Unearthing LTR Retrotransposon *gag* Genes Co-opted in the Deep Evolution of Eukaryotes

Jianhua Wang and Guan-Zhu Han*

Jiangsu Key Laboratory for Microbes and Functional Genomics, College of Life Sciences, Nanjing Normal University, Nanjing, Jiangsu, China

***Corresponding author:** E-mail: guanzhu@njnu.edu.cn.

Associate editor: Irina Arkhipova

Abstract

LTR retrotransposons comprise a major component of the genomes of eukaryotes. On occasion, retrotransposon genes can be recruited by their hosts for diverse functions, a process formally referred to as co-option. However, a comprehensive picture of LTR retrotransposon *gag* gene co-option in eukaryotes is still lacking, with several documented cases exclusively involving Ty3/Gypsy retrotransposons in animals. Here, we use a phylogenomic approach to systemically unearth co-option of retrotransposon *gag* genes above the family level of taxonomy in 2,011 eukaryotes, namely co-option occurring during the deep evolution of eukaryotes. We identify a total of 14 independent *gag* gene co-option events across more than 740 eukaryote families, eight of which have not been reported previously. Among these retrotransposon *gag* gene co-option events, nine, four, and one involve *gag* genes of Ty3/Gypsy, Ty1/Copia, and Bel-Pao retrotransposons, respectively. Seven, four, and three co-option events occurred in animals, plants, and fungi, respectively. Interestingly, two co-option events took place in the early evolution of angiosperms. Both selective pressure and gene expression analyses further support that these co-opted *gag* genes might perform diverse cellular functions in their hosts, and several co-opted *gag* genes might be subject to positive selection. Taken together, our results provide a comprehensive picture of LTR retrotransposon *gag* gene co-option events that occurred during the deep evolution of eukaryotes and suggest paucity of LTR retrotransposon *gag* gene co-option during the deep evolution of eukaryotes.

Key words: **LTR retrotransposon, co-option, phylogenetics.**

Introduction

Transposable elements (TEs), generally thought to be genomic parasites, are major components of the eukaryote genomes (Brandt, et al. 2005; Wicker, et al. 2007; Etchegaray, et al. 2021); for instance, ~45% of the human genome and ~85% of the maize genome are comprised of various TEs (Schnable et al. 2009; Lander et al. 2012). Based on their transposition mechanisms, TEs are typically classified into two major classes, class I (retrotransposons) and class II (DNA transposons) (Wicker et al. 2007). Among retrotransposons, long terminal repeat (LTR) retrotransposons characterized by the presence of LTRs at 5′- and 3′-termini encode two common genes, *gag* and *pol*, required for retrotransposition (Llorens et al. 2009; Naville et al. 2016; Sanchez et al. 2017). LTR retrotransposons can be further divided into several superfamilies, including Ty3/Gypsy, Ty1/Copia, and Bel-Pao retrotransposons as well as retroviruses/endogenous retroviruses (ERVs) (Wicker et al. 2007).

Most of LTR retrotransposons have been thought to be neutral or deleterious, and are removed by recombination between LTRs or inactivated and degraded by accumulating disruptive mutations (Stoye 2012; Jangam et al. 2017; Johnson 2019; Etchegaray et al. 2021). On occasion, coding or regulatory regions of LTR retrotransposons can be repurposed for diverse cellular functions in hosts, a process formally termed as co-option, domestication, or exaptation (Feschotte 2008; Kokošar and Kordiš 2013; Hoen and Bureau 2015; Chuong et al. 2016; Naville et al. 2016; Chuong et al. 2017; Jangam et al. 2017; Wang et al. 2019; Wang and Han 2020; Etchegaray et al. 2021). To date, more than 100 independent retroviral *gag* gene co-option events have been documented in literature (Campillos et al. 2006; Pastuzyn et al. 2018; Skirmuntt and Katzourakis 2019; Wang et al. 2019; Wang and Han 2020), as exemplified by *Fv1* gene that serves as a restriction factor to inhibit the replication of diverse retroviruses (Best et al. 1996; Yap et al. 2014; Boso et al. 2018). In contrast, only six cases of co-opted LTR retrotransposon *gag* gene occurred above the taxonomic family level have been identified (Campillos et al. 2006; Pastuzyn et al. 2018), all of which involve Ty3/Gypsy retrotransposons. 1) Activity-regulated cytoskeleton-associated proteins (Arc) in tetrapods and 2) dArc proteins in schizophoran flies were derived from *gag* genes of distinct Ty3/Gypsy retrotransposon lineages, and mediate intercellular RNA transfer in the nervous system (Ashley et al. 2018; Pastuzyn et al. 2018). 3) *Sushi-ichi retrotransposon homolog* (*SIRH*/*RTL*) family arose from *gag* gene of

Article

a sushi LTR retrotransposon, and two members of this family, *PEG10/RTL2* and *PEG11/RTL1*, are essential to the placenta development (Ono et al. 2006; Sekita et al. 2008). 4) *Paraneoplastic Ma antigen* (*PNMA*) gene family originated from Gypsy12_DR *gag* gene and might perform diverse functions; for example, *PNMA5* and *PNMA10* are associated with brain development (Takaji et al. 2009; Cho et al. 2011; Pang et al. 2018). 5) *SCAN* (*SRE-ZBP, CTfin-51, AW-1, Number 18 cDNA*) gene family was derived from C-terminal of Grm1-like LTR retrotransposon capsid (CA). Members of SCAN domains are usually associated with $(C_2H_2)_x$-type zinc fingers and Kruppel-associated box domains to form transcription factors, and function in many aspects of cell differentiation and development, for instance, regulating the transcription of growth factors (Collins et al. 2001; Sander et al. 2003; Edelstein and Collins 2005; Emerson and Thomas 2011). There were also sporadic documented LTR retrotransposon *gag* gene co-option events that occurred within the taxonomic family level; for example, *Gagr* is a Ty3/Gypsy retrotransposon *gag* gene co-opted within *Drosophila* (Nefedova et al. 2014; Makhnovskii et al. 2020).

Besides Ty3/Gypsy retrotransposon *gag* genes, various coding and regulatory regions of TEs can be repurposed for cellular functions (Chuong et al. 2017; Jangam et al. 2017; Etchegaray et al. 2021). RAG1 and RAG2 proteins, which are essential for the rearrangement of antigen receptors in vertebrates, originated through co-opting a DNA transposon known as ProtoRAG (Huang et al. 2016; Morales Poole et al. 2017). In angiosperms, several cases of co-option of class II TE transposase have been documented: *FAR1-related sequence* (*FRS*) and *MUSTANG* (*MUG*) gene families were derived from transposases of Mutator-like elements (MULEs), and *SLEEPER* gene family arose from *hAT* transposase (Oliver et al. 2013; Hoen and Bureau 2015; Joly-Lopez et al. 2016). *FRS* genes (e.g., *FHY3* and *FAR1*) perform diverse functions in plants, including acting as a light signal transducer, regulating the flowering time, and being involved in the division of chloroplasts (Lin et al. 2007; Wang and Wang 2015; Ma and Li 2018). *MUG* gene family plays crucial roles in plant growth, flowering time, and floral organ development (Joly-Lopez et al. 2012). *SLEEPER* genes regulate global gene expression and are crucial for the growth of plants (Bundock and Hooykaas 2005; Knip et al. 2012). In fission yeast, CENP-Bs (Cbh1, Cbh2, and Abp1) that were derived from transposases of pogo DNA transposons contribute to the silence of *Tf* retrotransposons (Cam et al. 2008). Moreover, *cis*-regulatory sequences of TEs can also be co-opted, shaping the evolution of host gene regulatory networks (Chuong et al. 2017).

To date, a comprehensive picture of LTR retrotransposon *gag* gene co-option in eukaryotes is still lacking. Little is known about the extent and diversity of LTR retrotransposon *gag* gene co-option in eukaryotes. In this study, we performed a comprehensive phylogenomic analysis to unearth LTR retrotransposon *gag* gene co-option events (RtGCEs) above the taxonomic family level across eukaryotes. We identified a total of 14 RtGCEs, seven, four, and three of which occurred in animals, plants, and fungi, respectively. We also analyzed the evolutionary history, expression pattern, and selective pressure for each co-opted LTR retrotransposon *gag* (*Crtg*) gene. Our study provides a snapshot of LTR retrotransposon *gag* gene co-option that occurred during the deep evolutionary history of eukaryotes.

Results

Mining Deep LTR Retrotransposon *gag* Gene Co-option in Eukaryotes

We used a similarity search and phylogenetic analysis combined approach to systematically identify *Crtg* genes above the taxonomic family level in eukaryotes (see Methods and Materials for the details; Wang and Han 2020). Our analyses included a total of 2,011 annotated proteomes of eukaryotes, which cover at least 743 families in eukaryotes (supplementary table S1, Supplementary Material online). The *Crtg* genes identified in this study fulfill two criteria: 1) The *Crtg* genes share similar synteny among the genomes of species across at least two families, and/or the *Crtg* phylogeny largely agree with the host phylogeny; and 2) the *Crtg* genes are subject to certain level of natural selection, implying potential cellular functionality (Graur et al. 2013; Jangam et al. 2017; Wang and Han 2020). Following these two criteria, we identified a total of 14 *Crtg* genes, referred to as *Crtg1* to *Crtg14* (figs. 1–5), seven (*Crtg1* to *Crtg7*), four (*Crtg8* to *Crtg11*), and three (*Crtg12* to *Crtg14*) of which were identified in the genomes of animals, plants, and fungi, respectively. Although six of these *Crtg* genes have been documented in animals, namely *Arc* (*Crtg2*), *dArc* (*Crtg3*), *RTL* (*Crtg4* and *Crtg5*), *PNMA* (*Crtg6*), and *SCAN* (*Crtg7*), the remaining eight *Crtg* genes were first reported in this study. It should be noted that the co-option events identified here represent the ones that occurred above the family level of taxonomy and, in general, during the deep evolution of eukaryotes.

To decipher the source of *Crtg* genes, we identified the LTR retrotransposons that were closely related to these *Crtg* genes. Phylogenetic analyses of reverse transcriptase (RT) suggest that nine (*Crtg1* to *Crtg7* in animals, *Crtg8* and *Crtg11* in plants), four (*Crtg9* and *Crtg10* in plants, *Crtg12* and *Crtg13* in fungi), and one (*Crtg14* in fungi) *Crtg*-related retrotransposons belong to Ty3/Gypsy, Ty1/Copia, and Bel-Pao retrotransposons, respectively (fig. 1). Whereas all the documented cases of retrotransposon *gag* gene co-option were derived from Ty3/Gypsy retrotransposons, our findings indicate that *gag* genes of all the three major LTR retrotransposon superfamilies (Ty3/Gypsy, Ty1/Copia, and Bel-Pao) could be co-opted during the evolutionary course of eukaryotes.

LTR Retrotransposon *gag* Gene Co-option in Animals

In animals, we identified a total of seven *Crtg* genes, including six previously known cases, namely *Arc* (*Crtg2*), *dArc* (*Crtg3*), *RTL* (*Crtg4* and *Crtg5*), *PNMA* (*Crtg6*), and *SCAN* (*Crtg7*). We further investigated or revisited the evolutionary history of these *Crtg* genes. We identified a novel *Crtg* gene, namely *Crtg1*, in invertebrates (fig. 2A). Synteny analysis suggests that *Crtg1* arose before the last common ancestor of Scarabaeidae and Lampyridae within Coleoptera (~286 Ma) (fig. 2A). Phylogenetic and similarity analyses of both

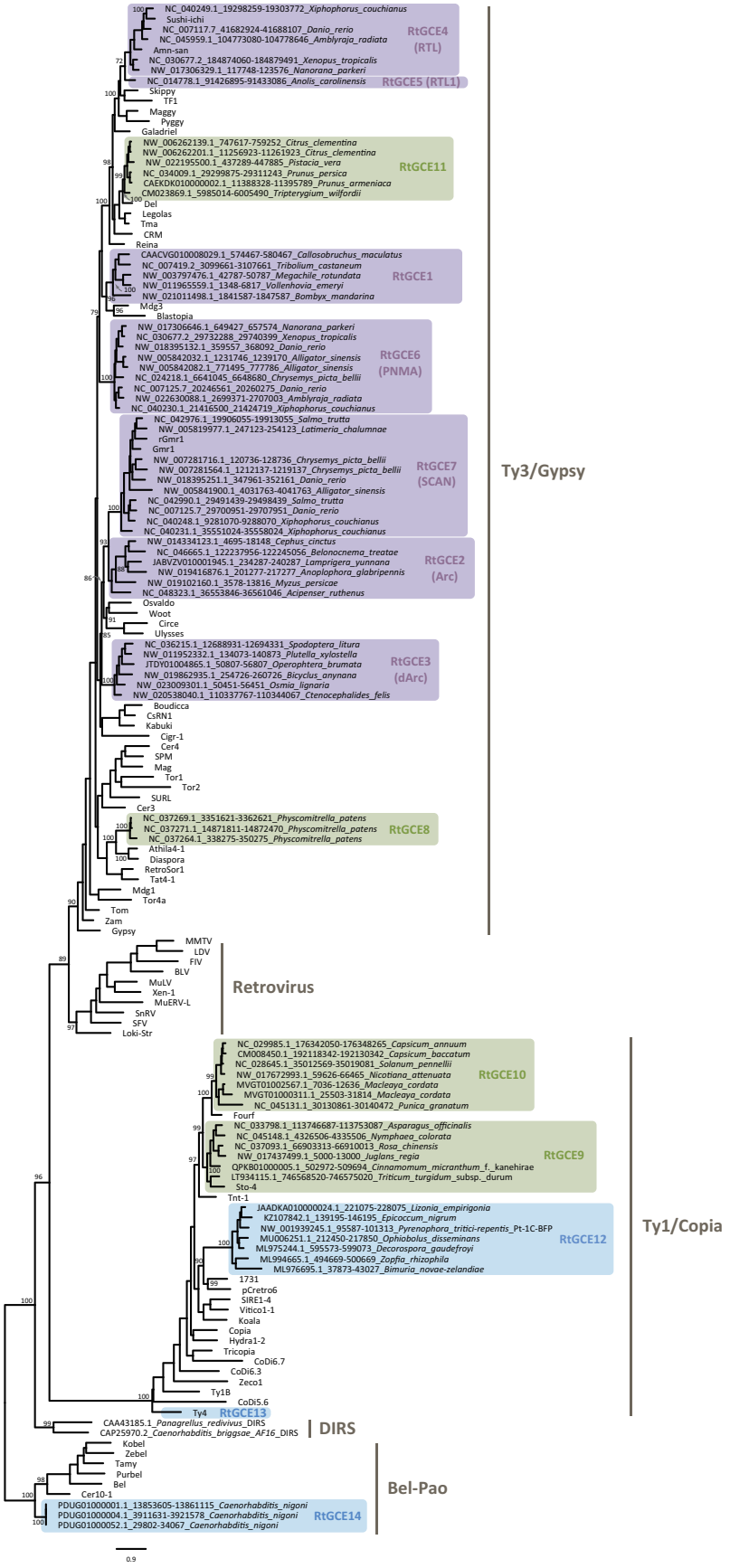

FIG. 1. The phylogenetic relationship between LTR retrotransposons that are closely related to Crtg proteins and representative LTR retrotransposons. This phylogenetic tree was reconstructed based on RT protein sequences. The LTR retrotransposons that are closely related to Crtg proteins in animals, plants and fungi were highlighted in purple, green, and blue, respectively.

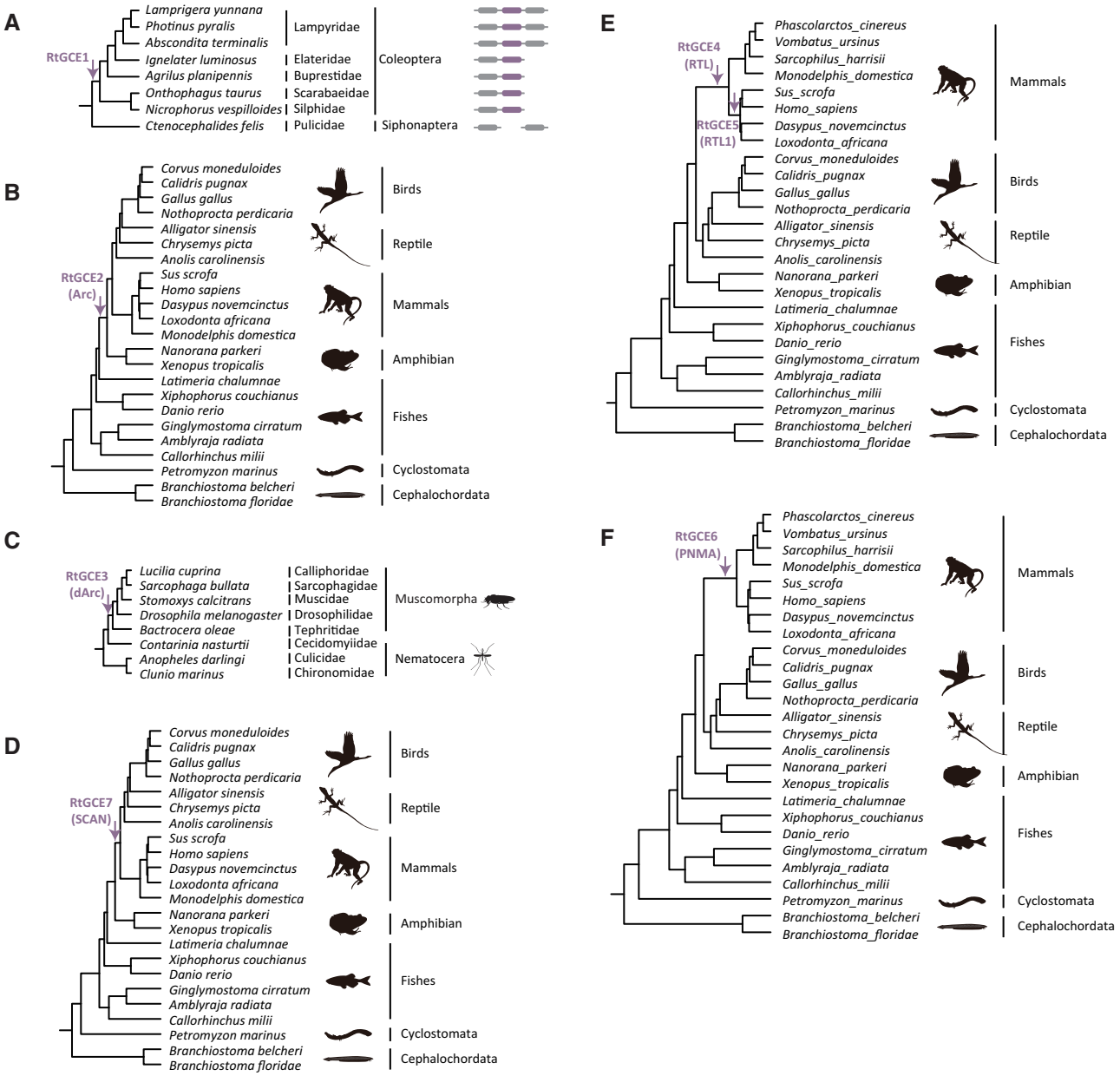

Fig. 2. The evolutionary history of *Crtg* genes in animals. The host phylogenetic relationship is based on TimeTree and literature (Wiegmann et al. 2011; Kumar et al. 2017; Zhang et al. 2018; Martin et al. 2019), and gene syntenies flanking each *Crtg* gene are shown near the corresponding species.

RT and Gag proteins show that *Crtg1* gene arose through repurposing the *gag* gene of a Mdg3-like retrotransposon within Ty3/Gypsy retrotransposons (fig. 1 and supplementary fig. S1*A*, Supplementary Material online). Consistent with previous studies, *Arc* and *dArc* originated independently: the co-option of *Arc* and *dArc* occurred before the last common ancestor of tetrapods (~352 Ma) and before the last common ancestor of Tephritidae and Drosophilidae within Muscomorpha (~126 Ma), respectively (Pastuzyn et al. 2018) (figs. 1 and 2*B* and 2*C*, and supplementary fig. S1*B*, S1*C*, Supplementary Material online). The known *RTL* genes appear to originate from sushi-like retrotransposons through two independent co-option events, with one occurring in the last common ancestor of mammals (~159 Ma) and the other occurring in the last common ancestor of placental mammals (~105 Ma) (figs. 1 and 2*E*, and supplementary fig. S1*D*, Supplementary Material online). *PNMA* genes arose through co-opting a Ty3/Gypsy retrotransposon *gag* gene before the last common ancestor of mammals (~159 Ma) (figs. 1 and 2*F*, and supplementary fig. S1*E*, Supplementary Material online). *SCAN* genes were derived from the *gag* gene of a Gmr1-like retrotransposon (Ty3/Gypsy) before the last common ancestor of tetrapods (~352 Ma) (Goodwin and Poulter 2002; Emerson and Thomas 2011) (figs. 1 and 2*D*, and supplementary fig. S2, Supplementary Material online). For *Ctrg1*, *Arc* (*Crtg2*), and *RTL-1* (*Crtg5*), most of them remain single copy in a species. Interestingly, *dArc* (*Crtg3*), *RTL* (*Crtg4*), *PNMA* (*Crtg6*), and *SCAN* (*Crtg7*) underwent extensive and complex gene duplication; for example, although *SCAN* gene remains single-copied in birds, *SCAN* genes underwent extensive

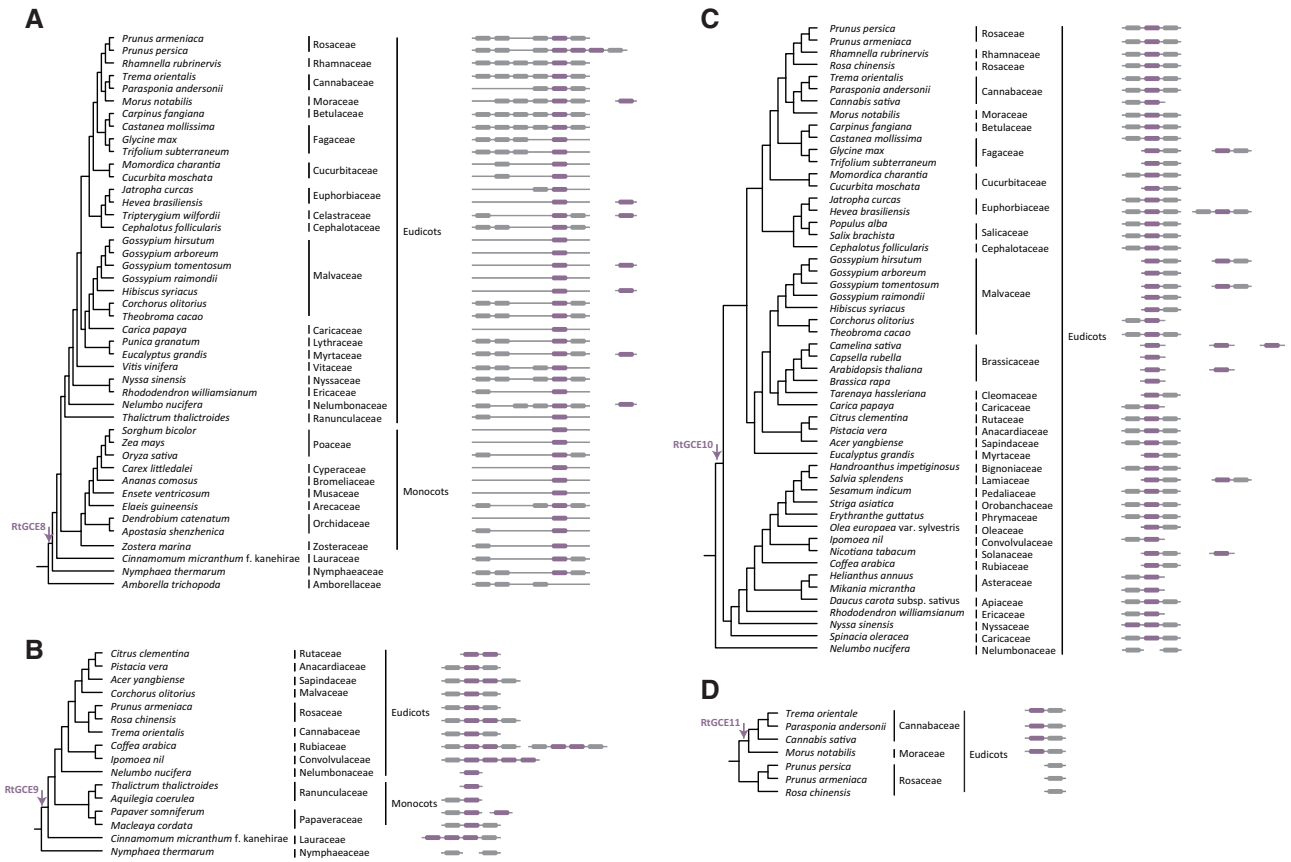

Fig. 3. The evolutionary history of *Crtg* genes in plants. The host phylogenetic relationship is based on TimeTree and literature (Kumar et al. 2017; Janssens et al. 2020), and gene syntenies flanking each *Crtg* gene are shown near the corresponding species.

duplication in reptiles and mammals independently (supplementary fig. S2, Supplementary Material online).

LTR Retrotransposon *gag* Gene Co-option in Plants

No retrotransposon *gag* gene co-option has been documented in plants before. In this study, we identified four novel retrotransposon *gag* gene co-option events in plants, generating *Crtg8* to *Crtg11* genes. Although *Crtg8* and *Crtg11* genes were derived from Ty3/Gypsy retrotransposon *gag* genes, the origin of *Crtg9* and *Crtg10* genes involved Ty1/Copia retrotransposon *gag* genes. Interestingly, two of them, *Crtg8* and *Crtg9*, originated during the early evolution of angiosperms. *Crtg8* genes are closely related to Athila-like retrotransposon *gag* genes, and the co-option occurred after the divergence of *Amborella trichopoda* from angiosperm (~175 Ma) (figs. 1 and 3A, and supplementary fig. S3A, Supplementary Material online). *Crtg9* genes appear to arise through co-opting a Tork-like retrotransposon *gag* gene, which occurred after the divergence of *Nymphaea thermarum* from angiosperms (~160 Ma) (figs. 1 and 3B, and supplementary fig. S3B, Supplementary Material online). The remaining co-option events took place during the evolutionary course of eudicots. *Crtg10* genes arose through co-opting a Tork-like retrotransposon *gag* gene, which occurred after the divergence of *Nelumbo nucifera* from eudicots (~117 Ma) (figs. 1 and 3C, and supplementary fig. S3C, Supplementary Material online). *Crtg11* genes are closely related to Del-like retrotransposon *gag* genes, and the co-option occurred before the common ancestor of Cannabaceae and Moraceae within eudicots (~86 Ma) (figs. 1 and 3D, and supplementary fig. S3D, Supplementary Material online). These *Crtg* genes underwent sporadic gene duplication, and notably *Crtg9* genes were tandemly duplicated in many species (fig. 3).

LTR Retrotransposon *gag* Gene Co-option in Fungi

No retrotransposon *gag* gene co-option has been documented in fungi before. In this study, we identified three retrotransposon *gag* gene co-option events in fungi, generating *Crtg12* to *Crtg14*. Although *Crtg12* and *Crtg13* appear to be derived from Ty1/Copia retrotransposon *gag* genes, *Crtg14* originated through repurposing a Bel-Pao retrotransposon *gag* gene (supplementary fig. S4, Supplementary Material online). The co-option of *Crtg12*, *Crtg13*, and *Crtg14* genes occurred before the last common ancestor of Lentitheciaceae and Lindgomycetaceae, before the last common ancestor of Pleosporales and Mytilinidiales (~242 Ma), and before the last common ancestor of Tremellales and Trichosporonales, respectively (fig. 4). All the *Crtg* genes in fungi are single copy (fig. 4).

Expression Pattern and Gene Structure of *Crtg* Genes

We used transcriptome sequencing (RNA-seq) raw data to explore whether the eight *Crtg* genes first identified in this study (*Crtg1*, *Crtg8* to *Crtg14*) are expressed. Strong evidence

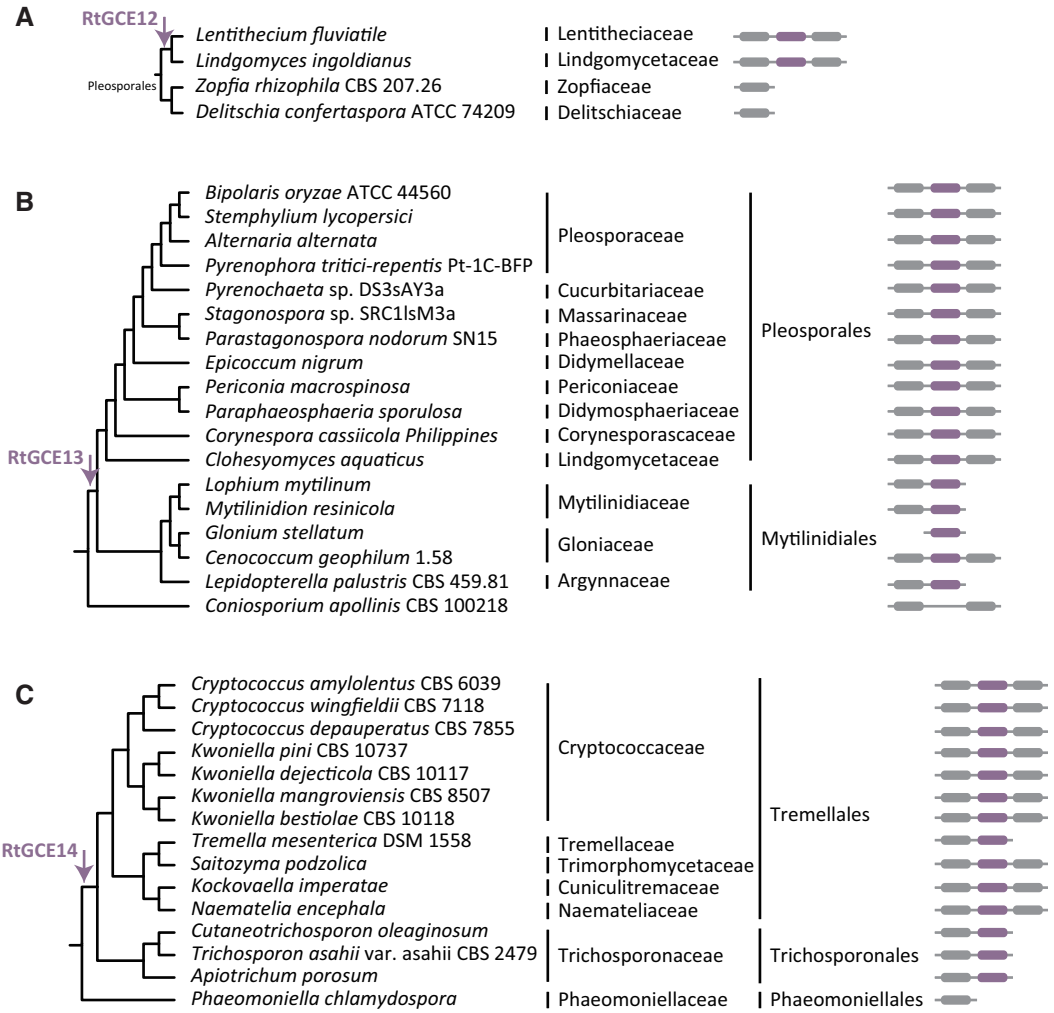

FIG. 4. The evolutionary history of *Crtg* genes in fungi. The host phylogenetic relationship is based on TimeTree and literature (Hirayama et al. 2010; Raja et al. 2011; Li et al. preprint; Tian et al. 2015), and gene syntenies flanking each *Crtg* gene are shown near the corresponding species.

for expression was found for almost all these eight *Crtg* genes (the largest transcript per million [TPM] value for each gene ranges from 1.61 to 72.23) (supplementary table S7, Supplementary Material online), except *Crtg12* with a TPM value of 0.41, which may be due to the poor quality of RNA-seq data (~60% reads are too short to map). *Crtg1* gene was found to be expressed during the development from larva to adult stages in *Agrilus planipennis* (supplementary table S7, Supplementary Material online). In plants, *Crtg8*, *Crtg9*, *Crtg10*, and *Crtg11* were found to be expressed in a wide range of tissues, including leaf, root, flower, and seed (supplementary table S7, Supplementary Material online). Because RNA-seq data are only available for a limited number of tissues and gene expression is of temporal and spatial specificity, our results do not necessarily indicate that the co-opted genes are not expressed in other tissues.

Interestingly, we found two *Crtg* genes, *Crtg10* and *Crtg14*, were fused with host genes (supplementary table S3, Supplementary Material online). *Crtg10* was fused with *KELP* gene, and the product of *KELP* gene is a transcriptional co-activator and binds movement protein (MP) of tomato mosaic virus to inhibit its cell-to-cell movement (Sasaki et al. 2009). *Crtg14* was fused with *PEX14* gene, which produces a peroxisomal membrane protein, Pex14p, involved in peroxisomal targeting signal-dependent protein import pathway (Albertini et al. 1997). Moreover, we found a putative intron with typical splicing sites GT-AG and the branch point (5′-YURAY-3′) (Thanaraj and Clark 2001) in the *gag*-derived region of *Crtg14* genes. Taken together, all the results of gene expression and gene structure analyses further support that the eight *Crtg* genes are co-opted *gag* genes.

Natural Selection Acting on *Crtg* Genes

Like host cellular genes, the co-opted retrotransposon genes should be subject to certain level of natural selection, implying potential cellular functionality (Graur et al. 2013; Jangam et al. 2017; Wang and Han 2020). To explore whether *Crtg* genes are subject to natural selection, we first calculated the dN/dS ratio for the eight *Crtg* genes (*Crtg1*, *Crtg8* to *Crtg14*), where dN represents the number of nonsynonymous substitutions per nonsynonymous site and dS represents the number of synonymous substitutions per synonymous site. The dN/dS ratio has often been used to detect signal of natural selection acting on genes (Daugherty and Malik 2012; Duggal

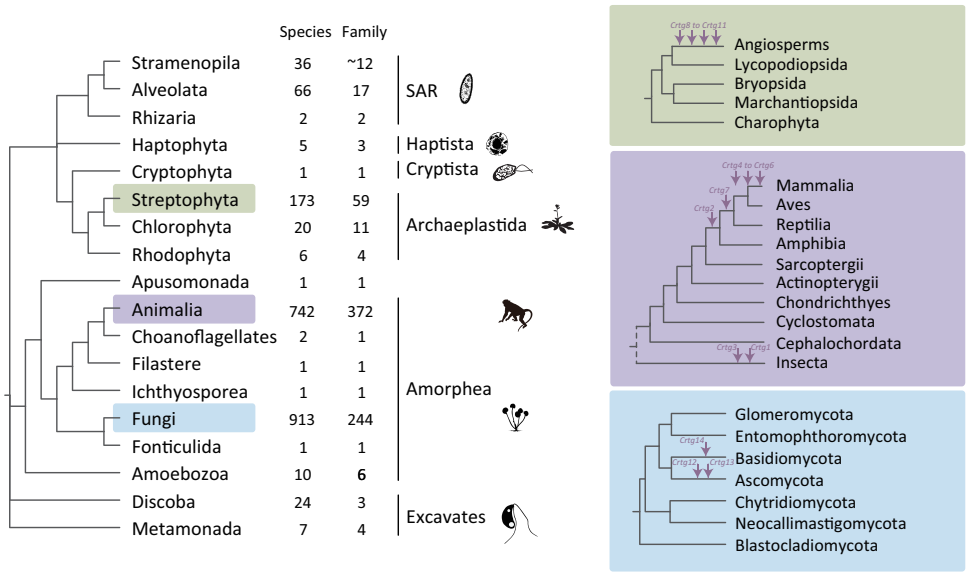

Fig. 5. The distribution of *Crtg* genes across eukaryotes. Phylogenetic tree of eukaryotes is based on literatures (Eichinger et al. 2005; Steenkamp et al. 2006; Burki et al. 2020).

and Emerman 2012; Sironi et al. 2015; Han 2019; Wang and Han 2020). We found that all the d*N*/d*S* ratios are less than one for all the eight genes (all but *Crtg12* are <0.44), indicating that they evolved under certain functional constraints (table 1).

Due to the conservative nature of d*N*/d*S* in detecting natural selection (Yang and Bielawski 2000), we further used two pairs of site models (M1a vs. M2a and M8a vs. M8) to detect sites subject to natural selection in *Crtg* genes. The results of M1a versus M2a show that all these eight *Crtg* genes possess a proportion of sites under purifying selection (table 1). Interestingly, evidence for positive selection was found in *Crtg10* gene (table 1). We also used the branch-site unrestricted statistical test for episodic diversification (BUSTED) method, which detects gene-wide diversifying selection (Murrell et al. 2015), and the fast, unconstrained Bayesian approximation for inferring selection (FUBAR) method, which identifies sites under purifying or positive selection (Murrell et al. 2013) to analyze natural selection in seven *Crtg* genes (*Crtg12* was excluded due to its limited number of sequences). The BUSTED method inferred at least one site or branch under positive selection in five *Crtg* genes (except *Crtg11* and *Crtg14*) (table 1). The FUBAR method identified many sites subject to purifying selection in all the seven genes and detected several positively selected sites in three plant *Crtg* genes, namely *Crtg9*, *Crtg10*, and *Crtg11* genes (table 1). Moreover, we divided each *Crtg* gene into several subsets at the taxonomic family level and tested the signals of natural selection for each subset. We detected several sites under positive selection in the Cryptococcaceae subset of *Crtg14* gene and found that the d*N*/d*S* ratio of the Fagaceae subset of *Crtg10* gene is greater than one (supplementary table S5, Supplementary Material online). Nevertheless, for each *Crtg* genes, the results of the subset analyses are similar to these of the whole data set analysis (supplementary table S5, Supplementary Material online). Overall, these results suggest that all the eight *Crtg* genes evolve under certain level of natural selection, indicating that they are likely to perform diverse cellular functions.

Discussion

Both coding and regulatory regions of TEs can be repurposed for diverse cellular functions (Feschotte 2008; Kokošar and Kordiš 2013; Hoen and Bureau 2015; Chuong et al. 2016; Naville et al. 2016; Chuong et al. 2017; Jangam et al. 2017; Wang et al. 2019; Wang and Han 2020; Etchegaray et al. 2021). LTR retrotransposons are highly abundant in the genomes of eukaryotes. However, a comprehensive picture of LTR retrotransposon *gag* gene co-option is still lacking, with only six cases documented in animals (Campillos et al. 2006; Pastuzyn et al. 2018). In this study, we used a phylogenomic approach to systematically mine co-opted LTR retrotransposon *gag* genes at the taxonomic family level across eukaryotes. We unearthed a total of 14 *Ctrg* genes, seven, four, and three of which were identified in animals, plants, and fungi, respectively. The LTR retrotransposon *gag* gene co-option events represent the ones occurred during the deep evolution of eukaryotes, because we only identified the co-option events occurring before the last common ancestor of at least two eukaryote families usually at the timescale of tens of millions of years. Among these cases of LTR retrotransposon *gag* gene co-option, eight have not been reported previously.

Across more than 740 eukaryote families, we only identified 14 LTR retrotransposon *gag* gene co-option events occurred above the taxonomic family level (fig. 5), indicating paucity of LTR retrotransposon *gag* gene co-option during the deep evolution of eukaryotes. Two scenarios could explain this pattern: 1) Co-option of LTR retrotransposon genes does occur at an extremely low frequency; and 2) co-option of LTR retrotransposon genes occurs frequently, but co-opted genes are frequently lost. In our previous study of retrovirus gene co-option, we also observed a similar pattern: retrovirus

Table 1. Selection Analyses of *Crtg* Genes.

Gene	No. of Sequences (used in selection analyses/total)	No. of Sites	dN/dS[a]	M1a Versus M2a					M8a Versus M8			BUSTED	FUBAR	
				2ΔlnL[b]	*P*-value[c]	P_0[d]	P_1[d]	P_2[d]	2ΔlnL[b]	*P*-value[c]	Codons with dN/dS > 1[e]	*P*-value	No. of Sites under Positive Selection	No. of Sites under Negative Selection
Crtg1	7/7	596	0.039	0	1	0.699	0.301	0	1.014	0.314	—	*P*-value = 0.005 ≤ .05	0	354
Crtg8	82/108	190	0.209	0	1	0.808	0.192	0	0.845	0.358	—	*P*-value = 0.000 ≤ .05	0	156
Crtg9	27/47	258	0.343	0	1	0.606	0.394	0	0.944	0.331	—	*P*-value = 0.000 ≤ .05	1	92
Crtg10	120/152	285	0.352	69.112	9.990×10^{-16}	0.573	0.386	0.041	28.113	0	27Q*; 233F*; 235S**; 236A**; 237N**; 241T*	*P*-value = ≤.05	1	161
Crtg11	4/4	259	0.44	0	1	0.64	0.36	0	0.861	0.353	—	*P*-value = 0.423 ≥ .05	1	13
Crtg12	2/2	217	0.996	0.001	0.999	0.005	0.995	0	0.001	0.972	—	—	—	—
Crtg13	62/63	293	0.115	0	1	0.844	0.156	0	0	1	—	*P*-value = 0.000 ≤ .05	0	285
Crtg14	14/22	236	0.113	0	1	0.71	0.29	0	0.032	0.859	—	*P*-value = 0.137 ≥ .05	0	145

[a]The dN/dS values of the *Ctrg* genes were estimated using the one-ratio model (M0) in PAML.
[b]2ΔlnL represents twice of the difference in the natural logs of the likelihoods of the pairs of models (M1a vs. M2a and M8a vs. M8) being compared.
[c]The *P*-value indicates the confidence with which the neutral models (M1a and M8a) can be rejected in favor of the positive selection models (M2a and M8), respectively.
[d]Proportion of sites with omega < 1 (P_0), omega = 1 (P_1), and omega > 1 (P_2).
[e]Codons under positive selection with a posterior probability > 95% and 99% by Bayes empirical Bayes (BEB) analysis are labeled with one and two asterisks, respectively.

gene co-option is relatively rare in the deep branches of vertebrates, which is likely due to frequent co-option and frequent loss (Wang and Han 2020). We think the paucity of LTR retrotransposon *gag* gene co-option during the deep evolution of eukaryotes could be explained by frequent co-option and frequent loss, and mining co-option events within the level of taxonomic family would help confirm this hypothesis. Our analyses came with several caveats: 1) We only mined annotated proteomes of eukaryotes, and many retrotransposon-related genes might not be well annotated. Thus, the number of co-opted LTR retrotransposon *gag* genes are underestimated in this study. 2) We only sampled ~740 eukaryote families. It is possible that many co-option events occurred recently, and these relatively recent co-option events might not be unearthed in this study. 3) Our data set is biased to animals, fungi, and plants, as most genome sequencing has been performed in these groups, which might result in underestimation of LTR co-option in protists. However, our study well covers the deep diversity of eukaryotes, and covers the major diversity of animals, plants, and fungi. If a co-option event occurred in deep past and the co-opted gene pass on to its descendants, and if the co-opted gene has been annotated in some of the descendants, our analysis could capture this event (Wang and Han 2020). It follows that we might not miss many co-option events occurred in deep past (especially within animals, plants, and fungi), such as the emergence of tetrapods or the emergence of angiosperms. Together, our results reveal paucity of LTR retrotransposon *gag* gene co-option during the deep evolution of eukaryotes and suggest that co-opted LTR retrotransposon *gag* genes might have not been maintained for extremely long periods of time.

Co-opted LTR retrotransposon *gag* genes and co-opted retrovirus genes have been known to perform diverse cellular functions in animals, ranging from mediating intercellular RNA transfer in the nervous system, to regulating developmental processes, to inhibiting viral infections (Ashley et al. 2018; Pastuzyn et al. 2018; Wang and Han 2020). In general, genes that regulate crucial physiological processes might be mainly subject to purifying selection, whereas genes that are involved in certain genetic conflicts mainly evolve under strong positive selection. For eight *Crtg* genes first identified in this study, we found these genes appear to evolve mainly under purifying selection. These genes are expressed in a wide range of tissues or developmental stages. Evidence of positive selection was detected for some *Ctrg* genes, especially *Ctrg10*. *Crtg10* gene was found to be fused with *KELP* gene, and *KELP* gene functions in the inhibition of tomato mosaic virus infection (table 1 and supplementary table S3, Supplementary Material online). It is possible that the fusion between a co-opted *gag* gene and *KELP* might participate in the evolutionary arms race between hosts and viruses. Overall, our results indicate *Crtg* genes might perform diverse cellular functions. Further experiments are still needed to explore the function of these *Crtg* genes.

Our study provides insights into the co-option of LTR retrotransposon *gag* genes. First, all the six co-opted LTR retrotransposon *gag* genes previously documented involve

Ty3/Gypsy retrotransposons (Campillos et al. 2006; Pastuzyn et al. 2018). In this study, we identified four and one *Ctrg* genes which were derived from *gag* genes of Ty1/Copia and Bel-Pao retrotransposons, respectively. Our study indicates that *gag* genes from diverse LTR retrotransposons can be repurposed for cellular functions (fig. 1). Second, all the previously reported co-option of LTR retrotransposon *gag* genes occurred in animals. In this study, we identified four and three LTR retrotransposon *gag* gene co-option events occurred in plants and fungi, respectively (fig. 5). Therefore, our study has expanded the range of LTR retrotransposon *gag* gene co-option. It follows that LTR retrotransposon *gag* gene co-option occurred more widely than previously appreciated. Our study provides a snapshot of LTR retrotransposon *gag* gene co-option in eukaryotes.

Materials and Methods

Identification of Co-opted LTR Retrotransposon *gag* Genes

We employed a similarity search and phylogenomic analysis combined approach (Wang and Han 2020) to identify co-opted LTR retrotransposon *gag* genes above the taxonomic family level across 2,011 eukaryotes. In brief, we first mined the homologs of LTR retrotransposon Gag proteins in 2,011 eukaryote genomes using the hmmsearch program in HMMER 3.3.1 with seven family in GAG-polyprotein clan (CL0523), Arc_C family, PNMA family, and SCAN family as queries and an *e* cut-off value of 0.1 (Eddy 2011) (supplementary tables S1 and S2, Supplementary Material online). The DUF4219 family in GAG-polyprotein clan was excluded, because its seed alignment is too short. Next, phylogenetic analyses of significant hits and representative Gag proteins of retroviruses and retrotransposons were performed (Llorens et al. 2008). Gag protein hits whose phylogenetic relationship largely agrees with their host above taxonomic family level were retrieved as co-opted Gag protein candidates. Finally, we verified the domain configuration for each co-opted Gag candidate using SMART and Conserved Domain (CD) search with default parameters (Lu et al. 2020; Letunic et al. 2021), and the candidates that encode these query domains were retrieved for further analyses. Protein sequences were aligned using MAFFT 7 (Katoh et al. 2002). Phylogenetic trees were reconstructed using an approximate maximum likelihood method implemented in FastTree 2.1.11 (Price et al. 2010). We used two criteria to define *Crtg* genes: 1) The *Crtg* genes share similar synteny among the genomes of species across at least two families, and/or the *Crtg* phylogeny largely agree with the host phylogeny; and 2) the *Crtg* genes are subject to certain level of selection. The syntenies flanking *Crtg* genes were based on genome annotation and/or domain annotation by CD search. The divergence time of hosts provides a minimum time estimate for the occurrence of co-option events. Host divergence time was based on TimeTree (Kumar et al. 2017).

Analysis of the Evolutionary History of Co-opted *gag* Genes

To explore the evolutionary history for each *Crtg* gene, we used the TBlastN algorithm to search 2,011 eukaryote genomes with a representative protein sequence for each *Crtg* gene as the query and an *e* cut-off value of 10^{-5}. Phylogenetic analysis of significant hits and representative Gag proteins was performed to confirm the distribution of *Crtg* genes, and identify LTR retrotransposons that are closely related to *Crtg* genes. The significant hits were bidirectionally extended to identify classic structure of LTR retrotransposons (supplementary table S4, Supplementary Material online). LTR_Finder and LTRharvest were used to identify LTRs, and CD search was used to annotate protein domains (Xu and Wang 2007; Ellinghaus et al. 2008; Lu et al. 2020). We failed to identify LTR retrotransposons that are closely related to *Crtg13* genes. But phylogenetic analysis of Gag proteins shows that it is closely related to Ty1/Copia retrotransposons (supplementary fig. S4*B*, Supplementary Material online).

Phylogenetic Analyses

For each *Crtg* gene, we performed phylogenetic analysis of Crtg proteins, Gag proteins of LTR retrotransposons closely related to Crtg, and representative LTR retrotransposons. We only used Crtg7 protein sequences with the length of >84 amino acids for phylogenetic analyses (Edelstein and Collins 2005). To explore the phylogenetic relationship between LTR retrotransposons closely related to Crtg proteins and representative LTR retrotransposons, we performed phylogenetic analysis of RT proteins of LTR retrotransposons closely related to Crtg proteins and representative LTR retrotransposons (supplementary tables S4 and S6, Supplementary Material online). Protein sequences were aligned using MAFFT 7 with the strategy of L-INS-I, and then manually refined (Katoh et al. 2002). Phylogenetic analyses were performed using a maximum likelihood method implemented in IQTREE 2.0 (Nguyen et al. 2015). ModelFinder in IQ-TREE 2.0 was used to select the best-fit models (Kalyaanamoorthy et al. 2017). The branch support values were assessed using the ultrafast bootstrap method with 1,000 replicates (Hoang et al. 2018).

Expression Pattern Analyses

The Illumina pair-end RNA-seq raw data for three fungi (*Lindgomyces ingoldianus*, *Alternaria alternata* SRC1lrK2f, and *Kockovaella imperatae* NRRL Y-17943), four developmental stages (larva, prepupae, pupae, and adult) of one animal (*A. planipennis*), and seven tissues (leaf, root, bud, ovule, flower, petiole, and seed) of four plants (*N. thermarum*, *Arabidopsis thaliana*, *N. nucifera*, and *Morus notabilis*) were retrieved from NCBI to analyze the expression pattern of *Crtg12* to *Crtg14*, *Crtg1*, and *Crtg8* to *Crtg11*, respectively (supplementary table S7, Supplementary Material online). First, we employed Trimmomatic v0.39 to trim and filter the RNA-seq raw data (Bolger et al. 2014). Next, we used STAR v2.5.4b to map reads to reference genomes (Dobin et al. 2013). To obtain the uniquely mapped reads for each gene, we used the –quantMode GeneCounts option. Read alignment files

generated by STAR v2.5.4b were sorted using Samtools v1.11 (Li et al. 2009). Finally, StringTie v2.1.5 was used to assemble transcripts and estimate the gene abundance through calculating TPM values (Kovaka et al. 2019). To further confirm that *gag*-derived regions of *Crtg10* and *Crtg14* are expressed, TBlastN algorithm was employed to BLAST against the corresponding RNA-seq data using the *gag*-derived regions as queries with an *e* cut-off value of 10^{-5}. 252 and 1712 raw reads mapped to the *gag*-derived regions of *Crtg10* and *Crtg14* with the identity of 100% and the query covery of 100%, respectively.

Selection Pressure Analyses

For each *Crtg* gene, we used sequences without any premature stop codon and frameshift mutation to perform selection analyses. The one ratio model (M0) in PAML 4.9j was used to estimate the overall dN/dS ratio (Yang 2007). Two pairs of site models, M1a versus M2a and M8a versus M8, in PAML 4.9j were used to detect sites under purifying selection and positive selection. For data sets with more than two nonidentical sequences, the BUSTED method and the FUBAR method implemented in HyPhy package were employed to identify gene-wide selection signal and sites under natural selection, respectively (Murrell et al. 2013, 2015). All the nucleotide sequences were aligned based on codons using MUSCLE, and the ambiguous regions were removed manually (Kumar et al. 2016). Phylogenetic trees used in selection analyses were reconstructed using IQ-TREE 2.0 (Nguyen et al. 2015).

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgements

This study was supported by National Natural Science Foundation of China (31922001).

Data Availability

No new data were generated in support of this research.

References

Albertini M, Rehling P, Erdmann R, Girzalsky W, Kiel JA, Veenhuis M, Kunau WH. 1997. Pex14p, a peroxisomal membrane protein binding both receptors of the two PTS-dependent import pathways. *Cell* 89(1):83–92.

Ashley J, Cordy B, Lucia D, Fradkin LG, Budnik V, Thomson T. 2018. Retrovirus-like gag protein Arc1 binds RNA and traffics across synaptic boutons. *Cell* 172(1–2):262–274.e11.

Best S, Le Tissier P, Towers G, Stoye JP. 1996. Positional cloning of the mouse retrovirus restriction gene Fv1. *Nature* 382(6594):826–829.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15):2114–2120.

Boso G, Buckler-White A, Kozak CA. 2018. Ancient evolutionary origin and positive selection of the retroviral restriction factor Fv1 in muroid rodents. *J Virol*. 92(18):e00850–18.

Brandt J, Veith AM, Volff JN. 2005. A family of neofunctionalized Ty3/gypsy retrotransposon genes in mammalian genomes. *Cytogenet Genome Res*. 110(1–4):307–317.

Bundock P, Hooykaas P. 2005. An Arabidopsis hAT-like transposase is essential for plant development. *Nature* 436(7048):282–284.

Burki F, Roger AJ, Brown MW, Simpson AGB. 2020. The new tree of eukaryotes. *Trends Ecol Evol*. 35(1):43–55.

Cam HP, Noma K, Ebina H, Levin HL, Grewal SI. 2008. Host genome surveillance for retrotransposons by transposon-derived proteins. *Nature* 451(7177):431–436.

Campillos M, Doerks T, Shah PK, Bork P. 2006. Computational characterization of multiple Gag-like human proteins. *Trends Genet*. 22(11):585–589.

Cho G, Lim Y, Golden JA. 2011. XLMR candidate mouse gene, Zcchc12 (Sizn1) is a novel marker of Cajal–Retzius cells. *Gene Expr Patterns*. 11(3–4):216–220.

Chuong EB, Elde NC, Feschotte C. 2016. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science* 351(6277):1083–1087.

Chuong EB, Elde NC, Feschotte C. 2017. Regulatory activities of transposable elements: from conflicts to benefits. *Nat Rev Genet*. 18(2):71–86.

Collins T, Stone JR, Williams AJ. 2001. All in the family: the BTB/POZ, KRAB, and SCAN domains. *Mol Cell Biol*. 21(11):3609–3615.

Daugherty MD, Malik HS. 2012. Rules of engagement: molecular insights from host–virus arms races. *Annu Rev Genet*. 46:677–700.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29(1):15–21.

Duggal NK, Emerman M. 2012. Evolutionary conflicts between viruses and restriction factors shape immunity. *Nat Rev Immunol*. 12(10):687–695.

Eddy SR. 2011. Accelerated profile HMM searches. *PLoS Comp Biol*. 7(10):e1002195.

Edelstein LC, Collins T. 2005. The SCAN domain family of zinc finger transcription factors. *Gene* 359:1–17.

Eichinger L, Pachebat JA, Glöckner G, Rajandream MA, Sucgang R, Berriman M, Song J, Olsen R, Szafranski K, Xu Q, et al. 2005. The genome of the social amoeba *Dictyostelium discoideum*. *Nature* 435(7038):43–57.

Ellinghaus D, Kurtz S, Willhoeft U. 2008. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* 9(1):18.

Emerson RO, Thomas JH. 2011. Gypsy and the birth of the SCAN domain. *J Virol*. 85(22):12043–12052.

Etchegaray E, Naville M, Volff JN, Haftek-Terreau Z. 2021. Transposable element-derived sequences in vertebrate development. *Mob DNA* 12:1.

Feschotte C. 2008. Transposable elements and the evolution of regulatory networks. *Nat Rev Genet*. 9(5):397–405.

Goodwin TJ, Poulter RT. 2002. A group of deuterostome Ty3/gypsy-like retrotransposons with Ty1/copia-like pol-domain orders. *Mol Genet Genomics*. 267(4):481–491.

Graur D, Zheng Y, Price N, Azevedo RB, Zufall RA, Elhaik E. 2013. On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. *Genome Biol Evol*. 5(3):578–590.

Han GZ. 2019. Origin and evolution of the plant immune system. *New Phytol*. 222(1):70–83.

Hirayama K, Tanaka K, Raja HA, Miller AN, Shearer CA. 2010. A molecular phylogenetic assessment of *Massarina ingoldiana* sensu lato. *Mycologia* 102(3):729–746.

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: improving the ultrafast bootstrap approximation. *Mol Biol Evol*. 35(2):518–522.

Hoen DR, Bureau TE. 2015. Discovery of novel genes derived from transposable elements using integrative genomic analysis. *Mol Biol Evol*. 32(6):1487–506.

Huang S, Tao X, Yuan S, Zhang Y, Li P, Beilinson HA, Zhang Y, Yu W, Pontarotti P, Escriva H, et al. 2016. Discovery of an active RAG transposon illuminates the origins of V(D). *J Recombin Cell*. 166(1):102–114.

Jangam D, Feschotte C, Betrán E. 2017. Transposable element domestication as an adaptation to evolutionary conflicts. *Trends Genet*. 33(11):817–831.

Janssens SB, Couvreur TLP, Mertens A, Dauby G, Dagallier LMJ, Vanden Abeele S, Vandelook F, Mascarello M, Beeckman H, Sosef M, et al. 2020. A large-scale species level dated angiosperm phylogeny for evolutionary and ecological analyses. *Biodivers Data J*. 8:e39677.

Johnson WE. 2019. Origins and evolutionary consequences of ancient endogenous retroviruses. *Nat Rev Microbiol*. 17(6):355–370.

Joly-Lopez Z, Forczek E, Hoen DR, Juretic N, Bureau TE. 2012. A gene family derived from transposable elements during early angiosperm evolution has reproductive fitness benefits in *Arabidopsis thaliana*. *PLoS Genet*. 8(9):e1002931.

Joly-Lopez Z, Hoen DR, Blanchette M, Bureau TE. 2016. Phylogenetic and genomic analyses resolve the origin of important plant genes derived from transposable elements. *Mol Biol Evol*. 33(8):1937–1956.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods*. 14(6):587–589.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res*. 30(14):3059–3066.

Knip M, de Pater S, Hooykaas PJ. 2012. The SLEEPER genes: a transposase-derived angiosperm-specific gene family. *BMC Plant Biol*. 12:192.

Kokošar J, Kordiš D. 2013. Genesis and regulatory wiring of retroelement-derived domesticated genes: a phylogenomic perspective. *Mol Biol Evol*. 30(5):1015–1031.

Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M. 2019. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol*. 20(1):278.

Kumar S, Stecher G, Suleski M, Hedges SB. 2017. TimeTree: a resource for timelines, timetrees, and divergence times. *Mol Biol Evol*. 34(7):1812–1819.

Kumar S, Stecher G, Tamura K. 2016. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol Biol Evol*. 33(7):1870–1874.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2012. Initial sequencing and analysis of the human genome. *Nature* 409:860–921.

Letunic I, Khedkar S, Bork P. 2021. SMART: recent updates, new developments and status in 2020. *Nucleic Acids Res*. 49(D1):D458–D460.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R.; 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078–2079.

Li YN, Steenwyk JL, Chang Y, Wang Y, James TY, Stajich JE, Spatafora JW, Groenewald M, Dunn CW, Hittinger CT, et al. 2021. A genome-scale. Curr Biol. 31:1–13. Available from: 10.1101/2020.08.23.262857.

Lin R, Ding L, Casola C, Ripoll DR, Feschotte C, Wang H. 2007. Transposase-derived transcription factors regulate light signaling in Arabidopsis. *Science* 318(5854):1302–1305.

Llorens C, Fares MA, Moya A. 2008. Relationships of gag-pol diversity between Ty3/Gypsy and Retroviridae LTR retroelements and the three kings hypothesis. *BMC Evol Biol*. 8(1):276.

Llorens C, Muñoz-Pomer A, Bernad L, Botella H, Moya A. 2009. Network dynamics of eukaryotic LTR retroelements beyond phylogenetic trees. *Biol Direct*. 4:41.

Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Marchler GH, Song JS, et al. 2020. CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res*. 48(D1), D265–D268.

Ma L, Li G. 2018. FAR1-related sequence (FRS) and FRS-related factor (FRF) family proteins in Arabidopsis growth and development. *Front Plant Sci*. 9:692.

Makhnovskii P, Balakireva Y, Nefedova L, Lavrenov A, Kuzmin I, Kim A. 2020. Domesticated gag gene of Drosophila LTR retrotransposons is involved in response to oxidative stress. *Genes (Basel)*. 11(4):396.

Martin GJ, Stanger-Hall KF, Branham MA, Silveira LFLDA, Lower SE, Hall DW, Li XY, Lemmon AR, Lemmon EM, Bybee SM. 2019. Higher-level phylogeny and reclassification of Lampyridae (Coleoptera: Elateroidea). *Insect Syst Divers*. 3(6):1–15

Morales Poole JR, Huang SF, Xu A, Bayet J, Pontarotti P. 2017. The RAG transposon is active through the deuterostome evolution and domesticated in jawed vertebrates. *Immunogenetics* 69(6):391–400.

Murrell B, Moola S, Mabona A, Weighill T, Sheward D, Kosakovsky Pond SL, Scheffler K. 2013. FUBAR: a fast, unconstrained Bayesian approximation for inferring selection. *Mol Biol Evol*. 30(5):1196–1205.

Murrell B, Weaver S, Smith MD, Wertheim JO, Murrell S, Aylward A, Eren K, Pollner T, Martin DP, Smith DM, et al. 2015. Gene-wide identification of episodic selection. *Mol Biol Evol*. 32(5):1365–1371.

Naville M, Warren IA, Haftek-Terreau Z, Chalopin D, Brunet F, Levin P, Galiana D, Volff JN. 2016. Not so bad after all: retroviruses and long terminal repeat retrotransposons as a source of new genes in vertebrates. *Clin Microbiol Infect*. 22(4):312–323.

Nefedova LN, Kuzmin IV, Makhnovskii PA, Kim AI. 2014. Domesticated retroviral GAG gene in Drosophila: new functions for an old gene. *Virology* 450–451:196–204.

Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*. 32(1):268–274.

Oliver KR, McComb JA, Greene WK. 2013. Transposable elements: powerful contributors to angiosperm evolution and diversity. *Genome Biol Evol*. 5(10):1886–1901.

Ono R, Nakamura K, Inoue K, Naruse M, Usami T, Wakisaka-Saito N, Hino T, Suzuki-Migishima R, Ogonuki N, Miki H, et al. 2006. Deletion of Peg10, an imprinted gene acquired from a retrotransposon, causes early embryonic lethality. *Nat Genet*. 38(1):101–106.

Pang SW, Lahiri C, Poh CL, Tan KO. 2018. PNMA family: protein interaction network and cell signalling pathways implicated in cancer and apoptosis. *Cell Signal*. 45:54–62.

Pastuzyn ED, Day CE, Kearns RB, Kyrke-Smith M, Taibi AV, McCormick J, Yoder N, Belnap DM, Erlendsson S, Morado DR, et al. 2018. The neuronal gene arc encodes a repurposed retrotransposon Gag protein that mediates intercellular RNA transfer. *Cell* 172(1–2):275–288.e18.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One* 5(3):e9490.

Raja HA, Tanaka K, Hirayama K, Miller AN, Shearer CA. 2011. Freshwater ascomycetes: two new species of *Lindgomyces* (Lindgomycetaceae, Pleosporales, Dothideomycetes) from Japan and USA. *Mycologia* 103(6):1421–1432.

Sanchez DH, Gaubert H, Drost HG, Zabet NR, Paszkowski J. 2017. High-frequency recombination between members of an LTR retrotransposon family during transposition bursts. *Nat Commun*. 8(1):1283.

Sander TL, Stringer KF, Maki JL, Szauter P, Stone JR, Collins T. 2003. The SCAN domain defines a large family of zinc finger transcription factors. *Gene* 310:29–38.

Sasaki N, Ogata T, Deguchi M, Nagai S, Tamai A, Meshi T, Kawakami S, Watanabe Y, Matsushita Y, Nyunoya H. 2009. Over-expression of putative transcriptional coactivator KELP interferes with Tomato mosaic virus cell-to-cell movement. *Mol Plant Pathol*. 10(2):161–173.

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, et al. 2009. The B73 maize genome: complexity, diversity, and dynamics. *Science* 326(5956):1112–1115.

Sekita Y, Wagatsuma H, Nakamura K, Ono R, Kagami M, Wakisaka N, Hino T, Suzuki-Migishima R, Kohda T, Ogura A, et al. 2008. Role of retrotransposon-derived imprinted gene, Rtl1, in the feto-maternal interface of mouse placenta. *Nat Genet*. 40(2):243–248.

Sironi M, Cagliani R, Forni D, Clerici M. 2015. Evolutionary insights into host-pathogen interactions from mammalian sequence data. *Nat Rev Genet*. 16(4):224–236.

Skirmuntt EC, Katzourakis A. 2019. The evolution of endogenous retroviral envelope genes in bats and their potential contribution to host biology. *Virus Res*. 270:197645.

Steenkamp ET, Wright J, Baldauf SL. 2006. The protistan origins of animals and fungi. *Mol Biol Evol*. 23(1):93–106.

Stoye JP. 2012. Studies of endogenous retroviruses reveal a continuing evolutionary saga. *Nat Rev Microbiol*. 10(6):395–406.

Takaji M, Komatsu Y, Watakabe A, Hashikawa T, Yamamori T. 2009. Paraneoplastic antigen-like 5 gene (PNMA5) is preferentially expressed in the association areas in a primate specific manner. *Cereb Cortex*. 19(12):2865–2879.

Thanaraj TA, Clark F. 2001. Human GC-AG alternative intron isoforms with weak donor sites show enhanced consensus at acceptor exon positions. *Nucleic Acids Res*. 29(12):2581–2593.

Tian, Q, Liu, JK, Hyde, KD, Wanasinghe, DN, Boonmee S, Jayasiri SC, Luo ZL, Taylor JE, Phillips AJL, Bhat DJ, et al. 2015. Phylogenetic relationships and morphological reappraisal of Melanommataceae (Pleosporales). *Fungal Divers*. 74(1):267–324.

Wang H, Wang H. 2015. Multifaceted roles of FHY3 and FAR1 in light signaling and beyond. *Trends Plant Sci*. 20(7):453–461.

Wang J, Gong Z, Han GZ. 2019. Convergent co-option of the retroviral gag gene during the early evolution of mammals. *J Virol*. 93(14):e00542–19.

Wang J, Han GZ. 2020. Frequent retroviral gene co-option during the evolution of vertebrates. *Mol Biol Evol*. 37(11):3232–3242.

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, et al. 2007. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet*. 8(12):973–982.

Wiegmann BM, Trautwein MD, Winkler IS, Barr NB, Kim JW, Lambkin C, Bertone MA, Cassel BK, Bayless KM, Heimberg AM, et al. 2011. Episodic radiations in the fly tree of life. *Proc Natl Acad Sci USA*. 108(14):5690–5695.

Xu Z, Wang H. 2007. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res*. 35(Web Server issue):W265–268.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 24(8):1586–1591.

Yang Z, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. *Trends Ecol Evol*. 15(12):496–503.

Yap MW, Colbeck E, Ellis SA, Stoye JP. 2014. Evolution of the retroviral restriction gene Fv1: inhibition of non-MLV retroviruses. *PLoS Pathog*. 10(3):e1003968.

Zhang SQ, Che LH, Li Y, Dan L, Pang H, Ślipiński A, Zhang P. 2018. Evolutionary history of Coleoptera revealed by extensive sampling of genes and species. *Nat Commun*. 9:205.