RNA BIOLOGY
2023, VOL. 20, NO. 1, 703–714
https://doi.org/10.1080/15476286.2023.2254985

Taylor & Francis
Taylor & Francis Group

BRIEF COMMUNICATION

OPEN ACCESS
Check for updates

A full repertoire of Hemiptera genomes reveals a multi-step evolutionary trajectory of auto-RNA editing site in insect *Adar* gene

Ling Ma[#], Caiqing Zheng[#], Shiwen Xu, Ye Xu, Fan Song, Li Tian, Wanzhi Cai, Hu Li, and Yuange Duan

Department of Entomology and MOA Key Lab of Pest Monitoring and Green Management, College of Plant Protection, China Agricultural University, Beijing, China

ABSTRACT

Adenosine-to-inosine (A-to-I) RNA editing, mediated by metazoan ADAR enzymes, is a prevalent post-transcriptional modification that diversifies the proteome and promotes adaptive evolution of organisms. The *Drosophila Adar* gene has an auto-recoding site (termed S>G site) that forms a negative-feedback loop and stabilizes the global editing activity. However, the evolutionary trajectory of *Adar* S>G site in many other insects remains largely unknown, preventing us from a deeper understanding on the significance of this auto-editing mechanism. In this study, we retrieved the well-annotated genomes of 375 arthropod species including the five major insect orders (Lepidoptera, Diptera, Coleoptera, Hymenoptera and Hemiptera) and several outgroup species. We performed comparative genomic analysis on the *Adar* auto-recoding S>G site. We found that the ancestral state of insect S>G site was an uneditable serine codon (unSer) and that this state was largely maintained in Hymenoptera. The editable serine codon (edSer) appeared in the common ancestor of Lepidoptera, Diptera and Coleoptera and was almost fixed in the three orders. Interestingly, Hemiptera species possessed comparable numbers of unSer and edSer codons, and a few 'intermediate codons', demonstrating a multi-step evolutionary trace from unSer-to-edSer with non-synchronized mutations at three codon positions. We argue that the evolution of *Adar* S>G site is the best genomic evidence supporting the 'proteomic diversifying hypothesis' of RNA editing. Our work deepens our understanding on the evolutionary significance of *Adar* auto-recoding site which stabilizes the global editing activity and controls transcriptomic diversity.

ARTICLE HISTORY
Revised 31 July 2023
Accepted 30 August 2023

KEYWORDS
A-to-I RNA editing; Adar auto-recoding; feedback loop; evolutionary trajectory; Hemiptera

Introduction

A-to-I RNA editing and the evolutionary significance

Adenosine-to-inosine (A-to-I) RNA editing is the most prevalent RNA modification in metazoans [1], fungi [2] and bacteria [3]. Thousands to millions of adenosines in the transcriptomes of different species are potentially editable [4,5], representing the high prevalence of this RNA modification. Adenosines in the double-stranded RNA (dsRNA) structures are deaminated to inosines by a specific enzyme family termed Adenosine Deaminase Acting on RNA (ADAR) (Figure 1a). Inosine is structurally similar to guanosine and therefore A-to-I RNA editing leads to similar effect as A-to-G mutation [6]. For example, A-to-I editing in coding sequence (CDS) might cause nonsynonymous mutations and alter the protein sequence. Nonsynonymous editing sites are also termed recoding sites.

However, one essential difference between RNA editing and DNA mutation is that RNA editing is controllable and could selectively take place in different tissues, cell types and developmental stages. Brains and nervous systems usually bear the most abundant RNA editing events [7]. This temporal-spatial flexibility of RNA editing avoids the pleiotropic effect of DNA mutation and makes RNA editing an advantageous mechanism that promotes adaptive evolution of organisms [8]. Two complementary hypotheses try to explain the adaptation of nonsynonymous RNA editing (recoding) during evolution. The 'diversifying hypothesis' believes that RNA editing confers adaptiveness by elevating the proteomic diversity in a flexible manner [9], increasing the sequence space of conserved genes and allowing organisms to adapt to changeable environment [10]. This diversifying role of RNA editing is especially typical in insects, where phenotypic divergence is prevalent across different developmental stages and hierarchical castes [11,12]. The transcriptomic and proteomic plasticity, which is the basis of phenotypic diversity in insects, is largely shaped by RNA editing.

Another hypothesis about adaptive RNA editing is the 'restorative hypothesis'. This hypothesis states that A-to-I(G) RNA editing is designed for correcting G-to-A DNA mutations to restore the ancestral G allele [13]. Under this restorative hypothesis, although the edited allele is no fitter than the ancestral G allele (because the editing level is usually lower than 100%), the editing mechanism itself still increases the

CONTACT Yuange Duan duanyuange@cau.edu.cn Department of Entomology and MOA Key Lab of Pest Monitoring and Green Management, College of Plant Protection, China Agricultural University, Beijing 100193, China

[#]Co-first author.

Supplemental data for this article can be accessed online at https://doi.org/10.1080/15476286.2023.2254985.

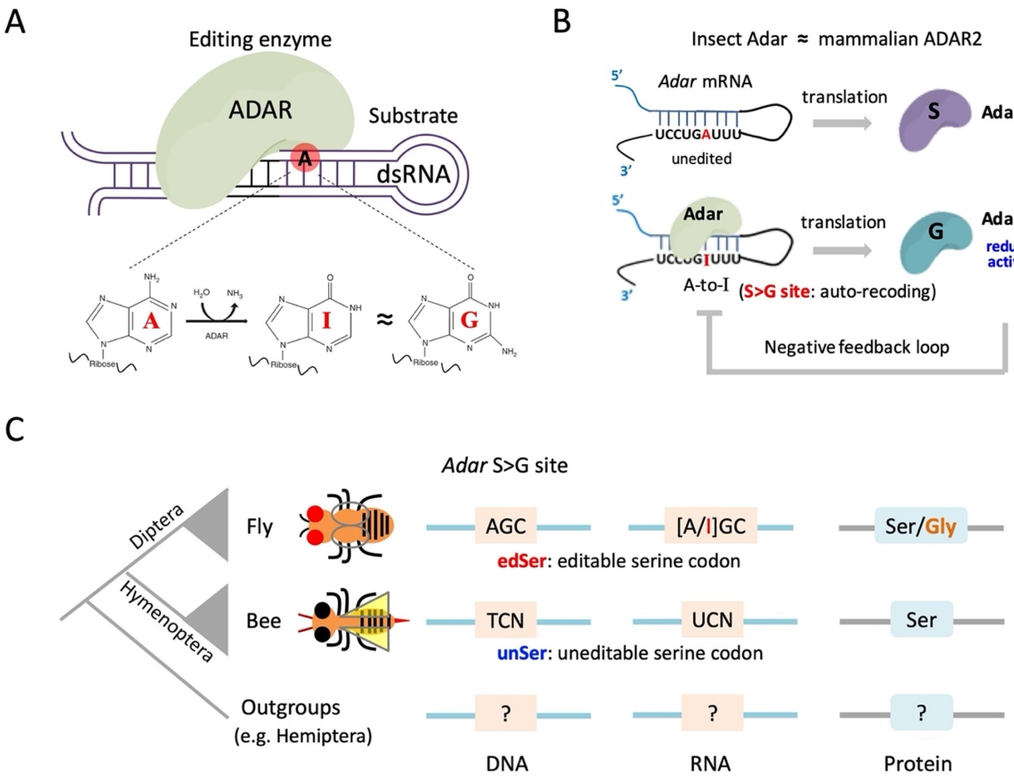


Figure 1. A-to-I RNA editing and the auto-recoding site in insect *Adar* gene. (a) The occurrence of A-to-I RNA editing. I is structurally similar to G. (b) The auto-feedback loop of *Adar* S>G site in *Drosophila*. The Adar^G isoform has lower editing activity than the Adar^S isoform. (c) Fixation of *Adar* S>G site in flies and bees. Flies have an editable serine codon (edSer) and bees have an uneditable serine codon (unSer). But the evolutionary trace in a larger scale is unclear.

fitness of the current species because the partially edited RNA pool (G > 0%) is certainly better than no editing at all (G = 0%). Thus, the current species still benefits from the RNA editing mechanism so that editing is still adaptive under restorative hypothesis. Moreover, in order to fully correct the DNA mutation, the RNA editing level should be as high as possible and avoid tissue-specificity. These stringent criteria limited the restorative hypothesis to a particular range of species where the unedited allele is apparently non-functional [14].

Auto-recoding site in *Drosophila Adar* gene regulates global RNA editing activity

The adaptive hypothesis of RNA editing emphasizes the flexible control of editing events in different tissues and stages. It requires that the occurrence of RNA editing is non-random and should be accurately regulated. Apart from the *cis* elements like the temperature-sensitive RNA structure and sequence context that could determine the 'editability' and editing level of a particular site [15,16], a stronger determinant is the *trans* factor (ADAR enzyme) that affects the global editing level. Mammals have three ADAR proteins, among which ADAR3 is inactive [17]. Insects have lost *ADAR1*, and the only *Adar* gene in insect genomes is orthologous to mammalian *ADAR2* [18]. In *Drosophila*, Adar can edit its own transcript and change a serine codon (AGC) to an IGC codon which is decoded as glycine [19]. This auto-recoding site is termed Ser>Gly (S>G) site (Figure 1b). The protein resulted from the Gly isoform (termed Adar^G) has lower activity than the genomically encoded Adar^S isoform. Therefore, this S>G site makes a negative feedback loop that stabilizes the total editing activity (as well as the global editing level in the transcriptome) (Figure 1b). This feedback loop mediated by auto-recoding of *Adar* gene only exists in insects and therefore the insect clade is the perfect representative for understanding this evolutionary phenomenon.

Interestingly, *Adar* S>G auto-recoding site is highly conserved in Diptera (Figure 1c), with editing levels ranging from 20%~40% in heads of different *Drosophila* species. This highlights the evolutionary importance of this auto-feedback mechanism that regulates Adar activity. One would intuitively expect that this editable serine codon (AGC/U, denoted as edSer) should be conserved across the entire insect clade. However, in the genomes of Hymenoptera species (such as honeybee *Apis mellifera*), the orthologous site is an 'uneditable' serine codon (UCN), denoted as unSer (Figure 1c). This unSer codon could not be edited to Gly so that the auto-regulatory mechanism of Adar is abolished in Hymenoptera.

Here comes several puzzles regarding the evolution and significance of this *Adar* auto-recoding site: (1) What is the orthologous codon in other insect orders like Lepidoptera, Coleoptera and Hemiptera? (2) Could this edSer codon be independently gained in other insects? (3) If the ancestral state is an unSer codon (UCN), how did this codon evolve to an edSer codon (AGC/U)? Is there an 'intermediate codon' during the transition? Answers to these questions would help us understand the origin and evolution of *Adar* auto-regulatory mechanism in insects.

Particularly, Hemiptera is of our interest as it represents the most successful incomplete metamorphosis insects [20]. Hemiptera is the fifth largest insect order, and hemipteran species has amazingly high phenotypic diversities at both inter-species level and intra-species level (plasticity). They have adapted to a wide range of habitats and evolved various feeding traits [21–23]. It is possible that the transcriptomic regulation like A-to-I RNA editing could be the molecular basis underlying the phenotypic diversity and plasticity across Hemiptera. There might be interesting findings in the *Adar* S>G auto-recoding sites in Hemiptera which reflects a general evolutionary route of this mechanism in insects.

Aims and scopes

In this study, we aim to unravel the evolutionary trajectory of *Adar* S>G site in insects. We retrieved the well-annotated genomes of 375 arthropod species including the five major insect orders Lepidoptera, Diptera, Coleoptera, Hymenoptera, Hemiptera and several outgroup species. We performed comparative genomic analysis on the *Adar* auto-recoding S>G site. We found that the ancestral state of insect S>G site was unSer and that this state was largely maintained in Hymenoptera. The edSer appeared in the common ancestor of Lepidoptera, Diptera and Coleoptera and was almost fixed in the three orders. Interestingly, Hemiptera species possessed comparable numbers of unSer, edSer and a few ‘intermediate codons’, demonstrating a multi-step evolutionary trace from unSer-to-edSer with non-synchronized mutations at three codon positions. We argue that the evolution of *Adar* S>G site is the best genomic evidence supporting the ‘diversifying hypothesis’ of RNA editing. Our work deepens our understanding on the evolutionary significance of *Adar* auto-recoding site that stabilizes the global editing activity and transcriptomic diversity.

Results

Clarifying the evolution of insect *Adar* S>G site across different orders

To determine the evolutionary dynamics of *Adar* S>G site in the whole insect class, we retrieved the annotated genomes of 375 arthropod species including the five major insect orders (in phylogenetic order) Lepidoptera (74 genomes), Diptera (113 genomes), Coleoptera (38 genomes), Hymenoptera (104 genomes), Hemiptera (32 species) and the outgroups (14 species) of the five orders (Figure 2a and **Supplementary Table S1**). We searched the *D. melanogaster* Adar protein sequence against the 375 arthropod genomes and obtained exactly one *Adar* gene for each species (**Materials and Methods**), supporting the notion that arthropods have a single *Adar* gene (which is orthologous to mammalian *ADAR2*) [18].

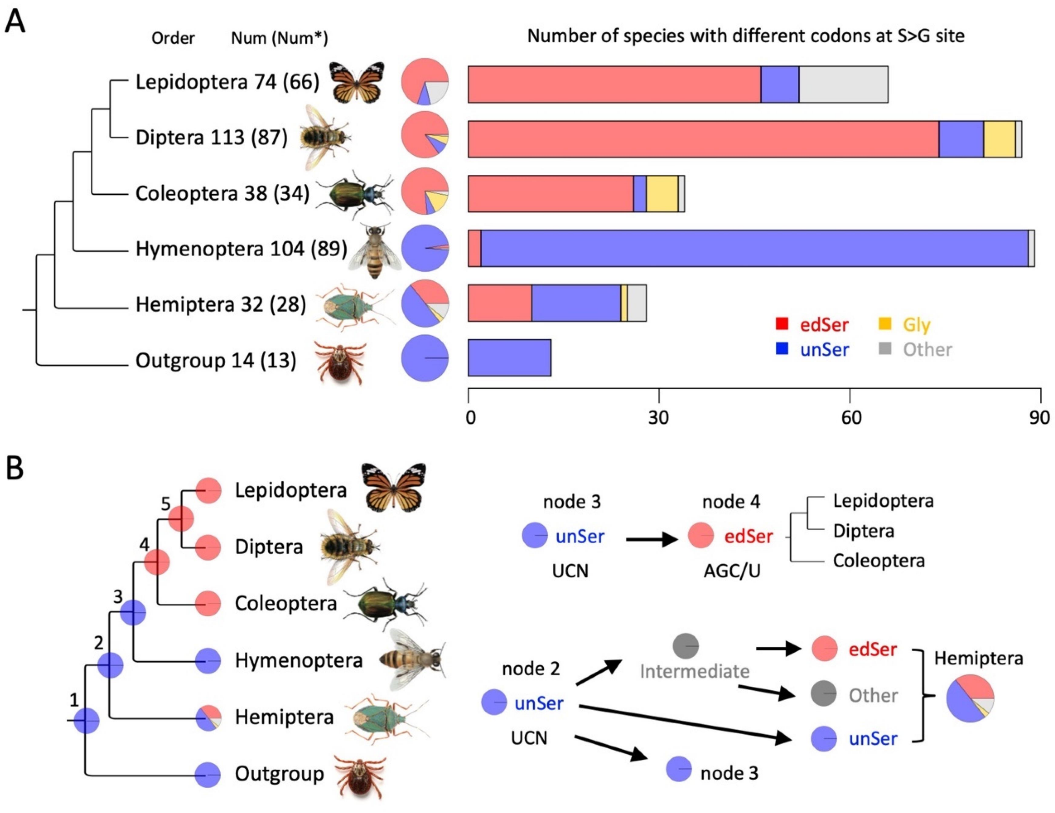


Figure 2. Evolution of Adar S>G site in five major insect orders. (a) Numbers and proportions of different codons at the orthologous S>G site. The numbers of available species were labeled for each order. The numbers in parentheses represented the number of species without a gap at the orthologous site in the alignment. The branch lengths were unscaled. (b) The most likely evolutionary trajectory from unSer to edSer. The transition from node 3 to node 4 lacked an intermediate codon(s) to bridge the unSer and edSer codons.

We aligned the CDS of *Adar* gene of the 375 species and extracted the corresponding position of S>G site from the alignment. Taken the encoded amino acids into account, these 375 codons were classified into editable serine (edSer), uneditable serine at the 1st position (unSer), glycine (Gly), gap and other codons. For the five major insect orders and the outgroup, we calculated the proportion of each codon class in these clades (Figure 2a). Interestingly, we found that (1) the majority of codons in the three most recent branches (Lepidoptera, Diptera and Coleoptera) were edSer; (2) the majority of codons in Hymenoptera and outgroups were unSer; (3) Hemiptera had considerable numbers of both edSer and unSer codons (Figure 2a).

This clearly raises a parsimonious evolutionary trajectory of this S>G site in major insect clades (Figure 2b): (1) The ancestral state of all insects is an unSer codon; (2) The common ancestor of Lepidoptera, Diptera and Coleoptera has obtained an edSer codon and this sequence has been maintained throughout the three orders. The well-studied S>G site in *D. melanogaster* is such an example; (3) Hemiptera has independently obtained the edSer codon, but there remains a plenty of hemipteran species bearing the unSer codon.

Here comes a puzzle regarding the evolution of S>G site in insects. From the ancestral unSer to the edSer in Lepidoptera, Diptera and Coleoptera, there lacks an 'intermediate codon' (Figure 2b). The unSer codons are UCA/C/G/U (UCN), and the edSer codons are AGC/U. There could not be a one-step transition (point mutation) from unSer to edSer, making the evolutionary trajectory of S>G site very obscure. Indeed, there were some 'other codons' in the orthologous site in Lepidoptera, Diptera and Coleoptera that might be the 'intermediate codon' (Figure 2a), but we have already inferred that the ancestral state of the three orders was edSer (Figure 2b) so that the so-called 'other codons' were actually derived codons that came from random drift. Therefore, to understand how an ancestral unSer codon has gradually evolved to an edSer codon, we need to find evidence from other clades, such as Hemiptera, with potential 'intermediate codons' (Figure 2b). These intermediate codons in Hemiptera, together with the presence of both unSer and edSer, might be helpful for unravelling the evolution of *Adar* S>G site.

Adar auto-recoding S>G site is independently gained in Hemiptera with clear evolutionary trajectory

Hemiptera is the most phenotypically diversified order in insects. Notably, on the orthologous site of *Adar* S>G site, Hemiptera had comparable unSer codons and edSer codons and meanwhile with a few 'other codons'. Together with the overall phylogeny (Figure 2), it suggests that the ancestor of all hemipteran species had an unSer codon at this site and the edSer codon was a derived state and that the few 'other codons' were likely to be the genuine 'intermediate codons'.

We constructed the phylogenetic tree of the 32 hemipteran species and mapped the S>G site to this phylogeny (Figure 3; also see **Materials and Methods**). Clearly, (1) the ancestral codon was the unSer codons (14 species) in Heteroptera, Auchenorrhyncha and part of Sternorrhyncha; (2) The intermediate codons were the Ala (3 species) and Gly (1 species) codons present in Coccoidea; and (3) The derived codon was the edSer codons in Aphididae (10 species) (Figure 3). There were four species showing gaps in the orthologous site in the alignment, potentially due to the fluctuating quality of the genome assembly across different loci.

With similar strategy, here we took a look at the alignment of S>G site in Hymenoptera, Coleoptera, Diptera and Lepidoptera (Figure 4). For Hymenoptera, as its ancestral state was unSer, it would be easily inferred from the tree that a small clade of two species has independently gained an edSer (Figure 4a). For Coleoptera, Diptera and Lepidoptera, as we have inferred that their ancestral state was an edSer, it is conceivable that a small clade in Coleoptera has re-obtained the unSer (Figure 4b). But for Diptera (Figure 4c) and Lepidoptera (Figure 4d), multiple independent gains of unSer occurred (although with very few species), this suggests that there might be ancestral polymorphism at this codon in Diptera and Lepidoptera (but the possibility of reverse mutation could not be ruled out). Taken all these five major insect clades, only Hemiptera has sufficient and comparable numbers of edSer and unSer codons at S>G site to make a solid analysis on its evolutionary trajectory.

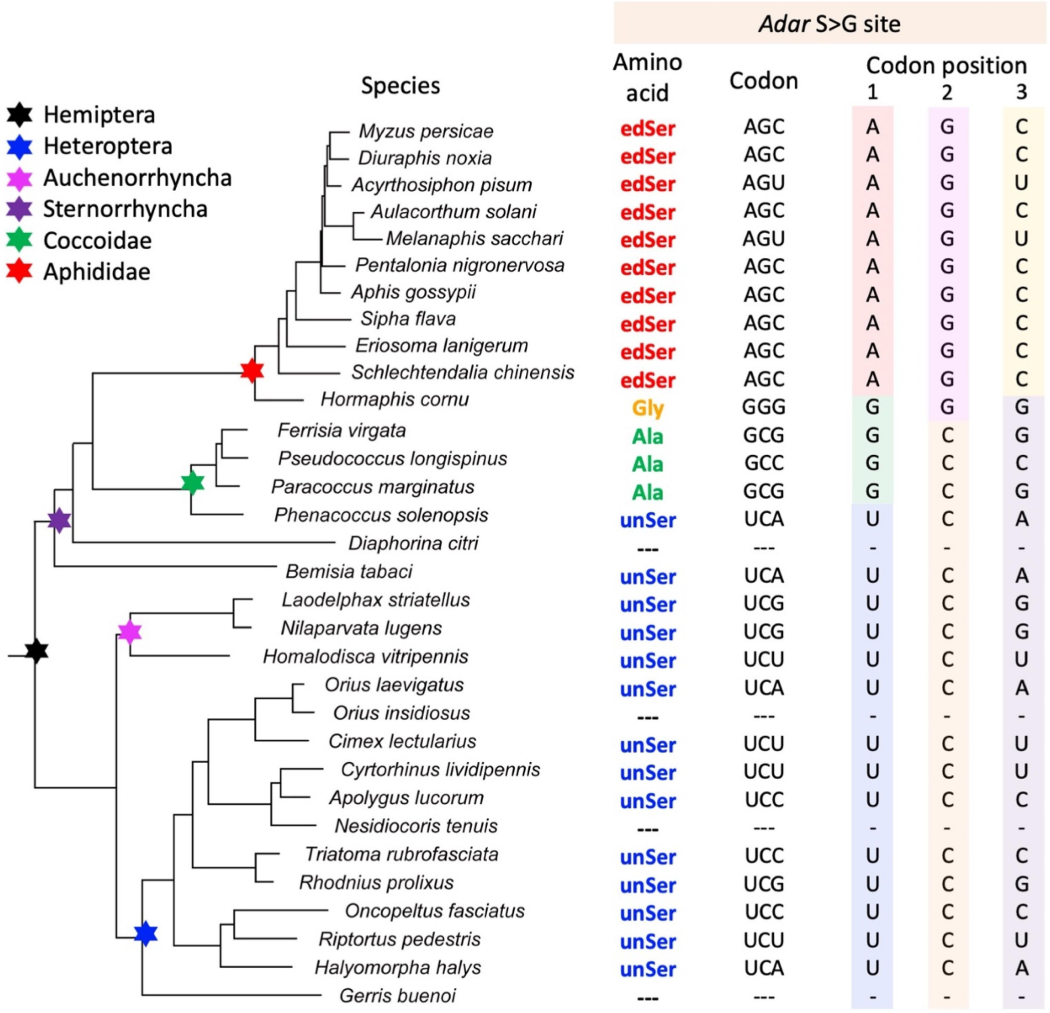
For the S>G site in Hemiptera, we looked at the three codon positions separately and the evolutionary trajectory became even clearer (Figure 3). Position 1 had a T(U) in the ancestral nodes and experienced a T>G mutation in an inner node and finally made a G>A change in the latest-split Aphididae branch (Figures 3 and 5). Position 2 was much simpler in evolution, where the ancestral state was C and the derived nucleotide was G (Figures 3 and 5). For the evolutionarily flexible position 3, all A/C/G/T(U) were present in the ancestral unSer codons, and the derived edSer codons had C and T(U) (Figures 3 and 5).

Notably, the changes of the three codon positions were not synchronized (Figure 5), suggesting that the change from an ancestral unSer codon to the most recent edSer codon was accumulated by multiple steps of point mutations. Without the availability of the intermediate codons in a few hemipteran species, the unSer-to-edSer transition could only be explained by an unusual simultaneous change at all codon positions (such as the transition from node 3 to node 4 in Figure 2). Therefore, Hemiptera has served as an ideal clade to unravel the evolutionary history of the auto-recoding S>G site in *Adar* gene.

The best evolutionary evidence for the proteomic diversification role of nonsynonymous RNA editing

So far, we have clarified the multi-step evolution of the *Adar* S>G site in Hemiptera, which successfully filled the gap between the ancestral unSer codon and the derived edSer codon. Unravelling this mystery would help us better understand the evolutionary significance of the auto-recoding mechanism of *Adar* gene and the importance of global regulation of A-to-I RNA editing.

More importantly, regarding the two hypotheses on the significance of nonsynonymous RNA editing (diversifying



	Adar S>G site				
Species	Amino acid	Codon	Codon position 1	2	3
Myzus persicae	edSer	AGC	A	G	C
Diuraphis noxia	edSer	AGC	A	G	C
Acyrthosiphon pisum	edSer	AGU	A	G	U
Aulacorthum solani	edSer	AGC	A	G	C
Melanaphis sacchari	edSer	AGU	A	G	U
Pentalonia nigronervosa	edSer	AGC	A	G	C
Aphis gossypii	edSer	AGC	A	G	C
Sipha flava	edSer	AGC	A	G	C
Eriosoma lanigerum	edSer	AGC	A	G	C
Schlechtendalia chinensis	edSer	AGC	A	G	C
Hormaphis cornu	Gly	GGG	G	G	G
Ferrisia virgata	Ala	GCG	G	C	G
Pseudococcus longispinus	Ala	GCC	G	C	C
Paracoccus marginatus	Ala	GCG	G	C	G
Phenacoccus solenopsis	unSer	UCA	U	C	A
Diaphorina citri	---	---	-	-	-
Bemisia tabaci	unSer	UCA	U	C	A
Laodelphax striatellus	unSer	UCG	U	C	G
Nilaparvata lugens	unSer	UCG	U	C	G
Homalodisca vitripennis	unSer	UCU	U	C	U
Orius laevigatus	unSer	UCA	U	C	A
Orius insidiosus	---	---	-	-	-
Cimex lectularius	unSer	UCU	U	C	U
Cyrtorhinus lividipennis	unSer	UCU	U	C	U
Apolygus lucorum	unSer	UCC	U	C	C
Nesidiocoris tenuis	---	---	-	-	-
Triatoma rubrofasciata	unSer	UCC	U	C	C
Rhodnius prolixus	unSer	UCG	U	C	G
Oncopeltus fasciatus	unSer	UCC	U	C	C
Riptortus pedestris	unSer	UCU	U	C	U
Halyomorpha halys	unSer	UCA	U	C	A
Gerris buenoi	---	---	-	-	-

Figure 3. Phylogenetic tree of the available 32 Hemiptera species. the information of orthologous site at *Adar* S>G site was shown next to each species.

hypothesis *versus* restorative hypothesis) [8,13,14], this *Adar* S>G site is undoubtedly the best genomic evidence (not experimental evidence) to prove the proteomic diversification role of nonsynonymous RNA editing in insects. The diversifying hypothesis stresses the advantage of editable status over the uneditable status (such as flexible regulation of proteomic diversity). This scenario, where an uneditable Ser codon has evolved to an editable Ser codon (and the editing on this codon leads to nonsynonymous mutation) (Figure 6a), is the direct comparison between the pre-editing protein version (Ser) *versus* the sum of pre-editing and post-editing protein versions (Ser + Gly). This is the most accurate observation of 'the gain of an editing site' during evolution (Figure 6a), but this case is not always available. The reasons go as follows:

(1) If the ancestral state of a nonsynonymous editing site is G, then this evolutionary trace would somehow support the restorative hypothesis (Figure 6b) and is not the gain of an editing site; (2) If the ancestral state of a nonsynonymous editing site is C or T(U), then the codons are not comparable at all due to the different AAs encoded by the ancestral (C/T), pre-editing (A) and post-editing codons (G) (Figure 6c). This is not the gain of an editing site either; (3) If the ancestral state of a nonsynonymous editing site is A, then it is more likely to be a conserved editing site in the ancestral node instead of an unedited adenosine (because no historical data were there to prove that the adenosine was unedited, especially when the sequence context of the ancestral adenosine was highly conserved to the current one). Under these situations, there are no direct transitions from the 'pure pre-editing version' to the editable status (Figure 6d) so this case is not the gain of an editing site either.

In fact, the only chance to achieve the 'pure pre-editing version' of the protein sequence is to find a synonymous codon without adenosine. This possibility only exists for the Ser and Arg which have six synonymous codons and the change from A-to-G causes a nonsynonymous mutation: for Ser, AGC/U are editable and UCA/C/G/U are uneditable at nonsynonymous positions (Figure 6a), and for Arg, AGA/G are editable and CGA/C/G/U are uneditable at nonsynonymous positions. Therefore, in Hemiptera and the entire insect clade, an uneditable Ser codon has been mutated to an editable Ser codon, and this transition has independently taken place for multiple times. This observation is the best genomic evidence to support the diversifying role of RNA editing.

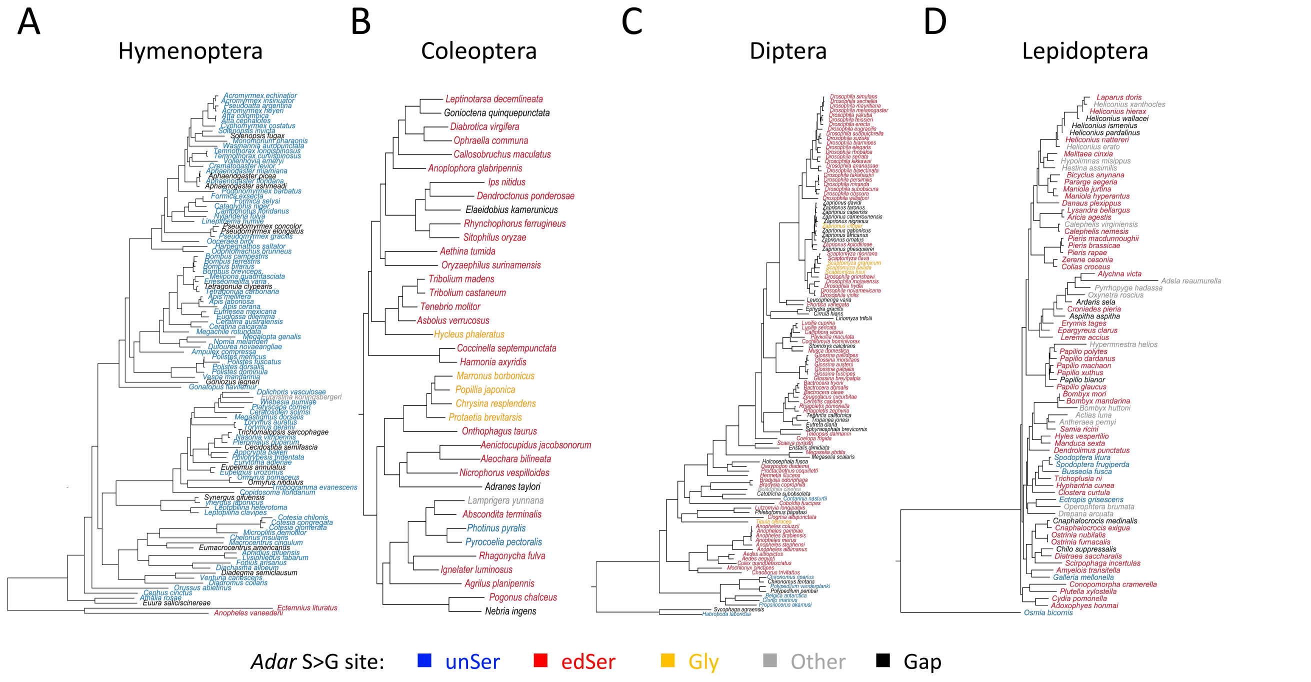


Figure 4. Phylogenetic tree of Hymenoptera, Coleoptera, Diptera and Lepidoptera. The amino acid and codon information of orthologous site at *Adar* S>G site was indicated by color. Blue represents unSer. Red represents edSer. Orange represents Gly. Grey represents other codons. Black represents gap.

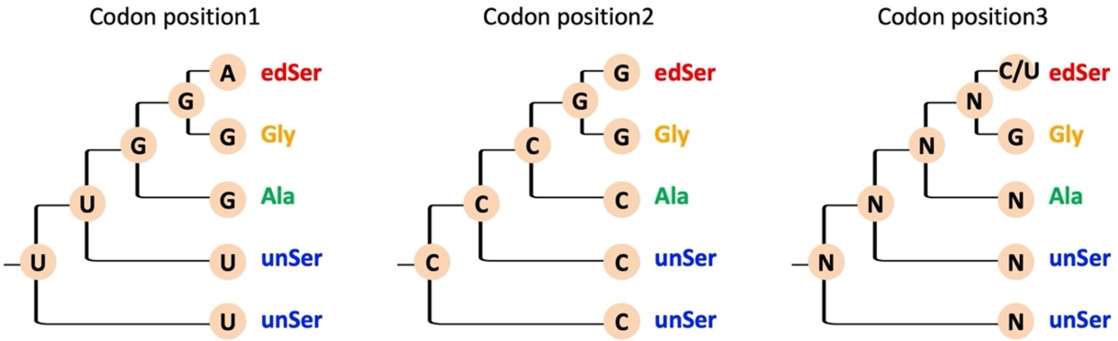


Figure 5. The evolutionary trajectory of the three codon positions of *Adar* S>G site in Hemiptera. N represents any nucleotide of A/C/G/U.

Long distant co-evolution between RNA binding domain and the S>G site in Adar?

Question comes that could we observe co-evolution between the RNA binding domain of Adar and its S>G site? Multiple studies revealed critical insights into the co-evolution of molecular interactions between RNA binding protein and RNA sequence [24,25]. This co-evolution became more interesting if the RNA binding protein would target its own mRNA (like Adar). For example, Gemin5 bound its own CDS to regulate protein translation and the substitution in the RNA binding domain would drastically reduce the binding affinity [24]. This suggests the collaboration between the binding domain and the RNA sequence. The same pattern might go for Adar.

The insect Adar protein typically consists of two dsRNA-binding domains at the N-terminal (AA positions 56–118 and 201–247) and a deamination (catalytic) domain at the C-terminal (AA positions 294–665). The auto-recoding S>G site locates in the deamination domain (Figure 7). Since we already know that the S>G site showed significant sequence divergence between Diptera and Hymenoptera, we wonder whether the RNA-binding domains show co-evolution in the two insect orders. We aligned the AA sequence of the two RNA-binding domains of Adar in Diptera and Hymenoptera. We have the following findings. While the Adar S>G site showed distinct divergence between Diptera and Hymenoptera (Figures 1c and 2), the RNA-binding domains showed relatively conserved AA sequences (Figure 7). However, we still observed that the intra-order difference was much lower than the inter-order difference. For domain 1, the pairwise AA identity among Diptera was 75.8% and the pairwise identity among Hymenoptera was 62.4%, and the between-order identity was 57.0%. Similarly, for domain 2, the pairwise identity among Diptera was 66.2% and the pairwise identity among Hymenoptera was 62.4%, but the between-order identity was only 43.8% (Figure 7). Therefore, the Diptera-Hymenoptera divergence at Adar RNA-binding

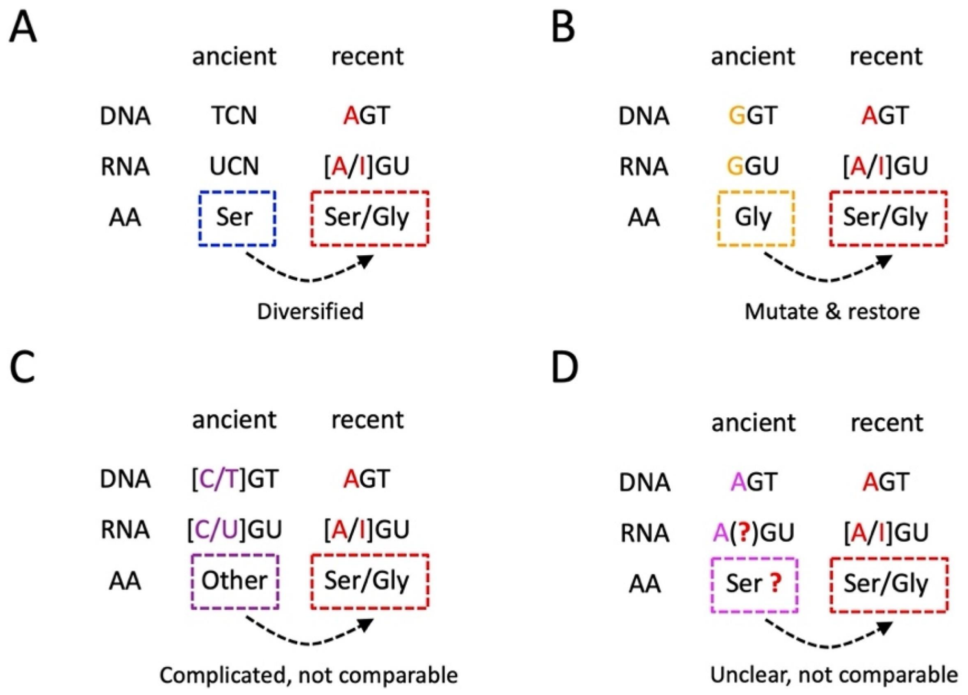


Figure 6. Different evolutionary trajectories of the *Adar* S>G site and their biological significance. (a) With an ancestral uneditable serine codon. This is what we observed in Hemiptera. (b) With an ancestral glycine codon. (c) With an ancestral C or T(U) at editing site. (d) With an ancestral adenosine at the editing site. Whether the ancestral adenosine was edited remains unclear.

domains was observable but was not as distinct as the divergence at Adar S>G site, these might serve as signals of co-evolution between RNA binding domain and the S>G site in insect Adar.

How unexpected it is to observe an unSer-to-edSer codon transition during evolution?

To further prove that the unSer-to-edSer transition is driven by the need for an auto-recoding mechanism (instead of random drift), we should quantitatively calculate how unexpected it is to observe this codon transition. This estimation could be achieved by parsing the genome-wide codon transitions from the unSer species to the edSer species. If this unSer-to-edSer transition is extremely rare across the genome (compared to, e.g., other synonymous mutations at inter-species level), then this will prove the selection-driven nature of the *Adar* auto-recoding site.

Indeed, we noticed that Arg codons also have both editable ones (AGA & AGG, denoted as 'edArg') and uneditable ones (CGN, denoted as 'unArg') at the first codon position. Editing at the first codon position of AGA/AGG would change Arg to Gly (GGA/GGG) and therefore this is a nonsynonymous editing site.

In contrast, although Leu also has six codons (UUA, UUG and CUN), the editing at the third codon position of UUA leads to a synonymous mutation. Thus, there is no point comparing the editable Leu codon UUA and the uneditable Leu codons CUN because this editing event does not increase proteomic diversity. However, these Leu codons could be used as negative controls to see whether there is a genome-wide tendency to switch from unSer (unArg) codons to edSer (edArg) codons. If the flexibility of nonsynonymous editing is advantageous, then the transitions from unSer (unArg) to edSer (edArg) should be more frequent than the transition from UUA/G to CUN (Leu).

In the CDS alignment of all orthologous genes in Hemiptera (see **Materials and Methods**), we defined two groups of species. Group 1 has 10 species from *M. persicae* to *S. chinensis*, representing the species with edSer codon at the *Adar* auto-recoding site (Figure 3). Group 2 has 18 species from *P. solenopsis* to *H. halys*, representing the species with unSer codons (or a few with gaps) at the *Adar* auto-recoding site (Figure 3). The species of intermediate codons were not considered. At genome-wide level, according to the sequence alignment (not including *Adar* gene), we found that 753 positions were all unSer codons in group 2 species, 2,155 positions were all edSer codons in group 1 species, but no overlap was found among the two sets of positions. Thus, the ratio of unSer-to-edSer transition is 0 (Figure 8) (not including the case in *Adar*). Similarly, we found that 15 positions were all unArg codons in group 2 species, 2,877 positions were all edArg codons in group1 species, but no overlap was found among the two sets of positions, either. The ratio of unArg-to-edArg transition is again 0 (Figure 8). In sharp contrast, for the Leu codons as negative control, 22 positions were all UUA/G codons in group 2 species, 4,473 positions were all CUN codons in group 1 species, and 4 positions were overlapped. Thus, the ratio of UU[A/G]-to-CUN transition was 4/22 = 0.182 (Figure 8). Apparently, the genome-wide unSer-to-edSer transition ratio (ratio = 0, not including *Adar*) is lower than the transition between Leu codons (which is a negative control). However, the *Adar* auto-

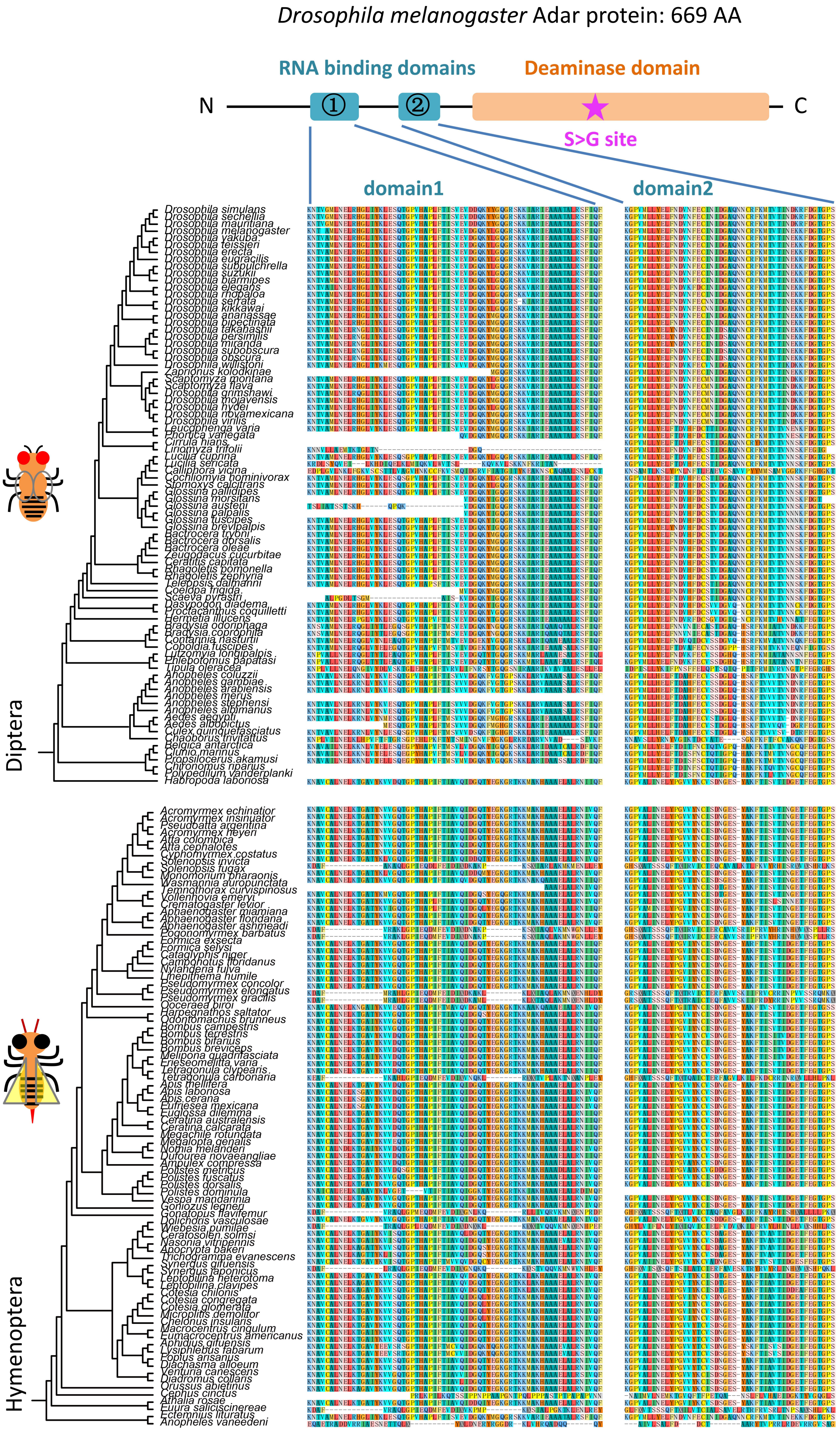


Figure 7. Protein sequence alignment of two RNA binding domains of Adar. Diptera and Hymenoptera species were used.

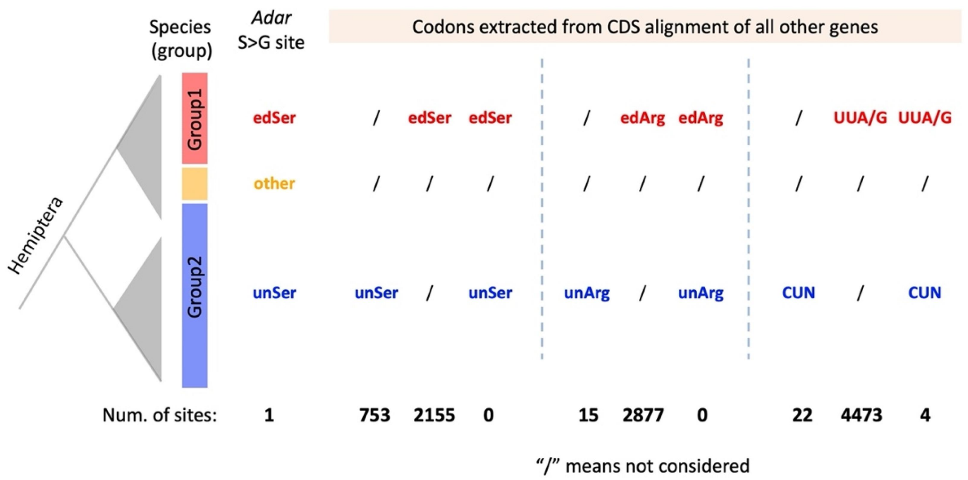


Figure 8. Codons extracted from the CDS alignment of 32 Hemipteran species. The numbers of sites of each category were shown below. "/" means not considered.

recoding site unexpectedly 'chose' to follow that unusual unSer-to-edSer trajectory during evolution, suggesting that there is selection pressure that prompts the emergence of an editable serine codon in the Adar catalytic domain.

Discussion

Summary of main findings

In this work, by using the well-annotated 375 genomes of arthropods, we found that the ancestral state of insect S>G site was unSer and that this state was largely maintained in Hymenoptera. The edSer appeared in the common ancestor of Lepidoptera, Diptera and Coleoptera and was almost fixed in the three orders. Notably, Hemiptera species possessed comparable numbers of unSer, edSer and a few 'intermediate codons', suggesting independent gain of *Adar* auto-recoding site in Hemiptera. It also demonstrates a multi-step evolutionary trace from unSer-to-edSer and we further found that this transition has been accomplished by non-synchronized mutations at three codon positions. As far as we know, this evolutionary pattern of *Adar* S>G site in insects is the best genomic evidence supporting the 'diversifying hypothesis' of RNA editing. Despite the significance of our work as we mentioned throughout the manuscript, there are several points that need to be discussed.

The diversifying hypothesis is not compatible with the restorative hypothesis

Notably, among the 32 available hemipteran species, only one species (*Hormaphis cornu*) in Aphididae had the Gly (GGG) codon (Figure 3). Its occurrence might be regarded as an accident (as the result of random drift). However, if one takes seriously about this single case of Gly, then we will provide stronger explanation that the evolutionary rationale of the edSer conforms with the diversifying hypothesis rather than the restorative hypothesis and that the two hypotheses were mutually exclusive. The site was originally an uneditable Ser; then, this Ser codon was mutated to a Gly codon at the ancestor node of Aphididae. Under the restorative hypothesis, this Gly version appeared to be more favourable than the Ser version. Normally, this advantageous Gly codon would be fixed in Aphididae. However, this Gly codon (GGN) was then mutated to a Ser codon again (AGC/U) in many (not a few) species, reducing the fitness of the host. This trace should be extremely rare. As a consequence, only RNA editing could rescue this DNA mutation and restore the Gly version. This scenario seems plausible under the restorative hypothesis. However, referring to the editing level in other species, this S>G site only had a 20%~30% level, which hardly convinces us that the purpose of editing was to restore the Gly version since >70% of the transcripts were unedited and producing the unfavourable Ser version.

On the other hand, under the diversifying hypothesis [8,9], the flexible regulation of the two protein versions (Ser + Gly) is more advantageous than having only one protein version (either Ser or Gly). Therefore, the evolution of *Adar* S>G site in Hemiptera nicely fits the diversifying hypothesis. The edSer is the derived allele, and unSer is the ancestral allele, while the Gly allele is very rare, which might appear in *Hormaphis cornu* by random drift.

Nevertheless, the diversifying hypothesis was recently verified by experimental evidence in fungi [26,27]. For a particular recoding site in *Fusarium graminearum*, researchers constructed different mutant strains that mimic the uneditable allele and fully edited G allele (hardwired). They found that the uneditable allele was more advantageous in asexual stage, while the fully edited allele was more advantageous in sexual stage [26]. Overall, the editable allele was the fittest due to its flexibility under different stages. The fitness of the fungi was comprehensively measured by the performance during ascus and ascospore formation [26].

Importance of the intermediate codons observed in Hemiptera

For the three species with intermediate Ala codons in Hemiptera (Figure 3), the unSer-to-edSer transition generally resulted from the non-synchronized mutations of the three codon positions. Although we do not know the relative fitness of the Ala allele (and it is irrelevant to both diversifying and restorative hypothesis), their existence is necessary as the unSer codon could not switch to edSer codon by one step. These Ala codons are valuable for inferring the evolutionary route from unSer to edSer, and they also consolidate the notion that the edSer is selectively favoured. If the edSer is neutral or even deleterious, then the Ala codons should be maintained after the unSer-to-Ala mutation. The re-appearance of Ser codons suggests the indispensability of the Ser allele but meanwhile it gained the editability to achieve a higher fitness with temporal-spatial flexibility. Taken together, without these 'Ala species' in Hemiptera, the transition from ancestral unSer (in outgroups) to the edSer (in Lepidoptera, Diptera and Coleoptera) would remain a mystery. However, as we have speculated, the Gly codon in the phylogeny might come from random drift. Since only one species supported the Gly codon, the strength of this speculation was weak. There needs more well-annotated Hemipteran genomes to build a better picture of the intermediate codons and the evolutionary trajectory of *Adar* S>G site.

Regarding our data collection, it is intuitive to think that identifying *Adar* gene from the genome sequence does not require the 'genome annotation file'. However, to build a phylogenetic tree, we need the CDS sequences of all orthologous genes so that the genome annotation is needed for each species. Without a phylogenetic tree, the evolution of *Adar* S>G site could not be studied. This is the reason we focused on genomes with annotation. For example, without requiring a genome annotation file, totally 42 Hemipteran species were available in NCBI, and we found 17 unSer codons, 14 edSer codons, 6 intermediate codons and 5 gaps. However, the number of species was not remarkably elevated, so it did not add much to our understanding on the evolutionary history of *Adar* S>G site in Hemiptera. Moreover, the limiting issue was that since the phylogenetic relationship of these species was not determined, we did not go deeper into this point.

Future perspectives

Since RNA editing is tissue-specific and is most abundant in brains/heads and nerve systems of insects [7], it would be necessary to verify whether the edSer codons (AGC/T) in these 10 hemipteran species were actually edited *in vivo*. Then, the high-quality transcriptome data of different developmental stages or tissues of hemipteran species (if any) would be helpful. The diversifying hypothesis predicts that the editing level on S>G site is not necessarily very high but should exhibit clear tissue-specificity. Even without searching the hemipteran transcriptome data, one would naturally expect a moderate editing level of S>G site in insect heads as inferred from the result of different *Drosophila* species. Moreover, the rationale that why the unSer codon has evolved to the edSer codon in the latest diverged hemipteran species might be answered by comparing the phenotypic divergence between the current unSer species *versus* the current edSer species in Hemiptera. The detailed design, methodology and analysis on these issues would require more efforts in the future.

In summary, we proposed that the evolution of *Adar* S>G site is the best genomic evidence supporting the 'proteomic diversifying hypothesis' of RNA editing. Our work deepens our understanding on the evolutionary significance of *Adar* auto-recoding site which stabilizes the global editing activity and transcriptomic diversity.

Materials and methods

Data availability

The genome accession IDs of the 375 arthropod species were provided as **Supplementary Table S1**. All selected species have both reference genome sequence and genome annotation file in NCBI. No additional filters were performed. Our collection included the five major insect orders (in phylogenetic order) Lepidoptera (74 genomes), Diptera (113 genomes), Coleoptera (38 genomes), Hymenoptera (104 genomes), Hemiptera (32 species) and the outgroups (14 species) of the five orders. The outgroups included all the Pterygota species of incomplete metamorphosis and the available non-insect arthropod Ixodida (**Supplementary Table S1**). The Adar protein sequence of *Drosophila melanogaster* was downloaded from FlyBase (https://flybase.org/) version dm6.04 with protein ID FBpp0308381 (corresponding transcript ID: FBtr0339272, gene ID: FBgn0026086).

Sequence alignment and phylogeny

Insects only have one *Adar* gene which is homologous to mammalian *ADAR2* [18,29]. The *Adar* sequence in model insect *D. melanogaster* is well annotated. We aligned the *D. melanogaster* Adar protein sequence to the CDS sequence of each arthropod species with tblastn [28]. Default parameters were used except we set E-value <1E−6. The hit with the lowest E value was regarded as *Adar* gene in each species. We claim that each species had exactly one *Adar* gene. This was clearly reflected by the fact that (1) most species only has one hit and the corresponding E values were extremely low; (2) for some species with multiple hits, the E values of the non-best hits were remarkably higher than the E value of the best hit, suggesting that the non-best hits were not reliable. In brief, the E-values of the best hits ranged from 0 to 1E−7 (with 2.5%–97.5% quantile from 0 to 2E−22, median = 9E−160), while the E-values for the non-best hits ranged from 5E−50 to 1E−6 (with 2.5%–97.5% quantile from 2.5E−34 to 2E−7, median = 2E−20). Since we set the E-value parameter to 1E−6 in the tblastn alignment, all reported hits would have E-value <1E−6 (otherwise they would not be reported). Therefore, they all fell in the 'significant range'. However, the E-values of best hits (median = 9E−160) and non-best hits (median = 2E−20) were remarkably different (140 orders of magnitude), suggesting that the best hits could be reliably regarded as the authentic *Adar* gene. The fact that the S>G sites of different species were successfully aligned further proves our appropriate selection of *Adar* gene.

The CDS sequences of *Adar* gene were translated into proteins and then aligned with MAFFT version 7.487 [29] with default parameters. The automatic option for selecting parameters and/or algorithms in MAFFT v7.487 (–auto) was used. The CDS sequences were aligned according to the protein alignment to avoid some out-of-frame misalignments [30]. The orthologous sites of the auto-recoding S>G site were extracted from the alignments according to their known positions in *D. melanogaster*.

The phylogeny of the five major insect clades was well studied, and their topology is highly acknowledged. The branch lengths in the topology was unscaled. For the phylogenetic tree of the 32 Hemiptera species, BUSCO sets are defined as collections of near-universal single-copy genes, which are rarely lost or duplicated. We collected 1,367 insect coding genes from database insecta_odb10 in BUSCO v5.4.7 [31]. Orthologous protein sequences of each candidate BUSCO group were aligned using MAFFT v7.487 with auto strategy [29]. Sequence alignments were trimmed and concatenated by TRIMAL v1.4 [36] and FASCONCAT-G v1.0.4 [32]. To reduce the possible systematic errors in large genomic data sets, we calculated the compositional heterogeneity of loci by BACOCA v1.1 [33]. Next, orthologous groups with single-copy orthologues present in 95% of the species were used for phylogenetic tree using IQTREE v2.2.0 under model identified by ModelFinder (−m MFP) [34]. The same strategy was used to build the tree of other insect orders. The phylogenetic tree was visualized by FigTree (http://tree.bio.ed.ac.uk/software/figtree/).

CDS alignment for codon extraction in Hemiptera

We retrieved the 1,367 insect coding genes from insecta_odb10 database and identified each gene in the genomes of 32 Hemipteran species using BUSCO v5.4.7 [31]. Next, gene merging and aligning, together with matrices generation, were executed in a custom script integrating MAFFT v7.487 with auto strategy [29] and FASCONCAT-G v1.0.4 [32]. We generated 95% complete matrices, and the completeness of a matrix represents the lowest ratio of taxa for all alignments. Finally, 710 genes existed in at least 95% of the 32 species. *Adar* was excluded in the alignment since this was independently done in the other sections.

Classification of editable and uneditable codons

In all arthropod species, the orthologous sites of the *Drosophila* S>G site were classified into five groups: editable Ser codons (at the 1st position) include AGT/C, uneditable Ser codons (at the 1st position) include TCN (*N* = A/C/G/T), post-edit version is GGN (Gly), unaligned regions are gaps, and the remaining is denoted as others. In our manuscript, T would be replaced with U when referring to mRNA codons.

Statistics and graphical works

Statistics and graphical works were accomplished in R language (version 3.6.3).

Abbreviations

A-to-I	adenosine-to-inosine.
ADAR	adenosine deaminase acting on RNA.
dsRNA	double-stranded RNA.
CDS	coding sequence.
edSer	editable serine codon.
unSer	uneditable serine codon.
AA	amino acid.

Acknowledgments

We thank the 2115 Talent Development Program of China Agricultural University for the financial support.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This study is financially supported by the National Natural Science Foundation of China (no. 31922012), and the 2115 Talent Development Program of China Agricultural University.

Authors' contributions

Conceptualization & supervision: W.C. and H.L.

Data analysis: Y.D., L.M. and C.Z.

Writing – original draft: Y.D., L.M. and C.Z.

Writing – review & editing: L.M., C.Z., S.X., Y.X., F.S., T.L, W.C., H. L. and Y.D.

Availability of data and materials

The genome accession IDs of the 375 arthropod species were provided as **Supplementary Table S1**. All selected species have both reference genome sequence and genome annotation file in NCBI. No additional filters were performed. Our collection included the five major insect orders (in phylogenetic order) Lepidoptera (74 genomes), Diptera (113 genomes), Coleoptera (38 genomes), Hymenoptera (104 genomes), Hemiptera (32 species) and the outgroups (14 species) of the five orders. The outgroups included all the Pterygota species of incomplete metamorphosis and the available non-insect arthropod Ixodida (**Supplementary Table S1**). The Adar protein sequence of *Drosophila melanogaster* was downloaded from FlyBase (https://flybase.org/) version dm6.04 with protein ID FBpp0308381 (corresponding transcript ID: FBtr0339272, gene ID: FBgn0026086).

ORCID

Yuange Duan http://orcid.org/0000-0003-2311-9859

References

[1] Zhang P, Zhu Y, Guo Q, et al. On the origin and evolution of RNA editing in metazoans. Cell Rep. 2023;42(2):112112. doi: 10.1016/j.celrep.2023.112112

[2] Bian Z, Ni Y, Xu JR, et al. A-to-I mRNA editing in fungi: occurrence, function, and evolution. Cell Mol Life Sci. 2019;76(2):329–340. doi: 10.1007/s00018-018-2936-3

[3] Liao W, Nie W, Ahmad I, et al. The occurrence, characteristics, and adaptation of A-to-I RNA editing in bacteria: a review. Front Microbiol. 2023;14:1143929. doi: 10.3389/fmicb.2023.1143929

[4] Bazak L, Haviv A, Barak M, et al. A-to-I RNA editing occurs at over a hundred million genomic sites, located in a majority of human genes. Genome Res. 2014;24(3):365–376. doi: 10.1101/gr.164749.113

[5] Porath HT, Knisbacher BA, Eisenberg E, et al. Massive A-to-I RNA editing is common across the metazoa and correlates with dsRNA abundance. Genome Biol. 2017;18(1):185. doi: 10.1186/s13059-017-1315-y

[6] Bass BL. RNA editing by adenosine deaminases that act on RNA. Annu Rev Biochem. 2002;71(1):817–846. doi: 10.1146/annurev.biochem.71.110601.135501

[7] Sapiro AL, Shmueli A, Henry GL, et al. Illuminating spatial A-to-I RNA editing signatures within the Drosophila brain. P Natl Acad Sci USA. 2019;116(6):2318–2327. doi: 10.1073/pnas.1811768116

[8] Gommans WM, Mullen SP, Maas S. RNA editing: a driving force for adaptive evolution? BioEssays. 2009;31(10):1137–1145. doi: 10.1002/bies.200900045

[9] Shoshan Y, Liscovitch-Brauer N, Rosenthal JJC, et al. Adaptive proteome diversification by nonsynonymous A-to-I RNA editing in coleoid cephalopods. Mol Biol Evol. 2021;38(9):3775–3788. doi: 10.1093/molbev/msab154

[10] Yablonovitch AL, Fu J, Li K, et al. Regulation of gene expression and RNA editing in Drosophila adapting to divergent microclimates. Nat Commun. 2017;8(1):1570. doi: 10.1038/s41467-017-01658-2

[11] Li Q, Wang Z, Lian J, et al. Caste-specific RNA editomes in the leaf-cutting ant Acromyrmex echinatior. Nat Commun. 2014;5(1):4943. doi: 10.1038/ncomms5943

[12] Porath HT, Hazan E, Shpigler H, et al. RNA editing is abundant and correlates with task performance in a social bumblebee. Nat Commun. 2019;10(1):10. doi: 10.1038/s41467-019-09543-w

[13] Jiang D, Zhang J. The preponderance of nonsynonymous A-to-I RNA editing in coleoids is nonadaptive. Nat Commun. 2019;10(1):5411. doi: 10.1038/s41467-019-13275-2

[14] Duan Y, Cai W, Li H. Chloroplast C-to-U RNA editing in vascular plants is adaptive due to its restorative effect: testing the restorative hypothesis. RNA. 2023;29(2):141–152. doi: 10.1261/rna.079450.122

[15] Zhang R, Deng P, Jacobson D, et al. Evolutionary analysis reveals regulatory and functional landscape of coding and non-coding RNA editing. PLoS Genet. 2017;13(2):e1006563. doi: 10.1371/journal.pgen.1006563

[16] Buchumenski I, Bartok O, Ashwal-Fluss R, et al. Dynamic hyper-editing underlies temperature adaptation in Drosophila. PLoS Genet. 2017;13(7):e1006931. doi: 10.1371/journal.pgen.1006931

[17] Savva YA, Rieder LE, Reenan RA. The ADAR protein family. Genome Biol. 2012;13(12):252. doi: 10.1186/gb-2012-13-12-252

[18] Keegan LP, McGurk L, Palavicini JP, et al. O'Connell MA: Functional conservation in human and Drosophila of metazoan ADAR2 involved in RNA editing: loss of ADAR1 in insects. Nucleic Acids Res. 2011;39(16):7249–7262. doi: 10.1093/nar/gkr423

[19] Savva YA, Jepson JE, Sahin A, et al. Auto-regulatory RNA editing fine-tunes mRNA re-coding and complex behaviour in Drosophila. Nat Commun. 2012;3(1):790. doi: 10.1038/ncomms1789

[20] Schuh RT, Weirauch C. True bugs of the world (Hemiptera: Heteroptera): classification and natural history. second ed. Rochdale: UK: Siri Scientific Press; 2020.

[21] Ye F, Kment P, Redei D, et al. Diversification of the phytophagous lineages of true bugs (Insecta: Hemiptera: Heteroptera) shortly after that of the flowering plants. Cladistics. 2022;38(4):403–428. doi: 10.1111/cla.12501

[22] Weirauch C, Schuh RT, Cassis G, et al. Revisiting habitat and lifestyle transitions in Heteroptera (Insecta: Hemiptera): insights from a combined morphological and molecular phylogeny. Cladistics. 2019;35(1):67–105. doi: 10.1111/cla.12233

[23] Li H, Leavengood JM Jr., Chapman EG, et al. Mitochondrial phylogenomics of Hemiptera reveals adaptive innovations driving the diversification of true bugs. Proc Biol Sci. 2017;284(1862):20171223. doi: 10.1098/rspb.2017.1223

[24] Francisco-Velilla R, Embarc-Buh A, Rangel-Guerrero S, et al. RNA-protein coevolution study of Gemin5 uncovers the role of the PXSS motif of RBS1 domain for RNA binding. RNA Biol. 2020;17(9):1331–1341. doi: 10.1080/15476286.2020.1762054

[25] Mallik S, Basu S, Hait S, et al. Translational regulation of ribosomal protein S15 drives characteristic patterns of protein-mRNA epistasis. Proteins. 2018;86(8):827–832. doi: 10.1002/prot.25518

[26] Xin K, Zhang Y, Fan L, et al. Liu H: Experimental evidence for the functional importance and adaptive advantage of A-to-I RNA editing in fungi. Proc Natl Acad Sci U S A. 2023;120(12):e2219029120. doi: 10.1073/pnas.2219029120

[27] Duan Y, Li H, Cai W. Adaptation of A-to-I RNA editing in bacteria, fungi, and animals. Front Microbiol. 2023;14:1204080. doi: 10.3389/fmicb.2023.1204080

[28] Camacho C, Coulouris G, Avagyan V, et al. BLAST+: architecture and applications. BMC Bioinf. 2009;10(1):421. doi: 10.1186/1471-2105-10-421

[29] Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30(4):772–780. doi: 10.1093/molbev/mst010

[30] Rice P, Longden I, Bleasby A. EMBOSS: the European molecular biology open software suite. Trends Genet. 2000;16(6):276–277. doi: 10.1016/S0168-9525(00)02024-2

[31] Simao FA, Waterhouse RM, Loannidis P, et al. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210–3212. doi: 10.1093/bioinformatics/btv351

[32] Kück P, Longo GC. FASconCAT-G: extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. Front Zool. 2014;11(1):81. doi: 10.1186/s12983-014-0081-x

[33] Kück P, Struck TH. BaCoCa–a heuristic software tool for the parallel assessment of sequence biases in hundreds of gene and taxon partitions. Mol Phylogen Evol. 2014;70:94–98. doi: 10.1016/j.ympev.2013.09.011

[34] Nguyen L-T, Schmidt HA, Von Haeseler A, et al. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32(1):268–274. doi: 10.1093/molbev/msu300