Genome Biology and Evolution Advance Access published April 2, 2014 doi:10.1093/gbe/evu068

The holozoan *Capsaspora owczarzaki* possesses a diverse complement of active transposable element families

Martin Carr*[1] and Hiroshi Suga[2]

[1]School of Applied Sciences, University of Huddersfield, West Yorkshire, United Kingdom

[2]Instituto de Biologia Evolutiva, Passeig Maritim de la Barceloneta, 37-49, 08003 Barcelona, Spain

*Author for correspondence: Martin Carr, School of Applied Sciences, University of Huddersfield, Huddersfield, West Yorkshire, United Kingdom, Telephone: +44 484-471608, email: M.Carr@hud.ac.uk

Abstract

Capsaspora owczarzaki, a protistan symbiont of the pulmonate snail *Biomphalaria glabrata*, is the centre of much interest in evolutionary biology due to its close relationship to Metazoa. The whole genome sequence of this protist has revealed new insights into the ancestral genome composition of Metazoa, in particular with regard to gene families involved in the evolution of multicellularity. The draft genome revealed the presence of 23 families of transposable element, made up from DNA transposon as well as LTR and non-LTR retrotransposon families.

The phylogenetic analyses presented here show that all of the transposable elements identified in the *C. owczarzaki* genome have orthologous families in Metazoa, indicating that the ancestral metazoan also had a rich diversity of elements. Molecular evolutionary analyses also show that the majority of families have recently been active within the *Capsaspora* genome. One family now appears to be inactive and a further five families show no evidence of current transposition. Most individual element copies are evolutionarily young, however a small proportion of inserts appear to have persisted for longer in the genome. The families present in the genome show contrasting population histories and appear to be in different stages of their life cycles. Transcriptome data have been analysed from multiple stages in the *C. owczarzaki* life cycle. Expression levels vary greatly both between families and between different stages of the life cycle, suggesting an unexpectedly complex level of transposable element regulation in a single celled organism.

Keywords: Filasterea, phylogenetics, population genetics, retrotransposons, DNA transposons

Introduction

The eukaryotic supergroup Opisthokonta contains, in the Metazoa and Fungi, two of the major multicellular eukaryotic lineages. In addition to the multicellular groups however, the opisthokonts also comprise the protistan groups Choanoflagellatea, Corallochytrea, Filasterea, Ichthyosporea, and the nuclearioid amoebae (Carr and Baldauf 2011). The evolutionary relationships within the opisthokonts are slowly becoming clear, with the group being divided into two major lineages. The Holomycota (Liu et al. 2009), are composed of Fungi and their sister group the nuclearioid amoebae, whilst the remaining opisthokont groups make up the Holozoa (Lang et al. 2002; Shalchian-Tabrizi et al. 2008).

Capsaspora owczarzaki is, along with two species of *Ministeria*, one of three protist taxa assigned to Filasterea (Shalchian-Tabrizi et al. 2008) and is the sole known representative of the genus *Capsaspora*. The species is of biological interest for two important reasons. Firstly, *C. owczarzaki* has a symbiotic, or parasitic, relationship with the snail *Biomphalaria glabrata*, which acts as a vector in the transmission of the human disease schistosomiasis (Sohn and Kornicker 1972). This chronic disease is estimated by the World Health Organization to affect over 240 million people and annually results in over 200,000 deaths (WHO Fact sheet N°115, 2013). Schistosomiasis is caused by the trematode worm *Schistosoma mansoni*, which uses *B. glabrata* as an intermediate host. The presence of *C. owczarzaki* may afford *B. glabrata* resistance to *Schistosoma* infection, as it preys upon the trematode larvae in the snail haemolymph (Owczarzak et al. 1980; Stibbs et al. 1979), which in turn has lead to speculation that *C. owczarzaki* has a potential role as a biological control of schistosomiasis (Ruiz-Trillo et al. 2006).

Secondly, *C. owczarzaki* is a close relative of the metazoans and the choanoflagellates (Shalchian-Tabrizi et al. 2008; Torruella et al. 2012) and is therefore an important study taxon in the transition from the unicellular protists to Metazoa (Ruiz-Trillo et al. 2008b; Sebé-Pedrós et al. 2011; Suga et al. 2013). The genome of *C. owczarzaki*, sequenced at the Broad Institute, Massachusetts under the Origins of Multicellularity Initiative (Suga et al. 2013), is 27.97Mb in length and contains over 8,000 predicted genes. The genome size is therefore similar to many fungal taxa and unicellular eukaryotes, although smaller than most metazoan genomes (Carr and Baldauf 2011).

Transposable elements are important components of eukaryotic genomes, acting as a source of deleterious mutations, genetic variability, phylogenetic markers, and beneficial domestication by the host, as well as being a component of genomic architecture (Biémont 2009; Capy et al. 2000; Carr et al. 2001; Casola et al. 2007; Eanes et al. 1988; Jordan et al. 2003). Transposable elements fall into two classes, defined by their mode of transposition. Class I elements are retrotransposons, either with or without long terminal repeats (LTRs), which transpose via an RNA intermediate. LTR elements are commonly found in their host genomes in two different forms. Full length elements (FLE) are composed of two LTRs that flank *gag* and *pol* open reading frames (ORFs), which encode structural and replication proteins respectively. The second form of LTR retrotransposon is the solo LTR. The two LTR sequences of a single element are capable of undergoing ectopic recombination with each other. This process leads to the excision of one LTR and the internal DNA, as an extrachromosomal circular element, and leaves a single LTR present in host chromosome. Transposition of all retrotransposons is facilitated by a Polyprotein (Pol), a multifunction protein which encodes reverse transcriptase and integrase domains. Non-LTR retrotransposons lack terminal repeats, however, like the LTR elements, they contain both *gag* and *pol* ORFs and transpose via an RNA intermediate. Daughter non-LTR elements are frequently 'dead on arrival', due to the propensity of non-LTR retrotransposons to undergo 5' truncations during transposition (Malik and Eickbush 1998)

Class II elements, the DNA transposons, exist solely as DNA. They usually employ a simple 'cut and paste mechanism' of transposition via their Transposase (Tnpase) protein, in which the entire element is excised from the host chromosome and inserted into a new genomic location.

Transposable elements have traditionally been considered as either selfish or junk DNA conferring no benefit to their host. This view has been challenged with clear examples of both beneficial individual insertions (Franchini et al. 2004; Schlenke and Begun 2004) and transposable element families (Biessmann et al. 1992). The idea that transposable elements are solely genomic parasites now appears overly simplistic; however the majority of insertions appear to be deleterious or neutral in a broad range of eukaryotes (Charlesworth et al. 1992; Jordan and McDonald 1999; Pereira 2004). Their deleterious nature results in natural selection playing an important role in restraining the proliferation of transposable elements in many host populations.

Most studies on transposable elements in eukaryotes have centred on the major multicellular groups, i.e. Metazoa, Fungi and plants, with relatively little known on the function and evolution of transposable elements in protists. To date only a single choanoflagellate, which are the sister group to metazoans, has been studied within the opisthokont protists with regard to the evolution of their transposable elements (Carr et al. 2008). *Monosiga brevicollis* was shown to harbour three transposable element families, all of which were LTR retrotransposons and active. Carr et al. (2008) also showed that transposable elements only constitute a very low fraction (~1%) of the *M. brevicollis* genome, with families in a state of constant turnover through ongoing transposition and loss possibly due to natural selection.

The *C. owczarzaki* genome allows further insight into the evolution of transposable elements in the opisthokont protists. The recently published *C. owczarzaki* draft genome identified 23 transposable element families, comprising five LTR retrotransposon, four non-LTR retrotransposon and 14 DNA transposon families. The 23 families were shown to belong to seven major superfamilies of transposable element and contributed to approximately 9% of the genome (Suga et al. 2013).

Presented here is a detailed characterization of the transposable elements present in the draft *C. owczarzaki* genome with an emphasis on their evolutionary biology. All of the families have been placed in a phylogenetic framework and are shown to cluster together to the exclusion of non-*Capsaspora* families. The *C. owczarzaki* families are generally isolated on long internal branches, however they do form phylogenetic associations with transposable elements from other opisthokont taxa. The data presented here give an insight into the ancestral transposable element composition of Metazoa, as all of the families identified in the *C. owczarzaki* genome have orthologous families in metazoan taxa. The seven superfamilies present in *C. owczarzaki* appear to have been present in the last common ancestor of metazoans and filastereans and subsequently retained in both lineages.

Through molecular evolutionary and phylogenetic analyses of the individual element copies, it can be seen that 22 of the 23 families have recently undergone transposition within the sequenced genome; however one of the families has subsequently lost the ability to transpose. The transposable element population of *C. owczarzaki* is dominated by young elements, however the presence of ancient inserts, as well as divergent subfamilies, highlight that many families are long term components of the genome. The families show contrasting population histories within *C. owczarzaki*; two families appear to be possible recent arrivals in the genome and five of the families show no evidence of current transposition.

Finally, the expression levels of each family have been determined in three different stages of the *C. owczarzaki* life cycle. The data suggest a strong relationship between the rates of expression and transposition in *Capsaspora* and, unexpectedly for a single celled organism, stage specific expression for families.

Methods

Extraction of individual element termini and transcripts

The termini of individual copies were extracted by Megablast similarity searches, using default parameters, on the *C. owczarzaki* Trace Archive using the reported sequences of each family, taken from the draft genome paper (Suga et al. 2013), as query sequences. The query sequences for the LTR retrotransposon families were the LTR sequences; 5' and 3' non-coding regions were used as queries for the non-LTR retrotransposons and the DNA transposon families. Query sequences were limited to a maximum of 300bp. The integration of a transposable element generally results in the duplication of the target site, with the same sequence being present at both the 5' and 3' termini. The termini for individual inserts were identified through the flanking DNA and target site duplications (TSDs) generated upon integration. The presence of both target site duplications in the Trace Archive allowed the two ends of individual inserts to be identified.

Transcriptome data were generated from three stages of the *C. owczarzaki* life cycle (Sebé-Pedrós et al. 2013), with three replicates produced for each stage, and are available in the NCBI BioProject PRJNA20341. The raw reads from the transcriptome were mapped to the full length sequences of each family using SMALT (Hannes Ponstingl, SMALT – Wellcome Trust Sanger Institute, http://www.sanger.ac.uk/resources/software/smalt/, 2013). The expressions of TEs were approximated by the normalized number of RNAseq reads, and visualised with a heatmap. The values were normalized by the maximum value (i.e. *Cocv1* in the adherent stage) and the heatmap was drawn in R 3.0.2 using its Bioconductor package (Gentleman et al. 2004). Since the expression levels of some families were extremely high, the same heatmap was also drawn by log scale, in which the minimum value is set to zero.

Differences in expression level between stages in the life cycle were tested using ANOVA for each family. For those families which showed significant variation across the three stages, the data were examined to see if one stage showed significantly elevated expression. To this end, for each family, the three replicates from the two stages with the highest combined expression level were compared using an unpaired t-test.

Phylogenetic and molecular evolution analyses

Superfamily phylogenies were created for all of the families in the *C. owczarzaki* genome using the amino acid sequences of Pol, in the case of retrotransposons, and Tnpase in the case of DNA transposons. In order to identify families closely related to the *C. owczarzaki* families, sequence similarity searches were performed with BLASTp and tBLASTn on the National Center for Biotechnology Information (NCBI) Nonredundant Proteins Sequences and Nucleotide Collection databases respectively using default parameters. Further searches were performed on the whole genome sequences of a taxonomically broad set of eukaryotes following the protocol set out in Suga et al. (2013) (Amoebozoa: *Dictyostelium discoideum*; Apusozoa: *Thecamonas trahens*; Choanoflagellatea: *M. brevicollis*, *S. rosetta*; Excavata: *Naegleria gruberi;* Fungi: *Laccaria bicolor*, *Rhizopus delemar*; Metazoa: *Amphimedon queenslandica*, *Drosophila melanogaster, Homo sapiens*, *Nematostella vectensis*, *Trichoplax adhaerens*) with both BLASTp and tBLASTn.

Amino acid sequences were aligned using MUSCLE 3.7 (Edgar 2004), on the EMBL-EBI server using default parameters, and modified by eye to minimize indel regions. In order to determine the appropriate amino acid substitution model for the maximum likelihood phylogenetic analyses, the alignments were analysed with ProtTest 2.4 (Abascal et al. 2005). Maximum likelihood analyses were performed using RAxML

7.2.6 (Stamatakis et al. 2005) on raxmlGUI2 (Silvestro 2012). The maximum likelihood phylogenies were initiated by 100 maximum parsimony trees and bootstrapped with 1,000 replicates. Bayesian inference phylogenies were created using MrBayes 3.1.2 (Ronquist and Huelsenbeck 2003) using a mixed amino acid model. The analyses were ran, using default temperatures, until the two parallel chains had reached convergence (average standard deviation of split frequencies <0.01) and consisted of a minimum of 1,000,000 generations with a sampling frequency of 100. The first 25% of sampled trees were discarded as burnin.

Phylogenies of the individual copies of each family were generated using the element termini sequences extracted through Megablast. Trees were created using maximum likelihood with RAxML and Bayesian inference with MrBayes, from nucleotide alignments created using MUSCLE. LTR retrotransposon phylogenies were generated using LTR sequences. DNA transposon and non-LTR retrotransposon phylogenies were created using 5’ inverted terminal repeat (ITR) and 3’ untranslated region (UTR) sequences respectively. The 3’ UTR sequences were chosen for the non-LTR retrotransposon families as there were fewer 5’ UTR sequences in the genome. The maximum likelihood trees were initiated by 100 maximum parsimony trees and bootstrapped with 1,000 replicates and performed using the GTRCAT model, as recommended by the program authors. MrBayes analyses were performed using the same protocol as with the amino acid phylogenies, other than substitution model (the GTR+I+Γ nucleotide substitution model was employed).

Values of π (Nei et al. 1987) and Tajima’s *D* (Tajima 1989) for DNA transposon and non-LTR retrotransposon families were generated from the 5’ ITR and 3’ UTR alignments used in the phylogenetics analyses. For the chromoviral families a single LTR sequence for each insert was used to determine values of Tajima’s *D*. Individual alignments with only one LTR per insert were created for all elements, full length elements and solo LTRs to determine values of π. Population genetics statistics were calculated using DnaSP v5.10.01 (Librado and Rozas 2009).

All of the alignments, with GenBank Accession Numbers and Trace Read Identification Numbers, used for each analysis are present in Supplementary Files 3 and 4.

Results

Phylogenetic Analyses of *C. owczarzaki* Transposable Element Families

The *C. owczarzaki* draft genome revealed the presence of 23 families of transposable element in the culture ATCC 30864 (Suga et al. 2013). The analyses of Suga et al. (2013) classified the families into seven superfamilies, these being chromovirus (LTR retrotransposon), *L1* (non-LTR retrotransposon), *Cobalt*, *CACTA*, *MULE*, *pogo* and *Tc1* (DNA transposon), but did not attempt to determine their phylogenetic relationships with transposable element families from other species. In order to understand the evolutionary origin of the *C. owczarzaki* transposable element families, phylogenies were created for each superfamily (Figures 1-3 and Supplementary Data Figures S1A-D).

Phylogenies were generated using the replication ORFs of each family, i.e. the Pol of the retrotransposons and the Tnpase of the DNA transposons, as they are the most highly conserved component of transposable element genomes. With the exception of the poorly resolved chromovirus phylogeny, all superfamily phylogenies clustered the *C. owczarzaki* families together with strong support (89-100% maximum likelihood bootstrap percentage (mlBP) and 1.00 Bayesian inference posterior probability (biPP)). In the chromoviral phylogeny *Cocv1-4* clustered together, however *Cocv5* has an unsupported position, clustering with Pol sequences from metazoans and fungi (Figure 1).

The grouping of the *C. owczarzaki* families together on isolated internal branches indicates that the families have a long-term evolutionary history within the filasterean lineage and have most probably radiated within it. Furthermore, all of the *C. owczarzaki* families were present in phylogenetic groups composed predominantly from other opisthokont transposable element families, a finding consistent with their vertical inheritance during the opisthokont radiation (Figures 1-3 and Supplementary Information Figures S1A-D).

It is becoming clear that close biological relationships, such as those of a parasite and host, may facilitate the horizontal transfer of transposable element families between species (Gilbert et al. 2010; Kuraku et al. 2012; Yoshiyama et al. 2001). As *C. owczarzaki* has an intimate relationship with both *B. glabrata* and *S. mansoni*, their genomes were screened for the presence of *C. owczarzaki* transposable elements. BLASTn screening of the *S. mansoni* genome (Berriman et al. 2009), as well as BLASTn and tBLASTn screening of the NCBI *B. glabrata* Whole Genome Shotgun Contigs (WGS) and Expressed Sequence Tags (EST) databases, failed to uncover close hits (>90% identity nucleotide/amino acid identity) of the *C. owczarzaki* transposable elements. These results indicate that there has been no successful horizontal transfer of transposable elements between *C. owczarzaki* and either species.

Target Site Insertion Patterns of *C. owczarzaki* Transposable Elements

The integration of transposable elements leads to the duplication of the target sequence and, for many families, the length of the target site is highly conserved (Craig 2002). Recent work has also highlighted that many of the transposable element families in *D. melanogaster* posses a level of conservation in the sequences of their target sites (Nefedova et al. 2011; Linheiro and Bergman 2012). Suga et al. (2013) determined the length of target sites for 22 of the 23 TE families in the *C. owczarzaki* genome (the exception being the non-LTR retrotransposon *CoL3*) and presented here are the nucleotide compositions of the TSDs for the 22 families (Figure S2). Within the DNA transposon families the *pogo*-like, *Tc1*-like and *Cobalt* families all have a highly conserved, two base, TA target site. TA target sites are a common feature of many DNA transposon families (Plasterk et al. 1999), however the *CoCACTA* and *Com* families all possess longer TSDs. The two *CACTA* families both have three base TSDs, whilst the two *Com* families have nine base TSDs; however, neither superfamily show conserved target sequences. As is commonly found with chromoviruses (Novikova et al. 2010), all five of the *C. owczarzaki* chromoviral families generate 5 base target site duplications. The families do not possess conserved target sequences, but there is a general preference for guanine or cytosine in both the terminal 5' and 3' positions of the target sequence. Target sites were identified for total of 169 LTRs, from across all five families, and the GC content for the first and fifth positions are 66.9% and 68% respectively. In contrast, the internal three nucleotides in the target sites are relatively GC poor (38.5%, 40.2% and 39.6% GC content). Three non-LTR retrotransposon families had the lengths of their TSDs identified by Suga et al. (2013), these being *CoL1*, *CoL2* and *CoL4*; however, the three families have no discernable conservation in TSD sequence (Figure S2).

Evidence For Recent Transposable Element Activity In The *C. owczarzaki* Genome

A number of eukaryotic species have been shown to possess families of transposable element that are no longer capable of transposition (Hellen and Brookfield 2011; Zou et al. 1995). The presence of multiple copies in a sequenced genome therefore does not confirm that a family is currently active within its host population. Twenty two of the families have multiple copies in the sequenced *C. owczarzaki* genome (Tables 1-3), with one family, the DNA transposon *Cobalt3*, being present as a single copy. Database mining of the *C. owczarzaki* genome and transcriptome was therefore employed to produce four lines of evidence, in the absence of observed transposition, to investigate element activity within the *C. owczarzaki* population. To this end, all terminal and flanking sequences present in the NCBI Trace Archive were extracted to identify the sequences of individual element copies. The Trace Archive was screened in favour of the assembled *C. owczarzaki* genome, as transposable element insertions are often collapsed in draft genome assemblies.

Expression of Transposable Element Families in C. owczarzaki

Transcription is an essential component of transposition, in order to produce catalytic proteins and, in the case of retrotransposons, RNA daughter elements. The presence of transposable element sequences in the RNASeq reads generated by the *C. owczarzaki* genome project (Suga et al. 2013; Sebé-Pedrós et al. 2013) would therefore provide indirect evidence that a family may be active. The complete RNASeq database contained 394,576,834 reads, of which 1,165,292 (0.3%) were of transposable element origin (Tables 1-3; Figure 4). Only one family, the LTR retrotransposon *Cocv4*, was essentially absent from the reads; this family presented a small number of short fragments, which did not cover the entire length of the known sequence, indicating that this family is no longer active in the sequenced culture of *C. owczarzaki*. The expression levels of the families vary by over three orders of magnitude and transcripts of *Cocv1* dominate, with 37.5% of reads being from this family. The number of RNASeq reads shows a strong positive correlation (*r*=0.846) with the number of identical paralogous copies in the genome for each family (Figure S3). This shows that the number of identical copies (which can be viewed as reflecting the transposition rate of a family) can be considered a good predictor of a family's expression in *C. owczarzaki*.

RNASeq was performed on three stages in the life cycle of *C. owczarzaki*, these being an amoeboid crawling stage, a free-floating stage and an aggregated cell stage (Sebé-Pedrós et al. 2013). Fifteen of the 23 transposable element families exhibited significant variation in expression level across the three stages and, of these, 9 showed significantly elevated expression in a particular stage (Supporting Information File 2). Three of the four expressed LTR retrotransposon families showed significantly raised expression in one stage in the life cycle of *C. owczarzaki*, however different families showed different expression patterns. In contrast, only one non-LTR retrotransposon family, *CoL1*, showed significant variation in expression across stages and a stage with an elevated expression level. Ten of the 14 DNA transposon families showed stage-specific expression patterns, with five (*Cobalt1*, *Com1*, *Cop1*, *Cop3* and *Cop4*) having a preferred stage for expression.

Patterns of Transposable Element Nucleotide Diversity in the C. owczarzaki *Genome*

The second line of evidence for element activity was restricted to the LTR retrotransposons. Their LTR sequences are believed to evolve in the fashion of a molecular clock, since they are identical at the time of integration and gradually diverge due to the accumulation of mutations (Bowen and McDonald 2001). Using unique TSDs, it was possible to identify the 5' and 3' LTRs from 40 individual inserts from across all five LTR retrotransposon families. Intra-element LTR nucleotide identity ranged between 99.3-100% (Table 1), demonstrating that all five families have been active recently and that ancient full length elements appear to be absent from the genome.

There are no analogous methods for using terminal repeats to age non-LTR retrotransposon and DNA transposon inserts. Two further methods were therefore used to obtain evidence for recent transposition in all 22 multicopy families. A number of recent studies have used Tajima's *D* statistic to look for the signature of transposition activity (Bartolomé et al. 2009; Carr et al. 2012; Maside et al. 2003; Sánchez-Gracia et al. 2005). Under such a model, most elements in a population will have very recent common ancestry and most nucleotide variants will be present at low frequency, resulting in a negative value of *D*. If a transposable element family has multiple active lineages in a genome, in other words it exhibits population substructure, some nucleotide variants may be present at an intermediate frequency, which could potentially result in a positive value of *D*. It is therefore important to generate phylogenies of the individual copies of a family in order to evaluate the results of Tajima's *D* tests. Recently transposed elements are expected to have short terminal branches in a phylogenetic tree, as daughter elements have had little time to accumulate unique mutations. The presence of identical paralogous elements can be considered as strong evidence for current element activity (Tables 1-3); short terminal branches indicate that a family has recently transposed, but do not provide any direct evidence that the family is currently active in the sequenced strains. Older inserts in the genome will have accumulated numerous unique mutations and therefore be present on long terminal branches.

All of the retrotransposon families exhibited negative values for *D*, consistent with all such families having recently undergone transposition; however the deviation from neutral expectation was not significant for the LTR retrotransposon *Cocv5*, as well as the non-LTR retrotransposons *CoL1* and *CoL2* (Tables 1-2). In phylogenies generated from LTR sequences the five chromoviral families all harbour identical paralogous copies, confirming that they have all undergone very recent transposition in the *C. owczarzaki* genome (Figure 5, Supplementary Data Figure S4A-D). Within the non-LTR retrotransposons all four families show copies with high nucleotide identity (>98.5% nucleotide identity), again highlighting that the families have recently been active; however only *CoL2* and *CoL4* show identical non-allelic copies (Supplementary Data Figures S4E-H).

Values for Tajima's *D* could not be determined for *Cobalt3* and *CoCACTA2*; the former only has a single copy in the genome, whilst the latter has too many overlapping indels to allow *D* to be calculated. However 7 of the remaining 12 DNA transposon families produced negative *D* values. Significantly negative values of *D* were obtained for *Cop1*, *Cop3* and *CoTc2*, whilst positive values of *D* were produced for

Cobalt1, *Cobalt2*, *CoCACTA1*, *Com1* and *Cop2* (Table 3). The phylogenies of these latter five families show multiple identical sequences, consistent with their current transposition (Supplementary Data Figure S4I-K, S4M and S4P). The families also have deep internal branching, a result of their population subdivision, which accounts for the positive values of *D*. Only three of the DNA transposon families do not show multiple copies with identical sequences, these being the single copy *Cobalt3* as well as *Com2* and *CoTc2* (Supplementary Data Figures S4N and S4U). However all multi-copy DNA transposon families have multiple inserts with high nucleotide (≥99%) identity, highlighting their recent transposition.

Bergman and Bensasson (2007) used terminal branch length as a proxy for retrotransposon age in the genome of *Drosophila melanogaster*, showing that LTR families were composed from younger individuals than non-LTR families. There is no such dichotomy in the *C. owczarzaki* genome, rather the phylogenies of all transposable element classes show that the *C. owczarzaki* genome is dominated by young copies. However 13 of the families contain older inserts (defined as copies with a branch length ≥0.05 substitutions/site – Figure 6) indicating that some individual elements can persist for a long time in the *C. owczarzaki* genome.

C. owczarzaki Transposable Element Families Show Differing Population Histories

Transposable element families can be considered to have their own life cycle within their host's genome (Brookfield 2005). The family has an origin (either through horizontal transfer between hosts or transposable element 'speciation' within a single host population) and subsequently proliferates in the genome. Individual elements in the new family accumulate null mutations, until no more functional copies exist and the family finally becomes a genomic fossil incapable of further transposition. The phylogenies of the 22 multicopy transposable element families show markedly different topologies

(Figure 5, Supplementary Data Figures S4A-U) and levels of nucleotide diversity within the families, based upon π, vary by over an order of magnitude (Tables 1-3). It is clear therefore that the families exhibit different population dynamics within the *C. owczarzaki* genome.

Within the chromoviral families *Cocv1* (Figure 5) and *Cocv3* (Figure S4B) both contain groups of long branch, presumably ancient, solo LTRs as well as clades of shorter branched sequences made up from both solo LTRs and full length elements. In contrast there are no long branch copies of *Cocv2* (Figure S4A) and the phylogeny has a star-like appearance, highlighting a recent common origin of all copies in the genome.

Carr et al. (2008) reported a greater level of nucleotide diversity among paralogous solo LTRs compared to LTRs from full length elements in the choanoflagellate *M. brevicollis*, as solo LTRs can persist in a population for a longer period of time. A similar situation exists in the case of *Cocv1*, *Cocv3* and *Cocv4* (Table 1). In *Cocv2* and *Cocv5* there are similar levels of nucleotide diversity in both solo LTRs and LTRs from full length elements, however their phylogenies highlight that the families have different population histories. *Cocv2* presents a star phylogeny with very little sequence divergence between copies (Supplementary Data Figure S4A). In contrast, *Cocv5* shows a higher degree of population substructure than other LTR retrotransposon families, with multiple apparently active lineages in the genome (Supplementary Data Figure S4D). Furthermore, no two of the 14 full length *Cocv5* elements in the genome have identical sequences (Figure S4D). This suggests that the family may have a lower recent rate of transposition in comparison to *Cocv1-3*, which all possess large numbers (>15) of identical full length copies. A putative low transposition rate is however not reflected by the expression level of *Cocv5*, which is the sixth most abundant transposable element family in the RNASeq database (Tables 1-3). Matsuda and Garfinkel (2009) showed that antisense RNA copies of *Ty1* in *Saccharomyces* can inhibit transposition, therefore the high expression of *Cocv5* and its putative low transposition rate are not incompatible.

In contrast to the other chromoviral families, which have a greater number of full length elements than solo LTRs, *Cocv4* consists mainly of long branched, presumably old, solo LTRs (Supplementary Data Figure S4C) and harbours the greatest level of nucleotide diversity within the chromoviruses (Table 1). A single full length element is present in the genome. This element appears to be the product of a recent transposition event, as no nucleotide substitutions have occurred in either of the LTRs. The element, and therefore the entire family, however appears to be no longer capable of transposition in the sequenced strains of *C. owczarzaki*, as the 3' LTR is truncated to 58bp in length by a large deletion. The lack of element activity is consistent with the dearth of RNASeq reads for this family, which only cover a small proportion of the *Cocv4* genome (data not shown).

Within the non-LTR retrotransposons, *CoL2* and *CoL4* show similarities in their phylogenies in that both families contain long branched inserts as well as clusters of short branch, presumably younger, copies (Supplementary Data Figures S4F and S4H). The two families also show multiple identical copies. The data indicate that both are old families that are currently transposing within the *C. owczarzaki* population. In contrast, although they both possess highly similar inserts, there are no identical non-allelic copies of either *CoL1* or *CoL3* (Supplementary Data Figures S4E and S4G). The data therefore suggest that both families have been active in the recent evolutionary history of *C. owczarzaki*, however their transposition may have currently ceased. Furthermore, both *CoL1* and *CoL3* have considerably lower expression levels than *CoL2* and *CoL4* (Table 2).

Within the DNA transposon families the level of nucleotide diversity varies by over an order of magnitude (Table 3). This is due to some families showing either deep or complex subdivision (e.g. *Com2*, *Cop2*, *Cop5* – Supplementary Data Figures S4N, S4P and S4S), whilst other families (*Cop1* and *CoTc1* – Supplementary Data Figures S4O and S4T) show limited subdivision with most copies having a single recent common ancestry. Seven of the DNA transposon families (*Cobalt1-2*, *CoCACTA1*, *Com1*, *Cop1-2* and *CoTc2* – Supplementary Data Figures S4I-J, S4M, S4O, S4P and S4U) are composed entirely from young elements (i.e. copies with a branch length <0.05 substitutions/site). However, all of these families show internal branching and population subdivision; highlighting that the families are older than the current population of copies in the genome.

Discussion

TE Family Diversity in the C. owczarzaki *Genome*

C. owczarzaki is the second holozoan protist, after *M. brevicollis*, to have a survey undertaken of its transposable elements. The transposable element complement of *C. owczarzaki* shows some similarities, but also many differences, to that of *M. brevicollis* and therefore affords us a greater understanding of transposable element evolution within the holozoans.

The most obvious difference between the two species is the greater diversity of families present in the genome of *C. owczarzaki*, in terms of both the number of families and the different classes of element present. *M. brevicollis* possesses a very limited range of transposable elements, with only three families, all of which are LTR retrotransposons (Carr et al. 2008). This contrasts sharply with the 23 families present in the *C. owczarzaki* genome, which can be assigned to seven superfamilies of DNA transposon, LTR retrotransposon and non-LTR retrotransposon. The genome sequence of *C. owczarzaki* was produced from a polymorphic and outbred culture, which suggests that we have captured much of the TE family diversity in this species, but it remains possible that additional families are present in the full population. The low number of families present in *M. brevicollis* would appear to be a result of major element loss and, as suggested by Carr et al. (2008), may be atypical for choanoflagellates. The closely related choanoflagellate *Salpingoeca rosetta* possesses transposable elements from four of the superfamilies present in the *C. owczarzaki* genome (chromovirus, *Cobalt*, *pogo,* MULE see Figures 1 and 3, Supplementary Data Figures S1A and S1D) and a third choanoflagellate, *Monosiga* sp. (ATCC 50635), has also been shown to possess DNA transposon, LTR and non-LTR retrotransposon families (Carr et al. 2008).

All seven superfamilies present in the *C. owczarzaki* genome are nested within clades composed predominantly of opisthokont families, consistent with their inheritance within the group by vertical transmission. This strongly suggests that the ancestral opisthokont also possessed a diverse complement of transposable elements and that widespread element loss and lineage sorting has occurred during the opisthokont radiation. The presence of all of the seven superfamilies in the genomes of metazoans indicates that the superfamilies were also probably present in the last common ancestor of Metazoa. These data are of particular importance in beginning to determine the evolution of transposable elements in Holozoa. None of the families present in the *M. brevicollis* genome cluster with holozoan sequences (Carr et al. 2008), so they were not informative for studies in the ancestral transposable element composition of either Holozoa or Metazoa. However, reconstructing the deep evolutionary history of transposable element superfamilies can be a difficult process, due to their rapid rate of evolution (Peterson-Burch and Voytas 2002) which often results in unresolved phylogenetic trees. As a result, phylogenies are often consistent with both vertical and horizontal inheritance. The LTR retrotransposon phylogeny (Figure 1) shows that all three screened species of holozoan protist (*C. owczarzaki*, *M. brevicollis* and *S. rosetta*) possess chromovirus families. Chromoviruses are also a common feature of fungal genomes and it has been suggested that they were a component of the ancestral opisthokont genome (Carr et al. 2008; Kordiš et al. 2005). This makes their apparent paucity in metazoan genomes, where they appear to be restricted to a small number of vertebrate lineages (Kordiš et al. 2005), something of an enigma. The chromoviral phylogeny has a poorly resolved backbone, so cannot rule out the vertical transmission and subsequent loss of the superfamily in the majority of metazoan lineages. Nevertheless, the phylogeny does not cluster the chromoviruses from Metazoa with those of their sister group, the choanoflagellates, as would be expected under vertical inheritance (Figure 1). Although the relationship is poorly supported, the metazoan chromoviruses are nested within those from the Dikarya fungi; therefore the phylogeny is also consistent with the potential horizontal transfer of chromoviruses to an ancestral vertebrate from a fungal lineage. The phylogeny is also consistent with the opisthokont last common ancestor having a highly diverse suite of chromovirus families, which have subsequently segregated in different lineages due to stochastic loss.

The increasing volume of available genomic data from opisthokont protist lineages, e.g. from multiple ichthyosporean, filasterean and nuclearioid taxa (Ruiz-Trillo et al. 2008a), is likely to soon bridge the large evolutionary distances between the transposable element families placed in the current phylogenies. These increased data should provide greater resolution to phylogenetic analyses and may allow questions on the inheritance of superfamilies within the opisthokonts to be answered.

TE Activity In The C. owczarzaki *Genome*

All three families in the *M. brevicollis* genome appear to be currently active (Carr et al. 2008). In contrast, *C. owczarzaki* has one putatively non-functional family in *Cocv4* and a further five families (*Cobalt3*, *CoL1*, *CoL3*, *Com2* and *CoTc2*) which show no phylogenetic evidence for current transposition. Inactive transposable element families are a common feature in a broad range of opisthokont genomes (Belshaw et al. 2005; Bergman and Bensasson 2007; Carr et al. 2012), therefore the presence of such families in the *C. owczarzaki* genome is not surprising. Furthermore, the transposable element families in the *C. owczarzaki* genome appear to be at different stages in their life cycles. The six families listed above appear to be either non-functional or potentially in the process of losing activity, whilst the majority of families appear to be active, long term inhabitants of the genome. *Cocv2* however appears to be a very young family. The family has no old copies present in the genome, has a very low level of nucleotide diversity and presents a star like phylogeny (Figure S4A and Table 1). In these respects it is very similar to *Ty2*, which has recently undergone a horizontal transfer from *S. mikatae* into the genome of *S. cerevisiae* (Carr et al. 2012). Similarly, the fact that there is only a single copy of *Cobalt3* in the genome is consistent with its own recent arrival in the genome. The single *Cobalt3* element is expressed in *C. owczarzaki* cells, albeit it at a low level, and encodes a conserved Tnpase protein (Suga et al. 2013) suggesting that the family has the potential to invade the genome. *Cocv2* and *Cobalt3* may therefore be new arrivals into the *C. owczarzaki* genome; however this cannot be confirmed in the absence of sequences from putative donor species and, despite a lack of evidence to show this, both families could be long term inhabitants in the genome,.

The transposable elements present in *M. brevicollis* and *C. owczarzaki* genome show a similarly low copy number, with all families harbouring less than 100 copies (Tables 1-3). This is in sharp contrast to many metazoan taxa, where families may be present in many thousands of copies (Robertson and Lampe 1995; Smit and Riggs 1996). The low copy numbers observed are likely to be, in part, due to the large population sizes of protists (Finlay and Fenchel 2004), which will facilitate more efficient purifying selection against transposable element induced deleterious mutations. It may also reflect a requirement of single celled organisms to maintain a streamlined genome in order to efficiently undergo rapid nuclear replication, cell division and reproduction.

Further similarities between *C. owczarzaki* and *M. brevicollis* are seen with regards to their complement of LTR retrotransposons. In both species, irrespective of the overall age of the families, full length elements are always young, with no single full length element showing more than 2.5% divergence between its two LTRs. This trait however is seen in a broad range of eukaryotes (Bowen and McDonald 2001; Gorinšek et al. 2004; Kordiš 2005; Xu et al. 2006), suggesting that full length LTR retrotransposons tend to be highly deleterious and rapidly removed from populations, either by the action of purifying selection or LTR-LTR recombination.

The apparent high rate of element turnover in the *C. owczarzaki* genome is not restricted to LTR retrotransposon families. The multicopy families are predominantly composed from short branched, presumably young copies, highlighting that the majority of copies have recently integrated into the genome. The presence of a small number of ancient inserts (Figure 6) and deep internal branching within phylogenies show that most families are long term inhabitants of the genome. Therefore it would appear that the majority of transposable element insertions are deleterious in the *C. owczarzaki* genome and that natural selection operates efficiently to remove individual copies. However, recent work by Blumenstiel et al. (2014) showed that the young age of many transposable elements copies in *D. melanogaster* is consistent with recent bursts of transposition, rather than continual negative selection and element turnover. There therefore is a possibility that the skew towards young elements in *C. owczarzaki* is a result of bursts of transposition in the families present in the genome. Testing this alternative scenario will require the discovery of additional populations of *C. owczarzaki*, in order to determine the allele frequency spectrum of insertions. A current obstacle to population studies in *C. owczarzaki* is the lack of multiple cultures, as only one laboratory culture has been established since the discovery of the organism in 1977. Until additional populations are isolated, it will be difficult to distinguish constant element turnover and recent bursts of transposition for the families present in *C. owczarzaki*.

Recent work on *S. cerevisiae* (Carr et al. 2012) has shown that long branch, ancient copies are predominantly fixed in the population, whereas short branch, young copies of transposable elements tend to be polymorphic. If this observation holds for *C. owczarzaki*, the transposable element allele frequency spectrum can be predicted to be highly variable within the host population, since most copies appear to be young. Under this scenario, transposable elements can be expected to be a major source of genetic diversity in the *C. owczarzaki* population.

The use of RNASeq reads presented here gives an additional insight into the biology of protistan transposable elements. There is a strong positive correlation between expression level and the number of newly transposed elements. Whilst such a relationship may be expected, post-translational regulation (Matsuda and Garfinkel 2009) may uncouple the rates of element expression and transposition. Indeed the relationship between expression and transposition does not hold for all families, as the LTR retrotransposon *Cocv5* and the non-LTR retrotransposon *CoL4* have high expression levels yet possess relatively few recently integrated elements.

The stage-specific transcriptome data presented here indicates that there are sophisticated regulatory interactions between *C. owczarzaki* and its transposable element complement. The majority of families show a level of variation in expression between different stages, with 9 families showing significantly elevated expression in one stage in the life cycle. Six of the families show significantly elevated expression in the floating stage, a stage seen in the laboratory as a response to overcrowding in cultures. This elevated expression may be a result of the host cell being under stress, due to overcrowding, as numerous transposable element families have been shown to be activated by host stress (Capy et al. 2000). It seems unlikely that a stress response is the sole cause of elevated transposable element expression, as other families are more highly expressed in the amoeboid adherent and aggregate stages.

The survey completed here sheds further light on the evolution of transposable elements within holozoan protist genomes. Data from additional genomes are required before general trends, if indeed there are any across such a broad range of organisms, can be identified. The relative simplicity of sequencing small protist genomes compared to those of metazoans should, in the near future, provide insights into the inheritance and population dynamics of protistan transposable elements at both the species and genome level.

Acknowledgements

The authors are grateful to Richard Carr for statistical advice and Louise Johnson for comments on the manuscript. The authors are also grateful for the input and comments from Casey Bergman and two anonymous referees.

Literature Cited

Abascal F, Zardoya R, Posada D. 2005. ProtTest: Selection of best-fit models of protein evolution. Bioinformatics 21:2104-2105.

Bartolomé C, Bello X, Maside X. 2009. Widespread evidence for horizontal transfer of transposable elements across *Drosophila* genomes. Genome Biol 10:R22.

Belshaw R et al. 2005. Levels of insertional polymorphism in the human endogenous retrovirus family HERV-K (HML2): Implications for present-day activity. J Virol 79:12507-12514.

Bergman CM, Bensasson D. 2007. Recent LTR retrotransposition insertion contrasts with waves of non-LTR insertion since speciation in *Drosophila melanogaster*. Proc Natl Acad Sci USA 104:11340-11345.

Berriman M et al. 2009. The genome of the blood fluke *Schistosoma mansoni*. Nature 460:352-358.

Biémont C. 2009. Are transposable elements simply silenced or are they under house arrest? Trends Genet 25:333-334.

Biessmann H et al. 1992. HeT-A, a transposable element specifically involved in "healing" broken chromosome ends in *Drosophila melanogaster*. Mol Cell Biol 12:3910-3918.

Blumenstiel JP, Chen X, He M, Bergman CM. 2014. An age-of-allele test of neutrality for transposable element insertions. Genetics 196: 523-538.

Bowen NJ, McDonald JF. 2001. *Drosophila* euchromatic LTR retrotransposons are much younger than the host species in which they reside. Genome Res 11:1527-1540.

Brookfield JFY. 2005. The ecology of the genome – mobile DNA elements and their hosts. Nat Rev Genet 6:128-136.

Capy P, Gasperi G, Biémont C, Bazin C. 2000. Stress and transposable elements: co-evolution or useful parasites? Heredity 85:101-106.

Carr M, Baldauf SL. 2011. The protistan origins of animals and fungi. In: Pöggeler S, Wöstemeyer J, editors *The Mycota, Vol. XIV* Springer Press. p 3-23.

Carr M, Bensasson D, Bergman CM. 2012. Evolutionary genomics of transposable elements in *Saccharomyces cerevisiae*. PLoS One 7:e50978.

Carr M, Nelson M, Leadbeater BSC, Baldauf SL. 2008. Three families of LTR retrotransposons are present in the genome of the choanoflagellate *Monosiga brevicollis*. Protist 159:579-590.

Carr M, Soloway JR, Robinson TE, Brookfield JF. 2001. An investigation of the cause of low variability on the fourth chromosome of *Drosophila melanogaster*. Mol Biol Evol 18:2260-69.

Casola C, Lawing AM, Betrán E, Feschotte C. 2007. *PIF*-like transposons are common in Drosophila and have been repeatedly domesticated to generate new host genes. Mol Biol Evol 24:1872-1888.

Charlesworth B, Lapid A, Canada D. 1992. The distribution of transposable elements within and between chromosomes in a population of *Drosophila melanogaster*. I. Element frequencies and distribution. Genet Res 60:103-114.

Craig NL. 2002. Mobile DNA: an Introduction. In: Craig NL, editor *Mobile DNA II*. Washington, D.C.: ASM Press. p 3-11.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32:1792-1797.

Eanes WF, Wesley C, Hey J, Houle D, Ajioka JW. 1988. The fitness consequences of *P* element insertion in *Drosophila melanogaster*. Genet Res 52:17-26.

Finlay BJ, Fenchel T. 2004. Cosmopolitan metapopulations of free-living microbial eukaryotes. Protist 155:237-244.

Franchini LF, Ganko EW, McDonald JF. 2004. Retrotransposon-gene associations are widespread among *D. melanogaster* populations. Mol Biol Evol 21:1323-1331.

Gentleman RC et al. 2004. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5:R80

Gilbert C, Schaak S, Pace II JK, Brindley PJ, Feschotte C. 2010. A role for host-parasite interactions in the horizontal transfer of transposons across phyla. Nature 464:1347-1352.

Gorinšek B, Gubenšek F, Kordiš D. 2004. Evolutionary genomics of chromoviruses in eukaryotes. Mol Biol Evol 21:781-798.

Hellen EHB, Brookfield JFY. 2011. Investigation of the origin and spread of a mammalian transposable element based on current sequence diversity. J Mol Evol 73:287-296.

Jordan IK, McDonald JF. 1999. Tempo and mode of Ty element evolution in *Saccharomyces cerevisiae*. Genetics 151:1341-1351.

Jordan IK, Rogozin IB, Glazko GV, Koonin EV. 2003. Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends Genet 19:68-72.

Kordiš D. 2005. A genomic perspective on the chromodomain-containing retrotransposons: Chromoviruses. Gene 347:161-173.

Kuraku S, Qiu H, Meyer A (2012) Horizontal transfers of Tc1 elements between teleost fishes and their vertebrate parasite, lampreys. Genome Biol and Evol 4:929-936.

Lang BF, O'Kelly C, Nerad T, Gray MW, Burger G. 2002. The closest unicellular relatives of animals. Current Biology 12:1773-1778.

Librado P, Rozas J. 2009. DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25:1451-1452.

Linheiro RS, Bergman CM. 2012. Whole genome resequencing reveals natural target site preferences of transposable elements in *Drosophila melanogaster*. PLoS One 7:e30008.

Liu Y et al. 2009. Phylogenomic analyses predict sistergroup relationship of nucleariids and Fungi and paraphyly of zygomycetes with significant support. BMC Evolutionary Biology 9:272.

Malik HS, Eickbush TH. 1998. The RTE class of non-LTR retrotransposons is widely distributed in animals and is the origin of many SINEs. Mol Biol Evol 15:1123-1134.

Maside X, Bartolomé C, Charlesworth B. 2003. Inferences on the evolutionary history of the *S*-element family of *Drosophila melanogaster*. Mol Biol Evol 20:1183-1187.

Matsude E, Garfinkel DJ. 2009. Posttranslational interference of Ty1 retrotransposition by antisense RNAs. Proc Natl Acad Sci USA 106:15657-11662.

Nefedova LN, Mannanova MM, Kim AI. 2011. Integration specificity of LTR-retrotransposons and retroviruses in the *Drosophila melanogaster* genome. Virus Genes 42:297-306.

Nei M. 1987. Molecular Evolutionary Genetics. Columbia University Press, New York

Novikova O, Smyshlyaev G, Blinov A. (2010) Evolutionary genomics revealed interkingdom distribution of Tcn1-like chromodomain-containing Gypsy LTR retrotransposons among fungi and plants. BMC Genomics 11:231.

Owczarzak A, Stibbs HH, Bayne CJ. 1980. The destruction of *Schistosoma mansoni* mother sporocysts in vitro by amoebae isolated from *Biomphalaria glabrata*: an ultrastructural study. J Invertbr Pathol 35:26-33.

Pereira V. 2004. Insertion bias and purifying selection of retrotransposons in the *Arabidopsis thaliana* genome. Genome Biology 5:R79.

Peterson-Burch BD, Voytas DF. 2002. Genes of the Pseudoviridae (Ty1/copia retrotransposons). Mol Biol Evol 19:1832-1845.

Plasterk RHA, Izsvák Z, Ivics Z. 1999. Resident aliens: the Tc1/*mariner* superfamily of transposable elements. Trends Genet 15:326-332.

Robertson HM, Lampe DJ. 1995. Recent horizontal transfer of a *mariner* transposable element among and between Diptera and Neuroptera. Mol Biol Evol 12:850-862.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19:1572 -1574.

Ruiz-Trillo I et al. 2008a. The origins of multicellularity: a multi-taxon genome initiative. Trends Genet 23:113-118.

Ruiz-Trillo I, Lane CE, Archibald JM, Roger AJ. 2006. Insights into the evolutionary origin and genome architecture of the unicellular opisthokonts *Capsaspora owczarzaki* and *Sphaeroforma arctica*. J Eukaryot Microbiol 53: 379-384.

Ruiz-Trillo I, Roger AJ, Burger G, Gray MW, Lang BF. 2008b. A phylogenomic investigation into the origin of Metazoa. Mol Biol Evol 25:664-672.

Sánchez-Gracia A, Maside X, Charlesworth B. 2005. High rate of horizontal transfer of transposable elements in *Drosophila*. Trends Genet 21:200-203.

Schlenke TA, Begun DJ. 2004. Strong selective sweep associated with a transposon insertion in *Drosophila simulans*. Proc Natl Acad Sci USA 101:1626-1631.

Sebé-Pedrós A et al. 2011. Unexpected repertoire of metazoan transcription factors in the unicellular holozoan *Capsaspora owczarzaki* Mol Biol Evol 28:1241-1254.

Sebé-Pedrós A, Irimia M, del Campo J, Parra-Acero H, Russ C, Nusbaum C, Blencowe BJ, Ruiz-Trillo I. 2013. Regulated aggregative multicellularity in a close unicellular relative of metazoa. eLife 2:e01287.

Shalchian-Tabrizi K et al. 2008. Multigene phylogeny of Choanozoa and the origin of animals. PLoS One 3:e2098.

Silvestro M. 2012. raxmlGUI: a graphical front-end for RAxML. Org Divers Evol 12:335-337.

Smit AFA, Riggs AD. 1996. *Tiggers* and other DNA transposon fossils in the human genome. Proc Natl Acad Sci USA 93:1443-1448.

Sohn IG, Kornicker LS. 1972. Predation of schistosomiasis vector snails by Ostracoda (Crustacea). Science 175:1258-1259.

Stamatakis A, Ludwig T, Meier H. 2005. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 21:456-463.

Stibbs HH, Owczarzak A, Bayne CJ, DeWan P. 1979. Schistosome sporocyst-killing amoebae isolated from *Biomphalaria glabrata*. J Invertebr Pathol 33:159-170.

Suga H et al. 2013. The *Capsaspora* genome reveals a complex unicellular prehistory of animals. Nat Commun 4:2325.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:393-407.

Torruella G et al. 2012. Phylogenetic relationships within the Opisthokonta based on phylogenomic analyses of conserved single copy protein domains. Mol Biol Evol. 29:531-544.

Xu J et al. 2006. The varying microsporidian genome: Existence of long-terminal repeat retrotransposon in domesticated silkworm parasite *Nosema bombycis*. Int J Parasitol 36:1049-1056.

Yoshiyama M et al. 2001. Possible horizontal transfer of a transposable element from host to parasitoid. Mol Biol Evol 18:1952-1958.

Zou S, Wright DA, Voytas DF. 1995. The *Saccharomyces* Ty5 retrotransposon family is associated with origins of DNA replication at the telomeres and the silent mating locus *HMR*. Proc Natl Acad Sci USA 92:920-924.

Figure Legends

Figure 1. Maximum likelihood phylogeny of chromoviral Pol amino acid sequences. The phylogeny was constructed from 630 aligned amino acid positions using the PROTCAT model with the LG substitution matrix. Values for maximum likelihood bootstrap percentages (mlBP) and Bayes posterior probabilities (biPP) are shown above and below the branches respectively. 100% mlBP and 1.00 biPP are both denoted by '*'. Values <50%mlBP or <0.70 biPP are denoted by "-". *C. owczarzaki* proteins are written in red bold font, metazoan proteins are written in dark blue, choanoflagellate proteins in light blue, fungal proteins in brown, plant proteins are in dark green and amoebozoan proteins are written in purple. The scale bar represents the number of amino acid substitutions per site.

Figure 2. Maximum likelihood phylogeny of *L1* Pol amino acid sequences. The phylogeny was constructed from 393 aligned amino acid positions using the PROTCAT model with the LG substitution matrix. Excavate proteins are in pink font, the formatting of the tree and labels is otherwise the same as in Figure 1.

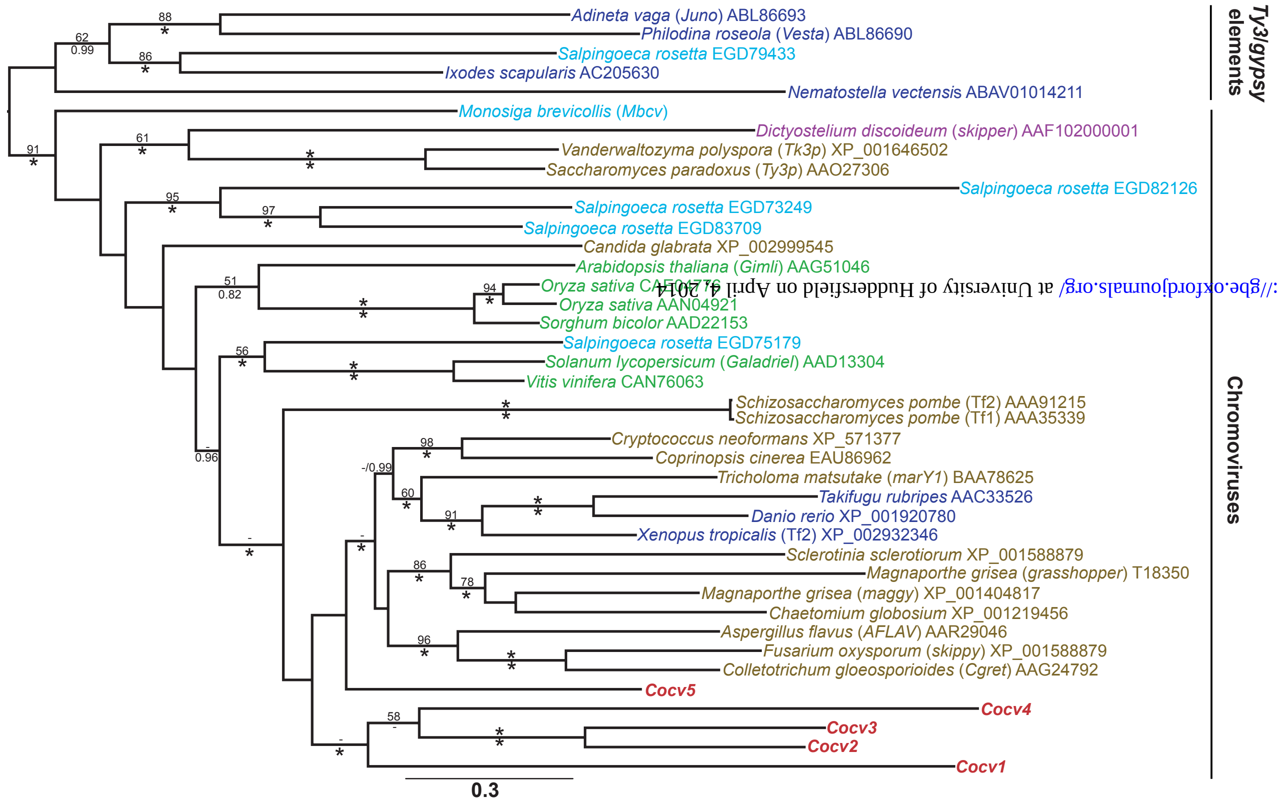
Figure 3. Maximum likelihood phylogeny of *pogo* Tnpase amino acid sequences. The phylogeny was constructed from 199 aligned amino acid positions using the PROTCAT model with the LG substitution matrix. Chromalveolate proteins in written in orange front, the formatting of the tree and labels is otherwise the same as in Figure 2.

Figure 4. *C. owczarzaki* transposable element expression. The expression levels of transposable elements were approximated by the normalized number of RNAseq reads and shown as a heatmap. (A) Values were normalized by the largest expression level,

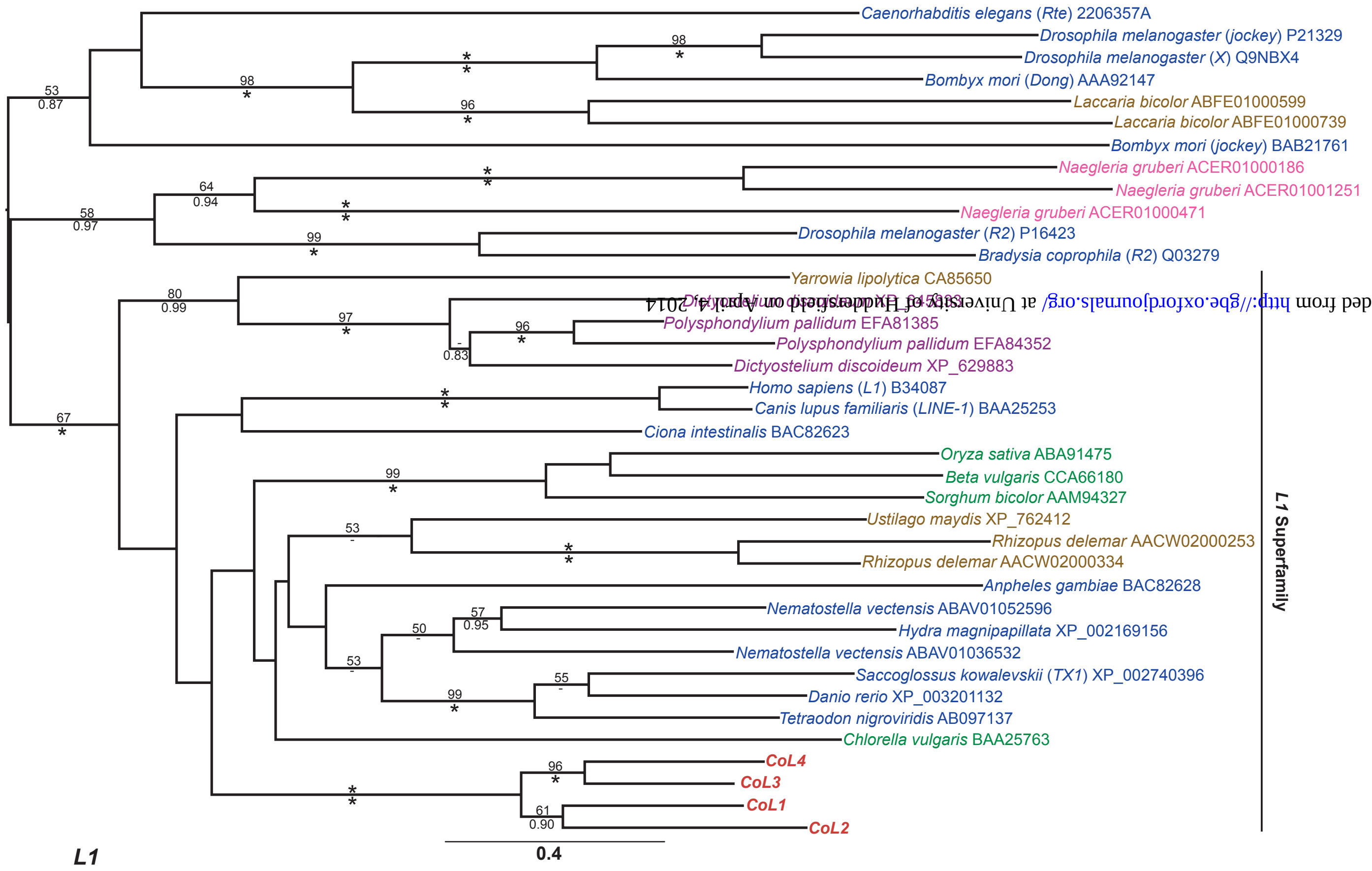
which is that of *Cocv1* in adherent cells. (B) The colour map is presented in log scale (values normalized to 0 - 1.0 range).

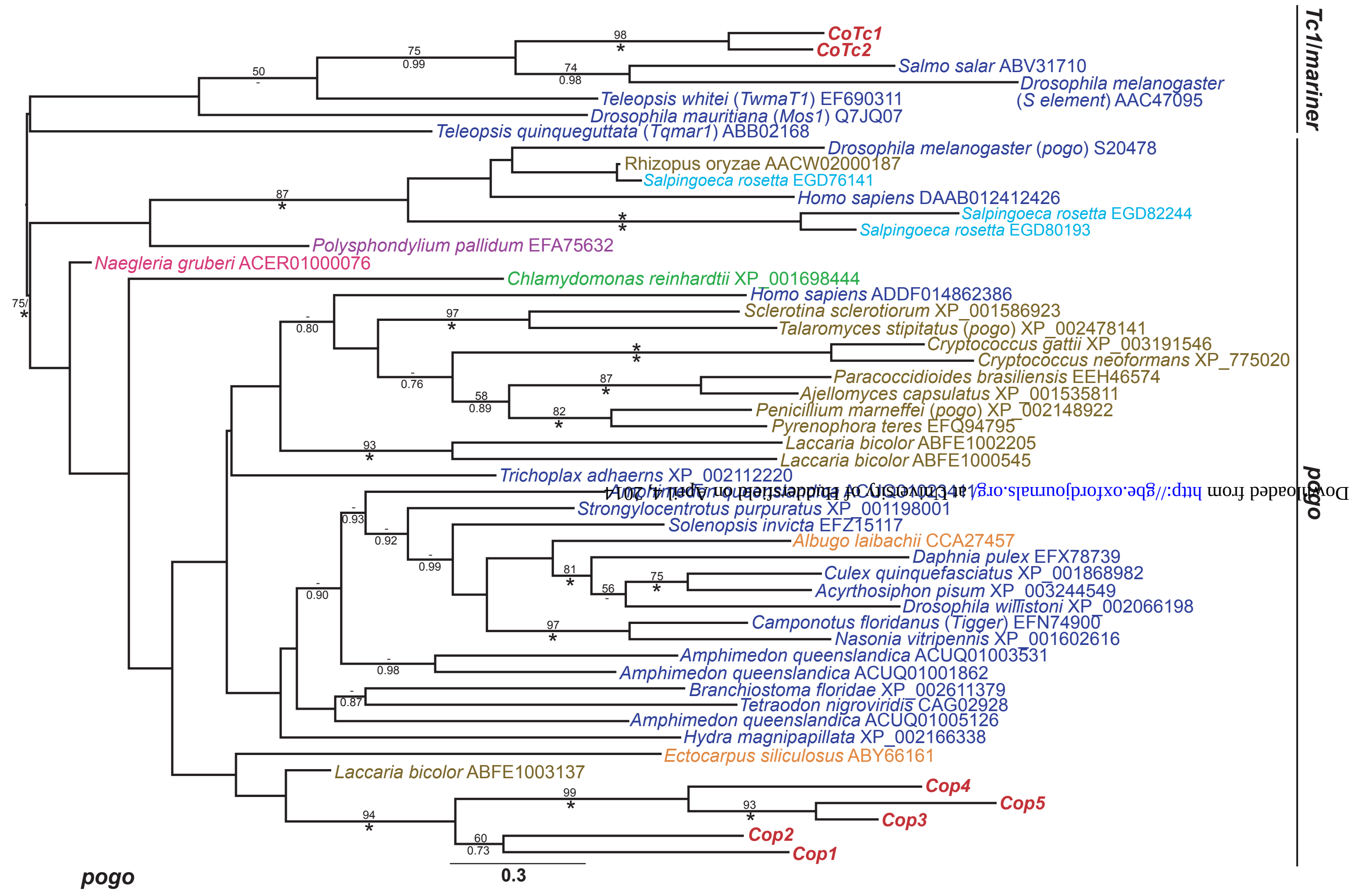
Figure 5. Maximum likelihood phylogenetic trees of individual element copies of *Cocv1*. The phylogeny was constructed from 171 aligned nucleotide positions using the GTRCAT model. Values for bootstrap percentages and Bayes posterior probabilities are shown above and below the branches respectively. 100% mlBP and 1.00 biPP are both denoted by '*'. Values <50%mlBP or <0.70 biPP are denoted by "-". 5' and 3' LTR sequences are shown in blue, solo LTR sequences are written in red. Terminal nodes are labelled with the flanking DNA sequence of the insert. The scale bar represents the number of nucleotide substitutions per site.

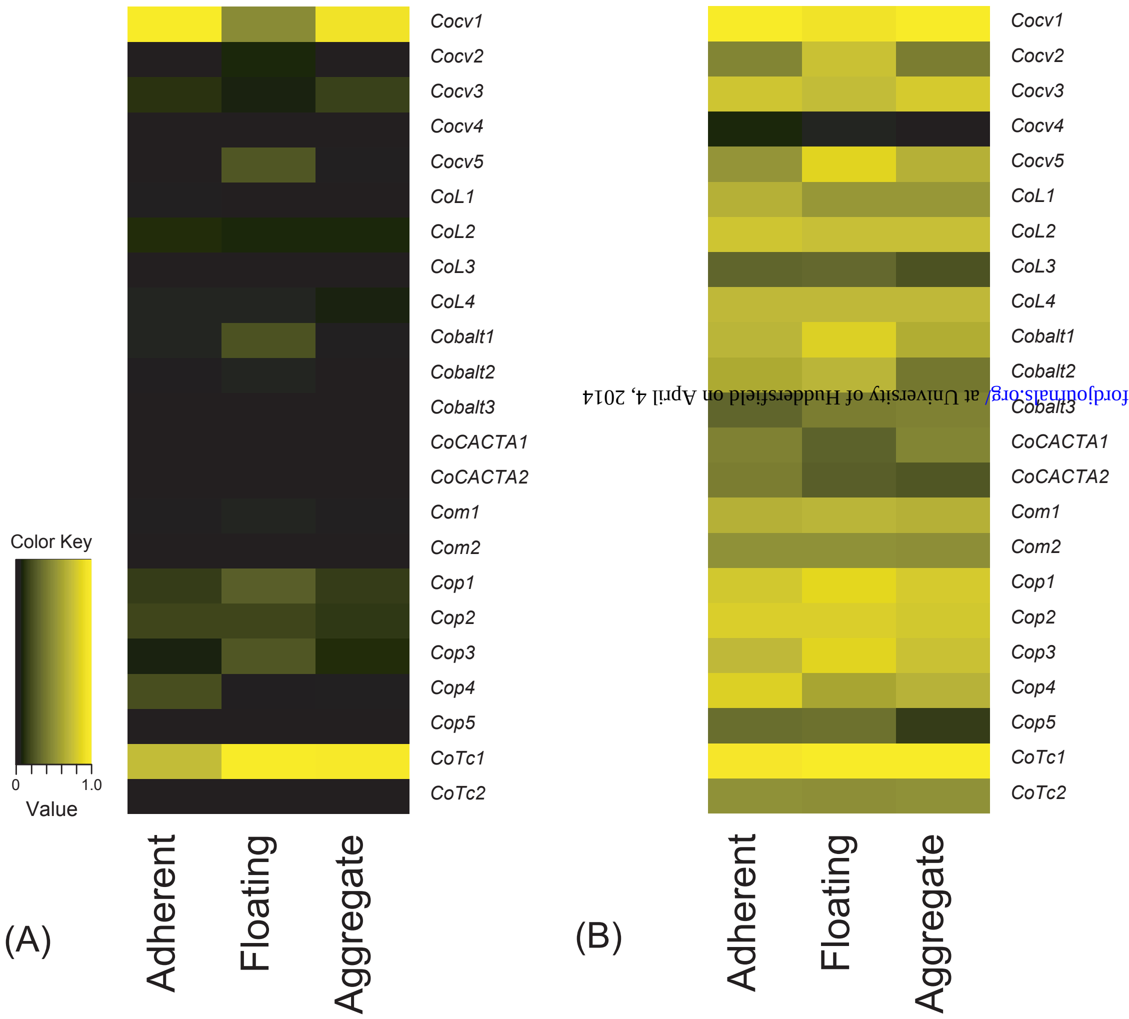
Figure 6. Terminal branch lengths of the 22 multicopy transposable element families in the *C. owczarzaki* genome. LTR retrotransposon, Non-LTR retrotransposon and DNA transposon families are represented by red, blue and green boxes respectively. Branch lengths for full length LTR retrotransposons were taken from the 5' LTR when this was present in the phylogeny; in its absence the 3' LTR was used. The filled boxes denote the interquartile range and the horizontal dark line represents the median branch length. The whiskers highlight 1.5 times the interquartile range from the median and the asterisks represent branch lengths outside this range.

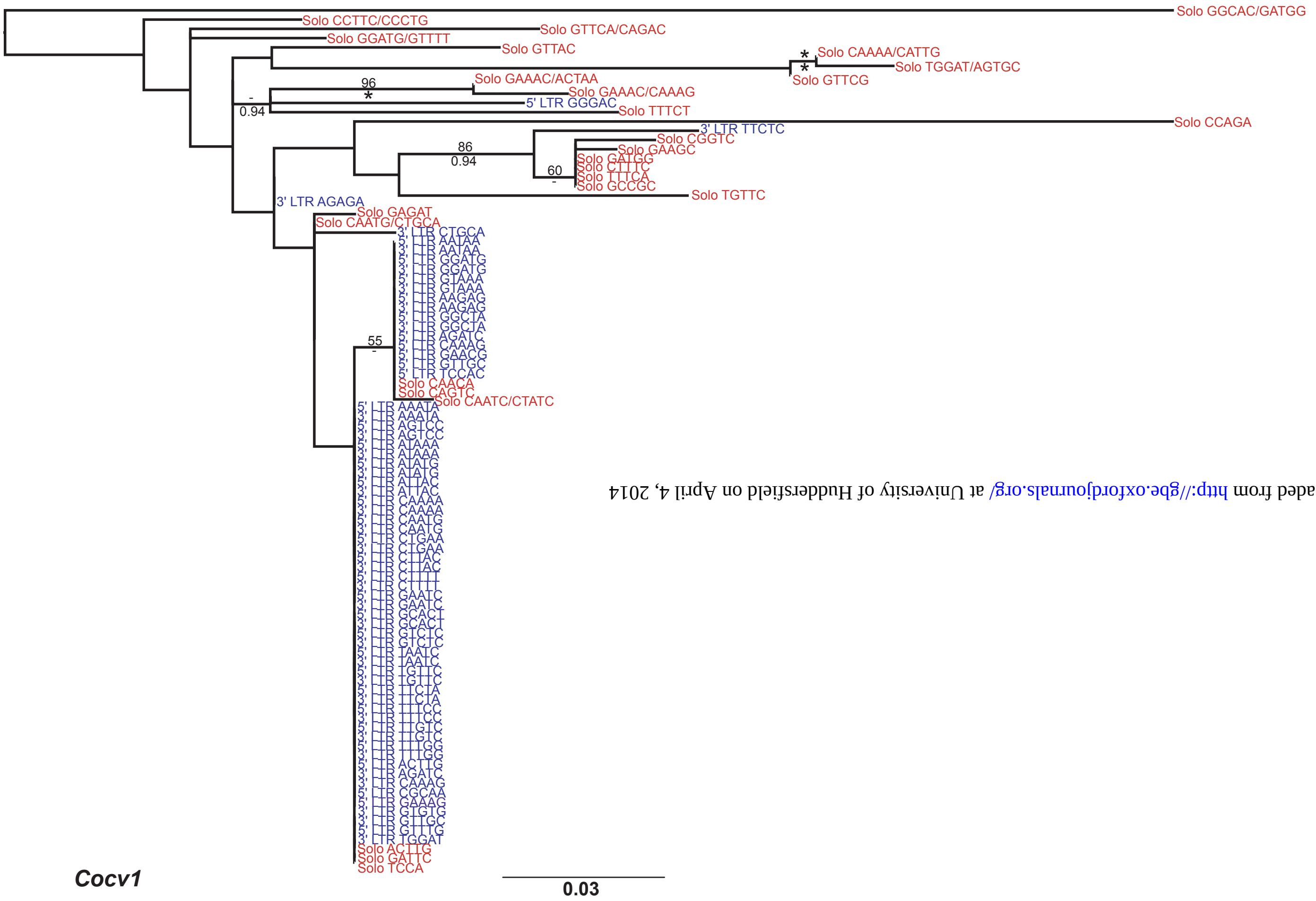

Adineta vaga (Juno) ABL86693
Philodina roseola (Vesta) ABL86690
Salpingoeca rosetta EGD79433
Ixodes scapularis AC205630
Nematostella vectensis ABAV01014211
Monosiga brevicollis (Mbcv)
Dictyostelium discoideum (skipper) AAF102000001
Vanderwaltozyma polyspora (Tk3p) XP_001646502
Saccharomyces paradoxus (Ty3p) AAO27306
Salpingoeca rosetta EGD82126
Salpingoeca rosetta EGD73249
Salpingoeca rosetta EGD83709
Candida glabrata XP_002999545
Arabidopsis thaliana (Gimli) AAG51046
Oryza sativa CAE04776
Oryza sativa AAN04921
Sorghum bicolor AAD22153
Salpingoeca rosetta EGD75179
Solanum lycopersicum (Galadriel) AAD13304
Vitis vinifera CAN76063
Schizosaccharomyces pombe (Tf2) AAA91215
Schizosaccharomyces pombe (Tf1) AAA35339
Cryptococcus neoformans XP_571377
Coprinopsis cinerea EAU86962
Tricholoma matsutake (marY1) BAA78625
Takifugu rubripes AAC33526
Danio rerio XP_001920780
Xenopus tropicalis (Tf2) XP_002932346
Sclerotinia sclerotiorum XP_001588879
Magnaporthe grisea (grasshopper) T18350
Magnaporthe grisea (maggy) XP_001404817
Chaetomium globosium XP_001219456
Aspergillus flavus (AFLAV) AAR29046
Fusarium oxysporum (skippy) XP_001588879
Colletotrichum gloeosporioides (Cgret) AAG24792
Cocv5
Cocv4
Cocv3
Cocv2
Cocv1
88
62
0.99
86
91
61
95
97
51
0.82
94
56
-
0.96
-/0.99
98
60
91
-
86
78
96
58
0.3
Ty3/gypsy elements
Chromoviruses

Chromoviruses


Caenorhabditis elegans (Rte) 2206357A
Drosophila melanogaster (jockey) P21329
Drosophila melanogaster (X) Q9NBX4
Bombyx mori (Dong) AAA92147
Laccaria bicolor ABFE01000599
Laccaria bicolor ABFE01000739
Bombyx mori (jockey) BAB21761
Naegleria gruberi ACER01000186
Naegleria gruberi ACER01001251
Naegleria gruberi ACER01000471
Drosophila melanogaster (R2) P16423
Bradysia coprophila (R2) Q03279
Yarrowia lipolytica CA85650
Dictyostelium discoideum XP_645833
Polysphondylium pallidum EFA81385
Polysphondylium pallidum EFA84352
Dictyostelium discoideum XP_629883
Homo sapiens (L1) B34087
Canis lupus familiaris (LINE-1) BAA25253
Ciona intestinalis BAC82623
Oryza sativa ABA91475
Beta vulgaris CCA66180
Sorghum bicolor AAM94327
Ustilago maydis XP_762412
Rhizopus delemar AACW02000253
Rhizopus delemar AACW02000334
Anpheles gambiae BAC82628
Nematostella vectensis ABAV01052596
Hydra magnipapillata XP_002169156
Nematostella vectensis ABAV01036532
Saccoglossus kowalevskii (TX1) XP_002740396
Danio rerio XP_003201132
Tetraodon nigroviridis AB097137
Chlorella vulgaris BAA25763
CoL4
CoL3
CoL1
CoL2
53
0.87
98
*
*
*
98
*
96
*
64
0.94
*
*
*
*
58
0.97
99
*
80
0.99
97
*
96
*
-
0.83
*
*
67
*
99
*
53
-
*
*
57
0.95
50
-
53
-
55
-
99
*
96
*
*
*
61
0.90
L1 Superfamily
0.4
L1


Tc1/mariner
pogo
CoTc1
CoTc2
Salmo salar ABV31710
Drosophila melanogaster (S element) AAC47095
Teleopsis whitei (TwmaT1) EF690311
Drosophila mauritiana (Mos1) Q7JQ07
Teleopsis quinqueguttata (Tqmar1) ABB02168
Drosophila melanogaster (pogo) S20478
Rhizopus oryzae AACW02000187
Salpingoeca rosetta EGD76141
Homo sapiens DAAB012412426
Salpingoeca rosetta EGD82244
Salpingoeca rosetta EGD80193
Polysphondylium pallidum EFA75632
Naegleria gruberi ACER01000076
Chlamydomonas reinhardtii XP_001698444
Homo sapiens ADDF014862386
Sclerotina sclerotiorum XP_001586923
Talaromyces stipitatus (pogo) XP_002478141
Cryptococcus gattii XP_003191546
Cryptococcus neoformans XP_775020
Paracoccidioides brasiliensis EEH46574
Ajellomyces capsulatus XP_001535811
Penicillium marneffei (pogo) XP_002148922
Pyrenophora teres EFQ94795
Laccaria bicolor ABFE1002205
Laccaria bicolor ABFE1000545
Trichoplax adhaerns XP_002112220
Amphimedon queenslandica ACUQ01002341
Strongylocentrotus purpuratus XP_001198001
Solenopsis invicta EFZ15117
Albugo laibachii CCA27457
Daphnia pulex EFX78739
Culex quinquefasciatus XP_001868982
Acyrthosiphon pisum XP_003244549
Drosophila willistoni XP_002066198
Camponotus floridanus (Tigger) EFN74900
Nasonia vitripennis XP_001602616
Amphimedon queenslandica ACUQ01003531
Amphimedon queenslandica ACUQ01001862
Branchiostoma floridae XP_002611379
Tetraodon nigroviridis CAG02928
Amphimedon queenslandica ACUQ01005126
Hydra magnipapillata XP_002166338
Ectocarpus siliculosus ABY66161
Laccaria bicolor ABFE1003137
Cop4
Cop5
Cop3
Cop2
Cop1
98
*
75
0.99
74
0.98
50
-
87
*
*
*
75/
*
-
0.80
97
*
*
*
-
0.76
87
*
58
0.89
82
*
93
*
-
0.93
-
0.92
-
0.99
81
*
56
-
75
*
97
*
-
0.90
-
0.98
-
0.87
99
*
93
*
94
*
60
0.73
0.3
pogo


Cocv1
Cocv2
Cocv3
Cocv4
Cocv5
CoL1
CoL2
CoL3
CoL4
Cobalt1
Cobalt2
Cobalt3
CoCACTA1
CoCACTA2
Com1
Com2
Cop1
Cop2
Cop3
Cop4
Cop5
CoTc1
CoTc2
Color Key
0
1.0
Value
(A)
Adherent
Floating
Aggregate
(B)


Solo GGCAC/GATGG
Solo CCTTC/CCCTG
Solo GTTCA/CAGAC
Solo GGATG/GTTTT
Solo GTTAC
*
Solo CAAAA/CATTG
*
Solo TGGAT/AGTGC
Solo GTTCG
96
*
Solo GAAAC/ACTAA
Solo GAAAC/CAAAG
-
0.94
5' LTR GGGAC
Solo TTTCT
Solo CCAGA
3' LTR TTCTC
Solo CGGTC
86
0.94
Solo GAAGC
Solo GATGG
60
-
Solo CTTTC
Solo TTTCA
Solo GCCGC
Solo TGTTC
3' LTR AGAGA
Solo GAGAT
Solo CAATG/CTGCA
3' LTR CTGCA
5' LTR AATAA
3' LTR AATAA
5' LTR GGATG
3' LTR GGATG
5' LTR GTAAA
3' LTR GTAAA
5' LTR AAGAG
3' LTR AAGAG
5' LTR GGCTA
3' LTR GGCTA
5' LTR AGATC
55
-
5' LTR CAAAG
5' LTR GAACG
5' LTR GTTGC
5' LTR TCCAC
Solo CAACA
Solo CAGTC
Solo CAATC/CTATC
5' LTR AAATA
3' LTR AAATA
5' LTR AGTCC
3' LTR AGTCC
5' LTR ATAAA
3' LTR ATAAA
5' LTR ATATG
3' LTR ATATG
5' LTR ATTAC
3' LTR ATTAC
5' LTR CAAAA
3' LTR CAAAA
5' LTR CAATG
3' LTR CAATG
5' LTR CTGAA
3' LTR CTGAA
5' LTR CTTAC
3' LTR CTTAC
5' LTR CTTTT
3' LTR CTTTT
5' LTR GAATC
3' LTR GAATC
5' LTR GCACT
3' LTR GCACT
5' LTR GTCTC
3' LTR GTCTC
5' LTR TAATC
3' LTR TAATC
5' LTR TGTTC
3' LTR TGTTC
5' LTR TTCTA
3' LTR TTCTA
5' LTR TTTCC
3' LTR TTTCC
5' LTR TTGTC
3' LTR TTGTC
5' LTR TTTGG
3' LTR TTTGG
5' LTR ACTTG
3' LTR AGATC
3' LTR CAAAG
5' LTR CGCAA
5' LTR GAAAG
3' LTR GTGTG
3' LTR GTTGC
5' LTR GTTTG
3' LTR TGGAT
Solo ACTTG
Solo GATTC
Solo TCCA
Cocv1
0.03

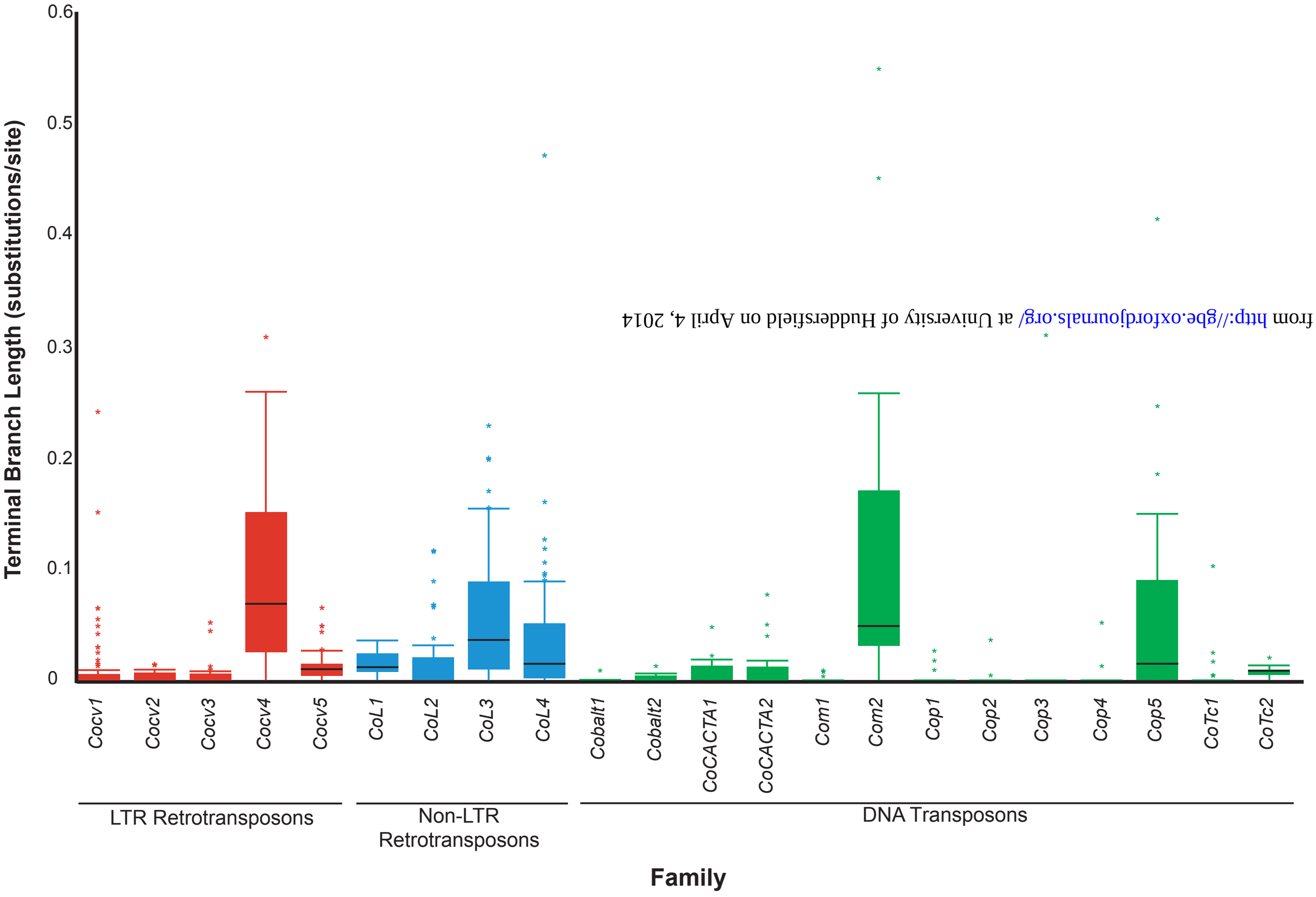

Terminal Branch Length (substitutions/site)
0.6
0.5
0.4
0.3
0.2
0.1
0
Cocv1
Cocv2
Cocv3
Cocv4
Cocv5
CoL1
CoL2
CoL3
CoL4
Cobalt1
Cobalt2
CoCACTA1
CoCACTA2
Com1
Com2
Cop1
Cop2
Cop3
Cop4
Cop5
CoTc1
CoTc2
LTR Retrotransposons
Non-LTR Retrotransposons
DNA Transposons
Family

Table 1. Characterization of the 5 identified families of LTR retrotransposon in the genome of *C. owczarzaki*.

Family	Copy Number (FLE/Solo)	No. of RNASeq Reads	No. of Identical Paralogous Copies	Intra-element LTR identity range (%)	Total LTR Diversity (π)	FLE LTR / Solo LTR Diversity (π)	Tajima's D^a
Cocv1	64 (39/25)	437,027	47	99.3-100	0.075	0.011/0.160	-2.439**
Cocv2	33 (23/10)	22,365	21	99.3-100	0.008	0.007/0.009	-2.396**
Cocv3	26 (16/10)	72,753	15	100	0.021	0.012/0.033	-2.118*
Cocv4	17 (1/16)	80	2	100	0.174	0.000/0.184	-2.103*
Cocv5	22 (14/8)	62,990	4	99.7-100	0.042	0.043/0.042	-1.526

a: Significance levels: *<0.05, **<0.01, ***<0.001

Table 2. Characterization of the 4 identified families of non-LTR retrotransposon in the genome of *C. owczarzaki*.

Family	Observed Copy Number	No. of RNASeq Reads	No. of Identical Paralogous Copies	Non-coding Diversity 3' UTR (π)	Tajima's *D*[a]
CoL1	12	15,275	0	0.048	-0.183
CoL2	51	63,284	16	0.069	-1.478
CoL3	30	777	0	0.164	-1.925*
CoL4	47	49,002	5	0.102	-2.186**

a: Significance levels: *<0.05, **<0.01, ***<0.001

Table 3. Characterization of the 14 identified families of DNA transposon in the genome of *C. owczarzaki*.

Family	Observed Copy Number (5'/3' ITR)	No. of RNASeq Reads	No. of Identical Paralogous Copies	Non-coding Diversity 5' ITR+UTR (π)	Tajima's *D*[a]
Cobalt1	17-34 (17/17)	29,065	16	0.013	0.326
Cobalt2	9-18 (9/9)	8,902	4	0.032	0.262
Cobalt3	1	805	0	-	-
CoCACTA1	18 (9/18)	2,629	4	0.028	0.210
CoCACTA2	35 (20/24)	1,344	9	0.086	n/a
Com1	19 (15/16)	26,930	12	0.015	1.669
Com2	12 (8/5)	4,941	0	0.197	-1.297
Cop1	28-51 (23/28)	65,112	20	0.004	-2.078*
Cop2	25-44 (19/25)	56,126	17	0.081	1.135
Cop3	15-27 (12/15)	46,819	11	0.047	-2.003*
Cop4	19-37 (19/18)	32,779	17	0.011	-1.405
Cop5	16-31 (16/15)	457	5	0.171	-1.518
CoTc1	42-83 (41/42)	164,264	34	0.007	-2.669***
CoTc2	8-14 (6/8)	1,566	0	0.022	-1.124

a: Significance levels: *<0.05, **<0.01, ***<0.001