Trade-off between transcriptome plasticity and genome evolution in cephalopods

Noa Liscovitch-Brauer[1], Shahar Alon[2], Hagit T. Porath[3], Boaz Elstein[3], Ron Unger[3], Tamar Ziv[4], Arie Admon[4], Erez Y. Levanon[3], Joshua J. C. Rosenthal[5,6*†] and Eli Eisenberg[1,7*]

[1]Raymond and Beverly Sackler School of Physics and Astronomy, Tel Aviv University, Tel Aviv 69978, Israel.

[2]Media Lab and McGovern Institute, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA.

[3]Mina and Everard Goodman Faculty of Life Sciences, Bar-Ilan University, Ramat Gan 52900, Israel.

[4]Smoler Proteomics Center and Faculty of Biology, Technion - Israel Institute of Technology, Haifa 32000, Israel.

[5]Eugene Bell Center, Marine Biological Laboratory, Woods Hole, MA 02543 USA.

[6]Institute of Neurobiology, University of Puerto Rico Medical Sciences Campus, San Juan, 00901 Puerto Rico.

[7]Sagol School of Neuroscience, Tel Aviv University, Tel Aviv 69978, Israel.

*Correspondence to: jrosenthal@mbl.edu (JJCR) or elieis@post.tau.ac.il (EE).

† Lead Contact

Summary:

RNA editing, a post-transcriptional process, allows the diversification of proteomes beyond the genomic blueprint; however it is infrequently used among animals. Recent reports suggesting increased levels of RNA editing in squids thus raise the question of their nature and effects in these organisms. We here show that RNA editing is particularly common in behaviorally sophisticated coleoid cephalopods, with tens of thousands of evolutionarily conserved sites. Editing is enriched in the nervous system affecting molecules pertinent for excitability and neuronal morphology. The genomic sequence flanking editing sites is highly conserved, suggesting that the process confers a selective advantage. Due to the large number of sites, the surrounding conservation greatly reduces the number of mutations and genomic polymorphisms in protein coding regions. This trade-off between genome evolution and transcriptome plasticity highlights the importance of RNA recoding as a strategy for diversifying proteins, particularly those associated with neural function.

Keywords: Epitranscriptome, RNA modifications, RNA editing, ADAR, Neural plasticity, Cephalopods, Genome evolution, Proteome diversity

Introduction:

It is generally assumed that genetic information passes faithfully from DNA to RNA to proteins. Proteome complexity, however, depends on a diverse set of post-transcriptional processes that modify and enrich genetic information beyond the genomic blueprint. RNA editing is one such process. Adenosine deamination to inosine by the ADAR family of enzymes is the most common form of editing among animals (Bass, 2002; Nishikura, 2015). Because inosine is recognized as guanosine during translation (Basilio et al., 1962), this process has the capacity to recode codons and fine-tune protein function. However, it seldom does so. Transcriptome-wide screens have revealed that only ~3% of human messages and 1-4% of those from Drosophila harbor a recoding site (Ramaswami and Li, 2014; St Laurent et al., 2013; Xu and Zhang, 2014; Yu et al., 2016). Even more surprising is the limited extent to which this process is conserved. There are only about 25 human transcripts that contain a recoding site that is conserved across mammals (Pinto et al., 2014), and only about 65 recoding sites conserved across the Drosophila lineage (Yu et al., 2016). In *C. elegans*, only a few putative recoding sites have been identified, some of which were not validated (Goldstein et al., 2016; Washburn et al., 2014; Whipple et al., 2015; Zhao et al., 2015). These data support the hypothesis that recoding by RNA editing is mostly neutral or detrimental, and only rarely adaptive (Xu and Zhang, 2014).

Recently, we reported an apparent exception: squid contain an unusually high level of recoding, with the majority of mRNAs in the nervous system harboring at least one event (Alon et al., 2015). This intriguing but anecdotal result raised fundamental questions about the nature of recoding in these organisms. Does the massive RNA-level recoding translate into proteome diversification? Is it simply a neutral byproduct of a promiscuous ADAR tasked with another

function, or adaptive, providing a functional advantage? Finally, is it related to behavioral sophistication?

Cephalopods are diverse, and can be divided into the behaviorally complex coleoids, consisting of squid, cuttlefish and octopus, and the more primitive nautiloids. In this report we show that in neural transcriptomes extensive A-to-I RNA editing is observed in the behaviorally complex coleoid cephalopods, but not in nautilus. The edited transcripts are translated into protein isoforms with modified functional properties. By comparing editing across coleoid taxa, we found that, unlike the case for mammals, many sites are highly conserved across the lineage, and undergo positive selection, resulting in a sizable slow-down of their genome evolution.

Results:

Extensive recoding is an invention of coleoid cephalopods

To assess the level of recoding via A-to-I RNA editing in cephalopods, we analyzed matching DNA and RNA samples of individual animals from species that span the cephalopod evolutionary tree. We studied four members of the coleoid cephalopod subclass (soft bodied cephalopods): two octopuses (*Octopus vulgaris* and *Octopus bimaculoides)*, a squid (*Doryteuthis pealeii*), and a cuttlefish (*Sepia oficianalis)*, as well as a nautiloid (*Nautilus pompilius*) and a gastropod mollusk (*Aplysia californica)* as an evolutionary outgroup. Cephalopods emerged in the late Cambrian period, roughly at ~530mya, and the divergence of nautiloids from coleoides is estimated to have occurred at 350-480mya (Kröger et al., 2011). The coleoides diverged to Vampyropoda (octopus lineage) and the Decabrachia (squid and cuttlefish lineage) at ~200-350mya (Albertin et al., 2015; Kröger et al., 2011). Divergence of squid from Sepiida is estimated to have occurred at 120-

220mya (Checa et al., 2015). The two octopus species used in this study, *Octopus vulgaris* and *Octopus bimaculoides*, have been shown to be closely related using mitochondrial DNA, and are in some cases even indistinguishable, depending on the geographical origins of the specimens (Söller et al., 2000). The divergence time between the gastropod species *Aplysia californica* and cephalopods is estimated to be 520-610mya (Kröger et al., 2011). A general representation of the phylogenetic relations between the species is shown at **Fig. 1a**.

A full genome sequence is not available for the cephalopod species used in this study (except for *Octopus bimaculoides*; see below). Thus, we used a genome-independent method (Alon et al., 2015) to detect editing sites, one that focuses specifically on the coding regions of the transcriptome. Briefly, RNA-Seq data **(**174-366 million reads per species; **Table S1, also see** (Alon et al., 2015)**)** was utilized to assemble a de-novo transcriptome (Grabherr et al., 2013) and the coding sequences were identified by comparison with Swiss-Prot (Bairoch et al., 2005) open reading frames (**Table S2**). RNA and DNA reads were then aligned to the assembled transcriptome (using Bowtie2 (Langmead and Salzberg, 2012) with local alignment configuration and default parameters). To detect editing events, we looked for systematic mismatches between RNA and DNA reads within the coding part of the transcriptome, filtering out those that stem from sequencing errors or genomic polymorphisms (see **Methods** for more details). The A-to-G DNA-to-RNA mismatches that are identified by this process could result from A-to-I RNA editing, while other types of mismatches provide an estimate of our false-detection rate.

For sepia, squid, and the two octopus species, most mismatches (>80%) detected by the above approach were A-to-G mismatches, and the noise level, estimated by the number of G-to-A

mismatches, is rather low - 2-3% (**Fig. 1B**). Furthermore, the residues surrounding the detected A-to-G sites exhibit a sequence pattern consistent with the known preferences for ADARs (Eggington et al., 2011; Kleinberger and Eisenberg, 2010) (**Fig. 1C**). We thus attribute these mismatches to A-to-I RNA editing events, and obtain 80-130 thousand editing sites in protein-coding regions (**Tables S3-S4**). Remarkably, results from nautilus and aplysia are in sharp contrast. First, we found only 1150 and 933 A-to-G mismatches for these species, much less than for the octopus, squid and sepia. Moreover, there is no excess of A-to-G mismatches over other events (**Fig. 1B, Tables S3-S4**), and the residues surrounding the detected A-to-G sites do not exhibit any sequence preference (**Fig. 1C**). Thus, the A-to-G mismatches found in nautilus and aplysia are likely to be (mostly) noise, with very few, if any, editing sites. Accordingly, editing within the coding sequence of these species is orders of magnitudes lower than for the octopus, squid and sepia. These data suggest that extensive recoding through RNA editing evolved along the coleoid lineage. As all of the cephalopod groups that separate coleoids and nautiloids are now extinct (e.g. belemnites and ammonoids), it will be difficult to pinpoint a more exact time for the emergence of extensive RNA editing.

Proteomic validation of recoding sites

Sanger-sequencing validation of the sites detected by the present scheme were previously reported (Alon et al., 2015). Here, we employed mass spectrometry analysis to further test whether the multitude of novel RNA isoforms created by extensive RNA editing are translated into proteins, resulting in extensive proteome diversification by recoding. We analyzed squid giant axon and stellate ganglion samples, looking for peptides translated from RNA that include editing sites. To simplify the analysis, we considered only peptides that include a single non-

synonymous (recoding) editing site, and checked whether the edited, non-edited or both versions of the peptide were observed. For squid stellate ganglion, a total of 74,146 unique peptides were detected, 4,115 of which harbor 5,617 recoding sites, and 3,204 peptides that include a single predicted site. Of these, 320 sites (10.0%) were shown to be edited (174 cases where both the pre-edited and edited versions are observed, and 146 found only in the edited version), including most of the sites predicted to be edited at high levels. Similarly, for squid giant axon 58,403 unique peptides were detected, 3,579 of which harbor 4,956 predicted recoding sites, and 2,741 peptides included a single predicted site. Of these, 283 sites (10.3%) were shown to be edited (160 cases where both pre-edited and edited versions are observed, and 123 found only in the edited version). Altogether, this experiment validated 432 protein recoding sites. The fraction of sites validated correlated very well with the editing level predicted from RNA-seq data (**Fig. 2)**.

Note that the shotgun proteomics method used here provides only partial coverage of the tryptic peptides generated by the proteolysis (Michalski et al., 2011). This is demonstrated by the fact that ~90% of the recoded amino acids are completely missing from our data, regardless of their editing state. Accordingly, lack of peptide evidence for an edited or unedited form of a given site cannot be considered as evidence for this isoform not being present. However, it is possible that some of the editing sites are not translated, or do not produce a stable protein, possibly due to deleterious effect of editing on the protein structure.

Protein recoding accounts for a sizable fraction of ADARs editing in neural tissues of *Octopus bimaculoides*

For most organisms, A-to-I editing is markedly depleted from the protein coding regions of the transcriptome. The question then arises whether the extensive recoding in cephalopods is accompanied by extra-ordinary editing of the non-coding transcriptome. Recently, a genome was published for *Octopous bimaculoides*, the first from a cephalopod (Albertin et al., 2015), allowing us to use genome-dependent methods (Picardi and Pesole, 2013; Ramaswami et al., 2012) to study the full editome, including editing in non-coding sequences, as well as a comparison with the genome-free method for the coding regions. Analyzing RNA-seq data of the same four neural tissues studied in the transcriptome-based approach resulted in 800,941 editing sites, 105,380 of them in annotated coding sequences (compared to 76,862 sites in coding sequence identified by the genome-free pipeline), 49,483 of these were also found using the transcriptome-based genome-free approach (**Fig. 3A**). Differences between the two methods are due to the different de-novo transcriptomes used, and the different methods employed to filter out random mismatches (see **Methods**). These results suggest that the genome-free method provides a reasonable coverage of the editing signal in coding sequence, and that the number of editing sites outside the coding region is likely to be an order of magnitude higher than the number within the coding sequence for the other cephalopods studied here.

Analyzing RNA from 12 different tissues, including non-neural ones, we found 903,742 editing sites in the transcriptome (**Table S5)**, 12% of which reside in coding regions (**Fig. 3B-C**). In mammals, editing mostly occurs within genomic repeats (Bazak et al., 2014a; Levanon et al., 2004; Neeman et al., 2006). In primates specifically, most RNA editing sites are found in Alu

repeats, whose sequence facilitates the creation of a double-stranded RNA structure that promotes ADAR binding. Similarly, editing in *Octopus bimaculoides* is enriched in repeats regions (303,414/903,742 sites, 34%; 159,005 of them in annotated repeats). The "editing index", a robust measure of editing activity (Bazak et al., 2014b) defined as the editing level averaged over all adenosines (edited and unedited) weighted by expression level, is calculated to be 0.21% in octopus repeats for the panel of 12 tissues studied, which is comparable to the index observed in human Alu repeats (Bazak et al., 2014a). Unlike primates, though, there is not one specific repeat family which was found to contain the majority of sites, and SINEs are not edited more than other repeats (**Fig. 3D)**. Therefore, as the repeat editing index in octopus is calculated over all repeats (~1.3Gbp), and editing in repeats accounts for only 21-38% of all editing events in octopus mRNAs (compared to >95% in primates), overall the number of editing events reflected in mRNA sequencing data is roughly an order of magnitude higher in *Octopus bimaculoides* compared to primates. Furthermore, in neural tissues ~11-13% of these events result in amino-acid modification, compared with <1% in mammals (Bazak et al., 2014a). RNA editing is known to be important in neural function (Rosenthal and Seeburg, 2012) and abnormal editing patterns or ADAR function have been shown to underlie several neural conditions (Slotkin and Nishikura, 2013). Indeed, we find that editing in non-neural tissues of *Octopous bimaculoides* is roughly two-fold lower, and recoding events are even more strongly suppressed (**Fig. 3B**). Consistently, GO analysis of edited transcripts shows enrichment of neuronal and cytoskeleton functions in all four species (**Table S6**).

An intriguing result from the recently reported *Octopus bimaculoides* genome was that the protocadherin gene family was greatly expanded (Albertin et al., 2015). In the mammalian brain

these proteins are important for mediating combinatorial complexity in neuronal connections and play a role in diversifying neural circuitry (Chen and Maniatis, 2013). We found a large number of protocadherins in the assembled transcriptomes for the four coleoid species (127-251 ORFs), but not in nautilus (28 ORFs) (**Fig. 3E)**. Interestingly, protocadherins are significantly enriched in editing sites and are edited at higher levels in all four coleoid species, but not in nautilus (**Fig. 3F-G)**.

Signs for positive selection of recoding sites in coleoid cephalopods

Mammalian editing events in the coding region (and the editing levels) are negatively correlated with the importance of a site or gene - essential genes, and genes under strong functional constraints, tend to harbor lower numbers of editing sites and exhibit lower editing levels (Xu and Zhang, 2014). Furthermore, nonsynonymous editing sites are suppressed, compared with synonymous ones, and the fraction of editing sites that are conserved across mammals is minute. These and other observations have led to the conclusion that while a few mammalian recoding sites are clearly beneficial, overall recoding by RNA editing is nonadaptive in mammals, presumably resulting from tolerable promiscuous targeting by the ADAR enzymes (Xu and Zhang, 2014).

To conduct a comparable analysis of the recoding repertoire in cephalopods, we first identified the non-synonymous editing sites. About 65% of cephalopod edits within coding sequences are nonsynonymous, leading to 54,287-86,230 recoding sites in 6,688-8,537 ORFs (**Table S7**), orders of magnitude more recoding than any other species. In sharp contrast with mammals, thousands of recoding sites are shared between species (**Fig. 4A-B**). As expected, the fraction of

conserved sites is higher for species that are evolutionary closer (**Fig. 4C)**, but unlike the picture observed in other evolutionary lineages (**Fig. 4D)**, editing in coding sequences is, to a large extent, conserved. Interestingly, 1146 editing sites (in 443 proteins) are conserved and shared by all four coleoid cephalopod species **(Fig. 4E)**. A large fraction of proteins are recoded at multiple sites, and many proteins harbor multiple conserved and highly-edited (>10% in at least one species) recoding sites (**Fig. 4F, Table S8**). Notably, even the editing levels in the shared sites are conserved, and exhibit significant and sizable correlations between evolutionarily distant species **(Fig. 4G)**.

Overall, the nonsynonymous to synonymous (N/S) ratio for cephalopod edits is 65/35=1.9, as expected under neutrality taking into account the ADAR target motif (Alon et al., 2015). However, the N/S ratios increases to much higher values as editing levels increase (**Fig. 5A**), signaling positive selection for the highly edited sites. Conserved sites show an even stronger pattern (**Fig. 5B)**, where almost all highly edited, conserved, sites are nonsynonymous. Consistently, and in stark contrast with mammals, the higher the editing levels, the more sites are conserved (**Fig. 5C-D**). Furthermore, editing is over-represented in highly conserved regions of the transcriptome (>95% identity between species) (**Fig. 5E**). Taken together, these results suggest that recoding by RNA editing is commonly adaptive in coleoid cephalopods, with many thousands of recoding sites under positive selection.

Functional implications of recoding sites

We next tested whether species-specific and conserved recoding events can affect protein function. We studied sepia, squid and *Octopus vulgaris* K_v2 potassium channel orthologs, whose

messages are abundantly edited (34-55 sites per species; five sites shared between all species; **Fig. S2 and Table S9**). Voltage-dependent potassium channels of the K_v2 subfamilly, also known as "delayed rectifiers", are expressed across the metazoa. In the mammalian central nervous system, they regulate excitability, action potential duration, and repetitive firing (Murakoshi and Trimmer, 1999). As with most voltage-dependent potassium channels, they are predominantly closed at negative membrane potentials and open at positive ones. When switched between negative and positive potentials, they open or close with characteristic rates. At positive potentials, channels will also spontaneously close after opening, a process known as "inactivation". The kinetics of these three processes play a vital role in determining how the channels regulate electrical signaling.

To measure the effects of editing on functional properties, we expressed all channels in *Xenopus* oocytes and studied them using the Cut-Open Oocyte Vaseline Gap Voltage Clamp technique (Taglialatela et al., 1992). The unedited versions of the channels open over a similar range of voltages, but have different opening, closing and inactivation kinetics (**Fig. S3** and **Fig. 6B**). To examine the effects of editing, we first looked at the sepia-specific editing site I529V. **Fig. 6Ai** shows superimposed current traces, obtained in response to a voltage step from -80mV to +40mV, for the unedited and edited (I529V) versions of sepia K_v2. Clearly, the edited channel inactivates more quickly, at all voltages tested (**Fig. 6Aii**; there is also a very modest slowing of channel closure upon bring the voltage back to -80mV). Editing had no effect on voltage sensitivity and channel opening (data not shown). We next looked at a common editing site (squid I579V, Sepia I630V and Octopus I632V) and found that it predominantly affects the channels' closing rates. Interestingly, the direction of the effect is species-dependent. First we

analyzed the tail currents in the unedited squid, sepia and octopus channels by recording currents at a negative membrane voltage of -80mV following a brief activating pulse to a positive potential, (**Fig. 6Bi**). Each closes at distinct, species-specific, rates, with squid the fastest and octopus the slowest. However, upon introduction of the common editing event, the channels converge on a similar rate (**Fig. 6Bii**); in squid, editing this site slows closing, while in octopus and sepia, it speeds it. This effect on closing kinetics was consistent at all voltages tested (**Fig. 6Biii-iv**). Based on these data from K_v2 orthologs, and the fact that editing is exceptionally abundant in ion channels and proteins involved in synaptic vesicle release and recycling, the overall influence of RNA editing on neurophysiology is likely profound and complex.

Positive selection of editing events slows down genome evolution

To edit a specific adenosine within an RNA, ADAR enzymes require surrounding dsRNA structures. These structures are often large, spanning hundreds of nucleotides (Morse et al., 2002). If editing is under positive selection, maintaining these structures would require elevated sequence conservation in the vicinity of editing sites (Herb et al., 1996; Higuchi et al., 1993). As this sequence conservation stems from constraints related to RNA structure, rather than its coding capacity, it should affect synonymous and non-synonymous changes equally. Indeed, we see a marked depletion of inter-species mutations (**Figs. 7A and S4a-f**) and intra-species genomic polymorphisms (**Figs. 7B and S4g**), synonymous and non-synonymous alike, up to ~100nt on each side of a recoding site. These regions show an elevated GC content (**Figs. 7C and S4h**), consistent with the requirement for the formation of stronger secondary structures.

The cumulative effect of this evolutionary constraint is considerable. Due to the large number of recoding sites and the extended range of the associated genomic rigidity, the local constraints observed in the vicinity of the recoding sites translate into a substantial global effect on genome evolution. These 200nt windows around recoding sites cover a sizable fraction of all protein coding sequences: 23-41%, depending on the coleoid species.

One may also quantify the effect of purifying selection in these regions by studying the fraction of inter-species mutations that were avoided, presumably due to maintaining the dsRNA structures required for editing. We analyzed the inter-species mutation rates (in orthologous parts of the respective transcriptomes) as a function of the distance to the closest conserved recoding site, and found again that the rates are considerably lowered in the vicinity of editing sites, compared with the baseline rate observed far from any editing site (**Fig. S5A)**. Attributing the difference between the observed mutations rate and the baseline to effects of editing on genome evolution, and integrating this difference over the entire transcriptome, we estimate that 3-15% of all transcriptomic inter-species mutations are purified (numbers vary for the for the various species-pairs), apparently due to constraints imposed by editing. Similarly, we find that the actual number of SNPs in cephalopod coding sequences is 10-26% lower than what would be seen in the absence of SNPs suppression in the vicinity of recoding sites **(Fig. S5B)**. Thus, the purifying selection against inter-species mutations and intra-species genomic polymorphisms residing in proximity to recoding sites results in a sizable reduction in the global number of mutations and polymorphisms in these species, revealing an unanticipated genome rigidity required to maintain the extensive transcriptome recoding.

Discussion:

Seminal studies on RNA editing focused on recoding events and their functional outcomes (Burns et al., 1997; Higuchi et al., 1993). Later, with the advent of deep sequencing technologies and accompanying computational advances, transcriptome-wide screens showed that recoding is extremely rare. For example, there are millions of editing sites in the human transcriptome, but almost all of these reside in untranslated regions (Bazak et al., 2014a). This distribution implies some fundamental principles about RNA editing by ADARs. First, there is an active mechanism for excluding editing sites from coding regions; otherwise they would be far more common. Second, although there are clear exceptions for individual editing sites, the overall purpose of editing is not to recode (Liddicoat et al., 2015; Mannion et al., 2014). This point is reinforced by the fact that most mammalian recoding sites are neutral at best (Xu and Zhang, 2014). The abundant recoding in coleoids reported here runs contrary to these ideas.

We presented evidence that high-level recoding was invented by coleoids, or an extinct ancestor, after the divergence of the nautiloids. It is plausible that protein recoding may not be the primary function of editing in cephalopods. Perhaps there are other purposes for robust ADAR activity, such as its potential use in innate immunity (Liddicoat et al., 2015; Mannion et al., 2014). As with any mutation, promiscuous ”off-target” edits would sometimes be advantageous and therefore selected. However, many other organisms, such as humans, edit abundantly, producing multiple promiscuous edits. What is unique about coleoid cephalopods is that they appear not to exclude editing from protein coding regions, leading to many thousands of recoding sites being recruited and conserved across distant species. Regardless of the primary motivation for editing, this unique phenomenon clearly has an enormous effect on the proteome.

The extensive recoding activity in cephalopods might suggest that there are underlying mechanistic novelties in their editing process, compared with other organisms. For example, cephalopod ADARs may have evolved to increase their catalytic activity or decrease their specificity. Previous studies have shown that squid express a splice variant of ADAR2 with an extra dsRNA binding domain and this feature increases its affinity for dsRNA, leading to a higher activity (Palavicini et al., 2009, 2012). Although cephalopods do express ADAR1 orthologs (Albertin et al., 2015; Alon et al., 2015), no functional studies have been conducted on them, nor on any invertebrate ADAR1 for that matter. They too may possess unique activities. Finally, one might expect the introduction of thousands of editing sites to be accompanied by undesirable side effects. For example, messages which contain so many mutations might often translate into dysfunctional, or even toxic, proteins. To accommodate this burden, cephalopods may have evolved unique mechanisms for protein folding and quality control. These ideas require further study.

Recoding in coleoid cephalopods is something of an enigma. Unlike the case for mammals, inter-species conservation and the higher than expected frequencies of non-synonymous changes suggest that a sizable fraction of events were recruited during the course of cephalopod evolution. Why would the coleoids choose to alter genetic information within RNA rather than hardwire the change in DNA? There are several potential advantages to making changes within RNA. First of all, the changes are transient. Thus an organism can choose to turn them on or off, providing phenotypic flexibility, a quality that is particularly useful for environmental acclimation (Garrett and Rosenthal, 2012; Rieder et al., 2015). In addition, RNA-level changes can better augment

genetic diversity. With DNA, an organism is limited to two alleles. With RNA, all messages need not be edited, and thus the pool of mRNAs can include edited or unedited versions at given sites. When a message contains more than one site, complexity can increase exponentially. Future proteomic experiments will be necessary to determine whether the combinatorial complexity is realized in neural proteins, and whether editing contributes to neuron-specific diversity, or the ability of the nervous system to respond to environmental cues. If the thousands of editing sites do indeed lead to independent functional outcomes, then the regulation of the editing process would be necessarily complex.

Among invertebrates, the nervous system of coleoids is uniquely large and complex. For example, with half a billion neurons, *Octopus vulgaris* has ~5 times the number of a mouse (Young, 1971). Coleoids have brain lobes dedicated to learning and memory (Hochner et al., 2003; Shomrat et al., 2008, 2015; Young, 1961), and exhibit a range of complex and plastic behaviors. Nautiloid brains are simpler, containing fewer neurons, and lack specific lobes dedicated to learning and memory (Young, 1965). The association of massive recoding with the nervous system, and the fact that it is unique to the coleoids and not observed in nautilus, hint at its relationship with the exceptional behavioral sophistication of the coleiods. This idea is reinforced by the high density of editing in transcripts that encode proteins directly involved in excitability.

What is most surprising about cephalopod recoding is its effect on genome evolution. From a mechanistic standpoint, this makes sense. In order to edit a specific adenosine, ADAR requires surrounding RNA structures. Even single nucleotide substitutions within these structures can

abolish editing (Reenan, 2005; Rieder et al., 2013). If an editing site is advantageous, the structure must be preserved. Abundant editing requires abundant structures that can span a large fraction of the genomic coding sequence. Thus, while extensive recoding presents the species with a route towards proteome complexity, it comes with its own price-tag. The constraints required to preserve thousands of recoding sites reduce the accumulation of mutations at positions in the proximity of an editing site, slowing down the rate of conventional, DNA-level, evolution. The nervous system is one of the most important targets for natural selection, as subtle changes can lead to behavioral advances. For coleoid cephalopods, the need to make specific A-to-I changes within the neural transcriptome is sufficiently important to forego standard pathways of neuronal evolution.

Author contributions:

JJCR and EE conceived the study and designed the experiments, with the help of EYL and NLB. NLB analyzed the data. SA developed some of the scripts used for analysis. HTP did the hyper-editing analysis. JJCR performed all experiments, except for the MS, done by TZ and AA. BE and RU analyzed the MS data. EE supervised the bioinformatics work, with input from EYL and JJCR. NLB, JJCR and EE wrote the paper.

Acknowledgments:

The authors thank Dr. Michal Barak for assisting the analyses, Tamar A. Carmeli for the artwork in Fig. 1a, the Smoler Proteomics Center at the Technion for using its facilities, Dr. Roger Hanlon for providing specimens of Sepia oficianalis, and Dr. Robyn Crook for help in obtaining a sample of Nautilus pompilius. NLB was supported by a post-doctoral scholarship from the Center for Nanoscience and Nanotechnology, Tel-Aviv University. The research of RU is supported by the Israel Science Foundation (772/13). The research of EYL was supported by the European Research Council (311257) and the Israel Science Foundation (1380/14). The research of JJCR was supported by the National Institutes of Health [1R0111223855, 1R01NS64259], the National Science Foundation (HRD-1137725), and the Frank R. Lillie and Laura and Arthur Colwin Research Fellowships from the Marine Biological Laboratory in Woods Hole. The work of JJCR and EE was supported by grant No 094/2013 from the United States-Israel Binational Science Foundation (BSF).

The authors declare no conflict of interest.

References

Albertin, C.B., Simakov, O., Mitros, T., Wang, Z.Y., Pungor, J.R., Edsinger-Gonzales, E., Brenner, S., Ragsdale, C.W., and Rokhsar, D.S. (2015). The octopus genome and the evolution of cephalopod neural and morphological novelties. Nature *524*, 220–224.

Alon, S., Garrett, S.C., Levanon, E.Y., Olson, S., Graveley, B.R., Rosenthal, J.J.C., and Eisenberg, E. (2015). The majority of transcripts in the squid nervous system are extensively recoded by A-to-I RNA editing. Elife *2015*.

Bairoch, A., Apweiler, R., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., et al. (2005). The Universal Protein Resource (UniProt). Nucleic Acids Res. *33*.

Basilio, C., Wahba, A.J., Lengyel, P., Speyer, J.F., and Ochoa, S. (1962). Synthetic polynucleotides and the amino acid code, V. Proc. Natl Acad. Sci. USA 613–616.

Bass, B.L. (2002). RNA editing by adenosine deaminases that act on RNA. Annu. Rev. Biochem. *71*, 817–846.

Bazak, L., Haviv, A., Barak, M., Jacob-Hirsch, J., Deng, P., Zhang, R., Isaacs, F.J., Rechavi, G., Li, J.B., Eisenberg, E., et al. (2014a). A-to-I RNA editing occurs at over a hundred million genomic sites, located in a majority of human genes. Genome Res. *24*, 365–376.

Bazak, L., Levanon, E.Y., and Eisenberg, E. (2014b). Genome-wide analysis of Alu editability. Nucleic Acids Res. *42*, 6876–6884.

Burns, C.M., Chu, H., Rueter, S.M., Hutchinson, L.K., Canton, H., Sanders-Bush, E., and Emeson, R.B. (1997). Regulation of serotonin-2C receptor G-protein coupling by RNA editing. Nature *387*, 303–308.

Checa, A.G., Cartwright, J.H.E., Sánchez-Almazo, I., Andrade, J.P., and Ruiz-Raya, F. (2015).

The cuttlefish Sepia officinalis (Sepiidae, Cephalopoda) constructs cuttlebone from a liquid-crystal precursor. Sci. Rep. *5*, 1–13.

Chen, W. V, and Maniatis, T. (2013). Clustered protocadherins. Development *140*, 3297–3302.

Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics *10*, 48.

Eggington, J.M., Greene, T., and Bass, B.L. (2011). Predicting sites of ADAR editing in double-stranded RNA. Nat. Commun. *2*, 319.

Garrett, S., and Rosenthal, J.J.C. (2012). RNA Editing Underlies Temperature Adaptation in K+ Channels from Polar Octopuses. Science (80-.). *335*, 848–851.

Goldstein, B., Agranat-Tamir, L., Light, D., Ben-Naim Zgayer, O., Fishman, A., and Lamm, A.T. (2016). A-to-I RNA editing promotes developmental-stage-specific gene and lncRNA expression. Genome Res. gr.211169.116.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2013). Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nat. Biotechnol. *29*, 644–652.

Hedges, S.B., Dudley, J., and Kumar, S. (2006). TimeTree: A public knowledge-base of divergence times among organisms. Bioinformatics *22*, 2971–2972.

Herb, A., Higuchi, M., Sprengel, R., and Seeburg, P.H. (1996). Q/R site editing in kainate receptor GluR5 and GluR6 pre-mRNAs requires distant intronic sequences. Proc. Natl. Acad. Sci. U. S. A. *93*, 1875–1880.

Higuchi, M., Single, F.N., Köhler, M., Sommer, B., Sprengel, R., and Seeburg, P.H. (1993). RNA editing of AMPA receptor subunit GluR-B: A base-paired intron-exon structure determines

position and efficiency. Cell *75*, 1361–1370.

Hochner, B., Brown, E.R., Langella, M., Shomrat, T., and Fiorito, G. (2003). A learning and memory area in the octopus brain manifests a vertebrate-like long-term potentiation. J. Neurophysiol. *90*, 3547–3554.

Ishihama, Y., Rappsilber, J., Andersen, J.S., and Mann, M. (2002). Microcolumns with self-assembled particle frits for proteomics. In Journal of Chromatography A, pp. 233–239.

Ishihama, Y., Rappsilber, J., and Mann, M. (2006). Modular stop and go extraction tips with stacked disks for parallel and multidimensional peptide fractionation in proteomics. J. Proteome Res. *5*, 988–994.

Kleinberger, Y., and Eisenberg, E. (2010). Large-scale analysis of structural, sequence and thermodynamic characteristics of A-to-I RNA editing sites in human Alu repeats. BMC Genomics *11*, 453.

Kröger, B., Vinther, J., and Fuchs, D. (2011). Cephalopod origin and evolution: A congruent picture emerging from fossils, development and molecules. BioEssays *33*, 602–613.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods *9*, 357–359.

Levanon, E.Y., Eisenberg, E., Yelin, R., Nemzer, S., Hallegger, M., Shemesh, R., Fligelman, Z.Y., Shoshan, A., Pollock, S.R., Sztybel, D., et al. (2004). Systematic identification of abundant A-to-I editing sites in the human transcriptome. Nat. Biotechnol. *22*, 1001–1005.

Li, J.B., Levanon, E.Y., Yoon, J.-K., Aach, J., Xie, B., Leproust, E., Zhang, K., Gao, Y., and Church, G.M. (2009). Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. Science (80-.). *324*, 1210–1213.

Li, L., Stoeckert, C.J., and Roos, D.S. (2003). OrthoMCL: Identification of ortholog groups for

eukaryotic genomes. Genome Res. *13*, 2178–2189.

Liddicoat, B.J., Piskol, R., Chalk, A.M., Ramaswami, G., Higuchi, M., Hartner, J.C., Li, J.B., Seeburg, P.H., Walkley, C.R., Danecek, P., et al. (2015). RNA editing by ADAR1 prevents MDA5 sensing of endogenous dsRNA as nonself. Science *349*, 1115–1120.

Liman, E.R., Tytgat, J., and Hess, P. (1992). Subunit stoichiometry of a mammalian K+ channel determined by construction of multimeric cDNAs. Neuron *9*, 861–871.

Lockery, S.R., and Goodman, M.B. (1998). Ion Channels Part B.

Mannion, N.M., Greenwood, S.M., Young, R., Cox, S., Brindle, J., Read, D., Nellåker, C., Vesely, C., Ponting, C.P., McLaughlin, P.J., et al. (2014). The RNA-Editing Enzyme ADAR1 Controls Innate Immune Responses to RNA. Cell Rep. *9*, 1482–1494.

Michalski, A., Cox, J., and Mann, M. (2011). More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. J. Proteome Res. *10*, 1785–1793.

Morse, D.P., Aruscavage, P.J., and Bass, B.L. (2002). RNA hairpins in noncoding regions of human brain and Caenorhabditis elegans mRNA are edited by adenosine deaminases that act on RNA. Proc. Natl. Acad. Sci. U. S. A. *99*, 7906–7911.

Murakoshi, H., and Trimmer, J.S. (1999). Identification of the Kv2.1 K+ Channel as a Major Component of the Delayed Rectifier K+ Current in Rat Hippocampal Neurons. J. Neurosci. *19*, 1728–1735.

Neeman, Y., Levanon, E.Y., Jantsch, M.F., and Eisenberg, E. (2006). RNA editing level in the mouse is determined by the genomic repeat repertoire. RNA *12*, 1802–1809.

Nishikura, K. (2015). A-to-I editing of coding and non-coding RNAs by ADARs. Nat. Rev. Mol. Cell Biol. *17*, 83–96.

Palavicini, J.P., O'Connell, M. a, and Rosenthal, J.J.C. (2009). An extra double-stranded RNA binding domain confers high activity to a squid RNA editing enzyme. RNA *15*, 1208–1218.

Palavicini, J.P., Correa-Rojas, R.A., and Rosenthal, J.J.C. (2012). Extra double-stranded RNA binding domain (dsRBD) in a squid RNA editing enzyme confers resistance to high salt environment. J. Biol. Chem. *287*, 17754–17764.

Picardi, E., and Pesole, G. (2013). REDItools: High-throughput RNA editing detection made easy. Bioinformatics *29*, 1813–1814.

Pinto, Y., Cohen, H.Y., and Levanon, E.Y. (2014). Mammalian conserved ADAR targets comprise only a small fragment of the human editosome. Genome Biol. *15*, R5.

Porath, H.T., Carmi, S., and Levanon, E.Y. (2014). A genome-wide map of hyper-edited RNA reveals numerous new sites. Nat. Commun. *5*, 4726.

Ramaswami, G., and Li, J.B. (2014). RADAR: a rigorously annotated database of A-to-I RNA editing. Nucleic Acids Res. *42*, D109-13.

Ramaswami, G., Lin, W., Piskol, R., Tan, M.H., Davis, C., and Li, J.B. (2012). Accurate identification of human Alu and non-Alu RNA editing sites. Nat. Methods *9*, 579–581.

Reenan, R.A. (2005). Molecular determinants and guided evolution of species-specific RNA editing. Nature *434*, 409–413.

Rieder, L.E., Staber, C.J., Hoopengardner, B., and Reenan, R.A. (2013). Tertiary structural elements determine the extent and specificity of messenger RNA editing. Nat. Commun. *4*, 2232.

Rieder, L.E., Savva, Y. a, Reyna, M. a, Chang, Y.-J., Dorsky, J.S., Rezaei, A., and Reenan, R. a (2015). Dynamic response of RNA editing to temperature in Drosophila. BMC Biol. *13*.

Rosenthal, J.J.C., and Seeburg, P.H. (2012). A-to-I RNA Editing: Effects on Proteins Key to Neural Excitability. Neuron *74*, 432–439.

Shomrat, T., Zarrella, I., Fiorito, G., and Hochner, B. (2008). The Octopus Vertical Lobe Modulates Short-Term Learning Rate and Uses LTP to Acquire Long-Term Memory. Curr. Biol. *18*, 337–342.

Shomrat, T., Turchetti-Maia, A.L., Stern-Mentch, N., Basil, J.A., and Hochner, B. (2015). The vertical lobe of cephalopods: an attractive brain structure for understanding the evolution of advanced learning and memory systems. J. Comp. Physiol. A Neuroethol. Sensory, Neural, Behav. Physiol. *201*, 947–956.

Slotkin, W., and Nishikura, K. (2013). Adenosine-to-inosine RNA editing and human disease. Genome Med. *5*, 105.

Söller, R., Warnke, K., Saint-Paul, U., and Blohm, D. (2000). Sequence divergence of mitochondrial DNA indicates cryptic biodiversity in Octopus vulgaris and supports the taxonomic distinctiveness of Octopus mimus (Cephalopoda: Octopodidae). Mar. Biol. *136*, 29–35.

St Laurent, G., Tackett, M.R., Nechkin, S., Shtokalo, D., Antonets, D., Savva, Y. a, Maloney, R., Kapranov, P., Lawrence, C.E., and Reenan, R. a (2013). Genome-wide analysis of A-to-I RNA editing by single-molecule sequencing in Drosophila. Nat. Struct. Mol. Biol. *20*, 1333–1339.

Taglialatela, M., Toro, L., and Stefani, E. (1992). Novel voltage clamp to record small, fast currents from ion channels expressed in Xenopus oocytes. Biophys. J. *61*, 78–82.

Vizcaino, J.A., Csordas, A., Del-Toro, N., Dianes, J.A., Griss, J., Lavidas, I., Mayer, G., Perez-Riverol, Y., Reisinger, F., Ternent, T., et al. (2016). 2016 update of the PRIDE database and its related tools. Nucleic Acids Res. *44*, D447–D456.

Washburn, M.C., Kakaradov, B., Sundararaman, B., Wheeler, E., Hoon, S., Yeo, G.W., and Hundley, H.A. (2014). The dsRBP and Inactive Editor ADR-1Utilizes dsRNA Binding to

Regulate A-to-I RNA Editing across the C.elegans Transcriptome. Cell Rep. *6*, 599–607.

Whipple, J.M., Youssef, O.A., Aruscavage, P.J., Nix, D.A., Hong, C., Johnson, W.E., and Bass, B.L. (2015). Genome-wide profiling of the C. elegans dsRNAome. RNA *21*, 786–800.

Xu, G., and Zhang, J. (2014). Human coding RNA editing is generally nonadaptive. Proc. Natl. Acad. Sci. U. S. A. *111*, 3769–3774.

Young, J.Z. (1961). Learning and discrimination in the octopus. Biol Rev Camb Philos Soc. 32–96.

Young, J.Z. (1965). The central nervous system of Nautilus. Philos. Trans. B *249*, 1–25.

Young, J.Z. (1971). The Anatomy of the Nervous System of Octopus vulgaris (Clarendon Press).

Yu, Y., Zhou, H., Kong, Y., Pan, B., Chen, L., Wang, H., Hao, P., and Li, X. (2016). The landscape of A-to-I RNA editome is shaped by both positive and purifying selection. PLOS Genet. *12*, e1006191.

Zhao, H.-Q., Zhang, P., Gao, H., He, X., Dou, Y., Huang, A.Y., Liu, X.-M., Ye, A.Y., Dong, M.-Q., and Wei, L. (2015). Profiling the RNA editomes of wild-type C. elegans and ADAR mutants. Genome Res. *25*, 66–75.

Figure Legends

Figure 1: Extensive recoding is an invention of coleoid cephalopods.

(A) The species studied span the cephalopod evolutionary tree, as well as sea hare (*Aplysia californica*) as an outgroup (top). For comparison, a representative tree for vertebrates is shown (bottom), constructed based on divergence times estimated in (Hedges et al., 2006).

(B) Tens-thousands of A-to-I editing sites (identified as A-to-G DNA-RNA mismatches) are detected in squid, sepia and the two octopus species (see **Tables S1-S4** for more details). The noise level (estimated by the number of G-to-A mismatches) is rather low. In contrast, in nautilus and sea hare no enrichment of A-to-G mismatches is observed (inset).

(C) The nucleotides neighboring the detected editing sites, show a clear pattern consistent with known ADAR preference (Alon et al., 2015; Eggington et al., 2011; Kleinberger and Eisenberg, 2010) for the extensively recoded coleoid species – squid, sepia, and the two octopus species – but not in nautilus or sea hare. The motif is characterized by under-representation of G upstream to the editing site (relative location -1) and over-representation of G in the downstream base (The height of the entire stack of letters represents the information content in bits, the relative height of each letter represents its frequency).

Figure 2: Proteomic validation of recoding by RNA editing

We analyzed peptides identified by mass spectrometry analysis of two squid tissues, looking for evidence of recoding. For each site covered by one or more peptides, we marked whether the edited, non-edited or both versions of the peptide are observed. The distribution is presented, binned by the predicted RNA editing level (as measured from RNA-seq data). In parentheses are the numbers of recoding sites analyzed in each editing-level bin. The proteomic recoding level

follows closely the predicted RNA editing level. Altogether, this experiment validated protein recoding in 432 sites in two tissues:

(A) Squid stellate ganglion, where 320 of the 3,204 single-site peptides (10.0%) were shown to be edited.

(B) Squid giant axon (giant fiber lobe), where 283 of the 2,741 single-site peptides (10.3%) were shown to be edited.

Figure 3: Editing in Octopus bimaculoides.

(A) A-to-I editing sites were found within coding sequences of *Octopus bimaculoides* using three methods: the genome-free method (alignment to de-novo transcriptome), the genome-dependent approach using REDItools (Picardi and Pesole, 2013), and identification of hyper-edited reads (Porath et al., 2014). Overall, the three methods identified 170,825 unique AG sites in *Octopus bimaculoides* coding sequences (38,066 hyper-editing sites do not overlap those found by the other methods). See **Methods** for analysis of the differences between the results of the first two methods.

(B) RNA editing levels, measured across the whole transcriptome (see **Table S5**) by the editing index (weighted average of editing levels over all editing sites identified in the transcriptome, see **Methods**). Levels vary across tissues and are highest for neural tissues (see **Table S6**). Unlike mammals, a sizable fraction of editing events (11-13% in neural tissues) results in recoding events. Annotation of transcripts and repeats is based on (Albertin et al., 2015). (CNS= central nervous system; ANC=Axial nerve cord; OL=Optic Lobe; Sub=Subesophageal ganglia; Supra=Supraesophageal ganglia; PSG=posterior salivary gland; ST15=stage 15 embryo)

(C) The number of editing sites in coding region is comparable to the number found in introns.

(D) Unlike the case in mammals, editing is not exceptionally enriched in specific repeat families in *Octopus bimaculoides,* as measured by the editing index (here defined as the editing level averaged over all, edited and unedited, adenosines in each specific repeat family). **(E)** Protocadherins is a gene family known to be principally expressed in the brain, important for mediating combinatorial complexity in neuronal connections and are thought to play a role in diversifying neural circuitry (Chen and Maniatis, 2013). It was impressively expanded in *Octopus bimaculoides* (Albertin et al., 2015). A large number of protocadherins are found in the assembled transcriptomes for the four coleoid species (127-251 open reading frames), but not in nautilus (28 open reading frames).

(F-G) Protocadherins contain significantly higher numbers of AG sites **(F)** and are edited at higher levels (editing level summed over all sites and normalized by ORF length), in all four coleoid species but not in nautilus **(G)**.

Figure 4: Extensive recoding is conserved across coleoid cephalopods.

(A) Tens-thousands sites are conserved across species (see **Table S7)**. The closer the species are evolutionarily, the higher the number of conserved sites.

(B) Virtually all (97.5-99%) mismatches conserved across species are A-to-G, resulting from A-to-I editing. Manual inspection of the few non-A-to-G mismatches appearing in multiple species suggests that they either result from systematic erroneous alignments, or they are actually editing sites that were mistakenly identified as G-to-A mismatches due to insufficient DNA coverage.

(C) The majority of editing sites is conserved between the two octopus species, and even the most distant species share a sizable fraction of their sites.

(D) In contrast, only 36 human recoding sites (1-2% of human recoding sites) are shared by mouse, and a similar number is shared between *Drosophila melanogaster* and *D. mojavensis* (Yu et al., 2016) (diverged at later times than squid-sepia).

(E) Interestingly, 1146 AG modification sites (in 443 proteins) are conserved and shared by all four coleoid cephalopod species. Of these, 887 are recoding sites and 705 are highly edited (>=10% editing) recoding sites (in 393 proteins).

(F) Some proteins include multiple highly-edited recoding sites (see **Table S8**). Of note are Uromodulin, α Spectrin (previously reported to harbor the highest number of recoding sites in squid (Alon et al., 2015)), and Calcium-dependent secretion activator 1 (CAPS1) with 14, 8 and 7 strong shared recoding sites, respectively. Recoding in CAPS1 was found to be conserved in vertebrate species from human to zebrafish (Li et al., 2009).

(G) Not only are the locations of editing sites conserved, but their editing levels are correlated as well. Editing levels in 887 recoding sites shared by all species are highly, positively and significantly correlated in all pairs of coleoid cephalopod species ($p<1e-75$ for all pairs; see **Supp. Fig. 1** for three additional pairs). Correlation is higher the closer the species are to each other in evolutionary terms, with Pearson rho = 0.95 for the two octopus species.

Figure 5: Signs for positive selection of recoding by editing.

(A) The fraction of recoding sites among all editing sites in coding region increases with editing levels (top), as well as the fraction of recoding sites among all conserved sites (bottom). Red horizontal dashed line represents the recoding fraction expected assuming neutrality.

(B) Editing levels are higher in conserved recoding sites. Distributions of editing levels in four groups of putative A-to-I editing sites: recoding and conserved (Rec+, Cons+), recoding and

non-conserved (Rec+, Cons-), conserved sites that cause a synonymous change (Rec-, Cons+), and non-conserved synonymous sites (Rec-, Cons-). Horizontal red lines mark the median level, and yellow diamonds mark the mean. Conservation and non-synonymity are both positively correlated with higher editing levels, as well as their interaction (ANOVA, p-value<1.0e-162). Data presented here for squid (conserved sites are conserved in sepia), but the results are similar and significant for all species.

(C) In contrast with the case in humans, highly edited sites tend to be more conserved: the fraction of conserved sites rises with the editing level for all species pairs, but more dramatically for the closely related octopuses and the sepia-squid pair.

(D) Highly conserved regions of the transcriptome are enriched in editing sites, further attesting for positive selection of RNA editing. Density of editing sites (number of AG sites normalized by length) is higher for 112 recoding regions that are highly-conserved across the four species (>95% identity; average length 1382bp), compared with all other, less conserved, regions (Wilcoxon p-value<0.001 for all species). Error bars represent the S.E.M.

Figure 6: Conserved and species-specific editing sites affect protein function. Unedited (wt) and singly-edited versions of the voltage-dependent K^+ channels of the K_v2 subfamily were studied under voltage-clamp (see **Table S9**).

(A) (i) Current traces resulting from a voltage step from -80 mV to 40 mV for the wt Sepia $K_v2.1$ and the same construct containing the sepia-specific I529V edit, lying within the 4^{th} transmembrane domain (green), showing that I529V accelerates the rate of slow inactivation. (ii) Time constants for slow inactivation determined by fitting single exponentials to traces similar to those in panel (i) at different activating voltages (Vm).

(B) (i) Tail currents measured at a voltage (Vm) of -80mV, following an activating pulse of +20 mV for 25 ms. Traces are shown for the wt $K_v2.1$ channels from squid, sepia and Octopus vulgaris. (ii) Tail currents for the same channels edited at the shared I-to-V site in the 6^{th} transmembrane span, following the same voltage protocol. (iii) Time constants from single exponential fits to tail currents obtained at various negative voltages (Vm) (following an activating pulse to 20 mV for 25 ms) show that the unedited channels close at distinct rates, (iv) but the edited versions close at similar rates. N = 5 ± s.e.m. for all data plotted in this figure.

Figure 7: RNA editing slows down cephalopod genome evolution.

(A) Inter-species mutations are purified from genome loci surrounding conserved recoding sites (data shown for sites shared by squid and sepia). Depletion of mutations extends up to ~100bp of shared recoding sites (left). As a control, we show the mutations density (mutations/bp) around random non-edited adenosines from the same transcripts (right). Yellow – synonymous change; light green – non-synonymous; dark green – deletions.

(B) Genomic polymorphisms are depleted near editing/recoding/conserved-recoding sites in squid, attesting to reduced genome plasticity. Effect is stronger for recoding sites, and even more so for the conserved recoding sites.

(C) GC-content is elevated near editing sites in squid, allowing for more stable double-stranded RNA structures. The effect is even stronger in conserved sites. Dashed line represents the baseline GC level in the entire ORFome, and error bars represent the S.E.M. See **Supp. Fig. 3** for analyses similar to those presented in panels A-C in other species.

Supplemental Figure Legends

Figure S1, related to Figure 4:

Editing levels of shared sites are correlated across coleoid species

Editing levels in 887 recoding sites shared by all species are highly, positively and significantly correlated in all pairs of coleoid cephalopod species ($p<1.0e-75$ for all pairs). Correlation is higher the closer the species are to each other in evolutionary terms, with Pearson rho = 0.95 for the two octopus species.

Figure S2, related to Figure 6:

Sequence alignments for coleoid $K_v2.1$ channels, and their editing sites.

Orthologous cephalopod $K_v2.1$ channel sequences are abundantly edited. Editing sites predicted by our analysis are indicated by small boxes around specific amino acids. Conserved transmembrane spans S1-S6 are enclosed in large boxes. The unique editing site for Sepia in S4 (I529V) and the common editing site in S6 (Squid I579V, Sepia I630V and *Octopus vulgaris* I632V) were those studied using electrophysiology in Figure 6 of the main test. The alignment was generated using Vector NTI software.

Figure S3, related to Figure 6:

Basic electrophysiological properties of genomically encoded (wt) cephalopod $K_v2.1$ channels.

(A) Examples of current traces recorded on fast (top) and slow (bottom) time-bases following activating pulses to +60 mV from a holding potential of -80 mV. These traces illustrate the variable activation and inactivation rates of the genomically encoded channels.

(B) The channels' voltage sensitivities are similar. Relative conductance was measured from normalized tail-current amplitudes following 150 ms activating pulses to various activation voltages (V_m).

(C) Inactivation time constants (τ) taken from single exponential fits to current traces measured during a 1-second activation step of V_m (similar to the traces presented in panel A).

(D) The half-times for 150 ms activating pulses were measured from traces similar to those presented in panel A. They measure the time it takes for the currents to reach one half of their maximum values following an activating voltage step. N=6 ± S.E.M. for all graphs.

Figure S4, related to Figure 7:

Reduction in intra- and inter-genomic variability around editing locations.

(A-F) Species-to-species mutations are suppressed within ~100bp of a recoding site shared by the two species (left panels). No effect is seen when taking random non-edited sites from conserved regions within the same transcripts (right panels) **(A)** Oct.bim.-Oct.vul. **(B)** Oct.bim.-Sepia **(C)** Oct.bim.-Squid **(D)** Sepia-Oct.vul. **(E)** Squid-Oct.vul. **(F)** Squid-Sepia. (yellow – synonymous change, light green – non-synonymous; dark green – deletions; Mutations density is the number of mutations found at a given distance from an editing site divided by the number of such sites, i.e. mutations per base-pair)

(G) Intra-species genomic polymorphisms are suppressed in genomic loci surrounding editing sites. Genomic polymorphisms are depleted near editing/recoding/conserved-recoding sites, attesting for a decrease in genome plasticity. Effect is stronger for recoding sites, and even more so for the conserved recoding sites.

(H) Higher GC content in the vicinity of editing sites. GC content is elevated near editing sites. This effect persists up to ~100bp, and is therefore unlikely to be explained by the local ADAR preferences. Rather, it may facilitate formation of stronger and more stable double-stranded RNA structures required for ADAR binding and editing. Blue: all editing sites; green: all recoding sites; red: conserved recoding sites; dashed black line: GC content averaged over all ORFs. Error bars represent the S.E.M. The effect is not seen in for AG mismatches found in nautilus, expected to be mostly false-positives.

Figure S5, related to Figure 7:

Positive selection of editing sites slows down genome evolution.

(A) Mutation rates (averaged over all orthologous parts of the transcriptome) are presented as a function of the distance of the genomic base to the closest conserved recoding site. The baseline mutation rate (red) represents the rate measured for all sites at distances >1000bp from a recoding site (including mutations in ORFs harboring no editing site). Integrating the difference between the baseline and the observed rate for genomic location at a distance shorter than 1000bp provides an estimate of the number of avoided mutations, apparently due to the constraints imposed by editing. We estimate the fraction of avoided mutations as the number of avoided mutations divided by (#avoided+#observed), and find this fraction (marked as A in the different panels) to be 3-15%. This is an underestimation, as sites not conserved between the two species, or sites in transcriptome regions not conserved between the two species, are not taken into account.

(B) Similar to the analysis presented in **(A)**, we looked at the SNPs density (number of SNPs per bp) as a function of the distance of the genomic base to the closest recoding site. The baseline SNPs density (red) is the density observed at distances of >500bp from a recoding site (including

SNPs in ORFs harboring no editing site). Integrating the difference between the baseline and the observed rate for genomic location at a distance shorter than 500bp provides an estimate of the number of avoided SNPs, apparently due to the constraints imposed by editing. Here too, we estimate the fraction of avoided SNPs as the number of avoided SNPs divided by (#avoided+#observed), and find this fraction (marked as A in the different panels) to be 10-26%. This is an underestimation, as the contributions of sites in non-conserved transcriptome regions, are not taken into account.

STAR Methods

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Joshua Rosenthal (jrosenthal@mbl.edu), Eugene Bell Center, Marine Biological Laboratory, Woods Hole, MA 02543 USA.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

An adult male *Octopus vulgaris* was captured in October of 2014 at the mouth of the Laguna de Condado, San Juan, Puerto Rico. An adult male *Sepia officinalis* was provided by Dr. Roger Hanlon from the Marine Biological Laboratory in Woods Hole. This individual was raised in the Marine Biological Laboratory's Marine Resource Center from a fertilized egg that was collected in the English Channel off of Christchurch, Dorset England in 2014. The Stellate ganglia (SG), the optic lobe (OL) and a portion of the sperm sack were manually dissected from these specimens. One sample of *Nautilus pompilius* originating from the Philippines was obtained from "SeaDwelling Creatures" in Los Angeles. The optic lobe and the supraesophogeal ganglia were manually dissected from the sample. All samples destined for RNA extraction were immersed in chilled, filtered seawater and immediately preserved in RNA later. The samples intended for DNA extraction were flash frozen in liquid nitrogen. All samples were then stored at -80C°. RNA from all tissues was extracted with the RNAqueous kit (Life Technologies, Carlsbad, CA), and genomic DNA was extracted from the sperm sack using Genomic Tip Columns (Qiagen, Venlo, Limburg, The Netherlands).

We added previously described RNA and DNA sequencing data from the squid species *Doryteuthis pealeii* (Alon et al., 2015) (PRJNA255916), as well as published data from another

Octopus species, *Octopus bimaculoides* (PRJNA270931, PRJNA285380), whose genome was recently sequenced (Albertin et al., 2015). We also added available RNA and DNA samples from the mollusk *Aplysia californica*, as an evolutionary outgroup (PRJNA13635, PRJNA77701).

METHOD DETAILS

Library preparation and sequencing

The genomic DNA sequencing library for *Octopus vulgaris*, sepia and nautilus were prepared using the TruSeq DNA Sample Prep kit, as described by the manufacturer (Illumina, San Diego, CA), and sequenced using three lanes of the Illumina HiSeq 2000 instrument. The RNA-Seq libraries for all the samples were prepared using the TruSeq Stranded mRNA Sample Prep Kit, as described by the manufacturer (Illumina), and were sequenced using one lane for each sample of Illumina HiSeq 2000 instrument.

Ilumina sequencing was utilized to generate paired-end, 151 nt reads, using RNA from OL and SG tissues for *Octopus vulgaris* and sepia, and OL and subesophageal ganglia for nautilus. For DNA sequencing, 101 nt reads were produced. The number of reads generated for each tissue is presented in **Table S1**.

Editing site validation

Validation of editing sites predicted by our bioinformatics pipeline was previously performed and reported (Alon et al., 2015). In brief, direct Sanger sequencing confirmed editing at 40/40 A-to-G squid recoding sites, and deep-sequencing validated 120/143 A-to-G recoding sites but none of the 12 non A-to-G sites tested. In this work we take validation a step further and examine editing at the protein level.

In gel proteolysis and mass spectrometry analysis

We applied proteomic mass spectroscopy (MS) analysis to examine the extent to which these RNA modifications are translated into the proteome. Squid giant axon and stellate ganglion samples were separated by a SDS-PAGE and the gel was stained with Coomassie Blue and sliced to 5 slices. The slices were processed for tryptic digestion by first reducing the disulfides with 3 mM DTT in 100 mM ammonium bicarbonate for 20 min at 60ºC. Next, the sulfhydryl were carboxymethylated with 10 mM iodoacetamide in 100 mM ammonium bicarbonate in the dark for 30 min at room temperature. The proteins were in-gel digested with modified trypsin (Promega) in 10% acetonitrile and 10 mM ammonium bicarbonate at a 1:10 enzyme-to-substrate ratio overnight at 37°C. Additional trypsinization was done for 4 hours.

The resulting tryptic peptides were desalted on C18, Stage-Tip (Ishihama et al., 2006) and resolved by reverse-phase chromatography on 0.075 X 200-mm fused silica capillaries (J&W) packed with reversed phase Reprosil-C18-Aqua (Dr. Maisch GmbH, Ammerbuch-Entringen, Germany) as in (Ishihama et al., 2002). The peptides were eluted with linear 105 minutes gradient of 5% to 28% acetonitrile with 0.1% formic acid in water, followed by 15 minutes gradient from 28% to 90% acetonitrile and 15 minutes at 90% acetonitrile at flow rates of 0.15 μl/minutes. MS was performed by a Q-Exactive-Plus mass spectrometer (Thermo Fisher Scientific) in a positive ion mode using repetitively full MS scan followed by Higher-energy Collision Dissociation (HCD) of the 10 most dominant ions, selected from the first MS scan. A dynamic exclusion list was enabled with an exclusion duration of 20s.

The MS data was analyzed using Proteome Discoverer 1.4 software with the Sequest (Thermo Fisher Scientific) algorithm against the specific databases, combining all 5 fractions of each sample in one search. Minimal peptide length was set to six amino acids and a maximum of

two miscleavages was allowed. Mass tolerance of 15 ppm for the precursor masses and for the fragment ions. Peptide- and protein-level false discovery rates (FDRs) were filtered to 1% using the target-decoy strategy. Semi quantitation was done by calculating the peak area of each peptide based its extracted ion currents (XICs), and the area of the protein is the average of the three most intense peptides from each protein.

Expression and recording of cephalopod K_v2 channels in Xenopus oocytes

Full-length Octopus vulgaris, Sepia and Squid $K_v2.1$ constructs, to be expressed in Xenopus oocytes, were based on the unedited amino acid sequences deduced from the transcriptome assemblies. The sequence for the two Octopus species are almost identical (**Supp. Table 4**), so we studied only one of them. Codon optimized versions for Xenopus laevis were synthesized using gene blocks and cloned into the Xenopus expression vector pGEMHE (Liman et al., 1992) using a Gibson Assembly. Single RNA editing sites were added to these clones by standard oligonucleotide-based site-directed mutagenesis using the Quickchange Lightening Site-Directed Mutagenesis Kit (Agilent Technologies). Capped, polyA-tailed cRNA from each clone was transcribed using the T7 mMessage mMachine Kit (Thermo Fisher) and injected into stage V and V1 Xenopus oocytes at approximately 750 pg/oocyte. Currents were recorded 2-3 days after injections using the Cut-Open Oocyte Vaseline Gap technique (Lockery and Goodman, 1998), using a relatively slow sampling rate. The external solution contained: 20mM K-Glutamate, 100mM N-Methyl-D-Glucamine-Glutamate, 2.5mM MgCl2, 2.5mM CaCl2, 10mM HEPES, pH=7.4. Oocytes were permeabilized using 0.3% Saponin in the internal solution. Voltage was controlled using a CA-1B High Performance Oocyte Clamp (Dagan Corporation). Analog currents were digitized at 100 kHz, and voltage commands were made, using an SBC6711 A/D D/A board (Innovative Integration, Simi Valley CA). Signals were filtered at 5 kHz. To avoid

series resistance errors, only traces exhibiting less than 10 mA current were used for analysis. Data collection and clamp command was made using GPATCH M software, kindly provided by Dr. F Bezanilla (University of Chicago). Leak currents were subtracted using a linear P/4 procedure. Data were analyzed using the ANALYSIS software, also provided by Dr. F. Bezanilla. Deactivation kinetics were measured42 by fitting a single exponential of the form $y = Ae^{-\tau/T} + B$ to the traces where A= current amplitude, τ = the time constant, T = time, and B =a constant baseline. Inactivation kinetics were measured42 by fitting a single exponential of the form $y = Ae^{-\tau/T}$ to the traces where A= current amplitude, τ = the time constant, and T = time. Relative conductance was measured using peak amplitude of tail currents. N = 6 for all data points ± SEM.

QUANTIFICATION AND STATISTICAL ANALYSIS

Transcriptome assembly

The species studied, with the exception of *Octopus bimaculoides*, do not have a reference genome. We therefore utilized the RNA-Seq data to assemble a transcriptome using the Trinity de novo assembly package, Version: Trinity-r2012-10-05 (Grabherr et al., 2013). We found that Trinity does not perform well for reads longer than 100bp, and therefore all reads were trimmed (symmetrically, on both sides) to this length prior to assembly. For consistency, we assembled the transcriptome for *Octopus bimaculoides* in the same manner, using data from the four neural tissues out of the 12 tissues available: optic lobe (OL), supraesophogeal ganglia (supra), subesophogeal ganglia (sub), and the axial nerve cord (ANC). For each gene, we kept only the longest isoform for downstream analysis. The assembly statistics are summarized in **Table S2**.

Detection of editing sites

In order to detect editing sites, using the matched RNA and genomic DNA samples, we employed a method similar to that described in (Alon et al., 2015). Briefly, RNA and DNA reads were separately aligned against the assembled transcriptome using Bowtie2 with local alignment configuration and default parameters (Langmead and Salzberg, 2012). Reads that were not uniquely aligned were discarded. We identified potential open reading frames (ORFs) in the assemblies by locating components that were found to be significantly similar (Blastx E-value<1e-6) to the Swiss-Prot proteins dataset (Bairoch et al., 2005). Each ORF was extended until either a stop codon or the end of the Trinity component was met. **Table S2** summarizes the properties of the ORFs found for each species. To detect editing events, we applied a binomial test to locate significant modifications between RNA reads and the Swiss-Prot ORFs and distinguish them from sequencing errors or SNPs. For more detailed description of the editing detection method, see (Alon et al., 2015). Two important modifications were introduced here with respect to the scheme presented in (Alon et al., 2015): (1) we discarded all mismatches that occur up to 6bp from alignments' ends (2) we discarded all reads that contained more than one mismatch type (e.g. A-to-G mismatch and A-to-C mismatch is the same read), or more than two mismatches (of any type) altogether, as these are suspected to be misaligned. The number of reads discarded ranges between 0.6 and 1.7 million, for the species studied. Of these, 200-800 thousand reads might have exhibited two editing events. Unlike (Alon et al., 2015), we did not distinguish between “weak” and “strong” sites. Rather, the genomic strand was determined by the DNA reads (and in the absence of DNA coverage, by the majority of RNA reads). In case of conflicting DNA reads, sites were discarded.

Editing in Octopus bimaculoides

Our genome-free detection scheme is limited by the quality of the assembled de-novo transcriptome, and the focus on coding sequences only. The recent sequencing of *Octopus bimaculoides* genome (Albertin et al., 2015) enabled us to compare our genome-free scheme to a genome-based one, obtain an independent assessment of the true extent of editing in coding sequences, and estimate the full picture of the octopus editome outside of coding sequences.

The REDITools package (V-1.0.3) (Picardi and Pesole, 2013) was used to locate RNA-DNA differences in the octopus genome, using RNA-seq data from the same four neural tissues that were used in the genome-free scheme. Editing sites annotation was based on the transcriptome provided in (Albertin et al., 2015), which differed from the one we used in the genome-free method.

We found 800941 AG sites, 105380 of them in annotated coding sequences (compared to 76862 sites in coding sequence identified by our transcriptome-based pipeline), 49483 of these sites were also found using our pipeline (see **Table S5**). Therefore, we see that our pipeline provides a reasonable coverage of the editing signal in coding sequence, and that the number of editing sites outside the coding region is likely to be an order of magnitude higher than the number within the coding sequence for the other cephalopods studied here. The differences between the two methods stem from the different de-novo transcriptome used, and the different parameters used to assess the mismatches observed (REDItools uses cutoffs on the number of reads, while the genome-free approach applied a binomial analysis). Using BLAST to compare the two transcriptomes, we find that ~90% of *Octopus bimaculoides* transcripts detected in our pipeline are covered by the transcriptome described in (Albertin et al., 2015), but only ~68% of the latter transcripts are covered by our detected ORFs, which explains the majority of the sites

missed by our pipeline. Of the ~56K sites found by REDItools but not by the genome-free method, 35911 (64%) are located within transcripts missing from our de-novo transcriptome and the rest are seen by the genome-free method but fail to achieve statistical significance (e.g. cases where coverage is very high, and only a few G's are observed). Out of the ~27K sites found using the genome-free method but not by REDItools, about 14k (52%) were missed due to too strict parameters employed by REDItools, and ~13K sites (48%) reside in sequences that do not exist in the transcriptome assembled in (Albertin et al., 2015).

In order to profile the editome in all available tissues, we re-did the analysis using all available RNA data from 12 tissues available: axial nerve cord (ANC), optic lobe (OL), subesophageal brain (Sub), supraesophageal brain (Supra), Ovaries, posterior salivary gland (PSG), retina, skin, stage 15 embryo (ST15), suckers, testes and viscera (heart, kidney and hepatopancreas), leading to a total of 903,742 sites.

As editing is so abundant in these organisms, it is expected that many sites will reside in reads that are extensively edited, to the point they are not aligned to the reference genome/transcriptome by the standard alignment tools. Thus, we also applied the method suggested in (Porath et al., 2014) to identify hyper-edited reads, leading to 38066 additional sites in coding regions, not found by the other methods (**Fig. 3A**).

The editing index was used to compare editing activity across different tissues (**Fig. 3B).** It is defined as the number of 'G's in RNA-reads nucleotides that were aligned to the predicted editing sites, divided by the total number of read-nucleotides that align to these positions ('A's and 'G's). In order to compare editing between different repeats, we used the repeats editing index, calculated in the same way over all genomic adenosines within repeats (number of A-to-G mismatches in RNA-reads nucleotides that were found in repeats, divided by the total number of

read-nucleotides that align to genomic adenosines in repeats). The higher the index, the more editing occurs in the specific repeat family element (**Fig. 3D**).

Functional analysis of edited ORFs

To test for functional enrichment, we ranked the genes by cumulative editing levels (editing levels summed over all sites within the gene, normalized by ORF length), and used the online tool GOrilla (Eden et al., 2009). As a control, the genes were also ranked by expression levels measured by FPKM and analyzed in the same manner (**Supp. Table 6**).

Finding orthologous editing sites

We used OrthoMCL (Li et al., 2003) to identify orthologous pairs of ORFs, using only best-two-way hits. This covered 77% of the transcriptomes. Then, for each editing location in the query species of each pair, we screened for the best possible aligned region in the ortholog using BLAST alignment scores, and found the matching nucleotide and amino acid. A conserved editing site is a case where the same mismatch occurs at the exact orthologous nucleotide, leading to the same amino acid substitution.

Identifying SNPs

In order to minimize false positives, we took a conservative approach, and called a genomic location a SNP if the DNA reads aligned to this location show exactly two types of nucleotides, with at least five reads supporting each type, and each of the types being supported by 30-70% of the reads. Obviously, this scheme is not meant to exhaust the list of SNPs in these species, which is anyway impossible using a single animal per species, but suffices to estimate the relative depletion of SNPs around recoding sites.

DATA AND SOFTWARE AVAILABILITY

The data used in the study are publicly available at the Sequence Read Archive (SRA), accessions PRJNA300723. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE (Vizcaino et al., 2016) partner repository with the dataset identifier PXD005827.The de-novo constructed transcriptomes used for the analysis are available at http://www.tau.ac.il/~elieis/squid.

Supplementary tables (separate Excel files):

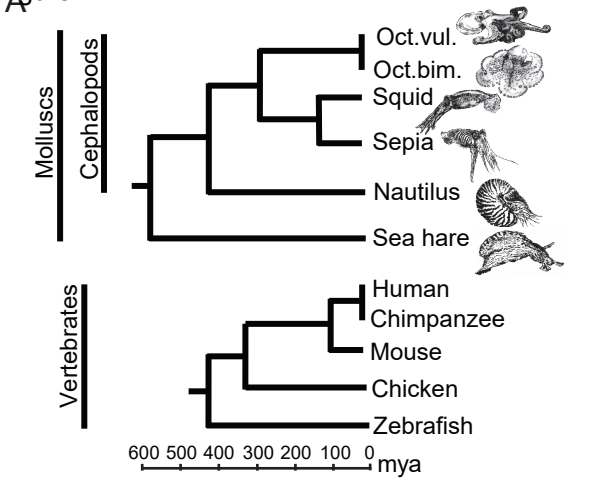
Table S4. Editing sites detected (genome-free method, all species), related to Figure 1.

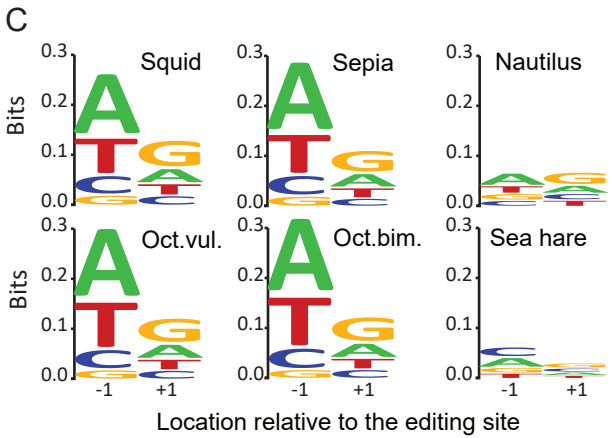
Table S5. Editing sites detected (genome alignment, Oct. Bim.), related to Figure 3.

Table S6. GO enrichment data (Oct. Bim.), related to Figure 3.

Table S7. Editing sites conserved across species, related to Figure 4.

Figure 1





B

Click here to download Figure Figure 1.pdf

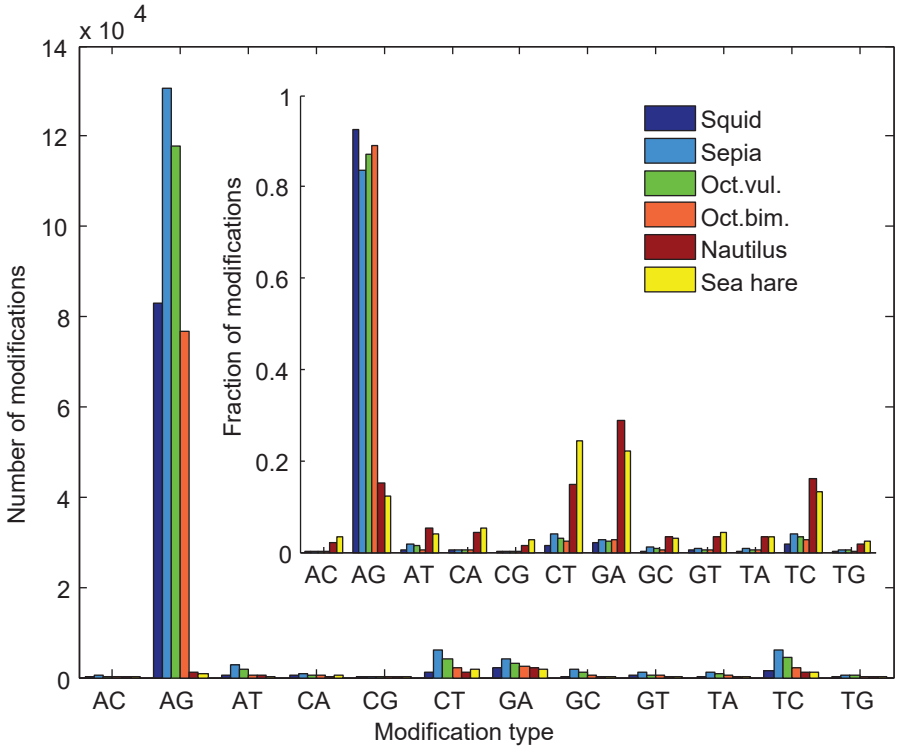


Figure 2

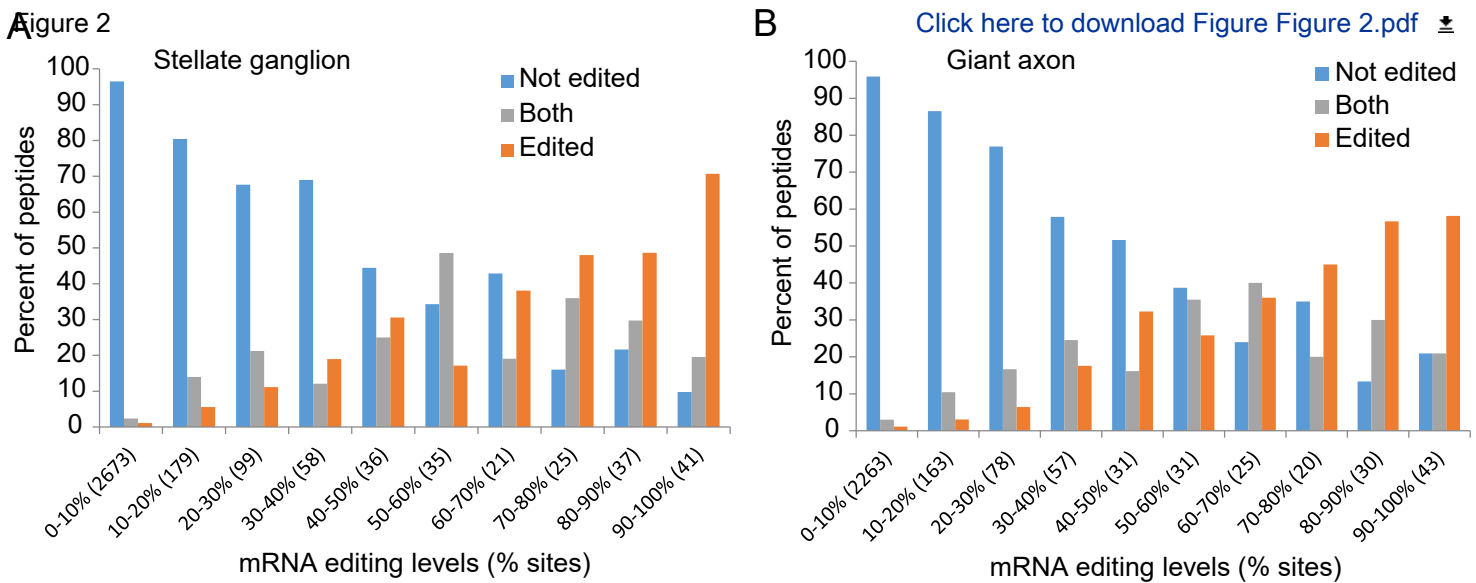

A
Stellate ganglion
Not edited
Both
Edited
Percent of peptides
100
90
80
70
60
50
40
30
20
10
0
0-10% (2673)
10-20% (179)
20-30% (99)
30-40% (58)
40-50% (36)
50-60% (35)
60-70% (21)
70-80% (25)
80-90% (37)
90-100% (41)
mRNA editing levels (% sites)
B
Click here to download Figure Figure 2.pdf
Giant axon
Not edited
Both
Edited
Percent of peptides
100
90
80
70
60
50
40
30
20
10
0
0-10% (2263)
10-20% (163)
20-30% (78)
30-40% (57)
40-50% (31)
50-60% (31)
60-70% (25)
70-80% (20)
80-90% (30)
90-100% (43)
mRNA editing levels (% sites)

Figure 3

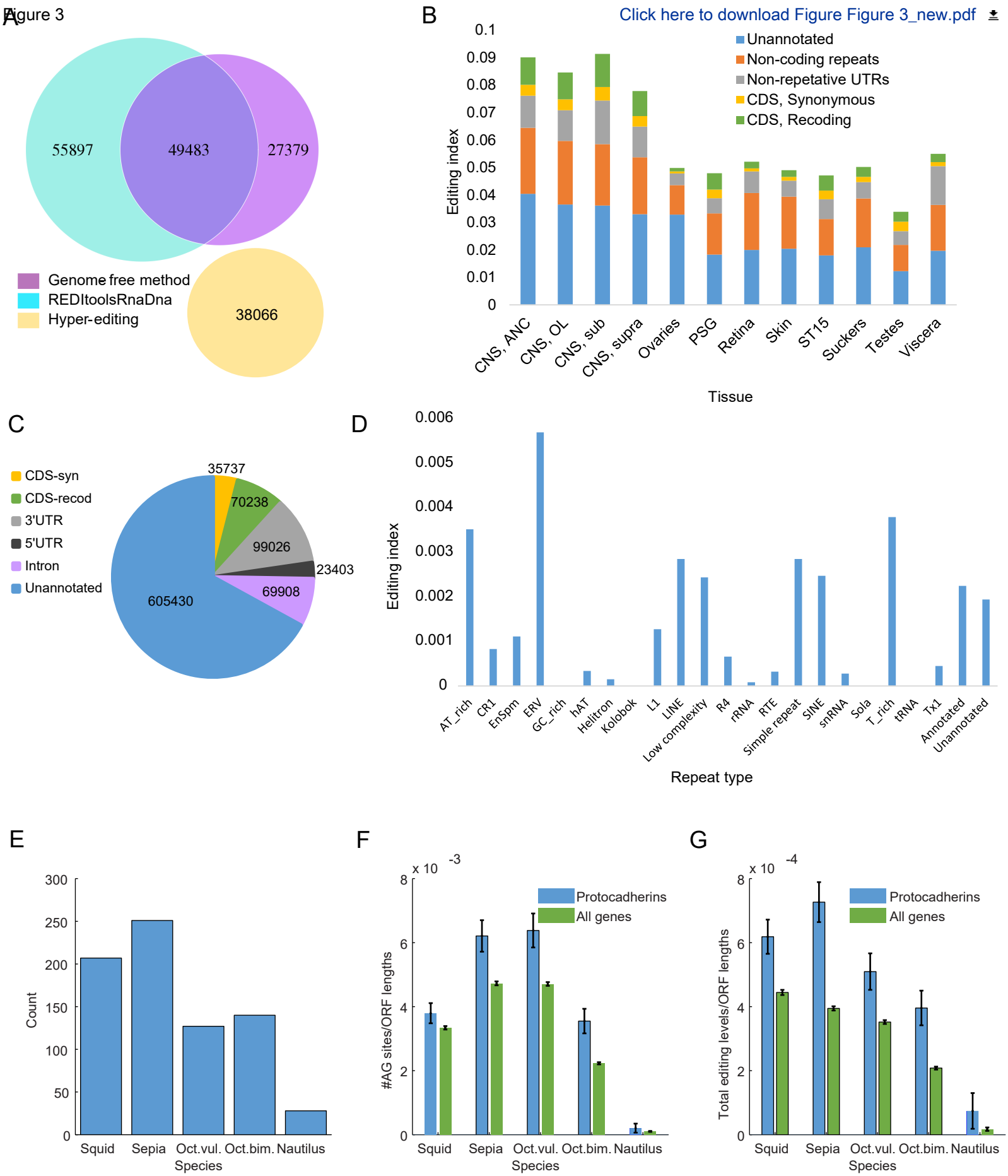

A
55897
49483
27379
38066
Genome free method
REDItoolsRnaDna
Hyper-editing
B
Click here to download Figure Figure 3_new.pdf
Unannotated
Non-coding repeats
Non-repetative UTRs
CDS, Synonymous
CDS, Recoding
Editing index
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
CNS, ANC
CNS, OL
CNS, sub
CNS, supra
Ovaries
PSG
Retina
Skin
ST15
Suckers
Testes
Viscera
Tissue
C
CDS-syn
CDS-recod
3'UTR
5'UTR
Intron
Unannotated
35737
70238
99026
23403
69908
605430
D
Editing index
0.006
0.005
0.004
0.003
0.002
0.001
0
AT_rich
CR1
EnSpm
ERV
GC_rich
hAT
Helitron
Kolobok
L1
LINE
Low complexity
R4
rRNA
RTE
Simple repeat
SINE
snRNA
Sola
T_rich
tRNA
Tx1
Annotated
Unannotated
Repeat type
E
Count
300
250
200
150
100
50
0
Squid
Sepia
Oct.vul.
Oct.bim.
Nautilus
Species
F
x 10 -3
#AG sites/ORF lengths
8
6
4
2
0
Protocadherins
All genes
Squid
Sepia
Oct.vul.
Oct.bim.
Nautilus
Species
G
x 10 -4
Total editing levels/ORF lengths
8
6
4
2
0
Protocadherins
All genes
Squid
Sepia
Oct.vul.
Oct.bim.
Nautilus
Species

Figure 4

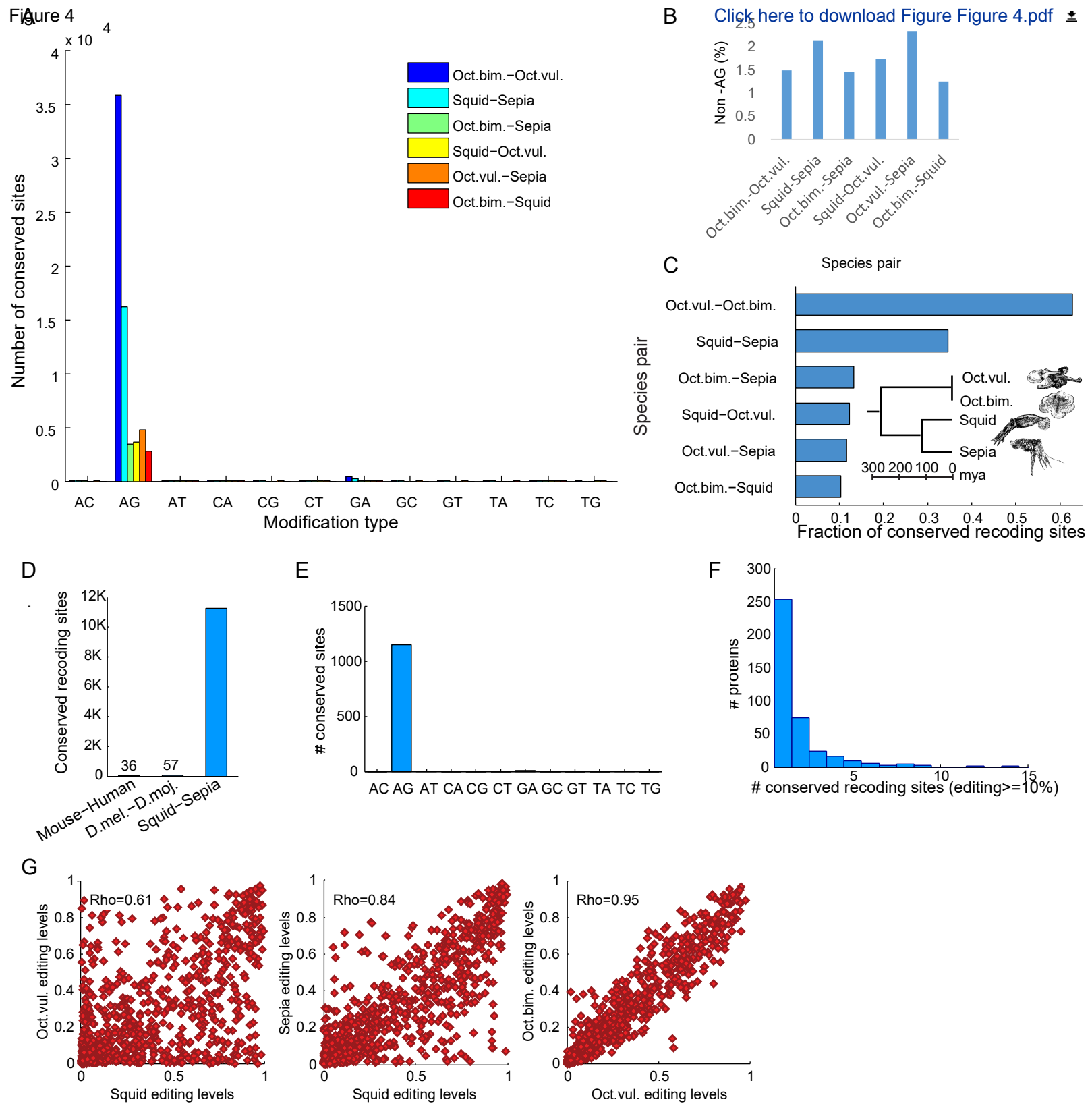
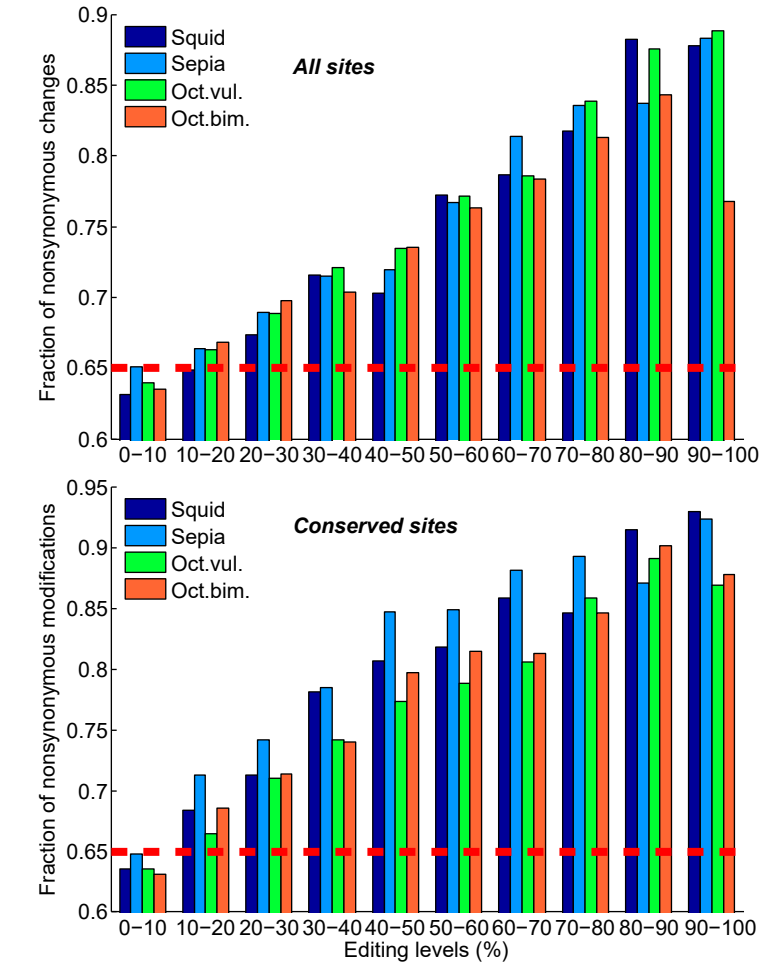
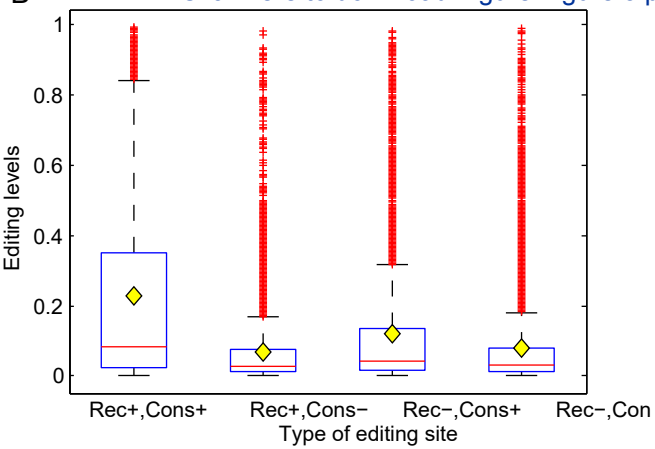
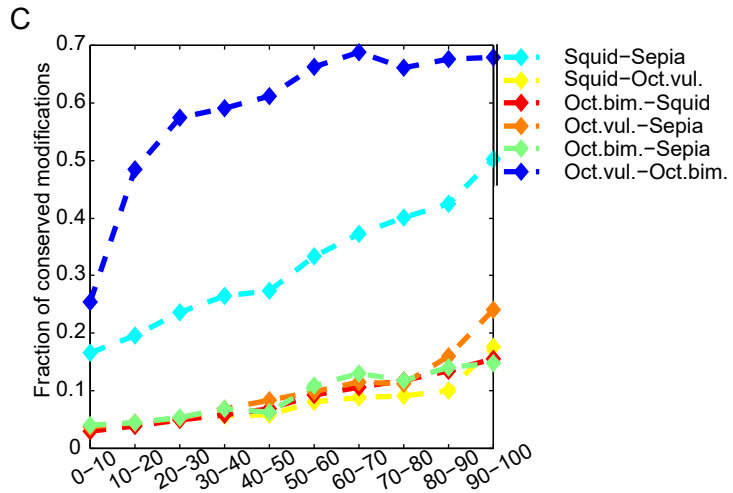
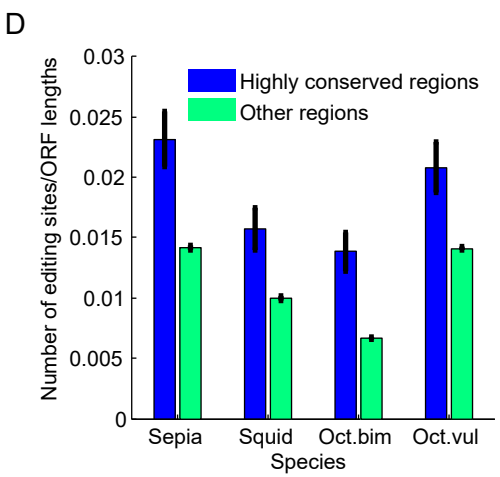

A
x 10^4
Number of conserved sites
Modification type
AC AG AT CA CG CT GA GC GT TA TC TG
Oct.bim.−Oct.vul.
Squid−Sepia
Oct.bim.−Sepia
Squid−Oct.vul.
Oct.vul.−Sepia
Oct.bim.−Squid
B
Click here to download Figure Figure 4.pdf
Non -AG (%)
Oct.bim.-Oct.vul.
Squid-Sepia
Oct.bim.-Sepia
Squid-Oct.vul.
Oct.vul.-Sepia
Oct.bim.-Squid
C
Species pair
Oct.vul.−Oct.bim.
Squid−Sepia
Oct.bim.−Sepia
Squid−Oct.vul.
Oct.vul.−Sepia
Oct.bim.−Squid
Oct.vul.
Oct.bim.
Squid
Sepia
300 200 100 0
mya
Fraction of conserved recoding sites
D
Conserved recoding sites
36
57
Mouse−Human
D.mel.−D.moj.
Squid−Sepia
E
conserved sites
AC AG AT CA CG CT GA GC GT TA TC TG
F
proteins
conserved recoding sites (editing>=10%)
G
Rho=0.61
Oct.vul. editing levels
Squid editing levels
Rho=0.84
Sepia editing levels
Squid editing levels
Rho=0.95
Oct.bim. editing levels
Oct.vul. editing levels

Figure 5

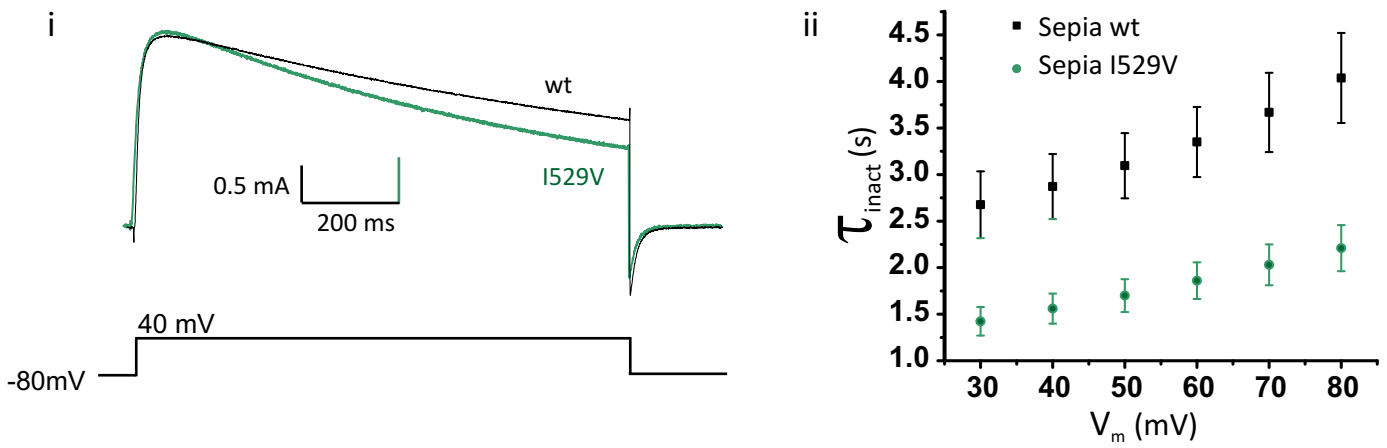

A
Squid
Sepia
Oct.vul.
Oct.bim.
All sites
Fraction of nonsynonymous changes
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0–10 10–20 20–30 30–40 40–50 50–60 60–70 70–80 80–90 90–100
Squid
Sepia
Oct.vul.
Oct.bim.
Conserved sites
Fraction of nonsynonymous modifications
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0–10 10–20 20–30 30–40 40–50 50–60 60–70 70–80 80–90 90–100
Editing levels (%)


B
Click here to download Figure Figure 5.pdf
Editing levels
1
0.8
0.6
0.4
0.2
0
Rec+,Cons+
Rec+,Cons−
Rec−,Cons+
Rec−,Con
Type of editing site


C
Fraction of conserved modifications
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0–10 10–20 20–30 30–40 40–50 50–60 60–70 70–80 80–90 90–100
Squid−Sepia
Squid−Oct.vul.
Oct.bim.−Squid
Oct.vul.−Sepia
Oct.bim.−Sepia
Oct.vul.−Oct.bim.


D
Highly conserved regions
Other regions
Number of editing sites/ORF lengths
0.03
0.025
0.02
0.015
0.01
0.005
0
Sepia
Squid
Oct.bim
Oct.vul
Species

A. Sepia $K_v2.1$ specific editing site



B.Sepia, squid and oct. vulg. common editing site

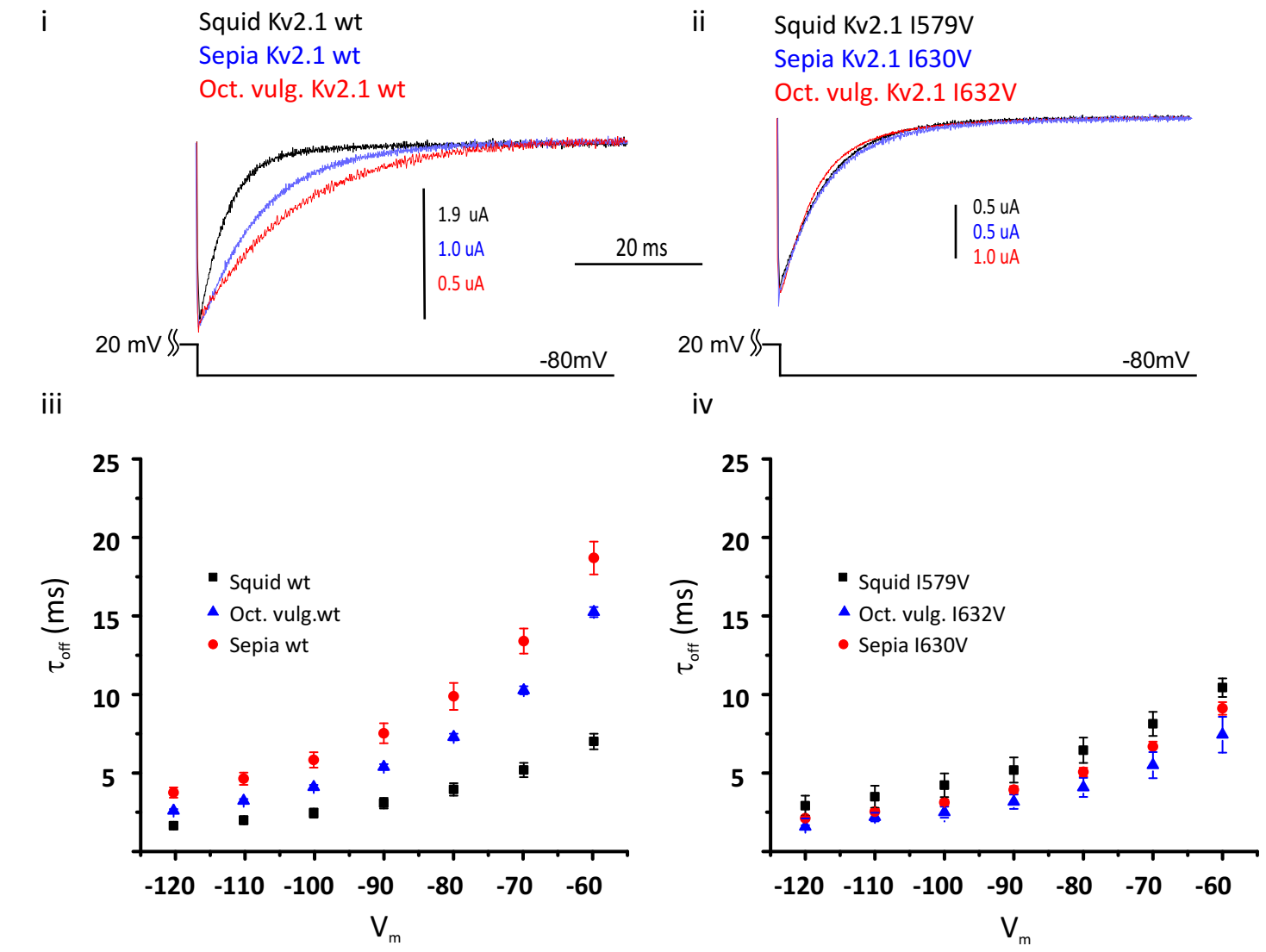
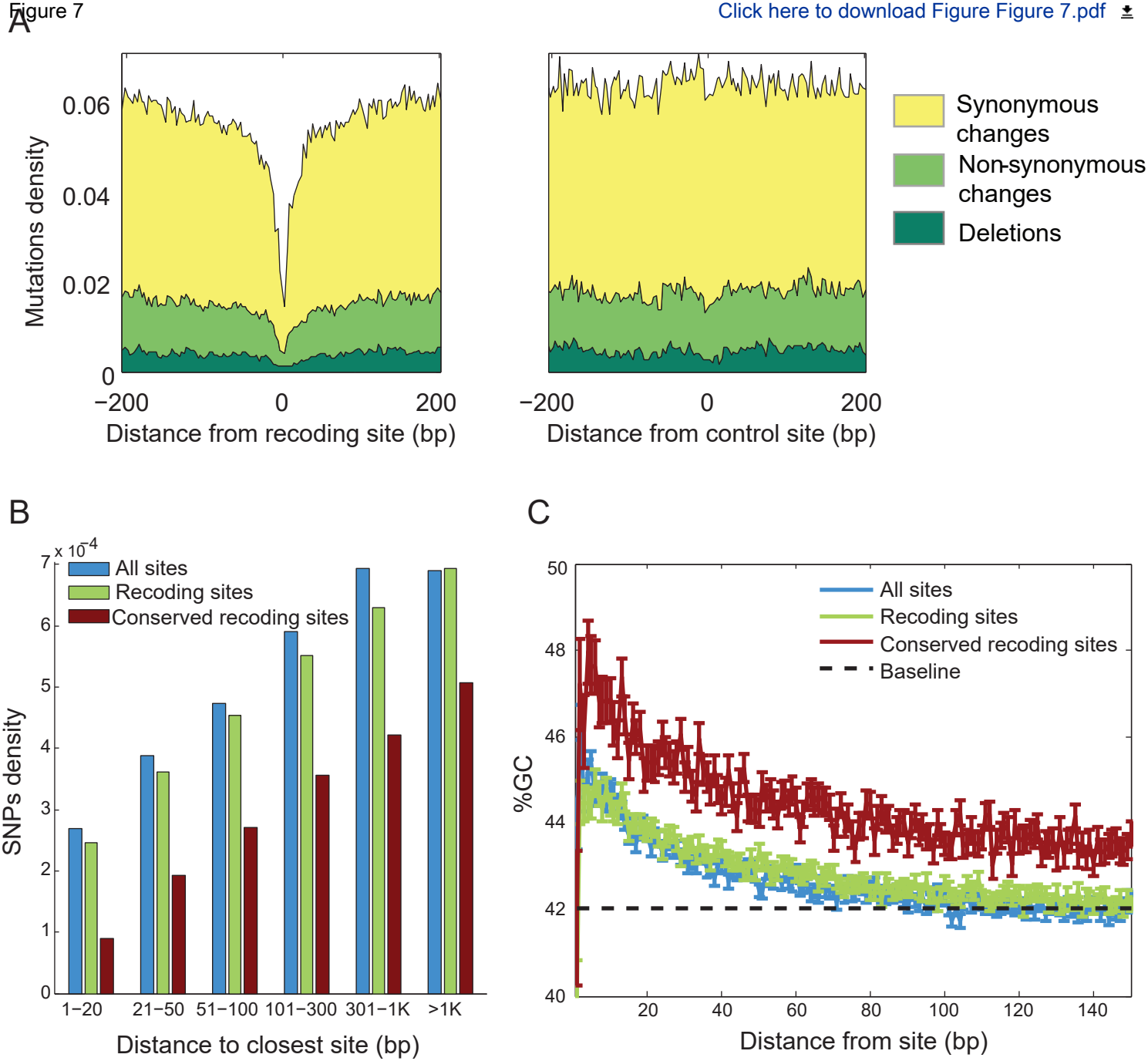


Figure 7

Click here to download Figure Figure 7.pdf



Supplemental Tables

Table S1. Sequencing data summary, related to STAR methods

Sequencing data summary for the RNA-seq and DNA-seq results.

Animal	Type	tissue	paired/single	#reads	read length	%GC
Oct.vul.	RNA-seq	OL	paired	127049529	151x2	39
Oct.vul.	RNA-seq	SG	paired	126573775	151x2	39
Sepia	RNA-seq	OL	paired	159933775	151x2	39
Sepia	RNA-seq	SG	paired	131031103	151x2	40
Nautilus	RNA-seq	OL	paired	188232900	151x2	41
Nautilus	RNA-seq	Supra	paired	178490917	151x2	41
Oct.vul.	DNA-seq	Sperm sac	paired	498212805	101x2	35
Sepia	DNA-seq	Sperm sac	paired	500285022	101x2	33
Nautilus	DNA-seq	Sperm sac	paired	503446649	101x2	35

Table S2. Transcriptome assembly and ORF statistics, related to STAR methods

Transcriptome assembly statistics, and ORF statistics (based on the longest isoform found for each Trinity "gene"), for the different species.

	Squid	Sepia	Oct.vul.	Oct.bim.	Nautilus	Aplysia
Total Trinity 'transcripts'	151477	240756	201414	271576	458827	256514
Total Trinity 'genes'	107025	168878	150616	207439	282627	196908
Contig N50	755	687	784	749	992	1802
Median contig length	334	323	334	340	352	387
Mean contig length	573.65	543.02	579.98	574.31	662.03	869.75
Total assembled bases	61395149	91704707	87354553	119133693	187107820	223102930
Number of ORFs	12218	14954	13005	17786	11139	22951
Number of unique proteins	9385	10967	10218	12852	8485	17283
Mean ORF length	1369	1155	1317	1253	1392	1147
Median ORF length	1026	780	969	882	999	663
Total ORF length	16725027	17272626	17129859	22292061	15501153	26321973

Table S3. Extraordinary extent of A-to-I RNA editing in coleoid cephalopods, related to Figure 1.

In all species but Oct.bim., the majority of ORFs found contained recoding events. The higher number of ORFs detected and the somewhat lower fraction of edited ORFs in Octopus bim. is due to the different composition of sub-tissues used for transcriptome assembly (OL, sub, supra and ANC, compared with OL and SG for the other species). Re-assembling a transcriptome using only reads derived from the OL sample of Octopus bim. one finds recoding fraction consistent with the other species. This suggests a considerable variation of editing level across neural sub-tissues.

Animal	***Tissues used***	***# AG sites***	***#recoding sites***	***# ORFs found***	***#ORFs recoded***
Oct.vul.	*OL, SG*	*117842*	*76639 (65%)*	13005	*7953 (61%)*
Oct.bim.	*OL, ANC, sub, supra*	*76862*	*50079 (65%)*	17786	*6964 (39%)*
Oct.bim.	*OL*	*74436*	*48181 (65%)*	10974	*6717 (61%)*
sepia	*OL, SG*	*130636*	*86230 (66%)*	14954	*8537 (57%)*
squid	*OL, GFL*	*82975*	*54287 (65%)*	12218	*6688 (55%)*

Table S8. Multiply recoded proteins, related to Figure 4

List of 24 proteins containing 5 or more highly edited (>=10% editing levels) recoding sites that are shared by all 4 coleoid cephalopod species.

UniProt ID	Protein name	# highly edited, 4-species conserved, recoding sites
O54774	AP-3 complex subunit delta-1	5
O60271	C-Jun-amino-terminal kinase-interacting protein 4	5
O95359	Transforming acidic coiled-coil-containing protein 2	6
P09482	Neuronal acetylcholine receptor subunit alpha-4	5
P13395	Spectrin alpha chain	8
P23468	Receptor-type tyrosine-protein phosphatase delta	8
P41823	Synaptotagmin-1	6
P55162	Membrane-associated protein Hem	6
Q01484	Ankyrin-2	8
Q4KM31	LIM domain-containing protein 2	5
Q5U239	Transmembrane protein 145	5
Q6GLR7	Calcium-dependent secretion activator 1	7
Q6GPD0	Rho GTPase-activating protein 32	5
Q6VNB8	WD repeat and FYVE domain-containing protein 3	8
Q6ZPF3	T-lymphoma invasion and metastasis-inducing protein 2	6
Q7Z3G6	Prickle-like protein 2	5
Q80U22	Iporin	5
Q862Z3	Uromodulin	14
Q8N2Q7	Neuroligin-1	5
Q96RL7	Vacuolar protein sorting-associated protein 13A	12
Q9I8C7	Neuronal acetylcholine receptor subunit alpha-10	7
Q9I8D1	Unconventional myosin-VI	9
Q9NGC3	Centaurin-gamma-1A	9

Table S9. Predicted editing sites in squid, sepia and octopus vulgaris Kv2 channel, related to Figure 6.

List of editing sites for cephalopod Kv2.1 orthologs. Numbers in parentheses refer to specific Trinity assembly from our transcriptome. Fractional editing is based on estimates from our pipeline using OL and SG. These are the same channels as those in Fig. 3 and Fig. S10.

Sepia Kv2 (comp267321_c0_seq19)					Octopus Kv2 (comp182915_c0_seq17)					Squid Kv2 (comp141576_c1_seq4)			
nt	unedited codon	edited codon	Fractional Editing		nt	unedited codon	edited codon	Fractional Editing		nt	unedited codon	edited codon	Fractional Editing
1273	S	G	0.064		672	Q	R	0.022		1221	Q	R	0.062
2612	Y	C	0.781		765	H	R	0.517		1225	Q	Q	0.082
2352	Q	Q	0.256		789	H	R	0.869		1227	Q	R	0.116
2909	K	R	0.095		802	P	P	0.277		1231	Q	Q	0.068
2371	I	V	0.081		808	L	L	0.292		1320	H	R	0.525
2786	N	S	0.05		824	T	A	0.031		1332	H	R	0.065
2782	R	G	0.112		845	T	A	0.044		1341	H	R	0.267
2386	K	E	0.068		952	A	A	0.017		1344	Q	R	0.069
1142	H	R	0.063		961	A	A	0.051		1354	P	P	0.114
1329	P	P	0.356		994	A	A	0.035		1357	P	P	0.677
3921	L	L	0.154		1028	R	G	0.914		1376	T	A	0.104
1133	H	R	0.606		1046	S	G	0.142		1444	E	E	0.482
2235	K	K	0.101		1055	S	G	0.11		1460	S	G	0.218
1351	S	G	0.254		1069	V	V	0.103		1513	P	P	0.5
1789	M	V	0.083		1386	Y	C	0.028		1535	S	G	0.321
1040	Q	R	0.075		1418	M	V	0.035		1604	I	V	0.2
1148	H	R	0.104		1477	L	L	0.024		1616	S	G	0.882
1167	P	P	0.026		1479	E	G	0.019		1660	Q	Q	0.121
3179	K	R	0.069		1685	I	V	0.072		1726	R	R	0.182
2500	S	G	0.082		1753	E	E	0.036		1729	E	E	0.603
3705	I	M	0.095		1981	Q	Q	0.132		1737	K	R	0.037
1257	E	E	0.146		1982	I	V	0.382		1767	K	R	0.037
2362	I	V	0.214		2116	I	M	0.046		1773	E	G	0.053
1476	Q	Q	0.061		2241	Y	C	0.374		1799	M	V	0.07
2493	L	L	0.06		2303	I	V	0.704		1816	L	L	0.032
2780	K	R	0.056		2411	R	G	0.031		1827	K	R	0.029
2385	L	L	0.049		2539	K	K	0.065		1973	M	V	0.198
1157	Q	R	0.221		2621	N	D	0.016		1995	E	G	0.12
1542	R	R	0.085		2765	N	D	0.077		2114	R	G	0.463
1189	T	A	0.026		2771	M	V	0.098		2115	E	G	0.302
1237	K	E	0.083		2788	K	K	0.057		2240	I	V	0.921
1096	T	A	0.078		3042	K	R	0.611		2296	E	E	0.047
2056	I	V	0.459		3206	R	G	0.25		2338	V	V	0.041
1545	E	E	0.514		3227	S	G	0.091		2419	K	K	0.078
2644	S	G	0.134							2527	R	R	0.041
1044	Q	Q	0.066							2535	Q	R	0.038
2796	I	M	0.051							2536	Q	Q	0.066
2670	L	L	0.079							2537	I	V	0.831
3928	R	G	0.167							2545	R	R	0.039
1930	K	E	0.23							2648	M	V	0.1
1615	M	V	0.024							2818	K	K	0.033
2674	I	V	0.883							2828	S	G	0.385
1170	P	P	0.117							2854	L	L	0.074
1931	K	R	0.128							2858	I	V	0.979
3004	K	E	0.044							2980	I	M	0.03
1154	H	R	0.164							3093	K	R	0.305
1118	H	R	0.056							3240	E	G	0.038
1192	T	A	0.053							3342	N	S	0.033
2487	I	M	0.054							3971	S	G	0.054
1811	E	G	0.031							4154	S	G	0.062
1432	S	G	0.842										
1038	Q	Q	0.088										
1059	V	V	0.036										
1274	N	S	0.688										
1411	S	G	0.242										

Figure S1

Click here to download Supplemental Figure Figure S1.pdf

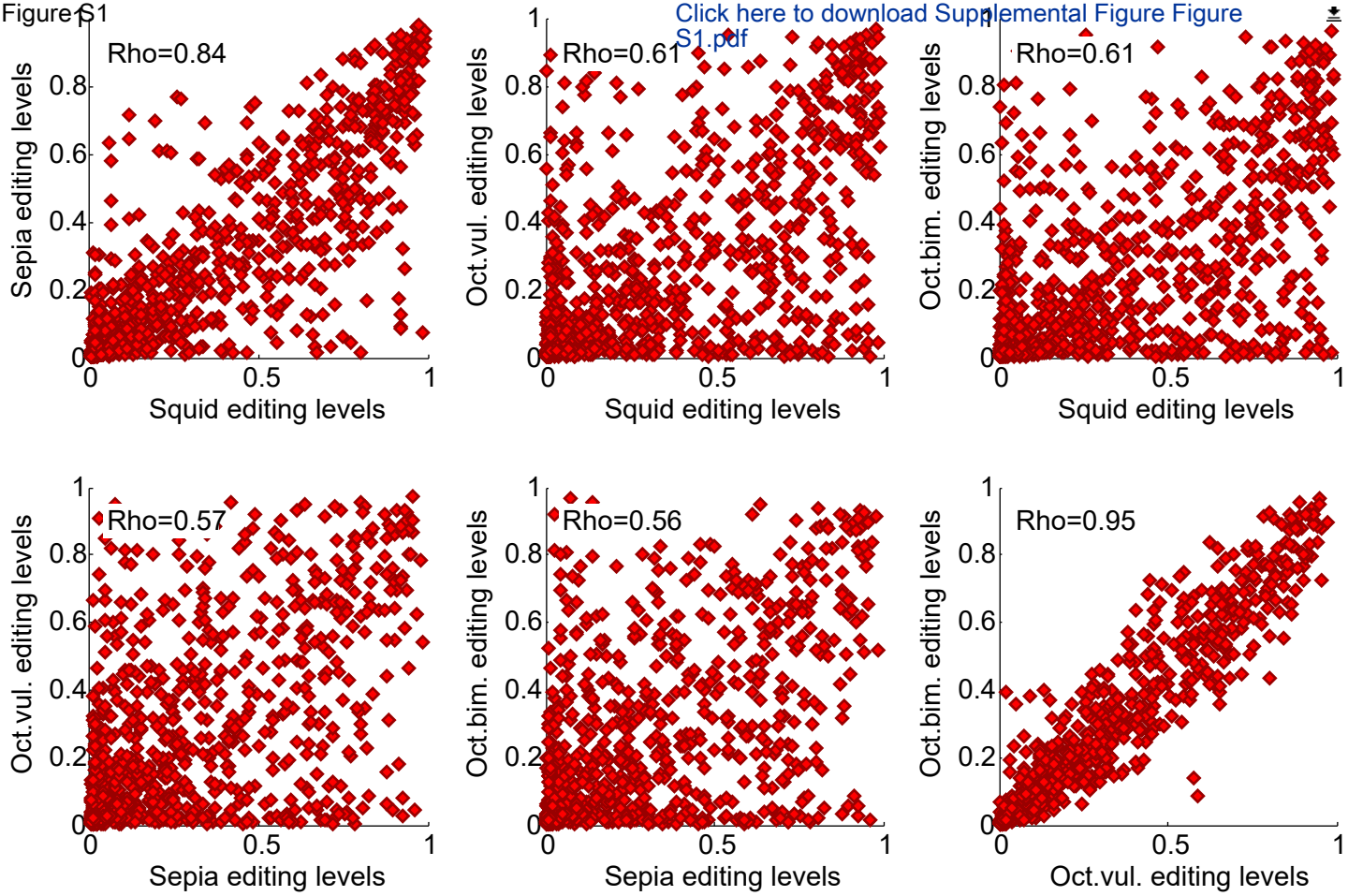


Figure S2

Click here to download Supplemental Figure Figure S2.pdf

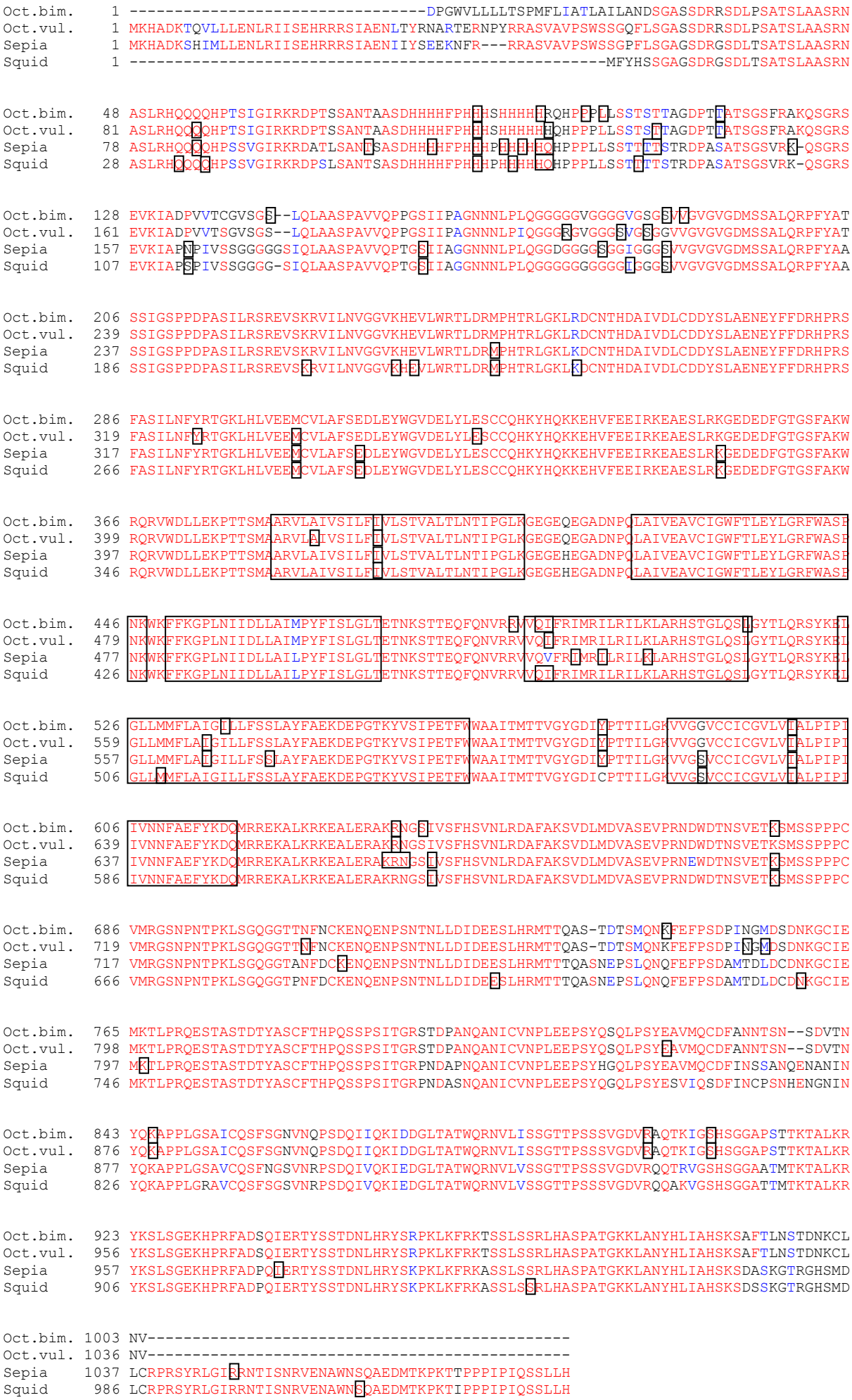


Figure S3

Click here to download Supplemental Figure Figure S3.pdf

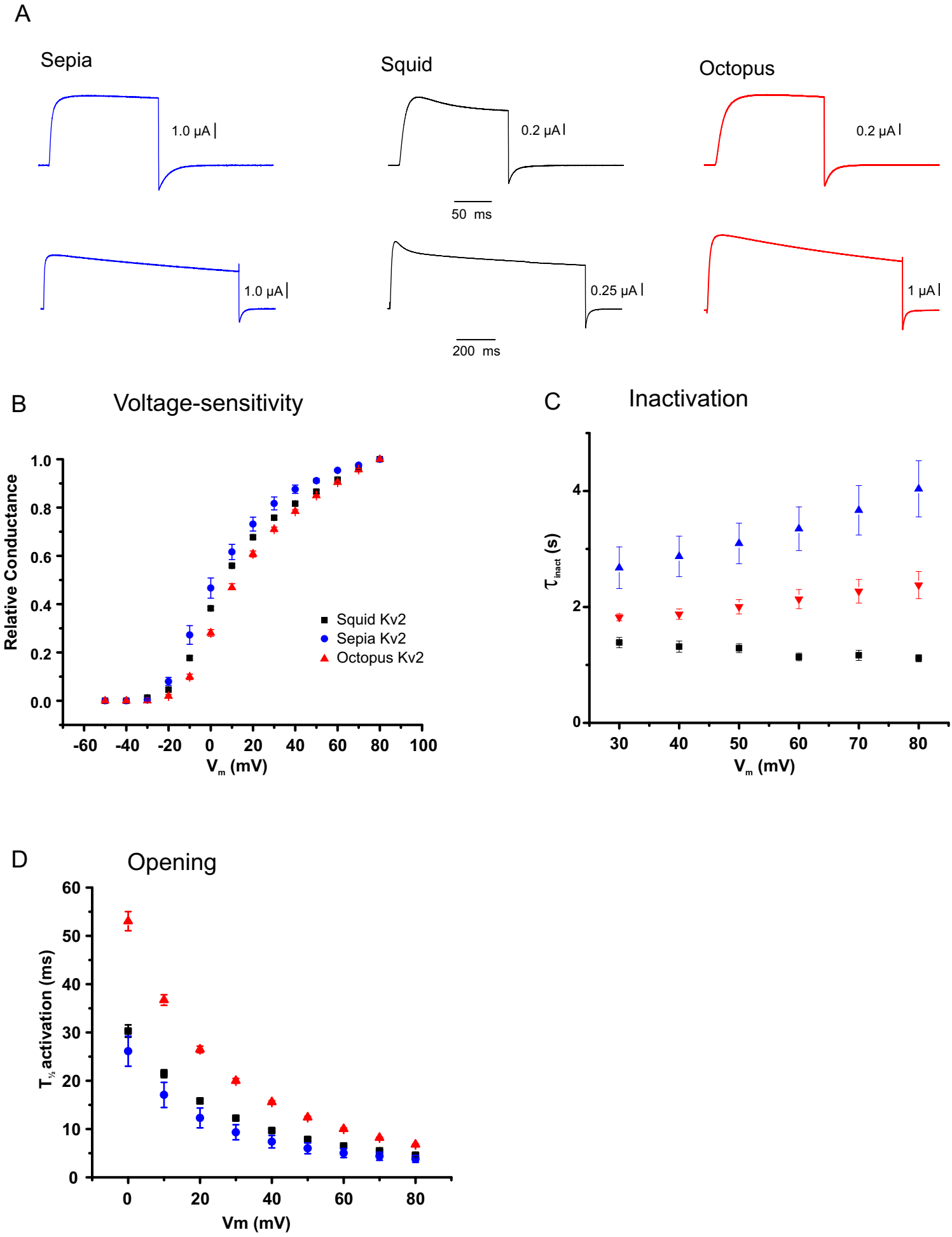


Figure S4

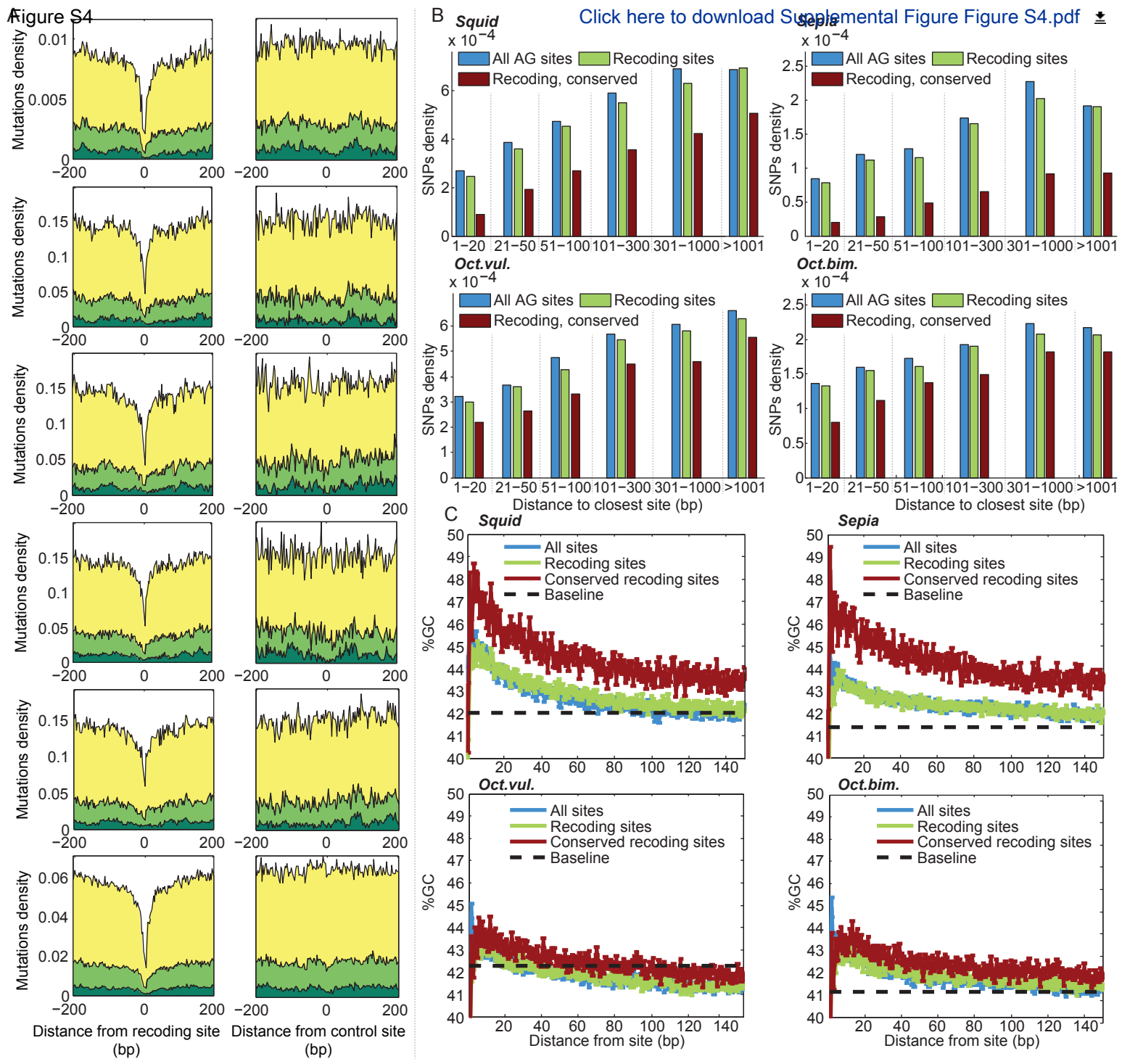


Figure S5

Click here to download Supplemental Figure Figure S5.pdf