do not in general have a selective advantage they will be eventually lost over time when sequence erosion and drift leads to a higher rate of loss than birth or reinfection occurs through zygote formation and horizontal transfer. Therefore it is expected that strictly asexual organisms will eventually be purged of TEs. The bdelloid rotifers are a widely diverged group of animals where meiosis, males, and sex have never been observed. Studies investigating TE distribution and diversity in the genome of bdelloid rotifers reveal the presence of a diverse assortment of DNA transposons and 2 families of retrovirus-like elements, but an apparent lack of non-LTR retrotransposons (Arkhipova and Meselson 2000; Arkhipova and Meselson 2005). Because DNA transposons and retroviruses, but not non-LTR retrotransposons, are thought to be prone to HT, this biased TE composition was interpreted to support the theory that the lack of sex indeed result in a general purging of TEs that cannot be reintroduced horizontally. Recent studies from the same group show that bdelloid rotifers have been subject to recurrent and massive horizontal transfer events, a finding that may lend support to the explanation that the biased TE composition of their genomes is driven by the presumed asexuality of these organisms.

However, breeding systems do not seem to be an overriding force in determining the diversity of TEs in other organisms. A survey of TEs in the ~20-Mb genomes of 4 related Entamoeba species reveals that a diverse set of TEs (including retrotransposons and DNA transposons) contribute to between 5% and 8% of these genomes (Pritham et al. 2005). Entamoeba are unicellular amoebas that are either parasitic or free living and have a genome size of ~20 Mb. They reproduce asexually using binary fission and no sexual stage has even been observed, however, a cryptic sexual stage has not been formally ruled out. Yet different species of Entamoeba display dramatic variation in the type and success of TEs populating their genomes. The closely related extracellular, human parasites E. bistolytica and E. dispar genomes are packed with non-LTR retrotransposons, whereas the genomes of the reptilian parasite, E. invadens and the free living, E. moshkovskii have relatively few retrotransposons but host a wide diversity of DNA transposons. No evidence of horizontal transfer was detected and phylogenetic analysis suggested that many of the TE lineages detected in these 4 species were present in their common ancestor and most likely have been vertically inherited. Demographic factors like population bottlenecks are expected to play a role in TE diversity both due to alterations in the efficacy of natural selection as well as the impact of genetic drift in changing genetic diversity. In addition, it has been suggested that the mechanism of transposition that leads to differential accumulation, such as is seen between retrotransposons (copy and paste) and DNA transposons (cut and paste) predisposes them to differential success in an effective population size dependent manner (Lynch and Conery 2003). Therefore, it may be that E. moshkovskii and E. invadens have an effective population size sufficient to allow for DNA transposon accumulation, whereas E. dispar and E. bistolytica do not. Indeed, studies have revealed that E. dispar and E. bistolytica have gone through recent population bottlenecks (Ghosh et al. 2000). It is clear that a single intrinsic or extrinsic factor is not sufficient to explain TE diversity within a genome.

Are TEs Ubiquitous Components of Eukaryotic Taxa?

The publication of several, unicellular eukaryotic genome papers fail to report the presence of TEs in the respective genomes. Included in the list are the red alga Cyanidioschyzon merolae, supergroup Plantae (16.5 Mb) Matsuzaki, Misumi et al. 2005, the Apicomplexans: Babesia bovis (9.4 Mb) Brayton et al. 2007 Cryptosporidium hominis (9.2 Mb) (Xu et al. 2004), C. parvum (9.09 Mb) Abrahamsen et al. 2004,
Several of the 8 species in the list inhabit the cell of another organism and their existence depends on exploitation of that organism’s resources. The eighth organism, *Cyanidioschyzon merolae*, is an extremophile, inhabiting an acidic hot-spring (Misumi et al. 2005). Intracellular parasites might be expected to be under a strong selective constraint to maintain cell size in order to occupy the cellular niche, effectively. In addition, genome reduction is a general feature of intracellular pathogens. Most bacterial intracellular pathogens, in concert with having a reduced genome are also depauperate in TEs, with the notable exception of some *Rickettsiales* species (Masui et al. 1999; Duron et al. 2005; Simser et al. 2005; Sanogo et al. 2007; Cordaux 2008). A key to this puzzle may be that in addition, to providing resources, perhaps the intracellular host environment acts as a shield from exposure to the vectors that are necessary for TE horizontal transfer, as mentioned above. Therefore, it might be reasonable to expect that an intracellular pathogen might initially lose its TEs through selection and drift, and then maintain, a TE-free genome, as a side effect of the protection from vectors, afforded by the inhabitation of the intracellular environment. Careful examination of the genomes of unicellular eukaryotes will provide a better picture of the relative importance of different factors in explaining the pattern seen.

**Some Unicellular Eukaryotic Genomes are Populated by TEs**

Multiple phylogenetically diverse extracellular pathogens have had their genomes sequenced and display a variety of TEs that have accumulated with varied levels of success. For example, TEs have been identified in the genomes of *Leishmania major*, *Trypanosoma brucei*, *T. cruzi* (Kinetoplastids), *Trichomonas vaginalis* (Trichomonad), *Giardia lamblia* (Diplomonad) (Figure 2B), *Entamoeba dispar*, *E. bistolytica*, *E. invadens*, *E. moebkowskii*, and *E. terrapina* (Archamoebae; Figure 2D) and *Perkinsus marinus* (Alveolate; Figure 2E) (unpublished data). Most of these organisms are parasites and all display genomes <200 Mb in size. Taking a closer look at unicellular eukaryotic genomes and in particular very small ones with seemingly atypical life history traits (e.g., absence of sex, obligate intracellular parasites, . . .), may be an excellent means to begin to decipher the relative importance of molecular, genetic, environmental, and population level forces in defining TE composition and success.

**Conclusions**

Recent advances in genome sequencing have led to a vast accumulation of TE data and allowed many interesting observations to be made. For example, a remarkable diversity of TEs has been uncovered, some with striking relationships to viruses, revealing a dynamic relationship between TEs and viruses. Consideration of the genome sequencing projects in a phylogenetic context reveals that despite the hundreds of eukaryotic genomes that have been sequenced, a strong bias in sampling exists. There is a general under-representation of unicellular eukaryotes and a dearth of genome projects in many branches of the eukaryotic phylogeny especially, in the supergroup, Rhizaria.
This bias in sampling warrants an embargo on generalizations concerning the distribution and behavior of TEs in eukaryotes. Indeed, 9 eukaryotic genomes are apparently devoid of TEs altogether, upending the long held idea that TEs are ubiquitous in eukaryotic taxa.

The population biology of the genome is reminiscent of an ecosystem where TEs are born, replicate and die and new populations emerge, migrate, and colonize other locations—as well as become extinct. This metaphor is not entirely artificial as chromosomes and even regions of chromosomes can be compared with different ecological niches to which TEs are more or less well adapted. Migration of TEs can occur between chromosomes, as well as between individuals. In addition, there is competition between TEs for genomic resources (host factors) as well as transposition proteins, because the latter are generally encoded by a relatively small fraction of TEs within a given genome. The factors that govern the TE diversity and richness of a genome are complex and are likely to be a combination of properties intrinsic to the TE itself as well as extrinsic to the host. Natural selection is acting on those TEs that are the most fit, where fitness is a measure of the number of copies produced without adversely affecting the fitness of the host and/or the ability to colonize new environments. Understanding the role of these factors in influencing the diversity and differential success of TE in genomes may allow observed patterns in extant genomes to provide a window into the past demographic history of the species in question. Small unicellular eukaryotic genomes are an excellent substrate to study the role of various factors in promoting TE diversity and composition.

Funding

National Institute of Health, National Institute of Allergy, and Infectious Disease (NIH 5R01AI068908-02).

Acknowledgments

I thank Michael Lynch for the invitation to participate in the 2007 American Genetics Association Symposium: Mechanisms of Genome Evolution, and Cedric Feschotte for critical review and comments on the manuscript.

References


Divine SE, Boeke JD. 1996. Integration of the yeast retrotransposon Tyl is targeted to regions upstream of genes transcribed by RNA polymerase III. Genes Dev. 10:626–633.


Corresponding Editor: Michael Lynch