Evolutionary Analyses of RNA Editing and Amino Acid Recoding in Cephalopods

Mingye (Christina) Wang and Erik Møhlhenrich

Princeton International School of Mathematics and Science, Spring 2019

Abstract - RNA editing is a post-transcriptional modification process that alters nucleotides of mRNA and consequently the amino acids of the translated protein without changing the original DNA sequence. In human and other mammals, amino acid recoding from RNA editing is rare, and most edits are non-adaptive and provide no fitness advantage (1). However, recently it was discovered that in soft-bodied cephalopods, which are exceptionally intelligent and include squid, octopus, and cuttlefish, RNA editing is widespread and positively selected (2). To examine the effects of RNA editing on individual genes, we developed a “diversity score” system that quantitatively assesses the amount of diversity generated in each gene, incorporating combinatorial diversity and the radicalness of amino acid changes. Using this metric, we compiled a list of top 100 genes across the cephalopod species that are most diversified by RNA editing. This list of candidate genes provides directions for future research into the specific functional impact of RNA editing in terms of protein structure and cellular function on individual proteins. Additionally, considering the connection of RNA editing to the nervous system, and the exceptional intelligence of cephalopod, the candidate genes may shed light to the molecular development of behavioral complexity and intelligence. To further investigate global influences of RNA editing on the transcriptome, we investigated changes in nucleotide composition and codon usage biases in edited genes and coleoid transcriptome in general. Results show that these features indeed correlate with editing and may correspond to causes or effects of RNA editing. In addition, we characterized the unusual RNA editing in cephalopods by analyzing ratio of radical to conservative amino acid substitutions (R/C) and distribution of amino acid recoding from editing. Our results show that compared to model organisms, editing in cephalopods have significantly decreased R/C ratio and distinct distribution of amino acid substitutions that favor conservative over radical changes, indicating selection at the amino acid level and providing a potential mechanism for the evolution of widespread RNA editing in cephalopods.

TABLE OF CONTENTS

1. INTRODUCTION
2. METHODS
3. RESULTS
4. DISCUSSION
5. ACKNOWLEDGEMENTS
6. REFERENCES
7. SUPPLEMENTARY MATERIALS

INTRODUCTION

According to the central dogma, DNA is transcribed into messenger RNA, which is translated into protein. In this process, there is a direct correspondence between the DNA sequence and the amino acid sequence. One exception to the central dogma is direct RNA editing, which alters the nucleotides of mRNA and potentially the amino acids of the translated protein without changing the original DNA sequence. A-to-I editing (herein simply referred to as RNA editing) is the most common type of RNA editing in vertebrates, in which adenosine is deaminated to form inosine, which behaves as guanine in translation (3). This process is carried out by the enzyme ADAR (Adenosine Deaminase Acting on RNA), which binds to double-stranded RNA with preferred sequences. Overall, RNA editing is rare in mammals. In humans, only about 3% of mRNA are edited, of which only 25 editing sites are conserved across mammals (2). Studies have shown that the vast majority of edits in human are non-adaptive and likely the result of promiscuous editing by ADAR (4). Furthermore, most editing occurs in non-coding transcripts encoding Alu repeat elements.

In contrast to mammals, a 2017 study demonstrated that mRNA is extensively edited in coleoids, the group of soft-bodied cephalopods that include squid, octopus, and cuttlefish (2). In *Octopus bimaculoides*, more than 900,000 editing sites were found by analyzing RNA from 12 different tissues, of which 12% are in coding regions. Furthermore, 65% of these editing sites in coding region are non-synonymous (resulting in an amino acid substitution, known as protein recoding), and conserved sites tend to have higher editing levels and a higher proportion of non-synonymous substitutions. In total, 1146 editing sites were found to be conserved over all four coleoid species

studied (*Octopus vulgaris*, *Octopus bimaculoides*, *Doryteuthis pealeii*, *Sepia officinalis*), which diverged 350-480 mya, in sharp contrast with the 25 transcripts conserved across mammals, which diverged only 150-200 mya (5).

Previous works have shown that RNA editing is crucial in regulating the nervous system in mammals. Mammalian ADARs have the highest expression level in the central nervous system (1). Although protein recoding by RNA editing in mammals is rare, several of the few recoding editing events involve proteins related to the nervous system, including glutamate receptor subunit GluR2 and potassium channel Kv1.1. Deletion of ADAR2 in mice, for example, cause frequent epileptic seizures due to neuronal death, and the effect seems to be due to the lack of only one essential edit (6). The connection between RNA editing nervous system is also discovered in cephalopods and is even more pronounced (2). Protein recoding (editing at non-synonymous sites) from RNA editing in coleoids is most significant in the nervous system. In *Octopus bimaculoides*, editing in neural tissues is 2-fold higher and protein recoding events are more frequent. In addition, the closely related nautilus species, which have simpler brains and are behaviorally less complex than coleoids, have significantly fewer RNA editing events, while coleoids with extensive RNA editing have the largest nervous system in invertebrates and are known for their exceptional intelligence (2). Together, these discoveries suggest an important role for protein recoding by RNA editing in neural and behavioral complexity.

To better understand the effects of RNA editing in cephalopods, we developed a quantitative metric to identify edited genes that are most impacted by RNA editing. One benefit of RNA editing over regular DNA mutations is the flexibility it allows for expression of different versions of a transcript at different time periods, potentially in response to shifting environmental conditions, thus offering another level of temporal regulatory control (6). Another major benefit of RNA editing is the diversity it creates: since not all mRNA transcripts may be edited, this allows different versions of the same sequence (both edited and unedited) to be expressed in the same cells (2). In fact, for genes with multiple editing sites, if the combinatorial editing is realized, the number of different variants for each encoded protein would increase exponentially as the number of editing sites within a gene increases (7). Therefore, we aimed to quantify this diversity and use it as the main evaluation for identifying candidate genes by developing a quantitative measure we named “diversity score” that ranks all genes containing at least one editing site by the level of protein diversity generated by RNA editing. The main factors that affect the amount of molecular diversity include the editing level and the specific amino acid change that is caused by the edited site. Editing level measures the relative amount of RNA transcripts that are edited, and vary from nearly 0% to

100%. A value near 50% achieves maximal diversity, as both types of transcripts are expressed evenly inside the cells, similar to the concept of species evenness for ecological systems (e.g. in Shannon diversity index), and is thus taken into account when assessing diversity (8). For the amino acid change, two aspects were considered in the diversity score. Chemically, amino acid substitutions between two dissimilar amino acids (i.e. a radical substitution, such as when a positively charged amino acid is substituted for a negatively charged amino acid) are more likely to impact the proteins significantly in structure and function than substitutions between two chemically similar amino acids (i.e. a conservative substitution), generating higher diversity (9). Therefore, chemical properties of the amino acid change caused by an editing site are taken into account in the diversity score.

Although physicochemically radical substitutions are more likely to impact protein structure and function on average, this may not be true at all residues in a protein. Conversely, the constraints on a particular residue may cause even conservative substitutions to have substantial effects. For example, a conservative substitution from arginine to histidine in the ATPase ATP7A causes Menkes disease, a fatal illness in infants (10). The constraints at a specific residue can be accounted for by incorporating evolutionary information - if having a certain amino acid at a site has a damaging effect on the protein structure and function, then it is expected that the particular amino acid would be very rare across matching protein sequences of numerous species, as natural selection tends to eliminate deleterious substitutions (11).

Here, we incorporate editing level with measures of evolutionary and physicochemical properties of amino acid substitutions to develop a protein diversity score that quantitatively assesses the amount of combinatorial diversity generated by RNA editing. Using this metric, we calculate diversity scores for each gene in the four cephalopod species and compile a list of top 100 genes across the cephalopods. We further analyze this list of candidate genes with gene ontology enrichment, disease annotations, protein-protein interaction networks, and comparisons to genes edited in model organisms. These results highlight connections between RNA editing and nervous system development, and provide a list of candidate genes for further studies into the role of RNA editing in neural and behavioral complexity in cephalopods.

In addition to specific molecular effects, widespread RNA editing may also lead to global changes in the transcriptome. For example, given that RNA editing in cephalopods acts on adenosine, one might expect the genome content of adenosine is increased relative to related species. Previously it has been shown that regions around editing sites (up to 100 nucleotides on each side) display elevated GC content, which facilitate formation of stable double-stranded regions required by ADAR enzymes to bind

and initiate editing (2). We therefore might expect increases in overall GC content in cephalopod species. Besides nucleotide composition, another potential consequence of widespread RNA editing is changes in codon usage bias, in which certain codons are used at higher frequencies than other synonymous codons (12). Codon usage bias has various influences on processes related to gene expression—for example, codon usage of highly expressed genes is selected to increase translation efficiency (13). Within the context of RNA editing in cephalopods, given synonymous codons that code for the identical amino acid, we might expect the codon containing a “A” to be preferred since it enables further editing that may diversify the protein. Conversely, the A-containing codon may also be actively avoided at crucial positions of a gene to prevent any editing. These changes in nucleotide composition and codon usage biases may affect many other fundamental processes, for example mutational biases and translation efficiency, adding to the cost-benefit trade-off of widespread RNA editing, in addition to the already discovered reduction of polymorphisms caused by the maintenance of necessary structures surrounding editing sites (2).

One major question arising from the original discovery of widespread RNA editing in cephalopods is the cause of its evolution in cephalopods and not other species. We hypothesize that overall genomic characteristics may play a part—initial changes in cephalopod genome may have facilitated RNA editing, which in turn led to more biases towards genomic content that favors RNA editing. Investigation into nucleotide composition and codon usage bias in cephalopods would thus not only address the effects of widespread RNA editing on global transcriptome characteristics, but also help elucidate the factors that led to the evolution of the process itself.

Therefore, we examined and compared nucleotide content and codon usage bias in coleoids with nautilus and aplysia. By building predictive models of editing, we show that nucleotide composition and codon usages indeed contain information on editing. Further analysis showed that contrary to expectation, edited genes do not appear to have different nucleotide content (specifically proportion of adenosine and GC) than non-edited genes, while coleoids in general do have higher levels of A, but lower levels of GC compared to nautilus and aplysia. Meanwhile, several codons show biases that follow expected trends from RNA editing and may warrant further attention.

Lastly, we analyzed the amino acid recoding from RNA editing to gain insights into the strength of natural selection acting on RNA editing and the potential causes of widespread editing. Previously it has already been shown that nonsynonymous edits are suppressed in mammals, one evidence that the majority of editing in mammals is non-adaptive and not under selective pressure, likely results of promiscuous editing by the ADAR enzyme (4). In cephalopods, nonsynonymous-to-synonymous ratio (N/S) is

neutral overall, but the ratio increases as editing levels increase, indicating positive selection for highly edited sites (2). These N/S analysis are focused on the nucleotide level, and therefore we investigated whether if similar pressures act on the amino acid level by calculating the ratio of radical-to-conservative amino acid substitutions (R/C). A lower R/C ratio indicates stronger purifying selection, since radical substitutions are more likely to be deleterious and actively removed when selection is strong (9). Our results show that cephalopods editing have lower R/C ratios compared to human, mice, and fly, indicating strong purifying selection in addition to the positive selection previously discovered. To further characterize the amino acid recoding, we examined the distribution of amino acid substitutions by RNA editing, and cephalopods show a distinct pattern that more strongly bias conservative substitutions over radical substitutions, compared to patterns in human, mice, and fly. These results provide insights into the amino acid recoding from RNA editing and the selection pressure acting on it, and offer potential clues to the evolution of widespread RNA editing in cephalopods.

METHODS

I. Transcriptome and Editing Data Acquisition

Transcriptome and editing data of four coleoid species (*Octopus vulgaris*, *Octopus bimaculoides*, *Doryteuthis pealeii*, and *Sepia officinalis*; hereafter abbreviated as oct vul, oct bim, squid, and sepia), and transcriptome data of a nautiloid (*Nautilus pompilius*; hereafter referred to as nautilus) and a gastroid mollusk (*Aplysia californica*; hereafter referred to as aplysia) were acquired from Liscovitch-Brauer et al., 2017 (see Supp. I for a sample section of the dataset). Transcriptome data of the four coleoid species were modified to reflect unedited sequences by matching editing sites with editing level greater or equal to 50% to the transcriptome nucleotide and mutating it from G to A. In total, 117842 sites in oct vul, 76862 sites in oct bim, 82975 editing sites in squid, and 130636 sites in sepia were used. Editing data of mouse (mm9) and fly (dm3) were acquired from the Rigorously Annotated Database of A-to-I RNA Editing (RADAR) (14). Editing data of human was acquired from both RADAR and REDportal from the GTEx project (15).

II. General Data Processing

Data cleaning, transformations, and calculations were performed using the online bioinformatics platform Galaxy (16) and other tools including RStudio, Unix shell, and Microsoft Excel. The software package EMBOSS was used for sequence translation (17), and the software package CAT was used for nucleotide composition and codon usage analysis (12).

III. Diversity Score Calculation

The diversity score is composed of the editing level score and the amino acid score. Diversity score of a gene is calculated as the product of the editing level score and the amino acid score summed over all editing sites in the gene.

Editing level score of each editing site is calculated as to reflect the larger diversity achieved with editing values close to 50%:

$$\text{Editing level score} = 2 \times (1 - \text{editing level}) \times \text{editing level}$$

In this way, the diversity score of one gene represents the mean distance, measured by the sum of amino acid scores per editing site, between two randomly selected transcripts, each containing a portion of the amino acids substituted, with distribution as according to the editing level. Two issues arising from data limitation should be noted: first, the distribution used in this method assumes all editing sites to be independent, while in reality not all combinatorial diversity may be achieved in the actual cellular environment: previous studies have shown that certain editing sites are correlated (18), and therefore some transcripts may not be realized or occur at a lower frequency than expected. Second, the editing level data was pooled over multiple tissues, and therefore it is possible that only certain sites are edited in one tissue, so the actual diversity is lower than expected. Thus, the diversity score calculated here represents only a theoretical maximum diversity that may be achieved.

Two different amino acid scores were calculated, one focusing on the general chemical effect of the substitution, and the other incorporating site-specific evolutionary information. The universal evolutionary index (UEI) was used as a proxy for the physicochemical significance of an amino acid substitution (19). UEI was derived from comparisons of sequences across a large number of proteins in various species, and describes the likelihood that one amino acid is substituted for another across evolutionary time. Since more chemically radical substitutions are on average more deleterious than conservative substitutions and more likely to be removed by natural selection (9), UEI scores serve as an indication of the general radicalness of an amino acid substitution chemically. The other amino acid score utilizes PSSM scores (Position Specific Scoring Matrix) from delta-BLAST, which searches in a database of sequences from various species to find sequences that are similar to the edited protein in cephalopods (20). PSSM is position specific, providing different scores for the same amino acid substitution at different positions. Higher scores correspond to more frequent amino acids (i.e. many species have the particular amino acid at the particular site of the protein) and often indicate critical functional residues. Editing that produces two amino acids with highly disparate PSSM scores (i.e. one

amino acid has a very high score, indicating it is commonly found at this location in other species, while the other amino acid has a low score, indicating it is rarely found) indicates the coexistence of a highly conserved, canonical form of the protein, as well as a novel form, and thus was taken to be highly diverse in the diversity score.

Numerically, the UEI score is defined as the reciprocal of the UEI value because in the original formulation of the UEI, the higher the value, the more conservative or chemically similar an amino acid change. The PSSM score is defined as the absolute value of the difference between the scores for the edited amino acid and the unedited version. PSSM values were calculated by searching cephalopod protein sequences against the CDD delta database with E-value = 1e-5 as the threshold for matches. Diversity score-UEI is defined as *editing level score* $\times$ *UEI score*, summed over all editing sites per gene. Similarly, diversity score-PSSM is defined as *editing level score* $\times$ *PSSM score* summed over all sites. An additional measure we named “novelty score” was calculated as the product of the absolute editing level and the directional difference of the PSSM scores (i.e. PSSM score of the edited amino acid subtracted by the score of the original amino acid, taken the negative to account for the sign, and if negative is replaced by 0 since an edit should not diminish novelty) summed over all sites. This score identifies sites that approximate a novel mutation, where the high editing level indicates strong selection for the edited version of the protein, and the large directional difference in the PSSM score (i.e. the unedited amino acid has a high score and the edited amino acid has a low score) indicates the edit transforms a common, canonical form of the protein into a rare, novel form.

Using these measures, the diversity scores (UEI, PSSM, and novelty) of all edited genes in each of the 4 cephalopod species were calculated. The scores are combined across the 4 species: the genes are matched across the cephalopod species by using their matches to human genes (Swiss Prot) and the diversity scores from each species were averaged. Due to the existence of paralogs, multiple genes in one cephalopod species may be mapped to a single human gene. In this case, the mean score of all possible combinations was taken. Since UEI and PSSM diversity scores differ in numerical scales, these two scores are scaled by each measure using the maximum UEI or PSSM score in the entire data set and summed together to generate an overall ranking of edited genes (which we refer to as the “overall” gene list). In addition, novelty scores of individual editing sites in each species, as well as conserved sites across all 4 species, were used to rank the sites for specific editing site candidates.

IV. Functional Analyses of Edited Genes

The online database STRING was used to generate protein-protein association networks of candidate genes (21). Experiments, databases, co-expression, neighborhood, gene fusion, and co-occurrence were used as sources of interactions, and medium confidence (0.400) was used for the minimum required interaction score. Interaction networks were generated for the top 100 and 200 candidate genes overall and were then analyzed to identify highly connected genes we call “hub” genes. Specifically, genes that connected to at least 5 genes in the top 100 list, and at least 10 in the top 200 list, were highlighted as hub genes.

A set of genes associated with human intelligence was acquired from a GWAS study (22). This intelligence gene set was compared to genes edited in cephalopods, and candidate genes that are included in this gene set are highlighted.

Gene ontology enrichment results were calculated using the online tool GOrilla (23) with a single list of genes ranked by diversity score, with p-value threshold = 1E-3.

Disease annotations for edited genes were acquired from the curated gene-disease associations dataset on DisGeNET (24). A cutoff of 0.6 was used for filtering out weak associations. Disease enrichment was calculated using the minimum hypergeometric score method used in gene ontology enrichment calculations in GOrilla with a single list of genes ranked by diversity score (23). Diseases associated with fewer than 10 genes in the entire dataset were removed.

For analyses requiring gene names, UniProt IDs were mapped to gene names using the UniProt website (<https://www.uniprot.org/mapping/>).

V. Codon Usage Bias Analysis

Nucleotide composition and codon usage analysis were performed on the four coleoid species (oct vul, oct bim, squid, and sepia) and two outgroups (nautilus, aplysia) using the software package CAT with translated transcriptome data (12). Edited genes from the four coleoid species were extracted and separate codon usage bias values were calculated.

To evaluate the hypothesis that codon usage bias profiles of edited genes are distinct from unedited genes, machine learning algorithms were used to predict a) whether if a gene is edited and b) cumulative editing level normalized by length, using observed levels of all 61 codons (excluding 3 stop codons), nucleotide composition (GC and AG proportions overall and at positions 1, 2, 3 in codon), and codon deviation coefficient (CDC), an overall measure of codon usage bias calculated by CAT, for a total of 70 features. For both tasks, transcriptome data from only *Octopus bimaculoides* were used to avoid potential interference from global differences between species. The data was split into 75% for training and 25% for testing. For

classifying if a gene is edited, a simple logistic regression model (R package glm) was enough to achieve substantial accuracy, while for predicting editing level, random forest was used (R package randomForest). Both are implemented through the R package caret (for random forest, RMSE from bootstrapping was used to select hyperparameter mtry = 5).

Next, to further investigate the correlation between RNA editing and nucleotide composition and codon usages, we compared nucleotide content and specific codon usage biases between a) edited genes and transcriptome average, and b) coleoid species and nautiloids (the other cephalopod group that does not show high levels of RNA editing and is less behaviorally complex) and aplysia (a more distantly-related species). Specifically, we calculated and compared mean proportions of adenosine and GC content by gene. For codon usage, to account for the expected level of different codons, we calculated adjusted bias of a specific codon per gene as:

$$g = \text{adjusted bias per gene} = \frac{\text{observed} - \text{expected}}{\text{expected}}$$

adjusted bias of a specific codon per species as:

$$\text{adjusted bias per species} = N_{10\%} - N_{-10\%}$$

where $N_{10\%}$ = number of genes with $g > 10\%$,
and $N_{-10\%}$ = number of genes with $g < -10\%$

Within the four coleoid species, adjusted bias per species values were compared between edited genes (i.e. genes containing at least one editing site) and the entire transcriptome of the corresponding species as a proxy to examine the differences between edited and unedited genes. Next, adjusted bias per species values were compared between the four coleoid species and the two outgroups (nautilus and aplysia) to investigate any difference in codon usage bias that evolved during the coleoid lineage.

VI. R/C Ratio and Amino Acid Recoding

To assess the extent of purifying selection on RNA editing in cephalopods, we calculated R/C ratios of amino acid recoding that result from editing in the four cephalopod species, as well as edits conserved in cephalopods, and edits in human, mice, and fly for comparison (using editing site data from RADAR). We classified amino acids into different groups using on 3 different metrics (from Hanada 2007 classifications A, B, and C; see Table I for an example), and counted a substitution as conservative if the change occurs within a category, and radical if the change crosses categories. For each metric, the number of radical recodings and conservative recodings are counted and their ratio is calculated. These ratios are averaged across the metrics to acquire the final R/C values for comparison.

To gain a better understanding of the R/C results, we analyzed the distribution of specific amino acid substitutions from RNA editing. To account for inherent differences caused by the structure of the genetic code, we calculated the expected frequency of an amino acid recoding

by finding the proportion of all possible amino acid changes resulting from a single A-to-I (or to G) change in a codon that cause the specific amino acid recoding. Observed frequencies of amino acid recoding by RNA editing are then compared to these expected values, and fold differences were used to compare between different amino acid substitutions and across species.

Category	Amino Acids
Acidic	D E
Neutral and no aromaticity	Q A V L I C S T N G P M
Neutral and aromaticity	F Y W
Basic	K R H

TABLE I
EXAMPLE CLASSIFICATION SCHEME OF AMINO ACIDS BY CHARGE AND AROMATICITY

RESULTS

I. Diversity Score

Validation of Diversity Score Metrics

To validate that the diversity score accounts for both editing level and amino acid substitutions, without one aspect being the dominating factor that determines the diversity score, we compared the diversity scores per species with (i.e. the regular diversity scores) or without (i.e. only the editing level score) the amino acid score. Since the primary function of diversity score is to identify top candidate genes, we focused the comparison on the number of overlaps in the top 100 gene list identified by scores. In general, for the 4 cephalopod species and 3 metrics (UEI, PSSM, and novelty), the proportion of overlap is approximately 50~60% (Table II), indicating that the amino acid score and the editing level each contribute to about half of the overall diversity score.

	Oct vul	Oct bim	Squid	Sepia
UEI	61	61	56	59
PSSM	67	57	59	55
Novelty	56	48	54	51

TABLE II
EFFECT OF AMINO ACID ON OVERALL DIVERSITY SCORE METRICS: EACH CELL COUNTS THE NUMBER OF OVERLAPPED GENES IN THE TOP 100 GENE LISTS IDENTIFIED USING THE SPECIFIED DIVERSITY SCORE IN EACH SPECIES WITH OR

WITHOUT THE AMINO ACID SCORE COMPONENT.

Top 100 Candidate Gene List

Table III shows the first 5 genes of the top 100 list of candidate genes as identified by the aggregate diversity score using both UEI and PSSM scores (see Supp. II for the all top 100 genes). Genes are noted for being highly connected in protein-protein interaction networks ("hub genes"), containing at least one nonsynonymous editing site in other model species (human, mice, or fruit fly), or associated with human intelligence as according to a recent GWAS study (22).

Rank	Gene	Network	Conserved	Intelligence
1	SCN9A			
2	HUWE1	✓		
3	MACF1			
4	DOP1A			
5	UBR4	✓	Human	

TABLE III
TOP 5 CANDIDATE GENES IDENTIFIED FROM DIVERSITY SCORE: INCLUDES ANNOTATIONS FROM NETWORK ANALYSES (WHETHER IF THE GENE IS IDENTIFIED AS A HUB GENE); COMPARISON WITH EDITED GENES IN OTHER SPECIES (WHETHER IF THE GENE CONTAINS AT LEAST ONE NONSYNONYMOUS MUTATION IN HUMAN, MICE, OR FLY); AND WHETHER THE GENE HAS BEEN ASSOCIATED WITH HUMAN INTELLIGENCE.

Novel Editing Sites

In addition to theoretical molecular diversity that focuses on even representation of different types of transcripts, we developed another metric to identify genes with large amounts of novel editing sites. In this case, the direction of editing is important; an edit that substitutes a commonly found amino acid into a rare, novel amino acid at high editing level is taken to be highly novel, as reflected in the score.

Using this metric, the top 100 genes with the highest novelty generated were identified. Among them, 84 genes are included in the overall top 100 gene list (from combined UEI and PSSM). This overlap is expected since genes with a large number of editing sites will be ranked high in both measures.

Despite the similarities, a handful of genes were identified with high novelty scores but low overall scores. Supp. III includes top 5 genes that are ranked top 200 in the novelty list but have lower rankings in the overall list.

Studying these genes at the particular editing sites with high novelty scores help understand functions of RNA editing not primarily used for creating high levels of combinatorial diversity, but focused on creating highly novel amino acid substitutions. These sites may have played a role in generating highly novel cephalopod proteins that led to significant cephalopod developments, especially in the nervous system, through evolution.

In addition to genes, the novelty score was used to identify specific editing sites that may lead to novel amino acid substitutions. Supp. IV shows the top 5 sites in each of the four cephalopod species by novelty score, and Table IV below shows the top 5 sites that are conserved across all four species. Given the high editing level, the rarity of the amino acid substitution, and the conservation across species, these editing sites may serve to generate novel, functionally significant proteins. These identified novel sites provide specific candidates for future experimental studies on the impact of RNA editing on protein functionalities.

Rank	Gene	Oct Bim Sequence	Location in Oct Bim	Amino Acid Change
1	CHRNA9	comp163 911_c0_s eq1	361	Y -> C
2	VPS13A	comp166 111_c0_s eq4	683	R -> G
3	CNGA2	comp153 103_c0_s eq1	556	R -> G
4	LRRK2	comp166 025_c2_s eq1	8914	R -> G
5	PRICKLE2	comp160 160_c0_s eq1	591	Y -> C

TABLE IV

TOP NOVEL EDITING SITES CONSERVED ACROSS CEPHALOPODS. EDITING SITES CONSERVED ACROSS ALL 4 CEPHALOPOD SPECIES (SQUID, SEPIA, OCT BIM, AND OCT VUL) IN THEIR RESPECTIVE PROTEINS WERE RANKED BY THE NOVELTY SCORE PER SITE TO IDENTIFY INDIVIDUAL NOVEL SITES. THE TABLE SHOWS THE CORRESPONDING SEQUENCE AND LOCATION OF THE SITES IN OCT BIM.

Identification of Hub Genes From Protein-Protein Interaction Networks

To further investigate relationships and interactions between the candidate genes, protein-protein interaction networks of the top 100 and 200 candidate genes were produced using STRING (Fig. I). Connections inferred from experiments (protein-protein interactions), databases (KEGG pathways), co-expression, neighborhood, gene fusion, and co-occurrence were considered. Both networks have significantly more interactions than expected (PPI enrichment p-value < 1e-16), indicating that highly edited genes are involved in the same or interacting pathways.

From the networks, highly connected genes we call “hub genes” were identified, which are taken to be potential key players in a diverse set of pathways and may be crucial targets of RNA editing in cephalopods. These genes are highlighted in the top 100 list (Supp. II).

(a)

FIGURE I

PROTEIN-PROTEIN INTERACTION NETWORKS OF (A) TOP 100 AND (B) TOP 200 CANDIDATE GENES. NETWORK NODES REPRESENT PROTEINS, EDGES REPRESENT INTERACTIONS: KNOWN INTERACTIONS FROM CURATED DATABASES (LIGHT BLUE), EXPERIMENTALLY DETERMINED (PINK); PREDICTED INTERACTIONS FROM GENE NEIGHBORHOOD (GREEN), PREDICTED GENE FUSIONS (RED), GENE CO-OCCURRENCE (DARK BLUE); TEXT-MINING (YELLOW-GREEN), COEXPRESSION (BLACK), PROTEIN HOMOLOGY (LIGHT PURPLE). GENES WITHOUT ANY CONNECTIONS WERE REMOVED FROM THE GRAPHS.

Conservation with Model Species

Editing of top candidate genes in human, mice, and fly were investigated using data from REDiportal (15). In total, 11 genes were found to contain nonsynonymous editing sites in both mice and across the four cephalopod species, 22 were found for fly (*Drosophila melanogaster*) and cephalopods, and 323 were found for human and cephalopods. Comparing average diversity scores, genes also edited in mice have the highest mean score, followed by fly and human (Table V). Among the top 100 candidate gene list, 2 genes contained nonsynonymous edits in mice and fly respectively, and 20 genes contained nonsynonymous edits in human. These genes are highlighted in the overall top 100 list (Table III and Supp. II). GO analysis shows that genes edited in both cephalopods and humans are enriched in

cytoskeleton-related processes, compared to genes only edited in cephalopods (Supp. V). Given the vast distance between cephalopods and these species, the conservation of editing in these genes, although not specifically by site, may indicate broad functional significance or unique structural features that facilitate their editing in a variety of species.

Editing of Human Intelligence Genes

The candidate genes were matched to a list of genes implicated in human intelligence through GWAS and other studies (22). In total, 225 edited gene were identified in this intelligence set, of which 9 are in top 100 (highlighted in Table III and Supp. II). The mean score and median rank of these intelligence-implicated genes are both higher than values of the whole set (Table V), indicating a higher degree of potential diversification in these intelligence-related genes and the potential role RNA editing plays in developing complex behavior in cephalopods.

Editing of Protocadherin Genes

Previous studies on the octopus genome have discovered a unique expansion of protocadherins genes through gene duplication, which regulate the development of neurons and synapses (26). In chordates, protocadherins are also diversified using extensive alternative splicing to generate unique molecules on cell surfaces for self-recognition, a key component for the development of complex brains (27). In octopus protocadherins, one large exon encodes all cadherin domains, while editing is reported to be enriched in protocadherins, suggesting RNA editing, in conjunction with gene duplication, as an alternative mechanism to RNA splicing for generation of diverse protocadherin molecules in the nervous system (27).

To further assess the possibility that editing of protocadherins contributes to the development of neural complexity, we computed diversity scores of protocadherin genes across the four cephalopod species. As shown in Table V, edited genes in the clustered protocadherin family show a slight increase in diversity score (compared to the average diversity score across all genes), while edited genes in the atypical cadherin family do have notably higher diversity scores than expected (Table V). In fact, two atypical cadherins were ranked top 100 in the candidate list: FAT1, involved in cell migration and cell-cell adhesion, and DCHS1, which regulates neuroprogenitor cell proliferation and differentiation (28, 29). Interestingly, mutations in DCHS1 have been shown to cause Van Maldegem syndrome 1, an autosomal recessive disease characterized by intellectual disability and other malformations (30). Together, these results demonstrate the potential role of RNA editing in diversifying protocadherins and thus promoting the development of complex nervous systems.

Cephalopod Edited Genes	Number of genes	Mean score $\pm$ SD	Mean rank $\pm$ SD
All genes	3118	0.098 $\pm$ 0.132	1560 $\pm$ 900
Edited in human	323	0.115 $\pm$ 0.152	1464 $\pm$ 900
Edited in mice	11	0.263 $\pm$ 0.237	641 $\pm$ 711
Edited in fly	21	0.147 $\pm$ 0.198	1535 $\pm$ 1159
Associated with human intelligence	225	0.117 $\pm$ 0.133	1380 $\pm$ 906
Clustered protocadherins	9	0.065 $\pm$ 0.028	1491 $\pm$ 497
Atypical cadherins	6	0.419 $\pm$ 0.285	199 $\pm$ 161

TABLE V

DIVERSITY SCORE VALUES OF VARIOUS GROUPS OF GENES EDITED IN CEPHALOPODS (ALL EDITED GENES, GENES EDITED IN MODEL ORGANISMS, GENES ASSOCIATED WITH HUMAN INTELLIGENCE, AND PROTOCADHERIN-RELATED GENES)

Gene Ontology Analysis

Gene ontology analysis was used to determine functional enrichment of highly diversified genes. Table VI shows the 5 most significant GO terms identified from the diversity score ranked list, of which four are nervous system-related processes, indicating the diversification of RNA editing has its main effects in the nervous system. These results also reaffirm the previously reported connection between RNA editing and the nervous system in cephalopods.

GO Term	Description	P-value
GO:0010975	regulation of neuron projection development	1.88E-10
GO:0042391	regulation of membrane potential	5.45E-10
GO:0045664	regulation of neuron differentiation	7.92E-10

GO:0051960	regulation of nervous system development	1.06E-09
GO:0120035	regulation of plasma membrane bounded cell projection organization	2.01E-09

TABLE VI

TOP GO PROCESSES OF HIGHLY DIVERSIFIED GENES AND ASSOCIATED P-VALUES. GO ENRICHMENT AND ASSOCIATED P-VALUES WERE CALCULATED USING GORILLA WITH A SINGLE LIST OF GENES RANKED BY DIVERSITY SCORE.

Diseases Associated with Highly Diversified Genes

Diseases enriched in highly diversified genes were investigated using DisGeNET gene-disease associations (24). Table VII shows top enriched diseases related to the nervous system, which include both behavioral-related disorders (e.g. autism) and molecular-based diseases (e.g. glioblastoma). These associations (with significant p-values) again highlight the connection between diversification from RNA editing and the nervous system, and also demonstrate the potential of candidate genes to aid understanding of human intelligence and related diseases.

Disease ID	Disease Name	P-value
C0236969	Substance-Related Disorders	1.92E-04
C0004352	Autistic Disorder	2.57E-04
C1263846	ADHD	5.15E-03
C0442874	Neuropathy	8.50E-03
C0017636	Glioblastoma	8.80E-03

TABLE VII

TOP NEURAL DISEASES ENRICHED IN HIGHLY DIVERSIFIED GENES AND ASSOCIATED P-VALUES; DISEASE ENRICHMENT WAS CALCULATED USING DISGENET GENE-DISEASE ASSOCIATIONS DATASET. P-VALUES WERE CALCULATED FROM THE MINIMUM HYPERGEOMETRIC SCORE.

II. Codon Usage Bias

Assessing Codon Usage Bias

Transcriptome data of four coleoid species (oct vul, oct bim, squid, and sepia) and two outgroup species (nautilus and aplysia) were used to calculate observed and expected value of all codons (except for stop codons), GC and AG content, and the codon deviation coefficient, a measure of

overall level of codon usage bias, for every individual gene (2, 12). The observed and expected values were used to calculate adjusted bias per gene in each species, which was used to compare codon usage bias across species.

Predicting Editing with Nucleotide Composition and Codon Usages

To evaluate whether nucleotide composition and codon usage contains information regarding editing in general, we tested the predictive power of these information on editing in *Octopus bimaculoides*.

First, we set out to classify whether if a gene is edited using a simple logistic regression trained on genes in the transcriptome. The model achieves a decent accuracy of 81% on the test set ($\kappa = 0.62$). Next, we attempted to predict the extent of editing (measured using cumulative editing level, normalized by length). We built a random forest model, which achieved an R-sq value of 0.23 on the test set (i.e. the model, or overall nucleotide content and codon information, can explain 23% of the variation in the extent of editing), which is not high but is expected since extent of editing depends not only on overall properties of the gene, but also specific sequence information. In fact, previous studies have been able to predict specific editing sites from SNVs with very high accuracy, but these models rely on the specific sequence surrounding the site and have only been tested on data from human and *Drosophila* (31). The predictive ability of our models, given that they only consider gene-level properties, thus indicate that indeed overall nucleotide content and codon usages are related to RNA editing.

Comparison of Global Nucleotide Composition

After affirming the correlation between overall nucleotide and codon information and editing, we examined the specific nucleotide composition patterns and its relationship to RNA editing in cephalopods. Given that RNA editing (specifically A-to-I editing) acts on adenosine, we expected edited genes to have higher proportions of A compared to species average, and coleoid transcriptome in general to have higher A than nautilus and aplysia. Supp. VI(a) shows the observed mean proportions of A by gene across the species: similar to our expectation, coleoids show slightly elevated proportions of A compared to nautilus and aplysia. However, edited genes do not show higher proportions of A compared to the transcriptome averages. In all cases, as expected, the range and distribution are wide. These results suggest that coleoid transcriptomes might have gone through changes that lead to elevated A, potentially facilitating RNA editing, but within each coleoid transcriptome edited genes on average maintain the same A content as other genes.

Previous studies have found regions around editing sites display higher GC content, which form stronger double-stranded regions required by ADAR enzymes for

editing (2). Therefore, we hypothesized that edited genes would have higher GC content on average, and the widespread editing in coleoid species may have led to an increased GC content overall compared to other species. Supp. VI(b) shows the GC content per gene: contrary to our hypothesis, edited genes do not show higher GC content than all genes, and coleoid species do not have higher GC content--in fact, the median GC content is slightly elevated in nautilus, and aplysia shows significantly higher GC content than the coleoid species. Gene length does not account for this difference in GC content ($R\text{-sq} < 0.01$). One possible explanation is that although GC content may be higher near editing sites, other regions of the genome (e.g. crucial sites of a gene at which an edit would likely reduce functioning) may have evolved lower GC content in order to actively avoid editing.

Comparison of Codon Usage Bias

Given that RNA editing requires an A site and a double-stranded region near the editing site, and ADAR enzyme shows biases for nucleotides immediately upstream and downstream the A site (2), we hypothesized that 1) edited genes show codon usage bias that facilitate editing (e.g. for synonymous codons, the A-containing codon will be more represented to allow editing, and G-containing codon will be less represented since a codon with A can code for the same amino acid but with more flexibility), and 2) widespread RNA editing in coleoid species may have influenced codon usage in the overall genome.

To assess these potential differences, we first calculated the CDC coefficient, which measures the overall codon usage bias in a gene (12), and perhaps surprisingly aplysia has the largest CDC value per gene, while nautilus has about the same level of CDC as the coleoid species. Edited genes have slightly lower median CDC values, but the difference seems negligible. One possible explanation for such results is gene length: shorter genes are slightly biased towards higher CDC values (12), and indeed aplysia genes have shorter lengths than other species, while edited genes tend to be longer (Supp. VII), and there is a negative correlation between gene length and CDC value in the dataset ($R\text{-sq} = 0.25$).

Next, we analyzed biases of individual codons across the species. Comparing edited genes with the overall species as control, 23 codons are consistently more biased in edited genes for all four coleoid species, and 65% of them contain at least an A, while 48% contain G (the neutral expectation is 57%); 21 codons are consistently less biased in edited genes, of which 52% contain A, and 67% contain G. In addition, A at codon position 2 is more biased in edited genes for all four species. Together, these results supporting our hypothesis that codons with A may be biased in edited genes for allowing potential editing. Among these codons, CAG and TCT are consistently more biased by at least 10% (see Methods for details on calculations) in edited genes for

all four coleoid species. As an A-containing codon, CAG codes for glutamine, which can be edited to CGG for arginine, an important type of editing that underlies a crucial RNA editing site in mice (Q/R in GRIA2, a neurotransmitter receptor that if unedited, leads to early death in mice)(1).

Comparing coleoid species and nautilus, the non-coleoid cephalopod with little RNA editing and less behavioral complexity, A is more biased in all four coleoid species, while G is less biased, again supporting our hypothesis that RNA editing leads to biases towards A-containing codons over G. Interestingly, C is more biased in coleoids compared to nautilus, and T is less biased, potentially to preserve similar GC content. Several A-containing codons are more biased by at least 10% in all four coleoid species compared to nautilus, including ATG, TTA, CAA, and CGA, of which two may lead to amino acid recoding through editing (ATG for M/V, CAA for Q/R). Codons AGA and AGG are less biased by at least 10%, which can be edited from AAA and AAG, a K/R amino acid change that is conservative and over-represented in cephalopods (see next section). The codon AAG and its edit to AGG also corresponds to the cephalopod ADAR preference for an upstream A and a downstream G around the editing site (2). Continuing these comparisons to between coleoids and aplysia, several trends still hold true: ATG, TTA, and CAA are more biased by at least 10% in all coleoids compared to aplysia as well. G at position 2 and codon AGA are more biased by at least 10% in both aplysia and nautilus compared to the coleoids.

Totaling these comparisons, codons ATG, GAA, and CAA among others are consistently more biased by 5% in 1) edited genes compared to overall species average, and 2) coleoids compared to nautilus and aplysia. A at position 2 and codon AAA are consistently more biased in the aforesaid levels, while their counterparts G at position 2 and codon AGA (corresponding to the potential edit K/R) are consistently less biased. Together, these results demonstrate the changes in codon usage bias that correspond to potential effects or causes of widespread RNA editing.

III. Natural Selection and Amino Acid Recoding

Assessing Natural Selection with R/C Ratio

To assess the level of natural selection (specifically, purifying selection) on cephalopod RNA editing, R/C ratios were calculated from amino acid recoding of nonsynonymous editing sites in the cephalopod species as well as human, mice and fly for comparison. Results (Table VIII) show that cephalopods have clearly lowered R/C ratios compared to human, mice, and fly, and within each group the R/C ratios are similar. This indicate that cephalopods face similar levels of purifying selection on RNA editing that are notably stronger than model species. Furthermore, edits conserved across cephalopods show even lower R/C ratio, providing evidence that conserved edits undergo

stronger selection and are more likely to have some form of functional significance. Interestingly, there is no clear relationship between R/C and editing level (Supp. VIII), unlike N/S, which has been shown to increase as editing level increases, potentially because measures incorporating R/C mostly correlates with purifying selection, while N/S is also influenced by positive selection, which may be strong for edits with higher editing levels (2, 9).

Another potential explanation for the reduced level of radical amino acid substitutions from RNA editing in cephalopods is that through changes of ADAR preferences and transcriptome sequence composition, the RNA editing mechanism is biased towards potential editing sites that give rise to conservative amino acid substitutions. Since conservative substitutions are on average less likely to be deleterious, this would lower the selective cost of RNA editing and promote the evolution of widespread RNA editing (9).

Species	R/C
Cephalopod conserved	0.52
Oct bim	0.55
Oct vul	0.57
Squid	0.58
Sepia	0.60
Human	0.79
Fly	0.79
Mice	0.83

TABLE VIII

R/C RATIOS OF AMINO ACID SUBSTITUTIONS RESULTING FROM RNA EDITING IN CEPHALOPODS AND MODEL SPECIES

Distribution of Amino Acid Recoding by RNA Editing

Given the reduced R/C ratios observed in cephalopods, we hypothesized that cephalopod RNA editing may have evolved distinct amino acid recoding preferences that favor conservative over radical substitutions, thus decreasing the potential for deleterious effects of edits. To test this hypothesis, we calculated the distribution of amino acid substitutions (measured in fold difference from expected frequencies of amino acid substitutions based on the setup of the genetic code) by RNA editing in the cephalopod species, again in comparison to human, mice, and fly. As shown in Figure IV, all four cephalopod species display similar distributions: for example, K/R and S/G substitutions are heavily over-represented in cephalopods,

two conservative amino acid changes. Compared to human, fly, and mice, cephalopod amino acid recoding shows a weaker preference for radical substitutions (for example, we see reductions in K/E and E/G substitutions, both of which are radical). These results thus support our hypothesis that cephalopod RNA editing tends to result in conservative substitutions, reducing negative impacts of potential off-target edits.

FIGURE IV

DISTRIBUTION OF AMINO ACID SUBSTITUTIONS FROM NONSYNONYMOUS RNA EDITING SITES IN CEPHALOPODS AND HUMAN, MICE, AND FLY

DISCUSSION

Recently it was discovered that A-to-I editing, a rare and generally non-adaptive process in humans, leads to widespread protein recoding in soft-bodied cephalopods, a group of marine animals with large nervous systems and complex behaviors (2). Furthermore, neural tissues were the most highly edited and edited genes are enriched in various neural processes, suggesting a connection between RNA

editing and the development of sophisticated behavior in soft-bodied cephalopods. In order to better understand the protein diversity created through RNA editing, we developed a quantitative metric that assesses the level of combinatorial diversity generated in edited proteins. This metric identifies top targets of RNA editing that may be further studied experimentally to assess the specific functional impact of the diversity brought by RNA editing. In addition, the novelty score highlights individual recoding substitutions that may be novel to cephalopods and under selective pressure, offering specific candidate sites for further investigations. As an example of experimental investigations, RNA editing on the Kv2 potassium channel orthologs in cephalopods (ranked in top 150 of the diversity score list) was recently studied by expressing the unedited and edited versions and assessing the electrical properties of the channels experimentally (2). Results showed that single amino acid substitutions (squid I579V, sepia I630V, and octopus I632V) lead to changes in the channels' closing rates. Similar studies may be performed on the highlighted candidate genes and novel editing sites to help understand the molecular and cellular effects of RNA editing.

Several genes in the top 100 list may warrant further attention: ranked second in the list, the gene HUWE1 is a ubiquitin-protein ligase that mediates protein degradation (29). HUWE1 is also identified as a hub gene from network analyses, suggesting that it may play a central role in diverse pathways. HUWE1 is known to be involved in the regulation of neural differentiation and proliferation, and is associated to intellectual disability and severe mental retardation and developmental delays in children (Juberg-Marsidi Syndrome) (33-35). These connections to neural development and intelligence suggest that RNA editing of HUWE1 may play an important role in the cephalopod nervous system. Another gene of interest is PTPRD, ranked 43rd in the diversity score metric, and also included in the gene set associated with human intelligence from a GWAS study (22). The gene codes for a protein tyrosine phosphatase that induces neuron differentiation (29). The gene also contains 17 editing sites conserved in all 4 cephalopod species, one of which is ranked top 5 in novel conserved edits (see Table 3). One final gene to highlight is HTT, or the Huntingtin protein, ranked 22nd in the list. In humans, expansions of the polyglutamine domain in HTT lead to Huntington's Disease, a progressive neurodegenerative disorder (36). HTT is also widely expressed throughout human development and is predicted to be highly pleiotropic; however, much is not known about its normal functions (36-38). Investigation of HTT protein diversity in cephalopods could provide a unique perspective on the function and structure of this protein.

Results of the diversity score highlight several interesting patterns and further support the connection between RNA editing and nervous system development. Disease annotations show associations of highly diversified

genes with various neurological and behavioral disorders, including autism, ADHD, neuropathy, and glioblastoma. GO analyses showed highly diversified genes are implicated in neuron differentiation and neuronal projection development, a pattern noted in previous studies of RNA editing (2). Along with the higher levels of protein diversity in atypical cadherins and axon guidance genes (see below), these results suggest a special role for RNA editing in neuron projection development and guidance.

In addition to providing targets for future studies, these results also offer a reference dataset of protein diversity values generated by RNA editing in cephalopods. Research in cephalopod proteins may consult this data to assess the level of diversity from editing for any specific protein of interest. For example, a recent study identified the gene *SLC6A4* as a receptor for drug substances that alter social behaviors in cephalopods (39). The level of diversity created by RNA editing on this protein can be checked using the diversity score results: in this case, *SLC6A4* is edited at a negligible level, indicating that RNA editing has minimal impact on the structure and function of this protein.

Previous studies on the octopus genome have identified 15 axon guidance genes known to play an important role in the development of complex neural circuitry (26). Here, the diversity score data provide a perspective on the diversification of these molecules: calculations show that the mean rank of these genes in the diversity score list is significantly higher than expectation, and two genes from the set are in the top 100 list: *ROBO2*, a transmembrane receptor for molecular guidance cues in cell migration, and *PLXNA4*, a co-receptor which modulates affinity for axon growth guidance molecules (29). Interestingly, decreased expression of both genes were found in patients with autism (40). Interestingly, *PLXN* genes were also found to be uniquely duplicated in parrot genomes compared to all other birds, indicating *PLXN* gene diversity may be a more general mechanism for the evolution of advanced cognition (41).

Cephalopod Intelligence: RNA editing, Autism, and Aging

Amodio et al. (2019) discuss the evolution of cephalopod intelligence and compare cephalopod biology with that of the large-brained vertebrate groups (cetaceans, corvids, primates). In contrast to the vertebrate groups, cephalopods seem to be unique in their lack of sociality. Some species aggregate into large anonymous schools, but cephalopods are generally solitary and do not engage in complex social bonds (42).

Similar to the asociality of cephalopods, Autism spectrum disorders in humans are characterized by dysregulation of social cognition, impaired communication, and stereotyped and repetitive behavior (43, 44). The high prevalence and heritability of autism spectrum disorders have led to a variety of evolutionary hypotheses on their

origin and maintenance (45). One such hypothesis is the “solitary forager hypothesis”, which proposes that the autistic behavioral profile would have been advantageous for hunting and gathering in scarce ancestral environments (46, 47). Cephalopods are also characterized by a solitary lifestyle and complex foraging niche (42), suggesting that there may be cognitive and neural similarities between autistic humans and cephalopods. This idea is supported by the results of our disease enrichment analysis that shows autism-risk genes are highly diversified by RNA editing across four cephalopod species (see Supp. IX for further analysis). Interestingly, ADHD-related genes also show enrichment. Mental traits associated with ADHD have also been conceptualized as adaptations for the dangerous and complex foraging environments faced throughout human evolution (46, 48, 49).

Previous work has shown that social behavior in bees is mediated by autism-related genes (50-52). Taken together with our results, this indicates a common genetic toolkit that modulates sociality and foraging-related cognitive traits across both vertebrates and invertebrates. Further support for a shared genetic toolkit comes from a study showing positive selection of autism risk genes in a cave-dwelling fish that displays autistic-like traits such as asociality, hyperactivity, imbalanced attention, and repetitive behaviors (53).

The high degree of RNA editing in these genes in cephalopods also suggests a more specific mechanism in which a large, asocial brain can be produced. Autism-risk alleles are also known to be associated with neurogenesis and enhanced brain size, and it has been hypothesized that the widespread positive selection of these genes in humans is due to this association (43, 54). Positive selection on RNA editing in these genes may have played a similar role in the development of a large brain in cephalopods. In addition, evidence in humans suggests that the etiology of autism is mediated to significant degree by effects of variation in gene copy number (55-58). RNA editing may have played a similar role to gene duplication in creating a diversity of gene products and enhancing transcriptional complexity. Overall, this hints that modulation of a few core genetic pathways through either RNA editing or gene duplication can affect both brain size and social cognition across vertebrates and invertebrates. Support for this notion comes from recent work demonstrating dysregulation of RNA editing in autistic brains (59). Additionally, the autism-risk genes shown to modulate sociality in bees are enriched for RNA editing functions (51). Further study of the connection between RNA editing and autism in cephalopods and humans may be a promising direction for future research.

Amodio et al. (2019) also highlight the uniqueness of the cephalopod’s fast life histories (lifespans of 0.5-2 years, semelparity, absence of parental care) in contrast to the slow life histories of the large-brained vertebrate groups. They suggest that cephalopod intelligence and the fast life

history evolved as a response to the loss of an external shell, which in turn led to an environment with high predation and cognitively challenging foraging niches. It is interesting to consider the connection between RNA editing, lifespan/aging, and cognition in this context. The link between RNA editing and cognition is well established on both a phylogenetic level and in specific spatiotemporal scales (2, 60-63). It is also apparent that RNA editing comes with a significant cost in the form of random off-target edits by the ADAR enzyme and the general potential for dysregulation. A significant amount of work has highlighted the role of RNA editing in neurodegenerative diseases and aging more generally (59, 63-67). We propose that this “balancing act of RNA editing” (64) was particularly salient during cephalopod evolution. Speaking generally, primates, corvids, and cetaceans can take their time to develop large brains and advanced cognition. Because of their long lifespans, there is also selective pressure to make a brain that is built to last. The use of RNA editing as a tool for brain development was important for these groups, but ultimately limited by the aforementioned costs. The deployment of RNA editing was largely limited to non-coding edits (68, 69) in order to limit the load of potentially damaging proteins with off-target recoding events. Cephalopods faced very different evolutionary pressures during their evolution. In this scenario, high predation favored the evolution of a short lifespan, but there was a concurrent pressure to rapidly develop neural and cognitive complexity. Additionally, this shift in environmental pressures caused by the loss of a shell may have happened over a relatively quick evolutionary timescale. Normal mechanisms of genome evolution (coding sequence evolution, gene duplication, expansion of non-coding regulatory DNA) may not have been able to meet this challenge of rapidly evolving a large brain that can rapidly develop over the course of a cephalopod’s short life. Widespread RNA editing enriched for protein recoding may have been the only option for developing the requisite protein diversity and transcriptional malleability (70). However, this came with a two-fold cost. The fact that ADAR enzymes can only edit dsRNA slowed down genome evolution and possibly constrained long-term evolutionary outcomes (2). The more immediate cost was the significant protein load caused by dangerous off-target protein recoding events. This protein load causes more rapid neurodegeneration and aging in cephalopods. Given the fact that there was already selective pressure to evolve a faster life history, this did not incur quite as significant of a cost and thus was evolutionary permissibly. The metaphor here would be someone who needs a house as soon as possible, but only for a short time, and therefore they build it quickly but shoddily. Research on cephalopod aging is limited, but all species studied have showed a characteristic period of senescence following sexual maturation (71-74). Older cephalopods show deficiencies in learning, memory, and behavior flexibility that are accompanied by

neurodegeneration (72). Further elucidation of the neurological and molecular signatures of aging in cephalopods may yield novel insights into the evolutionary costs and benefits of widespread RNA editing and protein recoding.

Future Directions and Conclusion

Improvement and refinement of the diversity score using additional data could yield valuable insights into the mechanisms and functions of RNA editing in cephalopod nervous systems. Generation and analysis of diversity score by specific neural tissues (the current data is pooled over multiple neural tissues) could provide insights into variations in RNA editing and protein diversity across tissues. Another direction is the acquisition of more data on RNA editing across multiple individuals for each cephalopod species. These data will help establish reliable patterns of editing and diversity and elucidate intraspecies variation in RNA editing and protein diversity. Another crucial component of the overall diversity created by RNA editing is temporal diversity, or the variation of produced proteins through time in response to varying environmental conditions or stages of development (6). This diversity may offer organisms greater flexibility in adapting to a changing environment but is currently not considered in the diversity score due to a lack of data on RNA editing at different time points. Acquisition of such data and the development of a temporal diversity score may help elucidate the function of RNA editing in environmental adaptation and development. In addition, the current calculations assume independence between editing sites in one RNA transcript, as the diversity score represents a theoretical maximum of the combinatorial diversity. However, a previous study has discovered substantial linkage between editing sites in *Drosophila* and cephalopods (18). Incorporations of these correlations between edits in one transcript may improve the diversity score to better reflect the realized protein diversity.

In summary, we develop a diversity score metric that quantitatively assesses the combinatorial protein diversity created by RNA editing in cephalopods. Results highlight connections between RNA editing and nervous system development, particularly neuron projection development and guidance, and identify highly diverse genes as candidates for future study. More generally, these results support the functional significance of RNA editing in cephalopods and suggest further research on this phenomenon could help uncover the mechanisms behind their neural and behavioral complexity.

Our results regarding nucleotide composition and codon usage biases show that they in general contain information on editing, and several specific shifts in edited genes and coleoid transcriptomes correspond to facilitating RNA editing. These changes may act together with previously shown differences in cephalopod ADAR preferences (compared to e.g. humans) for sequences near

the editing site to facilitate RNA editing and also influence the resulting type of amino acid substitutions, possibly lowering the selective cost of RNA editing (2). Potentially, initial changes in nucleotide composition and codon usages promoted RNA editing, which in turn leads to further selective pressure for codon usage biases and changes in nucleotide composition that favor editing, creating a positive feedback loop that may have prompted the evolution of widespread RNA editing in cephalopods. The lowered R/C ratio we discovered indicates stronger purifying selection at the amino acid level, and we found distinct amino acid preferences in cephalopod RNA editing favoring conservative over radical substitutions. These changes in amino acid recoding may be a result of upstream changes in nucleotide content and codon usage biases, a connection that may warrant further inspection. Analyses of ADAR preferences and their relation to nucleotide composition, codon usage bias, and amino acid recoding in the future could test these potential scenarios and potentially help unlock the mystery of the origin of widespread, adaptive RNA editing in cephalopods.

ACKNOWLEDGEMENTS

I would like to thank my mentor Mr. Erik Mohlhenrich for his guidance and encouragement throughout the project. I thank Dr. Sergey Samsonau and Mr. Joseph Li for their help on the project. I thank Dr. Joshua Rosenthal of University of Chicago and Dr. Eli Eisenberg of Tel Aviv University for their valuable advice. I thank PRISMS for providing me with the opportunity and support for research. (by M. Wang)

REFERENCES

1. K. Nishikura, A-to-I editing of coding and non-coding RNAs by ADARs. *Nat. Rev. Mol. Cell Biol.* **17**(2), 83–96 (2016).
2. N. Liscovitch-Brauer *et al.*, Trade-off between Transcriptome Plasticity and Genome Evolution in Cephalopods. *Cell*. **169**, 191–202 (2017).
3. B. Zinshteyn, K. Nishikura, Adenosine-to-inosine RNA editing. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **1**(2), 202–209 (2009).
4. G. Xu, J. Zhang, Human coding RNA editing is generally nonadaptive. *Proc. Natl. Acad. Sci.* **111**(10), 3769–3774 (2014).
5. S. Kumar, G. Stecher, M. Suleski, S. B. Hedges, TimeTree : A Resource for Timelines, Timetrees, and Divergence Times. *Mol. Biol. Evol.* **34**(7), 1812–1819 (2017).
6. Kazuko Nishikura, Functions and regulation of RNA editing by ADAR deaminases. *Annu. Rev. Biochem.* **79**, 321–349 (2010).
7. N. Paz-Yaacov *et al.*, Adenosine-to-inosine RNA editing shapes transcriptome diversity in primates. *Proc. Natl. Acad. Sci.* **107**(27), 12174–12179 (2010).
8. B. Smith, J. B. Wilson, A consumer's guide to evenness indices. *Oikos*. **76**, 70–82 (1996).
9. N. G. C. Smith, Are Radical and Conservative Substitution Rates Useful Statistics in Molecular Evolution. *J. Mol. Evol.* **57**, 467–478 (2003).
10. L. B. Møller *et al.*, Identification and analysis of 21 novel disease-causing amino acid substitutions in the conserved part of ATP7A. *Hum. Mutat.* **26**(2), 84–93 (2005).
11. P. C. Ng, S. Henikoff, Predicting Deleterious Amino Acid Substitutions. *Genome Res.* **11**(5), 863–874 (2001).
12. Z. Zhang *et al.*, Codon Deviation Coefficient : a novel measure for estimating codon usage bias and its statistical significance. *BMC Bioinformatics*. **13**, 43 (2012).
13. I. Frumkin *et al.*, Codon usage of highly expressed genes affects proteome-wide translation efficiency. *Proc. Natl. Acad. Sci.* **115**(21), E4940–E4949 (2018).
14. G. Ramaswami *et al.*, Genetic mapping uncovers cis-regulatory landscape of RNA editing. *Nat. Commun.* **6**, 8194 (2015).
15. E. Picardi, A. M. D. Erchia, C. Lo Giudice, G. Pesole, REDIPortal: a comprehensive database of A-to-I RNA editing events in humans. *Nucleic Acids Res.* **45**, D750–D757 (2017).
16. B. Giardine *et al.*, Galaxy: A platform for interactive large-scale genome analysis. *Genome Res.* **15**, 1451–1455 (2005).
17. P. Rice, The European Molecular Biology Open Software Suite EMBOS: The European Molecular Biology Open Software Suite. *Trends Genet.* **16**(6), 276–7 (2000).
18. Y. Duan *et al.*, Linkage of A-to-I RNA Editing in Metazoans and the Impact on Genome Evolution. *Mol. Biol. Evol.* **35**(1), 132–148 (2017).
19. H. Tang, G. J. Wyckoff, J. Lu, C. Wu, A Universal Evolutionary Index for Amino Acid Changes. *Mol. Biol. Evol.* **21**(8), 1548–1556 (2004).
20. G. M. Boratyn *et al.*, Domain enhanced lookup time accelerated BLAST. *Biology Direct*. **7**, 12 (2012).
21. D. Szklarczyk *et al.*, The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* **45**, D362–D368 (2017).
22. J. E. Savage *et al.*, Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nat. Genet.* **50**, 912–919 (2018).
23. E. Eden, R. Navon, I. Steinfeld, D. Lipson, Z. Yakhini, GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*. **10**, 48 (2009).
24. J. Piñero *et al.*, DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. **45**, D833–D839 (2017).
25. K. Hanada, S. H. Shiu, W. H. Li, The nonsynonymous/synonymous substitution rate ratio versus the radical/conservative replacement rate ratio in the evolution of mammalian genes. *Mol. Biol. Evol.* **24**(10), 2235–2241 (2007).
26. C. B. Albertin *et al.*, The octopus genome and the evolution of cephalopod neural and morphological novelties. *Nature*. **524**, 220–224 (2015).
27. Z. Y. Wang, C. W. Ragsdale, Seminars in Cell & Developmental Biology Cadherin genes and evolutionary novelties in the octopus. *Semin. Cell Dev. Biol.* **69**, 151–157 (2017).
28. H. Y. Gee *et al.*, FAT1 mutations cause a glomerulotubular nephropathy. *Nat. Commun.* **7**, 10822 (2016).
29. The UniProt Consortium, UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45**, D158–D169 (2017).
30. S. Cappello *et al.*, Mutations in genes encoding the cadherin receptor-ligand pair DCHS1 and FAT4 disrupt cerebral cortical development. *Nat. Genet.* **45**, 1300–1308 (2013).
31. Z. Ouyang *et al.*, Accurate identification of RNA editing sites from primitive sequence with deep neural networks. *Sci. Rep.* **8**, 6005 (2018).
32. J. Romiguier, V. Ranwez, E. J. P. Douzery, N. Galtier, Contrasting GC-content dynamics across 33 mammalian genomes: Relationship with life-history traits and chromosome sizes. *Genome Res.* **20**, 1001–1009 (2010).
33. X. Zhao *et al.*, The HECT-domain ubiquitin ligase Huwe1 controls neural differentiation and proliferation by destabilizing the N-Myc oncoprotein. *Nat Cell Biol.* **10**(6), 643–653 (2008).
34. S. Moortgat *et al.*, HUWE1 variants cause dominant X-linked intellectual disability: A clinical study of 21 patients. *Eur. J. Hum. Genet.* **26**(1), 64–74 (2018).

35. M. J. Friez *et al.*, HUWE1 mutations in Juberg-Marsidi and Brooks syndromes: the results of an X-chromosome exome sequencing study. *BMJ Open*. **6**, e009537 (2016).
36. J. Schulte, J. T. Littleton, The biological function of the Huntingtin protein and its relevance to Huntington's Disease pathology. *Curr. Trends Neurol.* **5**, 65–78 (2011).
37. Q. Guo *et al.*, The cryo-electron microscopy structure of huntingtin. *Nature*. **555**, 117–120 (2018).
38. J. Ochaba *et al.*, Potential function for the Huntingtin protein as a scaffold for selective autophagy. *Proc. Natl. Acad. Sci.* **111**(47), 16889–16894 (2014).
39. E. Edsinger, G. Dölen, A Conserved Role for Serotonergic Neurotransmission in Mediating Social Behavior in Octopus. *Current Biology*. **28**(9), 3136–3142.e4 (2018).
40. S. Suda *et al.*, Decreased expression of axon-guidance receptors in the anterior cingulate cortex in autism. *Mol. Autism*. **2**, 14 (2011).
41. M. Wirthlin *et al.*, Parrot Genomes and the Evolution of Heightened Longevity and Cognition. *Curr. Biol.* **28**, 4001–4008 (2018).
42. P. Amodio *et al.*, Grow Smart and Die Young: Why Did Cephalopods Evolve Intelligence? *Trends Ecol. Evol.* **34**(1), 45–56 (2019).
43. B. J. Crespi, Autism as a disorder of high intelligence. *Front. Neurosci.* **10**, 300 (2016).
44. J. E. Reser, Solitary mammals provide an animal model for autism spectrum disorders. *J. Comp. Psychol.* **128**(1), 99–113 (2014).
45. A. Ploeger, F. Galis, Evolutionary approaches to autism- an overview and integration. *McGill J. Med.* **13**(2), 38–43 (2011).
46. R. M. Nesse (2015). Evolutionary Psychology and Mental Health. In D. M. Buss (Ed.), *The Handbook of Evolutionary Psychology* (pp. 903–927). Hoboken, NJ, USA: John Wiley & Sons, Inc. doi:10.1002/9781119125563.evpsych243
47. J. E. Reser, Conceptualizing the autism spectrum in terms of natural selection and behavioral ecology: The solitary forager hypothesis. *Evol. Psychol.* **9**(2), 207–238 (2011).
48. D. Eisenberg, B. Campbell, The Evolution of ADHD Social Context Matters. *Med. Midlife*. 21–22 (2011).
49. J. Williams, E. Taylor, The evolution of hyperactivity, impulsivity and cognitive diversity. *J. R. Soc. Interface*. **3**, 399–413 (2006).
50. B. J. Crespi, Shared sociogenetic basis of honey bee behavior and human risk for autism. *Proc. Natl. Acad. Sci.* **114**(36), 9502–9504 (2017).
51. H. Y. Shpigler *et al.*, Deep evolutionary conservation of autism-related genes. *Proc. Natl. Acad. Sci.* **114**(36), 9653–9658 (2017).
52. M. C. Saul *et al.*, Cross-species systems analyses reveal a shared brain transcriptional response to social challenge. *bioRxiv*, 219444 (2018).
53. M. Yoshizawa *et al.*, The evolution of a series of behavioral traits is associated with autism-risk genes in cavefish. *BMC Evol. Biol.* **18**, 89 (2018).
54. R. Polimanti, J. Gelernter, Widespread signatures of positive selection in common risk alleles associated to autism spectrum disorder. *PLoS Genet.* **13**(2), e1006618 (2017).
55. R. M. Cantor, D. H. Geschwind, Schizophrenia: Genome, Interrupted. *Neuron*. **58**, 165–167 (2008).
56. B. Crespi, P. Stead, M. Elliot, Comparative genomics of autism and schizophrenia. *Proc. Natl. Acad. Sci.* **107**, 1736–1741 (2010).
57. B. Crespi, K. Summers, S. Dorus, Genomic sister-disorders of neurodevelopment: An evolutionary approach. *Evol. Appl.* **2**, 81–100 (2009).
58. D. Moreno-De-Luca *et al.*, Deletion 17q12 is a recurrent copy number variant that confers high risk of autism and schizophrenia. *Am. J. Hum. Genet.* **87**, 618–630 (2010).
59. S. S. Tran *et al.*, Widespread RNA editing dysregulation in brains from autistic individuals. *Nat. Neurosci.* **22**(1), 25–36 (2019).
60. J. B. Li, G. M. Church, Deciphering the functions and regulation of brain-enriched A-to-I RNA editing. *Nat. Neurosci.* **16**(11), 1518–1522 (2013).
61. J. S. Mattick, M. F. Mehler, RNA editing, DNA recoding and the evolution of human cognition. *Trends Neurosci.* **31**(5), 227–233 (2008).
62. J. S. Mattick, The central role of RNA in human development and cognition. *FEBS Lett.* **585**, 1600–1616 (2011).
63. T. Hwang *et al.*, Dynamic regulation of RNA editing in human brain development and disease. *Nat. Neurosci.* **19**, 1093–1099 (2016).
64. S. Farajollahi, S. Maas, Molecular Diversity through RNA Editing: A Balancing Act. *Trends Genet.* **26**(5), 221–230 (2010).
65. N. Schonrock *et al.*, The RNA modification landscape in human disease. *Rna*. **23**, 1754–1769 (2017).
66. M. F. Mehler, J. S. Mattick, Noncoding RNAs and RNA Editing in Brain Development, Functional Diversification, and Neurological Disease. *Physiol. Rev.* **87**, 799–823 (2007).
67. P. Sebastiani *et al.*, RNA editing genes associated with extreme old age in humans and with lifespan in *C. elegans*. *PLoS One*. **4**(12), e8210 (2009).
68. H. T. Porath, B. A. Knisbacher, E. Eisenberg, E. Y. Levanon, Massive A-to-I RNA editing is common across the Metazoa and correlates with dsRNA abundance. *Genome Biol.* **18**, 185 (2017).
69. M. H. Tan *et al.*, Dynamic landscape and regulation of RNA editing in mammals. *Nature*. **550**, 249–254 (2017).
70. S. Hussain, Shaping and Reshaping Transcriptome Plasticity during Evolution. *Trends Biochem. Sci.* **42**(9), 682–684 (2017).
71. R. C. Anderson, J. B. Wood, R. A. Byrne, Octopus Senescence: The Beginning of the End. *J. Appl. Anim. Welf. Sci.* **5**(4), 275–83 (2002).
72. M. P. Chichery, R. Chichery, Behavioural and neurohistological changes in aging Sepia. *Brain Res.* **574**, 77–84 (1992).
73. J. A. Mather, Behaviour Development: A Cephalopod Perspective. *Int. J. Comp. Psychol.* **19**, 98–115 (2006).
74. K. Roumbedakis, Á. Guerra (2019). Cephalopod Senescence and Parasitology. In: Gestal C., Pascual S., Guerra Á., Fiorito G., Vieites J. (eds) *Handbook of Pathogens and Diseases in Cephalopods*. Springer, Cham
75. Y. Jiang, M. D. Ehlers, Modeling Autism by SHANK Gene Mutations in Mice. *Neuron*. **78**, 8–27 (2013).
76. L. Boccuto *et al.*, Prevalence of SHANK3 variants in patients with different subtypes of autism spectrum disorders. *Eur. J. Hum. Genet.* **21**, 310–316 (2013).

SUPPLEMENTARY MATERIALS

Trinity name	SwissProt name	Human SwissProt name ("- if not detected)	Editing location inside sequence	Codon changes	RNA reads: G	RNA reads: A	Editing levels [%]	DNA reads: G	DNA reads: A
comp141867_c0_seq4	UAFA_STAS1	-	2696	IM	24	4737	0.504095778	0	10
comp140834_c2_seq1	HERC1_HUMAN	HERC1_HUMAN	1392	syn	8	394	1.990049751	0	10
comp141481_c3_seq6	RHG32_XENLA	RHG32_HUMAN	2829	TA	15	398	3.631961259	0	14
comp141407_c0_seq11	PTPRD_HUMAN	PTPRD_HUMAN	604	MV	70	4573	1.507645919	0	7
comp141386_c0_seq15	WNK3_MOUSE	WNK1_HUMAN	5651	HR	213	48	81.6091954	0	15

SUPP. I

SAMPLE ROWS OF EDITING SITE DATASET IN CEPHALOPODS: TRINITY NAME, HUMAN SWISSPROT NAME, EDITING LOCATION, CODON CHANGES, AND EDITING LEVELS WERE USED IN THIS STUDY.

SUPP. II: TOP 100 GENES BY DIVERSITY SCORE (SEE LINK)

Gene	Novelty Score Rank	Overall Diversity Score Rank
ARHGAP6	129	461
SYT1	168	497
HID1	165	477
CARMIL1	162	390
SLC12A2	199	408

SUPP. III

GENES WITH HIGHLY NOVEL EDITING SITES NOT IDENTIFIED FROM OVERALL DIVERSITY SCORE: GENES RANKED TOP 200 IN NOVELTY SCORES WERE COMPARED TO THEIR RANKS FROM OVERALL DIVERSITY SCORE. THE TABLE DISPLAYS THE TOP 5 GENES WITH THE LARGEST DIFFERENCE BETWEEN THE TWO RANKS.

Rank	Gene	Oct Bim Sequence	Location	Amino Acid
1	ADK	comp162290_c2_seq2	1591	N -> D
2	IDH3B	comp96997_c0_seq1	939	N -> S
3	VPS13A	comp166353_c0_seq3	8736	Y -> C
4	NPAS3	comp164774_c0_seq1	1285	Y -> C
5	MYO7A	comp164883_c0_seq19	5979	S -> G
Rank	Gene	Oct Vul Sequence	Location	Amino Acid
1	H6PD	comp179088_c1_seq1	2012	Y -> C
2	ADK	comp171017_c1_seq1	1061	N -> D
3	PTPRM	comp178865_c0_seq1	543	R -> G
4	IDH3B	comp178788_c0_seq1	1298	N -> S

5	IGF1R	comp183541_c2_seq12	1153	Y -> C
Rank	Gene	Squid Sequence	Location	Amino Acid
1	ZNF570	comp110213_c2_seq1	268	H -> R
2	MYO10	comp140327_c5_seq3	939	Y -> C
3	UGGT1	comp140978_c0_seq1	3070	Y -> C
4	H6PD	comp132684_c0_seq1	2031	K -> R
5	SMARCA5	comp140138_c12_seq1	954	Y -> C
Rank	Gene	Sepia Sequence	Location	Amino Acid
1	WNT8A	comp259845_c0_seq2	651	Y -> C
2	ZNF300	comp259368_c2_seq1	370	H -> R
3	AARS	comp255397_c0_seq1	1063	Y -> C
4	TOP2B	comp255128_c0_seq1	2720	Y -> C
5	OTOF	comp267591_c2_seq12	1725	Y -> C

SUPP. IV

TOP 5 NOVEL EDITING SITES IN OCT BIM, OCT VUL, SQUID, AND SEPIA.

GO Term	Description	P-value
GO:0007010	cytoskeleton organization	1.51E-04
GO:0022402	cell cycle process	2.90E-04
GO:0031589	cell-substrate adhesion	5.33E-04
GO:0006996	organelle organization	9.36E-04
GO:0000910	cytokinesis	9.72E-04

SUPP. V

GO PROCESSES ENRICHED IN GENES EDITED IN BOTH CEPHALOPODS AND HUMANS

SUPP. VI

ADENOSINE (A) AND GC CONTENT (B) PER GENE OF EDITED GENES, COLEOID SPECIES, AND OUTGROUPS.
OUTLIERS ARE REMOVED FROM THE BOXPLOT

SUPP. VII

MEAN GENE LENGTHS OF EDITED GENES, COLEOID SPECIES, AND OUTGROUPS. OUTLIERS ARE REMOVED FROM THE BOXPLOT

SUPP. VIII

R/C RATIOS OF EDITING SITES BY EDITING LEVEL IN CEPHALOPOD SPECIES

SFARI Autism Score	Number of genes matched in cephalopods	Average Diversity Score Rank
1	11	573
2	30	1015
3	52	900
4	106	1103
5	37	1470
6	3	1568

SUPP. IX

ANALYSIS OF SFARI AUTISM GENES IN DIVERSITY SCORE DATASET

Further analysis of known autism genes in our diversity score dataset supports the results of the disease enrichment analysis. The SFARI database (<https://gene.sfari.org/database/human-gene/>) provides a list of genes connected to autism and provides a score that indicates the level of support for this connection (ranked 1-6, with 1 being the strongest evidence). The table above shows the number of genes in the SFARI database that has orthologs edited in cephalopods, broken down by the SFARI score, and their average rank by diversity score (with the neutral expectation being 1561). The elevated rank for genes with strong SFARI scores suggest that these autism-related genes are highly diversified by RNA editing in cephalopods. The top ranked gene in diversity score, SCN9A, has a score of 2 in the SFARI database. In total, 27 of the top 100 are autism-risk genes. Among them is the gene SHANK2 (SFARI score 2), ranked 82 in diversity score. The SHANK family, which organize proteins at the postsynaptic density, has been suggested as causative genes of ASD (75). In particular, SHANK3 is a leading autism candidate gene (score 1), with studies reporting copy number variations and points mutations in SHANK3 discovered in ASD patients (76). SHANK2 mutations have also been associated with ASD (75).