Genome Biology and Evolution Advance Access published August 26, 2016 doi:10.1093/gbe/evw208

Genetic drift, not life history or RNAi, determine long term evolution of transposable elements

Amir Szitenberg[1]*, Soyeon Cha[2], Charles H. Opperman[2], David M. Bird[2], Mark L. Blaxter[3], David H. Lunt[1]

[1]Evolutionary Biology Group, School of Environmental Sciences, University of Hull, Hull, UK
[2]Department of Plant Pathology, North Carolina State University, Raleigh, NC, USA
[3]Institute of Evolutionary biology, School of Biological Sciences, University of Edinburgh, Edinburgh, UK

*Corresponding author: E-mail: szitenberg@gmail.com

Abstract

Transposable elements (TEs) are a major source of genome variation across the branches of life. Although TEs may play an adaptive role in their host's genome, they are more often deleterious, and purifying selection is an important factor controlling their genomic loads. In contrast, life history, mating system, GC content, and RNAi pathways, have been suggested to account for the disparity of TE loads in different species. Previous studies of fungal, plant, and animal genomes have reported conflicting results regarding the direction in which these genomic features drive TE evolution. Many of these studies have had limited power, however, because they studied taxonomically narrow systems, comparing only a limited number of phylogenetically independent contrasts, and did not address long-term effects on TE evolution. Here we test the long-term determinants of TE evolution by comparing 42 nematode genomes spanning over 500 million years of diversification. This analysis includes numerous transitions between life history states, and RNAi pathways, and evaluates if these forces are sufficiently persistent to affect the long-term evolution of TE loads in eukaryotic genomes. Although we demonstrate statistical power to detect selection, we find no evidence that variation in these factors influence genomic TE loads across extended periods of time. In contrast, the effects of genetic drift appear to persist and control TE variation among species. We suggest that

variation in the tested factors are largely inconsequential to the large differences in TE content observed between genomes, and only by these large-scale comparisons can we distinguish long-term and persistent effects from transient or random changes.

Keywords: Nematoda, transposable elements evolution, RNA interference, mating system, parasitism

Introduction

Transposable elements (TEs) are mobile genetic entities found in the genomes of organisms across diverse branches of life, and which are a major source of genetic variation (Kidwell & Lisch 1997; Bennett et al. 2004; Charlesworth et al. 1994). TEs comprise approximately two thirds of the human genome (de Koning et al. 2011), and in other plants and animals may account for up to 85% of all DNA (Schnable et al. 2009; Marracci et al. 1996). In stark contrast, other eukaryotic genomes contain only 1-3% TE-derived sequence within their typically much smaller genomes (Ibarra-Laclette et al. 2013; Burke et al. 2015). The mechanisms that create this variability are not fully understood.

TE insertions are a significant source of deleterious mutation causing gene disruption (Kidwell & Lisch 1997; Biémont et al. 1997), double-strand breaks (Gasior et al. 2006; Hedges & Deininger 2007), ectopic recombination (Charlesworth et al. 1997), gene expression change (Lerman et al. 2003), and other types of mutagenesis (Kidwell & Lisch 2001). In humans, deleterious TE activity contributes to approximately 0.3% of genetic disease (Cordaux & Batzer 2009; Callinan & Batzer 2006). Some TE insertions, however, have only weak deleterious effects, increasing their likelihood of survival and expansion (Kim et al. 1998; Zou et al. 1996; Leem et al. 2008; Gao et al. 2008; Pritham 2009; Hellen & Brookfield 2013). Given sufficient time, a small proportion of these may be co-opted for protein-coding or regulatory functions by

the host genome, and thus become very important components of organismal evolution (Lerman et al. 2003; Kojima & Jurka 2011; Keren et al. 2010). Despite being a key player in organismal evolution, the evolutionary forces determining the TE composition in genomes are far from clear. We have selected the phylum Nematoda, for its phylogenetic diversity of available genomes, as a system in which to investigate TE variation in a phylogenetically-controlled design. While other studies have examined the correspondence between life history or other traits with TE evolution (Hess et al. 2014; Cutter et al. 2008; Fierst et al. 2015; Campos et al. 2012, 2014), these often muster relatively few phylogenetically independent contrasts, and a relatively recent evolutionary time scale. Examining evolutionary events across the entire phylum Nematoda gives a broad perspective where the balance of evolutionary forces will have had time to work.

Substantial efforts to characterise the forces and processes shaping genome evolution have given rise to explanations for the divergence in TE loads among species, including the effects of mating system and recombination, life history, genome GC content, and transposition suppression systems such as RNAi. These factors influence TEs both directly, by affecting their possibility for spread or removal, and indirectly, by modifying the effective population size and probability of fixation (eg. Charlesworth & Charlesworth 1983). The effects of mating system and recombination have been much discussed, with conflicting predictions for either an increase (Wright & Schoen 2000; Montgomery et al. 1987) or decrease (Bestor 1999; Wright & Finnegan 2001; Nordborg 2000; Boutin et al. 2012; Arunkumar et al. 2015) in TE loads in selfing species. Duret, et al. (2000) found that non-recombining genomic regions are less TE rich than recombining regions in *Caenorhabditis elegans*, when considering DNA TEs. Also in *Caenorhabditis,* Cutter, et al. (2008) predicted lower TE loads in selfing compared to outcrossing species. In contrast, TE spread was positively associated with recombination in

Drosophila (Campos et al. 2012, 2014), although this was not recovered in a subsequent study (Bast et al. 2015). A mating system effect on genome size (and thus likely TE load), was reported in plants (Govindaraju & Cullis 1991; Albach & Greilhuber 2004; Wright et al. 2008), but subsequent studies accounting for phylogenetic associations in the data did not recover these effects (Whitney et al. 2010; Ågren et al. 2015; Fierst et al. 2015). Analysis of the evolution of TE loads in the Nematoda, where several independent shifts in mating system have occurred (Figure 1), may aid in better understanding the evolutionary forces and genomic processes operating.

Adoption of a parasitic lifestyle can reduce the effective population size, and thus the effectiveness of recombination and natural selection. Parasites may be subdivided into infrapopulations within hosts, and this population subdivision reduces the effective population size compared to free-living species (Criscione & Blouin 2005). Increased TE counts were found in ectoparasitic *Amanita* fungi compared to free living *Amanita* species (Hess et al. 2014), where the authors suggested the effective population size effects of parasitism as a cause for the difference. As Nematoda contain several independent transitions to parasitism, this hypothesis can also be further tested (Figure 1).

Genome nucleotide bias (GC content) has been shown to influence a wide variety of cellular processes, and especially the rates and patterns of molecular evolution. These effects include tRNA abundance and codon usage (Knight et al. 2001; Ikemura 1981, 1985; Muto & Osawa 1987), mutational patterns (Lobry 1996; Sueoka 1999), gene expression (Gouy & Gautier 1982; Holm 1986; Sharp et al. 1986; Sharp & Devine 1989; Stenico et al. 1994; Andersson & Kurland 1990), protein and RNA structure and composition (Zama 1989; Gambari et al. 1989; D'Onofrio et al. 1991; Huynen et al. 1992; Zama 1996; Collins & Jukes 1993; Gupta et al. 2000), and translational efficiency (Berg & Kurland 1997). The tight integration of TEs with cellular

processes will mean that they will also be affected by differential nucleotide biases, as has been examined by Hellen and Brookfield (Hellen & Brookfield 2013), who demonstrated the accumulation and persistence of human *Alu* elements was favoured in GC-rich regions. Again, diversity in GC content across nematode genomes offers power to detect the effects of GC on TE load evolution.

The host genome is engaged in defending itself against TE insertions, with RNA interference (RNAi) pathways a key cellular processes silencing TEs in eukaryotes (Tabara et al. 1999; Aravin et al. 2001; Sijen & Plasterk 2003; Chung et al. 2008; Czech et al. 2008; Ghildiyal et al. 2008; Slotkin et al. 2009; Kawamura et al. 2008). RNAi pathways variation is thus suggested to be key to TE evolution (Obbard et al. 2009; Rebollo et al. 2012; Matzke et al. 2000; Bossdorf et al. 2008; Richards 2008). In nematodes a variety of mechanisms of TE silencing have been characterised at the molecular level (Aravin et al. 2007; Das et al. 2008; Bagijn et al. 2012; Sarkies et al. 2015), with different pathways operating in different clades (Fig 1). This variation permits examination of the role of alternate TE silencing pathways in explaining genome-wide TE loads.

The importance of non-deterministic processes in shaping TE evolution has been long recognized by population geneticists (Le Rouzic et al. 2007; Charlesworth & Charlesworth 1983; Whitney et al. 2010; Lynch & Conery 2003) with the efficiency of selection and TE silencing likely to be greatly influenced by the effective population size. If differences in TEs between lineages are not determined by processes such as mating system or life history then a null model of genome evolution, one which is shaped by non-deterministic processes such as mutation and drift (Lynch 2007). Here we conduct correlation and ANOVA tests of deterministic forces previously proposed to affect TE evolution, with phylogenetically independent contrasts of TE counts in species from across the phylogenetic diversity of Nematoda (Blaxter et al. 1998)

as the dependant variable. We find no evidence for a deterministic effect of life history, GC content or RNAi pathway variation on TE load variation. Furthermore, our data strongly suggest that stochastic changes are the major genome-wide determinant of TE diversity.

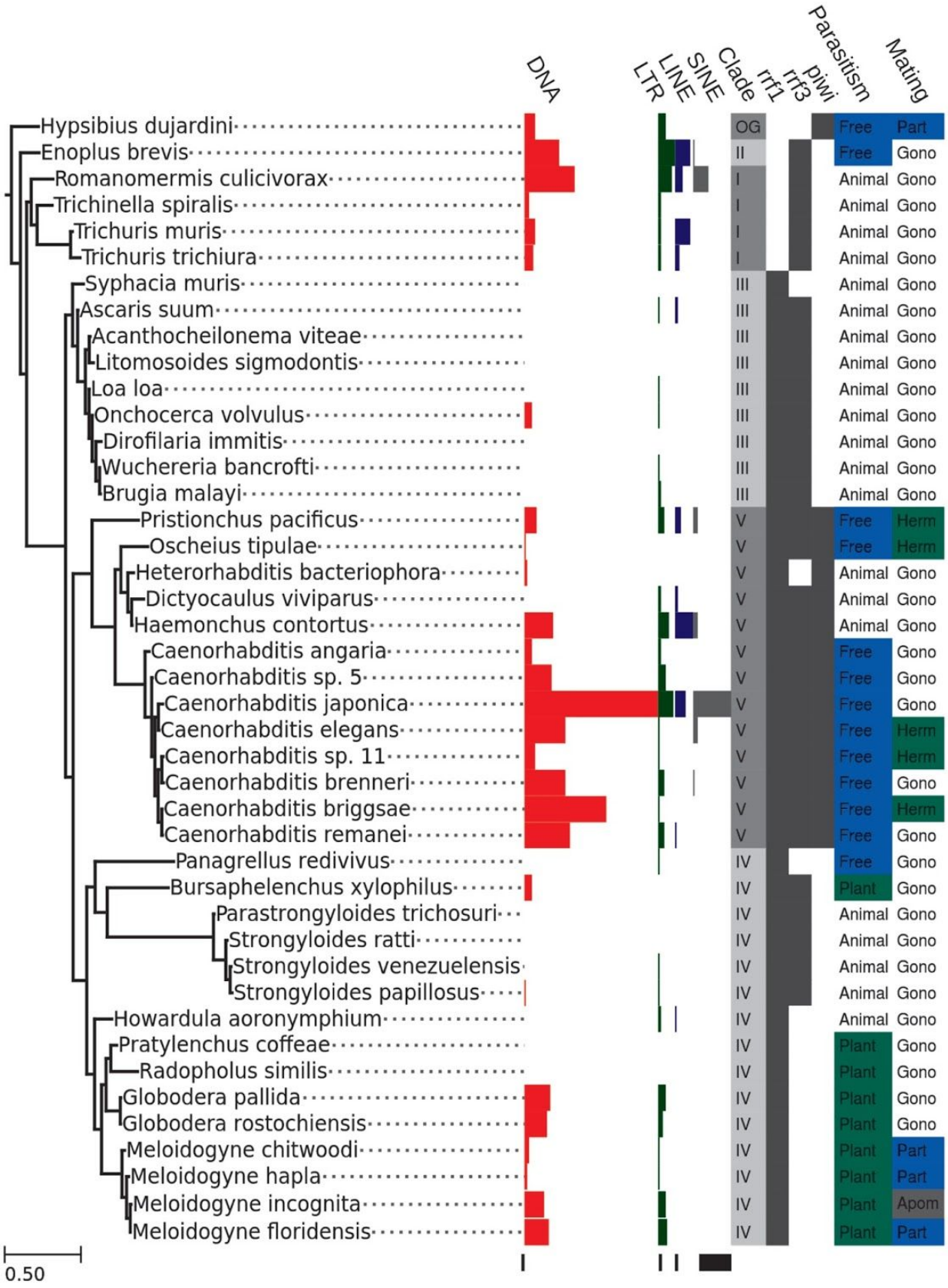

Fig 1. TE loads in Nematoda by class.

SSU-rRNA phylogenetic tree of Nematoda with TE load information by class. The columns represent (left to right) DNA, LTR, LINE and SINE element loads (numerical values are given in S2 Table), the phylogenetic clade *sensu (Blaxter et al. 1998)*, presence or absence of RNAi pathway proteins (RRF1, RRF3 and PIWI), parasitism (animal parasite, plant parasite, or free living), and mating system (parthenogenic, gonochoristic, hermaphroditic or apomictic). Black scales at the bottom of each bar-chart represent 2500 TEs. Sources for life history information are in S1 Table.

Results

TE loads in Nematoda

To test the effect of mating system, parasitic lifestyle, GC content, RNAi and transposition mechanism on TE evolution, TEs were identified and classified in 43 genome assemblies representing the five major nematode lineages and the tardigrade *Hypsibius dujardini* (Figure 1, S1 Table, S1 Methods, sections 1 to 7). Three quantifiers of TE loads, namely TE counts, coverage of the genome assembly by TEs, and the proportion of genome assembly covered by TEs, were strongly correlated with one another ($0.72 < r < 0.9$, p-value < 0.005). None of these measures correlated with genome assembly quality (represented by N50 values), although the TE counts and the total length of TEs did correlate with assembly length, asserting that TE prediction is robust to assembly quality differences (S1 Methods, section 12, S1 Figure). The correlation with assembly length was lost for almost all TE superfamilies under consideration of the phylogenetic relationships among the nematode species (S1 Methods, section 13, S1 Results, section 4). Since the different measures of TE content were shown to be strong proxies of one another, we focused our analyses on TE counts. We expect TE counts to represent TE related evolutionary rates (i.e. rate of change in TE content) more linearly than their assembly

coverage or its proportion of the genome assembly, because of the differences in sequence length among TEs from different TE groups.

High TE loads have a patchy distribution among species in Nematoda, with hotspots observed in the Dorylaimia (Clade I of Blaxter et al. (1998)) and Enoplia (Clade II), in Rhabditina (Clade V), and in the Tylenchomorpha genera *Meloidogyne* and *Globodera* (part of Clade IV) (Figure 1). DNA elements were usually the most abundant, followed by LTR elements, while LINE and SINE elements were quite scarce (Figure 1). When classes were broken down into families (Figure 2, S2 Table, S2 Figure), a large proportion of the variation among species, for 'cut and paste' DNA elements, was contributed by variation in loads of *TcMar* element families, which are scarce in Dorylaimida (Clade I) and abundant in Rhabditina (Clade V). *hAT* families followed a similar pattern, but with less extreme differences among species. *Onchocerca volvulus* (Spirurina; Clade III) had high loads of Helitron elements (5372 copies), and hardly any other TEs, a very different pattern from its relatives in Clade III. Among LTR superfamilies, *Gypsy* elements predominated, with *Copia* and *Pao* elements also prevalent, though a large proportion of the elements were unclassified. The predominant LINE elements were *Penelope* and RTE. SINE elements, although more abundant in a few Rhabditina (Clade V) species than in others, were generally scarce (< 500 in most species, S2 Table). The composition of the consensus TE library is described in S1 Results.

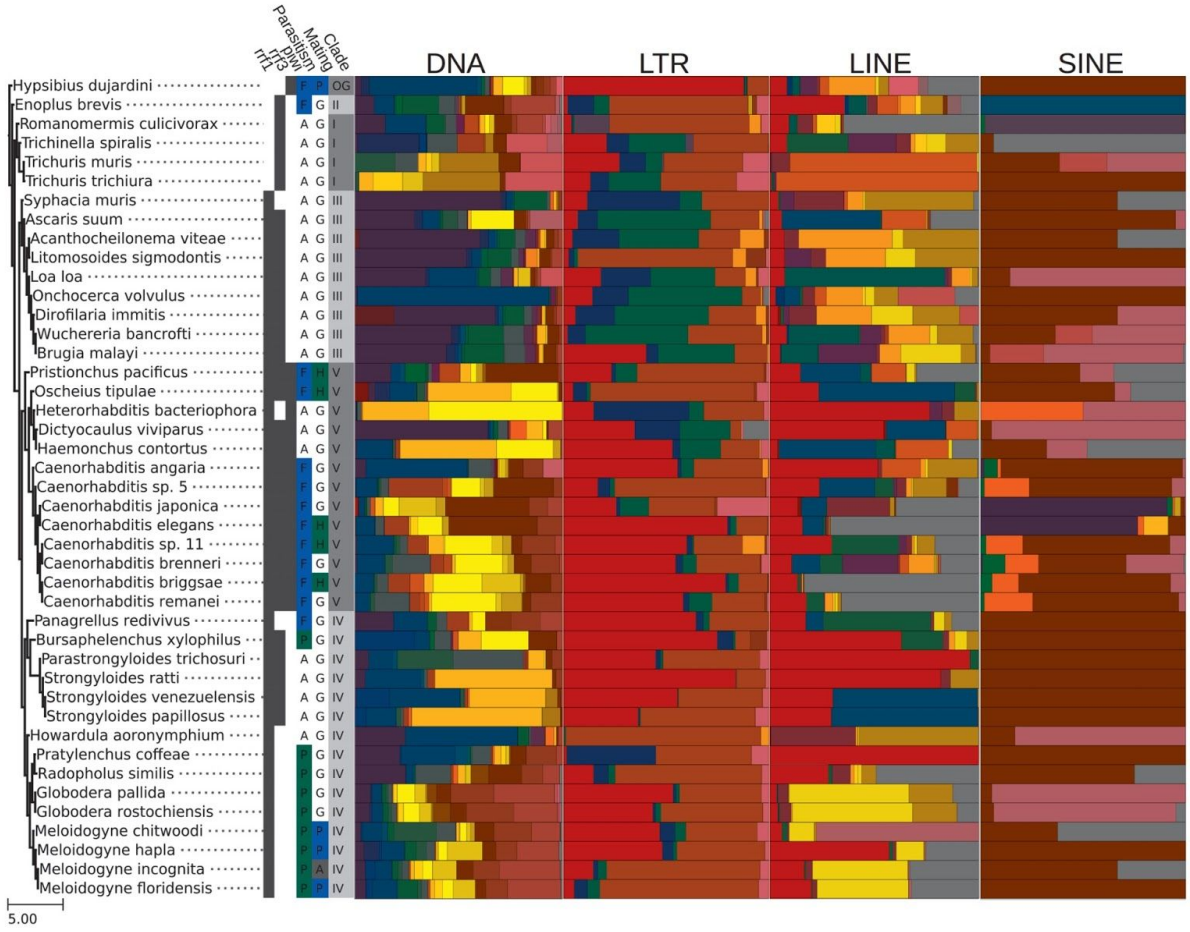

Fig 2. TE loads in Nematoda by superfamily.

SSU-rRNA phylogenetic tree of Nematoda with TE loads information by superfamily. The columns represent (left to right) the presence or absence of RRF1, RRF3 and PIWI RNAi pathway proteins, parasitism (A-animal parasite, P-plant parasite, or F-free living), and mating system (P-parthenogen, G-gonochoric, H-hermaphroditic or A-apomictic), the phylogenetic clade *sensu (Blaxter et al. 1998)* and the proportions of DNA, LTR, LINE and SINE element superfamilies within each of the classes (numerical values in S2 Table, colour key in S1 Fig).

Phylogenetic signal in TE load

According to our null hypothesis, TE loads evolve neutrally and change (in rates and patterns) is expected to be congruent with the topology and branch lengths of the species tree. This can be assessed via phylogenetic transformations of observed TE loads (Pagel 1994). To account for

phylogenetic uncertainty while computing such transformations, we generated a Bayesian posterior distribution of SSU-rRNA phylogenetic species trees. Tree transformation values of the TE counts were computed with each of the trees in the posterior distribution, and for each of the TE classes (DNA, LTR, LINE and SINE; Figure 3A). Transformation value distributions were also computed for each superfamily (S3 Figure), within each class of TEs, and the median value of each superfamily was recorded across the superfamilies in a given class (Figure 3B). We did not include SINE element superfamily medians, since SINE elements were too sparse to compute a meaningful distribution (Figure 3B).

The λ transformation (Pagel 1994) provides an estimate of the degree to which traits are predicted by the phylogenetic relationships, with $\lambda = 1$ indicating a strong fit. At the class level, DNA, LTR and LINE element load variations are strongly correlated with the species phylogenetic relationships ($\lambda > 0.5$; Figure 3A). For many superfamilies the median λ was greater than 0.5, indicating that high fit to the phylogeny is a general characteristic of TEs, and not only a feature of a few large superfamilies (Figure 3B). For SINE elements, in part due to their low abundance and phylogenetic uncertainty, this correlation was not recovered. The strong fit with the phylogenetic tree demonstrates that intraspecific variation in TE loads is not an important source of noise in our results. A second phylogenetic transformation κ, provides an estimate of the correspondence between the branch lengths and the rate of change of a trait (Pagel 1994). $\kappa > 1$ indicates a higher rate of change in longer branches, $\kappa = 1$ indicates that the rate of change of the trait conforms with the general evolutionary rate, and $\kappa < 1$ indicates that the trait is more conserved than expected from neutrality. The κ value distribution for nematode DNA TE loads showed that DNA TE evolution depends less on the organismal evolutionary rate than other TE classes, at the class level ($\kappa < 1$; Figure 3A). The pattern persisted for most superfamilies when considering κ median values at the DNA element superfamily level (Figure

3B). Lastly, the δ transformation estimates the tree depth at which non-neutral evolutionary events occurred, where $\delta < 1$ suggests ancient events and $\delta > 1$ indicates that the trait diversified recently. For DNA elements, δ was greater than 1, indicating that recent events explain their current TE load patterns, while for LTR elements, δ was less than 1, suggesting that ancient events explain their load patterns. δ was not determined for LINE and SINE elements due to phylogenetic uncertainty. Only for LTR elements did these patterns persist when the median δ values of individual superfamilies were considered (Figure 3B), where all of the LTR superfamilies underwent important early events (median $\delta < 0.3$).

The effect of life cycle, RNAi pathway, and genome GC content variation on TE evolution

Primary literature was surveyed in order to determine the mating system of each species and to identify parasites of plants and animals (S1 Table). Key proteins involved in RNA silencing of transposons (RRF1, RRF3 and PIWI) were identified in the genome assembly data using reference sequences (from (Sarkies et al. 2015), S1 Results, section 2, S1 Methods, section 8), and genome assembly N50, span and GC content were calculated (S1 Methods, section 1). The reproductive mode, parasitic status and RNAi pathway for each nematode species is summarized in Figure 1 and S1 Table. The presence and absence of RNAi pathway proteins for the most part conformed with the predictions made by Sarkies et al. (Sarkies et al. 2015), with a few exceptions. *Syphacia muris*, (Oxyuridomorpha, Spirurina in Clade III), lacks the expected RdRP RRF3 protein that is found in other Spirurina species. Since the genome assembly has high N50 values (60,730 bp), and much supporting transcriptome data (S1 Methods, section 8.9), it is highly likely that this species lacks RRF3 (or possess a very divergent RRF3 orthologue). The *Heterorhabditis bacteriophora* (Rhabditomorpha; Clade V) genome lacked an

RRF3 locus although RRF3 is expected in Rhabditomorpha species (Sarkies et al. 2015). Given the relatively high quality of the *H. bacteriophora* genome assembly (N50 of 33,765 bp), RRF3 is again likely absent (or very divergent) in this species. No RRF3 were found in any of the 9 Tylenchomorpha species (Clade IV), regardless of their N50 values (3,348 bp to- 121,687 bp), in keeping with expectations (Sarkies et al. 2015).

For each TE class and superfamily, we tested the effect of mating system, parasitic lifestyle, and variation of RNAi pathways on TE loads at terminal nodes using an ANOVA of phylogenetically independent contrasts (S1 Results, section 5). No significant effect was detected following Holm–Bonferroni correction (Holm 1979). In the absence of any p-value correction for multiple tests, the loads of only two superfamilies were significantly affected by mating system variation, but this is an expected rate of type I error, or an extreme minority of cases if these are true positives..(Holm 1979) We also explored the correlation between genome assembly GC contents and TE loads (S1 Methods section 10.25, S1 Results, section 3) and found no significant results following a Holm–Bonferroni correction (Holm 1979). Prior to this correction, a weak correlation was found in only two superfamilies. On the whole, neither the ANOVA tests or the correlation tests revealed an effect of either of the tested factors. It is unlikely that our results are biased by our taxonomic sampling, as such bias would usually cause false positive results, and such do not occur.

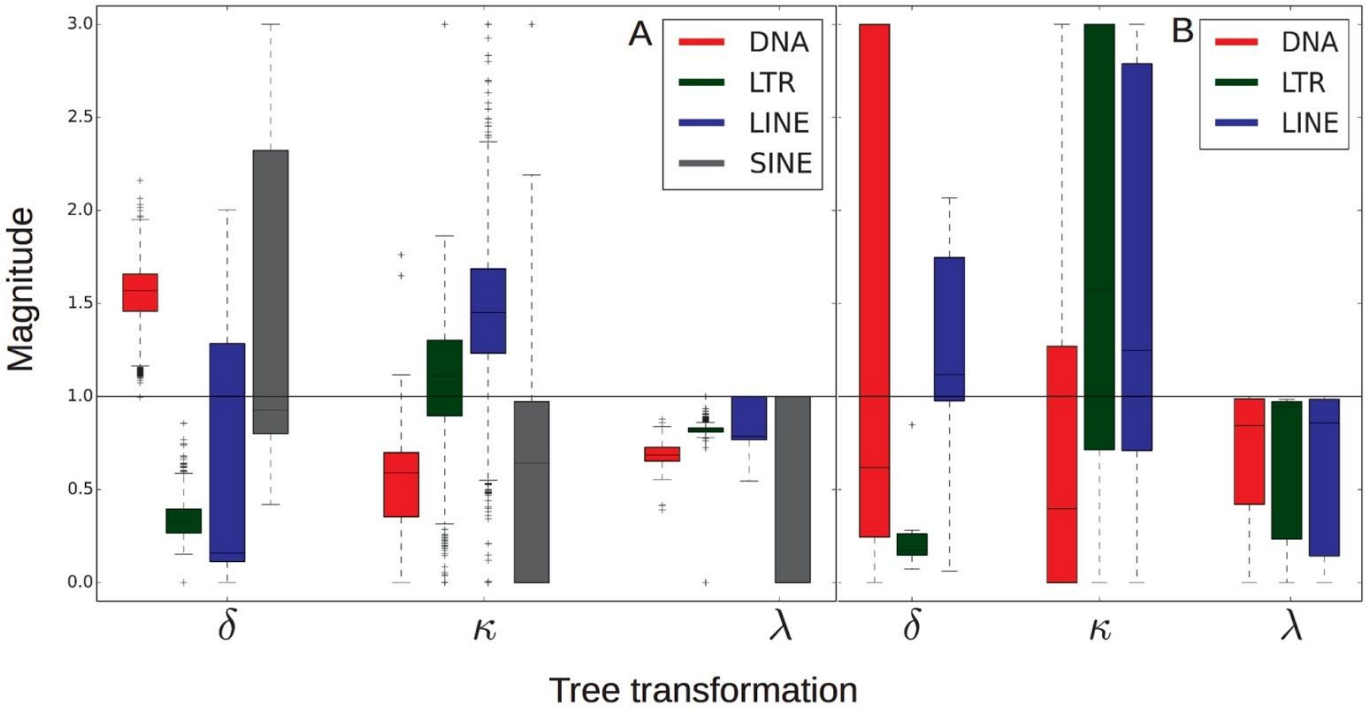

Fig 3. Phylogenetic transformations of TE loads.

The *δ, κ*, and *λ* transformations of TE loads, representing the fit between the TE loads and the tree's topology (*λ*), branch-lengths (*κ*) and root-tip distance (*δ*). (A) The distribution of transformation values across the posterior distribution of most likely phylogenetic trees for each element class (DNA, LTR, LINE and SINE). (B) The distribution of median transformation values of each superfamily of elements within each of the classes. Only superfamilies where the distance between the first and third quartiles was smaller than 0.2 for *λ* and smaller than 0.5 for *κ* and *δ* are included (i.e., superfamilies with an unresolved transformation value are excluded). SINE elements are not shown because the distributions cover the whole range of values. Per-superfamily distribution of the *λ*, *κ* and *δ* transformations across the posterior distribution of trees is shown in S3 Figure.

Changes of TE loads at ancestral nodes

To understand long term processes in TE evolution, we reconstructed the TE loads for each element superfamily at each node in the Nematoda phylogeny, and derived the median change in TE loads at each node compared to its ancestor (Figure 4). For all the four TE classes (DNA, LTR, LINE, and SINE) the evolutionary process was characterized by a trend towards contraction of TE loads, with only very few events of stable expansions, except for shallow nodes where the nature of change was less predictable. Contraction in deep nodes appeared to

have been more constant for LTR elements than other classes (Figure 4B), in agreement with the δ value in this class (Figure 3), and LTR elements were also the most dynamic in shallow nodes, with shallow expansion hotspots within Onchocercidae (Clade III), Strongylida and *Caenorhabditis* (Clade V), and *Strongyloides*, *Globodera* and *Meloidogyne* (CladeIV). Other classes (Figure 4A, C and D), also showed recent expansions, but only in a subset of these taxa.

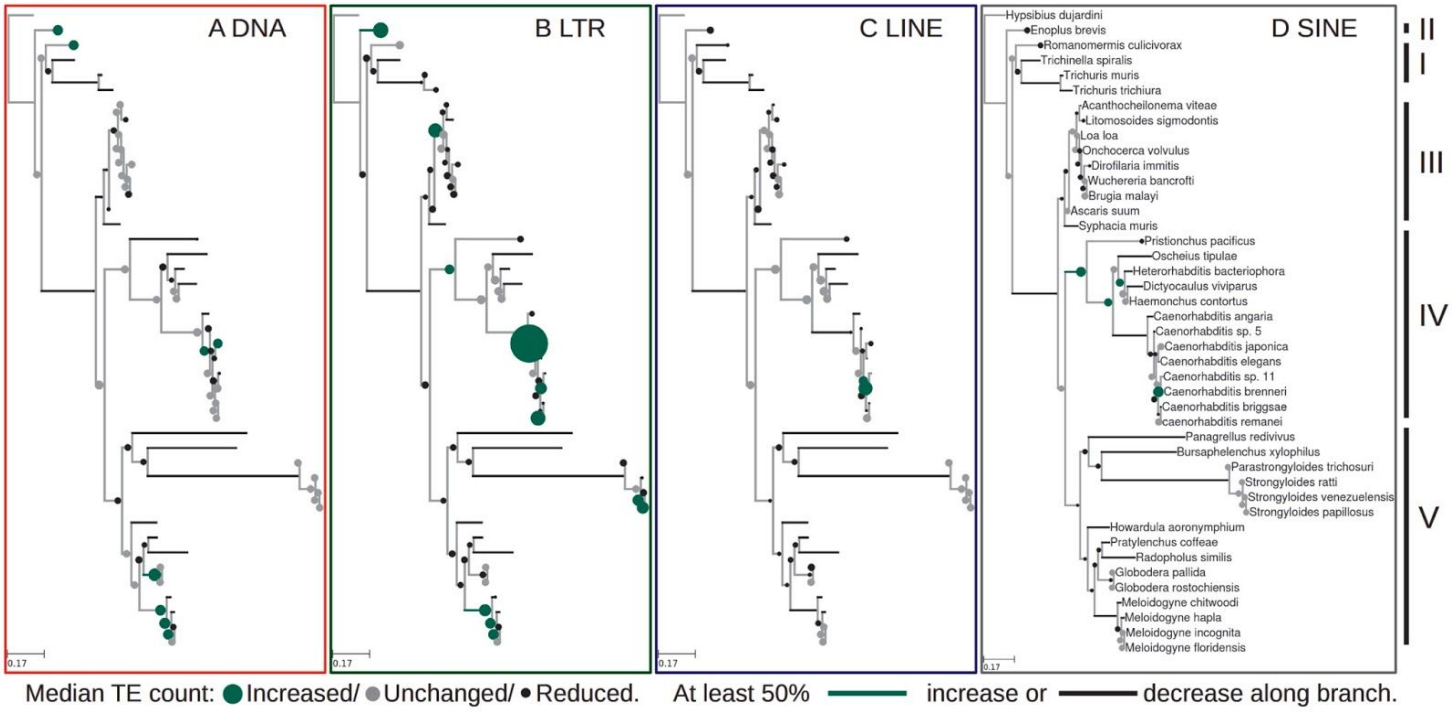

Fig 4 Median TE load change at ancestral and terminal nodes

The median load change of DNA (A), LTR (B), LINE (C), and SINE (D) superfamilies. Ancestral states were reconstructed for each superfamily. Then, the proportion of change, compared to the ancestral node, was computed for each superfamily, at each node. The median change proportions are presented for each class. Green nodes represent an increase compared to the most recent ancestor, with larger nodes representing a greater increase. Black nodes represent a decrease compared to the most recent ancestor, with smaller nodes representing a greater decrease. Where no bullet is visible, there has been a large decrease in TE counts. Long branches (0.06 or longer) along which at least 50% change in TE loads has occurred are green, gray or black to indicate an increase, stability or decrease of TE median loads along the branch. Since increase is not inferred, green branches do not ultimately occur.

Detection of adaptive processes and convergent evolution

To identify adaptive processes in the evolution of TEs, we tested the fit of the TE loads with the Ornstein Uhlenbeck (OU) model, using BayesTraits 2 (Pagel 1997). Since a strong phylogenetic background provides power to detect selection as a consistent deviation from it, and given the high λ values characterising TE evolution in Nematoda, we predicted high power to detect α, the selection strength parameter in the OU process, as illustrated in Figure 5A. This figure demonstrates that given the Nematoda phylogenetic tree and the phylogenetic pattern of the TE loads, low α values can be detected. Assuming no stochastic interference, α values significantly greater than 1 were detected in fourteen, seven, three, and one superfamilies from the classes of DNA, LINE, LTR and SINE elements respectively (S4 Figure, S1 Methods, section 10.18). These families were analysed for convergent evolution, fitting the most likely extended model, also allowing shifts in the selective optimum (θ) as well as stochastic change (σ^2). Since transposition can increase the TE loads even when it is not adaptive, α and θ may also represent neutral or slightly deleterious transposition events rather than positive selection, although the balance between α and σ^2 can help to distinguish between stochastic and deterministic trajectories, regardless of the true nature of α. Convergent evolution (i.e., polyphyletic lineages possessing the same selective optimum θ) was detected for all these elements. However, shifts in θ were only identified in terminal or otherwise shallow nodes, with the exception of a θ increase for two LTR element superfamilies at the base of the Rhabditina (Clade V), and never coincided with shifts in mating system, parasitism, or RNAi pathways (S5 Figure). Moreover, the σ^2 (drift) values were overwhelmingly higher than α (selection) for most of the superfamilies, as the example shown in Figure 5B, illustrating the stochasticity of the evolutionary trajectories of TE loads. Therefore, this analysis reveals stochastic evolutionary trajectories with no deterministic effect of the tested factors.

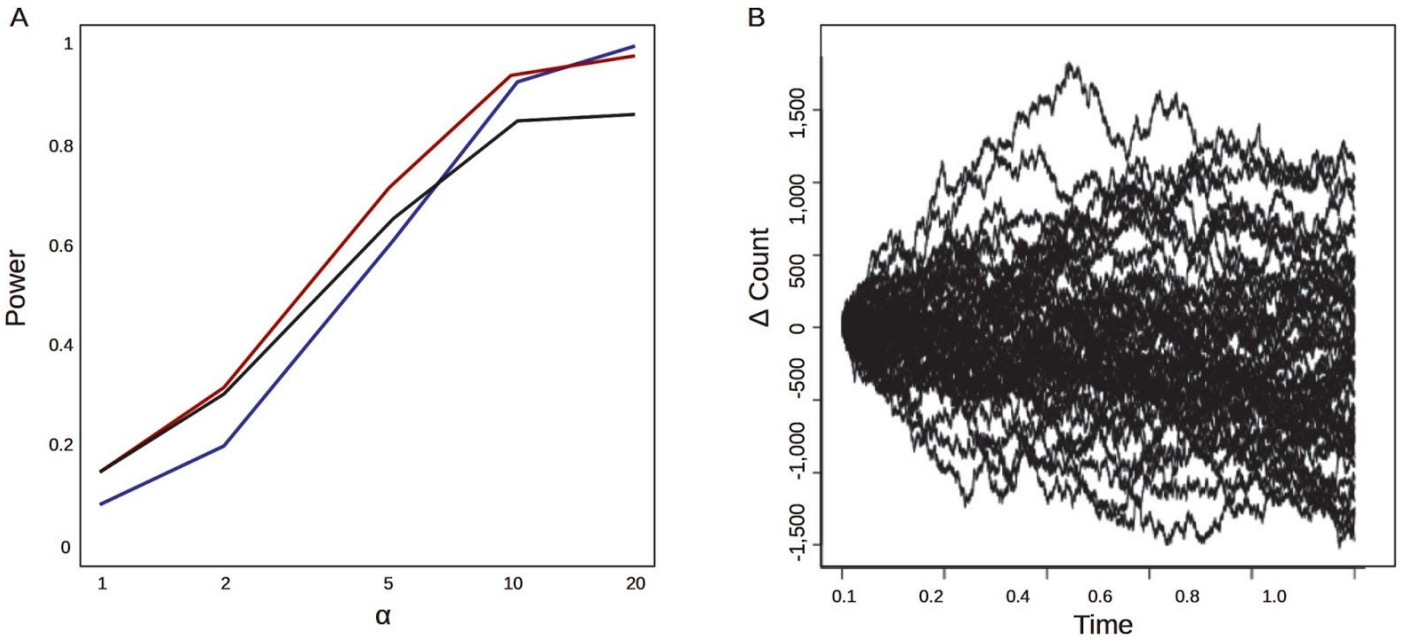

Fig 5. OU-model fitting to detect selection.

Power to detect the selection strength parameter alpha (A), under a gamma transformation value of 0.5 (black), 0.8 (red) and 1 (blue), and simulations of the evolutionary trajectory of the DNA/TcMar-Tc2 TE superfamily loads (B) under the OU model fitted to this superfamily (σ^2= 1*10^9, α=4*10^5 for 10^5 generations and 50 replications).

Discussion

The common ancestor of Nematoda dates back to the Cambrian radiation (Vanfleteren et al. 1994), 550 million years ago, and thus the genome sequences of nematodes that have become available in the last decade provide a unique opportunity for comparative genomics analyses of the long term forces shaping evolution. This contrasts with most previous studies which were only able to analyse recent periods (e.g., Duret et al. 2000; Cutter et al. 2008; Fierst et al. 2015; Campos et al. 2012, 2014; Hess et al. 2014; Albach & Greilhuber 2004; Wright et al. 2001; Ågren et al. 2015, 2014; but see Whitney et al. 2010). We present analyses of the long term evolution of TEs, exploring the roles and importances of multiple deterministic forces in a phylogenetic design. Our results establish that diversification in TE loads is recent and

independent of GC content, life history, and RNAi, and is best understood as a stochastic process. We also find a consistent reduction in TE loads at ancestral nodes across the Nematoda tree, most notably at the base of clade III+IV+V. It would thus seem that the consequence of genetic drift and purifying selection, known to shape TE loads at the population level (Hua-Van et al. 2011; Lynch & Conery 2003), endure to shape TE load variation in the phylum level, with little effect of other potentially deterministic forces. A caveat is that the constant reduction in TE loads in ancestral nodes may reflect the long term reduction in genome size as a whole, rather than a specific reduction in TE load.

Long term GC content variation does not determine TE loads

Genome GC content can change gradually along the phylogenetic tree. We therefore used an analytical procedure that accounts for the ancestral character states of both TE load and total GC content traits and tested for correlation between the two through the evolutionary history of nematodes. In humans, purifying selection against TE loss in gene rich regions of the genome is the main driver of variability in *Alu* element loads between GC-rich and -poor genomic regions (Brookfield 2001; Hellen & Brookfield 2013). However, although local GC content variation in the genome may explain the distribution of TEs inside that genome, the total GC content of the genome does not predict TE load differences between species as we did not find substantial correlation between the TE loads and the total GC content of the nematode genome assemblies. While the local GC content may indeed influence the number of insertions fixed in a given locus, it is not a limiting factor on TE loads in the genome as a whole.

Recent variation in RNAi pathways and life history is not a predictor of TE evolution

TE load variation is independent of recent shifts in the species' life history or RNAi pathways involved in TE silencing. Less than one percent of the ANOVA tests examining the effect of RNAi, parasitism, and mating system on TE loads suggested significant associations between traits and TE loads, and did not exceed an acceptable type I error rate. Since we cannot determine historic character states for RNAi pathways and life history, it is impossible to rule out completely that they would explain TE loads, but this limitation is shared by studies that do suggest a significant effect of these factors. In addition, the OU models fitted to the data were stochastic, and not directional. Even though a selection component is directly assumed in this model, and therefore always found, it was never strong enough to counteract stochasticity in our simulations. Deterministic models of TE load evolution thus have little or no support, and instead, the variation of TE loads among the extant species is consistent with a stochastic model. Exclusion of this wide range of direct possible deterministic explanations for TE load variation means that complex interactions among such forces must be postulated to retain strong effect for these proposed mechanisms. This is not to say that such deterministic effects are absent, only that they are short lived due to the genetic drift that often counteracts them. That we find genetic drift to be a dominant evolutionary force for TEs is not unexpected as, drift has been suggested to play a key role in the evolution of multicellular organisms due to their long-term and ancient reduction in effective population size (Lynch & Conery 2003).

TE load contraction has prevailed in TE evolution

Ancestral state reconstructions reveals a consistent contraction in TE loads through time on our tree. Furthermore, long branches in the phylogeny, including terminal ones, often coincide with

reduction of at least 50% in TE loads (Figure 4) and almost never with an increase. We thus suggest that a component of long term purifying selection, acting at the population level (Hua-Van et al. 2011), in addition to the more recent effect of drift discussed above, should be included in a realistic model at this scale. In such a model purifying selection prevails in the long run, over genetic drift that might increase or preserve TE loads temporarily. If this is true, the co-occurrence of increased purifying selection of LTR elements with their increased expansions in terminal nodes, suggests that, on average, LTR element loads have a tendency to increase faster than other elements and are therefore more deleterious and exposed to stronger purifying selection, in accordance with previous predictions (Brookfield 1995; Kidwell & Lisch 2001).

An increased strength of purifying selection that might be experienced by LTR elements could result from either their possible indiscriminate targeting of genic regions (Pritham 2009; Finnegan 1992) or from their suggested role in induction of increased ectopic recombination (Montgomery et al. 1987). It may be that LTR elements have not been able to evolve to efficiently target non-genic regions of the genome (McDonald et al. 1997; Zuker et al. 1984). One signature of increased ectopic recombination as a driver of purifying selection on LTR TEs would be an inverse correlation between the median sequence length of TE families and their loads in a given species (Petrov et al. 2003), but we did not detect such inverse correlation (S1 Methods, section 11). Still, additional sampling is required to pin down the cause of TE contraction in ancestral node.

Conclusions

A wide body of literature has sought biological explanations for the observed patterns of variation of TE loads in eukaryotic genomes, invoking explanatory variables such purifying selection, mating system, parasitic lifestyle, genome-wide GC content and RNAi pathways for silencing TEs. Our analysis of the evolution of TEs on a long time scale – across the entire

phylum Nematoda – shows high statistical power to detect directional selection, yet reveals that these variables do not, in fact, explain TE load variation among species, with the possible exception of purifying selection, given time. Instead, variation in TE loads is largely stochastic, explained by genetic drift, with little or no consistent effect of life history or genomic explanatory variables. We acknowledge that other characteristics, such as horizontal gene transfer, or recurrent activation and deactivation of TEs might also be stochastic. However, the strong congruence of the TE counts with the phylogenetic tree suggest that the variability in TE loads within species is smaller than between species, and thus the observed counts are close to fixation by drift. We also emphasize here that our results do not reject the importance of these or other factors, for an individual or a population, over relatively short time scales. Our inference is that in the long run they will not determine the evolutionary trajectories of TE loads, due to strong stochastic effects, and ultimately purifying selection. We also stress that although genetic drift and selection are processes that occur in populations, their signature can additionally be observed in speciation and over phylogenetic scales. We suggest that only studies that examine TE load across a large number of life history transitions and over large timescales will be able to provide power to reliably distinguish between stochastic and deterministic forces, and quantify the balance of evolutionary processes shaping this major component of eukaryotic genomes.

Methods

Genome assemblies

Genome assemblies of species from phylum Nematoda, representing the five major clades (Blaxter et al. 1998) were obtained from different sources (S3 Table). The assemblies included four species from Dorylaimida (Clade I), one from Enoplia (Clade II), nine from Spirurina (Clade III), fifteen from Tylenchina (Clade IV) and thirteen from Rhabditina (Clade V). For Dorylaimia

and Enoplida (Clades I and II) we analysed all the available genome assemblies. The genome of the tardigrade *Hypsibius dujardini* was used as outgroup. To compare the completeness of the genome assemblies, the N50 metric (S1 Methods, section 1) was calculated for each (S3 Table). The GC content of each genome assembly was calculated (S1 Methods, section 1).

TE identification

We conducted TE searches in the genome assemblies rather than in sequence read data, which are not publicly available for many of the target species. To mitigate the biases associated with this approach, we have also utilized complementary methods of TE searches. One of the approaches was homology based searches using reference DNA sequences of elements in a *de-novo* constructed library, representing a wide taxonomic range within phylum Nematoda. RepeatModeler 1.0.4 (Smit & Hubley 2010b) was used to identify repeat sequences in each genome assembly using RECON (Bao & Eddy 2002), RepeatScout (Price et al. 2005) and TRF (Benson 1999). RepeatModeler uses RepeatMasker (Smit & Hubley 2010a) to classify the consensus sequences of the recovered repetitive sequence clusters. The identification stage employed RMBlast (Camacho et al. 2009) and the Eukaryota TE library from Repbase Update (Jurka et al. 2005). The consensus sequences from all the species were pooled, and the uclust algorithm in USEARCH (Edgar 2010) was used to make a nonredundant library, picking one representative sequence for each 80% identical cluster. Additional classification of the consensus sequences was performed with the online version of Censor (Jurka et al. 1996). Classifications supported by matches with a score value larger than 300 and 80% identity were retained. The script used to construct this library is in S1 Methods, sections 2-3.

RepeatMasker (Smit & Hubley 2010a) was used to search for repeat sequences in the Nematoda genome assemblies and that of *H. dujardini*, using this *de-novo* Nematoda library (S1 Methods, section 4). To eliminate redundancies in RepeatMasker output, we used One Code to

Find Them All (Bailly-Bechet et al. 2014), which assembled overlapping matches with similar classifications, and retained only the highest scoring match of any remaining group of overlapping matches (S1 Methods, section 5). Alternative approaches to identify TEs were also employed. TransposonPSI B (http://transposonpsi.sourceforge.net/), which searches for protein sequence matches in a protein database thus allowing acurate identification of shorter fragments, and LTRharvest (Ellinghaus et al. 2008), which identifies secondary structures (S1 Methods, section 6), were used to screen the target genomes. . For TransposonPSI searches, only chains with a combined score larger than 80 were retained, while we retained only matches that were at least 2000 bp long and 80% similar to the query from LTRharvest searches. Where matches from the three approaches overlapped, we retained only the longest match (S1 Methods, section 7).

Characterization of RNAi pathways

Three key proteins, distinguishing the three RNAi pathways discussed in Sarkies et al. (2015), were searched for in the genome assemblies (S1 Methods, section 8.1), using the program Exonerate (Slater & Birney 2005). Sequences from the supplementary files Data S1 and S2 from Sarkies et al. (2015) were used as queries to identify homologues of PIWI, an Argonaute (AGO) subtype, and RNA-dependent RNA polymerase (RdRP ; specifically subtypes RRF1 and RRF3) respectively. Only matches at least 100 amino acids (aa) long and at least 60% similar to the query were retained. In addition, only the best scoring out of several overlapping matches was used (S1 Methods, section 8.2). The matches and the queries were used to build two phylogenetic trees, one of PIWI (and other AGO) sequences and the other of RdRP sequences, to verify the identity of the matches (S1 Methods, section 8.3). Each of the datasets was aligned using the L-ins-i algorithm in MAFFT 7 (Katoh & Standley 2013), and cleared of positions with a missing data proportion of over 0.3 using trimAl 1 (Capella-Gutiérrez et al. 2009). In the

resulting alignment, only sequences longer than 60 aa were retained. Maximum Likelihood (ML) trees were reconstructed using RAxML 8 (Stamatakis 2014) with sh-like branch supports. Species that occurred at least once in any of the three clades (PIWI in the first tree and RRF1 and RRF3 in the second), were scored as possessing that gene (Figure 1). Where a species did not have a representative sequence in one of the clades, a directed search for the specific protein was conducted in the sequences that did not pass the filter (i.e., the best match had lower score and length than the set cutoff). The identity of sequences retrieved in this way was examined in a second pass of phylogenetic reconstruction. This step did not yield additional phylogenetically validated matches and confirmed the validity of the cutoff set in the filtering step.

Phylogenetic reconstruction of the Nematoda using small subunit ribosomal RNA (SSU-rRNA) sequences

To control for phylogenetic relationships within the TE counts dataset we inferred a species phylogenetic tree using the SSU-rRNA gene. This locus is considered to be reliable for the reconstruction of the phylogeny of Nematoda, and produces trees that tend to agree with previous analyses (Meldal et al. 2007; Holterman et al. 2006; Blaxter et al. 1998; van Megen et al. 2009). First, we identified SSU-rRNA genes with BLAST+ 2.2.28 (Camacho et al. 2009), in each of the genome assemblies, where for each species the query was an SSU-rRNA sequence of the same or closely related species, taken from the Silva 122 database (Quast et al. 2013). Matches shorter than 1,400 bp were not selected and the query sequence was retained instead, providing it was identical to the match. Species for which the SSU-rRNA sequence could not be recovered and was not available online were excluded from further analysis. Since unbalanced taxon sampling may reduce the accuracy of the phylogenetic

reconstruction (Heath et al. 2008), we also included additional sequences from Silva (Quast et al. 2013), representing the diversity of Nematoda. ReproPhylo 0.1 (Szitenberg et al. 2015) was used to ensure the reproducibility of the phylogenetic workflow (S2 Results). A secondary structure aware sequence alignment was conducted using SINA 1.2 (Pruesse et al. 2012), and the alignment was then trimmed with trimAl 1 (Capella-Gutiérrez et al. 2009) to exclude positions with missing data levels that lie above a heuristically determined cutoff. An ML search was conducted with RAxML 8 (Stamatakis 2014) under the GTR-GAMMA model and starting with 50 randomized maximum parsimony trees. Branch support values were calculated from 100 thorough bootstrap tree replications. After tree reconstruction, nodes that did not represent a genome assembly (either the blast match or the Silva sequence substitute) were removed from the tree programmatically using ETE2 (Huerta-Cepas et al. 2010). To characterize the phylogenetic uncertainty, we generated a posterior distribution of trees using Phylobayes 3 (Lartillot et al. 2009). Two chains were computed, using the trimmed ML tree as a starting tree and the GTR - CAT model (*sensu* (Lartillot & Philippe 2004)). The analysis was continued until the termination criteria were met (specifically, maxdiff and rel_diff < 0.1, and effsize > 100), with a burnin fraction of 0.2 and by sampling each 10th tree. The same subsample of trees was used to generate a consensus tree. The reconstruction of the SSU rRNA tree is detailed in S1 Methods, section 1.

The effect of life cycle, RNAi and percent GC variation on TE loads

Primary literature was surveyed to determine the mating system of each species and to identify parasites of plants and animals (S1 Table). The effect of these factors on the TE loads was tested with an ANOVA of phylogenetically independent contrasts, using the R package Phytools (Revell et al. 2008). Species were classified into the four mating systems dioecy, androdioecy, facultative parthenogenesis (including both species that fuse sister gametes and species that

duplicate the genome in the gametes) and strict apomixis. Species that had both hermaphroditic and gonochoric life cycle stages were classified as gonochoric (e.g. *Heterorhabditis bacteriophora*, (Poinar 1975)). We conducted three tests, in the first of which the four levels were tested, in the second the parthenogenetic and androdioecious species were pooled, and in the third, species were divided into dioecious and non-dioecious.

To test the effect of parasitism, free living species, plant parasites and animal parasites were first tested as three separate groups, and then plant and animal parasites were pooled into a single group for a second test. The necromenic lifestyle of *Pristionchus pacificus* was classified as free living because this species is not reported to depend on any host function, only on the organisms that build up on its carcass (Dieterich et al. 2008).

ANOVA of phylogenetically independent contrasts was also used to test the effect of the variation in RNAi pathways on the TE loads. Six groups of species were determined based on the presence or absence of PIWI, RRF1 and RRF3 proteins. In addition, for each of the three proteins, the effect of their presence was tested independently of the other proteins. Finally, dependency between GC content of genome assemblies and their TE loads was tested by a regression of the squared contrasts of TE counts and the estimates of GC contents in ancestral nodes (Revell et al. 2008). The execution of ANOVA and correlation tests is detailed in S1 Methods sections 10.23-10.25.

Phylogenetic signal in the TE data

The phylogenetic transformations λ, κ and δ (Pagel 1994) were calculated with BayesTraits (Pagel 1997) over the subsample of trees produced with Phylobayes (see above), to account for phylogenetic uncertainty (S1 Methods, sections 10.4-10.10). They were estimated for the pooled classes of DNA, LTR, LINE and SINE elements, as well as for individual superfamilies

that occurred in at least 15 nematode species. The proportion of individual TEs that were included in this analysis is depicted in S2 Figure (bottom).

Detection of selection and convergent evolution of TE loads

The Ornstein Uhlenbeck (OU) process (Gardiner 1985) was originally suggested as an approach to model the evolution of continuous traits based on phylogenies (Felsenstein 1985). Building upon this process, Hansen (1997) has developed a method to study changes in selection regimes, on the macroevolutionary scale, neglecting stochastic effects on the process. In the OU process, a change in character state depends on the strength of selection (α) and its distance and direction from the current selection optimum (θ). Goodwin later (Goodwin et al. 2003) added a Brownian Motion (BM) component to the model (σ^2), recognizing the confounding effect that stochastic events related to demography might have on selection. The R package PMC (Boettiger et al. 2012) was used to assess the power of our data to detect OU processes in the evolution of TEs. The OU parameter α was estimated with Bayestraits (Pagel 1997), neglecting stochastic effects, in superfamilies occurring in at least 15 nematode species. Where a significant α was detected (p-value < 0.05, in the posterior distribution of trees), indicating selection, we examined the possibility of convergent evolution between species with similar life cycle or RNAi status with the R package SURFACE (Ingram & Mahler 2013). In these analyses, selection optima shifts are detected in the trees' branches through a heuristic search, which uses AIC test results as the optimization criterion. Then, further improvement to the fit of the model is attempted by unifying optimum shifts. Where the unification of two or more optimum shifts improved the AIC score of the model, convergent evolution is inferred. SURFACE uses the R package OUCH (Butler & King 2004) to fit OU models , and unlike Bayestraits, includes a stochastic component (σ^2), expressed by BM, in the OU model. The steps described in this paragraph are detailed in S1 Methods section 10.12-10.22.

Magnitude of change at ancestral nodes

To identify nodes in the species tree that were hotspots of change in TE loads, we reconstructed the ancestral character states for a subset of elements using an ML analysis (Revell et al. 2008). Since the phylogenetically independent contrast of a root node is also the maximum likelihood estimate of its character state (Felsenstein 1985), this analysis sequentially treats each node as the root, in order to compute the TE load at this node. The element subset included only classified elements from superfamilies that occurred in at least 15 species. Within each of the three groups of "cut and paste", LTR, LINE, and SINE elements, we calculated the median change magnitude across the superfamilies in the group, and for each node. The magnitude of change was expressed as the proportion of the load of a given element superfamily at node X out of the load of the same superfamily at the parent of node X. The steps described here are detailed in S1 Methods section 10.23, 10.26-10.27.

Supplementary Material

S1 Table. Life history information. Code: Species name abbreviation, corresponds with species names in Figure 1. Species**:** corresponds with species names in Figure 1. Clade: *sensu* Blaxer et. al (Blaxter et al. 1998). Mating: The mating system. M4: Mating system information with four classes (gonochoristic, androdioecious, facultative parthenogen, apomict). M3: Mating system information with three classes (gonochoristic, facultative sexual, apomict). M2: Mating system information with two classes (gonochoristic, non-gonochoristic). Parasitism: the parasitic or free living life-style of the nematode. P3: Parasitic lifestyle information with three classes (plant parasite, animal parasite, free living). P2: Parasitic lifestyle information with two classes (parasitic, free living).

S2 Table. Transposable element counts in Nematoda genome assemblies. Species

abbreviations correspond with the species names in Figure 1.

S3 Table. Genome assemblies used in this study. Code: Species name abbreviation, corresponds with species names in Figure 1. Genus**:** corresponds with genera names in Figure 1. Species**:** corresponds with species names in Figure 1. Clade: *sensu* Blaxer et. al (Blaxter et al. 1998). N50 length: the length of the longest contig among the shortest contigs that amount to half the assembly length. See S1 Methods, section 1, for detailed computation. Publisher: the assembly's publishing institute. Version/ Accession: The Assembly's version or genbank accession numbers.

S1 Fig. Correlation between TE loads and genome assembly statistics. The correlation between TE loads and the N50 length, the genome assembly length and the genome size.

S2 Fig. Transposable element loads in Nematoda genomes. Species name abbreviation, corresponds with species names in Fig 1.

S3 Fig. Phylogenetic transformations of Nematoda transposable element counts. Transposable elements superfamilies are sorted by abundance, with most abundant on the right. The transformation values from top to bottom are λ, κ and δ. They are explained in the Results section "Phylogenetic signal in TE load" and their computation is demonstrated in S1 Methods, sections 1

.4-10.10.

S4 Fig. Distribution of alpha in a deterministic OU model. Transposable elements superfamilies are sorted by abundance, with most abundant on the right. The α parameter is explained in the Results section "Detection of adaptive processes and convergent evolution" and its computation is demonstrated in S1 Methods, section 10.18.

S5 Fig. Median change in selective optima. Black nodes represent close to stable median selective optimum. Blue and red nodes represent an increase or decrease in the median

selective optimum, respectively. The size of the node represents the magnitude and direction of the change in the median selective optimum. The integer at each node denotes the number of TE superfamilies for which the selective optimum shifts at this node. The OU model is explained in the Results section "Detection of adaptive processes and convergent evolution" and its computation is demonstrated in S1 Methods, section 10.18.

S1 Methods. The code for the analyses carried out in the study. Static Jupyter notebooks containing the code used to carry out the analyses in this manuscript. https://dx.doi.org/10.6084/m9.figshare.2056101.v3. Also available on github along with all the intermediate and output files DOI:10.5281/zenodo.55462, https://github.com/HullUni-bioinformatics/Nematoda-TE-Evolution

S1 Results. 1)Taxonomic composition of the TE consensus library. Species codes correspond with those appearing in S1 Table. 2) RNAi pathway protein phylogenetic trees. 3-5) Results of Anova and correlation at ancestral nodes.

S2 Results. Phylogenetic analysis report. A ReproPhylo generated report describing the data, methods and results of the phylogenetic analysis used to produce Fig 1. https://dx.doi.org/10.6084/m9.figshare.2056107.v3

Acknowledgement

We thank Dr. Beth Hellen for her valuable comments, and Dr. Peter Sarkies, Dr. Arvid Ågren and Prof. Carl Boettiger for assistance with the analysis. The following funding sources supported this study: The Science of the Environment Council grant (http://www.nerc.ac.uk/) NE/J011355/1 was awarded to DHL and MLB. The Science of the Environment Council grant (http://www.nerc.ac.uk/) R8/H10/56 was awarded to GenPool, University of Edinburgh. The Medical Research Council grant (http://www.mrc.ac.uk/) G0900740 was awarded to GenPool,

University of Edinburgh. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Literature cited

Ågren JA et al. 2014. Mating system shifts and transposable element evolution in the plant genus *Capsella*. BMC Genomics. 15:602.

Ågren JA, Greiner S, Johnson MTJ, Wright SI. 2015. No evidence that sex and transposable elements drive genome size variation in evening primroses. Evolution. 69:1053–1062.

Albach DC, Greilhuber J. 2004. Genome size variation and evolution in *Veronica*. Ann. Bot. 94:897–911.

Andersson SG, Kurland CG. 1990. Codon preferences in free-living microorganisms. Microbiol. Rev. 54:198–210.

Aravin AA et al. 2001. Double-stranded RNA-mediated silencing of genomic tandem repeats and transposable elements in the *D. melanogaster* germline. Curr. Biol. 11:1017–1027.

Aravin AA, Hannon GJ, Brennecke J. 2007. The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. Science. 318:761–764.

Arunkumar R, Ness RW, Wright SI, Barrett SCH. 2015. The evolution of selfing is accompanied by reduced efficacy of selection and purging of deleterious mutations. Genetics. 199:817–829.

Bagijn MP et al. 2012. Function, targets, and evolution of *Caenorhabditis elegans* piRNAs. Science. 337:574–578.

Bailly-Bechet M, Haudry A, Lerat E. 2014. 'One code to find them all': a perl tool to conveniently parse RepeatMasker output files. Mob. DNA. 5:13.

Bao Z, Eddy SR. 2002. Automated *de novo* identification of repeat sequence families in sequenced genomes. Genome Res. 12:1269–1276.

Bast J et al. 2015. No accumulation of transposable elements in asexual arthropods. Mol. Biol. Evol. msv261.

Bennett EA, Coleman LE, Tsui C, Pittard WS, Devine SE. 2004. Natural genetic variation caused by transposable elements in humans. Genetics. 168:933–951.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27:573–580.

Berg OG, Kurland CG. 1997. Growth rate-optimised tRNA abundance and codon usage. J. Mol. Biol. 270:544–550.

Bestor TH. 1999. Sex brings transposons and genomes into conflict. Genetica. 107:289–295.

Biémont C, Tsitrone A, Vieira C, Hoogland C. 1997. Transposable element distribution in *Drosophila*. Genetics. 147:1997–1999.

Blaxter ML et al. 1998. A molecular evolutionary framework for the phylum Nematoda. Nature. 392:71–75.

Boettiger C, Coop G, Ralph P. 2012. Is your phylogeny informative? Measuring the power of comparative methods. Evolution. 66:2240–2251.

Bossdorf O, Richards CL, Pigliucci M. 2008. Epigenetics for ecologists. Ecol. Lett. 11:106–115.

Boutin TS, Le Rouzic A, Capy P. 2012. How does selfing affect the dynamics of selfish transposable elements. Mob. DNA. 3:5.

Brookfield J. 1995. Transposable elements as selfish DNA. Mobile genetic elements. Oxford University Press, New York, NY. 130–153.

Brookfield JFY. 2001. Selection on Alu sequences? Curr. Biol. 11:R900–R901.

Burke M et al. 2015. The plant parasite *Pratylenchus coffeae* carries a minimal nematode genome. Nematology. 17:621–637.

Butler MA, King AA. 2004. Phylogenetic comparative analysis: a modeling approach for adaptive evolution. Am. Nat. 164:683–695.

Callinan PA, Batzer MA. 2006. Retrotransposable elements and human disease. Genome Dyn. 1:104–115.

Camacho C et al. 2009. BLAST+: architecture and applications. BMC Bioinformatics. 10:421.

Campos JL, Charlesworth B, Haddrill PR. 2012. Molecular evolution in nonrecombining regions of the *Drosophila melanogaster* genome. Genome Biol. Evol. 4:278–288.

Campos JL, Halligan DL, Haddrill PR, Charlesworth B. 2014. The relation between recombination rate and patterns of molecular evolution and variation in *Drosophila melanogaster*. Mol. Biol. Evol. 31:1010–1028.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 25:1972–1973.

Charlesworth B, Charlesworth D. 1983. The population dynamics of transposable elements. Genet. Res. . 42:1–27.

Charlesworth B, Langley CH, Sniegowski PD. 1997. Transposable element distributions in *Drosophila*. Genetics. 147:1993–1995.

Charlesworth B, Sniegowski P, Stephan W. 1994. The evolutionary dynamics of repetitive DNA in eukaryotes. Nature. 371:215–220.

Chung W-J, Okamura K, Martin R, Lai EC. 2008. Endogenous RNA interference provides a

somatic defense against *Drosophila* transposons. Curr. Biol. 18:795–802.

Collins DW, Jukes TH. 1993. Relationship between G+ C in silent sites of codons and amino acid composition of human proteins. J. Mol. Evol. 36:201–213.

Cordaux R, Batzer MA. 2009. The impact of retrotransposons on human genome evolution. Nat. Rev. Genet. 10:691–703.

Criscione CD, Blouin MS. 2005. Effective sizes of macroparasite populations: a conceptual model. Trends Parasitol. 21:212–217.

Cutter AD, Wasmuth JD, Washington NL. 2008. Patterns of molecular evolution in *Caenorhabditis* preclude ancient origins of selfing. Genetics. 178:2093–2104.

Czech B et al. 2008. An endogenous small interfering RNA pathway in *Drosophila*. Nature. 453:798–802.

Das PP et al. 2008. Piwi and piRNAs act upstream of an endogenous siRNA pathway to suppress Tc3 transposon mobility in the *Caenorhabditis elegans* germline. Mol. Cell. 31:79–90.

Dieterich C et al. 2008. The *Pristionchus pacificus* genome provides a unique perspective on nematode lifestyle and parasitism. Nat. Genet. 40:1193–1198.

D'Onofrio G, Mouchiroud D, Aïssani B, Gautier C, Bernardi G. 1991. Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins. J. Mol. Evol. 32:504–510.

Duret L, Marais G, Biémont C. 2000. Transposons but not retrotransposons are located preferentially in regions of high recombination rate in *Caenorhabditis elegans*. Genetics. 156:1661–1669.

Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 26:2460–2461.

Ellinghaus D, Kurtz S, Willhoeft U. 2008. LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. BMC Bioinformatics. 9:18.

Felsenstein J. 1985. Phylogenies and the Comparative Method. Am. Nat. 125:1–15.

Fierst JL et al. 2015. Reproductive mode and the evolution of genome size and structure in *Caenorhabditis* nematodes. PLoS Genet. 11:e1005323.

Finnegan DJ. 1992. Transposable elements. Curr. Opin. Genet. Dev. 2:861–867.

Gambari R, Nastruzzi C, Barbieri R. 1989. Codon usage and secondary structure of the rabbit alpha-globin mRNA: a hypothesis. Biomed. Biochim. Acta. 49:S88–93.

Gao X, Hou Y, Ebina H, Levin HL, Voytas DF. 2008. Chromodomains direct integration of retrotransposons to heterochromatin. Genome Res. 18:359–369.

Gardiner CW. 1985. *Stochastic methods*. Springer-Verlag, Berlin--Heidelberg--New

York--Tokyo.

Gasior SL, Wakeman TP, Xu B, Deininger PL. 2006. The human LINE-1 retrotransposon creates DNA double-strand breaks. J. Mol. Biol. 357:1383–1393.

Ghildiyal M et al. 2008. Endogenous siRNAs derived from transposons and mRNAs in *Drosophila* somatic cells. Science. 320:1077–1081.

Goodwin TJD, Butler MI, Poulter RTM. 2003. Cryptons: a group of tyrosine-recombinase-encoding DNA transposons from pathogenic fungi. Microbiology-SGM. 149:3099–3109.

Gouy M, Gautier C. 1982. Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res. 10:7055–7074.

Govindaraju DR, Cullis CA. 1991. Modulation of genome size in plants: the influence of breeding systems and neighbourhood size. Evol. Trends in Plants (United Kingdom).

Gupta SK, Majumdar S, Bhattacharya TK, Ghosh TC. 2000. Studies on the relationships between the synonymous codon usage and protein secondary structural units. Biochem. Biophys. Res. Commun. 269:692–696.

Hansen TF. 1997. Stabilizing selection and the comparative analysis of adaptation. Evolution. 51:1341–1351.

Heath TA, Hedtke SM, Hillis DM. 2008. Taxon sampling and the accuracy of phylogenetic analyses. J. Syst. Evol. 46:239–257.

Hedges DJ, Deininger PL. 2007. Inviting instability: Transposable elements, double-strand breaks, and the maintenance of genome integrity. Mutat. Res. 616:46–59.

Hellen EHB, Brookfield JFY. 2013. Alu elements in primates are preferentially lost from areas of high GC content. PeerJ. 1:e78.

Hess J et al. 2014. Transposable element dynamics among asymbiotic and ectomycorrhizal *Amanita* fungi. Genome Biol. Evol. 6:1564–1578.

Holm L. 1986. Codon usage and gene expression. Nucleic Acids Res. 14:3075–3087.

Holm S. 1979. A simple sequentially rejective multiple test procedure. Scand. Stat. Theory Appl. 6:65–70.

Holterman M et al. 2006. Phylum-wide analysis of SSU rDNA reveals deep phylogenetic relationships among nematodes and accelerated evolution toward crown Clades. Mol. Biol. Evol. 23:1792–1800.

Hua-Van A, Le Rouzic A, Boutin TS, Filee J, Capy P. 2011. The struggle for life of the genome's selfish architects. Biol. Direct. 6:19.

Huerta-Cepas J, Dopazo J, Gabaldón T. 2010. ETE: a python environment for tree exploration.

BMC Bioinformatics. 11:24.

Huynen MA, Konings DAM, Hogeweg P. 1992. Equal G and C contents in histone genes indicate selection pressures on mRNA secondary structure. J. Mol. Evol. 34:280–291.

Ibarra-Laclette E et al. 2013. Architecture and evolution of a minute plant genome. Nature. 498:94–98.

Ikemura T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. 2:13–34.

Ikemura T. 1981. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. J. Mol. Biol. 151:389–409.

Ingram T, Mahler DL. 2013. SURFACE: detecting convergent evolution from comparative data by fitting Ornstein-Uhlenbeck models with stepwise Akaike Information Criterion. Methods Ecol. Evol. 4:416–425.

Jurka J et al. 2005. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res. 110:462–467.

Jurka J, Klonowski P, Dagman V, Pelton P. 1996. CENSOR--a program for identification and elimination of repetitive elements from DNA sequences. Comput. Chem. 20:119–121.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30:772–780.

Kawamura Y et al. 2008. *Drosophila* endogenous small RNAs bind to Argonaute 2 in somatic cells. Nature. 453:793–797.

Keren H, Lev-Maor G, Ast G. 2010. Alternative splicing and evolution: diversification, exon definition and function. Nat. Rev. Genet. 11:345–355.

Kidwell MG, Lisch D. 1997. Transposable elements as sources of variation in animals and plants. Proc. Natl. Acad. Sci. U. S. A. 94:7704–7711.

Kidwell MG, Lisch DR. 2001. Perspective: Transposable elements, parasitic DNA, and genome evolution. Evolution. 55:1–24.

Kim JM, Vanguri S, Boeke JD, Gabriel A, Voytas DF. 1998. Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome sequence. Genome Res. 8:464–478.

Knight RD, Freeland SJ, Landweber LF. 2001. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. Genome Biol. 2:research0010.

Kojima KK, Jurka J. 2011. Crypton transposons: identification of new diverse families and ancient domestication events. Mob. DNA. 2. doi: 10.1186/1759-8753-2-12.

de Koning APJ, Gu W, Castoe TA, Batzer MA, Pollock DD. 2011. Repetitive elements may comprise over two-thirds of the human genome. PLoS Genet. 7:e1002384.

Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics. 25:2286–2288.

Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. 21:1095–1109.

Leem Y-E et al. 2008. Retrotransposon Tf1 is targeted to Pol II promoters by transcription activators. Mol. Cell. 30:98–107.

Lerman DN, Michalak P, Helin AB, Bettencourt BR, Feder ME. 2003. Modification of heat-shock gene expression in *Drosophila melanogaster* populations via transposable elements. Mol. Biol. Evol. 20:135–144.

Le Rouzic A, Boutin TS, Capy P. 2007. Long-term evolution of transposable elements. Proc. Natl. Acad. Sci. U. S. A. 104:19375–19380.

Lobry JR. 1996. Asymmetric substitution patterns in the two DNA strands of bacteria. Mol. Biol. Evol. 13:660–665.

Lynch M. 2007. The frailty of adaptive hypotheses for the origins of organismal complexity. Proc. Natl. Acad. Sci. U. S. A. 104 Suppl 1:8597–8604.

Lynch M, Conery JS. 2003. The Origins of Genome Complexity. Science. 302:1401–1404.

Marracci S, Batistoni R, Pesole G, Citti L, Nardi I. 1996. Gypsy/Ty3-like elements in the genome of the terrestrial salamander *Hydromantes* (Amphibia, Urodela). J. Mol. Evol. 43:584–593.

Matzke MA, Mette MF, Matzke AJ. 2000. Transgene silencing by the host genome defense: implications for the evolution of epigenetic control mechanisms in plants and vertebrates. Plant Mol. Biol. 43:401–415.

McDonald JF et al. 1997. LTR retrotransposons and the evolution of eukaryotic enhancers. In: Evolution and Impact of Transposable Elements.Contemporary Issues in Genetics and Evolution Springer Netherlands pp. 3–13.

van Megen H et al. 2009. A phylogenetic tree of nematodes based on about 1200 full-length small subunit ribosomal DNA sequences. Nematology. 11:927–S27.

Meldal BHM et al. 2007. An improved molecular phylogeny of the Nematoda with special emphasis on marine taxa. Mol. Phylogenet. Evol. 42:622–636.

Montgomery E, Charlesworth B, Langley CH. 1987. A test for the role of natural selection in the stabilization of transposable element copy number in a population of *Drosophila melanogaster*. Genet. Res. 49:31–41.

Muto A, Osawa S. 1987. The guanine and cytosine content of genomic DNA and bacterial evolution. Proc Natl Acad Sci U S A. 84:166–169.

Nordborg M. 2000. Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self-fertilization. Genetics. 154:923–929.

Obbard DJ, Gordon KHJ, Buck AH, Jiggins FM. 2009. The evolution of RNAi as a defence against viruses and transposable elements. Philosophical Transactions of the Royal Society B-Biological Sciences. 364:99–115.

Pagel M. 1994. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. Proc R Soc B. 255:37–45.

Pagel M. 1997. Inferring evolutionary processes from phylogenies. Zool. Scr. 26:331–348.

Petrov DA, Aminetzach YT, Davis JC, Bensasson D, Hirsh AE. 2003. Size Matters: Non-LTR retrotransposable elements and ectopic recombination in *Drosophila*. Mol. Biol. Evol. 20:880–892.

Poinar GO. 1975. Description and biology of a new insect parasitic Rhabditoid, *Heterorhabditis Bacteriophora* N. Gen., N. Sp. (Rhabditida; Heterorhabditidae N. Fam.). Nematologica. 21:463–470.

Price AL, Jones NC, Pevzner PA. 2005. *De novo* identification of repeat families in large genomes. Bioinformatics. 21:i351–i358.

Pritham EJ. 2009. Transposable elements and factors influencing their success in eukaryotes. J. Hered. 100:648–655.

Pruesse E, Peplies J, Glöckner FO. 2012. SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. Bioinformatics. 28:1823–1829.

Quast C et al. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 41:D590–6.

Rebollo R, Romanish MT, Mager DL. 2012. Transposable elements: an abundant and natural source of regulatory sequences for host genes. Annu. Rev. Genet. 46:21–42.

Revell LJ, Harmon LJ, Collar DC. 2008. Phylogenetic signal, evolutionary process, and rate. Syst. Biol. 57:591–601.

Richards EJ. 2008. Population epigenetics. Curr. Opin. Genet. Dev. 18:221–226.

Sarkies P et al. 2015. Ancient and Novel Small RNA Pathways Compensate for the Loss of piRNAs in Multiple Independent Nematode Lineages. PLoS Biol. 13:e1002061.

Schnable PS et al. 2009. The B73 maize genome: complexity, diversity, and dynamics. Science. 326:1112–1115.

Sharp PM, Devine KM. 1989. Codon usage and gene expression level in *Dictyosteiium discoidtum*: highly expressed genes do [prefer [optimal codons. Nucleic Acids Res. 17:5029–5040.

Sharp PM, Tuohy TMF, Mosurski KR. 1986. Codon usage in yeast: cluster analysis clearly

differentiates highly and lowly expressed genes. Nucleic Acids Res. 14:5125–5143.

Sijen T, Plasterk RHA. 2003. Transposon silencing in the *Caenorhabditis elegans* germ line by natural RNAi. Nature. 426:310–314.

Slater GSC, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics. 6:31.

Slotkin RK et al. 2009. Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. Cell. 136:461–472.

Smit A, Hubley R. 2010a. *RepeatMasker Open-1.0. 1996-2010*.

Smit A, Hubley R. 2010b. *RepeatModeler Open-1.0. 2008-2010*.

Stamatakis A. 2014. RAxML Version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. btu033.

Stenico M, Lloyd AT, Sharp PM. 1994. Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases. Nucleic Acids Res. 22:2437–2446.

Sueoka N. 1999. Two aspects of DNA base composition: G+ C content and translation-coupled deviation from intra-strand rule of A= T and G= C. J. Mol. Evol. 49:49–62.

Szitenberg A, John M, Blaxter ML, Lunt DH. 2015. ReproPhylo: An environment for reproducible phylogenomics. PLoS Comput. Biol. 11:e1004447.

Tabara H et al. 1999. The rde-1 gene, RNA interference, and transposon silencing in *C. elegans*. Cell. 99:123–132.

Vanfleteren JR et al. 1994. Molecular genealogy of some nematode taxa as based on cytochrome c and globin amino acid sequences. Mol. Phylogenet. Evol. 3:92–101.

Whitney KD et al. 2010. A role for nonadaptive processes in plant genome size evolution? Evolution. 64:2097–2109.

Wright S, Finnegan D. 2001. Genome evolution: Sex and the transposable element. Curr. Biol. 11:R296–R299.

Wright SI, Le QH, Schoen DJ, Bureau TE. 2001. Population dynamics of an Ac-like transposable element in self-and cross-pollinating *Arabidopsis*. Genetics. 158:1279–1288.

Wright SI, Ness RW, Foxe JP, Barrett SCH. 2008. Genomic consequences of outcrossing and selfing in plants. Int. J. Plant Sci. 169:105–118.

Wright SI, Schoen DJ. 2000. Transposon dynamics and the breeding system. In: Transposable elements and genome evolution.Georgia Genetics Review 1 Springer Netherlands pp. 139–148.

Zama M. 1989. Codon usage and secondary structure of mRNA. In: Nucleic acids symposium series. pp. 93–94.

Zama M. 1996. Translational pauses during the synthesis of proteins and mRNA structure. In:

Nucleic acids symposium series. pp. 179–180.

Zou S, Ke N, Kim JM, Voytas DF. 1996. The *Saccharomyces* retrotransposon Ty5 integrates preferentially into regions of silent chromatin at the telomeres and mating loci. Genes Dev. 10:634–645.

Zuker C, Cappello J, Lodish HF, George P, Chung S. 1984. *Dictyostelium* transposable element DIRS-1 has 350-base-pair inverted terminal repeats that contain a heat shock promoter. Proc. Natl. Acad. Sci. U. S. A. 81:2660–2664.