Department of
BIOLOGY

A17753S1 PROJECT DISSERTATION

Candidate Number	1033961
Supervisor	Professor Aris Katzourakis
Title	A virus-host fusion gene might underlie the evolution of megabat antiviral immunity
Word Count	8267 = 6999 + 1268

DO NOT ENTER YOUR NAME ANYWHERE IN THIS DOCUMENT.

DO NOT ENTER YOUR CANDIDATE NUMBER ANYWHERE ELSE.

Department of BIOLOGY

UNIVERSITY OF OXFORD

TABLE OF CONTENTS

ABSTRACT

The unique antiviral immunity of bats is linked to their contribution to zoonotic disease outbreaks such as COVID-19 and Ebola. However, it is unknown how immunological genetic innovations that resulted from ancient host-virus interactions have shaped the evolution of bats' antiviral immunity. To address this question, I studied Gag Co-option Event 32 (*GCE32*) – an endogenous viral element-derived fusion gene that was captured and repurposed by the host – due to its conservation in megabats and unique combination of virus- and host-derived sequences. Using comparative genomic and bioinformatic approaches, I found that *GCE32* acquired three host gene-related domains from the megabat common ancestor's *RBM33, CRKL and PEG3 genes*. *GCE32*-related EVEs allow the structural reconstruction of ancient viral proteins, which are similar to current ones. *GCE32* is mainly conserved in evolution but positively selected in specific domains and megabat lineages. It is also currently expressed in multiple megabats in immune-privileged tissues such as the liver and the testis. These findings suggest that *GCE32* might have taken up immunological functions such as maintaining immune privilege in its evolution. They also provide insights into ancient bat-virus interactions, which might underlie bat immune system evolution. This study adds to our current understanding of bat antiviral immunity from an evolutionary perspective, which could contribute to the prevention and control of zoonotic diseases of bat origin.

INTRODUCTION

Bats are unique mammals with powered flight, exceptional longevity for their size, and the ability to harbour viral infections with minimal pathogenesis. Their unique physiology underlies their status as a major reservoir for zoonotic disease outbreaks (Irving *et al*, 2021). They harbour a higher diversity of viruses than other mammals and have the highest zoonotic disease potential per species among all mammalian orders (Letko *et al*, 2020; Olival *et al*, 2017). They have been attributed as a natural reservoir of zoonotic viruses such as SARS-CoV, SARS-CoV-2, Ebolavirus and Marburg virus (Irving *et al*, 2021; Letko *et al*, 2020). With increasing human-wildlife conflict, bats' reservoir status to zoonotic viruses has contributed to emerging infectious disease outbreaks and remains a significant threat to public health.

Although much is unknown about how bats' immune system allows them to harbour viruses lethal to other mammals without significant clinical symptoms, new mechanistic understandings have been made in recent years. This ability is possibly linked to the balance between the protective and pathological responses of their immune system (Irving *et al*, 2021). On the one hand, the protective responses to viral infections are strong in bats. For instance, the constitutive expression of interferons in bats offers stronger protection against viral infections (Shaw *et al*, 2017). Some stress-induced proteins, such as heat shock proteins and ABCB1, a transporter protein of foreign substances, are significantly elevated in expression in bats compared to other mammals (Chionh *et al*, 2019; Koh *et al*, 2019). These mechanisms protect their cells from DNA damage and make them highly tolerant to infection-induced stress factors. On the other hand, dampened inflammatory responses in bats also contribute to higher tolerance to viral infections (Banerjee *et al*, 2020). For example, bats have lost the ALR gene family that encodes sensors of double-stranded DNA – a signal of pathogen infiltration – and initiates inflammatory responses (Goh *et al*, 2020). Their caspase-1 protein leads to reduced processing and maturation of interleukin-1β, further limiting their inflammatory responses (Goh *et al*, 2020). Moreover, the activation of tumour necrosis factor and the inflammasome sensor protein NLRP3 in immune cells is significantly dampened in bats compared to mice in response to viral challenges (Ahn *et al*, 2019; Banerjee *et al*, 2017). This combination of host defence mechanisms and immune tolerance can contribute to bats' ability to harbour viruses and act as a significant viral reservoir.

Although the differences in universal immunological components between bats and other mammals have reshaped their antiviral immune responses, bats could also have

gained novel genes in their own evolutionary history that contributed to the evolution of their immune responses. For example, gene duplication can lead to divergent immune system evolution among bat species. While the Egyptian fruit bat (*Rousettus aegyptiacus* G.) has 46 type I interferon genes, the Jamaican fruit bat (*Artibeus jamaicensis* L.) possesses only 6; moreover, different types of interferons are activated upon viral infection (David *et al*, 2022). Furthermore, *TRIM5* and *TRIM22* – restriction factors directly interacting with the viral capsid – have independently duplicated multiple times among bats, resulting in differing copy numbers (Fernandes *et al*, 2022). While processes other than gene duplication – such as virus-to-host horizontal gene transfer and gene fusion – can also contribute to bat immunity's evolution, their role remains largely unknown (Chandrasekaran & Betrán, 2008). Understanding these processes can help us evaluate how host-virus interactions in evolution have shaped bats' antiviral immunity and provide insights into translating bats' inflammation-controlling mechanisms to humans. It would also assist with better prediction and prevention of zoonotic diseases from bats through understanding how some viruses, but not others, persist in bats. Thus, further studies are necessary to identify and characterise such genetic innovations in bats' evolutionary history.

One promising direction to investigate this question is through the lenses of paleovirology – the study of ancient viruses. Some viruses heritably integrate their genome into their host's genome in their life cycle, resulting in endogenous viral elements (EVEs) in the host genome. EVEs form a genomic fossil record of ancient infections and are widely identified; a diverse range of EVEs similar to retroviruses, filoviruses, bornaviruses and parvoviruses have been identified in bats (Skirmuntt *et al*, 2020). While most EVEs become genomic relics without function, some were captured and repurposed by the host for new functions – termed gene co-option or exaptation (True & Carroll, 2002; Gould & Vrba, 1982). A notable function of co-opted EVEs is EVE-derived immunity (EDI), in which EVEs are co-opted as components of the host's immunity (Aswad & Katzourakis, 2012). EDI genes have been identified and characterised in many vertebrates. They restrict the virus life cycle by blocking cell entry, interfering with viral integration, or mediating immune tolerance (Aswad & Katzourakis, 2012). For example, Friend virus susceptibility 1 (Fv1), a co-opted retroviral group-specific antigen (Gag) protein, interferes with gammaretrovirus replication in mice by interacting with incoming viral capsid proteins (Yap *et al*, 2014). In chicken, endogenous avian leukosis viruses mediate receptor blocking (Tikhonenko & Lomovskaya, 1990). Potential EDI genes were also identified in bats. Ebolavirus- and Marburg virus-related EVEs are conserved in different bat genomes; they might contribute to these viruses' decreased virulence in bats (Taylor *et al*, 2010; Skirmuntt *et al*, 2020). Apart from host defence, EVEs also participate in immune system evolution. Chuong *et al* (2016) found that endogenous

viruses in different mammalian lineages have led to the evolution of lineage-specific interferon enhancers – potentially due to responses to different viral infections. Thus, considering how EVEs can participate in antiviral immunity, studying potential co-opted EVE-derived genes in bat genomes holds a two-fold benefit. Not only can we identify novel EDI genes underlying bats' antiviral immunity and add to our understanding of bats' zoonotic virus reservoir status, but we can also characterise how ancient host-virus interactions have shaped the evolution of this immunity.

Among retroviral genes, the *gag* gene – which contains the retrovirus's structural proteins and mediates its assembly – accounts for a significant proportion of EVE co-option events in vertebrates and shows the propensity to act as EDI genes through viral restriction, with *Fv1* as an example (Wang & Han, 2020; Yap *et al*, 2014). However, while Skirmuntt & Katzourakis (2019) have recently investigated endogenous retroviral envelope (*env*) genes in bats, co-opted retroviral *gag* genes in bats have not been specifically studied. Thus, I focused on characterising putative *gag*-derived EDI genes in bats in this study.

I report that Gag Co-option Event 32 (*GCE32*), a virus-host fusion gene conserved in megabats (family Pteropodidae), is an actively expressed putative EDI gene that helps us understand ancient host-virus interactions. First, I screened the bat-specific putative co-opted *gag* genes identified by Wang & Han (2020), who searched for such genes in over 700 genomes. I singled out *GCE32* – a fusion gene in megabats – as a promising EDI gene candidate. I then focused on *GCE32*, confirming its conservation in all available megabat genomes, studying its evolutionary history, and modelling ancient host-virus interactions with structural predictions of reconstructed viral proteins. I further investigated the selective pressure in *GCE32* evolution and characterised its tissue-specific expression pattern in multiple megabat species. These analyses suggest that *GCE32* might function in antiviral immunity and shed light on complicated host-virus interactions that might have shaped this immunity in evolution.

METHODS

Overview

Firstly, to select putative EDI gene candidates, I screened all previously identified bat-specific putative co-opted *gag* genes (Wang & Han, 2020). After identifying *GCE32* as the primary candidate, I used various computational methods to investigate *GCE32's* evolutionary history and potential function (**Figure 1**). I used similarity search tools to annotate orthologous EVE insertions on the *GCE32* locus in megabats. To investigate *GCE32*'s evolutionary history, I built phylogenetic trees of *GCE32* with related viral and host genes with maximum-likelihood approaches. I also identified *GCE32*-related EVEs – such as *GCE32a* – and mapped conserved protein domains on these EVEs with character mapping approaches. To model *GCE32*-related ancient host-virus interactions, I reconstructed retroviral protein sequences with similar sequences from *GCE32*-related EVEs. I used machine learning-based tools to predict their structures and align them with present-day viral proteins. I also used maximum-likelihood approaches to evaluate the selective pressure acting on *GCE32*. To characterise *GCE32*'s expression pattern, I examined publicly available transcriptomic datasets in different megabat species.

Screening for *gag*-derived EDI candidates in bat genomes

To identify *gag*-derived gene candidates co-opted for EDI in bats, I screened the list of *gag* co-option events identified by Wang & Han (2020). To identify more orthologs of these events in other bat species, I conducted a BLASTn web search for each co-option event. Then, to confirm the orthology of my results, I conducted two synteny analyses. I examined (1) whether these elements' nearest coding genes upstream and downstream are orthologs and (2) whether these elements have very similar 2000-base-pair-long flanking sequences upstream and downstream (>90% sequence identity). These analyses complement each other and provide extra certainty if one fails due to insufficiently long genomic scaffolds or genomic reorganisation. After identifying orthologous sequences for each *gag* co-option event, I used the NCBI ORF finder to identify their coding sequence (Wheeler *et al*, 2003). I also used the RepeatMasker webserver to confirm their retroviral origin (Smit *et al*, 2013-2015; Strategy – speed: quick; DNA source: general mammal). To identify the conserved protein domains in each co-option event, I used NCBI conserved domain search and recorded significant hits (Lu *et al*, 2020).

Annotating *GCE32* orthologs and the retroviral insertion on the *GCE32* locus

After picking *GCE32* as the candidate for further analyses, I noticed 2-nucleotide-long deletions in the putative *GCE32* coding sequence in *Pteropus alecto* T., *Eidolon helvum* K., and *Eonycteris spelaea* D., resulting in a frameshift. Since sequencing error is likely for these frameshifts due to the conservation of the coding sequence after these frameshifts, I excluded them from further analyses.

To understand more about the original retroviral insertion into the genome on the *GCE32* locus (referred below as *GCE32* EVE sequence), I defined the insertion's boundary by comparing megabat genomes to two non-megabat genomes (*Hipposideros armiger* H. and *Rhinolophus ferrumequinum* S.). First, I identified the orthologous loci in these species by taking the sequence flanking the retroviral insertion (±10kb). Then, I aligned these orthologous loci and extracted the sequences uniquely found in megabats as the *GCE32* EVE sequence (see **Table 2B**).

To systematically identify viral gene relics in the *GCE32* EVE sequence, I used Exonerate v2.4.0, a pairwise sequence alignment tool (Slater & Birney, 2005). For the protein subject, I used the *gag*, *pol* (polymerase) and *env* genes of the human chronic myeloid leukemia-associated retrovirus (HCML-ARV) (see **Appendix Table 2**). For the nucleotide query, I reconstructed an ancestral sequence of the viral insertion with IQ-TREE v1.6.12, with the *GCE32* EVE sequences in megabats as input (Nguyen *et al*, 2015). I used the resulting *gag*, *pol* and *env*-related sequences to conduct downstream phylogenetic analysis and viral protein reconstruction.

Phylogenetic analysis

I used different computational tools to construct phylogenetic trees depending on sequence types and sizes (**Table 1**). Firstly, I aligned candidate sequences with MAFFT v7.490 (Katoh & Stadtley, 2013). I then trimmed the sequences with TrimAl 1.2rev57 before manual trimming and inspection (Capella-Gutiérrez *et al*, 2009). Finally, I constructed phylogenetic trees with IQ-TREE v1.6.12 or PhyML 3.3, using either the IQ-TREE-derived ModelFinder or modeltest-ng v0.1.7 to identify the best nucleotide/amino acid substitution model (Nguyen *et al*, 2015; Guindon *et al*, 2010; Kalyaanamoorthy *et al*, 2017; Darriba *et al*, 2020). I also conducted traditional or ultrafast bootstrap analysis depending on alignment size (Felsenstein, 1985; Hoang *et al*, 2018).

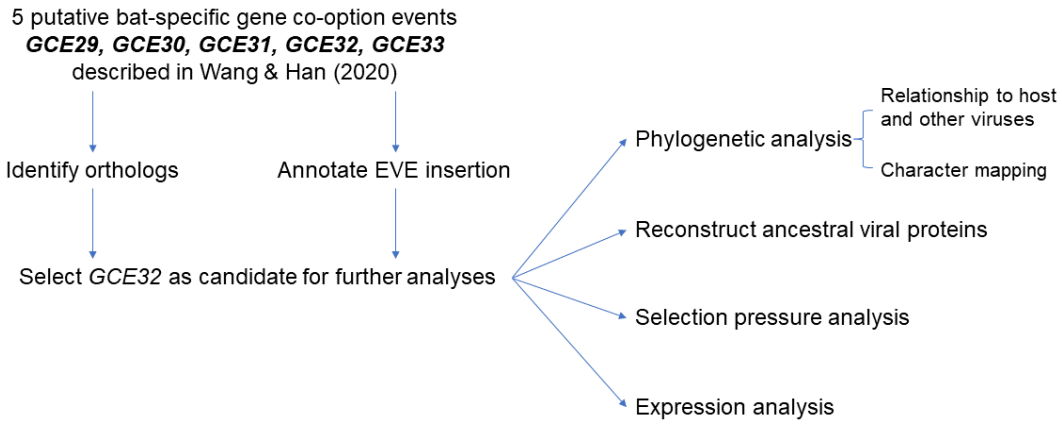


Figure 1: the workflow of this study. First, I isolated *GCE32* from a list of 5 putative bat-specific gene co-option events after preliminary analyses. Then, I used multiple computational analysis methods to probe into ancient host-virus interactions and speculate on the potential function of *GCE32*.

Table 1: different computational tools were used to construct phylogenetic trees depending on sample types and sizes. The specific options used for the programmes are denoted in parentheses; the version of these programmes are recorded in the main text. AIC, Akaike information criterion; AICc, size-corrected Akaike information criterion.

Number of sequences	<50	>50
Alignment	MAFFT (L-ins-i/G-ins-i depending on nature of sequences)	MAFFT (FFT-NS-2)
Trimming	TrimAl (-gt 0.5)	TrimAl (-gt 0.5: phylogenetic analysis, -gt 0.1: viral protein reconstruction, -gt 0.3: character-mapping)
Model finding	modeltest-ng (best model based on AICc)	IQ-TREE ModelFinder for phylogenetic analysis; modeltest-ng (best model based on AIC) for viral protein reconstruction and character mapping
Tree construction	PhyML	IQ-TREE, up to 1000 SH-aLRT branch tests
Bootstrapping	100 bootstraps	1000 ultrafast bootstraps

Identifying *GCE32*-related EVEs

I used similarity search methods to identify *GCE32*-related EVEs. Firstly, I used BLASTn to search the genome of megabat species in the RefSeq_genome database with the putative *GCE32* coding sequence in *R. aegyptiacus* as the query (Altschul *et al*, 1990). *GCE32a* was identified as a sequence highly similar to *GCE32* in two of its domains. I also recovered and confirmed its orthologs in megabats using similarity search and synteny analyses described above (genomic coordinates in **Table 3A-B**; see **Figure 3A-B**).

To identify more EVEs similar to the *env* gene in the *GCE32* EVE sequence, I conducted a systematic tBLASTn web search in all mammals. I downloaded the top 100 results displayed and filtered alignments with 10^{-10} as the e-value cut-off. I isolated these alignments' upstream and downstream sequences (±10kb) if the genomic scaffold was sufficiently long (Appendices, get_coors_char.ipynb). To ensure I captured no identical sequences, I re-extracted the *env* gene sequences from these downloaded sequences with Exonerate v2.4.0, filtering out EVEs with *env* sequences identical to previously-identified sequences (Slater & Birney, 2005; Appendices, noid.ipynb). I used these *env* sequences for further phylogenetic analysis as described below.

Character mapping

To investigate the gain of conserved protein domains of the *GCE32* EVE sequence in evolution, I mapped the presence of conserved domains in *GCE32*-related EVEs that I identified. To do this, I first identified the existing conserved protein domains with NCBI Batch Web Conserved Domain Search Tool with these EVEs' 6-frame translation (Lu *et al*, 2020). Then, to confirm the statistically significant presence of these domains, I used these sequences' 6-frame translation for an HMMsearch in HMMER v3.3.2, with the same nucleotide sequence as the subject and the Hidden Markov Model for each conserved domain from Pfam 35.0 as the query (Appendices, parse_trees.ipynb; Mistry *et al*, 2021; Eddy, 2011; see **Figure 3D**). I considered results with an e-value smaller than 10^{-3} as positive results. After that, I mapped the results on a phylogenetic tree of these EVEs that I constructed with IQ-TREE v1.6.12, using aligned sequences from *env* genes (Appendices, parse_trees.ipynb).

Viral protein reconstruction and alignment

To predict the structures of Gag, Pol and Env viral polyproteins of the *GCE32* EVE sequence, I extracted gene sequences from *GCE32*-related EVEs and reconstructed their ancestral sequences. Firstly, I used the *gag* sequence from *GCE32* and the *pol* and *env* sequences from *GCE32a* for a tBLASTn web search to identify mammalian *GCE32*-related EVEs. Then, I downloaded the top 500 results displayed and selected the alignment with the smallest e-value from each result if it was lower than 10^{-10}. I isolated these alignments' upstream and downstream sequences (±10kb) if the genomic scaffold was sufficiently long (Appendices, get_coors_recon.ipynb). After that, I extracted putative viral genes from these sequences with Exonerate v2.4.0, with HCML-ARV *gag*, *pol*, and *env* as templates (see **Appendix Table 2**). Finally, I aligned these sequences with the HCML-ARV genes (Slater & Birney, 2005; Appendices, parse_vulgar.ipynb).

To reconstruct the ancestral sequence of these multiple-sequence alignments, I added an outgroup to root the phylogenetic tree (Reticuloendotheliosis virus *gag*, *pol*, and *env*; see **Appendix Table 2**). I then aligned and trimmed them as described above. Next, I used modeltest-ng v0.1.7 to select substitution models (AICc criterion) (Darriba *et al*, 2020). Finally, I used the ancestral sequence reconstruction feature in IQ-TREE v1.6.2 (-asr) with the specified substitution model to reconstruct ancestral sequences and retrieved the sequence from the ancestral node of all ingroup sequences (Nguyen *et al*, 2015).

To pin down the exact boundary of reconstructed viral proteins used for structural prediction, I combined NCBI conserved domain search results, alignment with exogenous viral proteins with previously-resolved crystal structures, and manual inspection (Lu *et al*, 2020). I predicted these proteins' structures with ColabFold v1.3, an online server version of AlphaFold 2 (Jumper *et al*, 2020; Mirdita *et al*, 2021). I generated five prediction models with 48 recycles and Assisted Model Building with Energy Refinement (AMBER) for each sequence (Hornak *et al*, 2006). For multimers longer than 1000 amino acids, I generated five structural prediction models with 24 recycles. I predicted the proteins' multimeric structure with SWISS-MODEL (Waterhouse *et al*, 2018). To align protein structures, I used the MUSTANG algorithm in YASARA v21.12.18 (Konagurthu *et al*, 2006; Krieger & Vriend, 2014).

Selective pressure analysis

I used PAML 4.9j to infer dN/dS ratios for *GCE32* coding sequences (Yang, 2007). dN/dS ratio is the ratio of the number of non-synonymous substitutions per non-synonymous site (assumed to be experiencing selection) to the number of synonymous substitutions per synonymous site (assumed to be neutral) in a coding sequence (Kryazhimskiy & Plotkin, 2008). It measures the intensity of selective pressure acting on a sequence in a specified time. dN/dS is expected to exceed one if the sequence undergoes positive selection, below one if conserved, and equal one if it evolves neutrally (but dN/dS = 1 does not prove neutrality since it can be averaged out from positively and negatively selected sites).

I used two types of models – site and branch models – to answer different questions on *GCE32*'s selective pressure. To investigate the dN/dS ratios of *GCE32* domains and infer positively selected amino acid sites, I used site models that fix dN/dS among phylogenetic tree branches but allow variance among codon sites (Yang *et al*, 2000). I first attempted to reject the sequence is neutrally evolving with a simple site model assuming dN/dS = 1. I then conducted likelihood ratio tests on two pairs of site models (NSSites M1a-M2a and M7-M8) to infer dN/dS and positively selected sites as they ensure the robustness of the results (Yang *et al*, 2000; Yang *et al*, 2005; see **Table 4A-B**). To infer the selective pressure on different phylogenetic tree branches, I used branch models that fix dN/dS among sites and allow dN/dS to vary among branches (Yang, 1998; Yang & Nielson, 1998). I used likelihood ratio tests to guide model selection (see **Table 4A, C**).

Gene expression analysis

To characterise *GCE32*'s expression pattern, I examined all available transcriptomic datasets in the NCBI Sequence Read Archive (SRA) of megabats from multiple tissues and species, including 11 *R. aegyptiacus* tissue types (Lee *et al*, 2015; see **Appendix Table 4**). First, I retrieved the accession numbers of all transcriptomic datasets in these species. Next, I conducted an SRA-BLAST search for all searchable megabat transcriptomic datasets with the *GCE32* sequence from the respective species as the query (Altschul *et al*, 1990). I then recorded if *GCE32*-matching transcripts were observed in these datasets, dividing them into three categories: no expression (no matching reads with ≥98% sequence identity), low-coverage positive [≥1 read(s) with ≥98% sequence identity], and high-coverage positive (≥2x coverage of reads with ≥98% sequence identity on ≥50% of the entire sequence) (see **Table 5B-C**). I assembled the *GCE32*-matching transcripts *de novo* with TransAbyss v2.0.1 (Robertson *et al*, 2010).

RESULTS

GCE32 is a virus-host fusion gene that evolved from extensive host-virus interactions

Gag Co-option Event (*GCE*) *29-33* are five bat-specific *gag* co-option events identified by Wang & Han (2020). I conducted preliminary analyses to investigate their basic characteristics, including synteny, coding sequence conservation, and conserved protein domains (**Appendix Table 1**). Among these genes, *GCE32* was described as a *gag* co-option event that is 'fused with host genes' in five megabat species (Wang & Han, 2020). I chose it for further analyses because it is well-conserved in megabats, consists of non-viral conserved domains that might provide insights into host-virus interactions, and might function as an EDI gene.

GCE32 is uniquely conserved in megabats with open reading frames longer than 2,500 nucleotides. I examined the orthologous loci of 10 megabat species, three non-megabat bat species, and the pig, identifying *GCE32* and the corresponding EVE in all megabat genomes. I confirmed their orthology in 8 out of 10 species by identifying neighbouring genes and flanking sequences. The genomic scaffold was too short to identify neighbouring genes in the remaining two species, but I confirmed their orthology with flanking sequences. In contrast, I did not find *GCE32* in non-megabat species, only identifying empty pre-integration sites at their orthologous loci (**Figure 2A**; **Table 2A**).

The EVE insertion on the *GCE32* locus, including the *GCE32* coding sequence, has several conserved domains similar to viral or host genes (**Figure 2B**; **Table 2B**). I investigated these domains' evolutionary history by comparing them with similar viral or host gene sequences. The *env* gene of the insertion is closely related to exogenous and endogenous gammaretroviruses. Its closest relatives are human endogenous retrovirus-E (HERV-E) and HCML-ARV (bootstrap support 64%; **Figure 2C**; **Appendix Table 2**).

Apart from viral domains, the *GCE32* coding sequence has three other domains highly similar to host genes *RBM33*, *CRKL*, and *PEG3,* respectively (sequence identity 76.77%, 79.61%, and 75% in the aligned region) (**Figure 2B**). To investigate whether the *GCE32* sequence was derived from these genes but not other genes, I constructed phylogenetic trees with these genes' orthologs in multiple mammalian species (**Appendix Table 3A-C**). The aligned *GCE32* domains cluster with the megabat orthologs in all phylogenetic trees (**Figure 2D-F**). These results suggest that these *GCE32* domains were likely captured from the corresponding host genes.

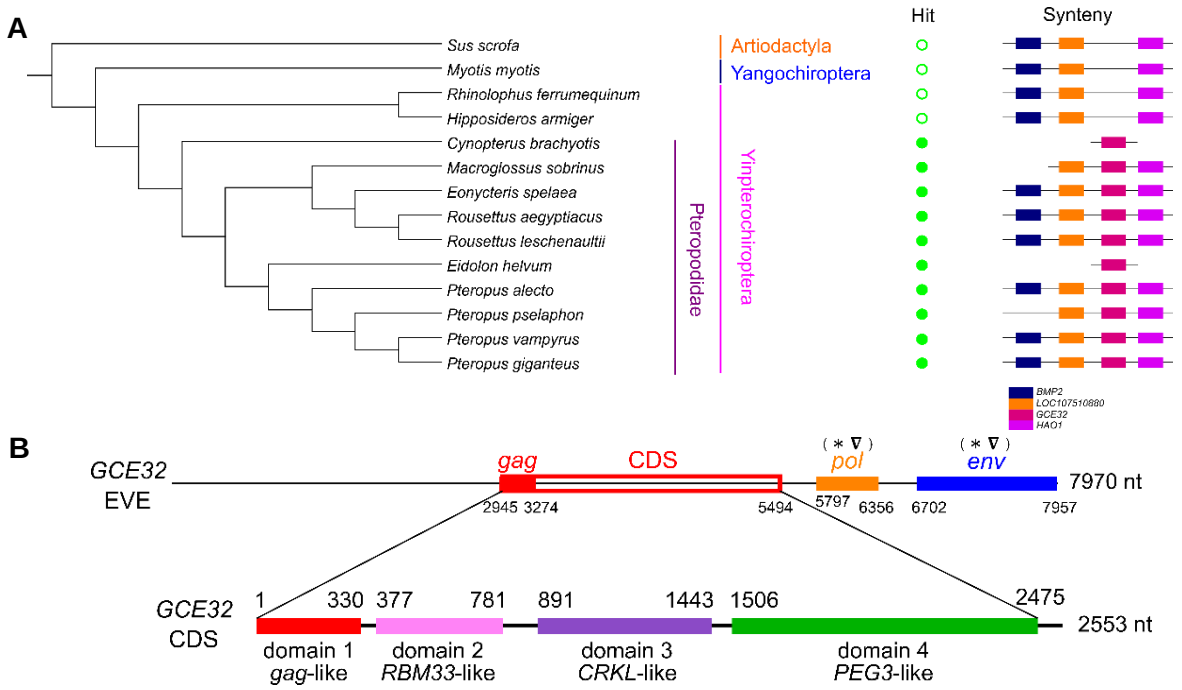


Figure 2: A, *GCE32* orthologs are present uniquely in megabats (family Pteropodidae). The presence or absence of *GCE32* and its neighbouring genes on the same genomic scaffold in 10 megabat species, three bat species of other families, and the pig are shown as solid rectangles on black lines. Shorter black lines indicate that the scaffold is not long enough to capture the neighbouring gene(s). *BMP2*, bone morphogenetic protein 2; *HAO1*, hydroxy-acid oxidase 1; *LOC107510880*, uncharacterised locus 107510880. Maximum-likelihood phylogeny is adapted from Almeida *et al* (2020) and Lei & Dong (2016). **B**, the EVE on the *GCE32* locus contains multiple conserved domains similar to host and viral sequences. The *pol* and *env* domains are interrupted by mutations and are thus inactive. nt, nucleotides; CDS, coding sequence. Inactive gene remnants are marked by symbols (* ∇). **C** (next page), the phylogenetic relationship between the *env*-derived sequence in *GCE32* orthologs and aligned *env* sequences from selected exogenous and endogenous gammaretroviruses (see **Appendix Table 2** for full names). The exogenous gammaretrovirus group highlighted in blue is non-exhaustive. *Env*-derived sequences were used since they were best conserved among the viral gene sequences. **D** (next page), the phylogenetic relationship between *GCE32* domain 2 and *RBM33* orthologs. *RBM33*, RNA-binding motif protein 33. **E** (next page), the phylogenetic relationship between *GCE32* domain 3 and related proteins. *CRKL*, *CRK*-like. **F** (next page), the phylogenetic relationship between *GCE32* domain 4 and related proteins. *PEG3*, paternally-expressed gene 3; ZNF568, zinc finger protein 568; *INSM1*, insulinoma-associated protein 1; *CTCF*, CCCTC-binding factor; *CTCFL*, *CTCF*-like; *AEBP2*, adipocyte enhancer-binding protein 2. Numbers at tree nodes are bootstrap values. Scale bars show phylogenetic distances between lineages measured by amino acid substitutions per site. Phylogenetic trees are outgroup-rooted in **C** and **D** and midpoint-rooted in **E** and **F**.

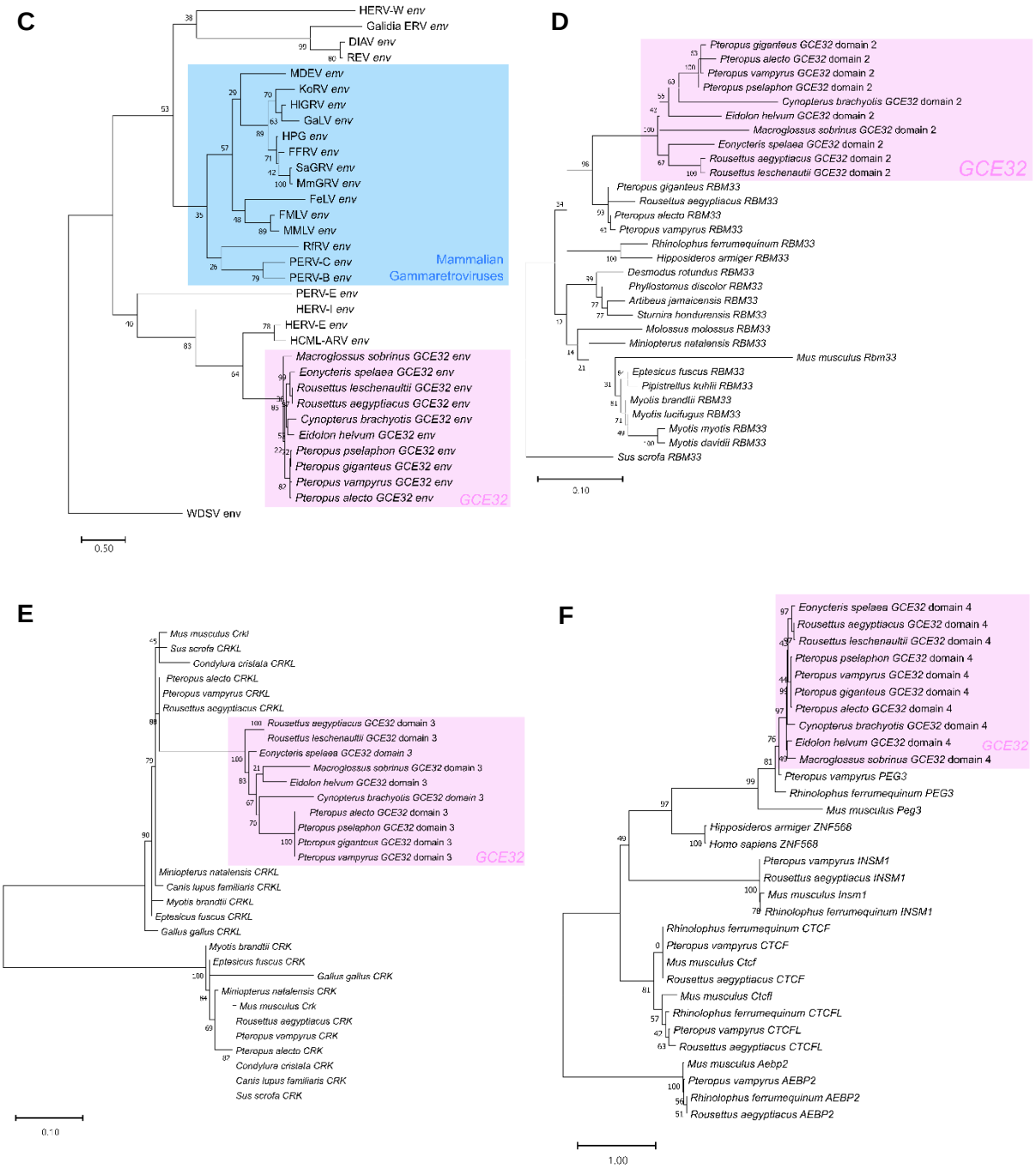

C
HERV-W env
Galidia ERV env
DIAV env
REV env
MDEV env
KoRV env
HIGRV env
GaLV env
HPG env
FFRV env
SaGRV env
MmGRV env
FeLV env
FMLV env
MMLV env
RfRV env
PERV-C env
PERV-B env
Mammalian Gammaretroviruses
PERV-E env
HERV-I env
HERV-E env
HCML-ARV env
Macroglossus sobrinus GCE32 env
Eonycteris spelaea GCE32 env
Rousettus leschenaultii GCE32 env
Rousettus aegyptiacus GCE32 env
Cynopterus brachyotis GCE32 env
Eidolon helvum GCE32 env
Pteropus pselaphon GCE32 env
Pteropus giganteus GCE32 env
Pteropus vampyrus GCE32 env
Pteropus alecto GCE32 env
GCE32
WDSV env
0.50
D
Pteropus giganteus GCE32 domain 2
Pteropus alecto GCE32 domain 2
Pteropus vampyrus GCE32 domain 2
Pteropus pselaphon GCE32 domain 2
Cynopterus brachyotis GCE32 domain 2
Eidolon helvum GCE32 domain 2
Macroglossus sobrinus GCE32 domain 2
Eonycteris spelaea GCE32 domain 2
Rousettus aegyptiacus GCE32 domain 2
Rousettus leschenaultii GCE32 domain 2
GCE32
Pteropus giganteus RBM33
Rousettus aegyptiacus RBM33
Pteropus alecto RBM33
Pteropus vampyrus RBM33
Rhinolophus ferrumequinum RBM33
Hipposideros armiger RBM33
Desmodus rotundus RBM33
Phyllostomus discolor RBM33
Artibeus jamaicensis RBM33
Sturnira hondurensis RBM33
Molossus molossus RBM33
Miniopterus natalensis RBM33
Mus musculus Rbm33
Eptesicus fuscus RBM33
Pipistrellus kuhlii RBM33
Myotis brandtii RBM33
Myotis lucifugus RBM33
Myotis myotis RBM33
Myotis davidii RBM33
Sus scrofa RBM33
0.10
E
Mus musculus Crkl
Sus scrofa CRKL
Condylura cristata CRKL
Pteropus alecto CRKL
Pteropus vampyrus CRKL
Rousettus aegyptiacus CRKL
Rousettus aegyptiacus GCE32 domain 3
Rousettus leschenaultii GCE32 domain 3
Eonycteris spelaea GCE32 domain 3
Macroglossus sobrinus GCE32 domain 3
Eidolon helvum GCE32 domain 3
Cynopterus brachyotis GCE32 domain 3
Pteropus alecto GCE32 domain 3
Pteropus pselaphon GCE32 domain 3
Pteropus giganteus GCE32 domain 3
Pteropus vampyrus GCE32 domain 3
GCE32
Miniopterus natalensis CRKL
Canis lupus familiaris CRKL
Myotis brandtii CRKL
Eptesicus fuscus CRKL
Gallus gallus CRKL
Myotis brandtii CRK
Eptesicus fuscus CRK
Gallus gallus CRK
Miniopterus natalensis CRK
Mus musculus Crk
Rousettus aegyptiacus CRK
Pteropus vampyrus CRK
Pteropus alecto CRK
Condylura cristata CRK
Canis lupus familiaris CRK
Sus scrofa CRK
0.10
F
Eonycteris spelaea GCE32 domain 4
Rousettus aegyptiacus GCE32 domain 4
Rousettus leschenaultii GCE32 domain 4
Pteropus pselaphon GCE32 domain 4
Pteropus vampyrus GCE32 domain 4
Pteropus giganteus GCE32 domain 4
Pteropus alecto GCE32 domain 4
Cynopterus brachyotis GCE32 domain 4
Eidolon helvum GCE32 domain 4
Macroglossus sobrinus GCE32 domain 4
GCE32
Pteropus vampyrus PEG3
Rhinolophus ferrumequinum PEG3
Mus musculus Peg3
Hipposideros armiger ZNF568
Homo sapiens ZNF568
Pteropus vampyrus INSM1
Rousettus aegyptiacus INSM1
Mus musculus Insm1
Rhinolophus ferrumequinum INSM1
Rhinolophus ferrumequinum CTCF
Pteropus vampyrus CTCF
Mus musculus Ctcf
Rousettus aegyptiacus CTCF
Mus musculus Ctcfl
Rhinolophus ferrumequinum CTCFL
Pteropus vampyrus CTCFL
Rousettus aegyptiacus CTCFL
Mus musculus Aebp2
Pteropus vampyrus AEBP2
Rhinolophus ferrumequinum AEBP2
Rousettus aegyptiacus AEBP2
1.00

Table 2A: the genomic coordinates and synteny of the *GCE32* coding sequence in 10 megabat species. Asterisks (*) mark the cases in which the genomic scaffold is too short to identify synteny by upstream and downstream neighbouring genes.

Gene	Species	Accession	Start	End	Synteny
GCE32	*Rousettus aegyptiacus* G.	NW_023416287	26269142	26289602	*BMP2-GCE32-HAO1*
GCE32	*Pteropus vampyrus* L.	NW_011888788	4571821	4594507	*BMP2-GCE32-HAO1*
GCE32	*Pteropus giganteus* B.	NW_024342471	416962	436924	*BMP2-GCE32-HAO1*
GCE32	*Pteropus alecto* T.	NW_006442491	13126387	13145463	*BMP2-GCE32-HAO1*
GCE32	*Pteropus pselaphon* L.	BMBI01006887	648836	668902	*LOC107510880-GCE32-HAO1**
GCE32	*Eidolon helvum* K.	AWHC01227563	3091	23760	*
GCE32	*Rousettus leschenaultii* D.	BNJM01000014	1703910	1724216	*BMP2-GCE32-HAO1*
GCE32	*Eonycteris spelaea* D.	PUFA01000110	4234666	4255111	*BMP2-GCE32-HAO1*
GCE32	*Macroglossus sobrinus* A.	PVKZ01000650	481912	502279	*LOC107510880-GCE32-HAO1**
GCE32	*Cynopterus brachyotis* M.	SSHV01047299	3133	22991	*

Table 2B: the genomic coordinates of the EVE insertion on the *GCE32* locus in 10 megabat species based on the comparison to two non-megabat species (R_ferru, *Rhinolophus ferrumequinum*; H_armiger, *Hipposideros armiger*). Asterisk-marked (*) coordinates are used to isolate the EVE sequence and reconstruct the ancestral sequence to generate the coordinates in **Figure 2B**.

Species	Accession	R_ferru_start	R_ferru_end	H_armiger_start	H_armiger_end
Cynopterus brachyotis	SSHV01047299	14228*	23266*	15894	23265
Eidolon helvum	AWHC01227563	16074*	24024*	16630	24009
Eonycteris spelaea	PUFA01000110	4234389*	4242265*	4234392	4242265
Macroglossus sobrinus	PVKZ01000650	481632*	489388*	481639	489388
Rousettus aegyptiacus	NW_023416287	26268865*	26276787*	26268866	26276787
Rousettus leschenaultii	BNJM01000014	1703633*	1711796	1703634	1714165*
Pteropus alecto	NW_006442491	13139326*	13146589*	13139326	13146574
Pteropus giganteus	NW_024342471	429927*	437198*	429927	437198
Pteropus pselaphon	BMBI01006887	648563*	655839*	648578	655839
Pteropus vampyrus	NW_011888788	4571548*	4578818*	4571563	4578818

GCE32-related EVEs provide insights into ancient viruses

To further probe into *GCE32*'s evolutionary history, I searched for *GCE32*-related EVEs in the megabat genome. *GCE32a* is another EVE uniquely conserved in megabats highly similar to *GCE32* (>85% identity) (**Figure 3A**; **Table 3A-B**). It has better-conserved *pol* and *env* gene relics and SH2 and SH3 conserved domain remnants, a characteristic of the *CRKL*-like domain in *GCE32* (**Figure 3B**). *GCE32a* also contains a CC2-LZ conserved domain typically found in the NF-κB essential modulator (NEMO), which regulates NF-κB expression (Grubisha *et al*, 2010).

To understand how *GCE32* gained its distinctive host-derived domains in evolution, I used the *GCE32 env*-derived sequence to systematically search for *GCE32*-related EVEs in all mammalian species. I identified their conserved domains and constructed a phylogenetic tree with them, mapping the presence of non-viral conserved domains onto it (**Figure 3C**). All non-virus conserved domains identified in *GCE32* and *GCE32a* – SH2, SH3, and CC2-LZ – are clustered in a megabat-specific monophyletic branch, suggesting they were acquired uniquely in megabats. To understand the conserved domain arrangement of the original virus that integrated at the *GCE32* locus, I also identified the *gag* capsid domain in some of these EVEs. This domain is absent in *GCE32* and *GCE32a*, which only have the *gag* matrix domain. Notably, by mapping the arrangement of these domains in some *GCE32*-related EVEs, I found that the host-derived domains (SH2, SH3, CC2-LZ) fall closely between the virus-derived *gag* matrix and capsid protein domains (**Figure 3D**). This finding suggests that the ancient virus might have already contained these host-derived domains as both matrix and capsid proteins are required for viral replication. In contrast, there are no *RBM33*- or *PEG3*-like conserved domains in these *GCE32*-related EVEs.

To identify if any other *GCE32*-related EVEs might be active in megabat genomes, I further conducted an open reading frame (ORF) analysis. There are 7 ORFs longer than 900 base pairs among the EVEs in **Figure 3C** (marked 1-7; genomic coordinates in **Table 3C**). Apart from *GCE32* and *GCE32a*, two other ORFs in *P. alecto* contain at least part of the SH2 and SH3 domains (Nos. 2, 3 in **Table 3C**). These two sequences were annotated as pseudogenes (LOC112478140, LOC112478141), but neither sequence has transcriptomic coverage. This finding suggests that there might be more active members of the *GCE32*-related gene family, but it is unknown under which conditions they are expressed.

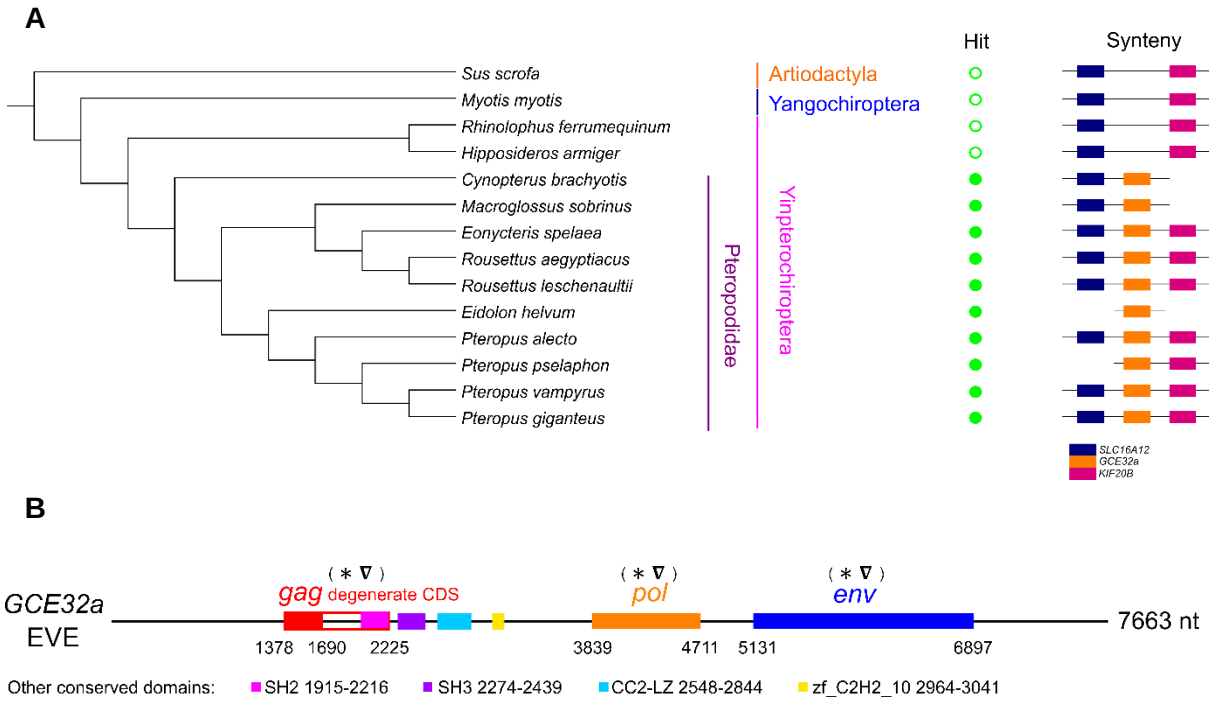
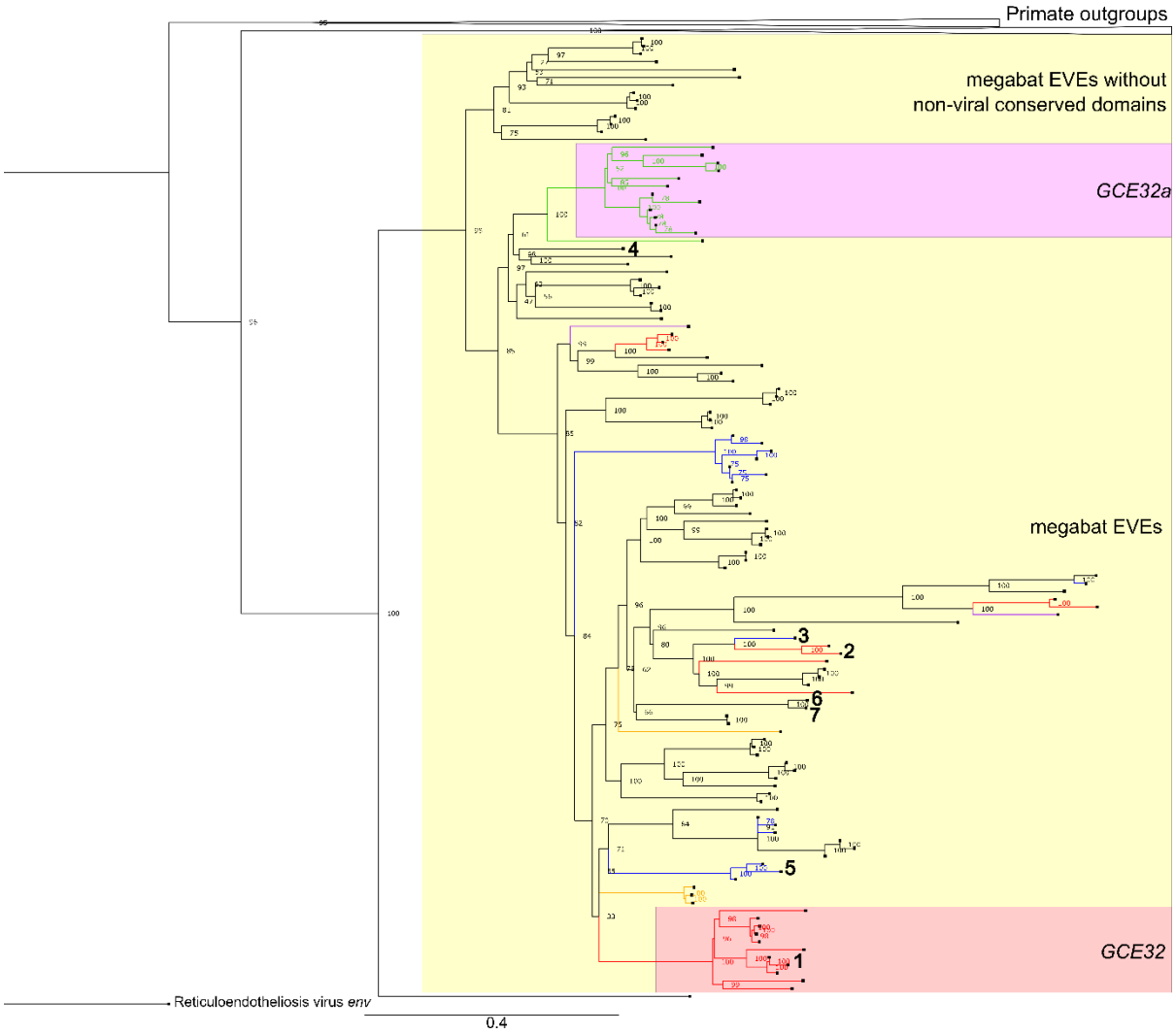


Figure 3: A, the conservation and synteny of *GCE32a* in megabat and non-megabat species. The presence or absence of *GCE32a* and its neighbouring genes on the same genomic scaffold in 10 megabat species, three bat species of other families, and the pig are shown as solid rectangles on black lines. Shorter black lines indicate that the scaffold is not long enough to capture the neighbouring gene(s). *SLC16A12*, solute carrier family 16 member 12; *KIF20B*, kinesin family member 20B. **B**, the coordinates of *gag*, *pol*, *env* viral gene domains, and other conserved protein domains identified within *GCE32a* EVE. nt, nucleotides. Degenerate CDS refers to a premature stop codon compared to *GCE32*. zf_C2H2_10 is a conserved domain of a zinc finger protein. Inactive gene remnants are marked by symbols (* ∇). **C** (next page), the phylogenetic tree of *GCE32*-related EVEs in megabats and primates, with the presence of SH2, SH3 and CC2-LZ conserved domains in different *GCE32*-related EVEs in megabats marked by different colours (blue = SH2 only, purple = SH3 only, orange = CC2-LZ only, red = SH2 and SH3, lavender = SH2 and CC2-LZ, green = SH2, SH3, and CC2-LZ). The phylogenetic tree was generated from the aligned *env* sequences of these EVEs, with the aligned sequence from the reticuloendotheliosis virus as the outgroup (distance not shown). The scalebar shows phylogenetic distances between EVEs measured by amino acid substitutions per site. **D** (next page), the genomic arrangement of conserved domains identified in selected *GCE32*-related EVEs and *GCE32a*, showing the host domains (SH2, SH3, and CC2-LZ) falling between the viral *gag* matrix and capsid proteins.

C



D

Species	Accession	Genomic coordinates		Conserved domains (drawn to scale)
		Start	End	
Rousettus aegyptiacus	NW_023416287	50902709	50922278	
Rousettus aegyptiacus	NW_023416306	82144399	82165256	
Pteropus giganteus	NW_024354550	417066	437743	
Pteropus vampyrus	NW_011889107	2244124	2264459	
Pteropus alecto	NW_006430462	23933	43028	
Pteropus vampyrus	NW_011889032	1838766	1858969	
Pteropus vampyrus	NW_011888788	1219310	1239914	
Pteropus alecto	NW_006442491	9660592	9680885	
Rousettus aegyptiacus	NW_023416287	41244309	41263485	
Pteropus giganteus	NW_024351546	2458501	2478785	
Rousettus aegyptiacus	NW_023416309	*GCE32a*		

gag matrix SH2 SH3 CC2-LZ *gag* capsid

Table 3A: the genomic coordinates and synteny of the longest coding sequence in the *GCE32a* EVE in 10 megabat species. Asterisks (*) mark the cases in which the genomic scaffold is too short to identify synteny by upstream and downstream neighbouring genes.

Gene	Species	Accession	Start	End	Synteny
GCE32a	*Cynopterus brachyotis*	SSHV01005295	136795	137241	*GCE32a-KIF20B**
GCE32a	*Rousettus aegyptiacus*	NW_023416309	32243510	32244360	*SLC16a12-GCE32a-KIF20B*
GCE32a	*Rousettus leschenaultii*	BNJM01000102	167252	168102	*SLC16a12-GCE32a-KIF20B*
GCE32a	*Pteropus alecto*	NW_006494611	16670796	16671655	*SLC16a12-GCE32a-KIF20B*
GCE32a	*Pteropus pselaphon*	BMBI011006047	389876	390735	*GCE32a-KIF20B**
GCE32a	*Pteropus giganteus*	NW_024344378	28066399	28067258	*SLC16a12-GCE32a-KIF20B*
GCE32a	*Pteropus vampyrus*	NW_011888849	3717749	3718608	*SLC16a12-GCE32a-KIF20B*
GCE32a	*Eidolon helvum*	AWHC01191639	2096	2954	*
GCE32a	*Eonycteris spelaea*	PUFA01000278	104032	104882	*SLC16a12-GCE32a-KIF20B*
GCE32a	*Macroglossus sobrinus*	PVKZ01003010	12430	13283	*SLC16a12-GCE32a**

Table 3B: the genomic coordinates of the EVE insertion on the *GCE32a* locus in 10 megabat species based on the comparison to two non-megabat species (R_ferru, *Rhinolophus ferrumequinum*; H_armiger, *Hipposideros armiger*). Asterisk-marked (*) coordinates are used to isolate the EVE sequence and reconstruct the ancestral sequence to generate the coordinates in **Figure 3B**.

Species	Accession	R_ferru_start	R_ferru_end	H_armiger_start	H_armiger_end
Cynopterus brachyotis	SSHV01005295	135393	142295	133962*	142898*
Eidolon helvum	AWHC01191639	Not found	8348	701*	8481*
Eonycteris spelaea	PUFA01000278	98593	105629	98458*	106290*
Macroglossus sobrinus	PVKZ01003010	11286	18780*	10427*	18648
Rousettus aegyptiacus	NW_023416309	32238109	32245062	32237974*	32245768*
Rousettus leschenaultii	BNJM01000102	161855	168842	161719*	169501*
Pteropus alecto	NW_006494611	16669412*	16677169*	16670092	16677040
Pteropus giganteus	NW_024344378	28065710	28072644	28064819*	28072773*
Pteropus pselaphon	BMBI01006047	389173	396123	388283*	396252*
Pteropus vampyrus	NW_011888849	3716945	3723990	3716158*	3724119*

Table 3C: the genomic coordinates, length, and conserved domains of the open reading frames (ORFs) over 900 base pairs long identified in *GCE32*-related EVEs (numbered 1-7 on the phylogenetic tree in **Figure 3C**). ORF coordinates refer to the position within the specified EVE. The No. 1 sequence is *R. aegyptiacus GCE32*.

No.	Species	Accession	EVE coordinates	ORF coordinates	ORF length	Conserved domains
1	*Rousettus aegyptiacus*	NW_023416287	26259949-26279224	13894-11348	2547	gag_ma, SH2, SH3, zf-C2H2, Ribosomal L40e
2	*Pteropus alecto*	NW_006434824	185090-204254	14116-13148	969	gag_ma, SH2, SH3
3	*Pteropus alecto*	NW_006434824	64232-84891	5781-6872	1092	gag_ma, SH2, SH3
4	*Pteropus alecto*	NW_006438113	13769639-13788941	13899-12904	996	pepsin-retropepsin, RT-like
5	*Pteropus alecto*	NW_006435779	2119665-2140348	16904-17821	918	None
6	*Pteropus giganteus*	NW_024355205	38889836-38909426	13008-14177	1170	None
7	*Pteropus vampyrus*	NW_011889077	614021-634431	4976-3756	1221	No significant conserved domains

Reconstructed viral proteins from the *GCE32* insertion are structurally highly similar to present-day viral proteins

To characterise the nature of host-virus interactions when the ancient virus on the *GCE32* locus integrated into the megabat genome, I reconstructed its original *gag, pol* and *env* genes using sequences of other more conserved *GCE32*-related EVEs. I predicted four viral protein structures from these sequences – Gag matrix protein, Gag capsid protein, Pol reverse transcriptase (RT) N-terminal domain, and Env receptor-binding domain (RBD) – and aligned them to their respective exogenous viral proteins. I used the Root-mean-square deviation of the alpha carbon* atomic coordinates (C-alpha RMSD) to measure the similarity between aligned structures. Generally, C-alpha RMSD smaller than 2.5Å is considered a good fit in structural modelling (Tsai *et al*, 2004).

I first inspected the predicted Gag matrix protein. The initially predicted sequence did not include the first 13 amino acids of the *GCE32*-derived *gag* sequence since the query sequence from HCML-ARV lacked these amino acids. Thus, I added them back to the reconstructed Gag matrix protein sequence and predicted the structure of the region that aligns with the Moloney murine leukemia virus (MoMLV) matrix protein (RCSB: 1MN8). Although the amino acid sequences are divergent (38.27% sequence identity), the predicted protein aligns well with the MoMLV matrix protein, with all secondary structural alignment observed (**Figure 4A-B**; C-alpha RMSD 1.189Å, 81 aligned residues).

For the capsid protein, Qu *et al* (2018) observed that murine leukemia virus (MLV) capsid protein multimerize into pentamers or hexamers depending on the virus's life cycle stages (RCSB: 6HWX hexamer, 6HWY pentamer). Homology prediction revealed that the multimeric structures of the reconstructed capsid protein sequence are highly similar to MLV's resolved capsid protein structures (**Figure 4C-D**; hexamer: C-alpha RMSD 0.151Å, 3669 aligned residues, 52.98% sequence identity; pentamer: C-alpha RMSD 0.213Å, 4099 aligned residues, 52.70% sequence identity). I also predicted a high-confidence monomer structure of the capsid protein (**Figure 4E-F**).

To further model ancient host-virus interactions, I predicted the structure of the Env RBD and the Pol RT N-terminal domain of the reconstructed virus since their resolved crystal structures are available. The alignment between the reconstructed Env RBD with Friend murine leukemia virus (FMLV) Env RBD was of high quality despite low sequence similarity (**Figure 4G**; RCSB: 1AOL; C-alpha RMSD 1.940Å, 110 aligned residues, 20.91% sequence identity). However, despite the overall high quality, the rapidly evolving host-specific domain is poorly modelled. In comparison, the Pol reverse transcriptase N-terminal domain is more conserved as modelled to its MoMLV counterpart, and the alignment quality

*The first carbon atom in an amino acid attached to a functional group.

is also high (**Figure 4H**; RCSB: 1I6J; C-alpha RMSD 0.939Å, 235 aligned residues, 67.23% sequence identity).

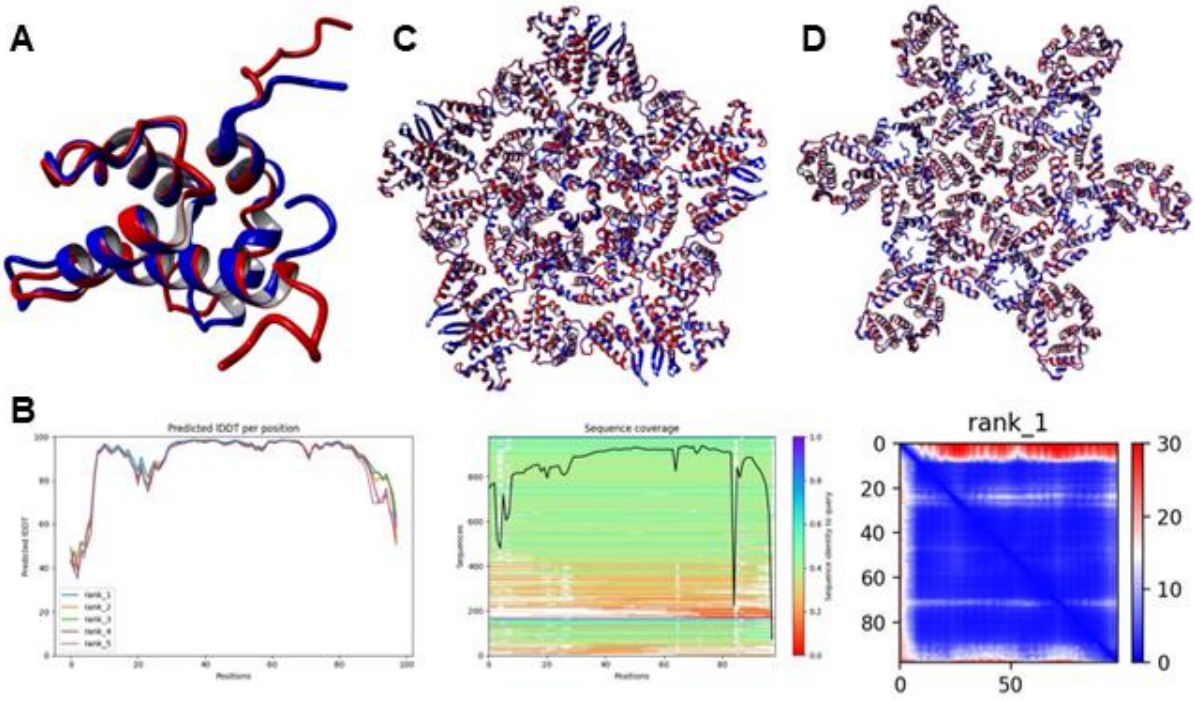


Figure 4: A, the highest-quality predicted structure of the reconstructed Gag matrix protein out of five predictions (red) overlaid with MoMLV matrix protein (RCSB: 1MN8, blue). **B**, quality statistics of the structural prediction in **A**. From left to right, the panels show the predicted local distance difference test (lDDT) statistics of all five models predicted (models ranked 2-5 are not shown in this figure), sequence coverage, and predicted alignment error of the best model (rank_1). A higher lDDT, a higher sequence coverage, and a lower predicted alignment error indicate better prediction quality. **C**, homology-modelled pentamer structure of the reconstructed Gag capsid protein (red) aligned with the pentamer structure of the mature MLV capsid protein (RCSB: 6HWY, blue). **D**, homology-modelled hexamer structure of Gag capsid protein (red) aligned with the hexamer structure of the mature MLV capsid protein (RCSB: 6HWX, blue).

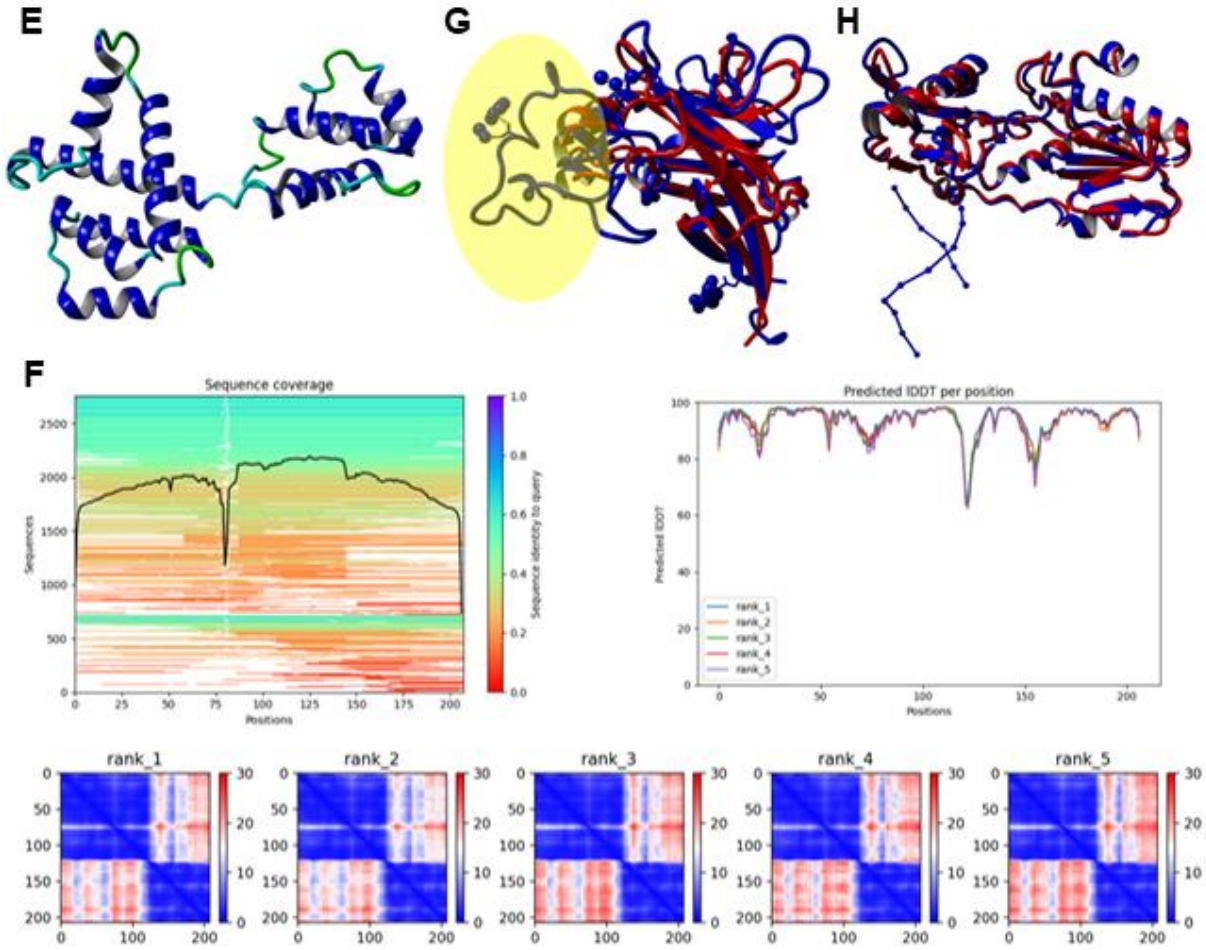


Figure 4 (continued): E, the highest-quality predicted monomeric structure of the reconstructed Gag capsid protein. **F**, quality statistics of the structural prediction in **E**. Panels from left to right show sequence coverage, predicted lDDT, and predicted alignment error for all five models (described in detail in **B**). Panel **E** shows the rank_1 model. **G**, the highest-quality predicted structure of the reconstructed Env RBD (red) aligned with the structure of FMLV Env RBD (RCSB: 1AOL, blue). The host-specific domain modelled in low quality is highlighted in yellow. **H**, the highest-quality predicted structure of Pol RT N-terminal domain (red) aligned with the structure of MoMLV Pol RT N-terminal domain (RCSB: 1I6J, blue). The balls and sticks on the left-hand corner are DNA, depicting protein-DNA interactions.

Selective pressure acts differently on different *GCE32* domains and orthologs

Since *GCE32* is conserved in all megabat species examined, I investigated the selective forces that have driven its evolution by elucidating different aspects of the *GCE32* coding sequence's dN/dS ratio (see Methods). While *GCE32* undergoes slightly purifying selection as a whole (dN/dS ≈ 0.57), its different domains are under disparate selective pressures. Among them, domains 1 and 3 (*gag*-like and *CRKL*-like) are conserved, domain 2 (*RBM33*-like) is positively selected, and domain 4 (*PEG3*-like) domain undergoes near-neutral evolution (**Figure 5A**; **Table 4A**). I also identified 10 positively selected sites in domain-by-domain analysis (**Table 4B**).

I also quantified the selective pressure acting on the entire *GCE32* coding sequence in different branches of the megabat phylogenetic tree. While most branches undergo purifying selection, the *P. pselaphon* terminal branch undergoes highly positive selection (dN/dS = 8.45408) (**Figure 5B**; **Table 4C**). Compared to the *GCE32* coding sequence at the branch's internal node, the *P. pselaphon* sequence has 4 non-synonymous substitutions and 0 synonymous substitutions, resulting in a high dN/dS ratio.

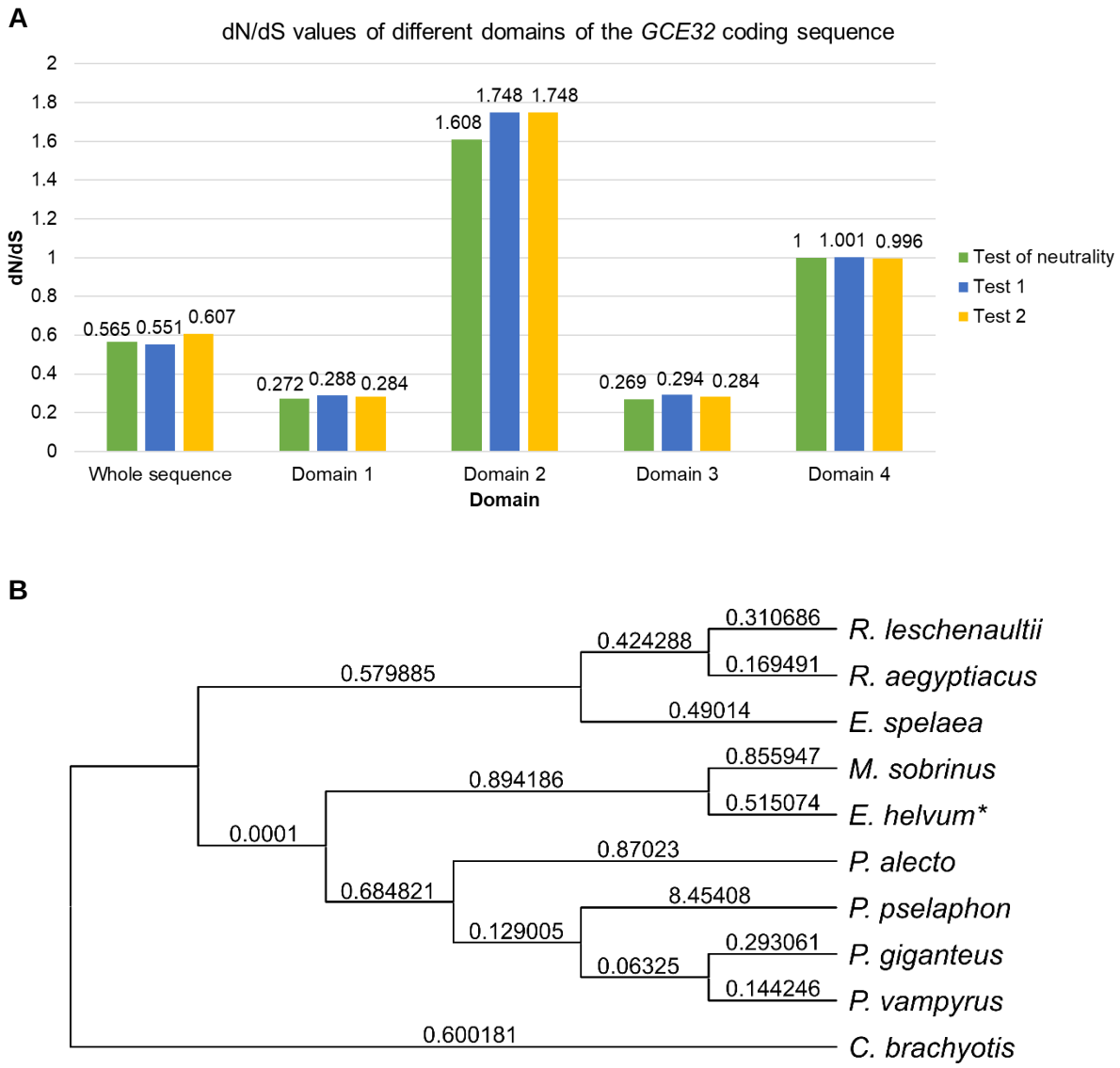


Figure 5: A, dN/dS values of different domains of the *GCE32* coding sequence calculated with three different site models (see **Table 4A** for details on model selection guided by likelihood ratio tests). **B**, dN/dS values of *GCE32* in different branches of the megabat phylogeny based on the free-ratio branch model (see **Table 4B**), rooted with *C. brachyotis* as the most basal megabat species. *The *E. helvum* branch is inconsistent with the phylogenetic tree from Almeida *et al* (2020); the correct position is basal to all *Pteropus* species.

Table 4A: different *GCE32* domains undergo radically different selective pressure. I first conducted a test of neutrality (a one-ratio model with fixed dN/dS = 1 versus a normal one-ratio model) to test whether the sequence is neutrally evolving. Then, I carried out two likelihood ratio tests (M1a vs M2a and M7 vs M8) to investigate the selective pressure on different parts of the *GCE32* coding sequence (Yang *et al*, 2000; Yang *et al*, 2005). While M1a and M2a are discrete models dividing the dN/dS of different sites into classes, M7 and M8 model the dN/dS of different sites with a continuous distribution governed by a parameter determining the shape of this distribution. Combining these tests ensure that the average dN/dS and any identified positively selected sites are robust. dN/dS > 1 is consistent with positive selection, dN/dS = 1 is consistent with neutral selection, and dN/dS < 1 is consistent with purifying selection. lnL, the natural logarithm of the likelihood ratio; df, degree(s) of freedom. M0, M1a, M2a, M7 and M8 are different site models that allow dN/dS to vary among different sites, as described in the PAML user guide (version 4.9j).

		Test of neutrality			**Test 1 of positive selection**			**Test 2 of positive selection**		
		M0 ω = 1 fixed	**M0**	**result (df = 1)**	**M1a**	**M2a**	**result (df = 2)**	**M7**	**M8**	**result (df = 2)**
Whole sequence	**dN/dS**	1	**0.5652**	reject fixed M0 (p < 0.0001)	**0.5513**	0.6068	do not reject M1a (p = 0.05004)	0.5594	**0.6068**	reject M7 (p = 0.0274)
	lnL	-6967.49	-6943.38		-6919.56	-6916.56		-6920.17	-6916.57	
Domain 1	**dN/dS**	1	**0.2715**	reject fixed M0 (p < 0.0001)	**0.2878**	0.2878	do not reject M1a (p > 0.999)	**0.2844**	0.2877	do not reject M7 (p = 0.9770)
	lnL	-859.309	-844.408		-841.035	-841.035		-841.059	-841.036	
Domain 2	**dN/dS**	1	**1.608**	reject fixed M0 (p = 0.0097)	0.9935	**1.7481**	reject M1a (p = 0.0017)	1	**1.7482**	reject M7 (p = 0.0017)
	lnL	-1169.94	-1166.59		-1169.93	-1163.56		-1169.94	-1163.56	
Domain 3	**dN/dS**	1	**0.2686**	reject fixed M0 (p < 0.0001)	**0.2942**	0.2942	do not reject M1a (p > 0.999)	**0.2836**	0.2836	do not reject M7 (p > 0.999)
	lnL	-1505.83	-1477.39		-1473.44	-1473.44		-1473.3	-1473.3	
Domain 4	**dN/dS**	**1**	0.835	do not reject fixed M0 (p = 0.2371)	0.6388	**1.001**	reject M1a (p = 0.0004)	0.6046	**0.9956**	reject M7 (p = 0.0003)
	lnL	-2511.77	-2511.07		-2501.14	-2493.3		-2501.33	-2493.3	

Table 4B: positively selected sites and their dN/dS values within the *GCE32* coding sequence. These positively selected sites are based on Bayes Empirical Bayes analysis by M8 (beta distribution allowing positively selected sites) (Yang *et al*, 2005). Pr (ω>1), the probability that dN/dS is greater than 1; ω refers to the dN/dS ratio; SE, standard error.

Domain 2 (*RBM33*-like)					**Domain 4 (*PEG3*-like)**				
Position in domain	Position in the whole sequence	Amino acid	Pr (ω>1)	Posterior mean ± SE for ω	Position in domain	Position in the whole sequence	Amino acid	Pr (ω>1)	Posterior mean ± SE for ω
15	145	Serine	0.987	2.882 ± 0.752	33	537	Arginine	0.978	5.417 ± 2.228
47	177	Proline	0.969	2.840 ± 0.791	49	556	Leucine	0.954	5.302 ± 2.313
49	179	Alanine	0.961	2.821 ± 0.803	80	588	Arginine	0.984	5.446 ± 2.203
70	212	Cysteine	0.963	2.826 ± 0.800	82	590	Cysteine	0.958	5.306 ± 2.291
109	243	Glutamine	0.965	2.831 ± 0.798					
114	248	Glutamic acid	0.96	2.816 ± 0.807					

Table 4C: test statistics of the likelihood ratio test (LRT) between the one-ratio and the free-ratio branch models on *GCE32* selective pressure. This LRT tests whether the dN/dS ratios are different among lineages (Yang, 1998). np, number of parameters; df, degrees of freedom; ΔlnL, the difference of natural logarithm likelihoods between the two models. The significance level is p = 0.05.

Model	lnL	np	df	2ΔlnL	p-value	result
one-ratio model	-6943.379646	19	16	29.65382	0.0199	reject one-ratio model
free-ratio model	-6928.552736	35				

GCE32 expression is tissue-specific

To identify and characterise *GCE32*'s expression, I searched for *GCE32*-like sequences in 263 publicly available transcriptomic datasets from 7 megabat species (**Table 5A**; **Appendix Table 4**). I identified high-coverage *GCE32* expression in one *P. alecto* spleen sample, four *P. alecto* liver samples, and one *R. aegyptiacus* testis sample. While over 10 tissue types were analysed, *GCE32* expression is uniquely pronounced in the *R. aegyptiacus* testis, with 25.4972 reads per 10 million (**Table 5B**). *De novo* assembly of the matching reads in the *R. aegyptiacus* testis dataset gives out five 224~449-nucleotide-long sequences, all identical to the *GCE32* coding sequence (**Figure 6**; **Appendix Table 5**). I also identified low-coverage matches to *GCE32* in 23 other samples (**Table 5C**), providing evidence that *GCE32* is expressed in different tissues and species.

Figure 6: the alignment of the *R. aegyptiacus GCE32* genomic sequence (bold) with the sequences reconstructed by *de novo* transcriptomic assembly of the matching reads in the *R. aegyptiacus* testis transcriptome (non-bold, SRR2914372). '|' signifies a match. The sequence is 2547 nucleotides in length.

Table 5A: summary of the megabat species examined for *GCE32* expression and the number of datasets examined for each species.

Species	**Number of samples**
Eidolon helvum	9
Eonycteris spelaea	6
Pteropus alecto	93
Pteropus giganteus	6
Pteropus vampyrus	15
Rousettus aegyptiacus	130
Rousettus leschenaultii	4
TOTAL	263

Table 5B: the relative expression level of *GCE32* in transcriptomic datasets where high-coverage expression is observed. Fold enrichment refers to the fold increase of expression level in the sample relative to SRR8858518 as the reference dataset.

Species	**Accession**	**Tissue**	**Treatment**	**Expression level (reads per 10 million)**	**Fold enrich-ment**
Pteropus alecto	SRR8858518	Spleen	Control	0.6958	1
Pteropus alecto	SRR8858519	Liver	Control	1.693	2.433
Pteropus alecto	SRR8858520	Liver	Control	1.755	2.522
Pteropus alecto	SRR8858521	Liver	Control	3.926	5.642
Pteropus alecto	SRR8858522	Liver	Control	7.178	10.316
Rousettus aegyptiacus	SRR2914372	Testis	Control	25.497	36.644

Table 5C: the species and treatment condition of the 23 transcriptomic datasets in which low-coverage *GCE32* expression is observed.

Species	**Accession**	**Tissue type**	**Treatment**
Eonycteris spelaea	SRR8836180	Liver	Control
Eonycteris spelaea	SRR8836183	Liver	Control
Eidolon helvum	SRR6134352	Cell line	Ebolavirus infection 24hrs
Eidolon helvum	SRR6134357	Cell line	Ebolavirus infection 0hrs
Pteropus alecto	SRR1531161	Cell line	Hendra virus infection 24hrs
Pteropus alecto	ERR3619156	Cell line	Control
Pteropus alecto	SRR8858515	Spleen	Control
Pteropus alecto	SRR8858516	Spleen	Control
Pteropus alecto	SRR8858517	Spleen	Control
Pteropus alecto	SRR8859514	Cell line	Interferon regulatory factor 1 (*IRF1*) knockout + Melaka virus infection 24hrs
Rousettus aegyptiacus	SRR2914059	Liver	Control
Rousettus aegyptiacus	SRR2914063	Ovary	Control
Rousettus aegyptiacus	SRR2914369	Liver	Control
Rousettus aegyptiacus	SRR6453213	Cell line	Marburg Virus infection 7hrs
Rousettus aegyptiacus	SRR7548030	Cell line	Sendai Cantell virus infection 3hrs
Rousettus aegyptiacus	SRR7609220	Cell line	Sendai virus infection 3hrs
Rousettus aegyptiacus	SRR7609222	Cell line	Sendai virus infection 8hrs
Rousettus aegyptiacus	SRR7609227	Cell line	Control 3hrs
Rousettus aegyptiacus	SRR7609231	Cell line	Control 8hrs
Rousettus aegyptiacus	SRR7609233	Cell line	Control 24hrs
Rousettus aegyptiacus	SRR7609234	Cell line	Control 24hrs
Rousettus aegyptiacus	SRR11148689	Cell line	Universal interferon 4hrs
Rousettus leschenaultii	SRR6796668	Cochlea	Control

DISCUSSION

The investigation of *GCE32*, an active co-opted retroviral *gag* gene in megabats, provides multiple insights into its evolutionary history and function. *GCE32* encodes virus- and host-derived domains and is uniquely conserved in megabats (**Figure 2**). The recovery of *GCE32*-related EVEs sheds light on *GCE32*'s evolutionary history (**Figure 3**) and allows us to predict ancient viral protein structures similar to present exogenous viral proteins (**Figure 4**). While *GCE32* is generally conserved in evolution, it is positively selected in some species and domains (**Figure 5**). *GCE32* is also significantly expressed in the spleen, liver, and the testis, suggesting it still plays a functional role (**Table 5**). These findings provide insights into the timescale of *GCE32* evolution, host-virus interactions in *GCE32* evolution, and the possible functions of *GCE32*.

GCE32 evolved from host-virus interactions over 58.5 million years ago

Host-virus gene exchange and virus-derived genes are significant drivers of evolutionary change, taking part in critical biological processes like placenta evolution and mammalian antiviral immunity (Irwin *et al*, 2022; Katzourakis & Aswad, 2017). Moreover, endogenous gammaretroviruses are highly diverse in bats, providing ample materials for gene co-option by the host (Cui *et al*, 2012). Considering this long-lasting evolutionary relationship, it is not surprising that *GCE32* – a co-opted virus-host fusion gene – is still active and has possibly shaped megabats' antiviral immunity in their evolution. Since *GCE32* is conserved in all megabat species examined, we can trace its origin to before the last common ancestor of megabats but after megabats split from other bats – 58.5~62.6 million years ago in bats' early evolutionary history (Lei & Dong, 2016).

GCE32 evolution provides insights into ancient host-virus interactions

GCE32's evolutionary history provides a glimpse into ancient host-virus interactions similar to present ones. The reconstructed viral proteins provide evidence that some structural features of 60-million-year-old gammaretroviruses are highly conserved despite sequence dissimilarity, indicating that the virus's life cycle was possibly similar to that of present exogenous gammaretroviruses. However, it remains hard to reconstruct the details of some virus life cycle processes, such as binding with host cell surface receptor, polymerisation, assembly, and budding, since they involve complicated interactions with host

cell components. These questions are essential for understanding how viruses infect host cells; they could be addressed by *in vivo* experiments that reconstruct viruses and use them to infect cells.

Apart from the typical gammaretrovirus *gag*, *pol* and *env* genes, the ancient retrovirus ancestral to *GCE32* possibly contained extra components. Since the *CRKL*-characteristic SH2 and SH3 domains are widely identified in these EVEs, it is likely that this ancestral retrovirus already possessed these domains and integrated into the host genome multiple times. These domains participate in protein-protein interactions: SH2 allows proteins to dock to phosphorylated tyrosine residues, while SH3 mediates peptide binding by recognising a consensus peptide sequence (Birge *et al*, 2009). At least 4 *GCE32*-related EVEs, including *GCE32a*, also possess the CC2-LZ domain typically found in the NF-κB essential modulator gene (NEMO) (Grubisha *et al*, 2010). In contrast, no *RBM33*-like and *PEG3*-like sequences in *GCE32*-related EVEs were identified apart from *GCE32*; this could be due to *GCE32* acquiring them during its own co-option process.

The presence and arrangement of these domains provide insights into the virus ancestor's mode of action. The SH2 and SH3 domains are present in transformative avian sarcoma viruses. For example, the v-Crk protein first identified in avian sarcoma virus CT10 contains a *gag* domain fused to SH2 and SH3 domains captured from its chicken host (Birge *et al*, 2009). The virus transforms chicken embryo fibroblasts by inducing excessive tyrosine phosphorylation (Mayer *et al*, 1988). The gammaretrovirus genus also contains multiple transformative viruses, such as MLV, feline leukemia virus, and simian sarcoma virus. In contrast, the CC2-LZ domain plays a different function. It is essential for NEMO function; disrupting it would inhibit NF-κB activation and downregulate the expression of various pro-inflammatory cytokines and chemokines (Grubisha *et al*, 2010; Liu *et al*, 2017). These domains are widely identified in *GCE32*-related EVEs and are arranged closely between the Gag matrix and capsid proteins (see **Figure 3D**), suggesting that the original retrovirus possibly contained these domains. It probably had transformative potential with the SH2 and SH3 domains, with the CC2-LZ domain adding to its ability to induce inflammatory responses. Thus, given its mode of action, this virus might have elicited intense host immune responses. It would be advantageous for the host to capture this virus's sequence and co-opt it as a viral restriction factor; *GCE32* might be an example. The co-opted gene might also be involved in cancer pathways, potentially suppressing their incidence and contributing to bats' extremely low cancer incidence (Seluanov *et al*, 2018).

Apart from the virus's mode of action, these domains also inform a possible evolutionary trajectory of *GCE32* evolution (**Figure 7**). Firstly, a generic gammaretrovirus acquired SH2 and SH3 domains from the host *CRKL*. This acquisition can be achieved through retrogene formation or recombination, both of which involve reverse transcriptase activity (Pan & Zhang, 2009; Geuking *et al*, 2009). This potentially transformative retrovirus integrated with the host genome multiple times, resulting in numerous *GCE32*-related EVEs in the megabat genome. Then, in one scenario, this retrovirus acquired the CC2-LZ domain from the host and subsequently integrated into the host genome. The EVEs then became inactive, as seen in some *GCE32*-related EVEs such as *GCE32a*. In another scenario, the virus integrated into the host genome first, and its *gag* gene was co-opted by the host while its *pol* and *env* domains lost function and slowly decayed. Then, it acquired *RBM33*-like and *PEG3*-like domains through processes such as recombination, completing the co-option process for a novel function.

Potential functions of *GCE32*

Several pieces of evidence support the co-option of *GCE32* as a virus-derived gene. It consists of a virus-derived *gag* matrix protein domain and other host domains in a single open reading frame. Moreover, the *env* and *pol* domains of the EVE on the *GCE32* locus are functionless relics, implying that only the *gag*-derived *GCE32* gene is currently active. It has faced disparate selective pressures and is expressed in multiple tissues. These observations indicate that *GCE32* might have played different functions throughout its evolutionary history.

We can speculate the change in *GCE32*'s function in its evolutionary history by analysing the selective pressure acting on it. In the branch model (see **Figure 5B**), an ancestral branch uniting six species and connecting to the most ancestral *C. brachyotis* has a dN/dS ratio of 0.0001, implying strong purifying selection in the context of intense host-virus interactions (although a small sample size requires cautious treatment). Thus, *GCE32* might have actively participated in EDI when the original virus was active, acting as a virus-specific restriction factor like Fv1 that targets essential viral life cycle processes (Yap *et al*, 2014). It was thus strongly conserved. Experimental studies investigating *GCE32*'s effect on viral challenges from reconstructed and exogenous bat retroviruses could help test this hypothesis. In contrast, while still undergoing purifying selection, *GCE32* is less conserved in terminal tree branches (average dN/dS ≈ 0.47 in branches except for *P. pselaphon*). This observation might have resulted from the loss of virus-specific EDI – a common evolutionary outcome for EDI genes resulting from evolutionary arms races – and secondary co-option for a new function (Aswad & Katzourakis, 2012).

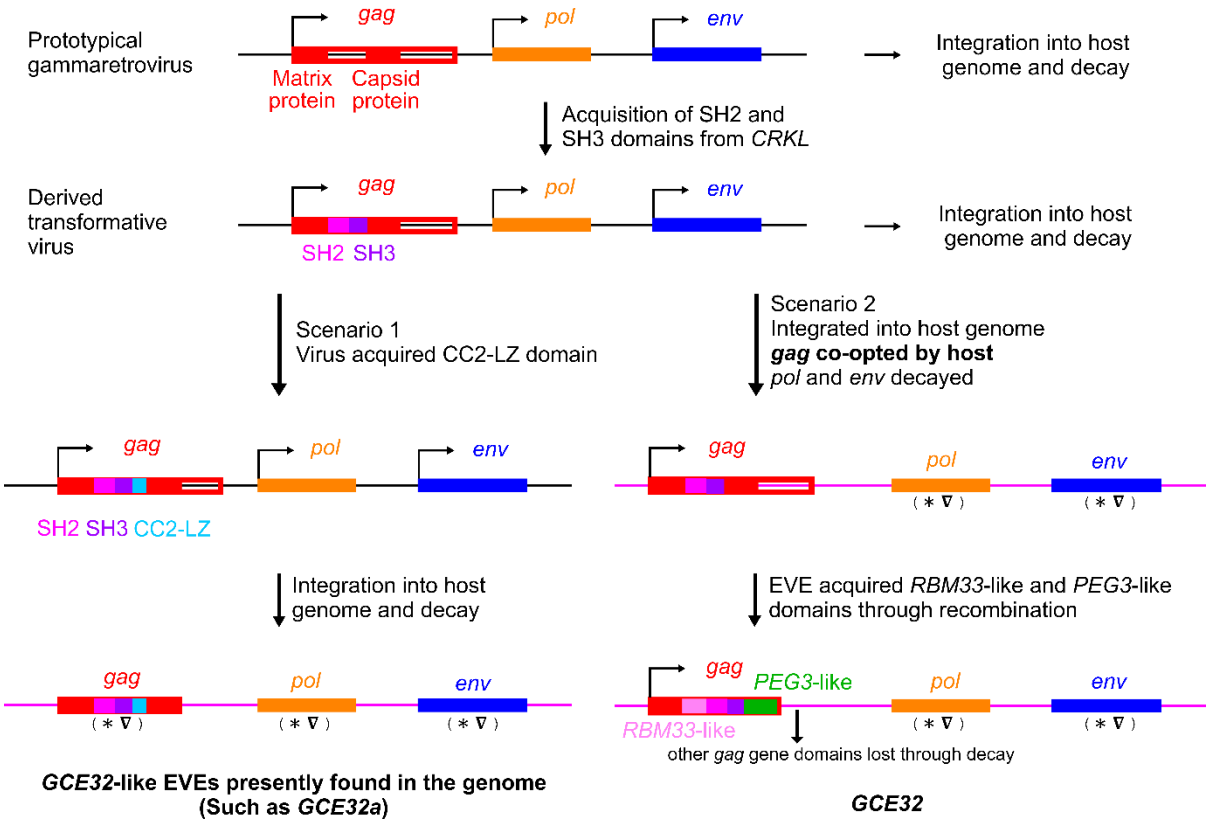


Figure 7: a postulated evolutionary model in the megabat common ancestor that led to the current presence of *GCE32*-related EVEs in megabats. The prototypical gammaretrovirus first gained the SH2 and SH3 domains from the host *CRKL* gene and then could have undergone two evolutionary trajectories. In the first scenario, the virus further acquired the CC2-LZ domain potentially from *IKBKG*, the host gene that codes for NEMO. It then integrated into the host, eventually becoming an inactive virus. This scenario is seen in several *GCE32*-related EVEs, including *GCE32a*. In the second scenario, which gave rise to *GCE32*, the virus first integrated into the host genome. Then, its *gag* gene was co-opted by the host, while its *pol* and *env* genes decayed and lost function. After that, it acquired two more domains from two host genes (*RBM33* and *PEG3*) separately through processes such as retroposition or recombination, resulting in the *GCE32* coding sequence as seen today. ORFs are marked by arrows extending above the genes. Inactive gene remnants are marked by a lack of ORF and symbols (* ∇); the black genomic backbone indicates the genes are in the viral genome, while the magenta genomic backbone indicates the genes are in the host genome.

Another intriguing observation is that positive selection acts on *GCE32*'s multiple sites, its *RBM33*-like domain, and its ortholog in *P. pselaphon*. Positive selection is a hallmark of host-pathogen evolutionary arms races and is not normally observed in genes (Daugherty & Malik, 2012). Thus, it is very likely that part of *GCE32* functions in EDI. It is also likely to be secondarily co-opted for a new function in *P. pselaphon*, an endangered species of about 150 individuals (Tani *et al*, 2019). While bottleneck effects and genetic drift might underlie this observation, there is also possibly an ongoing evolutionary arms race between *P. pselaphon* and a pathogen circulating in its population. It is thus worth investigating the *GCE32* sequence in related species and at the *P. pselaphon* population level from a wildlife conservation perspective.

GCE32's conserved domains allow us to speculate on its present mechanisms of function based on sequence homology. Apart from the *gag*-derived domain, *GCE32* has domains similar to three host genes – *RBM33*, *CRKL*, and *PEG3*. While *GCE32*'s nucleotide sequence matches that of *RBM33*, there is a frameshift resulting in different amino acid sequences, making us unable to predict the function of the *RBM33*-like domain (**Figure 8**). The *CRKL*-like domain contains the SH2 and SH3 conserved domains that typically transmit intracellular signals by binding to tyrosine-phosphorylated proteins (Birge *et al*, 2009). *PEG3* is a DNA-binding transcription factor with 12 zinc finger domains; it acts upstream of NF-κB activation and mediates the apoptosis pathway activated by DNA damage, which can also function as an antiviral strategy against pathogens (Thiaville *et al*, 2013).

The presence of these domains indicates several possible scenarios for *GCE32*'s function. Firstly, *GCE32*'s expression and purifying selection suggest that it might function at the RNA level – the mode of action for some other restriction factors such as small interfering RNAs. Since *GCE32* has a conserved ORF, it is likely translated and functions as a protein, either by interacting with other proteins or mediating other proteins' expression as a transcription factor. Since different *GCE32* domains undergo different selective pressure, they might be proteolytically cleaved after translation and function differently – as a signal transducer by the *CRKL*-like domain and a transcription factor by the *PEG3*-like domain. These functions might underlie a coordinated EDI response that recognises viral infection by the *gag*-like domain, transduces this signal by the *CRKL*-like domain, and regulates downstream gene expression by the *PEG3*-like domain. Further experimental studies, such as two-hybrid screening and chromatin immunoprecipitation sequencing, would help elucidate other proteins it interacts with and its involvement in protein-DNA interactions.

Figure 8: a frameshift between the original *RBM33* sequence and the *RBM33*-derived sequence in the *P. vampyrus GCE32* sequence leads to two completely different sets of amino acid sequences. Codons are marked alternatively in bold and normal font, and amino acids are marked in red (*RBM33*) or blue (*GCE32*). There is a high similarity between the two sequences on the codon level but no significant similarities on the amino acid level.

Apart from *GCE32*'s functioning mechanisms, its expression profile also provides insights into its potential involvement in biological processes. Its expression in the testis might be a by-product of epigenomic reprogramming, a reset of the host genome's epigenomic status in germ cells that allows the propagation of virus-derived elements and transposons (Schumann *et al*, 2019). However, it is also expressed under other circumstances, such as the liver. We can thus also postulate that *GCE32* is a marker of immunologically 'privileged' organs such as the liver and the testis, in which materials are protected from autoimmune attack (Forrester *et al*, 2008). These organs' immune privilege is based on blood-tissue barriers and tightly mediated regulatory immune cells (Fijak & Meinhardt, 2006; Crispe *et al*, 2006). It provides a mechanism for tissue-specific long-term persistence of viruses – for example, Ebolavirus can persist in human testes, leading to persistence and new outbreaks (Schindell *et al*, 2018). Thus, *GCE32* might also act as a counter-retrovirus EDI that persist in immunologically privileged organs. A more detailed single-cell transcriptomic study of *GCE32* expression would help elucidate the cell types in which *GCE32* is expressed and identify any viruses persisting in these organs.

Addressing this study's caveats

This study has several caveats. Firstly, I regarded the 2nt-long insertions in the *GCE32* coding sequence as sequencing errors, while they might be genuine insertions. This could be tested through a PCR experiment provided adequate samples and laboratory resources. Secondly, I constructed the phylogenies of *GCE32*-related EVEs with *env* gene sequences (**Figures 2C, 3C**). While *env* is the best conserved of the three genes in *GCE32*, it is common for viruses to recombine, a genetic process that disrupts tree-like phylogeny. A co-phylogenetic analysis to determine whether these genes' phylogenies represent mirror images would address this caveat, but it was not performed due to time limits.

CONCLUSION

GCE32 is a potentially co-opted, virus-derived fusion gene in megabats. Multifaceted computational analyses of its evolutionary history, selective pressure, and expression allow us to probe into an ancient host-virus interface and speculate on *GCE32*'s potential functions. Future studies could focus on two aspects. Firstly, by investigating *GCE32*'s restriction activity to ancient and present viruses in bats, we can elucidate its possible interactions with viruses. Secondly, to probe further into its current function, we can investigate its involvement in immune privilege by integrating the results from protein interaction experiments and single-cell transcriptomic studies in different bat cell types.

This study shows the two-fold benefit of investigating a co-opted EVE-derived gene in megabat genomes. Firstly, I discovered that *GCE32* is a virus-derived genetic innovation that might underlie bats' unique antiviral immunity. Moreover, as a by-product of intensive host-virus interactions, studying *GCE32* allowed us to probe into these interactions, which can provide insights into how they have taken part in the evolution of this immunity. Furthermore, it is also attractive to investigate the potentially different functions it might have had in its evolutionary history in response to host-virus interactions. All these insights can shed light on the greater significance of bat immunology – the investigation of their exceptional ability to harbour viral infections, the potential to adapt bats' immune strategies to control virus-induced inflammation in humans, and the prevention of zoonotic diseases they might cause.

ACKNOWLEDGEMENTS

I am extremely grateful to my supervisor Professor Aris Katzourakis, who has given me continued, detailed guidance to steer my way through the project. I would also like to extend my gratitude to José Gabriel Niño Barreat, who often listened to my progress, helped address the technical difficulties and inspired me with new thoughts. I must also thank Charles Reuben de Souza for the daily discussions and informal conversations. I am also deeply indebted to my family and my partner, Wang Chuying, for their love and support throughout the project, without whom this project would have been impossible. I would also like to thank many of my friends for my conversations with them.

BIBLIOGRAPHY

Ahn M, Anderson DE, Zhang Q, *et al*. Dampened NLRP3-mediated inflammation in bats and implications for a special viral reservoir host. *Nat Microbiol*. 2019;4(5):789-799.

Almeida FC, Simmons NB, Giannini NP. A Species-Level Phylogeny of Old World Fruit Bats with a New Higher-Level Classification of the Family Pteropodidae. *American Museum Novitates,* 2020(3950):1-24.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990;215(3):403-410.

Aswad A, Katzourakis A. Paleovirology and virally derived immunity. *Trends Ecol Evol*. 2012;27(11):627-636.

Banerjee A, Baker ML, Kulcsar K, Misra V, Plowright R, Mossman K. Novel Insights Into Immune Systems of Bats. *Front Immunol*. 2020;11:26.

Banerjee A, Rapin N, Bollinger T, Misra V. Lack of inflammatory gene expression in bats: a unique role for a transcription repressor. *Sci Rep*. 2017;7(1):2232.

Birge RB, Kalodimos C, Inagaki F, Tanaka S. Crk and CrkL adaptor proteins: networks for physiological and pathological signaling. *Cell Commun Signal*. 2009;7:13.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*. 2009;25(15):1972-1973.

Chandrasekaran C, Betrán E. Origins of new genes and pseudogenes. *Nature Education*. 2008;1(1):181.

Chionh YT, Cui J, Koh J, *et al*. High basal heat-shock protein expression in bats confers resistance to cellular heat/oxidative stress. *Cell Stress Chaperones*. 2019;24(4):835-849.

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

Crispe IN, Giannandrea M, Klein I, John B, Sampson B, Wuensch S. Cellular and molecular mechanisms of liver tolerance. *Immunol Rev*. 2006;213:101-118.

Cui J, Tachedjian M, Wang L, Tachedjian G, Wang LF, Zhang S. Discovery of retroviral homologs in bats: implications for the origin of mammalian gammaretroviruses. *J Virol*. 2012;86(8):4288-4293.

Darriba D, Posada D, Kozlov AM, Stamatakis A, Morel B, Flouri T. ModelTest-NG: A New and Scalable Tool for the Selection of DNA and Protein Evolutionary Models. *Mol Biol Evol*. 2020;37(1):291-294.

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

David Q, Schountz T, Schwemmle M, Ciminski K. Different but Not Unique: Deciphering the Immunity of the Jamaican Fruit Bat by Studying Its Viriome. *Viruses*. 2022;14(2):238.

Eddy SR. Accelerated Profile HMM Searches. *PLoS Comput Biol*. 2011;7(10):e1002195.

Felsenstein J. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*. 1985;39(4):783-791.

Fernandes AP, Águeda-Pinto A, Pinheiro A, Rebelo H, Esteves PJ. Evolution of TRIM5 and TRIM22 in Bats Reveals a Complex Duplication Process. *Viruses*. 2022;14(2):345.

Fijak M, Meinhardt A. The testis in immune privilege. *Immunol Rev*. 2006;213:66-81.

Forrester JV, Xu H, Lambe T, Cornall R. Immune privilege or privileged immunity? *Mucosal Immunol*. 2008;1(5):372-381.

Geuking MB, Weber J, Dewannieux M, *et al*. Recombination of retrotransposon and exogenous RNA virus results in nonretroviral cDNA integration. *Science*. 2009;323(5912):393-396.

Goh G, Ahn M, Zhu F, *et al*. Complementary regulation of caspase-1 and IL-1β reveals additional mechanisms of dampened inflammation in bats. *Proc Natl Acad Sci U S A*. 2020;117(46):28939-28949.

Gould SJ, Vrba ES. Exaptation—a Missing Term in the Science of Form. *Paleobiology*. 1982;8(1):4-15.

Grubisha O, Kaminska M, Duquerroy S, *et al*. DARPin-assisted crystallography of the CC2-LZ domain of NEMO reveals a coupling between dimerisation and ubiquitin binding. *J Mol Biol*. 2010;395(1):89-104.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol*. 2010;59(3):307-321.

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol Biol Evol*. 2018;35(2):518-522.

Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, Simmerling C. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins*. 2006;65(3):712-725.

Irving AT, Ahn M, Goh G, Anderson DE, Wang LF. Lessons from the host defences of bats, a unique viral reservoir. *Nature*. 2021;589(7842):363-370.

Irwin NAT, Pittis AA, Richards TA, Keeling PJ. Systematic evaluation of horizontal gene transfer between eukaryotes and viruses. *Nat Microbiol*. 2022;7(2):327-336.

Jumper J, Evans R, Pritzel A, *et al*. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596(7873):583-589.

Katzourakis A, Aswad A. Evolution: Endogenous Viruses Provide Shortcuts in Antiviral Immunity. *Curr Biol*. 2016;26(10):R427-R429.

Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 2013;30(4):772-780.

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

Koh J, Itahana Y, Mendenhall IH, *et al*. ABCB1 protects bat cells from DNA damage induced by genotoxic compounds. *Nat Commun*. 2019;10(1):2820.

Konagurthu AS, Whisstock JC, Stuckey PJ, Lesk AM. MUSTANG: a multiple structural alignment algorithm. *Proteins*. 2006;64(3):559-574.

Krieger E, Vriend G. YASARA View – molecular graphics for all devices – from smartphones to workstations. *Bioinformatics*. 2014;30(20):2981-2982.

Kryazhimskiy S, Plotkin JB. The population genetics of dN/dS. *PLoS Genet*. 2008;4(12):e1000304.

Lee AK, Kulcsar KA, Elliott O, *et al*. De novo transcriptome reconstruction and annotation of the Egyptian rousette bat. *BMC Genomics*. 2015;16:1033.

Lei M, Dong D. Phylogenomic analyses of bat subordinal relationships based on transcriptome data. *Sci Rep*. 2016;6:27726.

Letko M, Seifert SN, Olival KJ, Plowright RK, Munster VJ. Bat-borne virus diversity, spillover and emergence. *Nat Rev Microbiol*. 2020 Aug;18(8):461-471.

Liu T, Zhang L, Joo D, Sun SC. NF-κB signaling in inflammation. *Sig Transduct Target Ther*. 2017;2:17023.

Lu S, Wang J, Chitsaz F, *et al*. CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res*. 2020;48(D1):D265-D268.

Mayer BJ, Hamaguchi M, Hanafusa H. A novel viral oncogene with structural similarity to phospholipase C. *Nature*. 1988;332(6161):272-275.

Mirdita M, Schütze K, Moriwaki Y, *et al*. ColabFold – Making protein folding accessible to all. *bioRxiv*. 2021. doi.org/10.1101/2021.08.15.456425

Mistry J, Chuguransky S, Williams L, *et al*. Pfam: The protein families database in 2021. *Nucleic Acids Res*. 2021;49(D1):D412-D419.

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

Olival KJ, Hosseini PR, Zambrana-Torrelio C, Ross N, Bogich TL, Daszak P. Host and viral traits predict zoonotic spillover from mammals. *Nature*. 2017;546(7660):646-650.

Pan D, Zhang L. Burst of young retrogenes and independent retrogene formation in mammals. *PLoS One*. 2009;4(3):e5040.

Qu K, Glass B, Doležal M, *et al*. Structure and architecture of immature and mature murine leukemia virus capsids. *Proc Natl Acad Sci U S A*. 2018;115(50):E11751-E11760.

Robertson G, Schein J, Chiu R, *et al*. *De novo* assembly and analysis of RNA-seq data. *Nat Methods*. 2010;7(11):909-912.

Schindell BG, Webb AL, Kindrachuk J. Persistence and Sexual Transmission of Filoviruses. *Viruses*. 2018 Dec 2;10(12):683.

Schumann GG, Fuchs NV, Tristán-Ramos P, Sebe A, Ivics Z, Heras SR. The impact of transposable element activity on therapeutically relevant human stem cells. *Mob DNA*. 2019;10:9.

Seluanov A, Gladyshev VN, Vijg J, Gorbunova V. Mechanisms of cancer resistance in long-lived mammals. *Nat Rev Cancer*. 2018;18(7):433-441.

Shaw AE, Hughes J, Gu Q, *et al*. Fundamental properties of the mammalian innate immune system revealed by multispecies comparison of type I interferon responses. *PLoS Biol*. 2017;15(12):e2004086.

Skirmuntt EC, Escalera-Zamudio M, Teeling EC, Smith A, Katzourakis A. The Potential Role of Endogenous Viral Elements in the Evolution of Bats as Reservoirs for Zoonotic Viruses. *Annu Rev Virol*. 2020;7(1):103-119.

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

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

Smit AFA, Hubley R, Green P. *RepeatMasker Open-4.0*. 2013-2015. http://www.repeatmasker.org

Tani T, Eitsuka T, Katayama M, *et al*. Establishment of immortalized primary cell from the critically endangered Bonin flying fox (*Pteropus pselaphon*). *PLoS One*. 2019;14(8):e0221364.

Taylor DJ, Leach RW, Bruenn J. Filoviruses are ancient and integrated into mammalian genomes. *BMC Evol Biol*. 2010;10:193.

Thiaville MM, Huang JM, Kim H, Ekram MB, Roh TY, Kim J. DNA-binding motif and target genes of the imprinted transcription factor PEG3. *Gene*. 2013;512(2):314-320.

Tikhonenko AT, Lomovskaya OL. Avian endogenous provirus (ev-3) *env* gene sequencing: implication for pathogenic retrovirus origination. *Virus Genes*. 1990;3(3):251-258.

Tsai HH, Tsai CJ, Ma B, Nussinov R. *In silico* protein design by combinatorial assembly of protein building blocks. *Protein Sci*. 2004;13(10):2753-2765.

True JR, Carroll SB. Gene co-option in physiological and morphological evolution. *Annu Rev Cell Dev Biol*. 2002;18:53-80.

Vargiu L, Rodriguez-Tomé P, Sperber GO, *et al*. Classification and characterisation of human endogenous retroviruses; mosaic forms are common. *Retrovirology*. 2016;13:7.

Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, Heer FT, de Beer TAP, Rempfer C, Bordoli L, Lepore R, Schwede T. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* 2018;46(W1):W296-W303.

Wang J, Han GZ. Frequent Retroviral Gene Co-option during the Evolution of Vertebrates. *Mol Biol Evol*. 2020;37(11):3232-3242.

Wheeler DL, Church DM, Federhen S, Lash AE, Madden TL, Pontius JU, Schuler GD, Schriml LM, Sequeira E, Tatusova TA, Wagner L. Database resources of the National Center for Biotechnology. *Nucleic Acids Res.* 2003;31(1):28-33.

Yang Z. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol*. 1998;15(5):568-573.

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

Yang Z, Nielsen R. Synonymous and non-synonymous rate variation in nuclear genes of mammals. *J Mol Evol*. 1998;46(4):409-418.

Yang Z, Nielsen R, Goldman N, Pedersen AM. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*. 2000;155(1):431-449.

Yang Z, Wong WS, Nielsen R. Bayes empirical bayes inference of amino acid sites under positive selection. *Mol Biol Evol*. 2005;22(4):1107-1118.

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

MANAGEMENT REPORT

The project has mostly been carried out according to plan, which was to single out bat-specific EDI candidates from a list of putative co-opted genes derived from retroviral *gag* genes. I successfully identified a gene candidate, conducting phylogenetic and selection analyses as planned. However, while I originally planned to perform experimental work to establish the expression pattern and restriction activity of the candidate gene, the lab space has not been made available to us in time as it is new to the research group. To compensate for this, I conducted further computational analysis on the candidate gene – namely viral protein structure reconstruction, conserved domain mapping, and expression analysis. These analyses should be treated with merit, as they provided insights into ancient host-virus interactions. Additionally, the computational gene expression analysis provided much more positive results than what would have been achieved by the originally proposed experiments for the time allowed, since it enabled us to examine *GCE32*'s expression pattern in multiple species and tissue types. These results would better guide further studies on investigating the function of *GCE32*.

In summary, despite unforeseeable changes in the infrastructure and an unavoidable change in the original plan as a result, I have made considerable progress on the project in the limited time provided. I expect that experimental studies will proceed just beyond the timescale of this project.

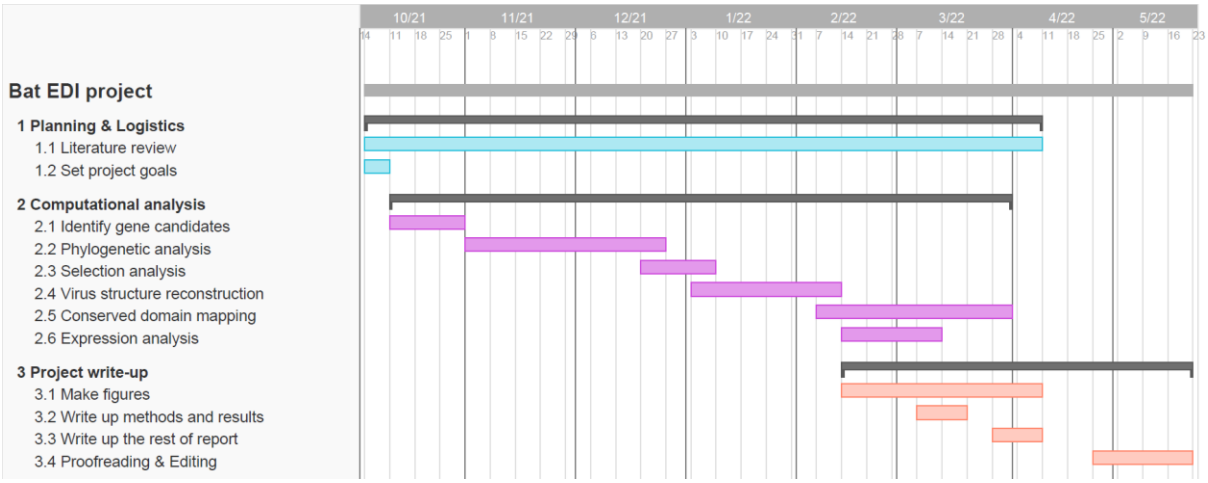


Figure 9: a Gantt chart summarising the work that has been done on this project.

APPENDICES

The related materials for this report, such as custom scripts, protein models, and search results, are available on an online GitHub repository at https://github.com/anonymous-ox/MBiol. Appendix Tables 1 to 5 as referred to in the main text are attached here, starting from the next page.

Appendix Table 1: the genomic location, synteny, nature of the sequence, and conserved domains of five putative *GCE* gene sequences (*GCE29, GCE30, GCE31, GCE32,* and *GCE33*) identified by Wang & Han (2020). Asterisks (*) in the 'synteny' column indicate that the genomic scaffold is too short to identify some or all neighbouring genes. 'Virus-derived' refers to the percentage of virus-derived sequence in the coding sequence of this gene, and the type of the most prevalent EVE in the sequence, as identified by RepeatMasker.

Gene	Species	Accession	Start	End	Synteny	Virus-derived	Conserved domains
GCE29	*Miniopterus natalensis*	NW_015504389	611358	615466	*ABHD4-gag**	ERVII-85.96%	Gag_p10/Gag_p24
GCE29	*Miniopterus schreibersii*	PVJG01001619	185103	189165	*ABHD4-gag**	ERVII-98.61%	Gag_p10/Gag_p24
GCE30	*Miniopterus natalensis*	NW_015504523	1291482	1297104	*LOC107536702-gag**	ERVI-90.77%	Gag_p30/RT_RnaseH/zf-H/rve/MLVIN_C
GCE30	*Miniopterus schreibersii*	PVJG01002394	1259	10063	*LOC107536702-gag**	ERVI-29.37%	Gag_p30/RT_RnaseH/zf-H/rve
GCE31	*Pteropus vampyrus*	NW_011888792	2560754	2579690	*LOC111732073-gag-ATG10*	ERVI-58.45%	Gag_p30/MLVIN-C/Gag_matrix
GCE31	*Pteropus alecto*	NW_006436298	14681986	14700991	*LOC111731452-gag-ATG10*	ERVI-35.80%	Gag_p30/MLVIN-C/Gag_matrix
GCE31	*Pteropus giganteus*	NW_024346062	3541128	3560037	*LOC111731452-gag-ATG10*	ERVI-41.09%	Gag_p30/MLVIN-C
GCE31	*Pteropus pselaphon*	BMBI01006582	195480	214366	*LOC111731452-gag-ATG10*	ERVI-54.69%	Gag_p30/MLVIN-C
GCE32	*Rousettus aegyptiacus*	NW_023416287	26269142	26289602	*BMP2-gag-HAO1*	ERVI-13.55%	Gag_MA/PRK07764/SH2 superfamily/SH3 superfamily/zf-C2H2/Ribosomal L40e
GCE32	*Pteropus vampyrus*	NW_011888788	4571821	4594507	*BMP2-gag-HAO1*	ERVI-14.35%	Gag_MA/SH2/SH3/Ribosomal L40e/zf
GCE32	*Pteropus giganteus*	NW_024342471	416962	436924	*BMP2-gag-HAO1*	ERVI-16.09%	Gag_MA/SH2/SH3/Ribosomal L40e/zf
GCE32	*Pteropus alecto*	NW_006442491	13126387	13145463	*BMP2-gag-HAO1*	ERVI-14.39%	Gag_MA/SH2/SH3/Ribosomal L40e
GCE32	*Pteropus pselaphon*	BMBI01006887	648836	668902	*LOC107510880-gag-HAO1**	ERVI-14.32%	Gag_MA/SH2/SH3/Ribosomal L40e/zf
GCE32	*Eidolon helvum*	AWHC01227563	3091	23760	*	ERVI-13.69%	Gag_MA/SH2/SH3/Ribosomal L40e/dnaA
GCE32	*Rousettus leschenaultii*	BNJM01000014	1703910	1724216	*BMP2-gag-HAO1*	ERVI-13.55%	Gag_MA/PRK07764/SH2 superfamily/SH3 superfamily/zf-C2H2/Ribosomal L40e
GCE32	*Eonycteris spelaea*	PUFA01000110	4234666	4255111	*BMP2-gag-HAO1*	ERVI-14.07%	Gag_MA/PRK07764/SH2 superfamily/SH3 superfamily/Ribosomal L40e
GCE32	*Macroglossus sobrinus*	PVKZ01000650	481912	502279	*LOC107510880-gag-HAO1**	ERVI-14.19%	Gag_MA/SH2/SH3/Ribosomal L40e
GCE32	*Cynopterus brachyotis*	SSHV01047299	3133	22991	*	ERVI-13.02%	Gag_MA/SH2/SH3/Ribosomal L40e/PRK07764
GCE33	*Pteropus alecto*	NW_006442103	18244907	18252356	*LOC102878255-gag-BOLA1*	ERVI-99.97%	RT_like/Gag_p30/retropepsin/RT_RNaseH
GCE33	*Rousettus aegyptiacus*	NW_023416292	27869958	27878180	*LOC102878255-gag-BOLA1*	ERVI-98.32%	RT_like/Gag_p30/retropepsin/RT_RNaseH/zf
GCE33	*Pteropus vampyrus*	NW_011889398	3669	10412	*LOC102878255-gag-BOLA1*	ERVI-99.94%	RT_like/Gag_p30/retropepsin/RT_RNaseH
GCE33	*Pteropus giganteus*	NW_024350924	10736022	10745181	*LOC102878255-gag-BOLA1*	ERVI-99.94%	RT_like/Gag_p30/retropepsin/RT_RNaseH
GCE33	*Rhinolophus ferrumequinum*	NC_046305	50351199	50360207	*LOC102878255-gag-BOLA1*	ERVI-80.25%	RT_like/Gag_p30/PHA03247/retropepsin/RT_RNaseH
GCE33	*Hipposideros armiger*	NW_017731971	34383	43521	*LOC102878255-gag-BOLA1*	ERVI-99.91%	RT_like/Gag_p30/retropepsin/RT_RNaseH
GCE33	*Pteropus pselaphon*	BMBI01007269	1355045	1364457	*LOC102878255-gag-BOLA1*	ERVI-99.94%	RT_like/Gag_p30/retropepsin/RT_RNaseH/PHA03247
GCE33	*Eonycteris spelaea*	PUFA01000202	1597126	1606180	*LOC102878255-gag-BOLA1*	ERVI-99.22%	RT_like/Gag_p30/retropepsin/RT_RNaseH
GCE33	*Macroglossus sobrinus*	PVKZ01001179	2771	11826	*LOC102878255-gag-BOLA1*	ERVI-98.17%	RT_like/Gag_p30/retropepsin/RT_RNaseH
GCE33	*Cynopterus brachyotis*	SSHV01003821	386992	394321	*LOC102889931-gag-SV2A**	ERVI-98.21%	RT_like/Gag_p30/retropepsin/RT_RNaseH/ZnF_C2HC
GCE33	*Eidolon helvum*	AWHC01155135	1	7068	*	ERVI-99.01%	RT_like/Gag_p30/retropepsin
GCE33	*Rousettus leschenaultii*	BNJM01000090	18089516	18098698	*LOC102878255-gag-BOLA1*	ERVI-98.32%	RT_like/Gag_p30/retropepsin/RT_RNaseH/ZnF_C2HC
GCE33	*Hipposideros galeritus*	PVLB01010137	6576	14836	*LOC102878255-gag-BOLA1*	ERVI-99.94%	RT_like/Gag_p30/retropepsin/RT_RNaseH2/RT_RNaseH
GCE33	*Megaderma lyra*	PVJL010013404	3932	11233	*LOC102878255-gag-BOLA1*	ERVI-86.98%	RT_like/RT_RNaseH_2/RT_RNaseH/Gag_p30

Appendix Table 2: the accessions or sources of viral gene sequences used for phylogenetic analysis of *GCE32*, ancestral sequence reconstruction, and rooting phylogenetic trees.

Virus	genome	*gag*	*pol*	*env*
Human chronic myeloid leukemia-associated retrovirus (HCML-ARV)	AY208746.1	AAP06676.1	AAP06677.1	AAP06678.1
Reticuloendotheliosis virus (REV)	NC_006934.1	YP_223870.1	YP_223871.1	YP_223872.1
Duck infectious anemia virus (DIAV)	KF313137.1	AGV92858.1	AGV92859.1	AGV92860.1
Galidia endogenous retrovirus (Galidia ERV)	KF313135.1	AGV92852.1	AGV92853.1	AGV92854.1
Flying fox retrovirus (FFRV)	MK040728.1	QDA02049.1	QDA02050.1	QDA02051.1
Hervey pteropid gammaretrovirus (HPG)	MN413610.1	QJT93246.1	QJT93247.1	QJT93248.1
Porcine endogenous retrovirus-B (PERV-B)	HQ540595.1	AEF12614.1	AEF12615.1	AEF12616.1
Porcine endogenous retrovirus-C (PERV-C)	HM159246.1	ADK35877.1	ADK35878.1	ADK35879.1
Gibbon ape leukemia virus (GaLV)	KT724047.1	ALV83299.1	ALV83300.1	ALV83301.1
Macroglossus minimus gammaretrovirus (MmGRV)	MN413611.1	QJT93249.1	QJT93250.1	QJT93251.1
Friend Murine Leukemia Virus (FMLV)	NC_001362.1	NP_040332.1	NP_040333.1	NP_040334.1
Feline leukemia virus (FeLV)	NC_001940.1	NP_047255.1	NP_047255.1	NP_047256.1
Syconycteris australis gammaretrovirus (SaGRV)	MN413612.1	QJT93252.1	QJT93253.1	QJT93254.1
Mus dunni endogenous virus (MDEV)	AF053745.1	AAC31804.1	AAC31805.1	AAC31806.1
Koala retrovirus (KoRV)	NC_039228.1	YP_009513210.1	YP_009513211.1	YP_009513212.1
Rhinolophus fer retrovirus (RfRV)	JQ303225.1	AFA52558.1	AFA52559.1	AFA52560.1
Hipposideros larvatus gammaretrovirus (HlGRV)	MN413613.1	QJT93255.1	QJT93256.1	QJT93257.1
Porcine endogenous retrovirus-E (PERV-E)	NC_003059.1	predicted from genome	predicted from genome	predicted from genome
Human endogenous retrovirus-E (HERV-E)	From reference sequences in Vargiu *et al* (2016)			
Human endogenous retrovirus-H (HERV-H)				
Human endogenous retrovirus-I (HERV-I)				
Human endogenous retrovirus-W (HERV-W)				
Walleye Dermal Sarcoma Virus (WDSV) – outgroup	NC_001867.1	NP_045938.1	NP_045937.2	NP_045939.1

Appendix Table 3A: *RBM33* orthologs used for phylogenetic analysis.

Gene ID	Gene symbol	Description	Scientific name	Common name	RefSeq Transcript accession	RefSeq Protein accession
381626	*Rbm33*	RNA binding motif protein 33	*Mus musculus*	house mouse	NM_028234.1	NP_082510.1
100512935	*RBM33*	RNA binding motif protein 33	*Sus scrofa*	pig	XM_021079037.1	XP_020934696.1
102254572	*RBM33*	RNA binding motif protein 33	*Myotis brandtii*	Brandt's bat	XM_014532259.1	XP_014387745.1
102422130	*RBM33*	RNA binding motif protein 33	*Myotis lucifugus*	little brown bat	XM_023744270.1	XP_023600038.1
102765519	*RBM33*	RNA binding motif protein 33	*Myotis davidii*	David’s myotis	XM_006767157.2	XP_006767220.2
102882763	*RBM33*	RNA binding motif protein 33	*Pteropus alecto*	black flying fox	XM_025044919.1	XP_024900687.1
103299601	*RBM33*	RNA binding motif protein 33	*Eptesicus fuscus*	big brown bat	XM_028132531.1	XP_027988332.1
105298926	*RBM33*	RNA binding motif protein 33	*Pteropus vampyrus*	large flying fox	XM_023536733.1	XP_023392501.1
107518864	*RBM33*	RNA binding motif protein 33	*Rousettus aegyptiacus*	Egyptian rousette	XM_036230607.1	XP_036086500.1
107524917	*RBM33*	RNA binding motif protein 33	*Miniopterus natalensis*	Nata long-fingered bat	XM_016196983.1	XP_016052469.1
109391496	*RBM33*	RNA binding motif protein 33	*Hipposideros armiger*	great roundleaf bat	XM_019659088.1	XP_019514633.1
112308357	*RBM33*	RNA binding motif protein 33	*Desmodus rotundus*	common vampire bat	XM_024564559.1	XP_024420327.1
114507489	*RBM33*	RNA binding motif protein 33	*Phyllostomus discolor*	pale spear-nosed bat	XM_036011068.1	XP_035866961.1
117018276	*RBM33*	RNA binding motif protein 33	*Rhinolophus ferrumequinum*	greater horseshoe bat	XM_033099199.1	XP_032955090.1
118620394	*RBM33*	RNA binding motif protein 33	*Molossus molossus*	Pallas's mastiff bat	XM_036247460.1	XP_036103353.1
118679428	*RBM33*	RNA binding motif protein 33	*Myotis myotis*	Greater mouse-eared bat	XM_036356174.1	XP_036212067.1
118712269	*RBM33*	RNA binding motif protein 33	*Pipistrellus kuhlii*	Kuhl's pipistrelle	XM_036425834.1	XP_036281727.1
118989378	*RBM33*	RNA binding motif protein 33	*Sturnira hondurensis*		XM_037049500.1	XP_036905395.1
119039278	*RBM33*	RNA binding motif protein 33	*Artibeus jamaicensis*	Jamaican fruit-eating bat	XM_037131970.1	XP_036987865.1
120582131	*RBM33*	RNA binding motif protein 33	*Pteropus giganteus*	Indian flying fox	XM_039837281.1	XP_039693215.1

Appendix Table 3B: *CRK* and *CRKL* orthologs used for phylogenetic analysis.

Gene ID	Gene symbol	Description	Scientific name	Common name	RefSeq Transcript accession	RefSeq Protein accession
12929	*Crkl*	v-crk avian sarcoma virus CT10 oncogene homolog-like	*Mus musculus*	house mouse	NM_007764.5	NP_031790.2
416939	*CRKL*	CRK like proto-oncogene, adaptor protein	*Gallus gallus*	chicken	XM_415233.7	XP_415233.1
608125	*CRKL*	CRK like proto-oncogene, adaptor protein	*Canis lupus familiaris*	dog	XM_844856.6	XP_849949.1
100525363	*CRKL*	CRK like proto-oncogene, adaptor protein	*Sus scrofa*	pig	XM_003132994.4	XP_003133042.1
101634890	*CRKL*	CRK like proto-oncogene, adaptor protein	*Condylura cristata*	star-nosed mole	XM_004694944.2	XP_004695001.1
102248665	*CRKL*	CRK like proto-oncogene, adaptor protein	*Myotis brandtii*	Brandt's bat	XM_005875508.2	XP_005875570.1
102898080	*CRKL*	CRK like proto-oncogene, adaptor protein	*Pteropus alecto*	black flying fox	XM_006908649.3	XP_006908711.1
103286991	*CRKL*	CRK like proto-oncogene, adaptor protein	*Eptesicus fuscus*	big brown bat	XM_008142633.2	XP_008140855.1
105289311	*CRKL*	CRK like proto-oncogene, adaptor protein	*Pteropus vampyrus*	large flying fox	XM_011355934.2	XP_011354236.1
107497859	*CRKL*	CRK like proto-oncogene, adaptor protein	*Rousettus aegyptiacus*	Egyptian rousette	XM_016120192.2	XP_015975678.1
107530545	*CRKL*	CRK like proto-oncogene, adaptor protein	*Miniopterus natalensis*	Natal long-fingered bat	XM_016204328.1	XP_016059814.1
12928	*Crk*	v-crk avian sarcoma virus CT10 oncogene homolog	*Mus musculus*	house mouse	NM_133656.5	NP_598417.2
100192444	*CRK*	CRK proto-oncogene, adaptor protein	*Sus scrofa*	pig	NM_001137636.2	NP_001131108.1
100684309	*CRK*	CRK proto-oncogene, adaptor protein	*Canis lupus familiaris*	dog	XM_038677667.1	XP_038533595.1
101627402	*CRK*	CRK proto-oncogene, adaptor protein	*Condylura cristata*	star-nosed mole	XM_004684803.2	XP_004684860.1
102252934	*CRK*	CRK proto-oncogene, adaptor protein	*Myotis brandtii*	Brandt's bat	XM_014529690.1	XP_014385176.1
102885553	*CRK*	CRK proto-oncogene, adaptor protein	*Pteropus alecto*	black flying fox	XM_006925056.3	XP_006925118.1
103291758	*CRK*	CRK proto-oncogene, adaptor protein	*Eptesicus fuscus*	big brown bat	XM_008147869.2	XP_008146091.1
105291866	*CRK*	CRK proto-oncogene, adaptor protein	*Pteropus vampyrus*	large flying fox	XM_011359888.2	XP_011358190.1
107054794	*CRK*	CRK proto-oncogene, adaptor protein	*Gallus gallus*	chicken	NM_001353939.1	NP_001340868.1
107507729	*CRK*	CRK proto-oncogene, adaptor protein	*Rousettus aegyptiacus*	Egyptian rousette	XM_016138954.2	XP_015994440.1
107535608	*CRK*	CRK proto-oncogene, adaptor protein	*Miniopterus natalensis*	Natal long-fingered bat	XM_016211113.1	XP_016066599.1

Appendix Table 3C: *PEG3*-related genes and orthologs used for phylogenetic analysis.

Gene ID	Gene symbol	Description	Scientific name	Common name	RefSeq Transcript accession	RefSeq Protein accession
664799	*Ctcfl*	CCCTC-binding factor (zinc finger protein)-like	*Mus musculus*	house mouse	NM_001355185.1	NP_001342114.1
105298777	*CTCFL*	CCCTC-binding factor like	*Pteropus vampyrus*	large flying fox	XM_023536614.1	XP_023392382.1
107500099	*CTCFL*	CCCTC-binding factor like	*Rousettus aegyptiacus*	Egyptian rousette	XM_016124948.2	XP_015980434.2
117015998	*CTCFL*	CCCTC-binding factor like	*Rhinolophus ferrumequinum*	greater horseshoe bat	XM_033094978.1	XP_032950869.1
13018	*Ctcf*	CCCTC-binding factor	*Mus musculus*	house mouse	NM_181322.3	NP_851839.1
105305721	*CTCF*	CCCTC-binding factor	*Pteropus vampyrus*	large flying fox	XM_011380480.2	XP_011378782.1
107520246	*CTCF*	CCCTC-binding factor	*Rousettus aegyptiacus*	Egyptian rousette	XM_016163783.2	XP_016019269.1
117034435	*CTCF*	CCCTC-binding factor	*Rhinolophus ferrumequinum*	greater horseshoe bat	XM_033127325.1	XP_032983216.1
53626	*Insm1*	insulinoma-associated 1	*Mus musculus*	house mouse	NM_016889.3	NP_058585.2
105299541	*INSM1*	INSM transcriptional repressor 1	*Pteropus vampyrus*	large flying fox	XM_023536812.1	XP_023392580.1
107511605	*INSM1*	INSM transcriptional repressor 1	*Rousettus aegyptiacus*	Egyptian rousette	XM_016146101.2	XP_016001587.1
117015631	*INSM1*	INSM transcriptional repressor 1	*Rhinolophus ferrumequinum*	greater horseshoe bat	XM_033094251.1	XP_032950142.1
374900	*ZNF568*	zinc finger protein 568	*Homo sapiens*	human	NM_198539.4	NP_940941.2
109378060	*ZNF568*	zinc finger protein 568	*Hipposideros armiger*	great roundleaf bat	XM_019634524.1	XP_019490069.1
18616	*Peg3*	paternally expressed 3	*Mus musculus*	house mouse	NM_008817.2	NP_032843.2
105304872	*PEG3*	paternally expressed 3	*Pteropus vampyrus*	large flying fox	XM_011379186.2	XP_011377488.1
117034566	*PEG3*	paternally expressed 3	*Rhinolophus ferrumequinum*	greater horseshoe bat	XM_033127625.1	XP_032983516.1

Appendix Table 4: the accession numbers (run IDs) of the sequence read archive (SRA) transcriptomic datasets used for analysing *GCE32* expression.

E. helvum	***P. alecto***		***R. aegyptiacus***		
SRR6134344	ERR3619156	SRR8859499	SRR10153049	SRR7548028	SRR6453208
SRR6134348	ERR3619157	SRR8859500	SRR10153050	SRR7548029	SRR6453209
SRR6134351	SRR1524840	SRR8859501	SRR10153051	SRR7548030	SRR6453210
SRR6134352	SRR1531161	SRR8859502	SRR10153052	SRR7609218	SRR6453211
SRR6134353	SRR1531544	SRR8859503	SRR10153053	SRR7609219	SRR6453212
SRR6134354	SRR2995111	SRR8859504	SRR10153054	SRR7609220	SRR6453213
SRR6134357	SRR2995136	SRR8859505	SRR10153055	SRR7609221	SRR6453214
SRR6134358	SRR2995138	SRR8859506	SRR11148658	SRR7609222	SRR6453215
SRR6134360	SRR350710	SRR8859507	SRR11148659	SRR7609223	SRR6453216
	SRR351237	SRR8859508	SRR11148660	SRR7609224	SRR11148699
E. spelaea	SRR5722761	SRR8859509	SRR11148661	SRR7609225	SRR11148700
SRR8836180	SRR5904911	SRR8859510	SRR11148662	SRR7609226	SRR11148701
SRR8836181	SRR5904912	SRR8859511	SRR11148663	SRR7609227	SRR11148702
SRR8836182	SRR5904913	SRR8859512	SRR11148664	SRR7609228	SRR11148703
SRR8836183	SRR5904914	SRR8859513	SRR11148665	SRR7609229	SRR11148704
SRR8836184	SRR5904915	SRR8859514	SRR11148666	SRR7609230	SRR11148705
SRR8836185	SRR5904916	SRR8859515	SRR11148667	SRR7609231	SRR11148706
	SRR5904917	SRR8859516	SRR11148668	SRR7609232	SRR11148707
P. giganteus	SRR5904918	SRR8859517	SRR11148669	SRR7609233	SRR11148708
ERR3348238	SRR5904919	SRR8859518	SRR11148670	SRR7609234	SRR11148709
ERR3348239	SRR5904920	SRR8859519	SRR11148671	SRR7609235	SRR11148710
ERR3348240	SRR5904921	SRR8859520	SRR11148672	SRR7735102	SRR11148711
ERR3348241	SRR5904922	SRR8859521	SRR11148673	SRR7909628	SRR11148712
ERR3348242	SRR5904923	SRR8859522	SRR11148674	SRR2913352	SRR11148713
ERR3348243	SRR5904924	SRR8859523	SRR11148675	SRR2913353	SRR11148714
	SRR5904925	SRR8859524	SRR11148676	SRR2913354	SRR11148715
P. vampyrus	SRR628071	SRR8859525	SRR11148677	SRR2913355	SRR11148716
ERR2012483	SRR8858515	SRR8859526	SRR11148678	SRR2913598	SRR11148717
ERR2012484	SRR8858516	SRR8859527	SRR11148679	SRR2914051	SRR11148718
ERR2012485	SRR8858517	SRR8859528	SRR11148680	SRR2914059	SRR11148719
ERR2012486	SRR8858518	SRR8859529	SRR11148681	SRR2914063	SRR11148720
ERR2012487	SRR8858519	SRR8859530	SRR11148682	SRR2914068	SRR11148721
ERR2012488	SRR8858520	SRR8859531	SRR11148683	SRR2914113	SRR11148722
SRR9719897	SRR8858521	SRR8859532	SRR11148684	ERR2012497	SRR11148723
SRR9719898	SRR8858522	SRR8859533	SRR11148685	ERR2012498	
SRR9719899	SRR8858523	SRR8859534	SRR11148686	ERR2012499	
SRR9719900	SRR8858524	SRR8859535	SRR11148687	ERR2012500	
SRR9719901	SRR8858525	SRR8859536	SRR11148688	ERR2012501	
SRR9719902	SRR8858526	SRR8859537	SRR11148689	ERR2012502	
SRR9719903	SRR8858527	SRR8859538	SRR11148690	SRR2914295	
SRR9719904	SRR8859491	SRR8859539	SRR11148691	SRR2914359	
SRR9719905	SRR8859492	SRR8859540	SRR11148692	SRR2914360	
	SRR8859493	SRR8859541	SRR11148693	SRR2914366	
R. leschenaultii	SRR8859494	SRR8859542	SRR11148694	SRR2914368	
SRR2153214	SRR8859495	SRR953498	SRR11148695	SRR2914369	
SRR6796666	SRR8859496		SRR11148696	SRR2914370	
SRR6796667	SRR8859497		SRR11148697	SRR2914371	
SRR6796668	SRR8859498		SRR11148698	SRR2914372	

Appendix Table 5: matches of *de novo* transcriptome assembly sequence to the original *GCE32* coding sequence in *R. aegyptiacus*.

No.	Sequence	Match statistics
1	AGCTCTACGAGTGCCCGGCATGTGGGGAGTGTTTTGTTCATGGC TCATTCCTCTTCGAGCATCAGAAAGTCCATGAGAAAGATCAGGTT TATGGTTATAGGAGGTATGATGAGCGTTTTGTGCAACCCTCGGTC ATTAACTCCCAGAGGCCTCATGCCCCACAGAAGAACCCTCCTCC AGGGGCGCTCCTTCAGTGTCACATGTGTGGACAAGATTTCATTCA TGGCTCTGTCCTTAACGACCAGATGACAGTTTATACTGGAGAAAA TTTACCAGAGCAGGGCCAGGGCAGTGACAATGCCATCAACCCAG AGTTGGCCCTCACCGAGTTACAGAGAAGTTGTGCCGAAGAGAAA CACTACAAAGGTGAAACCTGCGGAGAATCCTTCCTCAGTCAATCA GACCT	Match to 1889-2294, 406nt, 100% identity
2	GTGAGACAGCAGCCCGTCCGCAAGGTGACACTGACCAACGGGA CCCCGAGGAAGAACCTATCTTCCTACCAGACAGCCCTCCACCAT CGGTGCCTTTGCTGCCCCAACCTGCCCCAGAGCCCTGGCCCGG CACCATGTCCTCTGCCAGGTTCAGGTTCAACTCCTTAGACTGTTC TGCCTGGTACGCGGGGCTGGTATCTCGTCAGGAGGCACAGACT CGGCTCCAGGGCCAGCGCCACGGCATGTTCCTGGTCCGTGACT CCTCTACCCGACCTGGGGACTATGTGCTTTCCGTGTCCGAGAAC CTGCGGGTCTCCCACTACCTCATCAACTCACTGCCCAACCGCCA TTTTAAGATTGGGGACCTGGAGTTTGACCACTTGCCAGCTCTGCT GGAGTTCTACAAGGGTCACTACCTGGACACCACCACCCTGATCG AGCCTGCACCCAGGTATCCAGGCCCATCAGAATATGTACAGACT CTGTATGATTTTCCTGG	Match to 734-1232, 499nt, 100% identity
3	GGTGATTATAGAGAAGCCTGAGGAACAGTGGTGGAGGGCCCGG AACAAGGATGGCCGAATTGGGATGATTCCCGTCCCTTATGTTGAA AGGCTTGTGAGAGGCTTACCACATGGAAAACATAGAAACAGGAG TTCCAATAGTTACGGGATCTCAAAACCTGCTAAGACCTCTGTAAG TAAAAACTATGAACGATCTATCATTCGCAGCTTAGCTTCCACTGAT C	Match to 1275-1498, 224nt, 100% identity
4	GCCAAGAATGTCAGGAGTGTGGGCAATGCTTTGCGACTATTGAA GACCTCAGTGCGCATCAGAAGATCTATGTCCGCGAGGAGTTCTA TGGGGGGAAGCAGCTTGGAGACTCTGTGATTCAGGGCATGGGC CTGGATCGGCCTGAGCAAGATGAGCTAGAGGAGCGGGACAAAC AGGGTGATCCTGAGGACATGATCTATAGGTGCAAGGACTGTGGG CTCGGCTTCAGGGATTGCGCAGACCTTAAGGACCACCAAAAAGT GCATGGCAAAGAGTATCTCACTGACACTC	Match to 1529-1819, 291nt, 100% identity
5	GTGAGTTTGAATGGCCTACCTTTGGGGTTGGATGGCCCTCGGAA GGGACCCTAAACCTCCCCACTGTGGAAGCTGTGTACTGGGTAGT GACAAGGACCCCTGGACATCCAGACCAGTTTCCATACATAGCCT CATGGTTGCACATCGCCACTACACTGCCCCCTTGGATTCGGATCT GTGTGCATAGGCAGGGACAGAGTAAGGTACTTATGGCCCGACTG ACTTGGGGAAATGACAAGGAGGAGCCAACAACAATCCACCAAGG GGGCCCTTCAGCAGCTGCAGTTCCTGCCGGTGGGGGCCCACAT ACACTCAGCAGGCCCACCAGGCATCAAGAACGTCCAGGGACTTT ACCTGGCCAAGAAGGTCCTCGTGCACGGGAGAGGCCAGGGCAT GGCAGGGCAGATGGGCCGCGGGCGCCCAATGTCAA	Match to 116-545, 430nt, 100% identity