Skolkovo Institute of Science and Technology

HERITABLE MODIFICATIONS OF TRANSMITTED BIOLOGICAL INFORMATION AS POSSIBLE SOURCES OF ADAPTATION

Doctoral Thesis
by
MIKHAIL MOLDOVAN

DOCTORAL PROGRAM IN LIFE SCIENCES

Supervisor
Professor Mikhail S. Gelfand

Moscow - 2022

Declarations

I hereby declare that the work presented in this thesis was carried out by myself at Skolkovo Institute of Science and Technology, Moscow and has not been submitted for any other degree.

Candidate (Mikhail Moldovan)

Supervisor (Prof. Mikhail Gelfand)

Abstract

Heritable variability in proteome on the population scale mostly arises due to germline mutations causing amino acid substitutions. Along with that, variability of proteomes may stem from regular heritable changes of mRNA and protein molecules occurring by means of mRNA editing and post-translational modifications of proteins. Some of these changes may mimic the effects of direct mutations, such as the adenine-to-inosine editing of RNA, which mimics adenine-to-guanine substitution, and phosphorylation of proteins, which to some extent mimics emergence of negatively charged amino acids in protein sequences. These changes are correlated to a large extent, which results in formations of clusters of modifications along protein and RNA sequences. These clusters further add to the variability generated by heritable modifications. The aim of the work described in the present thesis is to assess the possible role of heritable modifications and their features such as clusterization on the expressed variability and rates of adaptation. In particular, we study the abundant A-to-I editing in soft-bodied cephalopods and protein phosphorylation in mammals. Our objectives include: 1) Constructing a theory of evolution of heritable modifications, 2) Assessment of the effects A-to-I RNA editing may have on the adaptation of soft-bodied cephalopods, 3) Assessment of the effects of protein phosphorylation on evolutionary patterns at modified sites, 4) Study of clusterization of A-to-I editing sites in soft-bodied cephalopods and phosphorylated amino acid residues in mammals and effects this clusterization may have on evolution.

We begin by constructing a general theoretical framework for the studies focusing on the effects of heritable modifications of biological information on the evolution of organisms. These effects are discussed in terms of various existing concepts such as evolvability, rates of adaptation, phenotypic plasticity, exaptation and other.

Next, we describe our published studies dedicated to adenine-to-inosine mRNA editing in coleoids, soft bodied cephalopods. In coleoids, adenine-to-inosine RNA editing, resulting in non-

synonimous changes of codons, is orders of magnitude more frequent than in any other studied group of organisms. Editing is heritable, as adenine-to-inosine editing requires, firstly, specific trinucleotide contexts, and secondly local RNA structures. By studying selection regimes of edited adenines in coleoids, we find that positive selection is acting concordantly with editing: in positions occupied by frequently edited non-conserved adenines, there is selection towards guanines.

Adenine-to-guanine RNA editing sites in coleoids tend to cluster. We show that their clustering contributes about a half to the variance generated by editing in general, thus further boosting the phenotypic variance. In an analogy between genetic variants and editing sites, this clustering corresponds to epistasis. We also show that editing sites tend to form clusters of three distinct size ranges, which correspond to three types of RNA secondary structures. These structures, when occurring around functional editing sites, tend to impose additional selective constraint on sequences, thus decreasing the genetic variance. Hence our findings point at a complex role of coleoid editing sites in general: whereas functional editing sites decrease genetic variability in their vicinity, non-functional sites are points of increased observed variability.

Finally, we describe our study of evolutionary patterns associated with phosphorylated amino acids in mammals. Phosphorylation is the most conserved and the best studied post-translational modification, which introduces local negative charge to protein globules, and phosphorylated serines have been shown to frequently mutate to or originate from negatively charged amino acids. The bulk of phosphorylated amino acids is represented by serines and threonines with a much lesser fraction of tyrosines and other amino acids such as histidines. Phosphorylation typically requires local contexts on the sequence, which are mostly grouped in families of acidic, basic, or proline contexts. In our study we employ the largest available dataset of phosphorylated residues and show that phosphorylated residues tend to cluster along the protein sequence. Interestingly, the major fraction of phosphorylated residues occurs in clusters and clustered residues exhibit a larger propensity to mutate to negatively charged amino acids and to have acidic contexts than the non-clustered ones. In addition, sites

phosphorylated in larger numbers of tissues tend to mutate to negatively charged amino acids more frequently than those phosphorylated in smaller numbers of tissues. This study points at a weaker selective constraint against mutations to negatively charged amino acids associated with clustered phosphorylation and contributes to our understanding of the evolution of negatively charged regions in proteins.

Publications

1. **Moldovan, M**., Chervontseva, Z., Bazykin, G., & Gelfand, M. S. (2020). Adaptive evolution at mRNA editing sites in soft-bodied cephalopods. PeerJ, 8, e10456. https://doi.org/10.7717/peerj.10456
2. **Moldovan MA**, Gelfand MS. (2020). Phospho-islands and the evolution of phosphorylated amino acids in mammals. PeerJ 8:e10436 https://doi.org/10.7717/peerj.10436
3. **Moldovan, M. A.,** Chervontseva, Z. S., Nogina, D. S., & Gelfand, M. S. (2022). A hierarchy in clusters of cephalopod mRNA editing sites. In Scientific Reports (Vol. 12, Issue 1, article 3447). Springer Science and Business Media LLC. https://doi.org/10.1038/s41598-022-07460-5

Conferences

1. Gaydukova, S; Moldovan. M; Gelfand, M. ITaS’2019

2. Moldovan, M. CephRes’2020

Acknowledgements

Here, I would like to thank Prof. Mikhail Gelfand for guidance and supervision, colleagues from the Gelfand and Bazykin laboratories for discussions and criticisms of my work, my friends and family for their support, and Skoltech for providing facilities and infrastructure for the research.

Table of contents

Declarations *2*
Abstract *3*
Publications *6*
Conferences *7*
Acknowledgements *8*
List of symbols, abbreviations *12*
List of figures *14*
List of tables *18*
Chapter 1. Introduction *19*
1.1. Relevance and significance of the work **19**
1.2. Personal contribution **20**
Chapter 2. Review of the literature *21*
2.1. Introduction **22**
2.2. Heritable modifiers of biological information **24**
2.3. Evolution of information modifiers **31**
2.4. Epistasis **35**
2.5. Polymorphism and polygenic traits **36**
2.6. Exaptation **40**
2.7. Phenotypic plasticity / Accommodation **43**
2.8. Genetic assimilation **48**
2.9. E-Variance-mediated adaptation **50**
2.10. Balancing selection **54**
2.11. Evolvability **56**
2.12. Expectations **59**
Chapter 3. Adaptive evolution at mRNA editing sites in soft-bodied cephalopods *61*
3.1. Introduction **62**
3.2. Methods **65**
3.2.1. Data. 65
3.2.2. Annotation of structured and unstructured regions. 65
3.2.3. Analysis of polymorphisms. 66
3.2.4. Alignments. 66
3.2.5. Context analysis. 66
3.2.6. Substitution matrix. 67
3.2.7. *R* and *Q* calculation for non-synonymous (NES) and synonymous (SES) editing sites. 68
3.2.8. Calculation of *dN/dS*. 69
3.2.9. Statistics. 70
3.2.10. Data availability. 70
3.3. Results **70**
3.3.1. Editing level is associated with the local and global sequence context. 70
3.3.2. Editing level is affected by secondary structure in adjacent RNA. 72

3.3.3. Edited adenines are often substituted by guanines. 75
3.3.4. Editing recapitulates substitutions that are positively selected. 77

3.4. Discussion 79
3.4.1. The hypothesis about the adaptivity of non-conserved editing sites is supported by our observations. 79
3.4.2. E-to-G substitutions versus G-to-E substitutions. 79
3.4.3. Positive selection in favor of E-to-G substitutions. 82
3.4.4. Conservation and function of editing. 84
3.4.5. Theoretical frameworks and alternative explanations 85

Chapter 4. A hierarchy in clusters of cephalopod mRNA editing sites 89

4.1. Introduction 89

4.2. Methods 94
4.2.1. Data 94
4.2.2. Calculation of *S* values 94
4.2.3. Control sets of adenines 94
4.2.4. Editing state co-occurence 95
4.2.5. Variance due to editing 95
4.2.6. RNA structural annotations 96
4.2.7. Structural mismatch annotations 97
4.2.8. Order of editing events 97
4.2.9. Statistics 98
4.2.10. Code availability 99

4.3. Results 99
4.3.1. Correlated editing 99
4.3.2. Dense editing site clusters (adjacent adenines) 102
4.3.3. Medium-range clusters of editing sites 104
4.3.4. Long-range clusters of editing sites 107
4.3.5. Directionality of editing 109

4.4. Discussion 113
4.4.1. Cooperativity of RNA editing 113
4.4.2. The range of influence of editing sites 115

Chapter 5. Phospho-islands and the evolution of phosphorylated amino acids in mammals 117

5.1. Introduction 117

5.2. Methods 120
5.2.1. Data 120
5.2.2. Alignments and Trees 120
5.2.3. Phosphorylation retention upon mutations 120
5.2.4. False-positive rates of phosphorylation identification by homologous propagation 121
5.2.5. Mutation matrices 122
5.2.6. Disordered regions and identification of phospho-islands 122
5.2.7. Phosphosite contexts 123
5.2.8. Local mutation matrices 124
5.2.9. Statistics 124
5.2.10. Code availability 124

5.3. Results 125
5.3.1. Conserved phosphosites 125
5.3.2. Phosphorylation islands 128
5.3.3. Mutational patterns of phosphorylated amino acids 131
5.3.4. Phosphosite contexts 134
5.3.5. Phosphorylation breadth 135
5.3.6. Mutation patterns in the proximity of phosphosites 137

5.4. Discussion 139
5.4.1. Clustered vs. individual phosphosites 139
5.4.2. Two types of mutations 141

5.4.3. Human phosphosites 142
5.4.4. Mutations to lysine 143
5.4.5. Evolution of non-studied phosphosite groups 144

Chapter 6. Conclusions ***145***

Bibliography ***148***

Supplementary materials ***168***

Supplementary materials for chapter 3: Adaptive evolution at mRNA editing sites in soft-bodied cephalopods. **168**
Extrapolation 168
Caveats 169
Supplementary Figures 170

Supplementary materials for chapter 4: A hierarchy in clusters of cephalopod mRNA editing sites. **176**
Supplementary Figures 176
Supplementary Tables 181

Supplementary materials for chapter 5: Phospho-islands and the evolution of phosphorylated amino acids in mammals. **184**
Supplementary Figures 184
Supplementary Tables 195

List of symbols, abbreviations

NCES – non-conserved editing site

CES – conserved editing site

NES – non-synonymous editing site

SES – synonymous editing site

N_1-to-N_2 – nucleotide substitution from N_1 to N_2

N_1-N_2 – nucleotide mismatch between homologous positions occupied by N_1 nucleotide in one sequence and N_2 nucleotide in another sequence.

X_1-to-X_2 – amino acid substitution from X_1 to X_2

X_1-X_2 – amino acid mismatch between homologous positions occupied by X_1 amino acid in one sequence and X_2 amino acid in another sequence.

EL – editing level of an editing site calculated as fraction of RNAseq reads mapped onto genomic homozygous nucleotide that contain nucleotide non-identical to the genomic one.

ADAR – Adenosine deaminases acting on RNA. A family of proteins converting genomically-encoded adenines in RNA molecules to inosines, informational analogues of guanines.

dN/dS – ratio of non-synonymous to synonymous substitution rates.

DC – dense editing site cluster.

S – linear between-site distance either in amino acids or in nucleotides on the primary structure.

NS – not significant.

PTMs – Protein post-translational modification.

STY amonoacids – serines, threonines, and tyrosines.

pSTY amonoacids – phosphorylated serines, threonines, and tyrosines.

IDR – intrinsically disordered region.

OR – ordered region.

NCA – negatively charged amino acid.

OGG – orthologous gene group.

HMR phosphosites – phosphorylated amino acids conserved between human and rat or human and mouse.

HMM – hidden Markov model.

LSM – local substitution matrice.

HIM – heritable information modifier.

List of figures

Figure 1 | Prevalent mRNA editing in coleoid molluscs.

Figure 2 | Contextual features of coleoid A-to-I mRNA editing sites.

Figure 3 | *R* and *Q* values.

Figure 4 | *dN/dS* values of adenine substitutions to guanines for various EL thresholds.

Figure 5 | Clustered mRNA editing in coleoid molluscs

Figure 6 | Correlations between various properties of editing sites.

Figure 7 | Properties of densely clustered A-to-I editing sites.

Figure 8 | Properties of medium-range clusters of editing sites

Figure 9 | Long-range editing site clusters.

Figure 10 | Directionality of dense clusters.

Figure 11 | Phosphosites considered in the study.

Figure 12 | Phospho-islands for the HMR phosphosite dataset.

Figure 13 | $pX_0 \rightarrow X_1$ substitution vectors.

Figure 14 | Phosphosite contexts and phosphorylation breadth.

Figure 15 | *Q* values of mutations near ST phosphosites with probabilities significantly different from the expected ones.

Supplementary Figure S1 | LOGOs of nucleotides adjacent to editing sites in all studied organisms at different editing levels.

Supplementary Figure S2 | LOGOs of nucleotides adjacent to conserved and non-conserved editing sites in the squid-cuttlefish pair.

Supplementary Figure S3 | Over- and underrepresented mismatch *R* values in the local context of non-conserved editing sites in the pair of *Octopus* species.

Supplementary Figure S4 | Pearson's correlation between the absolute value of difference of ELs in homologous sites and the number of mismatches in a window of a given size for different window sizes.

Supplementary Figure S5 | *A. californica* has less adenines in structured regions then other species.

Supplementary Figure S6 | The local secondary structure is more stable at edited adenines than at homologous, non-edited adenines.

Supplementary Figure S7 | Dependence of Pearson's correlation between the difference of structural Z-scores and the difference in ELs of homologous sites on the minimal considered difference in ELs.

Supplementary Figure S8 | Dependence of R_G and R_Y on the editing level considered separately for non-synonymous and synonymous sites.

Supplementary Figure S9 | Normalized *dN/dS* of editing site substitutions.

Supplementary Figure S10 | Mutational characteristics of editing sites.

Supplementary Figure S11 | The distributions of correlation coefficients of coleoid editing at two sites with respect to the distances between sites.

Supplementary Figure S12 | The distributions of correlation coefficients of *O. vulgaris* editing at two sites with respect to the distances between sites for different minimal editing level threshold values.

Supplementary Figure S13 | The dependence of the correlations of ELs on the *S* values for the considered coleoid A-to-I editing site datasets.

Supplementary Figure S14 | Results obtained with the context-aware expectation of the distribution of edited adenine.

Supplementary Figure S15 | Histograms of dense cluster sizes for the real coleoid editing site datasets and the corresponding randomly obtained ones.

Supplementary Figure S16 | Distributions of mismatches of adenines in double RNA helices.

Supplementary Figure S17| Structural properties of coleoid A-to-I editing sites.

Supplementary Figure S18 | Clustering of A-to-I editing sites in coleoid transcriptomes.

Supplementary Figure S19 | Distributions of the $r'(A_i, A_j)$ values calculated for the structurally close editing sites (red boxes) and for the control site pairs with no predicted secondary RNA structure between the sites in a pair (grey boxes), where both sites A_i and A_j are located in the same exon of *O. bimaculoides*.

Supplementary Figure S20 | Distributions of differences in ELs between down- and upstream editing site in two-adenine dense clusters obtained for three reading frames.

Supplementary Figure S21 | Stability of hidden Markov model clustered editing site predictions with respect to transitional probability values.

Supplementary Figure S22 | Phospho-island analysis in various datasets.

Supplementary Figure S23 | Comparison of the mutational patterns observed in HMR phosphosites located in intrinsically disordered and ordered regions.

Supplementary Figure S24 | Comparison of the mutational patterns observed in HMR phosphosites located in phospho-islands *vs*. individual ones.

Supplementary Figure S25 | Comparison of the mutational patterns observed in mouse phosphosites located in intrinsically disordered and ordered regions.

Supplementary Figure S26 | Comparison of the mutational patterns observed in mouse phosphosites located in phospho-islands *vs*. individual ones.

Supplementary Figure S27 | Comparison of the mutational patterns observed in human phosphosites located in intrinsically disordered and ordered regions.

Supplementary Figure S28 | Comparison of the mutational patterns observed in human phosphosites located in phospho-islands *vs*. individual ones.

Supplementary Figure S29 | *R* values of the pS-to-X mutations for different phosphosite sets.

Supplementary Figure S30 | Conservation of phosphorylated amino acids for various mouse datasets.

Supplementary Figure S31 | *R* values of mutations near ST phosphosites with probabilities significantly different from the expected ones for various HMR subsets.

Supplementary Figure S32 | Maximal DC sizes in randomly computed uniform expectations and the actual maximal editing DC size in *O. vulgaris*.

List of tables

Table 1 | Discussed information-modifying processes.

Supplementary table S1 | SRA identifiers of RNAseq libraries employed the study Moldovan et al., 2022.

Supplementary Table S2 | Variance in the transcriptome and proteome explained by editing and by correlations in editing events.

Supplementary Table S3 | Effect sizes and confidence intervals for the data presented on Fig. 10C.

Supplementary Table S4 | 95% Confidence intervals of differences in base-pairing probabilities between paired editing sites (EE) and three types of control AA-dinucleotides obtained by random sampling.

Chapter 1. Introduction

1.1. Relevance and significance of the work

Our studies are dedicated to, firstly, construction of a theoretical framework for evolutionary studies of biological information transmission and, secondly, to the analyses of several processes influencing this transmission. These lines of research are relevant in several aspects:

Firstly, a large fraction of studies in systems biology are dedicated to various types of modifications of transmitted information such as post-translational modifications of proteins, regulation of transcription, regulation of the pool of mRNA in the cell by, e.g., micro-RNAs, etc. While these studies are concerned with the value of these processes from the standpoints of regulation or physiology, we propose an additional view from the evolutionary standpoint. Indeed, these processes may gain an evolutionary value by including non-coding loci and polymorphisms they harbor in the production of phenotypes. Although this is most likely not the primary role of information-modifying processes, the phenotypes they generate still can enhance adaptation rates or the rates of neutral evolution.

Secondly, in our studies we consider amino acids in proteins as complex traits yielded, on the one hand, by sequences of respective codons and, on the other hand, by contexts needed for information-modifying machinery to operate. This gives us probably the largest set of complex traits ever studied. Complex traits have recently become an object for intensive studies dedicated to heritability yielded by non-coding polymorphism and the general non-coding constraint. There is generally three lines of research here: i. Structure of these traits, i.e. what are the processes behind the interaction of alleles, ii. Genetics of these traits, e.g., what is the fraction of epistatic variance or why is there polymorphism of complex traits, iii. Evolution of complex traits, i.e. how do complex traits evolve and what stands behind the changes in their distributions. At least for one information-modifying process, mRNA editing in coleoid cephalopods, we provide at least partial answers to all these questions.

1.2. Personal contribution

The analyses reported here have been performed by the author, except results presented in paragraphs 3.3.2 “Editing level is affected by secondary structure in adjacent RNA”, 4.3.3 “Medium-range clusters of editing sites”, 4.3.4 “Long-range clusters of editing sites,) and 4.3.5 “Directionality of editing”, which have been obtained by Zoe Chervontseva and are summarized in, respectively, Figures 2BCD, 8B, 9B and 10C.

Text presented here was written by the author and edited by the supervisor, except:

1. Paragraph 3.2.2 (“Annotation of structured and unstructured regions”, which was written by Zoe Chervontseva.
2. Paragraph 3.3.2 “Editing level is affected by secondary structure in adjacent RNA”, which was written jointly by Zoe Chervontseva and the author.
3. Paragraph 3.4.2 “Positive selection in favor of E-to-G substitutions”, which was written by Prof. Georgii Bazykin.
4. Paragraphs 4.2.6 “RNA structural annotations” and 4.2.7 “Structural mismatch annotations”, which were written jointly by Zoe Chervontseva and the author.

They are included for completeness of the exposition.

Conceptual work reported here was performed jointly by the author and the supervisor.

Chapter 2. Review of the literature

Advances in molecular biology allow us to study effects conveyed by different alleles in contexts of specific biological processes. Of particular interest are processes of realization of genetically encoded information, which have been shown to be influenced by such factors as the general noise inherent to all molecular systems and programmed modifications. Both of the named factors may influence the effects of genetic variants on the phenotype, hence being potential major contributors to the evolution of traits. Consequently, in the past few years it has been shown that the bulk of heritability of traits in well-studied populations may stem not from polymorphism in genes, but rather from the non-coding genome.

Biological information passes through some processes that may be studied in the context of evolution of traits. As these processes can generate ambiguous states as is the case with, e.g., alternative splicing, the affected traits arise not from the genotype itself, but from an ensemble of states genetically-encoded information takes. This adds an additional nuance to our view of evolution itself. Indeed, according to the classical scheme, evolution may be regarded as the loop of information with two steps: i. Genotype that is decoded and phenotype is produced, ii. Selection as well as random environmental processes influence the frequencies of phenotypes and hence alleles underlying the phenotypic values and the system returns to step (i). In our view, there is a third step between steps (i) and (ii), namely modifications genetically-encoded information can take. Regarding this step can greatly add to our understanding of biology behind evolutionary changes or evolutionary stability.

Just like virtually any trait, the variation in heritable changes of the information decoding has generally three parts: i. Genetic variation, which stems from some specific context of changes such as variation in the sequences of RNA editing sites or in gene promoters, ii. Dependence of genotype on environment, as is the case with differential regulation of gene expression and iii. Noise conveyed by the environment. Over more than a century of evolutionary thought, multiple theoretical frameworks concerning all of these aspects have been formulated and some are applicable here.

In the following chapter, we review some of the most studied processes influencing decoding of information as well as the theories, which may provide the frameworks for describing the effects of individual processes on the evolution of organisms. Finally, we find that various profound aspects of the effects of biological information modifications on the general evolutionary process may be studied using a small and largely redundant set of tests, which are listed in the concluding section. Thus, this review provides the most general theoretical framework for studying modifications of biological information.

2.1. Introduction

Organisms evolve by accumulating heritable changes in phenotype which are constantly supplied by the mutational process (Kimura 1983). If these changes are selected upon, organisms are said to adapt, and so adaptations, or any heritable changes at that matter, are impossible without random genetic changes (Dobzhansky, Hecht, and Steere 1968; Kimura 1983; Lynch and Walsh 1998). Consequently, the mutational process and genetic variation it generates have been some of the main subjects of biology in the past century (Kimura 1983; Seplyarskiy et al. 2021).

Both empirical and theoretical evolutionary studies are mainly focused on the effects of genetic variation arising in coding sequences of proteins or sequences coding for structural RNA. However, it recently has become clear that a large fraction of genetic variation may be explained not by variation in coding sequences or functional RNA sequences, but rather by variation in non-coding modifier loci (Karczewski et al. 2020; Seplyarskiy et al. 2021). And even if we look at such phenotypes as protein sequences encoded in sequences of codons, the pure direct genotype-to-phenotype perspective is a rather simplistic one, and the discoveries in the field of evolutionary molecular biology may provide a more accurate and nuanced picture, which includes the features of the transmission of information within a living organism from genomic sequences to phenotypes. Information transmission is also subject to multiple random changes, some of which may be heritable, hence being prone to neutral evolution and natural selection. Mainly, information within the cell is transmitted from DNA to RNA

and sometimes further from RNA to proteins (Crick 1958; Crick et al. 1961) and can be modified on each of these steps via, e.g. alternative RNA splicing (Mironov et al. 2021), RNA editing (Eisenberg and Levanon 2018), ambiguous codon decoding (Feketová et al. 2010), translation of upstream ORFs (Zhang et al. 2021) and further, e.g. by post-translational modifications of proteins (Huang et al. 2018). All these processes have the potential to introduce novel phenotypic variants, and (almost) all of them have a necessary heritable component in the form of specific contexts, thus having potential to introduce heritable changes in resulting phenotypes and hence may constitute some of the driving forces behind evolution.

Modifications of biological information such as alternative splicing or alternative gene expression give rise to a variety of regulated responses to environment (Ghalambor et al. 2015; Liu et al. 2022). And, as survival of organisms largely depends on the adequacy of responses to environmental stimuli, the heritable variance in information transmission can be viewed as a necessary part of adaptation, which has been suggested to include, on par with selection-driven genetic changes, induction of traits by environment. As West-Eberhard puts it: "*If it is the phenotype, not the genotype, that is the object of selection, then selection can proceed for generations without genetic variation and without an evolutionary effect, as long as there is developmentally significant environmental variation.*" (West-Eberhard 2005). In addition, modifications themselves may depend on the environment. And so, we argue that modification of transmitted biological information can be a source of heritable changes on multiple levels and should be regarded in studies aiming to understand the general evolutionary patterns.

Also, virtually all known organisms, especially complex ones, utilize multiple modifiers of biological information to diversify the transcriptome, regulate development and regulate responses to the environment (Koonin 2016; Duclos, Hendrikse, and Jamniczky 2019). This leads to a possibility of novel variants, such as proteins with additional sequences coded by newly acquired exons, arising purely by means of variation in information-modifying processes. Another characteristic of complex

organisms is a typically lower effective population size (Kimura 1983), which leads to less effective selection, which, in turn, may lead to evolution exploiting alternative sources of variation beyond standard non-synonymous substitution/indel/gene reassortation mutations within genes. Hence, evolution of organismal complexity to be at some degree deemed evolution of characters acquired via epigenetic modification of encoded information (Hoekstra and Coyne 2007).

We begin our review by listing the most studied information-modifying processes and possible sources of additional genetic variation they can introduce. Further, we list the existing theories focusing on the evolution of information-modifying processes. As we feel it impossible to ignore the colossal body of intellectual work focusing on various aspects of adaptation, namely genetic assimilation, exaptation, influence of epistasis, adaptation of multi-locus characters, evolvability and others, we further discuss our subject in terms of each of these fields. As the terminologies used to formulate some of these concepts, such as exaptation, are not readily applicable to the current omics-centered studies, we re-define some parts of the concepts where it is due. Next, we review the data on the evolution of various epigenetic modifications of biological information and suggest possible directions for future studies. We conclude with a list of remarks, ideas and speculations that fall in frames of our topic.

2.2. Heritable modifiers of biological information

Before we proceed with the discussion of effects on adaptation posed by heritable variation in information transmission pathways, a necessary remark should be made about the scope of the general theme of this thesis. We do not discuss modifications with no long-term heritability, e.g. heritable histone modifications, as short-term heritability means limited time when selection can operate (Charlesworth, Barton, and Charlesworth 2017), which renders this route of adaptation ineffective. For reviews and discussions see (Charlesworth, Barton, and Charlesworth 2017; Samhita 2021). Likewise, we do not discuss in detail processes influencing the genome, e.g. stress-induced mutagenesis, CRISPR-cas, or mutational process in general, however we do refer to some studies on these topics

where necessary. For reviews, see (Koonin and Wolf 2009; Koonin 2012; Charlesworth, Barton, and Charlesworth 2017). This leaves us with the following, not necessarily comprehensive, list of processes.

1. **Differential expression.** The rate of the bulk of chemical reactions catalyzed by enzymes and ribozymes, association of structural molecules and regulatory interactions between biological polymers mostly depends on two factors: concentration of proteins or other biologically-active molecules and the capacity of individual molecules to associate with some substrate or transform it (Wright 1929; Bisswanger 2014). The former is defined mostly by expression of a molecule, whereas the latter is defined by the properties of its chemical structure like the sequences of RNA or protein molecules (Michaelis and Menten 1913; Srinivasan 2021). Thus, the reaction a protein or an RNA is involved in can be evolutionary influenced either by changes in the protein or RNA sequence itself or by changes in promoter sequences. As promoters co-evolve with transcription factor sequences (Lynch and Hagner 2015), expression is also influenced by the portion of genes coding for transcription factors. One additional level on which mutational process may influence some given reaction or interaction may be posed by sequences of, e.g, RNA polymerases or histones which non-specifically influence transcription in general.
2. **Alternative splicing**. One of the characteristic features of Eukaryotes are *genes in pieces*, i.e. coding sequences comprising exons separated in their genomes by non-coding sequences: introns (Koonin, Csuros, and Rogozin 2013). Each exon-intron junction is defined by splicing sites, forming pairs on both sides of each intron: donor sites and acceptor sites with additional branching sites located in the vicinity of acceptor sites. The pattern of splicing can alternate between different tissues and lineages, being *par excellence* genetically determined (Merkin et al. 2012; Hiller and Platzer 2008; Mironov et al. 2021). Thus, the inclusion of exons in protein sequences is governed by splicing sites, and that gives us an additional source of variation in

protein sequences. In other words, the effects of alleles at protein-coding regions depend both on the nature of these alleles and on the efficiency of splicing sites, which govern the inclusion of focal coding sites in resulting mature mRNAs (Mironov et al. 2021). Apart from the splicing site-exon sequences pair, one could also regard splicing enhancer site-splicing site pair, where splicing signal at a given site partially depends on the splicing-enhancer sequence (Wang and Burge 2008). As in the previous case, additional non-specific trans-effects in the form of splicing machinery can be regarded as an additional layer influencing decoding of a given coding sequence.

3. **Alternative polyadenylation.** Polyadenylation is a process of adding multiple adenosine monomers to the 3' end of mRNA in the process of its maturation (Tian 2005). The number of added adenines is controlled by sequences of polyadenylation sites located downstream from coding sequences and regulates mRNA stability and translation rates of protein products from coding regions (Tian 2005; Zhang et al. 2021). At the population-level, apart from variations in poly-A tails caused by differences in sequences of specific polyadenylation sites, there may be additional variation caused by multiple alternative polyadenylation sites governing translation of some given gene (Tian and Manley 2017). Here, we have the following three types of sites: i) Coding regions in mature mRNAs ii) Polyadenylation sites iii) Sequences of factors involved in polyadenylation machinery.
4. **RNA editing**. Changes in translated RNA sequences are a widespread mechanism used by organisms on various branches on the tree of life to fight viruses, produce coding sequences from error-rich DNA templates, regulate gene activity or provide final stages to the genesis of biologically active RNA molecules, e.g. tRNAs (Eisenberg and Levanon 2018; Pecori et al. 2022). The most widespread types of RNA editing in metazoans are cytosine-to-uracil (C-to-U) editing performed by ABOBEC and adenine-to-inosine (A-to-I) editing performed by ADAR. Typically, RNA editing is performed on nucleotides located in double-stranded RNA

molecules with sometimes a specific nucleotide context surrounding the focal editing site, thus providing the basis for heredity of editing sites (Alon et al. 2012). This gives us another example of a context influencing decoding at a focal (editing) site along with the corresponding possibility of evolutionary changes of the editing process itself through changes in, e.g., APOBEC or ADAR sequences.

5. **micro-RNAs/siRNAs.** Micro (mi) RNAs and small interfering (si) RNAs are the most widespread regulators of gene expression at post-transcriptional level. These molecules function via complementary binding to small (approx. 20-22nt) regions on transcripts and attracting enzymes which destroy the transcript or render it non-translatable by depositing it in dense granules. Both miRNAs and siRNAs may act as cis-regulators if they are translated from intronic sequences of regulated gene or as trans-regulators in case of their location in other non-coding regions. In our terms, the general scheme here is very much similar to that in case of transcription regulators: regulatory sequence that has no phenotypic effect besides acting on phenotypic effects of other sequences. However, one important difference here is that regulation by means of mi-and siRNAs requires a portion of the target sequence to be complementary to regulating RNA molecules. Thus, both regulatory and target sequences govern the efficiency of regulation. Of course, apart from sequences of miRNAs and siRNAs influencing specific genes, sequences of enzymes that govern this mode of regulation such as DICER or RISC complexes provide an additional layer to evolution of these processes (Berezikov 2011).
6. **Programmed ribosomal frameshifting.** A single transcript does not necessarily need to harbor a single coding sequence in a single reading frame. Instead, proteins are in multiple instances translated from multiple reading frames joined by the events of ribosome shifting by a number of nucleotides that is not a multiple of three. Such events are known as ribosomal frameshifting, and a plethora of known organisms, especially viruses, utilize ribosomal

frameshifting for the sake of regulation. To be both inherited and regulated in specific sites, ribosomal frameshifting requires specific features to be present in the vicinity of a site where reading frames should be changed, i.e. context on the sequence and/or specific RNA structures (Atkins et al. 2016). The three types of sites here are constituted by: i) coding sequences located in multiple reading frames, ii) sequences of RNA structures or contextual sequences of frameshift sites, iii) sequences of proteins and RNA molecules comprising ribosomes.

7. **u-ORFs.** Upstream open reading frames (ORFs) are ORFs located in 5'-untranslated regions of transcripts that are capable of attenuating translation at downstream coding ORFs. As in the case of elements affecting transcription, here we have a sequence of a gene, its decoding affected by regulatory sequence and sequences of translation apparatus non-specifically affecting this process (Zhang et al. 2021).
8. **Context-dependent protein modifications.** Proteins are known to frequently undergo post-translational modifications (Huang et al. 2018; The UniProt Consortium 2019). Whereas the bulk of these modifications actually happening in cell are non-specific as is the case with chemical acetylation of proteins in mitochondria or ubiquitination (Baeza, Smallegan, and Denu 2016; Sadowski and Sarcevic 2010), other modifications are guided by contexts, e.g., protein phosphorylation (Villén et al. 2007). Although specific contexts are not known for a number of frequent modifications, such as acetylation, the consistency of modifications observed on specific residues points at the presence of some sequence or structural contexts (Zhang et al. 2009). Contextual sequences thus influence decoding of codons coding for modified amino acid residues, this influence being heritable. This adds to our growing list yet another trio: i) codons coding for modified amino acids, ii) contextual sequences guiding modification iii) sequences of phosphorylases, acetylases etc, that may be either site-specific, in which case they may be coupled with contextual sequences, or non-specific (pleiotropic) and influence the process non-specifically.

The discussed processes share two common traits:

1. Evolution of protein sequences that are produced as a result of these processes can take at least two forms: evolution of contexts or evolution of sequences *per se*. In addition, there may be another level: evolution of the molecular machinery performing modifications, e.g. ADAR complexes in cases of A-to-I RNA editing or snRNAs in case of alternative splicing. Although theoretically this third route is possible, we deem it generally unlikely due to the following reasons. Firstly, respective polymorphisms will have large pleiotropic effects, which would make these polymorphisms less likely to be neutral or beneficial. Secondly, in case of adaptive evolution there would likely be recombination-driven loss of associations between beneficial variants and the variants of information-modifying machinery. Nonetheless, as in the case of stress-induced loss of translation termination in yeast, general modulations of decoding machinery may produce beneficial phenotypes in some extreme conditions (discussed below, "E-variance mediated adaptation" and "Evolvability").
2. In each case, there is some level of variance in decodings, i.e. expressed states, associated with the process. This variance may be associated with the variation in levels of modifications conveyed by individual sites or in the emergence of new modifications along the genome, e.g. variation in transcription levels and noisy transcription of random parts of the genome. However, the emergence of variance is in these cases at least partially governed by genetic variants present at respective contexts, and thus these contexts and hence ensembles of decodings may change by means of mutation, selection and drift.

Now we clarify some of the terminology used below. Here, we call the primary targets of selection, i.e. protein or RNA sequences, primary sites, and the secondary targets, i.e. contexts of information modifiers, secondary sites. Sequences of entities governing information modifications in general, e.g., ribosomes, small nuclear RNAs or polymerases, we call tertiary sites. Respectively, we refer to

selection at these levels as primary, secondary or tertiary selection. To refrain from introducing a large body of novel terms and use the widely accepted ones instead, we refer to decodings of primary sites as traits. Consequently, as we are dealing here with ensembles of decodings and ambiguous expressions for each primary site, we also discuss our topic using the terms “penetrance” and “trait variance”. For the latter, as mentioned before, we also employ Fisher’s approach and presume it to be comprised of three summands: variance explained by genetic polymorphism (G-variance), variance arising from different expression of alleles under varying environmental conditions (GxE covariance) and variance explained by environment (E-variance). The latter is comprised of variances arising due to specific conditions such as the effects of different environments population lives in, family effects, maternal effects etc. and the variance not assignable to any effect (Lynch and Walsh 1998). The discussed processes and the corresponding primary, secondary and tertiary sites in each case are summarized in Table 1.

Table 1 | Discussed information-modifying processes

Process	**Primary sites**	**Secondary sites**	**Tertiary sites**
Differential expression	Protein / RNA sequences	Promoter binding sites / transcription factors	Polymerase / histone genes
Alternative splicing #1	Protein sequences	Splicing sites	snRNAs
Alternative splicing #2	Splicing sites	Splicing enhancers / RNA-binding proteins	snRNAs
RNA editing	Protein / RNA sequences / splicing sites	RNA structural nucleotides / sequence contexts	ADAR / APOBEC complexes
miRNAs / siRNAs	Protein sequences	Sequences of pre-mi / siRNAs	Complexes governing RNA-interference
uORFs	Protein sequences	uORF sequences	Ribosome
Protein modifications #1	Protein sequences	Contexts of modifications	Non-site-specific modifying enzymes

Protein modifications #2	Protein sequences	Contexts of modifications / site-specific modifying enzymes	Higher-level regulators
Alternative polyadenylation	mRNA sequences	Polyadenylation sites	Polyadenylation and splicing machinery
Ribosomal frameshifting	Protein sequences	Frameshifting contexts	Ribosome

2.3. Evolution of information modifiers

Modifiers of biological information are typically considered as means to produce controlled diversity of genes, which is used in regulation of organismal development and in regulation of responses to changing environmental conditions (Ghalambor et al. 2015; Eisenberg and Levanon 2018; Wang and Burge 2008; Wang et al. 2015; Wang and Cooper 2007). Consequently, genetically encoded information undergoes famously frequent modifications in organs with the most complex development and with the largest degrees of functional plasticity, e.g. in neural tissues of vertebrates and molluscs (Su and Tarn 2018; Liscovitch-Brauer et al. 2017; Eisenberg and Levanon 2018). Modifications are so common that major fractions of genes in higher metazoans are known to be modified at some point of the organism's life in some subset of organs by, e.g., alternative splicing (Mironov, Fickett, and Gelfand 1999). While regulation is clearly the prevailing role of encoded information modifications, some authors suggested that the initial functions of some of these modifications had been different. For instance, it could have emerged as a mechanism to cope with the invasion of introns in early eukaryotic life forms (Koonin 2006). Alternative splicing thus initially had to be noise generated by errors of the splicing apparatus. Due to the contextual nature of splicing, this noise was heritable, and thus there was a potential for emergence of controlled alternative splicing purely by means of mutational process and selection.

Perhaps the first theory describing evolution of alleles modifying information expressed by other alleles is the Fisher's theory of dominance (Fisher 1928; Wright 1929). Fisher has proposed that, although originally all mutations are semidominant, subsequent selection on dominance-modifying

alleles changes the degree of dominance at focal sites, rendering deleterious alleles recessive and making beneficial alleles dominant. Initially criticized by Wright and other authors, this theory is generally not accepted today, however there are a number of examples, where it may be applied (Wright 1929; 1934; Charlesworth 1979; Lynch and Walsh 1998). In particular, this theory may be applied in cases of polymorphisms under balancing selection: long-standing polymorphism gives modifier mutations enough time to emerge and spread to fixation (Charlesworth 1979). As information modifiers may produce ambiguously decoded states highly reminiscent of heterozygotes (see below, "Balancing selection"), which are maintained through long periods of time, similar ideas may be applied here when we consider beneficial vs. deleterious decodings. A minor difference is that under our terminology we should call selection on modifier alleles selection on variants at secondary sites.

Rupert Riedl, an Austrian zoologist and an advocate of the extension of the Modern Synthesis argued that in the inferences of the general features of biological evolution we cannot ignore the constraints imposed by development and genetic architecture (Riedl and Auer 1975; Wagner and Laubichler 2004). In his book *Die Ordnung des Lebendigen* ("Order in living organisms") and in commentary article he raises a question in the context of the adaptation rates: "*...what would happen if independent genetic units, the structural results of which have become functionally dependent, were also to become epigenetically dependent, for example, by adopting a superimposed genetic unit upon which both are dependent, as in the case of two structural genes dependent on an operator gene?*" (Riedl and Auer 1975; Riedl 1977). Riedl's answer is that, in cases of genetic dependencies, adaptation may be by mutations in both operator and focal gene sequences leading to similar phenotypes. Thus, the waiting time for a beneficial change generally decreases. In the context of processes that we discuss, evolution at secondary sites may produce beneficial phenotypes and influence selective pressure on the primary sites as is the case with, e.g. RNA editing, where edited states, being consequences of the sequences at secondary sites, may mimic substitutions at primary sites (Popitsch et al. 2020; Moldovan et al. 2020).

Some authors point out the role of the noise generated by epigenetics in boosting phenotypic variance. In a review on the evolution of A-to-I RNA editing, Gommans, Mullen and Maas write: "*Substantial transcriptome and proteome variability is generated by A-to-I RNA editing through site-selective post-transcriptional recoding of single nucleotides. We posit that this epigenetic source of phenotypic variation is an unrecognized mechanism of adaptive evolution.*" (Gommans, Mullen, and Maas 2009). They put forward a theory which they call Continuous Probing Hypothesis (CoP), which states that a large number of adenines is subjected to sporadic low-level A-to-I RNA editing, which allows organisms to explore the mutation space without generating novel variants with high penetrance via non-synonymous substitutions it: "*The genetic variation introduced through editing occurs at low evolutionary cost since predominant production of the wild-type protein is retained.*".

A number of critical comments are due here. (1) If the generated variation comes at a low evolutionary cost, the alternative variants should have minor effects on the proteome, hence beneficial variants may experience no effective selection. Indeed, consider an adenine with a non-synonymous substitution to guanine edited in 1% of transcripts due to the noise, which is a reasonable assumption for A-to-I editing (Liscovitch-Brauer et al. 2017). Then, if we expect efficient selection at secondary sites on variants giving this 1% editing and there is no dominance, the coefficient of selection for guanine at primary site has to be quite large – at least of the order of 200/Ne, which is about 0.01 for the human population. This would make the sites effectively selected for their editing, even theoretically, a rather rare phenomenon, unless some polymorphisms at secondary sites yield dramatic increase in penetrance of editing. (2) Gommans *et al*., similar to Riedl, view A-to-I editing as a source of evolvability: "*...molecular mechanisms that increase genetic variability and/or flexibility, e.g., may increase the evolvability of species if they increase an organism's ability to express novel phenotypic variation in response to changing environments. If so, then selection may favor systems with high levels of genetic variance and/or significant phenotypic plasticity*" (Gommans, Mullen, and Maas 2009). However, evolvability may not be the central theme here. Indeed, whereas evolution of

evolvability presumes that characters increasing polymorphism should evolve for exactly this property, evolutionary value of A-to-I editing could have arisen purely by chance, i.e. be non-evolved, and evolution simply acts on the variants editing generates instead of acting on the cause of variance (see below, “Evolvability”) (Gould and Lewontin 1979; Partridge and Barton 2000). (3) If editing is a major driving force of evolution, positively selected A-to-G substitutions should in a sufficiently large number of cases bear signatures of editing, which so far has been shown only indirectly (Popitsch et al. 2020; Moldovan et al. 2020). 4) As we discuss later, secondary sites may overlap with other secondary sites as is the case with overlapping promoter sequences or may overlap with primary sites like in case of A-to-I editing site contexts (Moldovan et al. 2022). If such overlap happens, polymorphism at secondary sites will influence multiple traits and thus will be pleiotropic, and pleiotropy, although quite widespread in nature, is generally known to decrease evolvability by imposing additional constraints on evolution at a single locus (Hughes and Leips 2017).

We believe that some other views heritable information modifiers (HIMs) evolution pose a reasonable addition to the body of theoretical investigations performed so far. In particular, we would like to emphasize that evolution of HIMs follows the same routes as evolution of any other trait or process and involves mutation, selection and drift, yet may provide some insights into features of evolutionary process with regard to the genome architecture of living organisms. In subsequent chapters, we will consider evolution of HIMs and its relatedness to the evolution at primary sites in contexts of multiple widely accepted theories and discuss the effects HIMs may have on the evolution of organisms.

2.4. Epistasis

The term *epistasis* usually invokes two main connotations. Firstly, epistasis is a component in the trait variance associated with interdependencies of phenotypic effects of genotypes or alleles at different loci. This value is calculated for a given population and can vary depending on, *e.g.*, frequencies of alleles in the population (Lynch and Walsh 1998). Secondly and consequently, epistasis is thought of as the nonlinearity of genotype-to-phenotype maps, this nonlinearity resulting in their complex shapes. In the case of genotype-to-fitness map, termed fitness landscape, epistasis makes the landscape generally less passable, as a larger portion of genotypes accessible to a population at any time results in deleterious phenotypes (Smith 1970; de Visser and Krug 2014; Usmanova et al. 2015; Kondrashov and Kondrashov 2015). The rest of mutations are either effectively neutral or mildly deleterious or beneficial and comprise evolutionary trajectories that a given population can traverse. Epistasis in the form of amplification of deleterious allelic effects, termed negative epistasis, thus results in the further general narrowing of these trajectories, *i.e.* to a smaller number of tolerated mutations at any time. As it has been shown that negative epistasis is the prevailing form of nonlinearity in fitness landscapes, high degrees of epistasis have been proposed to slow down adaptation and increase genetic load by multiplying effects of mildly deleterious alleles (Carter, Hermisson, and Hansen 2005; Kondrashov and Kondrashov 2015; Bendixsen, Østman, and Hayden 2017).

In the case of information modifiers discussed in this essay, there is an evident epistasis between the primary and secondary sites. For instance, damaging of splicing sites by the mutational process results in changes of effects of alleles in respective exons to zero. Consider a trait influenced by variants at an exonic biallelic locus on average by the value s. The alleles at this locus, A and a, are at frequencies p and 1–p. Further, if a polymorphism at a splicing site arises, there would be two splicing site alleles: B and b on frequencies q and 1–q. Consider a polymorphism arising in linkage with the A allele, such that B results in a fully functional splicing site, and b – in a completely non-

functional one. Then the epistatic variance, in the absence of dominance and under complete linkage, will be sq(1–q), i.e. non-zero. Note that under no linkage as may be the case for, e.g., micro-RNAs, variance will be still non-zero: spq(1–pq). The same, i.e. epistasis arising from context-primary site interdependence, can be shown for other considered processes.

Interestingly, epistasis discussed here, at least theoretically, opens novel adaptive trajectories rather than closes the existing ones. This is explained by novel polymorphisms at contextual sites, that might otherwise be completely neutral, having the possibility to influence adaptation. Mathematically, this results from the addition of novel dimensions to the genotype-to-phenotype map brought about by additional sites (Moldovan et al. 2020). Moreover, the perspective of epistasis yields another nuance: in the above example, there is a polymorphism in the presence of an exon with some phenotypic effect. If the exon is ancestral, the epistasis between the exon and the splicing site is in fact limiting adaptation by excluding variants present in the exon. Analogously, Gommans *et al.* propose that in the case of A-to-I editing, adaptation is facilitated by the inclusion of synonymous polymorphisms, arising in the A-to-I editing site contexts, in the evolutionary process, or, in other words, by the creation of epistatic links between synonymous and nonsynonymous polymorphisms (Gommans, Mullen, and Maas 2009). However, if a contextual polymorphism is non-synonymous, editing creates an epistatic link between polymorphisms with phenotypic effects, which is expected to generally limit adaptation due to the arising pleiotropy at the contextual site (Hoekstra and Coyne 2007; Hughes and Leips 2017).

2.5. Polymorphism and polygenic traits

Due to the epistasis arising between primary and secondary (contextual) sites, the number of sites with effectively independent mutations in the genome decreases due to emergence of alternative mutations with similar phenotypic effects (Moldovan et al. 2020). Also, as mentioned before, the number of possible non-neutral mutations generally increases, as some of them are affecting contexts. These two factors result in the enhanced rates of non-neutral mutations supplied by the mutational process, thus potentially affecting the speed of adaptation, which Maynard Smith defined as the rate

of accumulation of beneficial variants (Maynard Smith 1976). Here, we define polymorphism as the population mutation rate $\theta = 4Ne\mu$, where μ is the per-site mutation rate (Maynard Smith 1976; Kimura 1983). Also it should be noted that polymorphism mostly varies due to varying effective population size N_e, as μ varies to a much lesser degree (Zuckerkandl and Pauling 1965; Kimura 1983; Bromham and Penny 2003).

But can increase in the population's genetic variability influence the rates of adaptation? On the one hand, natural populations and mutation rates are so large that given enough time any beneficial variant should arise multiple times and be eventually fixed. Moreover, the time needed for a beneficial variant to be fixed is typically not that large on the evolutionary timescale (Kimura 1983). Consequently, the data-supported evidence for more rapid accumulation of beneficial variants in highly polymorphic populations relative to low-polymorphic ones is scarce (Bell 2013; Rousselle et al. 2020). On the other hand, at least theoretically, polymorphism should influence the rates of adaptation, as θ enters the expression for the rate of adaptive substitutions k with selection coefficient s: $k = \theta s$ (Maynard Smith 1976; Kimura 1983). Moreover, as the differences in θ are mainly a consequence of differences in effective population sizes, and the relation of selection coefficient of an allele to the inverse of effective population sizes determines the regime of the allele's evolution, namely selection or genetic drift (Kimura 1983), in more polymorphic populations selection is expected to act more efficiently on mildly beneficial variants with s of the order of $1/N_e$. In simulation experiments, Lynch and Hagner demonstrated that selection of complex traits as well as the number of mutational paths available to populations (selective sieve) are largely dependent on the effective size (Lynch and Hagner 2015; Lynch 2020). However, the applicability of this result to real life should largely depend on the distribution of fitness effects in nature. Only recently, Rousselle et al. proposed a resolution to this problem. By analyzing the rates of accumulation of adaptive substitutions in metazoan lineages they have shown that in very low-polymorphic populations, *e.g.*, human, the influx of random mutations should in fact influence the rate of adaptation (Rousselle et al. 2020). If, however,

polymorphism in a considered population is above approx. 1%, the increased mutation rate does not increase the rate of adaptation. Thus, if a population experiences contractions in size or strong founder effects, naturally occurring mutations may not provide enough genetic variance for efficient adaptation.

Another line of discussion here is as follows. Selection acts at the level of traits, not variants (Gould and Lewontin 1979), and many traits are known to be influenced by multiple genetic loci (Lynch and Walsh 1998). Thus, if a trait is selected upon, multiple alternative alleles can be supported by selection, however this depends on the interaction between the alleles (epistasis). Indeed, if a beneficial trait arises only if multiple specific mutations are present (positive epistasis), such trait would most likely never emerge, as the probability of fixation, and hence the inverse waiting time, of specific variants at multiple loci is negligibly small (Riedl 1977). Another extreme is a trait that arises via fixation of any of alternative alleles (negative epistasis). In this case the waiting time of fixation of beneficial traits decreases exponentially with the number of alleles (Riedl 1977). In the absence of epistasis, adaptation still should be faster than in the case of a single-locus trait, as efficiently selected trait-inducing variants are expected to fix more frequently due to their numerosity. In support of this idea, it was shown that traits in low-N_e populations evolve faster if influenced by multiple loci (Lynch 2020).

Thus, surplus genetic variation provided by information modifications may in fact increase the rate of adaptation by providing additional variation in the form of polymorphic secondary sites. In some cases, *e.g.*, in the case of polymorphism in promoters yielding alternative gene expression and hence alternative phenotypes, polymorphic secondary sites may provide additional venues for adaptation (Hoekstra and Coyne 2007). In other cases, such as the A-to-I RNA editing mimicking A-to-G polymorphism, selection at secondary sites is somewhat alternative to selection at primary sites and hence decreases the intensity of the latter. Under this scenario, specific genetic patterns emerging at secondary sites make the decoding of a primary site a *de facto* polygenic trait and may fully

compensate for low polymorphism, insufficient for quick responses to primary selection (M. Moldovan et al. 2020).

Finally, at the end of this chapter, as we will do at the end of subsequent chapters, we would like to speculate about the analyses needed to support the hypothesis, which is in this case the ability of heritable information modifiers to promote adaptation. Suppose a population is subjected to selective pressure and there is a certain level of polymorphism associated with both primary and secondary sites. What fraction of adaptive changes do we expect to come from changes of frequencies of alleles at secondary sites and how will these changes influence selection at primary sites? Evidently, selection at secondary sites may promote adaptations, as demonstrated by multiple studies (Berg and von Hippel 1988; Berg, Willmann, and Lässig 2004; Mustonen and Lassig 2005; Mustonen et al. 2008; Lynch and Hagner 2015; Singh et al. 2017). In the case of gene promoters, there is even a (criticized) notion about secondary selection for promoters being a larger factor in adaptation than the primary selection on the sequences of genes (Hoekstra and Coyne 2007). Thus, the problem amounts to the search of secondary selection, which may be not an easy task. Indeed, whereas methods for selection inference in coding sequences are numerous, highly precise, and in many cases do not even require large amounts of data, inference of selection at a given set of sites in a general case requires information about local mutation rates, rates of allele frequency changes, or distributions of allele frequencies on haplotypes (Walsh, Lynch, and Lynch 2018).

Another question that should be discussed here is whether selection-driven mutations at secondary sites influence the intensity of selection at primary sites and *vice versa*, and in particular whether selection at secondary sites may compensate selection at primary sites. This problem is more complex than the one discussed above, as its solution requires not only the assessment of the magnitudes of primary and secondary selection, but also tests for epistasis between respective polymorphisms. In the context of evolution of low-polymorphic populations it may be checked whether heritable variation at secondary sites, even under conditions of modification-generated

variants being less beneficial than the hypothetical variants at primary sites, is still able to promote adaptation.

2.6. Exaptation

As Darwin has pointed out, basically any organ can have multiple functions, and selection can act on any one of them at any time depending on the demands imposed by the environment on the population (Darwin 1872). For instance, feathers in birds are hypothesized to have emerged as insulators, but later took on an additional function when the ancestors of birds started to enjoy the benefits of flight. In addition, feathers may be used in mating rituals, hunting, fishing, etc. (Ostrom 1974; 1979; Gould and Vrba 1982).

Following Darwin's idea, researchers coined the term "preadaptation" to describe selection on secondary functions of organs (Limoges 1976). However, the term is inconsistent with the Darwinian view of evolution as the product of *bona fide* random mutational process, as it has a teleological ring to it, and, consequently, its usage has been largely criticized (Gould and Vrba 1982). Gould and Vrba recognized this problem stating "*...we traditionally apologize for "preadaptation" in our textbooks, and laboriously point out to students that we do not mean to imply foreordination, and that the word is somehow wrong (though the concept is secure)*" and proposed a solution in the form of a novel term – exaptation. It was suggested that "*...characters, evolved for other usages (or for no function at all), and later "coopted" for their current role, be called exaptations.*". However, one can easily notice a problem with the concept of exaptation as well.

In the evolutionary-genomic point of view, especially at the level of individual nucleotides, the concepts of "function" or "usage" become elusive at best. Indeed, whereas in some cases we can explicitly say that specific nucleotides at specific loci increase fitness relative to other variants by means of, *e.g.*, enhancing strength of binding of DNA sequences with transcription factors (Berg and von Hippel 1988; Mustonen et al. 2008), for the overwhelming majority of known positively selected variants no function can be established (Booker, Jackson, and Keightley 2017; Walsh and Lynch

2018). Thus, from the pure methodological perspective, the concept of exaptation is not straightforwardly applicable to site-centered studies. Further problems may stem from a more general issue: the concept of "evolution for [function] or [usage]" itself. Firstly, the word "function" in a non-mathematical sense typically refers to some sort of purpose of the object, and thus is itself somewhat teleological, as the very word *telos* (in Greek) refers to *purpose*. In particular, the phrase "*...evolved for ... function...*" sounds in this regard as hardly compatible with Darwinian evolution. Secondly, "function" is typically described by a narrative expressed in a natural language, that is, from the *philosophia naturalis* point of view, arbitrarily and hardly formalizable in the vast majority of studied cases of natural selection at the molecular level. This manifests, in particular, in our seeming inability to assign functional roles to the majority of selected nucleotide variants, mentioned before.

This problem was addressed by Riedl in *Strukturen der Komplexität* ("Structures of Complexity") (Riedl 2000). Riedl argues that the word "function" does not need to have a teleological connotation, as functions "do not necessarily harbor intentions". He further elaborates on this topic by comparing two German expressions that can both be translated as "purpose": *functionale Zwecke* (functional purposes) and *Absicht* (intent). Riedl states that the former is not as much teleological as the latter, as it does not presume a subject with possible intents. Another solution to this problem, according to Riedl, is the usage of Aristotle's term *entelechy* to describe the purposes behind biological functions. Entelechy is defined as "that which bears its goal within itself", *i.e.* the purpose and the object are presumed inseparable. Riedl's suggestions, however, do not provide a full solution to this problem, as they are fit solely for descriptions of particular states such as the list of functions of particular structures in particular populations. When talking about evolution, *i.e.* a process, we cannot say that the function which is selected upon is the goal that organism bears within itself, as evolution acts at the level of populations, not organisms (Kimura 1983). Although we fully appreciate the convenience of the usage of the term "function" in descriptions of evolutionary processes, we deem it

necessary to point out that this term cannot be used as a founding concept of an evolution-centered theory due to its teleological nature.

But does all this mean that the concept of exaptation should be abandoned altogether? Exaptation can be alternatively seen as referring to an ambiguous state or a character that at some point starts experiencing selection acting on one of the forms. In particular, the ambiguous information decoding may yield a variant on which selection starts acting at some point can be viewed as an instance of exaptation, just in a slightly modified definition. Indeed, if we substitute the term "function" in Gould and Vrba's definition with "phenotypic manifestation", and, accordingly, the word "character" with the word "locus", we get the new definition of exaptation: "Loci at which selection has acted on the phenotypic manifestation that has initially arisen due to inner workings of decoding machinery and experienced no selection". Note that under the novel definition, just like under the classical definition of exaptation, there are generally two possibilities: i. A site with a non-stationary selection pattern that had gained some additional phenotypic manifestation at some point, that later had become beneficial. ii. A [neutral] site with an ambiguous phenotypic manifestation, at which the initially secondary manifestation begins to be selected for, *e.g.* a site occupied by adenine experiencing A-to-I RNA editing may be selected for the presence of inosine at the site (Gould and Vrba 1982a; Gommans, Mullen, and Maas 2009; Popitsch et al. 2020).

Thus, virtually any coding site can be viewed as exaptation in the new sense due to at least the first noted possibility, much like virtually any character can be viewed as exaptation in the classical sense. As for the second possibility, it may be argued that selection for secondary manifestations in some cases, like A-to-I RNA editing, is identical to the selection on variants of the primary manifestation. However, this may not always be the case. Firstly, selection may promote proteome diversity, in which case the intensity of selection would depend on dominance and penetrance of the emerging variants (Eisenberg and Levanon 2018; Walsh and Lynch 2018). Secondly, as discussed above, exaptation in this case may be viewed as a route of adaptation for, *e.g.*, variants experiencing

strong selection or variants in a low-polymorphic population (Moldovan et al. 2020; Popitsch et al. 2020).

So, the general scheme of adaptation in the form of exaptation is, much like in the case of excessive polymorphism discussed above, the following. (1) At a locus, a heritable secondary manifestation arises due to mutational noise at secondary sites. For instance, an alternative splicing site emerges. (2) Initially, there is no selection for the secondary manifestation, however later it becomes selected upon. (3) The new manifestation becomes established, either on par with the primary one or is canalized. We return to the problem of canalization vs. variation later (see Genetic Assimilation). Thus, the search for exaptations in the sense presented here should amount to the search for non-stationary selection in secondary sites.

Probably here an important question must be asked: why do we mention exaptation here and go as far as to coin a new definition for it? Firstly, we believe that the term provides an important distinction between exaptation and adaptation, *i.e.* it allows for a deeper description and, consequently, better understanding of the evolutionary process. Secondly, as Gould and Vrba put it: *"Together, these two classes of characters, adaptations and nonaptations, provide an enormous pool of variability, at a level higher than mutations, for cooptation as exaptations."*, *i.e.* the characters that are initially not selected upon together with non-stationary selection represent an additional and underappreciated source of both evolutionary variability and restrictions imposed on the evolutionary process.

2.7. Phenotypic plasticity / Accommodation

Waddington and Schmalhausen have proposed, independently, that the development of traits has a general tendency to be canalized in evolution, either by means of negative selection acting on genes or through evolutionary constraints imposed on their regulators (Waddington 1942; Schmalhausen 1949). As a result, the variance of trait values and the ability of organisms to respond to environmental cues decreases (West-Eberhard 2003; 2005). However, if the variance of some fitness-related character decreases dramatically and little diversity of respective trait values is observed

in a given population, such population is expected to suffer from changing environments more than populations exhibiting diverse responses to changing environment, or phenotypic plasticity. Note that under the described model the variance behind phenotypic plasticity is the genotype-to-environment interaction covariance (GxE) (Lynch and Walsh 1998). In the framework of this essay, GxE arises mostly due to evolved responses of information-modifying systems to environmental stimuli, *e.g.* by differential expression or alternative splicing. However, there are some striking examples where phenotypic plasticity yields beneficial phenotypes even in the absence of apparent prior selection, *e.g.* the two-legged goat described by Slijper in 1942 (Slijper 1942), that, while having abnormal structure of the front legs, was nevertheless able to walk on hind legs due to profound adjustments of the skeleton and muscles. As the effects of GxE on adaptation are poorly studied, it has not yet been firmly incorporated in any evolutionary theory. However it is thoroughly discussed and there are four mutually excluding theoretical options (Ghalambor et al. 2007; Ghalambor et al. 2015; Markov and Ivnitsky 2016):

1. Phenotypic plasticity is adaptive and results in slower rates of evolution. If under novel conditions plastic response produces a fit phenotype, there can be simply no selection that would result in shifts in variant frequencies (Waddington 1961; Ancel 2000; Price, Qvarnström, and Irwin 2003; West-Eberhard 2005; Garland and Kelly 2006; Paenke, Sendhoff, and Kawecki 2007; Pfennig et al. 2010; Fitzpatrick 2012). This effect should mostly arise either in populations evolved to inhibit various or changing environments or upon small environmental changes. Here, plasticity acts as a phenotypic capacitor that shields organisms against the perils of existence (West-Eberhard 2005).
2. Plasticity is adaptive and results in enhanced rates of evolution. In this case, the adaptive plastic response results in lesser numbers of individuals in a population falling prey to the changing environment, resulting in the preservation of a larger portion of the genetic variance and lower waiting times of beneficial mutations (Waddington 1942; 1953b; 1959; Ancel 2000; Price,

Qvarnström, and Irwin 2003; Crispo 2007; Ghalambor et al. 2007; Ghalambor et al. 2015). Although this option can be tied with the Baldwin effect or Waddington's genetic assimilation (Ghalambor et al. 2015; Markov and Ivnitsky 2016), as we discuss later, this association may be a fallacy in the general case.

3. Plasticity is nonadaptive and results in slower rates of evolution than in the case of a non-plastic population. Plasticity also can generate disadvatageous phenotypes such as decreased fertility and negatively affect other fitness-related characters, which results in smaller numbers of surviving individuals (Minelli and Fusco 2010; Fitzpatrick 2012; Markov and Ivnitsky 2016). Although in this case the effect on the rate of evolution is somewhat indirect, one can argue that sufficiently extensive elimination of plastic non-adaptive individuals results in smaller genetic variance. And smaller genetic variance results in the possible absence of variants selection can act upon, which, in turn, results in lower rates of evolution. An alternative is, of course, that a plastically-nonadaptive population is pushed out of existence by competitors and ultimately stops evolving (Markov et al. 2016; Markov and Ivnitsky 2016).
4. Plasticity is nonadaptive and results in enhanced rates of evolution. Here, plasticity generates less adaptive phenotypes that put more selective pressure on existing beneficial variants. As a result, beneficial variants spread more rapidly and in addition more variants become efficiently selected. The terms "genetic compensation" and "counter-gradient variation" are usually used to describe this process (Grether 2005; Conover, Duffy, and Hice 2009; Ghalambor et al. 2015).

Two notes are in order here. Firstly, options (1) and (3) become absurd if by "rates of evolution" we presume rates in changes of characters, not genotypes (West-Eberhard 2005). Secondly, although the listed options are evidently mutually exclusive, they may all be realized for any given trait depending on the initial population and conditions the population is put into. For instance, if a population experiences minor environmental perturbations, no selection would arise, hence option (1) will be realized, otherwise evolution will follow any of the scenarios (2-4).

Information modifiers are known to produce phenotypic plasticity as a result of alternative modifications, *e.g.* alternative splicing. The latter may be controlled, and thus contribute to the complexity of responses demonstrated by various organisms. Two of many examples are alternative splicing controlling cadmium tolerance in *Arabidopsis* (Zhang et al. 2014; Liu et al. 2022) and RNA editing regulating temperature responses in coleoid cephalopods (Garrett and Rosenthal 2012). Thus, information modifiers seem to contribute beneficial phenotypic plasticity and thus may provide sources for large-scale studies of evolution of plasticity. For instance, alternative gene expression has been studied exactly in this regard (Ghalambor et al. 2015; Ho and Zhang 2018).

The problem with most studies of the adaptiveness of plasticity is that adaptiveness is verified solely from genetic assimilation of plastically-acquired characters. For instance, Ghalambor *et al.* counted as a sign of adaptive plasticity only concordant changes in gene expression upon introduction to a novel environment before and after adaptation. As a consequence, for only a small fraction of genes beneficial plasticity was established, the rest of the genes deemed to produce nonadaptive plasticity (Ghalambor et al. 2015). However, we know that evolution can take different routes, and adaptation is rarely restricted to a single sequence of beneficial changes. For the sake of clarity, we will take a non-molecular hypothetical example: behavior of amphibians in arid environments. If we take a frog population initially adapted to more humid areas and put it into a more arid environment, the frogs will have longer stays in water than in initial humid areas, demonstrating an evidently beneficial plastic response. Although the subsequent selective pressure may be on adaptation to the life in water, it may also follow another route: development of larger lungs and dryer skin. The result of the latter route will be frogs spending less time in water. If we look solely on one trait: percentage of time spent in water by frogs, we will name the initial plastic change non-adaptive, as subsequent evolution reversed this change. The lack of such intuition in molecular evolution studies may lead us to potentially false claims about adaptiveness of plasticity. The more routes adaptation can take in a particular instance or (possibly) the larger is the genetic variance in a population, the more plastic

changes will be labeled as non-adaptive. Thus, we propose that the benefits of phenotypic plasticity should be viewed separately from the phenomenon of genetic assimilation.

And so, we have two questions that need to be answered in the paradigm presented here: whether plasticity generated by information modifiers is adaptive and how it induces the rates of evolution. If such plasticity is adaptive, we should expect selection on secondary sites generating ambiguous states of primary sites. We should also see more of such selection on secondary sites in populations living in spatially or temporally varying environments, where we expect varying selection patterns in primary sites (Mustonen and Lassig 2010). With a necessary verification of modifications of secondary sites indeed producing differences in phenotypic manifestations of primary sites in different environments (reaction norms), we would be able to assess adaptive potential of phenotypic plasticity conferred by information modifiers. In addition, we may look into the benefits of increased plasticity upon introduction into completely novel environments by assessing competitiveness of plastic non-specialized population vs. canalized specialized ones.

Another question is posed by the influence of plasticity on rates of evolution at primary sites. We should note here that differential selection on primary sites, which we expect in varying environments, is expected to yield more selection in them, and thus studies which compare the rate of evolution in varying vs. constant environments may find simply a positive association between plasticity and the rates of evolution at primary sites, not the former leading to the latter. The causal relationship can be established here if we look at evolution of populations introduced to novel environments (Ghalambor et al. 2015). Like in the previous paragraph, one may assess evolution rates in initially canalized populations and in their initially plastic counterparts in novel environments.

2.8. Genetic assimilation

As we have noted before, genetic assimilation of an initially plastic character is an important special case of evolutionary scenarios involving beneficial phenotypic plasticity. As noted by Baldwin and Waddington, plastic changes of traits may be supported by selection acting on existing genetic variation, in which case the novel values of traits become the "default" values. This process is referred to as genetic assimilation of initially plastic characters, or simply as genetic assimilation (Waddington 1942). In addition, genetic assimilation presumes reduction in plasticity: assimilated traits no longer experience plastic responses, or at least experience them to a lesser extent than before the genotypic change. In his classical experiments, Waddington demonstrated that the initially rare environmentally-induced *crossveinless* phenotype in *Drosophila melanogaster* can be efficiently selected upon and become prevalent in a fruit fly population. Moreover, after the exposure to selection for several generations, this phenotype did not require environmental stimulus to appear even when selective pressure was lifted. Thus, selective pressure first assimilated and then canalized the *crossveinless* phenotype by making it no more plastic (Waddington 1953). However, it is worth noting that Waddington's results may be viewed in terms of exposure of hidden genetic variation by an environmental stimulus followed by negative, instead of positive, selection on the revealed variation (Schmalhausen 1949; Pigliucci, Murren, and Schlichting 2006). After Waddington's experiments, more examples of genetic assimilation of environmentally-induced traits have been obtained along with a number of counterexamples, for reviews see, *e.g.*, (West-Eberhard 2003; Pigliucci, Murren, and Schlichting 2006; Crispo 2007; Ghalambor et al. 2007; Levis and Pfennig 2016; Markov and Ivnitsky 2016).

By definition, genetic assimilation requires beneficial phenotypic changes to be induced by phenotypic plasticity (Waddington 1942; Ghalambor et al. 2007). But in this case beneficial plastic changes have to be additionally insufficient for adaptation: if plastic changes in a trait happen to be sufficient for trait-affecting variants to become neutral or deleterious, genetically-driven adaptation

cannot possibly occur due to the lack of selection (West-Eberhard 2005). Under this scenario, genetic drift remains the sole force with a potential to produce heritable changes in traits, and plastic changes may be fixed via neutral accumulation of trait-inducing variants supporting plastically acquired change. Additionally, one could, following Schmalhausen and Waddington, argue that adaptation can further decrease the variance of a trait, either by means of positive selection on a subset of alleles (Waddington 1942) or via negative selection on another subset of alleles (Schmalhausen 1949). However, note that if the trait variance in the form of, *e.g.*, genetic control over trait values is viewed as a separate trait, the same general line of reasoning as the one laid out above can be applied, *i.e.* there may be no selection on the phenotypic variance if plastic changes are sufficient to produce an [effectively] optimal range of phenotypic values.

Information modifiers can be, at least theoretically, subject to evolution via intensive genetic assimilation. Indeed, as in the previous chapter, consider an ensemble of modified states that arises as a programmed response to environmental stimuli. If the novel plastic ensemble is beneficial, evolution of the average ensemble composition can take five routes: (1) Plastic change assimilated or reinforced by secondary selection (selection on alternative splicing sites), (2) Plastic change assimilated or reinforced by primary selection (selection on A-to-G substitution at A-to-I editing site), (3) Plastic change assimilated by neutral mutations, (4) Plastic change is not assimilated, (5) Plastic change is reversed by an evolutionary process acting on another route of adaptation than the one posed by the plastic change. Evidently, the evolution of variances in ensemble composition has the same possibilities. If one considers the Baldwin and Waddington effects separately (for review see Crispo 2007), the observed effects should be the following.

For the Baldwin effect, i.e. genetic assimilation of changes in the trait mean with no Waddington's canalization, we should observe effects i-iii on the ensemble mean with no decreases in the GxE variance. For the Waddington effect, canalization of plastic traits, we should observe effects i-iii on the mean *and* a decrease of the GxE variance. However, as the assessment of gains and losses

in plasticity may pose quite a challenging task, the term *genetic accommodation* can be used here to describe the general synergy of plastic and adaptive effects (West-Eberhard 2003).

As noted, genetic assimilation requires beneficial plasticity, which has been discussed in the previous chapter. Given a beneficial plastic response, to verify the hypothesis presented in this chapter we should next test for selection that assimilates this response. For instance, given beneficial plastic changes in RNA editing patterns, we should see either selection reinforcing these changes or selection modifying the genomic blueprint in accordance with the plastic change. In general, we should see primary or secondary selection producing the same phenotypes as the plastic changes. Considering distinctions between the Baldwin and Waddington effects, one should additionally test for the preservation of GxE variance in the ensemble of informational molecules after the episode of selection.

2.9. E-Variance-mediated adaptation

In theory, organisms may adapt not only by changing the mean values of traits, but also by changing the variance. The variance of traits can decrease, thus canalizing the trait, a thoroughly studied process first predicted and described by Waddington and Schmalhausen (Waddington 1942; Schmalhausen 1949). But can the increase in variance of trait values promote adaptation? Here, it is important to distinguish between several types of trait variance: genetic variance, genotype-environment interaction covariance, and environmental variance (Fisher, 1918; Lynch and Walsh 1998).

If we are talking about genetic variance, the increase in variance par excellence means the increased potential to evolve due to larger fractions in trait variance explained by additive genetic variance (Lynch and Walsh 1998). Exceptions here can stem from enhanced genetic variance being explained mainly by non-linear effects, *i.e.* dominance and epistasis. Genetic variance is the main studied source for adaptation, and the link of heritable information modifiers with it has been previously discussed (see 2.4. Epistasis, 2.5. Polymorphism and 2.6. Exaptation) and will be in some form discussed later (see 2.10. Balancing selection and Evolvability).

The genotype-environment interaction covariance conveys specific modifications of phenotypes in response to environmental perturbations, either spatial or temporal (Ghalambor et al. 2007). The effects on environment-dependent phenotypes on adaptation were also discussed (see 2.7. Phenotypic plasticity, 2.8. Genetic assimilation).

Environmental variance and its potential to influence accommodation and adaptation are rather poorly discussed both in our essay and in general. A probable reason for that is the rather limited theoretical possibility of positive influence of the increase in environmental variance on adaptation. Indeed, consider a toy example with a simple uniform distribution of fitness in a population with respect to constraints imposed by environment on some trait: a range of phenotypic values of a trait available to an initial population ranging from 1 to 11 that matches the respective optimal value range of 1 to 11, and the fraction of individuals in population with optimal trait values is thus equal to 1. Let us further assume that the novel environment introduces a novel optimal value range, which is the old one shifted by 1 to the right, i.e. 2 to 12. Thus, the fraction of individuals with optimal values becomes 0.9. If further accommodation or adaptation increases the mean trait value or decreases the variance, more individuals will fall into the optimal range and the fitness of a population increases. If, however, the variance increases and the range of values in population becomes, *e.g.*, 0 to 12, the fraction of optimal individuals decreases reaching ⅚, which is smaller than the initial 0.9. Does it mean that the increase in variance, at least in our example, ultimately decreases fitness, and canalization of traits through decrease in environmental variance is the only theoretically possible mode of variance-affecting evolution? Continuing our example, let us assume that an optimal trait value range in a novel environment dramatically differs from that in the old one and is, e.g., 10 to 20. Thus, the fraction of fit individuals is reduced to 0.1. If, like in the previous case, the trait variance in population increases and the range becomes 0 to 12, the fraction of fit individuals is increased to ⅙. This points at the benefits of increased trait variance during substantial environmental changes. In our example, the small increase in variance becomes beneficial when the mean of the optimal value range is shifted by more

than 5 points. With regard to the real data, two main questions arise here: i. How frequent are large environmental perturbations that could require the environmental variance of some traits to be either inherently large or to increase under stressful conditions? ii. What is the frequency of variance-mediated adaptation relative to adaptation with the changes in mean values?

Both of the stated questions cannot be quantitatively answered at the moment, however there are a number of facts that should be mentioned here. Firstly, the system of stress-induced translation termination suppression in yeast seems to be an example of an environmental variance enhancer in protein sequences, as it introduces random non-heritable phenotypic changes (True and Lindquist 2000; Partridge and Barton 2000; Koonin 2012). The existence of such a system hints at the possibility of frequent and substantial environmental changes that affected molecular evolution of yeast to such an extent that a variance-enhancing system emerged. Secondly, there seems to be a certain level of noise associated with all biological information-transmission processes, even central ones such as transcription or translation (Libby and Gallant 1991; Ou et al. 2019). For instance, the degree of editing of RNA molecules does not follow the expected binomial distribution, but has also a large and significant component of added variance (Harjanto et al. 2016a). Also, the generally enhanced variance in phenotypic manifestations of some nucleotides creates non-heritable heterozygote-like states (Harjanto et al. 2016a; Alon et al. 2012; Eisenberg and Levanon 2018). These states may theoretically buffer organisms against balancing selection, which is known to ultimately increase segregational loads (Kimura 1983) (see below). Thus, although the criteria for variance-mediated adaptation are seemingly strict and organisms seem to evolve in the direction of decreasing variance rather than increasing one, variance-mediated adaptation could be a potent evolutionary force in some extreme cases.

Heritable information modifiers are mostly noisy. In other words, they produce traits with evident environmental components to their variance, and thus may act as sources of variance-mediated adaptation. Here, we are faced with a seeming absurdity, as we cannot say that anything in biological

systems is beneficial unless it can be inherited through multiple generations, and environmentally-driven noise is not heritable. While that is most certainly true, we should note that changes in secondary sites in information modifiers can affect the variance of a trait with little or no influence on its mean value, while being heritable by definition. For instance, mutations in a transcription factor may compromise precision of its binding with promoters (O. G. Berg and von Hippel 1988), multiple possible conformations of the local RNA structure may enhance variance of RNA editing, changes in snRNA sequences may cause them to act non-specific etc. An already mentioned stress-induced system in yeast that blocks translation termination and thus yields proteins with random C-ends is also a relevant example (True and Lindquist 2000; Koonin 2012), as it increases the general translation noise.

But how do we detect the discussed effect? At the level of organisms, we can postulate that there are certain systems that introduce random noise to the expressed information: (yeast prions/HSP-mediated variance/stress-induced mutagenesis to some degree etc.). At the level of individual primary sites, we may look for selection on mutations in secondary sites that increase the variance of expression of the former. While that would be a direct test, in a lot of cases it requires huge volumes of data, *e.g.*, data on variance in expression values depending on the sequence of transcription factor or its binding sites coupled with fitness advantage conferred by each version of transcription factor or binding site. Mostly such data is simply not there, and a non-direct analysis here would be to look at the variance in ensembles of primary sites themselves, *i.e.* at variances in expression levels, RNA editing levels, alternative splicing frequency etc. If we will see the tendency for increase in variance or evident added variance, as in the discussed case of RNA editing, that would be an indirect proof of benefits conferred by enhanced environmental variance. The alternative that should be reckoned with is the additional variance that emerges simply by chance due to, *e.g.*, dependencies in modifications, such as correlated gene expression or RNA editing.

2.10. Balancing selection

Balancing selection refers to selection on the frequency of alleles in a population (Walsh and Lynch 2018). It may arise generally via three processes: (i) as a consequence of direct heterozygote advantage over homozygotes, as is the case with sickle-cell anemia variants in the human sub-Saharan population (Allison 1954). (ii) As a consequence of frequency-dependent selection, like in case of various MHC alleles, where diversity is additionally maintained by different, yet repeating epidemics (Lewontin 1958; Penn, Damjanovich, and Potts 2002; Ejsmond, Babik, and Radwan 2010). (iii) As a consequence of selection varying in time or space (Mustonen and Lassig 2010; Bertram and Masel 2019).

Methods used to infer balancing selection rely mostly on two facts: i. Balancing selection generates stable standing genetic variation (Kimura 1983), hence loci with alleles that converge on times larger than the expected coalescent time for all loci, or even loci with standing polymorphism that holds between species, are considered to evolve under balancing selection (Bubb et al. 2006). ii. Balancing selection generates states with intermediate frequencies of minor alleles, and thus footprints of recently emerged selection can be detected as regions (haplotypes) with the excess of intermediate-frequency alleles (Navarro and Barton 2002; Charlesworth 2006; Walsh and Lynch 2018).

Although heterozygote advantage, termed overdominance, and frequency-dependent positive selection are theoretically plausible scenarios and multiple instances of balancing selection have been observed, a small number of sites have been shown to evolve under this regime. For instance Leffler et al. found only 125 candidate loci in the human genome with signatures of balancing selection (Leffler et al. 2013). This might be a consequence of balancing selection ultimately introducing segregational loads to populations (Kimura 1983). Indeed, in case of overdominance, if there is heterozygote advantage, some fraction of gamete matches would inevitably carry deleterious homozygous alleles (Crow 1958; Kimura 1960). If there is frequency-dependent selection, there will be inevitable deviations of allele frequencies from optimal values due to genetic drift (Kimura 1983).

These deviations will lead either to fixation of one of the alternative deleterious variants or to segregational loads (Kimura 1983). In case of space- or time-dependent selection, in a population there will inevitably be individuals that live in the wrong place or in the wrong time, hence, again, creating segregational loads. It means that even theoretically, balancing selection cannot be a major evolutionary force on par with, e.g., positive selection, as many loci evolving under balancing selection at one time will create segregational loads of the order of sum of loads conveyed by individual loci.

However, the fact that we do not see much balancing selection does not necessarily mean it does not arise frequently and is bypassed due to changes in local fitness landscapes arising from epistasis or from genotype-environment interactions. In particular, as processes such as, e.g., RNA editing generate heterozygote-like states due to incomplete penetrance of the vast majority of editing sites, balancing selection, at least in the form of heterosis, may theoretically be mimicked by positive selection on secondary sites. This holds for other discussed processes such as alternative splicing, where balancing selection on heterozygous splicing sites may be bypassed by expressing both splice forms at the same time.

As discussed earlier, balancing selection manifests as standing polymorphisms or excess of intermediate-frequency variants. Thus, to assess the propensity of information modifiers to act as substitutes for heterosis, one needs to estimate the probability of primary sites with ambiguous phenotypic manifestations to be homologous to sites evolving under balancing selection. In addition, positive selection on secondary sites yielding this ambiguous decoding should be tested.

Two final notes are due here. Firstly, there seems to be an obvious similarity between some theories concerning evolution of information modifiers (Eisenberg and Levanon 2018) and early ideas about heterozygote advantage, which were eventually debunked. In his influential "Ecological Genetics", Edmund Brisco Ford states that heterozygous states will have *"...nothing but advantage and be superior to homozygotes which will have both advantage and disadvantage."*. (Ford 1971). This notion about heterozygotes being beneficial simply by virtue of being heterozygotes can be

paralleled with most recent ideas about information modifiers introducing beneficial diversity to the transcriptome, which are discussed here (Eisenberg and Levanon 2018). The transcriptome diversity by itself is in this case mostly deemed advantageous much like heterozygous states were deemed advantageous in 1960-s. Secondly, we emphasize that even if the majority of ambiguous decodings of primary sites are advantageous, this advantage seemingly cannot be ultimately explained by ambiguous decoding and proteome diversification being beneficial in itself, i.e. that there is a heterozygote advantage. Indeed, although information modifiers provide for populations an elegant scenario of escape from the deleterious effects of overdominance, there are simply too few known cases of the latter (Leffler et al. 2013; Walsh and Lynch 2018). Thus, although some ambiguously decoded primary loci may be explained by the benefits of diverse proteomes, the bulk of them requires another explanation, e.g., benefits stemming from regulation or a neutral alternative: decoding noise or neutral evolution of secondary sites. This issue is a good subject for future research.

2.11. Evolvability

The current paradigm of evolutionary biology postulates that all organisms, both extant and extinct, are or have been adapted to certain environmental conditions through natural selection of variants influencing biological traits. On par with that, the bulk of mutations arises via neutral evolution (Gould and Lewontin 1979; Kimura 1983). Naturally, evolution is impossible without, on one hand, genetic variation translatable into phenotypic variation and, on the other hand, the existence of adaptive trajectories: arrays of subsequent mutations such that each mutational step results in a fitter phenotype (Smith 1970). Other factors, such as the time required for a beneficial variant to emerge in an evolving population, the number of adaptive trajectories available at every given moment, lengths of adaptive trajectories yielding a fitter phenotype, narrow-sense heritability of traits *etc*., constitute the propensity of organisms to evolve, termed evolvability (Pigliucci, Murren, and Schlichting 2006). Additionally, the propensity of populations to produce genetic and phenotypic variation is termed variability (Wagner and Altenberg 1996). Although multiple definitions for evolvability, focusing on

various aspects of evolutionary process named here, have been proposed, they all are not mutually exclusive and generally sum up to a number of parameters describing evolutionary process. In some sense, trait variance, epistasis, phenotypic plasticity, and exaptation discussed above may also be regarded in this context.

Under practically any definition, evolvability can be measured as genetic variance, rate of adaptation, number of adaptive states, *etc*., and its existence is not debatable, as evolution of organisms can be observed and has been observed (Walsh and Lynch 2018). What is a subject to a current debate is whether evolvability is itself evolvable (Riedl and Auer 1975; Partridge and Barton 2000; Earl and Deem 2004; Charlesworth, Barton, and Charlesworth 2017). Evolvability is mostly studied as variability defined as the rate of mutational process, and, if there are benefits to an increased mutation rate, we would observe selection for alleles associated with enhanced rates of production of novel variants (Wagner 1981; Pigliucci 2008). The problem here is that such alleles *per se* cannot be selected for unless associated with beneficial variants that are produced as a consequence of enhanced mutation rates (Wagner 1981; Partridge and Barton 2000). This presumes linkage between variance-enhancing variants and beneficial variants. Such linkage is present in the vast majority of prokaryotic genomes, where enhanced mutation rates were observed to be beneficial by association with novel beneficial variants (Sniegowski, Gerrish, and Lenski 1997) and where we see systems enhancing mutation rates under stressful conditions (Ram and Hadany 2014). However, in sexually reproducing organisms, where the uncoupling of variants through recombination is common, the probability of linkage between evolvability-enhancing variants and beneficial variants may be too small to yield efficient selection for the former (D. Charlesworth, Barton, and Charlesworth 2017). Thus, studies postulating selection for evolvability in sexually reproducing populations have been criticized.

Another question that arises here is, whether selection for beneficial variants that arise due to evolvability-enhancing variants, counts as selection for evolvability-enhancing variants. The answer is probably negative, as selection in this case acts on beneficial variants, not on their cause. Otherwise,

we would have to recon effectively any episode of selection as selection for the mutational process itself. However, we see two lines of discussion here.

Firstly, selection acts not at the level of variants, but on the level of phenotypes of their carriers – populations of organisms. If an organism has a beneficial phenotype, a part of which is an enhanced mutation rate, we may count enhanced mutation rate as beneficial, as it constitutes a part of a beneficial phenotype. This rhetoric is highly problematic, as it does not allow us to reduce selection on a phenotype to selection on specific traits and to selection on specific variants. If we do so, we will be led to the starting point of this discussion, *i.e.* whether we can postulate selection of evolvability-enhancing variants. Thus, evolution of evolvability seems in this context a meaningless question, and, following Ludwig Wittgeschtein's approach, should be dissolved rather than solved.

Secondly, selection does not necessarily act on variants in a single population. Consider two competing non-mating populations, one of which is capable of generating beneficial variants faster than its competitor by virtue of enhanced mutation rates (G. P. Wagner 1981). If these two populations were to compete in various environments and a more evolvable population would outcompete a less evolvable one in the majority of situations by more rapidly generating various sets of adaptive traits, this would mean that evolvability may in fact be an evolved trait. However, such analysis is highly complicated.

As discussed above (see 2.5. Polymorphism), heritable information modifiers may enhance the observed heritable and non-heritable variance of traits, thereby potentially contributing to evolvability of organisms. The source of this evolvability is, on the one hand, the enhanced numbers of mutations affecting phenotypes in the form of protein sequences, and, on the other hand, mutations in machinery behind the discussed processes, like snRNA complexes forming spliceosomes, ADAR complexes performing A-to-I RNA editing, ribosomal proteins and RNAs, *etc*. The sequences of elements of machineries behind heritable information modification are the third category of sites that one may consider here. The effects of secondary sites on evolvability have been discussed above, and so here

we should briefly discuss the idea about the evolution of evolvability conferred by information modifiers in general, *i.e.* tertiary sites. Information modifiers serve mostly as means of regulation of responses to various environmental cues, and hence seem highly unlikely to have emerged as enhancers of adaptation rates. If we are talking about adjustments to the optimal adaptation rates in times of high selectional pressure on populations (Eldredge and Gould 1997), there may be a degree of selection on tertiary sites (e.g. responsible for ribosome fidelity), *i.e.* on the weakening of constraints imposed on information decoding. However, such selection would lead to non-heritable states arising as consequences of enhanced E-variance (see above), and the enhancement of mutation rate seems in this regard a more plausible scenario. Generally, it seems that the question of evolvability is in this case restricted to adaptation arising simply by means of accidental beneficial mutations in secondary sites, and hence there should be no selection on general lowering of the fidelity of information decoding.

2.12. Expectations

In descriptions of analyses presented in the last paragraphs of previous sections there is a degree of redundancy. These analyses generally sum up to the establishment of a small number of facts, which we discuss here.

1. Does secondary selection exist? In each of the discussed processes this fact should be established separately. For instance, for promoter sequences secondary selection has been shown and thoroughly studied (Mustonen and Lassig 2005; Mustonen et al. 2008).
2. To what degree can secondary selection pose an alternative to the primary selection, *i.e.* is there negative epistasis? Theoretically, changes in expression patterns or splicing patterns may compensate for changes in protein sequences (see 2.4. Epistasis), however this remains to be estimated. This question deals directly with the debate about the main route of evolution being changes in proteins *vs.* changes in regulatory patterns (Hoekstra and Coyne 2007).

3. Secondary selection and polymorphism. Although it has been shown that low-polymorphic populations adapt faster if novel variants are randomly introduced to their genetic pool (Sniegowski, Gerrish, and Lenski 1997; Rousselle et al. 2020), the degree to which secondary selection may relieve populations from the burden of lag load in this case remains to be estimated. Here, we will need either to compare the pressure of secondary selection relative to the pressure of primary selection in low-polymorphic and in highly polymorphic populations, or to estimate the dynamics of primary-to-secondary selection ratio along phylogenies of organisms with changing levels of polymorphism.
4. Non-stationary secondary selection and exaptation. Do we see a pool of neutral variation in heritable information modifications, both genetic and environmental, that is utilized to population's adaptive advantage upon changes of selection patterns? In particular, here we may study, e.g., variable transcription that is canalized in novel environments.
5. Secondary selection and phenotypic plasticity. Just like in the previous case, here we need to study variation of modifications that is subsequently canalized. However, this variation should be specifically of the genotype-to-environment type (Ghalambor et al. 2015).
6. Secondary selection as a substitute of balancing selection. Here, we should observe orthologous positions with modified sites in some species and balanced polymorphisms in sister species.

Chapter 3. Adaptive evolution at mRNA editing sites in soft-bodied cephalopods

This chapter describes our study "Adaptive evolution at mRNA editing sites in soft-bodied cephalopods" published in PeerJ in 2020. Authors: Moldovan, M., Chervontseva, Z., Bazykin, G., & Gelfand, M. S.

The bulk of variability in mRNA sequence arises due to mutation – change in DNA sequence which is heritable if it occurs in the germline. However, variation in mRNA can also be achieved by post-transcriptional modification including mRNA editing, changes in mRNA nucleotide sequence that mimic the effect of mutations. Such modifications are not inherited directly; however, as the processes affecting them are encoded in the genome, they have a heritable component, and therefore can be shaped by selection. In soft-bodied cephalopods, adenine-to-inosine RNA editing is very frequent, and much of it occurs at nonsynonymous sites, affecting the sequence of the encoded protein. We study selection regimes at coleoid A-to-I editing sites, estimate the prevalence of positive selection, and analyze interdependencies between the editing level and contextual characteristics of editing site. We show that mRNA editing of individual nonsynonymous sites in cephalopods originates in evolution through substitutions at regions adjacent to these sites. As such substitutions mimic the effect of the substitution at the edited site itself, we hypothesize that they are favored by selection if the inosine is selectively advantageous to adenine at the edited position. Consistent with this hypothesis, we show that edited adenines are more frequently substituted with guanine, an informational analogue of inosine, in the course of evolution than their unedited counterparts, and for heavily edited adenines, these transitions are favored by positive selection. Our study shows that coleoid editing sites may enhance adaptation, which, together with recent observations on *Drosophila* and human editing sites, points at a general role of RNA editing in the molecular evolution of metazoans.

3.1. Introduction

The process of natural selection requires heritable variation to be present in a population and the absence of genetic variants selection could act upon is generally considered to be a factor hampering adaptation (Lush 1937; Smith 1976; Barton and Partridge 2000; Lanfear, Kokko, and Eyre-Walker 2014; Rousselle et al. 2020). Heritable variation is generated mainly by the mutational process (Lewontin 1964; Avery and Hill 1977; Lynch and Walsh 1998). Hence, the mutation rate may be a factor affecting the evolution rate, which we, following J. Maynard Smith, define here as the rate of accumulation of beneficial mutations (Maynard Smith 1976; Nam et al. 2017; Rousselle et al. 2020). As shown recently, in populations with low genetic variability the mutation rate is indeed correlated with the evolution rate (Rousselle et al. 2020). Thus, in order to adapt, a low-polymorphic population may need additional expressed genetic variability. Here, we test the hypothesis that a potential source of such variability could be introduced by heritable epigenetic modifications, specifically, mRNA editing (Bass and Weintraub 1988; Gommans, Mullen, and Maas 2009; Klironomos, Berg, and Collins 2013; Kronholm and Collins 2016).

We consider the A-to-I mRNA editing, where adenine (A) is modified to inosine (I) that is subsequently read by the translation machinery as guanine (G) (Bass and Weintraub 1988). In most of the studied organisms, the A-to-I editing affecting protein sequences is restricted to only a few thousand adenines, with the vast majority of edited adenines located in non-coding regions, e.g. in Alu-repeats (Kim et al. 2004; Ramaswami et al. 2012; Yablonovitch et al. 2017). Edited sites are poorly conserved between species, suggesting that most editing events are non-functional, with a few possible exceptions (Yang et al. 2008; Pinto, Cohen, and Levanon 2014; Yu et al. 2016). However, in coleoids, soft-bodied cephalopods, about 1% of adenines in the transcriptome are edited, and re-coding (i.e., affecting the amino acid sequence) and conserved sites comprise considerable fractions (Alon et al. 2015; Liscovitch-Brauer et al. 2017). One explanation for this phenomenon comes from the observation that the conserved editing sites tend to be edited in the nervous tissue, and editing may

contribute to the increased plasticity and complexity of the coleoid nervous system and behavior compared to other extant cephalopods (*Nautilus*) (Albertin et al. 2015; Alon et al. 2015; Liscovitch-Brauer et al. 2017; Eisenberg and Levanon 2018). This hypothesis is supported by analogous observations in other organisms (Pinto, Cohen, and Levanon 2014; Yu et al. 2016) and, although indirectly, by the finding that the A-to-I RNA editing has emerged approximately at the same time as the nervous systems of multicellular organisms have become more complex (Jin, Zhang, and Li 2009).

A-to-I editing is not absolutely efficient and, if it occurs at a non-synonymous site, would result in two non-identical proteins with a varying ratio (Gommans, Mullen, and Maas 2009; Liscovitch-Brauer et al. 2017; Yablonovitch et al. 2017). The efficiency of mRNA editing depends on the strength of the site motif and the local mRNA secondary structure (Morse, Aruscavage, and Bass 2002; Reenan 2005; Gommans, Mullen, and Maas 2009; Alon et al. 2015; Savva, Rieder, and Reenan 2012; Klironomos, Berg, and Collins 2013; Rieder et al. 2013; Liscovitch-Brauer et al. 2017). As the sequence and structure requirements seem to be relatively weak, mRNA editing sites have been proposed to constantly emerge at random points of the genome (Gommans, Mullen, and Maas 2009; Xu and Zhang 2014).

To date, four models of A-to-I editing site evolution have been proposed. (1) Most A-to-I editing sites generally are not adaptive and mainly arise at positions with tolerable, i.e. effectively neutral or mildly deleterious, A-to-G substitutions (Xu and Zhang 2014). (2) A-to-I editing is a mechanism of rescuing deleterious G-to-A substitutions (Jiang and Zhang 2019). (3) A-to-I editing, generating multiple protein variants, is important for the advantageous transcriptome diversification, and hence the individual sites should be conserved (Liscovitch-Brauer et al. 2017; Eisenberg and Levanon 2018). (4) The potential of A-to-I editing to mimic A-to-G substitutions is advantageous, and thus A-to-I editing sites function as transitory states when an advantageous mutation has not yet occurred (Popitsch et al. 2020).

Editing site evolution in *Drosophila* and human has been recently shown to adhere to model (4) (Popitsch et al. 2020), while editing sites in coleoids are largely considered as means for proteome diversification as in model (3) (Liscovitch-Brauer et al. 2017; Eisenberg and Levanon 2018) or be selectively neutral (Jiang and Zhang 2019). We attempt to resolve this controversy by detailed analysis of substitution patterns and selection regimes, taking into account the varying strength of A-to-I editing at different sites.

Generally, in a population with low genetic variability, one might expect evolutionary benefits of A-to-I editing consistent with model (4). Indeed, if there is a position in the genome occupied by an adenine, but guanine in this position would yield a fitter genotype, there are two evolutionary pathways for adaptation: through an A-to-G substitution at this site, or through emergence of a local sequence context yielding or reinforcing A-to-I editing of this site. If the selective benefit conferred by both pathways is comparable, which of them will be taken will depend on the probability of the corresponding mutation (Yampolsky and Stoltzfus 2001). A specific mutation is needed in the first scenario; by contrast, many different editing context-improving mutations could yield a fitter genotype, and the waiting time for any such mutation could be shorter (Durrett and Schmidt 2008). As a result, selection would lead to emergence of the adaptive editing phenotype.

We propose that non-conserved coleoid A-to-I mRNA editing sites, comprising the larger percentage relative to the conserved ones, could function as substitutes of beneficial A-to-G substitutions in low-polymorphic coleoid populations. We show that the levels of cephalopod A-to-I editing heavily depend on the sequence of adjacent regions, and hence are influenced by a multitude of possible mutations. Critically, we show that edited adenines are more frequently substituted in related species to guanines and less frequently, to cytosines or thymines, than non-edited ones. At strongly edited sites, the adenine-to-guanine transitions are favored by positive selection. Our results suggest that, while conserved coleoid editing sites could be functionally important *per se*, a large subset of non-conserved editing sites could play a role in the adaptive evolution by introducing, at least in a

fraction of transcripts, guanines that are beneficial at the given positions. When this study had been completed, a similar observation was made for *Drosophila* and human editing sites by analysis of genomic polymorphisms (Popitsch et al. 2020). This indicates that A-to-I editing could have similar, important evolutionary roles in multiple metazoan lineages.

3.2. Methods

3.2.1. Data.

Transcriptomes for all six considered species, *O. vulgaris*, *O. bimaculoides*, *S. esculenta*, *L. pealei*, *N. pompilius*, and *A. californica*, parameters of editing sites, and tables of conserved editing sites were taken from the online supplementary data of Liscovitch-Brauer et al. 2017 (Fig. 1). The sets of editing sites for the four coleoid species are consistent with respect to the distributions of editing levels (Fig. 1B). Genomic read data were downloaded from the SRA database. *S. esculenta* and *O. vulgaris* genomic read data were taken from bioproject PRJNA299756, *L. pealei*, from PRJNA255916, and *O. bimaculoides*, from PRJNA270931.

3.2.2. Annotation of structured and unstructured regions.

To estimate the structural potential of each position we used Z-score values obtained by the RNASurface program (Soldatov, Vinogradova, and Mironov 2014). Here, Z-score of a sequence is defined as $Z = (E - \mu)/\sigma$ where E is the minimal free energy of a biological sequence, μ and σ are the mean and standard deviation of the energy distribution of shuffled sequences with preserved length and average dinucleotide composition. The program was run with parameters maximal sliding window length 350 and minimal sliding window length 20. From the RNASurface output, structural potential of overlapping segments was inferred. Each position of each transcript was assigned the best (minimal) *Z-score* of all structured segments containing it, if it was less than –2, otherwise it was assigned null value. As a result, each transcript was divided into structured and unstructured regions with a *Z*-score

value assigned to all positions in the structured regions. The difference between the structural potential upon the A-to-G change (Fig. 2D) was considered if its absolute value exceeded 2.

3.2.3. Analysis of polymorphisms.

Genomic reads were mapped onto transcriptomes with bowtie2 (Langmead and Salzberg 2012) using the --sensitive-local run mode. After the sorting of the resulting read alignment files with the samtools package (Li et al. 2009), diploid genotypes were called with bcftools (Narasimhan et al. 2016). Next, we discarded all non-SNP variants and variants with the quality score below 20. We computed synonymous nucleotide diversity π_s with the pairwise haplotype comparison implemented in the PAML package (Yang 2007).

3.2.4. Alignments.

To construct multiple transcriptome alignments, we selected a transcriptome of one species and performed BLASTn (Altschul et al. 1990) with the E-value threshold of 10^{-15} against the transcriptomes of the remaining species. Resulting alignment was obtained by merging of the pairwise BLASTn alignments. The results showed only a negligible dependence on the choice of the seed species.

3.2.5. Context analysis.

Site LOGOs were built with the WebLOGO server (Crooks et al. 2004). *R* values for mismatches in contexts of non-conserved editing sites were defined as:

$$R^{\pm 1}_{N_1,N_2} = \frac{p(\mathrm{E}N_1, \mathrm{A}N_2)}{p(\mathrm{A}N_1, \mathrm{A}N_2)}$$

where N_1 and N_2 represent nucleotides in positions +1 and –1 relative to the considered adenine, p(EN_1, AN_2) is the probability of a mismatch at position +1 or –1 relative to the considered adenine that is edited in one of the two considered species and not edited an another, defined as:

$$p(\mathrm{E}N_1, \mathrm{A}N_2) = \frac{\#(\mathrm{E}N_1, \mathrm{A}N_2)}{\#(E, A)}$$

with $\#(E, A)$ and $\#(\mathrm{E}N_1, \mathrm{A}N_2)$ being the number of homologous A-E states and the number of contextual *N1*-*N2* mismatches associated with the A-E pairs, respectively. A and E are the edited and non-edited states of adenines, respectively.

p(A*N1*, A*N2*) is the respective probability when both homologous adenines are non-edited defined as:

$$p(\mathrm{A}N_1, \mathrm{A}N_2) = \frac{\#(\mathrm{A}N_1, \mathrm{A}N_2)}{\#(A, A)}$$

$\#(A, A)$ and $\#(\mathrm{A}N_1, \mathrm{A}N_2)$ being the number of homologous A-A states and the number of *N1*-*N2* mismatches adjacent to the A-A pairs, respectively.

The statistical significance of the *R* values was assessed by the chi-squared contingency test with the Bonferroni correction on the number of *N1*-*N2* mismatch types. A 2×2 Contingency matrix *S* used in the chi-squared test was constructed from the numbers used to define *p*(E*N1*, A*N2*) and *p*(A*N1*, A*N2*):

$$S = \begin{pmatrix} \#(\mathrm{E}N_1, \mathrm{A}N_2) & \#(\mathrm{A}N_1, \mathrm{A}N_2) \\ \#(E, A) & \#(A, A) \end{pmatrix}$$

3.2.6. Substitution matrix.

For a considered species, we considered its closest relative and an outgroup that could be either of the two remaining coleoids (Fig. 1A). Given the low number of available species, we used maximum parsimony (MP) to reconstruct ancestral states. Thus, for a position in the alignment, the ancestral state of nucleotide *N* was inferred if the closest relative and an outgroup had the same nucleotide N^{anc}; an ancestral adenine was considered to be edited if the homologous adenines in the closest relative and an outgroup were edited. The substitution matrix was thus comprised of counts inferred by MP, $\#(N^{\mathrm{anc}} \rightarrow N)$.

3.2.7. *R* and *Q* calculation for non-synonymous (NES) and synonymous (SES) editing sites.

When *R* measures were computed separately for SES and NES, we applied a modification of the expression for the *R* value. For substitutions at SES:

$$R_{\to N}^{\mathrm{syn}} = \frac{p(\mathrm{E}^{\mathrm{syn}}, N)}{p(\mathrm{A}^{\mathrm{syn}}, N)}$$

where $\mathrm{E}^{\mathrm{syn}}$ are synonymous editing sites, i.e. edited adenines that, when substituted to guanine, do not change the amino acid, and, similarly, $\mathrm{A}^{\mathrm{syn}}$ are synonymous unedited adenines. An analogous formula was applied for non-synonymous editing sites. The definitions of probabilities $p(\mathrm{E}^{\mathrm{syn}}, N)$ and $p(\mathrm{A}^{\mathrm{syn}}, N)$ are in this case analogous to those used in the context analysis, see above.

When we calculated $R_{N\to}$ separately for NES and SES, we applied another modification of the expression for the *R* value. For mutations to SES we have:

$$R_{N\to}^{\mathrm{syn}} = \frac{p^*(N^{\mathrm{syn}}, \mathrm{E}^{\mathrm{syn}})}{p^*(N^{\mathrm{syn}}, \mathrm{A}^{\mathrm{syn}})}$$

where $\mathrm{E}^{\mathrm{syn}}$ and $\mathrm{A}^{\mathrm{syn}}$ are defined just as above, and N^{syn} represents nucleotides, that, when substituted with adenine and with guanine would yield the same amino acid. p^* are conditional probabilities:

$$p^*(N, \mathrm{E}) = p(N, \mathrm{E}) \Big/ \frac{\#\mathrm{E}}{\#\mathrm{E} + \#\mathrm{A}}$$

$$p^*(N, \mathrm{A}) = p(N, \mathrm{A}) \Big/ \frac{\#\mathrm{A}}{\#\mathrm{E} + \#\mathrm{A}}$$

An analogous formula is applied to NES, with N^{non} representing nucleotides, that, when substituted with adenine and with guanine would yield different amino acids.

3.2.8. Calculation of *dN/dS*.

We estimated the strength of positive selection acting on substitutions to guanines and to pyrimidines separately by applying the *dN/dS* measure to edited adenines with a subsequent normalization by *dN/dS* of unedited adenines. Thus, for substitutions to G we applied the formula (Suppl. Fig. S9):

$$\frac{dN(\mathrm{E} \rightarrow \mathrm{G})}{dS(\mathrm{E} \rightarrow \mathrm{G})} \Big/ \frac{dN(\mathrm{A} \rightarrow \mathrm{G})}{dS(\mathrm{A} \rightarrow \mathrm{G})}$$

where *dN* were calculated for all codons and *dS*, for four- and six-fold degenerate codons. An analogous formula was used to estimate selection acting on E-to-Y substitutions. Next, we applied this measure separately for 10% editing level (EL) bins, counted Pearson's correlation coefficient and applied the *F* statistic to estimate the significance of the obtained correlation.

Positive selection on editing sites was estimated with the *dN/dS* ratio where non-synonymous substitutions were considered for edited adenines, and synonymous, for unedited adenines:

$$\frac{dN(\mathrm{E} \rightarrow \mathrm{G})}{dS(\mathrm{A} \rightarrow \mathrm{G})} = \frac{p(\mathrm{E}^{\mathrm{non}} \rightarrow \mathrm{G})}{p(\mathrm{A}^{\mathrm{syn}} \rightarrow \mathrm{G})} * \frac{\xi^{\mathrm{non}}}{\xi^{\mathrm{syn}}}$$

where ξ^{non} and ξ^{syn} are normalizing coefficients accounting for differences in codon probabilities and different probabilities of, respectively, synonymous and non-synonymous substitutions under the neutral evolution assumption. These coefficients are defined as:

$$\xi^{\mathrm{non}} = 1 \Big/ \left(\sum_{N_1N_2N_3 \in \{\mathrm{A,T,G,C}\}^3} f(N_1N_2N_3) \times K^{\mathrm{non}}(N_1N_2N_3, \mathrm{A} \rightarrow \mathrm{G}) \right)$$

$$\xi^{\mathrm{syn}} = 1 \Big/ \left(\sum_{N_1N_2N_3 \in \{\mathrm{A,T,G,C}\}^3} f(N_1N_2N_3) \times K^{\mathrm{syn}}(N_1N_2N_3, \mathrm{A} \rightarrow \mathrm{G}) \right)$$

where $f(N_1N_2N_3)$ is the codon frequency while K^{non} and K^{syn} are, respectively, the numbers of possible non-synonymous and synonymous A-to-G substitutions in a given codon.

3.2.9. Statistics.

For mutation frequency and *dN/dS* analysis, statistics were obtained from 10^5 random sets of mutation numbers sampled from the binomial distributions with the parameters equal to the observed substitution frequencies. For the analysis of parallel evolution, two-tailed confidence intervals were inferred from the binomial distribution. The binomial test was applied to compare fractions of conserved and not conserved editing sites in structured segments. For the analysis of changes in the secondary structure stability following A-to-G *in silico* substitutions, sizes of tails in the distribution of Z-score differences were compared using the binomial test, and to compare the results for different types of sites, random 100-sequence samples of each type were compared with the Wilcoxon signed-rank test.

3.2.10. Data availability.

Ad hoc scripts were written in Python. Graphs were built using R. All scripts and data analysis protocols are available online at https://github.com/mikemoldovan/coleoidRNAediting.

3.3. Results

3.3.1. Editing level is associated with the local and global sequence context.

We studied the A-to-I editing using available genomic read libraries, transcriptomes, and editing sites data for four coleoids, closely related octopuses *Octopus vulgaris* and *O. bimaculoides*, squid *Loligo pealei*, and cuttlefish *Sepia esculenta* (Liscovitch-Brauer et al. 2017). As outgroups, we considered nautiloid *Nautilus pompilius* and gastropod mollusk *Aplysia californica* (Liscovitch-Brauer et al. 2017) (Fig. 1A).

The action of editing sites as surrogates of beneficial A-to-G substitutions presumes advantageous enhancement of editing probabilities at individual sites. As A-to-I editing is to be affected by the local sequence context (Alon et al. 2012; Liscovitch-Brauer et al. 2017) and the RNA secondary RNA structure (Morse, Aruscavage, and Bass 2002; Reenan 2005; Gommans, Mullen, and

Maas 2009; Savva, Rieder, and Reenan 2012; Klironomos, Berg, and Collins 2013; Rieder et al. 2013), one would expect, firstly, contextual differences around weakly *vs*. heavily edited sites and, secondly, contextual mutations associated with changes in editing status. Indeed, we have observed a previously unnoted dependence of the editing level (EL) (Fig. 1B), defined as the percent of transcripts containing I at the considered site at the moment of sequencing (Fig. 2A, Suppl. Fig. S1), on the site context (±1 motif). Certain changes in the ±1 motif, specifically, an increase in the preference for G or T at the +1 position, are associated with the increase of EL, although its information content of the motif remains approximately the same. Although the ±1 motif of both weakly and strongly edited sites is consistent with the ADAR (adenosine deaminases acting on RNA) profile (Alon et al. 2012; Liscovitch-Brauer et al. 2017), this observation could point to the action of different ADAR enzymes or to different modes of action of the same enzyme on strongly and weakly edited sites. There also seem to be some differences between the motifs of conserved and non-conserved sites (Suppl. Fig. S2).

The analysis of non-conserved editing sites (NCES) in the octopus pair demonstrates overrepresentation of mismatches in the ±1 motif of the edited adenines reinforcing the local editing context compared to the homologous unedited adenines, for which the editing context is not observed (See Materials and Methods, Suppl. Fig. S3). Thus, both the editing status and the EL of a site are associated with substitutions in the ±1 motif. In the squid-cuttlefish pair, the higher number of mutations obscures this analysis.

To estimate the size of the region that affects editing, we have measured the correlation between the editing level difference in conserved editing sites (CES) in closely related species and the number of mismatches in variable-sized windows centered at edited adenines. The window size yielding the largest correlation coefficient shows the average span of the context affecting the ADAR performance. For the *Octopus* pair, the highest correlation has been obtained at the window size of ~100 nucleotides (Suppl. Fig. S4), consistent with previous estimates for the length of the region affecting editing (Liscovitch-Brauer et al. 2017).

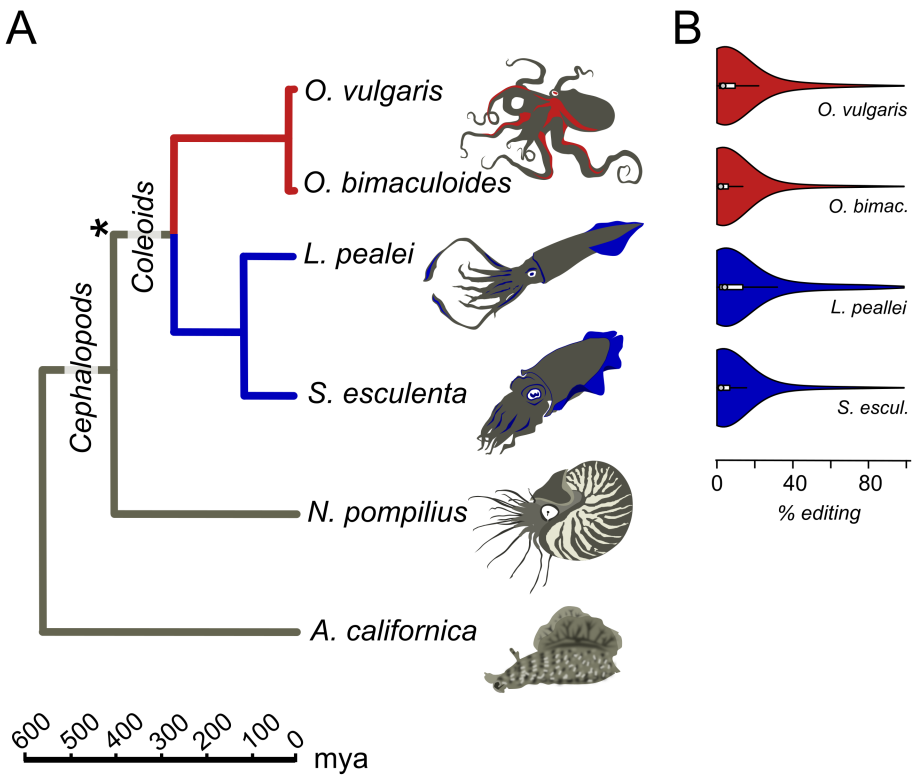


Figure 1 | Prevalent mRNA editing in coleoid molluscs. (**a**) Phylogenetic tree of the species taken from TimeTree (Hedges, Dudley, and Kumar 2006). The asterisk marks the putative beginning of editing site expansion (Liscovitch-Brauer et al. 2017). (**b**) Distributions of per-nucleotide editing levels of the predicted editing sites in the studied coleoids. Note that editing sites have quite low (<10%) editing levels on average, however the distributions of editing levels across sites have heavy right tails, which correspond to large sets of heavily edited sites.

3.3.2. Editing level is affected by secondary structure in adjacent RNA.

The A-to-I editing in model species depends on large RNA structures spanning hundreds of nucleotides in addition to the local sequence context (Morse, Aruscavage, and Bass 2002; Reenan 2005; Ensterö et al. 2009; Rieder et al. 2013; Kurmangaliyev, Ali, and Nuzhdin 2016) as the ADAR-mediated mRNA editing generally requires secondary RNA structures (Gommans, Mullen, and Maas 2009; Farajollahi and Maas 2010; G. Xu and Zhang 2014). Thus, we have assessed the link between RNA secondary structure and ELs of focal sites.

We have predicted structured segments in the transcripts of all six considered species. As the fraction of adenines located within structured segments is the same for all cephalopod species, including *Nautilus* (Suppl. Fig. S5), our secondary structure analyses are not systematically influenced by the GC-content of the studied genomes (Wang et al. 1984). Then we have assessed the contribution of mRNA secondary structure to the editing process by comparing structural contexts of edited and unedited adenines (Materials and Methods). The fraction of edited adenines located in putative structured regions is higher than the respective fraction for non-edited sites. Moreover, sites that are more highly edited (Fig. 2B) as well as sites conserved between more distant species (Fig. 2D) tend to be more structured.

To uncover the connection between the strength of a local secondary RNA structure and the editing status at individual sites, we have compared the fractions of non-conserved editing sites (NCES) located within structured segments in edited *vs*. non-edited states. We considered the *Octopus* pair and the squid–cuttlefish pair. For both pairs, we have compared CES and NCES. For NCES in both species pairs we have observed significantly more cases when the edited site in a pair is more structured than the unedited one while the control CES set shows no bias (binomial test $p<1.1\times10^{-3}$ for two species pairs: *O. vulgaris/O. bimaculoides* and squid/cuttlefish; Fig. 2C, Suppl. Fig. S6).

Not only the fact of editing, but the difference in editing levels is linked to local secondary structures. For the closely related *Octopus* pair, we have calculated correlations between differences in ELs of homologous edited adenines and differences in their structure Z-scores (Suppl. Fig. S7). Almost no correlation (r=0.1, t-test $p < 0.05$) is seen when the EL difference is small (>5%), whereas for large differences in ELs (>50%) the correlation is substantial (r=0.7, t-test $p < 0.05$). A likely explanation is that larger differences are indeed due to the strength of the local secondary structure, whereas small differences in ELs arise as consequences of random noise. Consistent with the observations above, if we consider structures around edited adenines and their unedited homologs,

setting the ELs of unedited adenines to 0, we observe a similar, although a weaker trend (Suppl. Fig. S7), with correlations reaching 0.4 (t-test $p < 0.05$) when the ELs of NCES are high.

The observations about local contexts, both the ±1 motif and RNA structures, imply that mutations near editing sites influence the editing status as well as the editing level.

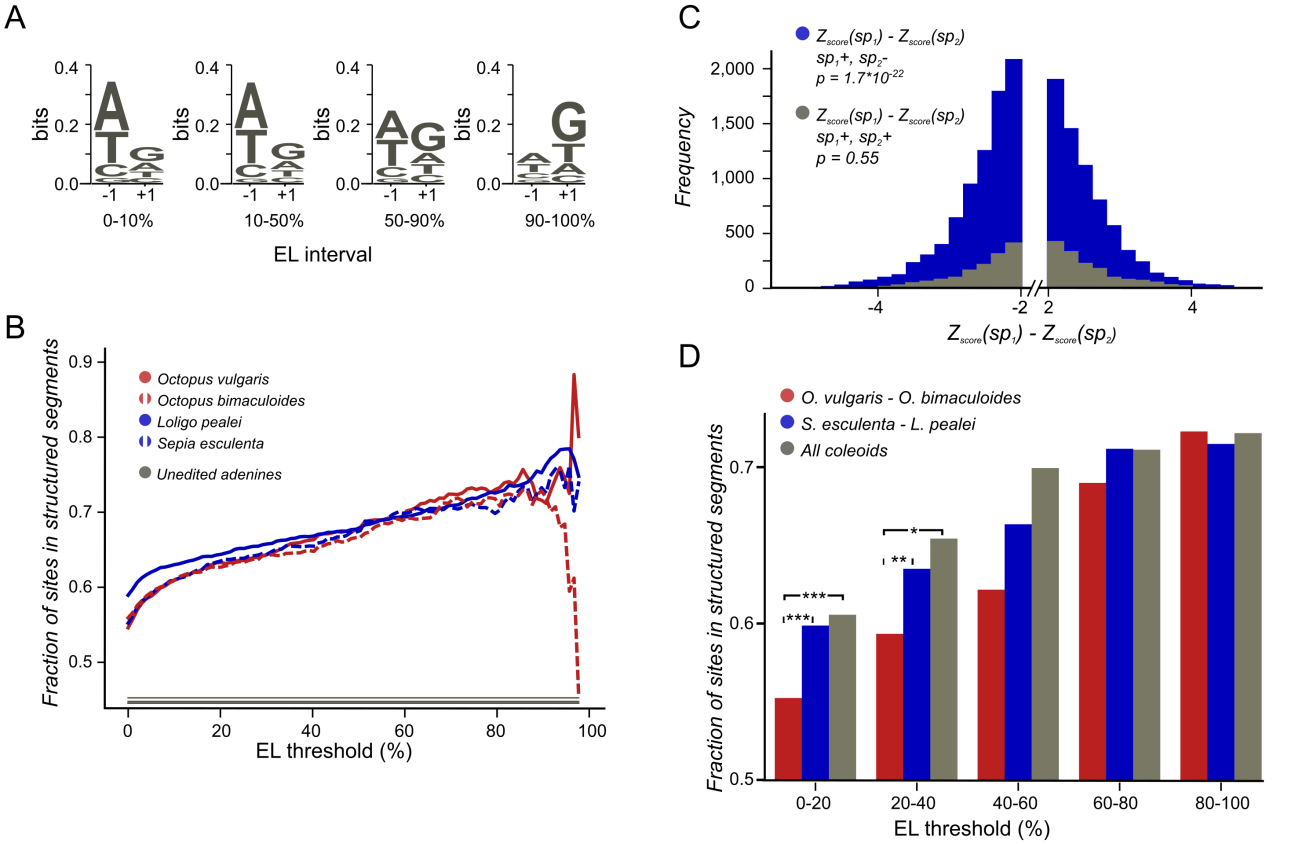


Figure 2 | (**a**) *O. bimaculoides* editing site context changes with the increase of editing level. The height of the letters represents the LOGO bit score of each nucleotide. (**b**) Highly conserved editing sites tend to be relatively more structured. The fraction of editing sites that are in structured segments is shown for different editing levels: red — *O. vulgaris*, red dashed — *O. bimaculoides*, blue — *L. pealei*, blue dashed — *S. esculenta*, grey — the constant showing the fraction of unedited adenines located in structured segments. The noisy pattern at the right is due to a low number of very highly edited sites. (**c**) The stability of the local secondary structure is higher at edited adenines than at homologous, non-edited adenines for the squid/cuttlefish pair. The distribution of the difference of the minimal free energy Z-score between homologous sites in squid and cuttlefish is shown in blue when

two homologous sites have different editing status (edited minus unedited) and in grey when both sites in a pair are edited. The left tail of the blue histogram is heavier than the right one ($p = 1.7\times10^{-22}$ versus 0.32 for the grey histogram), showing that the editing sites tend to regions with higher secondary structure stability. (**d**). Conserved editing sites tend to be more structured. The three groups of sites are those present in two of the four species *(O. vulgaris* and *O. bimaculoides*, red, or *L. pealei* and *S. esculenta,* blue), or in all four species (grey). Statistically significant differences are shown with brackets (*** $p<0.001$, * $p<0.05$).

3.3.3. Edited adenines are often substituted by guanines.

If edited adenines indeed frequently mimic the beneficial guanine state, the substitution patterns of edited and unedited adenines should differ, with edited adenines being more prone to substitutions to guanine and less prone to substitutions to cytosine or thymine (Popitsch et al. 2020). Firstly, we performed the analysis of the species pairs to infer the properties of A-G mismatches at editing sites. For a pair of considered species, we define *R* as the mismatch probability for an edited adenine divided by the probability of the same mismatch for an unedited adenine: $R_N=p(\mathrm{E},N)/p(\mathrm{A},N)$, where E and A are, respectively, edited and not edited adenines in one species, and *N* is the non-E, non-A nucleotide at the homologous site in the other organism. Similar formulas are applied when we consider directed substitutions instead of mismatches. If a pair of the ancestral and the descendant species is considered, we use notation $R_{\to N}$ to identify the directionality. $R_{\to N} = p(\mathrm{E}\to N)/p(\mathrm{A}\to N)$, where $p(\mathrm{E}\to N)$ and $p(\mathrm{A}\to N)$ are, respectively, the probabilities of the substitution of the edited and non-edited adenine to *N*. Similarly, notation $R_{N\to}$ is used when substitutions from ancestral *N* to E and A are considered: $R_{N\to}=p(N\to\mathrm{E})/p(N\to\mathrm{A})$. Higher values of $R_{N\to}$ imply that the ancestral nucleotide *N* is more likely to be substituted by an edited adenine, compared to an unedited one.

We have observed a striking dependence of the calculated mismatch probabilities on the editing status of the adenines and their ELs. In the *Octopus* pair, R_G and R_Y (Y denotes pyrimidine, C or T)

differ both in value and in the dependence on the EL (Fig. 3AB). Indeed, R_G is always higher than R_Y with R_G further increasing and R_Y decreasing as the EL increases. The probability for an adenine to be substituted by a guanine in the *O. vulgaris* lineage is ~8 times higher when the homologous adenine is strongly edited in *O. bimaculoides* than when it is not (Fig. 3A). For the more distantly related squid–cuttlefish pair, we observe a similar although less pronounced effect. For all distant pairs, that is, *Octopus*–squid/cuttlefish, R_G shows no or only a weak dependence on the EL.

We have calculated *R* values separately for non-synonymous editing sites (NES), which comprise between 64.6% and 65.7% of all detected coleoid editing sites, and for synonymous editing sites (SESs) which comprise the remaining 34.3% to 35.4%. NES (Suppl. Fig. S8AB) demonstrate the same pattern as described above for all sites, whereas for SES, we see no dependence of R_Y on the EL (Suppl. Fig. S8CD). NES demonstrate very low R_Y at high ELs. These patterns suggest that at highly edited nonsynonymous adenine sites, any nucleotide other than guanines are impeded by strong negative selection; whereas the guanine states at such sites are frequent. Thus, at non-synonymous NCES, the selection patterns differ from those at non-conserved adenines: Y mismatches with NCES experience stronger negative selection than Y mismatches with non-edited adenines, and stronger positive and/or weaker negative selection acting on E-to-G or G-to-E substitutions compared to A-to-G or G-to-A ones, respectively.

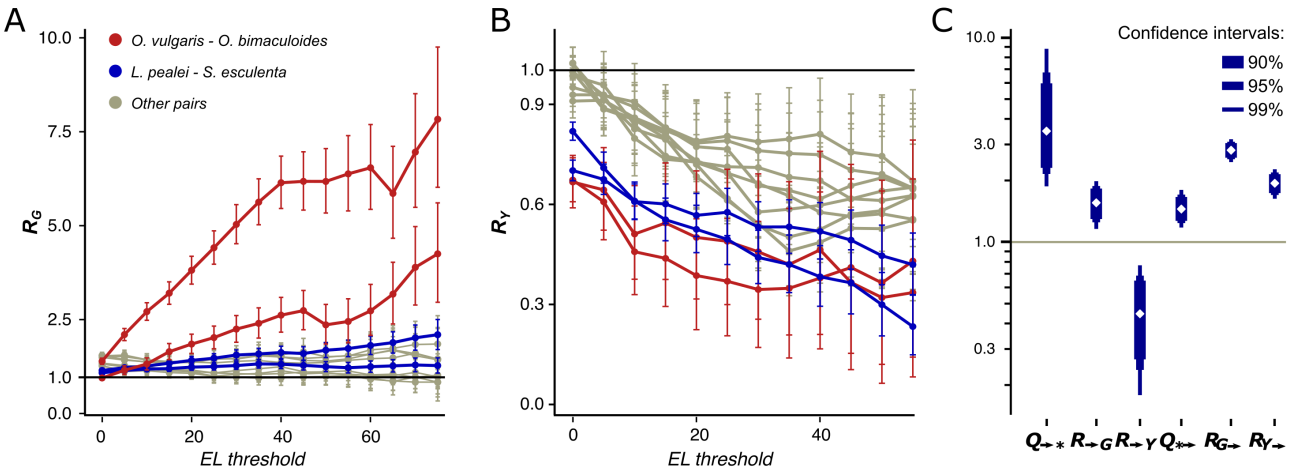


Figure 3 | *R* and *Q* values. (a.b) Dependence of R_G (a) and R_Y (b) on the editing level. Two curves for each pair are given, since R_N is calculated two times for each pair of species using one of them a a reference each time. The red curves correspond to the pair *O. vulgaris – O. bimaculoides*; the blue curves, to the pair cuttlefish–squid, the grey curves, to distant pairs. **(c)**. Mutational characteristics of editing sites for the squid–cuttlefish summary substitution matrix. Left to right: $Q_{\rightarrow *} >> 1$, $R_{\rightarrow G} > 1$, $R_{\rightarrow Y} < 1$, $Q_{*\rightarrow} > 1$, $R_{G\rightarrow} >> 1$, $R_{Y\rightarrow} > 1$.

3.3.4. Editing recapitulates substitutions that are positively selected.

To reveal the mode of selection at edited sites, we have calculated the *dN/dS* ratios separately for mismatches of edited and unedited adenines with guanines and with pyrimidines (Suppl. Fig. S9). For weakly edited adenines, the *dN/dS* values of mismatches with guanines and with pyrimidines are approximately the same as those for unedited adenines. However, at highly edited sites, the *dN/dS* ratio for substitutions to guanine is two- to threefold higher, compared to unedited adenines, while the respective ratio for pyrimidines is twofold lower. Thus, strongly edited sites evolve under weaker purifying selection against E-to-G and/or G-to-E transitions and stronger purifying selection against E-to-Y and/or Y-to-E substitutions.

To distinguish between positive selection and relaxation of negative selection at these sites, we have calculated *dN/dS* for A-G mismatches where *dN* and *dS* are calculated for edited and unedited adenines, respectively. It is larger than 1 at high ELs for the closely related octopus species pair (Fig. 4), indicating positive selection acting on the E-to-G transition: heavily edited adenines are positively selected for substitutions to guanine.

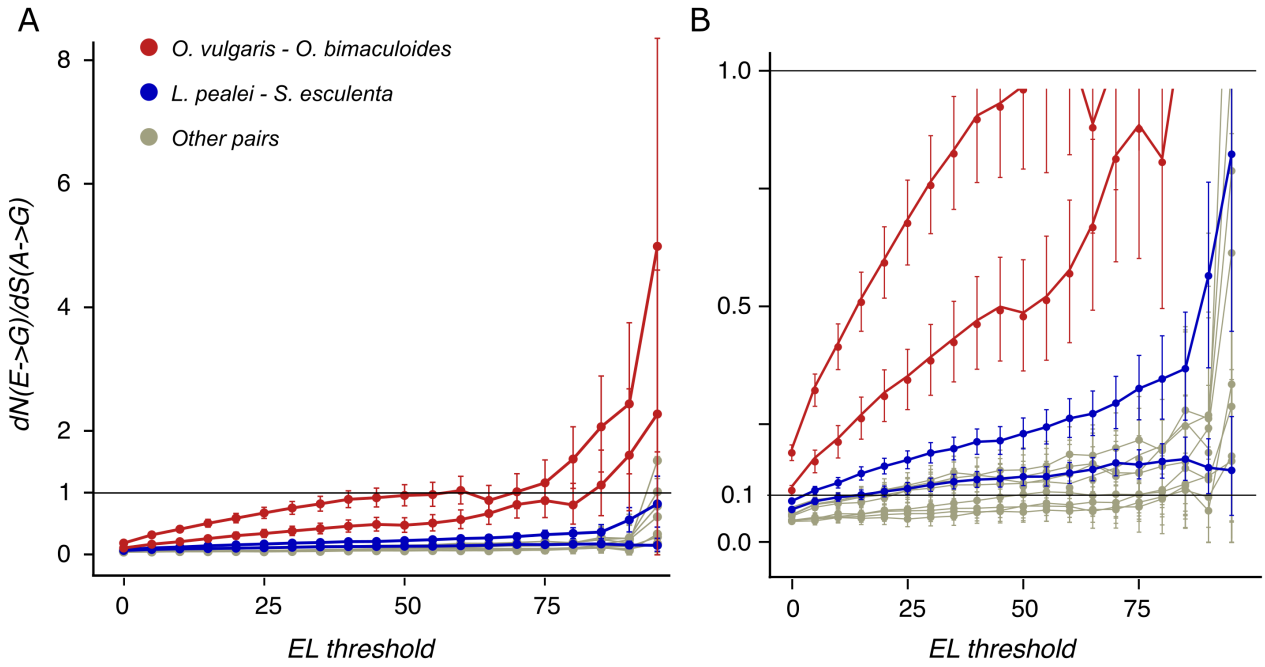


Figure 4 | *dN/dS* values of adenine substitutions to guanines for various EL thresholds. Non-synonymous substitutions are calculated for edited adenines, and synonymous, for unedited adenines. Error bars indicate the 95% probability value range. **a**. Plot for the whole range of *dN/dS* values. **b**. Truncated value range. Note the increase of *dN/dS* values at high EL values for all species pairs.

3.4. Discussion

3.4.1. The hypothesis about the adaptivity of non-conserved editing sites is supported by our observations.

Editing in coleoids is essential for transcriptome diversification, and results in a more complex phenotype (Liscovitch-Brauer et al. 2017; Eisenberg and Levanon 2018). Indeed, a considerable fraction of coleoid editing sites are conserved between even distantly related species, and a majority of heavily edited sites affect protein sequence (Liscovitch-Brauer et al. 2017). We propose that non-conserved coleoid editing sites could facilitate adaptation by extending selection to regions affecting editing if guanine is the beneficial variant at the editing site. This hypothesis is directly supported by our observations. Indeed, strong dependence of editing on the local context allows for selection of mutations in the vicinity of the editing site, hence extending the variety of beneficial mutations. On the other hand, edited adenines indeed tend to be substituted by guanines, and guanines are selected for if the editing levels of homologous adenines is high. This positive selection pattern is specific to guanine variants, as substitutions of edited adenines to cytosine or thymine are avoided.

An indirect observation also supports our hypothesis. Sizes of the effects such as the E-to-G substitution rate or the rate of positive selection on the guanine variant at editing sites are larger for heavily edited adenines compared to medium and weakly edited ones. This effect could be explained by beneficial A-to-G substitutions provoking selection on adjacent regions, which leads to the increased ELs and hence to the enhanced presence of the guanine-like variant. Indeed, if G is beneficial at a given site, it would manifest as both positive selection towards G at this site, and by mutations at adjacent sites yielding higher A-to-I editing level, and hence these two types of effects would be correlated.

3.4.2. E-to-G substitutions versus G-to-E substitutions.

In theory, two processes could lead to the increase in the observed *R* and *dN/dS* values of edited sites — the increased frequency of either E-to-G or G-to-E substitutions. To distinguish between these

possibilities, we use the procedure described in Materials and Methods to calculate the frequencies of all types of substitutions for each species since its closest ancestor. We also consider the more robust, averaged substitution frequencies for the *Octopus* pair and for the squid–cuttlefish pair. As the frequencies of substitutions to edited and non-edited adenines are calculated separately, we introduce the normalized, directional measure $Q_{\to *}$ reflecting the preference of edited adenines to substitute to guanine:

$$Q_{\to *} = \frac{R_{\to \mathrm{G}}}{R_{\to \mathrm{Y}}} = \frac{p(\mathrm{E} \to \mathrm{G})}{p(\mathrm{A} \to \mathrm{G})} \Big/ \frac{p(\mathrm{E} \to \mathrm{Y})}{p(\mathrm{A} \to \mathrm{Y})} = \frac{p(\mathrm{E} \to \mathrm{G})}{p(\mathrm{E} \to \mathrm{Y})} \Big/ \frac{p(\mathrm{A} \to \mathrm{G})}{p(\mathrm{A} \to \mathrm{Y})}$$

By this definition, the $Q_{\to *}$ measure is an indicator of the joint effect of the prevalence of E-to-G over A-to-G substitutions and of the underrepresentation of E-to-Y relative to A-to-Y substitutions. For the squid–cuttlefish clade, and for SESs and NESs considered separately, $Q_{\to *}$ ranges from 3.49 to 6.4 (Fig. 3C, Suppl. Fig. S10AB), in all cases being significantly higher than 1 expected under a neutral model (p<0.005). Hence, as in the case of pairwise comparison of extant species (Fig. 3AB), edited adenines have a substitution pattern strikingly different from that of unedited adenines, and are likely to mutate into guanines.

However, large values of $Q_{\to *}$ may be explained by two effects, high $R_{\to \mathrm{G}}$ of E-to-G substitutions or low $R_{\to \mathrm{Y}}$ of E-to-Y substitutions (Fig. 3C) both yielding $R_{\to \mathrm{G}}$ higher than $R_{\to \mathrm{Y}}$. $R_{\to \mathrm{G}}$ is higher than 1 (p<0.005), thus indicating that an edited adenine is more likely to be substituted by guanine than an unedited adenine. Combined with $R_{\to \mathrm{Y}}$ being smaller than 1 (p<0.005), this indicates that in fact both effects contribute to the observed $Q_{\to *}$ values. A similar pattern holds if we consider NES and SES separately: $R_{\to \mathrm{G}}$ is higher than $R_{\to \mathrm{Y}}$, although for NES high $Q_{\to *}$ can be almost entirely attributed to $R_{\to \mathrm{G}}$, and for SES, to $R_{\to \mathrm{Y}}$ (p<0.005) (Suppl. Fig. S10AB).

To analyze the directionality of the mutation process that affects editing states, we consider a similar function measuring the degree of prevalence of G-to-E substitutions:

$$Q_{*\rightarrow} = \frac{R_{\mathrm{G}\rightarrow}}{R_{\mathrm{Y}\rightarrow}} = \frac{p^*(\mathrm{G}\rightarrow\mathrm{E})}{p^*(\mathrm{G}\rightarrow\mathrm{A})} \Big/ \frac{p^*(\mathrm{Y}\rightarrow\mathrm{E})}{p^*(\mathrm{Y}\rightarrow\mathrm{A})} = \frac{p^*(\mathrm{G}\rightarrow\mathrm{E})}{p^*(\mathrm{Y}\rightarrow\mathrm{E})} \Big/ \frac{p^*(\mathrm{G}\rightarrow\mathrm{A})}{p^*(\mathrm{Y}\rightarrow\mathrm{A})}$$

where probabilities p^* are conditional probabilities of a nucleotide mutating to either edited or unedited adenine after taking into account differences in the E and A densities in the transcriptomes:

$$p^*(N\rightarrow\mathrm{E}) = p(N\rightarrow\mathrm{E}) \Big/ \frac{\#\mathrm{E}}{\#\mathrm{E}+\#\mathrm{A}}$$

$$p^*(N\rightarrow\mathrm{A}) = p(N\rightarrow\mathrm{A}) \Big/ \frac{\#\mathrm{A}}{\#\mathrm{E}+\#\mathrm{A}}$$

For both the *Octopus* pair and the squid–cuttlefish pair, $Q_{*\rightarrow}$ is larger than 1 (p<0.005) (Fig. 3c, Suppl. Fig. S10), thus suggesting that guanines tend to be substituted by edited rather than unedited adenines. However, this effect is on average twofold smaller than that for substitutions of edited adenines to guanines, suggesting that the process of G-to-E transitions is less directional than that for E-to-G transitions. If $R_{\mathrm{G}\rightarrow}$ and $R_{\mathrm{Y}\rightarrow}$ are considered separately, they both are larger than the expected value 1 (Fig. 3C) (p<0.005), which points to a generally faster generation of E sites from both G and Y nucleotides. As the observed effect is small, it could be attributed to weaker negative selection acting upon the G-to-E transition relative to G-to-A, as the edited adenine is a state closer to the guanine-only variant (Jiang and Zhang 2019).

The Q value defined as the ratio of undirected R values increases with the EL (as follows from Fig. 3AB). On the other hand, formally it is monotonic with respect to the directed $Q_{\rightarrow *}$ and $Q_{*\rightarrow}$ values (see Suppl. Mat. 1). Hence, even though we could not detect a significant dependence of $Q_{\rightarrow *}$ and $Q_{*\rightarrow}$ on EL due to insufficient data, at least one of them should increase with the EL. However, the effects observed for the E-to-G substitution are more pronounced compared with those for the G-to-E substitutions, hinting at A-to-I editing sites mimicking beneficial A-to-G substitutions rather than rescuing deleterious G-to-A substitutions.

3.4.3. Positive selection in favor of E-to-G substitutions.

Why would substitutions that recapitulate editing be adaptive? Conceivably, it could be that variability at the transcriptome level is advantageous by itself, contributing to the proteome diversity, similar to alternative splicing, alternative transcription and translation starts, *etc* (Raj and van Oudenaarden 2008; Gommans, Mullen, and Maas 2009; Pickrell et al. 2010). However, this scenario does not explain positive selection of substitutions to G.

Alternatively, editing might create an unconditionally beneficial variant, so that at an edited site, G is always better than A. Under this scenario, editing could recreate the G allele previously lost due to a deleterious G-to-A mutation, or produce a novel G variant which is favored by selection but has not been present at this site previously (Jiang and Zhang 2019). This scenario is supported by the observed selection favoring guanines at edited sites. In particular, deleterious G states may arise due to dC-to-dU deamination, which results in complementary dG-to-dA mutations and is known as one of the major mechanisms of mutagenesis in the human population (Seplyarskiy et al. 2021).

But why would selection in favor of G result in an increased A-to-I editing of a fraction of the transcripts, when a "direct" A-to-G genomic mutation at this site would lead to the same result in 100% of transcripts? One reason could be that mutations creating editing sites and/or increasing editing level are more numerous, and therefore more readily available. For a strongly advantageous mutation (with $4N_e s >> 1$) that does not preexist in the population, the time till its fixation equals $1/(4N_e s\mu)$, where N_e is the effective population size, s is selection in favor of the new mutation, and μ is the mutation rate, see eq. 3.22 in Kimura 1983. If two types of mutations can yield the desired phenotype, which of them would be the first to fix in an evolving lineage is determined by the product of the corresponding selection and mutation rates.

Let μ_1 be the rate of the direct mutation, and s_1, selection in its favor. Assume that an increase in the number of favored transcripts can also be achieved by M other mutations, each characterized by

rate μ_2 and selection s_2. The probability that the editing-enhancing mutation will be the first to occur then equals $M\mu_2 s_2/(\mu_1 s_1 + M\mu_2 s_2)$ (Yampolsky and Stoltzfus 2001). If $M\mu_2 s_2 > \mu_1 s_1$, the editing-increasing mutation will typically fix earlier than the direct mutation. As we show, many tens of sites may affect editing, making M large, and this scenario likely. For example, if the direct A-to-G substitution confers a 10% increase in fitness, but a 1% increase can be achieved by changes in editing by mutations at each of 20 other sites, then the editing-increasing change will be the first to occur with probability 2/3 if the mutation rates are uniform.

This reasoning only applies if the within-species variability level $N_e\mu$ is low ($<<1$); otherwise each site will carry a preexisting mutation, and the mutation rate will be less relevant (McCandlish and Stoltzfus 2014). Low variability is indeed a characteristic trait of the considered coleoid species, with synonymous-site pairwise divergence of 2.5×10^{-3} for *O. bimaculoides*, 2.2×10^{-3} for *O. vulgaris*, 1.8×10^{-3} for *S. eculenta*, and 4.5×10^{-3} for *L. pealei* (see Materials and Methods). These values imply $N_e\mu<<1$, suggesting that evolution can be indeed mutation-limited in this group of species. Low values of $N_e\mu$ characteristic of higher animals have been proposed to underlie many aspects of genomic complexity (Lynch 2007); they may also cause the high prevalence of RNA editing in coleoids.

When this study had been completed, Popitsch et al. published a population-genetic study of *Drosophila* and human A-to-I RNA editing sites, in which they showed a similar pattern of selection at editing sites, with the derived G state selected upon, whereas C and T variants being suppressed, indicating enhanced negative selection (Popitsch et al. 2020). That study indirectly supports our claim about coleoid A-to-I editing sites mimicking beneficial A-to-G substitutions. Furthermore, as coleoids possess many more conserved re-coding A-to-I editing sites than any other studied metazoan group, one might expect the bulk of coleoid editing, especially at heavily edited sites, to be important *per se*, *e.g.* for transcriptome diversification, which would result in suppression of any non-adenine variants in editing sites. On the contrary, we have observed positive selection in A-G mismatches, when adenines are heavily edited, with selection acting specifically on A-to-G transitions. Also, like Popitsch

et al., we have observed enhanced negative selection against A-to-C and A-to-T substitutions and mismatches at coleoid editing sites. The consistency of results obtained for coleoids, *Drosophila*, and human points towards a general role of A-to-I editing sites as imitations and precursors of A-to-G transitions in the evolution of metazoans with low-polymorphic populations.

3.4.4. Conservation and function of editing.

Earlier, it has been proposed that most editing sites result from tolerable promiscuous ADAR action (Xu and Zhang 2014). However, the A-to-I editing sites in coleoids are under positive selection if ELs are high (Fig. 4). Hence large ELs cannot result simply from the tolerance towards substitutions to guanines at these sites.

Coleoid editing sites are often considered to be important for complex regulation (Albertin et al. 2015; Alon et al. 2015; Liscovitch-Brauer et al. 2017; Eisenberg and Levanon 2018; Jiang and Zhang 2019). However, this applies only to conserved, and hence functional, editing sites. We propose that coleoid editing sites form two populations with different properties. Firstly, there are functional editing sites, which are important *per se* due to their ability to diversify protein products in various tissues and environmental conditions (Savva, Rieder, and Reenan 2012; Alon et al. 2015; Harjanto et al. 2016; Buchumenski et al. 2017; Duan et al. 2017; Liscovitch-Brauer et al. 2017; GTEx Consortium et al. 2017). As such sites should be retained over long periods of time, we may consider conservation as a good proxy for functionality. Conserved sites are surrounded by conserved regions (Liscovitch-Brauer et al. 2017), their ELs show dependence on the number of substitutions in adjacent regions (Suppl. Fig. S4), and they comprise up to about a half of A-to-I editing sites in a coleoid transcriptome (Liscovitch-Brauer et al. 2017b).

Secondly, there are non-functional sites; the proxy here are non-conserved sites, with a caveat that some recently emerged sites could be functional. Nonetheless, as the proportion of young functional sites should be minimal (Gommans, Mullen, and Maas 2009), the general properties of non-

conserved sites should reasonably well represent those of non-functional ones. Non-conserved sites are not flanked by conserved regions, their ELs show no correlation with the number of substitutions in adjacent regions, and their sequence contexts differ from those of the conserved ones (Suppl. Fig. S2). Our hypothesis that (non-conserved) editing sites have an intrinsic evolutionary value does not contradict the fact that some (possibly large) subset of editing sites are functional as editing sites *per se* from the physiological point of view.

Theoretically, our results could have been influenced by underprediction of editing sites. As the mean EL is about 5%, a site might be easily missed especially in transcripts with low expression levels (Bahn et al. 2012; Alon et al. 2012; Liscovitch-Brauer et al. 2017). However, the majority of our observations are obtained for highly edited adenines, which are predicted with greater accuracy (Bahn et al. 2012), and hence should not be influenced by missing weakly edited sites.

3.4.5. Theoretical frameworks and alternative explanations

Our results could be interpreted within several paradigms. Firstly, as noted above, the observations could mean that editing rescues deleterious G-to-A substitutions (Jiang and Zhang 2019). However, as also mentioned above, the estimates of Q values, which represent the mutation process directionality, indicate that E-to-G substitutions differ in terms of the transition/transversion rate from A-to-G ones to a much greater extent, than G-to-E substitutions differ from G-to-A ($Q_{\rightarrow *} >> Q_{* \rightarrow}$,); in addition, $Q_{* \rightarrow} < 1$ at non-synonymous sites (Fig. 3C), again supporting the idea that the E-to-G transitions contribute to the observed effects to a larger degree. Ultimately, this issue would be resolved when more data are available, allowing for the reconstruction of ancestral states of NCES.

Conrad Hal Waddington has proposed that a trait exhibiting extreme values in a novel environment due to the phenotypic variation present in a population will be canalized through subsequent evolution (Waddington 1953a; 1953b). Evidently, the trait has to be phenotypically plastic, i.e. to exhibit genotype×environment interaction covariance (Lynch and Walsh 1998; Ghalambor et

al. 2015; Ghalambor et al. 2007; Levis and Pfennig 2019). Over the years, the subject of phenotypic plasticity generally facilitating adaptation through genetic assimilation remained debatable (Ghalambor et al. 2015), and, despite multiple examples of genetic assimilation (Levis and Pfennig 2019), large-scale studies of differential expression of genes under environmental changes before introduction to a novel environment and in the novel environment before and after adaptation have shown that genetic changes tend to reverse rather than enhance the plastic ones (Ghalambor et al. 2015). Considering that editing may be influenced by environmental changes, one may imagine the environmental variance of editing, and, moreover, directed changes in the editing status in novel environments (Duan et al. 2017). Thus, positive selection acting on E-to-G transitions may be interpreted as genetic changes reinforcing phenotypic changes in the course of adaptation. However, the reinforcement of plastic changes by subsequent adaptation is notoriously difficult to prove, and the research design has to meet a number of specific criteria, such as the ability to show the very presence of genotype×environment interaction covariance, which we do not have sufficient data to test (Duan et al. 2017). This explanation is also contradicted by A-to-I editing being performed by a single small family of ADAR enzymes (Eisenberg and Levanon 2018), that for simple combinatorial reasons cannot provide a complex and specific response to a novel environment as differential expression. Nonetheless, editing in coleoids could be regulated by other proteins or processes such as the dependence of local RNA structures on temperature, and hence coleoid mRNA editing could be a good object for future studies of genetic assimilation.

Yet another possible explanation is in terms of preadaptation. The term «preadaptation» refers to a pre-existing structure that has changed its function in the course of evolution, a concept introduced by Charles Darwin (Darwin 1872; McLennan 2008; Cadotte et al. 2018; Stephen Jay Gould and Vrba 1982b; Ardila 2016; Casinos 2017), and currently it is applied to both morphological and molecular traits. This definition presumes that we can identify three stages of the structure's evolution: (1) structure with the ancestral function, (2) structure that has acquired a novel, derived function, but

retained the ancestral one, and (3) structure with only the derived function (Gould and Vrba 1982; McLennan 2008). Stage (1) is optional, as a structure that would be beneficial in the future could emerge by neutral evolution and have no specific ancestral function (McLennan 2008). In the case of non-conserved editing we seem to observe a rather similar pattern — an ancestral adenine that, through a transitory stage of an edited nucleotide, where two mRNA isoforms are present, is substituted with guanine. Under this hypothesis, we would expect edited adenines to be substituted more frequently and directionally to guanines, positive selection to act on such substitutions, and these effects to be more pronounced for highly edited sites and for more closely related species, as preadaptation has been shown to be better seen when closely related species are considered. These criteria are basically the same as those listed in the Introduction section, and hence the preadaptation scenario could be applied here. One problem is, that *de facto* we have not observed the complete chain of events, our findings being restricted only to finding that E-to-G transitions are selected for and to frequent emergence of editing sites from unedited adenines. Another problem with this explanation is that one cannot establish the function of every editing site, and, in order for the preadaptation explanation to be applicable here, we have to extend the meaning of "function" in the definition of preadaptation to a broader term, e.g. "phenotypic manifestation" of a nucleotide, which would include traits such as the occurrence of a specific amino acid at a specific position of a protein.

Hence, the most reasonable framework for our findings seems to be in terms of non-functional editing sites enhancing the expressed genetic variability, thus contributing to the acceleration of the evolutionary process at sites with beneficial A-to-G substitution. The Continuous Probing Hypothesis (Gommans, Mullen, and Maas 2009) states that editing sites, due to the lack of a strict context, constantly emerge at random points of the transcribed genomic regions. Hence, an adenine with a beneficial substitution to guanine could become edited if the editing context emerges around it purely by chance. The context can be further selected upon, resulting in the mimicking of the beneficial guanine variant.

A similar rhetoric can be applied to other cellular information transmission processes such as transcription and splicing. These processes depend on regulatory sites and contexts that change the quantity, dynamics (developmental stage, tissue-specificity, response to external conditions), and sequence of encoded proteins and hence are subject to selection (Raj and van Oudenaarden 2008; Pickrell et al. 2010). Hence a natural extension of this study would be to systematically assess the evolutionary advantage of noise in information transmission processes in low-polymorphic populations.

Chapter 4. A hierarchy in clusters of cephalopod mRNA editing sites

This chapter describes our study "A hierarchy in clusters of cephalopod mRNA editing sites" published in Scientific Reports in 2022. Authors: Moldovan, M., Chervontseva, Z., Nogina, D. & Gelfand, M. S.

RNA editing in the form of substituting adenine with inosine (A-to-I editing) is the most frequent type of RNA editing in many metazoan species. In most species, A-to-I editing sites tend to form clusters and editing at clustered sites depends on editing of the adjacent sites. Although functionally important in some specific cases, A-to-I editing usually is rare. The exception occurs in soft-bodied coleoid cephalopods, where tens of thousands of potentially important A-to-I editing sites have been identified, making coleoids an ideal model for studying of properties and evolution of A-to-I editing sites. Here, we apply several diverse techniques to demonstrate a strong tendency of coleoid RNA editing sites to cluster along the transcript. We show that clustering of editing sites and correlated editing substantially contribute to the transcriptome diversity that arises due to extensive RNA editing. Moreover, we identify three distinct types of editing site clusters, varying in size, and describe RNA structural features and mechanisms likely underlying formation of these clusters. In particular, these observations may explain sequence conservation at large distances around editing sites and the observed dependency of editing on mutations in the vicinity of editing sites.

4.1. Introduction

The mRNA editing process, where an adenine is substituted by inosine (A-to-I editing), is a widespread mechanism of transcriptome diversification in metazoans (Bass and Weintraub 1988; Reenan 2005; Yang et al. 2008; Ensterö et al. 2009; Morse, Aruscavage, and Bass 2002). Inosine is recognized by the cellular machinery as guanine (Xu and Zhang 2014; Wahba et al. 1963; Sommer et al. 1991; Kazuko Nishikura 2006; 2010; 2016), and hence the proteins translated from an edited

transcript may be re-coded, thus contributing to the proteome diversity (Alon et al. 2012; 2015; Liscovitch-Brauer et al. 2017; Eisenberg and Levanon 2018). A-to-I editing is performed by the family of ADAR enzymes, and mutations corrupting ADAR may cause reduction of fitness in model organisms and disease in humans (Kazuko Nishikura 2010; Eisenberg and Levanon 2018; Garrett and Rosenthal 2012; Feldmeyer et al. 1999; Brusa et al. 1995; Maas et al. 2006).

Still, A-to-I editing sites are rare in coding regions of most genomes studied so far, with only minor fractions of them being conserved or functionally important (Yang et al. 2008; Yablonovitch et al. 2017; Ramaswami et al. 2012; Kim et al. 2004; Pinto, Cohen, and Levanon 2014; Yu et al. 2016). However, in coleoids (soft-bodied cephalopods, Fig. 5A), not only A-to-I editing is frequent, but is also more functionally important than in other studied lineages, i.e. mammals and *Drosophila* (Liscovitch-Brauer et al. 2017; Eisenberg and Levanon 2018; Alon et al. 2015). Editing in coleoids involves up to 1% of all adenines in the transcriptomes and has been suggested to play an important role in proteome diversification, allowing for responses to many environmental cues, such as phenotypic adjustments to low temperatures (Liscovitch-Brauer et al. 2017; Eisenberg and Levanon 2018; Shoshan et al. 2021). Along with that, editing sites could have an evolutionary value by rescuing deleterious G-to-A substitutions (Chen 2013; Jiang and Zhang 2019) or by providing heritable phenotypes selection can act upon, thus enhancing the rate of adaptation (Moldovan et al. 2020; Popitsch et al. 2020).

To edit transcripts, ADAR enzymes require specific features of the sequence around editing sites (Reenan 2005; Ensterö et al. 2009; Morse, Aruscavage, and Bass 2002; Alon et al. 2012; Liscovitch-Brauer et al. 2017b; Savva, Rieder, and Reenan 2012). Along with the edited adenine itself, a specific nucleotide context is required at positions ±1 relative to the edited adenine. However, the consensus at these positions is rather weak (Liscovitch-Brauer et al. 2017; Alon et al. 2015; Moldovan et al. 2020; Eggington, Greene, and Bass 2011). The ADAR enzymes also require edited adenines to be located in RNA helices, which may form complex structures spanning over 1kb of linear nucleotide

sequence (Reenan 2005; Morse, Aruscavage, and Bass 2002; Nishikura et al. 1991; Morse and Bass 1999; Paz-Yaacov et al. 2010). Thus, editing at individual sites may be influenced by distant loci, which has been shown directly by the edQTL analysis (Kurmangaliyev, Ali, and Nuzhdin 2016). However, on average, the span of regions affecting editing at a particular site is about 200–400 nt (Liscovitch-Brauer et al. 2017), as shown by edQTL studies and analysis of sequence conservation in regions around editing sites in *Drosophila* (Kurmangaliyev, Ali, and Nuzhdin 2016) and coleoids (Liscovitch-Brauer et al. 2017), respectively.

Nonetheless, the ADAR requirements on sequence and structure to edit a particular site are rather weak, yielding multiple weakly edited adenines in every studied transcriptome. Consequently, editing sites have been proposed to form constantly at random points of the genome, especially in structured RNA segments (Gommans, Mullen, and Maas 2009). Adjacently located edited adenines tend to be edited simultaneously (Nishikura et al. 1991; Duan et al. 2018; Polson and Bass 1994; Zhang and Carmichael 2001; Prasanth et al. 2005). In human and *Drosophila*, such correlations are mainly due to the involvement of such sites in the same secondary RNA structures. Additionally, editing sites located in coding regions are clustered for *Drosophila* and leaf-cutter ants, with clusters arbitrarily defined as groups of editing sites where adjacent sites are located at most at 30–50 nt from each other (Li et al. 2014; Zhang et al. 2017).

Clustering of editing sites has been extensively studied in tandem, differently oriented Alu repeats, where formation of Alu-Alu double helices is common (Paz-Yaacov et al. 2010; Levanon and Eisenberg 2015; Athanasiadis, Rich, and Maas 2004; Barak et al. 2009). Editing of Alu repeats has been hypothesized to protect against negative effects of Alu repeats on the organism's fitness (Eisenberg and Levanon 2018) and Alu sequences may be used as indicators of the general editing activity in tissues (Paz-Yaacov et al. 2010). Extensive editing at Alu sequences also allows for the analyses of subtle features of ADAR-mediated editing such as the correlated editing at specific sites

(Paz-Yaacov et al. 2010) or establishment of preferential sequence of editing events along Alu-containing transcripts (Barak et al. 2009).

In coding sequences, clusters of A-to-I editing sites are also present and abundant, with clustered editing sites being on average more conserved and heavily edited than their individual counterparts (Liscovitch-Brauer et al. 2017; Duan et al. 2018). The enhanced conservation of clustered editing sites, their distance-dependent linkage, and dependencies of editing at one site on editing at another (Duan et al. 2018; Barak et al. 2009) suggest the importance not only of A-to-I editing *per se*, but also of dynamics of the editing process, so that editing tends to occur simultaneously at a multitude of sites in a given transcript. This hypothesis is supported by the observation that non-synonymous editing sites in protein-coding sequences are more clustered than synonymous ones (Duan et al. 2018).

By having a large number of A-to-I editing sites, coleoids are a perfect model for studying subtle evolutionary and statistical features of RNA editing. One relevant question is posed by the possible structure of A-to-I editing clusters and the processes underlying formation of clusters with specific structures. As coleoid editing sites demonstrate same contextual features and secondary RNA structure requirements as mammalian or *Drosophila* editing sites (Fig. 5BC) (Liscovitch-Brauer et al. 2017; Moldovan et al. 2020; Gommans, Mullen, and Maas 2009), studying coleoids as a convenient model we may enhance our understanding of the ADAR action in general and of the evolutionary and functional mechanisms involved in the emergence of new editing sites.

Here, we rely on four coleoid transcriptomes to show that the level of association between A-to-I editing at individual sites in coding regions strongly depends on the distances between sites. The underlying intuition here is as follows: closely located adenines should be similar in terms of local RNA structure, and if one of them is edited, the other one is more likely to have the necessary prerequisites for ADAR-mediated editing. Hence, we expect more closely located adenines to be edited simultaneously with a higher probability than more distantly located ones. with the highest correlation observed for immediately adjacent edited adenines (Fig. 5D).

By applying multiple and diverse approaches to analyze the distribution of editing sites along transcripts, we identified three distinct types of clusters of coleoid editing sites with sharply different characteristic size ranges. Analyzing local RNA structural features, we observe a tendency of editing sites to be located in putative loops, mismatches or bulges in secondary RNA structures, in agreement with observations of individual A-to-I editing sites that form A-C mismatches in RNA helices (Morse, Aruscavage, and Bass 2002; Morse and Bass 1999; Wong, Sato, and Lazinski 2001; Kallman 2003). In addition, we show that correlated editing in coding regions strongly contributes to transcriptome diversity driven by ADAR-mediated editing in general and that editing in clusters generally occurs in the 3'-to-5' direction.

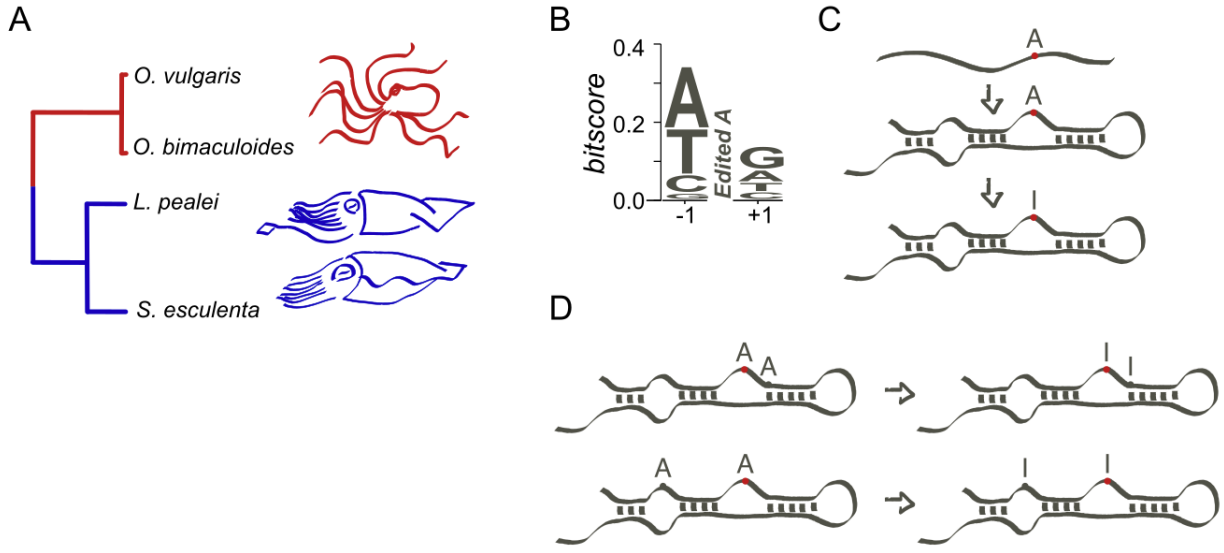


Figure 5 | (A) Phylogenetic tree of four mollusks (octopuses *Octopus vulgaris* and *O. bimaculoides*, squid *Loligo pealei*, and cuttlefish *Sepia esculenta*) considered in this study. The tree has been taken from TimeTree (Hedges, Dudley, and Kumar 2006). **(B)** Sequence context of coleoid A-to-I editing sites. **(C)** ADAR enzymes performing A-to-I editing require secondary RNA structures. **(D)** Editing at closely and at distantly located sites. See the text for details.

4.2. Methods

4.2.1. Data

We used previously published transcriptomes (Liscovitch-Brauer et al. 2017b) of *O. vulgaris*, *O. bimaculoides*, *S. esculenta*, and *L. pealei* along with the publicly available coleoid editing sites data (Liscovitch-Brauer et al. 2017). The corresponding transcriptomic read data, summarized in Supplementary Table S1, were downloaded from the SRA database. For each species, corresponding SRA files were pooled. For the analysis of exons, we used the publicly available genomic sequences and annotation of *O. bimaculoides* (Albertin et al. 2015).

4.2.2. Calculation of *S* values

S values were calculated as nucleotide distances between edited adenines on transcripts. Along with *S* values calculated for actual editing sites, we calculated *S* values for randomly selected adenines. To eliminate the biases caused by factors such as the higher accuracy of editing sites prediction in highly expressed transcripts or the general tendency of some transcripts to be edited more frequently than others, we have randomly selected in each transcript the number of adenines equal to the number of editing sites it contains (see 4.2.3. "Control sets of adenines"). *S** values were calculated as nucleotide distances between subsequent edited adenines, i.e., for pairs of editing sites with no edited adenines between them.

4.2.3. Control sets of adenines

As a uniform control, we constructed a random set of adenines as follows: in each transcript we selected the number of random adenines exactly equal to the number of editing sites in this transcript. By using this control, we address the issue of non-uniformity of adenine occurrence along the transcripts and, in particular, possible overrepresentation of AA dinucleotides.

In addition, we constructed a second set of random adenines such that the three-nucleotide context of randomly chosen adenine set matched that of editing sites (Fig. 5B). Thus, we addressed

the non-uniformity of distribution of adenines in specific contexts along the transcripts. The procedure was as follows: for each transcript and for each editing site, we considered its 3-nucleotide context. Next, we selected a random adenine in the same transcript in the same context. This procedure ensures that the number of sites in the control set is exactly the same as the number of editing sites and the contexts of the control set exactly matches that of the edited adenine dataset.

To check for the stability of constructed controls with respect to the sampling of adenines, we have constructed 1000 control sets of adenines for *O. vulgaris* and calculated the maximal DC size in these sets (Suppl. Fig. S32). Maximal DC size was 3 in 97% of cases and 4 – in the remaining 3%. This shows that the procedure is robust with respect to random sampling variance.

4.2.4. Editing state co-occurence

The reads were mapped onto the transcripts with the bowtie2 package (Langmead and Salzberg 2012) using the –sensitive-local alignment mode. We filtered out read alignments that did not contain regions of continuous read mappings larger than half the read length. The resulting alignment files were further processed with a set of ad hoc scripts and the numbers of editing state occurrences for each considered editing site pair were calculated.

To infer the tendency of edited states to co-occur in transcripts, we have calculated the Pearson correlation (Pearson 1895) of edited state occurrences for all pairs of edited adenines located within the window with the radius equal to the read length.

4.2.5. Variance due to editing

To estimate the variance in transcriptomes conferred by editing, at each position, we formally assigned values 0 and 1 respectively to edited and non-edited reads mapping to this position. Thus, the variance at a position is simply $EL(1–EL)$, where EL stands for the editing level at the considered position. Alternatively, this can be written as $f_i^A f_i^I$, as it is written in the Results section. The additive variance component is thus the sum of $f_i^A f_i^I$ over all edited adenines in the transcriptome. The net

variance is calculated as the sum of all variances and covariances in the form of $f_{i,j}^{AA} f_{i,j}^{II} - f_{i,j}^{AI} f_{i,j}^{IA}$ (see text for details). Additionally, all between-site covariances considered in this analysis had to be significant ($p < 0.05$, t-test with FDR correction for multiple testing), otherwise they were formally set to zero.

4.2.6. RNA structural annotations

To estimate the propensity of sequences to form RNA secondary structure, we have calculated the structural potential Z-scores for each nucleotide using the RNASurface program (Soldatov, Vinogradova, and Mironov 2014). *Z-score* is defined as $Z = (E - \mu)/\sigma$ with E, μ and σ being the minimal free energy of a considered cequence, mean and standard deviation of the free energy distribution of shuffled sequences with preserved length and average dinucleotide composition, respectively. RNASurface was run with the maximal and minimal sliding window length set to 350 and 20, respectively. For each position, *Z-score* was inferred as the minimal *Z-score* over all windows containing the position.

The base pairing probabilities were calculated with the plfold algorithm of the Vienna package (Lorenz et al. 2011) with –W and –L parameters set to sequence lengths and –cutoff parameter set to 0.0.

For the analysis of editing sites brought close by secondary RNA structures, all possible pairs of editing sites for each transcript were considered. For Fig. 10B, every such pair was assigned to one of the three groups: “close due to structure”, “distant, unstructured”, or “intermediate”. Two editing sites were considered close due to the structure if the distance between them in the structure was less than the distance between them taken by sequence, and, additionally, the distance in structure was less than eight nucleotides. The pair of sites were considered distant if the distance between the sites in the structure was equal to the distance by the sequence or was more than 40 nucleotides. The distance in the structure for the pair of sites was computed as the minimal distance between them in the graph of the transcript with all the potentially paired base pairs and all nucleotides adjacent in the sequence

connected by edges. The graph was obtained using the Vienna RNAplfold program with the 0.8 cut-off for the pairing probability (Lorenz et al. 2011).

4.2.7. Structural mismatch annotations

The tendency of edited adenines to mismatch with cytosines in RNA structures was estimated as follows. Given probabilities for every two nucleotides to be paired in a structure, we selected all adenines sandwiched between two base pairs with pairing probability higher than 0.7 and having one-nucleotide symmetrical bubble in between. In total, we obtained sets of 455 and 4298 edited and non-edited, respectively, adenines in such structures. Then we compared the distribution of such mismatch partners for the edited adenines and for the nonedited ones (Supplementary Figure S16). The probability of pairing was computed using RNAplfold (Lorenz et al. 2011) for all exon sequences as they would have been prior to editing events.

4.2.8. Order of editing events

In each 50 nt window in the read alignment on the transcriptome of *O. vulgaris*, we looked for at least 5 editing sites covered by at least 20 reads. Further, we required each transcript state, defined as the editing pattern in the read, to be supported by at least 5 reads, and the number of states to be at least 4 if one of the states represented the all-A (not edited) state and at least 3 otherwise. If the alignment lacked an all-A state, this state was added artificially as an outgroup. Next, trees were built upon the selected A-I sites treated as polymorphisms by the maximum-likelihood algorithm implemented in the IQ-TREE package (Nguyen et al. 2015). The All-A states were used as outgroups. Quasi-ancestral states representing editing intermediates were reconstructed with the TreeTime package (Sagulenko, Puller, and Neher 2018). To control for the quality of the tree reconstruction, we additionally filtered out 13% of the constructed trees where I-to-A substitutions were identified, as the A-to-I editing is irreversible.

An editing path is defined as the sequence of editing events that occur when travelling from the root (all-A sequence) to the leaves of the tree. Additionally, the weight of each A-to-I substitution event was defined as the number of reads representing descendants of this event. The weight of a path was then defined as the weight of the corresponding terminal leaf, which corresponded to the number of reads in one specific editing state. In other words, the weight of an editing path is the minimal weight of an event in this path and represents the number of edited reads that emerged as a consequence of a specific path of editing events. To obtain the resulting ensemble of editing paths, we pooled the constructed paths with corresponding weights. The resulting ensemble consisted of 1529 different editing paths with the total weight of 51326.

4.2.9. Statistics

The tendency of editing states to co-occur on the transcripts and correlations between the editing levels at pairs of sites were assessed with the Pearson's correlation (Pearson 1895). The confidence intervals and the significance of each correlation coefficient were inferred using the t-test with the Bonferroni correction (Bonferroni 1936) for multiple testing. The distributions of *S* values were compared using the two-sample Kolmogorov-Smirnov test (Kolmogorov 1933). Editing levels, the distributions of correlation coefficients, and the distributions of structural potential Z-scores were compared with the Mann-Whitney U test (Mann and Whitney 1947). The editing levels at upstream and downstream editing sites were compared with the Wilcoxon's signed-rank test (Wilcoxon 1945).

The grouping of *S* values with respect to the differences in correlations between edited states on transcripts was performed using the Mann-Whitney U test: for each pair of correlation arrays corresponding to different *S* value ranges, the Mann-Whitney statistic was calculated, and groups of *S* value ranges were further defined as the groups of sequential ranges differing insignificantly from each other.

4.2.10. Code availability

All data analyses were performed in Python 3.7. Scripts and data analysis protocols are available online at https://github.com/mikemoldovan/coleoidRNAediting2.

4.3. Results

4.3.1. Correlated editing

In model metazoan species, editing may be correlated if the sites are located sufficiently close to each other (Duan et al. 2018). The unusually large numbers of coleoid editing sites allowed us to assess the interplay between co-occurrences of editing states and the distances between editing sites at the single-nucleotide resolution. In our study, we used the available trascriptomes and editing site sets for four coleoids — two octopuses *Octopus vulgaris* and *O. bimaculoides*, *Sepia esculenta* (cuttlefish), and *Loligo pealei* (squid) (Liscovitch-Brauer et al. 2017). We used raw RNAseq data (Supplementary Table S1) to calculate the correlation of edited states for each pair of editing sites separated by at most the distance equal to read lengths in our dataset (~100–150 nt) (Supplementary Table S1). The correlation coefficients for a pair of edited adenines E_i and E_j given the RNAseq read mapping to transcripts is defined as in (Duan et al. 2018) (Fig. 6A, Supplementary Figures S11, S12): $r(E_i, E_j) = (f_{i,j}^{AA} f_{i,j}^{II} - f_{i,j}^{AI} f_{i,j}^{IA}) / \sqrt{f_i^A f_i^I f_j^A f_j^I}$, where $f_{i,j}^{N_1 N_2}$ are frequencies of co-occurrences of observed nucleotides N_1 and N_2 (A or I/G) at positions *i* and *j* in the RNAseq read data, and f_i^N are frequencies of nucleotide *N* in the read mapping data at position *i*. We compared the distributions of $r(E_i, E_j)$ for different inter-site distances, which we refer to as the *S* values, *S* defined as $j - i$ (Fig. 6A). The correlations were on average higher for immediately adjacent sites, with mean $r(E_i, E_j)$ values further decreasing with the increase of the *S* distance, consistent with observations in Duan et al., 2018.

The editing level (EL) of an A-to-I editing site is defined as the percentage of mapped reads in a sample containing inosine (read as guanine) at the considered site. As the editing levels of most sites

are rather low (<10%), one could speculate that the bulk of associations is lost in the above analyses due to missed low-EL sites that could not be retrieved from the data (Eggington, Greene, and Bass 2011). Indeed, if we consider sites with EL≥5% (Supplementary Figure S12a), the average $r(E_i, E_j)$ values increase almost twofold, reaching 0.43 for *S*=1. To check whether higher $r(E_i, E_j)$ values are not simply a property of efficiently edited sites, we calculated the $r(E_i, E_j)$ distributions for sites with the EL≥10% and obtained only slightly larger $r(E_i, E_j)$ values, as compared to sites with EL≥5% (Supplementary Figure S12b). Thus, the association between the A-to-I editing events is indeed strong, especially for adjacently located editing sites.

To check whether the editing state co-occurrence manifests as similarities between ELs, we assessed the correlations between the ELs at individual sites for a series of *S* values (Fig. 6B, Supplementary Figure S13). For immediately adjacent editing sites (*S*=1), this correlation turned out to be on average twofold larger than for any other *S* (p<0.001, the *t*–test). If adjacent sites are not considered, the correlations in ELs only slightly depend on *S*, being significant (p<0.05, the *t*–test) even for quite distantly located sites (*S*>500). Non-zero EL correlations at very large distances may be explained by some transcripts being edited to a higher overall degree than other transcripts. An alternative explanation is as follows. The general variance of ELs in the transcriptome may be decomposed into two summands: the between-transcript variance and the within-transcript variance, the former being the variance of the mean EL values in transcripts, and the latter being the variance of the deviations of ELs from the means in each transcript. If the between-transcript variance is non-zero due to, e.g. low average numbers of editing sites per transcript yielding the estimates of means with high variance, we would observe a baseline correlation for any *S* value, which is simply not defined for sites located in different transcripts.

In theory, correlated editing at different sites may enhance the transcriptome diversity defined as the number of possible states with respect to editing. So if there is one editing site, which can be either in inosine or adenine state, this number would be 2, if there are two sites – 4, *etc*.. (Barak et al.

2009). Here, the increase in transcriptome diversity due to correlated editing may seem counterintuitive, as dependencies in editing events should generally decrease the possible numbers of transcript variants in a given cell. However, the number of possible transcriptome states in coleoids even under complete linkage of editing events is still astronomically large, and thus hardly represents a bottleneck of transcriptome diversity: on average, about 8000 coleoid genes are edited, which yields 2^{8000}, or 10^{2408} transcriptome states.

An alternative approach here could be to assess the variance in transcriptome and proteome generated by editing. If we, following the definition of between-site correlation, define the variance in an editing site i as $f_i^A f_i^I$ and the covariance between two sites i and j as $f_{i,j}^{AA} f_{i,j}^{II} - f_{i,j}^{AI} f_{i,j}^{IA}$, we can calculate the net variance generated by editing to be up to 111 in transcriptomes and up to 92 in proteomes (Supplementary Table S2). In the context of populations, such variance can be generated by 888 and 736 two-allele polymorphisms with minor allele frequencies of 0.5 without dominance, which is quite a lot. Moreover, we find almost half of this variance to be explained by correlated editing at pairs of sites, namely, up to 46.3% of the transcriptome variance and up to 46.5% of the proteome variance due to non-synonimous editing. These percentages likely represent lower bounds, as covariances incorporated in this analysis have had to satisfy stringent statistical criteria, otherwise they have been set to zero (see 4.2. Methods).

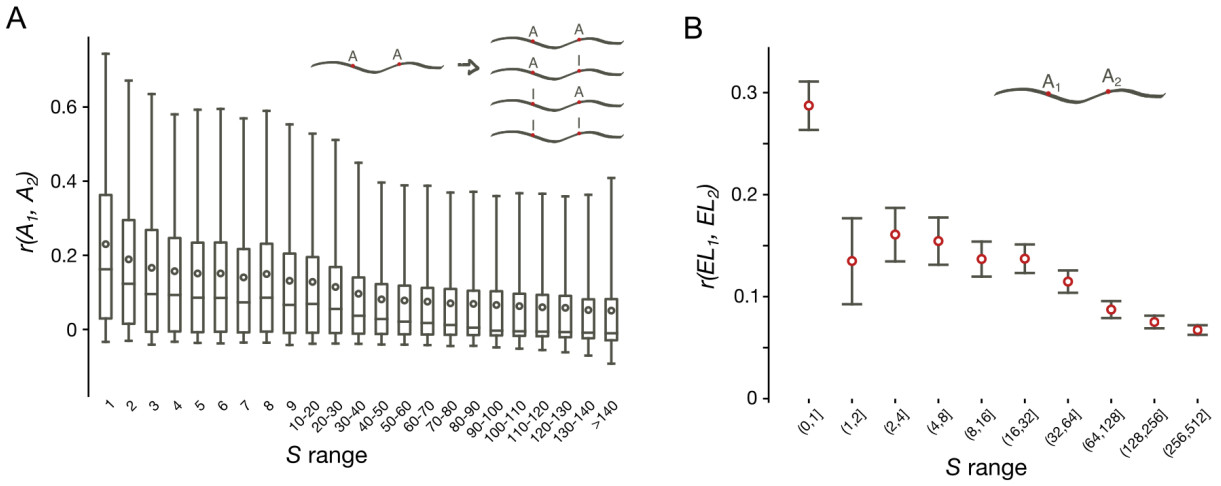


Figure 6 | Correlations between various properties of editing sites. (A) Distributions of correlation coefficients of *O. vulgaris* editing (*r)* at two sites with respect to the distances between sites (*S*). Boxes represent quartiles, red circles represent the means and the grey lines (whiskers) indicate 95% two-sided confidence intervals of distributions. **(B)** Dependence of correlations of ELs on the *S* distance, *O. vulgaris* dataset. Red circles mark values of correlation coefficients and grey lines represent Bonferroni corrected 95% two-sided confidence intervals obtained from the *t*-distribution.

4.3.2. Dense editing site clusters (adjacent adenines)

Notably, the correlation between ELs is by far the highest for immediately adjacent editing sites with *S*=1 (Fig. 6B). We consider these sites separately and refer to them as *dense editing site clusters* (DCs) in the general case, and as *paired editing sites* if there are only two adenines per cluster. The observed enhanced positive correlation of editing site co-occurrence for dense clusters (Fig. 6) hints at editing at a focal site being dependent on editing at the immediately adjacent adenine. This could lead to overrepresentation of DCs in the coleoid transcriptomes.

To check whether DCs are indeed overrepresented, we calculated the numbers of sites in DCs separately for each DC size across the coleoid transcriptomes (Fig. 7A, Supplementary Figure S14). As controls, we randomly selected adenines with and without regard to the local trinucleotide context (see Methods). The results obtained for the two control sets did not differ (Supplementary Figure S14)

and the results obtained for multiple permutation rounds did not differ (permutation $p < 10^{-3}$, Supplementary Figure S32). For all DC sizes, which ranged from two to eight consecutive adenines, the site count in the real datasets was larger than that in the control datasets, the effect being stronger for DCs with larger numbers of adenines (Fig. 7A, Supplementary Figure S14A, Supplementary Figure S15).

Given the observed stronger association of editing at heavily edited adenines compared to that of weakly edited ones (Fig. 6A, Supplementary Figure S12), one would expect enhanced editing levels of adenines in DCs. However, the enhanced levels of editing at clustered sites should also be taken into account. Thus, following Duan et al., 2018, we have divided editing sites into clustered sites with the between-site distance smaller than 100nt and individual sites, for which no editing is observed in the 100nt vicinity. To disentangle effects on editing conveyed by <100nt proximity and by location of sites in DCs, we further divided the set of clustered adenines into editing sites located in DCs and non-DC clustered sites, and compared the distributions of ELs in all three resulting categories of sites (Fig. 7B). The average ELs of sites in DCs were up to 1.67-fold larger than those of individual sites ($p<2.4\times10^{-7}$, the Mann–Whitney U–test) and up to 1.59-fold larger than average ELs of non-DC clustered sites ($p<9.8\times10^{-201}$, the Mann–Whitney U–test). Accordingly, the fraction of heavily edited sites (EL > 50%) in DCs is up to 3.42-fold larger than that in individual sites ($p<1.96\times10^{-45}$, the χ^2 contingency test) and up to 2-fold larger than in non-DC clustered sites ($p<2.26\times10^{-6}$, the χ^2 contingency test). Interestingly, we did not observe consistently significant differences between ELs at individual *vs.* non-DC clustered sites, which shows that the effects of clustering on EL observed by Duan et al. are largely conferred by densely clustered sites. However, non-DC clustered sites differ from individual ones when the fraction of heavily edited sites is considered, which is up to 1.65-fold higher in non-DC clustered sites ($p<1.7\times10^{-4}$, the χ^2 contingency test).

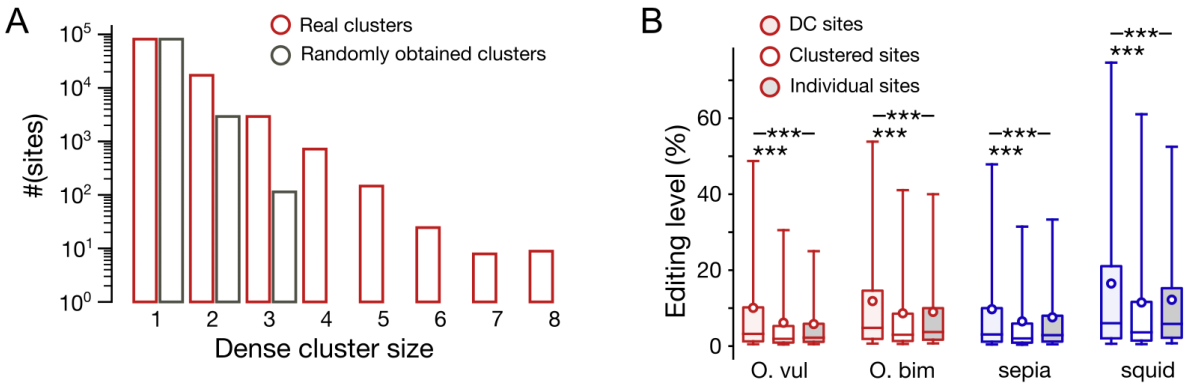


Figure 7 | Properties of densely clustered A-to-I editing sites. (A) Histogram of dense cluster sizes (nt) for the real *O. vulgaris* editing site dataset (red) and a matching random dataset (grey). **(B)** Comparison of editing levels in densely clustered (S=1, red and blue filled boxes), not densely clustered ($1<S<100$, white boxes) sites, and individual sites ($S\geq100$, grey-filled boxes). Three asterisks mark statistical significance of the differences in means ($p < 0.001$, the Mann–Whitney U–test).

4.3.3. Medium-range clusters of editing sites

Previous studies and the observed correlations in the editing state co-occurrence for S values larger than 1 (Fig. 6A) hint that A-to-I editing sites may cluster not only in the form of DCs (Morse, Aruscavage, and Bass 2002; Liscovitch-Brauer et al. 2017; Nishikura et al. 1991; Morse and Bass 1999; Duan et al. 2018; Kallman 2003). Thus, we checked how the distance to the nearest editing site affects the probability of adenine editing (Fig. 8A). We introduce the measure S^* defined as the distance between two edited adenines such that no other edited adenine is located between them, and consider the deviation of the observed S^* distribution from the expected one (Fig. 8A, Supplementary Figure S14B). The expected distributions were calculated on randomly generated datasets described above. For all considered coleoid species, the observed and expected S^* distributions differ significantly only for windows of up to 18 nucleotides ($p<0.01$, the χ^2 test with the Bonferroni correction), thus suggesting a direct dependence of editing events within the 18nt distance.

As noted above, A-to-I editing requires secondary RNA structures to be formed around the edited adenine (Reenan 2005; Morse, Aruscavage, and Bass 2002; Nishikura 2010; Moldovan et al. 2020; Nishikura et al. 1991; Levanon and Eisenberg 2015; Athanasiadis, Rich, and Maas 2004; Kallman 2003). Hence, the observed clustering of editing sites may be explained by common RNA structures at clustered sites. Thus, we have assessed the average size of a local secondary RNA structure by analyzing average base pairing probabilities of nucleotides around editing sites (Fig. 8B). To control for the accuracy of our predictions of RNA structures around editing sites, we checked for the presence of a well-known effect, where edited adenines tend to form A-C mismatches in RNA double helices more than their non-edited counterparts (Morse, Aruscavage, and Bass 2002; Morse and Bass 1999; Wong, Sato, and Lazinski 2001; Kallman 2003). Indeed, this effect was substantial (Supplementary Figure S16) and highly significant (p=5.1×10^{-34}, Fischer's exact test).

The average RNA structure size for each coleoid species is determined as the average width of peak in pairing probabilities of nucleotides centered at editing sites; the peak is defined at the region where the average base-pairing probabilities are greater than those of nucleotides distant from editing sites. So defined peaks for all four considered coleoid species fall in the range 32–45 nt, which is consistent with the above estimate of the distance at which an edited adenine influences the probability of editing of a neighboring adenine, which is 2×18 nt = 36 nt (Fig. 8A). Thus, the correlated editing of adenines located sufficiently close to each other indeed may be caused by common local secondary RNA structures. Moreover, as there is a higher probability of editing of adenines located in the vicinity of editing sites, editing sites should cluster along the transcript, forming what we call medium-range editing site clusters.

The result about editing sites being less likely involved in secondary RNA structures (Fig. 8B) seemingly contradicts earlier observations that these sites tend to reside within structured regions (Morse, Aruscavage, and Bass 2002; Moldovan et al. 2020; Morse and Bass 1999; Levanon and Eisenberg 2015; Athanasiadis, Rich, and Maas 2004; Kallman 2003). This controversy was resolved

by nucleotide-resolution structural analysis of regions around editing sites. For each edited adenine we sampled the nearest non-edited adenine as a control and assessed the site and control base-pairing probabilities (Supplementary Figure S17). The base-pairing probability of control sites turned out to be larger than that of editing sites, the effect being stronger for sites with large ELs (Supplementary Figure S17A). Moreover, the energy of the local secondary RNA structure was lower for editing sites compared to that of control ones (Supplementary Figure S17B), confirming that the RNA structure around editing sites is more stable on average than that at the editing sites themselves. The observed pattern suggests that editing sites generally tend to reside in loops or bulges, i.e. in non-paired regions surrounded by stable helices and are also likely to form A-C mismatches.

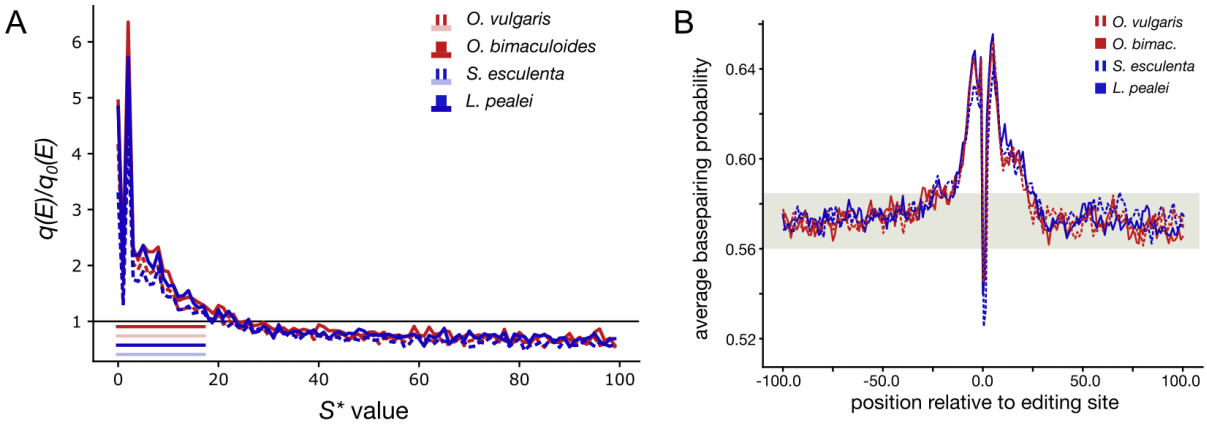


Figure 8 | Properties of medium-range clusters of editing sites. (A) Deviation of the editing probabilities of adenines located near editing sites ($q(E)$) from the respective expected probabilities ($q_0(E)$) as dependent on the S^* values. The colored stripes in the lower left corner represent the S^* value ranges on which $q(E)$ are significantly higher than $q_0(E)$ ($p < 0.01$, the χ^2 contingency test, Bonferroni corrected) **(B)** Average base-pairing probabilities in the regions centered at editing sites in four coleoid species. The gray stripe marks the base pairing probability range in regions distant from editing sites (>200 nt), considered as noise. The values above the noise (the central peak) describe the putative average RNA structure around editing sites; the width of the peak is the average size of the

structure. The dip in the middle is caused by generally low base-pairing probabilities of edited adenines.

4.3.4. Long-range clusters of editing sites

Earlier studies of coleoid editing sites demonstrated relatively higher sequence conservation in intervals of ±100–200 nt relative to conserved editing sites (Liscovitch-Brauer et al. 2017) and a correlation between differences in the editing levels at homologous sites and the number of mismatches in the ±100 nt region (Moldovan et al. 2020). These two consistent estimates indicate that editing at focal sites depends on ±100–200nt context, which exceeds the size of medium-range cluster sizes, established above as of 32–45 nt (Fig. 8).

Medium-range clusters have been identified by probability measures. A complementary approach is the comparison of real and expected *S* values, *S* being the distance (in nucleotides) between two edited adenines located in a single transcript, regardless of other possible editing sites between them. As with dense and medium-range clusters, the null models for *S* values were derived from random sets of adenines with the per-transcript number of editing sites preserved and with the tri-nucleotide context preserved (see Materials and Methods). We have observed that the distribution of distances, *S*, calculated for known coleoid editing site sets is bimodal with a high and distinct peak at 1, reflecting overrepresentation of edited adenines in dense clusters (Fig. 9A, red curve, Supplementary Figures S14C, S18). Having calculated distances *S* using the randomized set of adenines, we have observed strong and highly significant differences between the real and control *S* distributions ($p < 2.2\times10^{-308}$, the Kolmogorov–Smirnov test, Fig. 9A). At that, the differences are limited to distances *S* smaller than approx. 100–200 nt (Fig. 9, Supplementary Figure S17), consistent with the earlier observations mentioned above (Liscovitch-Brauer et al. 2017; Moldovan et al. 2020), and yields long-range editing site clusters at the scale of 200–400 nt.

To understand the mechanisms yielding long-range clusters, we applied a relaxed definition of RNA structure spanning over a pair of edited adenines. We considered pairs of adenines brought close

to each other in space by secondary RNA structure (see Methods). As a control, we considered pairs of sites such that no secondary RNA structure could be identified between them (Fig. 9B, Supplementary Figure S19). As a measure of co-operativity of editing, we employed the formula: $r'(A_i, A_j) = f_{i,j}^{II}/(f_i^I f_j^I)$, where $f_{i,j}^{II}$ is the frequency of co-editing at a pair of sites *i* and *j*, and f_i^I and f_j^I are the individual frequencies of editing at the respective sites. Editing sites brought close by secondary RNA structures were generally more co-operative (p=7.8×10^{-7}, the Mann-Whitney U-test) than the control sites, with the sites at distances 4–16 nt and 128–256 nt exhibiting significant increase in co-operativity when brought close by secondary RNA structure (p<0.05, the Mann-Whitney U-test with the Bonferroni correction for binning) (Fig. 9B, Supplementary Figure S19). This result indicates the effects on co-operativity at characteristic long-range cluster sizes to be brought about by secondary RNA structures. These structures could be expected to be rather weak on average, as the structural potentials of nucleotides at distances from editing sites larger than 36 are indistinguishable from noise (Fig. 8B).

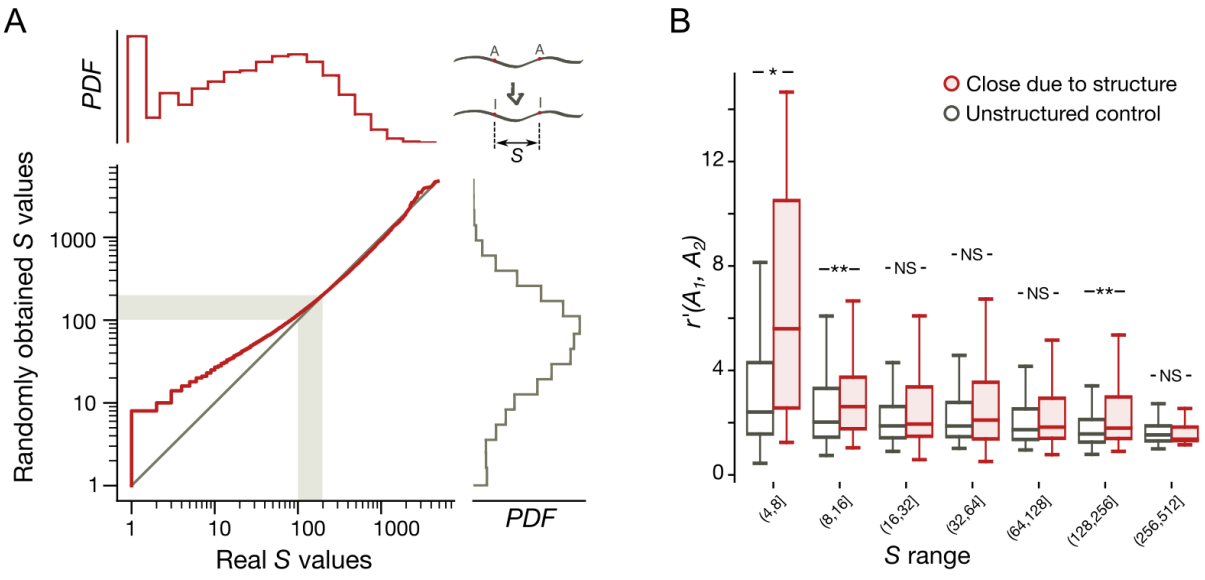


Figure 9 | Long-range editing site clusters. (A) Distribution of *S* distances for *O. vulgaris*. The real editing site set (red histogram) *vs*. randomly selected adenines (grey histogram), see the text for details. The red line is the plot of dependence between the real and the randomly obtained *S* values in arrays sorted by the distance *S*. The grey diagonal represents the expected dependence form $y=x$. Grey stripes represent the boundary of the possible span of regions affecting editing sites (Liscovitch-Brauer et al. 2017; Moldovan et al. 2020). PDF – probability density function. **(B)** Distributions of the $r'(A_i, A_j)$ values calculated for the structurally close editing sites (red boxes) and for the control site pairs with no predicted secondary RNA structure between the sites in a pair (grey boxes) (see the text for details). Asterisks mark statistical significance of differences of means calculated using the Mann-Whitney U-test with the Bonferroni correction for binning. Two asterisks indicate $p<0.01$; one asterisk, $p<0.05$, NS, not significant.

4.3.5. Directionality of editing

As noted above, the strongest association in terms of EL or the co-occurrence of edited states is observed for adjacent editing sites ($S=1$) (Fig. 6AB), with the two-adenine (AA) clusters comprising the vast majority of dense clusters (Fig. 7A). The observed effects may be due to co-operativity of editing, so that, if an adenine is edited, this would enhance the editing context for an adjacent adenine.

The editing context is asymmetric (Fig. 5B), hence we expect probabilities of editing of adenines located immediately up- and downstream from an editing site to differ. Moreover, the contextual features of editing sites were hypothesized to yield AI rather than IA as the preferred intermediate to the II dinucleotide in paired editing events, and consequently more AI-reads were observed in paired editing sites of coleoids (Duan et al. 2018). Indeed, the ELs at downstream sites are on average 4–6% higher than those of the upstream ones ($p<1.5\times10^{-80}$, the Wilcoxon signed-rank test) and this result does not depend on the position of the AA-cluster relative to the reading frame of the coding sequence (Fig. 10A, Supplementary Figure S20). Thus, the dynamics of editing of AA-clusters manifests as general differences in ELs at the up- and downstream adenines in DCs.

Re-coding (non-synonymous) A-to-I editing in coleoids might be beneficial, as it diversifies the proteome and, consequently, allows for appropriate phenotypic and evolutionary responses to novel environments (Liscovitch-Brauer et al. 2017; Eisenberg and Levanon 2018; Shoshan et al. 2021; Moldovan et al. 2020; Popitsch et al. 2020). In line with this reasoning coupled with the observation that downstream sites in AA-clusters were more prone to editing, we compared the fraction of sites with non-synonymous A-to-G substitutions among up- and downstream adenines in AA-clusters (Fig. 10B), where both adenines were edited, with the corresponding fractions in AA dinucleotides, where both adenines were not edited. The probabilities of the downstream sites to be re-coding was higher than those for the upstream sites ($p<3\times10^{-6}$, the binomial test) even accounting for differences in the probabilities of editing in AA dinucleotides.

The differences between ELs and the fractions of re-coding sites of up- and downstream paired edited adenines may be also explained by features of the local secondary RNA structure required for the ADAR action (Reenan 2005). We assessed the latter explanation by calculating the probabilities of each nucleotide to be involved in secondary RNA structures, which we refer to as the base-pairing probabilities (see Methods). For each paired editing site (EE-site), we considered the base-pairing probabilities of up- and downstream editing sites separately. As controls, we considered three sets of

AA dinucleotides located within ±20 nt windows around EE-sites: (i) pairs of non-edited adenines (AA-sites), (ii) downstream-edited and upstream-unedited adenines (AE-sites), and (iii) upstream-edited and downstream-unedited adenines (EA-sites). If none of the controls could be obtained for an EE-site, it was not considered further (Fig. 10C, Supplementary Tables S3, S4). As in the case with EE-sites, we considered base-pairing probabilities in control dinucleotides separately for up- and downstream nucleotides.

Firstly, we observed the base-pairing probabilities of downstream adenines in EE-sites to be significantly lower than those of upstream adenines (Wilcoxon $p < 2.6\times10^{-39}$). The dependency of base-pairing probabilities on the adenine position in a dinucleotide extends to the comparison of base-pairing probabilities of EE-sites with those of control dinucleotides (Fig. 10C, Supplementary Tables S3, S4), where the downstream adenine seems to be generally less structured than the upstream adenine. Additionally, positions of editing sites in the control sets largely and consistently affect the results: AE-sites are generally more structured than EA-sites (Fig. 10C). Thus, the downstream adenines in EE-sites are edited more frequently, are more likely to be re-coding if edited, and are less likely to be involved in secondary RNA structure.

These results suggest that editing at downstream sites is the primary event in DC editing, which may be followed by editing at upstream sites. To check this hypothesis, we reconstructed the temporal sequences of editing events in *O. vulgaris* transcriptome using an approach similar to the one of Barak et al., 2009 (see Methods). For coupled editing sites, we observed a significant tendency for the downstream sites to be edited prior to the upstream ones (Wilcoxon $p = 3.8\times10^{-35}$). One possible explanation for that would be a general tendency of ADARs to edit firstly down- and then upstream sites located nearby. To check it, we considered paths of editing events, where the pairs of editing sites are separated by more than one nucleotide ($S>1$). As in the case with coupled editing sites, we observed a significantly larger number of paths where downstream adenines were edited prior to the upstream ones (Wilcoxon $p = 1.4\times10^{-96}$). Thus, at least to some extent, the directionality in DC editing is

explained by the general directionality of editing. However, this result does not rule out an alternative possibility that changes in the local context of upstream sites introduced by editing at downstream sites induce editing at upstream sites, as suggested by the established editing site context, where the preferential downstream nucleotide for an edited adenine is guanine (Fig. 5B).

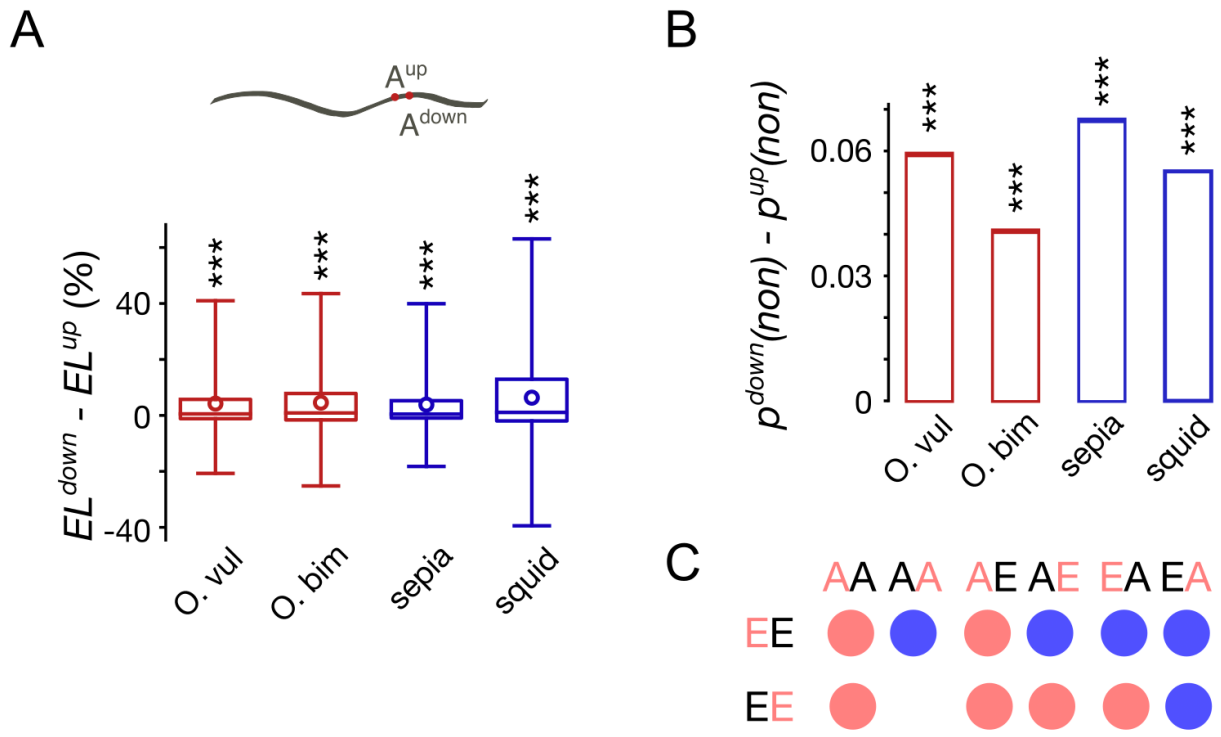


Figure 10 | Directionality of dense clusters. (A) Distributions of the differences in ELs between down- and upstream editing sites in two-adenine (AA) dense clusters. Three asterisks mark statistical significance of the differences in means (p<0.001, the χ^2 contingency test). **(B)** Differences between the probabilities of down- and upstream editing sites to be non-synonymous. Three asterisks mark statistical significance of the differences in means (p<0.001, the binomial test). **(C)** Differences in base-pairing probabilities between paired editing sites (EE) and three types of control AA-dinucleotides (see the text for details). Red color of a letter indicates the nucleotide in a dinucleotide, for which base-pairing probabilities are considered. Red and blue circles show significantly lower and higher base-pairing probabilities for the EE dinucleotide compared to the respective control (the Wilcoxon test p<0.05, Bonferroni corrected).

4.4. Discussion

4.4.1. Cooperativity of RNA editing

A-to-I editing sites in coleoid genomes tend to cluster. The strength of correlations in the editing state co-occurrence clearly depends on the distance between the sites. One explanation is provided by the common secondary RNA structure formed around closely located editing sites. However, the common RNA structures do not explain the inosine co-occurrence observed here (Fig. 6A) and in other studies (Morse, Aruscavage, and Bass 2002; Liscovitch-Brauer et al. 2017; Morse and Bass 1999; Duan et al. 2018; Eisenberg and Levanon 2018; Athanasiadis, Rich, and Maas 2004; Kallman 2003). Indeed, suppose an adenine is edited due to the local RNA structural features. The local structure would generally enhance the probability of editing of adjacent adenines (Gommans, Mullen, and Maas 2009), however, editing at an adjacent site would not depend on the editing at the considered site, unless the RNA structure has changed due to the first act of editing. Thus, no correlations would be observed. This prompts for a dynamic explanation based on changes of editing probabilities near the focal site introduced by editing at that site. We consider the following two scenarios: (i) ADAR enzymatic action at adjacent editing sites is co-operative, manifesting as simultaneous adenine editing dependent on the linear distance between the edited adenines and (ii) inosine produced by editing at one site stabilizes the existing local secondary RNA structure or even causes RNA to fold in a different manner, hence enhancing the probabilities of editing at nearby adenines.

The former explanation presumes that ADAR enzymes can edit multiple sites in a series of enzymatic acts, this ability being dependent on inter-site distances. This is indirectly supported by the fact that different ADAR subunits show enzymatic cooperativity for substrate binding (Valente and Nishikura 2007). Similar effects are observed e.g. in the case of co-operative phosphorylation of adjacent amino acids in proteins, where clusters of phosphorylated residues form due to the enzymatic features of phosphatases (Al-Khouri et al. 2005; Moldovan and Gelfand 2020; Schweiger and Linial

2010). That, however, does not explain the prevailing editing state co-occurrence in the adjacent adenines, as two ADAR subunits may not physically edit two consecutive adenines simultaneously (Stefl et al. 2010). But there may be slippage of the ADAR RNA-binding domain on the RNA sequence, resulting in editing of the adjacent adenine.

In the RNA-centered model, the seeming co-operativity of A-to-I editing of adjacent sites is attributed to the reinforcement of the local secondary RNA structure, which would increase the probabilities of editing at adjacent or closely positioned adenines. Inosines form base-pairs with cytosines, the I-C base pair being isosteric to, but slightly less stable than the G-C pair (Wright, Force, and Znosko 2018). Together with our observation about edited adenines forming frequent A-C mismatches in the local structure (Morse, Aruscavage, and Bass 2002; Morse and Bass 1999; Wong, Sato, and Lazinski 2001; Kallman 2003) (Supplementary Figure S16), this points to a possibility that editing at a focal site changes the local RNA structure pattern, reinforcing the propensity towards stronger secondary structure, and hence promotes editing at adenines in the vicinity. We could not test this explanation computationally due to insufficient data on structural features of inosines (Wright, Force, and Znosko 2018).

The editing of coupled adenines seems to be consistent with the RNA-centered model and follow the scenario involving two factors: dynamics of the sequence context (Duan et al. 2018) and dynamics of the local RNA structure. First, the downstream adenine in a pair is edited due to the upstream adenine being the preferred context (Fig. 5B) and due to the larger accessibility to ADAR as a non-structured element in a secondary structure (Fig. 10C). As a result, the context of the upstream site changes to upstream I instead of A. At that, guanine, an analogue to inosine is the preferred downstream context for editing (Fig. 5B). Along with that, the local RNA structure may be reinforced, specifically due to inosine pairing with cytosine. These two factors may pave the way for editing of the upstream adenine. This scenario implies editing of the upstream adenine to be largely a mechanistic consequence of editing of the downstream adenine. While this scheme may not be true in all cases, we

observe downstream adenines to be more frequently re-coding and hence possibly more frequently selected upon than their upstream counterparts. Thus, in a large number of cases, editing of upstream adenines may indeed be guided by contextual and structural changes induced by the editing at downstream sites. However, this does not explain the phenomenon of directional editing in non-DC clusters, which may be a consequence of specific ADAR activities.

4.4.2. The range of influence of editing sites

Previous studies have established the linear lengths of RNA structures associated with A-to-I mRNA editing to be of various sizes ranging from rather short structures (Savva, Rieder, and Reenan 2012) to complex formations spanning over large fragments of the transcript (Reenan 2005). In coleoid coding sequences, conserved regions around conserved editing sites span on average 100–200 nt in each direction (Liscovitch-Brauer et al. 2017). Accordingly, clustering of edited adenines obtained from the *S* value analysis and the analysis of structurally close edited adenines is observed at up 100–200 nt and up to 256 nt, respectively (Fig. 6A). However, the analysis of adenine editing probabilities in the vicinity of edited sites (Supplementary Figure S12) and the analysis of base-pairing probabilities in the regions around edited adenines (Fig. 8C) have yielded different and consistent estimates of 36 nt and 32–45 nt, respectively. This indicates a hierarchy in the cluster structure, with relatively large, diffuse clusters yielded possibly by weak secondary RNA structures associated with editing sites, which span up to 256 nt (Fig. 10). Smaller, however more stable structures spanning up to 45 nt yield the intermediate level of clustering (Fig. 9). Finally, the local features of RNA structure, e.g. loops, mismatches or bulges, confer the strongest association in terms of editing, which manifests as clusters of adjacent edited adenines (Fig. 6, Fig. 8C).

One important limitation of our and other similar studies is that the existence of introns is largely ignored. Indeed, editing involves unspliced transcripts, whereas one cannot infer the editing state of intronic adenines from the sequenced mRNA data. However, according to the annotation (Albertin et al. 2015), an average adenine in the transcriptome is expected to be located in a 1467 nt

exon, which is at least several fold larger than the distances considered here; hence, our observations should not be affected by the exon-intron structure to a considerable extent. Indeed, the analysis of relaxed long-range structures that considers the longest distances (Fig. 9B) yields the same results when only exons are considered instead of transcripts (Supplementary Figure S19). Nonetheless, the lack of data on the exon-intron structures in coleoids may explain an apparent discrepancy between a typical cluster size and the observations of editing eQTLs (Kurmangaliyev, Ali, and Nuzhdin 2016) and RNA secondary structures (Reenan 2005) spanning thousands of nucleotides. A simpler alternative, of course, is that large-scale statistical studies may not detect rare and long-range effects.

Chapter 5. Phospho-islands and the evolution of phosphorylated amino acids in mammals

This chapter describes our study "Phospho-islands and the evolution of phosphorylated amino acids in mammals" published in PeerJ in 2020. Authors: Moldovan, M., & Gelfand, M. S.

Protein phosphorylation is the best studied post-translational modification strongly influencing protein function. Phosphorylated amino acids not only differ in physico-chemical properties from non-phosphorylated counterparts, but also exhibit different evolutionary patterns, tending to mutate to and originate from negatively charged amino acids. The distribution of phosphosites along protein sequences is non-uniform, as phosphosites tend to cluster, forming so-called phospho-islands.

Here, we have developed a hidden Markov model-based procedure for the identification of phospho-islands and studied the properties of the obtained phosphorylation clusters. To check robustness of evolutionary analysis, we consider different models for the reconstructions of ancestral phosphorylation states.

Clustered phosphosites differ from individual phosphosites in several functional and evolutionary aspects including underrepresentation of phosphotyrosines, higher conservation, more frequent mutations to negatively charged amino acids. The spectrum of tissues, frequencies of specific phosphorylation contexts, and mutational patterns observed near clustered sites also are different.

5.1. Introduction

Protein post-translational modifications (PTMs) are important for a living cell (Schweiger and Linial 2010a; Kurmangaliyev, Goland, and Gelfand 2011; Studer et al. 2016; Huang et al. 2018). By changing physico-chemical properties of proteins, PTMs affect their function, often introducing novel biological features (Pearlman, Serber, and Ferrell 2011). To date, hundreds of thousands of PTMs in various organisms have been identified and various databases containing information about PTMs have been compiled (Ptacek and Snyder 2006; Huang et al. 2018).

Protein phosphorylation is likely both the most common and the best studied PTM (Ptacek and Snyder 2006; Schweiger and Linial 2010; Huang et al. 2018). Phosphorylation introduces a negative charge and a large chemical group to the local protein structure, hence strongly affecting the protein conformation (Pearlman, Serber, and Ferrell 2011; Nishi, Shaytan, and Panchenko 2014). As a result, diverse cellular signaling pathways are based on sequential phosphorylation events (Moses and Landry 2010; Pearlman, Serber, and Ferrell 2011; Ardito et al. 2017). In eukaryotes, phosphorylation sites (phosphosites) are mainly represented by serines, threonines, and tyrosines (which we here refer to as STY amonoacids), with only a minor fraction involving other amino acids, such as histidine (Fuhs and Hunter 2017; Huang et al. 2018).

Phosphosites are overrepresented in intrinsically disordered regions (IDRs) of proteins, i.e. in regions devoid of tertiary structure, usually located on the surface of a protein globule (Iakoucheva 2004). Hence, studies of the evolution of phosphosites have mainly concentrated on sites located in IDRs (Kurmangaliyev, Goland, and Gelfand 2011; Miao et al. 2018). In particular, it has been shown, that phosphosites tend to arise from negatively charged amino acids (NCAs) more frequently than their non-phosphorylated counterparts, and, in a number of cases, retain structural features initially maintained by NCAs (Kurmangaliyev, Goland, and Gelfand 2011; Miao et al. 2018). As phosphorylation is often highly conserved (Macek et al. 2008), experimental limitations on the number of model species with established phosphosites may be overcome in evolutionary studies by formally assigning phosphorylation labels to homologous sites (Kurmangaliyev, Goland, and Gelfand 2011; Huang et al. 2018). However, this approach requires a degree of caution when dealing with evolutionary trees of substantial depths, *e.g.* only a small fraction of yeast phosphosites are conserved between species separated by ~1400 My (million years), while about a half of phosphosites are conserved at a shorter time (~360 My) (Studer et al. 2016). At smaller distances, this method may be applied to infer some evolutionary properties of phosphosites, *e. g.* in *Drosophila* or in vertebrate

species, phosphosites tend to mutate to NCA (Kurmangaliyev, Goland, and Gelfand 2011; Miao et al. 2018).

Phosphorylation can be both a constitutive modification and a way to transiently modify the protein function (Landry et al. 2014). In the former case, the change of a phosphosite to NCA should not cause a significant fitness reduction, as physico-chemical properties are not strongly affected, whereas in the latter case a mutation would have dire consequences (Moses and Landry 2010; Landry et al. 2014).

In proteins, phosphosites often form co-localized groups called phosphorylation islands or phosphorylation clusters, and about a half of phosphorylated serines and threonines are located in such clusters (Schweiger and Linial 2010b). While individual phosphosites function as simple switches, phospho-islands are phosphorylated in a cooperative manner, so that the probability of a phosphorylation event at a focal site strongly depends on the phosphorylation of adjacent sites, and when the number of phosphorylated amino acids exceeds a threshold, the cumulative negative charge of the phosphate groups introduces functionally significant changes to the protein structure (Landry et al. 2014).

Accurate procedures for the identification of phosphosites and next-generation sequencing technologies yielded large numbers of well-annotated phosphosites (Altenhoff et al. 2018; Huang et al. 2018; The UniProt Consortium 2019) enabling us to develop an accurate automatic procedure for the identification of phosphosite clusters we call phospho-islands. We show that clustered phosphosites exhibit evolutionary properties distinct from those of individual phosphosites, in particular, an enhanced mutation rate to NCA and altered mutational patterns of amino acids in the phosphosite vicinity. Our study complements earlier observations on the general evolutionary patterns in phosphosites with the analysis of mutations in non-serine phosphosites and the demonstration of differences in the evolution of clustered and individual phosphorylated residues.

5.2. Methods

5.2.1. Data

The phosphosite data for human, mouse and rat proteomes were downloaded from the iPTMnet database (Huang et al. 2018). The phosphorylation breadth values for the mouse dataset were obtained from Huttlin et al. 2010. Human, mouse and rat proteomes were obtained from the UniProt database (The UniProt Consortium 2019). Vertebrate orthologous gene groups (OGGs) for human and mouse proteomes were downloaded from the OMA database (Altenhoff et al. 2018). Then, all paralogous sequences and all non-mammalian sequences were excluded from the obtained OGGs.

5.2.2. Alignments and Trees

We searched for homologous proteins in three proteomes with pairwise BLASTp alignments (Altschul et al. 1990). Pairs of proteins with highest scores were considered closest homologs. The information about closest homologs was subsequently used to predict phosphosites conserved between human and rat or human and mouse which we hereinafter refer to as HMR phosphosites. OGG were aligned by the ClustalO multiple protein alignment (Sievers et al. 2011) and, while the HMR phosphosites were identified based on Muscle pairwise protein alignments (Edgar 2004). The mammalian phylogenetic tree was obtained from Timetree (Kumar et al. 2017).

5.2.3. Phosphorylation retention upon mutations

After the identification of homologous protein pairs in human/mouse and mouse/rat proteomes and the proteome alignment construction, we identified homologous phosphosites as homologous STY residues which were shown to be phosphorylated in both species. We have shown that phosphorylation is retained on S-T and T-S mutation by comparing two pairs of retention probabilities (Fig. 11C): $p(pS\text{-}pS)$ with $p(pS\text{-}pT \mid S)$ and $p(pT\text{-}pT)$ with $p(pS\text{-}pT \mid T)$ (analogously for the phosphorylation of tyrosines), $p(pX\text{-}pX)$ being defined as the fraction of X amino acids phosphorylated in both considered species:

$$p(pX - pX) = \frac{\#(pX - pX)}{\#(pX - pX) + \#(pX - X)}$$

and $p(pX_1\text{-}pX_2)$, as the fraction of phosphorylated X_1 residues in one species given that in another species another amino acid residue (X_2) is also phosphorylated:

$$p(pX_1 - pX_2|X_1) = \frac{\#(pX_1 - pX_2)}{\#(pX_1 - pX_2) + \#(pX_1 - X_2)}$$

$$p(pX_1 - pX_2|X_2) = \frac{\#(pX_1 - pX_2)}{\#(pX_1 - pX_2) + \#(X_1 - pX_2)}$$

Homologous phosphosite lists from the human/mouse and human/rat pairs were merged to produce HMR phosphosite list of human phosphosites.

5.2.4. False-positive rates of phosphorylation identification by homologous propagation

We assessed the quality of the phosphorylation prediction via homologous propagation approaches by counting false-positive rates of phosphosite predictions in species with large phosphosite lists. As the numbers of predicted phosphosites drastically differed between species (Huang et al. 2018), we considered multiway predictions in each case as characteristics of the procedure performance. Hence, considering mouse phosphosites predicted by homology with known human phosphosites, we also considered human phosphosites predicted based on known mouse phosphosites. The false-positive rate was assessed as the proportion of incorrectly predicted phosphosites among the STY amino acids in one species homologous to phosphosite positions in other considered species.

When assessing the quality of phosphosite predictions based on phosphosites experimentally identified in at least two species, we considered human, mouse and rat and the lists of phosphosites homologous between human and mouse and between human and rat. In these cases, predictions were made for rat and mouse, respectively with the false-positive rate assessed by the same approach as in the previous case.

5.2.5. Mutation matrices

To obtain single amino acid mutation matrices, we first reconstructed ancestral states with the PAML software (Yang 2007). For the reconstruction, we used OGG alignments which did not contain paralogs and pruned mammalian trees retaining only organisms contributing to corresponding OGG alignments. The alignment of both extant and reconstructed ancestral sequences and the corresponding trees were then used to construct mutation matrices, where we distinguished the phosphorylated and non-phosphorylated states of STY amino acids. Here, the phosphorylation state was assigned to STY amino acids using the phosphorylation propagation approach described above. When calculating the mutation matrix, we did not count mutations predicted to happen on branches leading from the root to first-order nodes, as PAML did not reconstruct them well without an outgroup (Koshi and Goldstein 1996; Yang 2007). Tree pruning and calculating the mutation matrix count were implemented in *ad hoc* python scripts using functions from the ete3 python module.

5.2.6. Disordered regions and identification of phospho-islands

Intrinsically disordered protein regions (IDRs) are defined here, following (Xue et al. 2010), as regions of proteins lacking stable and well-defined three-dimensional structure. IDRs were predicted with the PONDR VSL2 software with default parameters (Xue et al. 2010). This algorithm was selected, firstly, as one of the best IDR predictors yielding results highly consistent with other top-IDR predictors (Peng and Kurgan 2012; Zhou et al. 2020), and, secondly, as the one efficiently predicting long IDRs, (Peng and Kurgan 2012), which is essential for the present study.

Phosphorylated amino acids were divided into those located in predicted IDRs and those located in ordered regions (ORs). On the HMR set construction, phosphosites with conflicting IDR/OR labels were excluded from the analysis. In the analyses of separate IDR/OR mutations, we considered IDR and OR labels of amino acids to be conserved along the mammalian tree and hence inferred the remaining extant and ancestral IDR/OR states from homology with both mouse and human ORs and IDRs. We consider the premise of conserved mammalian IDRs justified here, as it is known that protein tertiary structure elements, including IDRs, are evolving slowly (Chen et al. 2006; Toth-Petroczy et al. 2008).

Phospho-islands were identified by a hidden Markov model (HMM) built upon the distributions of distances between clustered and individual phosphosites. For that, the most likely clustered/individual phosphosite assignments were obtained with the Viterbi algorithm that is guaranteed to maximize the posterior probability (Viterbi 1967). The emission probabilities for the HMM were obtained as the ratio of density values in the decomposition of the distribution of amino acid distances between adjacent phosphosites in IDRs, *S* (the likelihood ratio normalized to 1) (Fig. 12a). To select the optimal transitional probability values, we performed a stability check by analyzing the dependence of the fraction of clustered phosphosites on the transitional probability values (Suppl. Fig. S21). The percentages of clustered phosphosites turned out to be extremely stable with respect to transitional probability values if the latter were smaller than 0.3. Hence, the transitional probabilities were set to 0.2 (Fig. 12b).

5.2.7. Phosphosite contexts

We employed the list of phosphosite contexts as well as the binary decision-tree procedure to define the context of a given phosphosite from Villén et al. 2007. The procedure is as follows. (i) Proline context is assigned if there is a proline at position +1 relative to the phosphosite. (ii) Acidic context is assigned if there are five or six E/D amino acids at positions +1 to +6 relative to the phosphosite. (iii) Basic context is assigned if there is a R/K amino acid at position –3. (iv) Acidic

context is assigned if there are D/E amino acids at any of positions +1, +2 or +3. (v) Basic context is assigned if there are at least two R/K amino acids at positions –6 to –1. Otherwise, no context is assigned and we denote this as the "O" (other) context. We consider tyrosine phosphosites separately and formally assign the with the "Y" (tyrosine) context.

5.2.8. Local mutation matrices

We computed local substitution matrices (LSMs) as the substitution matrices for amino acids located within a frame with the radius *k* centered at a phosphorylated serine or threonine. When computing LSMs, we did not count mutations of or resulting in STY amino acids to exclude the effects introduced by the presence and abundance of phospho-islands. We have set *k* to 1, 3, 5, and 7 and selected 3 as for this value we observed the strongest effect, that is, obtained the largest number of mutations with frequencies statistically different from those for non-phosphorylated serines and threonines.

5.2.9. Statistics

When comparing frequencies, we used the χ^2 test if all values in the contingency matrix exceeded 20 and Fisher's exact test otherwise. To correct for multiple testing, we used the Bonferroni correction (Bonferroni 1936) with the scaling factor set to 17 for the substitution vector comparison and to 17×17 for the comparison of substitution matrices with excluded STY amino acids. 95% two-tailed confidence intervals shown in figures were computed by the χ^2 or Fisher's exact test. The significance of obtained Pearson's correlation coefficients was assessed with the F-statistic.

5.2.10. Code availability

Ad hoc scripts were written in Python. Graphs were built using R. All scripts and data analysis protocols are available online at https://github.com/mikemoldovan/phosphosites.

5.3. Results

5.3.1. Conserved phosphosites

As protein phosphorylation in a vast majority of organisms has not been studied or has been studied rather poorly (Huang et al. 2018), the evolutionary analyses of phosphosites typically rely on the assumption of absolute conservation of the phosphorylation label assigned to STY amino acids on a considered tree (Kurmangaliyev, Goland, and Gelfand 2011; Miao et al. 2018). Thus, if, for instance, a serine is phosphorylated in human, we, following this approach, would consider any mutation in the homologous position of the type S-to-X to be a mutation of a phosphorylated serine to amino acid X (Fig. 11B). However, the comprehensive analysis of yeast phosphosites has shown low conservation of the phosphorylation label at the timescales of the order 100 My and more (Studer et al. 2016). Thus, we have considered only orthologous groups of mammalian proteins, present in the OMA database (Altenhoff et al. 2018). The mammalian phylogenetic tree is about 177 My deep (Kumar et al. 2017), which corresponds to about 50% of the phosphorylation loss in the 182 My-deep yeast *Saccharomyces-Lachancea* evolutionary path (Studer et al. 2016). The tree contains three organisms with well-studied phosphoproteomes: human (227834 sites), mouse (92943 sites), and rat (24466 sites) (Huang et al. 2018) (Fig. 11A).

Still, the expected 50% of mispredicted phosphosites could render an accurate evolutionary analysis impossible. This could be partially offset by considering phosphosites conserved in well-studied lineages. Thus, we compiled a set of human phosphosites homologous to residues phosphorylated also in mouse and/or rat, which we will further refer to as human-mouse/rat (HMR) phosphosites. The HMR set consists of 53437 sites covering 54.6% and 61.2% of known mouse and rat phosphosites, respectively, which is consistent with the above-mentioned observation about 50% phosphorylation loss in yeast on evolutionary distances similar to the ones between the human and rodent lineages (Fig. 11ab).

We consider the sites predicted by homology with the HMR set to be enriched in accurately identified phosphosites, as retaining only conserved phosphosites we substantially reduce the number of mispredictions. If we simply propagated human phosphorylation labels to mouse and *vice versa* we would get about 77.6% and 42.3% of false positive labels, respectively. However, sites conserved between human and rat or sites conserved between rat and mouse would yield about twofold lesser percentages of 41.9% and 19.9% of false positives in mouse and human, respectively. The obtained percentages can be considered as upper estimates of false positive rates, as current experimental phosphosite coverage in mammals cannot guarantee the identification of all conserved phosphosites (Huang et al. 2018). Thus, the HMR dataset is sufficiently robust for the prediction of phosphorylation labels in less-studied mammalian lineages.

Treatment of STY amino acids homologous to phosphorylated ones as phosphorylated yields another possible caveat, stemming from the possible loss of phosphorylation upon STY-to-STY mutations. To assess this effect, we compared the probabilities of phosphosite retention upon pSTY-to-STY mutation, pSTY indicating the phosphorylated state, and the respective probabilities in the situation when a mutation has not occurred for a pair of species with well-established phosphosite lists, i.e. human and mouse (Fig. 11C). We have observed only a minor, insignificant decrease of the probabilities of the phosphorylation retention in the cases of pS-pT and pS-pY mismatches relative to the pT-pT states in mouse and human, indicating the general conservation of the phosphorylation label upon amino acid substitution. An interesting observation here is that the pS-pS states appear to be the most conserved ones (Fig. 11C). Taken together, these results indicate the evolutionary stability of phosphorylation states upon mutation.

The increased evolutionary robustness of the pS state relative to the pT and pY states should manifest as overrepresentation of phosphoserines among phosphosites with respect to non-phosphorylated amino acid positions. Thus, we assessed the relative abundancies of pSTY amino acids in the HMR dataset relative to the established human phosphosite set and to the set of non-

phosphorylated STY amino acids. Serines and threonines, comprising the vast majority of the pSTY amino acids, are, respectively, over- and underrepresented in the phosphosite sets (Fig. 11CD). This effect is significantly more pronounced in the HMR dataset relative to the total human phosphosite dataset, further supporting the observation about lower conservation of pT relative to pS, as the HMR dataset is enriched in conserved phosphosites by design.

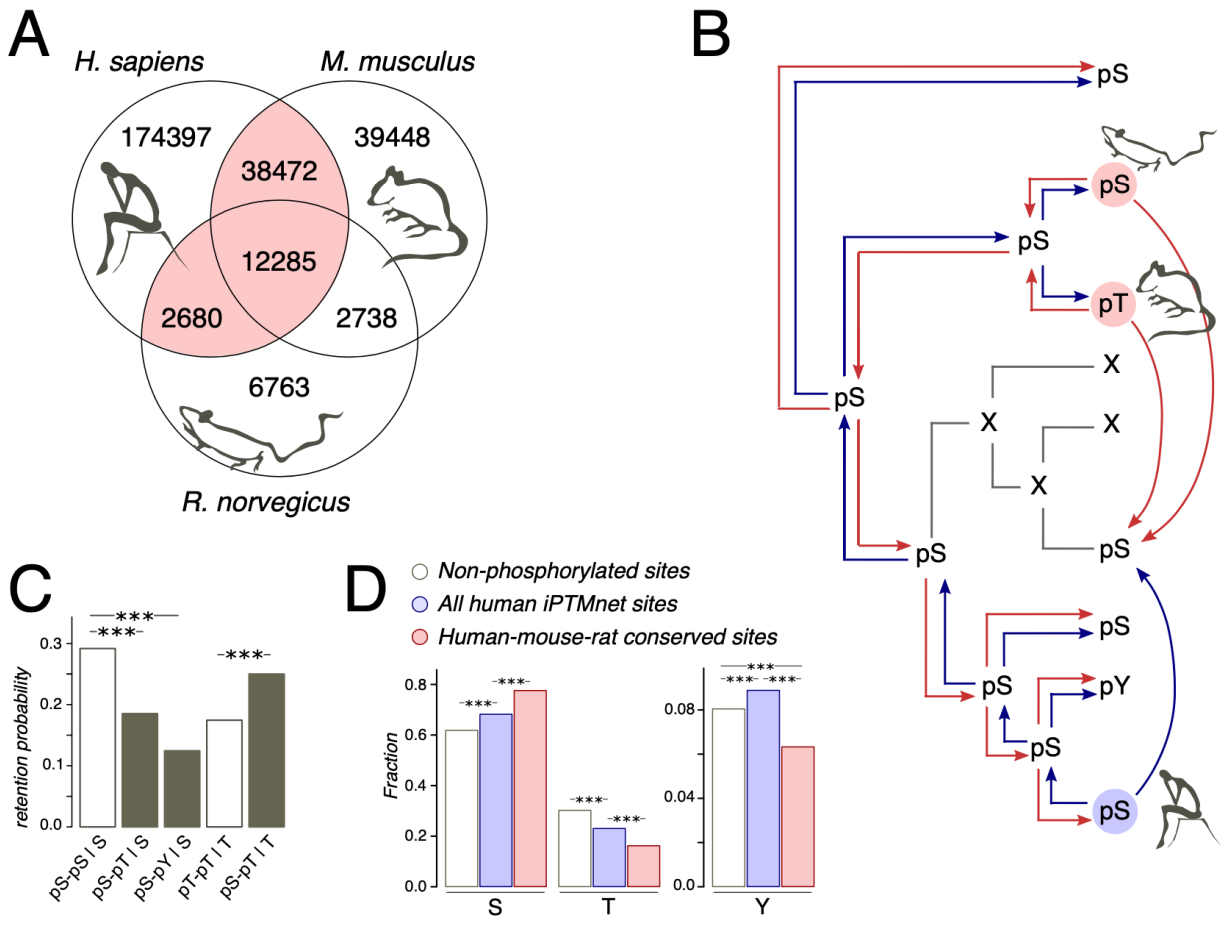


Figure 11 | Phosphosites considered in the study. (A) Venn diagram of iPTMnet HMR phosphosites. Intersections correspond to conserved phosphosites. The HMR phosphosite dataset is shown in pink. **(B)** Phosphosite assignment procedures. Given a tree of a mammalian orthologous gene group and a column in the respective alignment, we assign phosphorylation labels to ancestral and extant amino acids, firstly, by propagating labels from one species to all other species in the tree (shown as separate red and blue arrows) and, secondly, by propagating labels predicted both in the selected species (e.g., human, as shown) and in one of the remaining species (mouse and rat); this corresponds to blue and

red arrows entering a given node in the tree. Phosphosites obtained by the latter procedure are referred to as the HMR phosphosite dataset. In both procedures, phosphorylation is considered to be retained both for direct and indirect STY-to-STY mutations. **(C)** Retention of phosphorylation upon mutation. Bars represent the probability of a conserved modification for the human dataset in the case of mutation and if mutation has not occurred. The letter after the vertical bar is an amino acid over which the probability was normalized. Three asterisks represent $p < 0.001$ ($\chi 2$ test). **(D)** STY amino acid content of three groups of phosphosite datasets.

5.3.2. Phosphorylation islands

The distribution of distances between phosphosites is different from that of randomly chosen serines and threonines even accounting for the tendency of phosphosites to occur in intrinsiacally disordered regions (IDRs) (Schweiger and Linial 2010) (Fig. 12A). However, this observation depends on an arbitrary definition of phosphorylation islands as groups of phosphosites separated by at most four amino acids (Schweiger and Linial 2010b). We have developed an approach that reduces the degree of arbitrariness in the definition of phospho-islands which is based on a statistical model of the distances between phosphorylated residues in phosphor-islands and for individual phosphosites.

We consider only phospho-islands located in IDRs, due to two reasons. First, IDR phosphosites, being more abundant, yield reliable statistics. Second, ordered regions are largely non-uniform in terms of local structural features, e.g being localized in the protein hydrophobic core or at the surface (van der Lee et al. 2014). This would render construction of the null model of between-phosphosite distances impossible without considering all protein structures of the mammalian proteome, which is currently not feasible. Hence, we will hereinafter refer to phospho-islands located in IDRs simply as phospho-islands and to non-clustered phosphosites located in IDRs as individual phosphosites.

Let S be the distribution of amino acid distances between adjacent phosphosites in IDRs. The logarithm of S is not unimodal (Fig. 12A), and we suggest that it is a superposition of two distributions:

one generated by phosphosites in phospho-islands and the other reflecting phosphosites outside phospho-islands (left and right peaks, respectively). The latter distribution can be obtained from random sampling from IDRs of non-phosphorylated STY amino acids while preserving the amino acid composition and the sample size, as we expect individual phosphosites to emerge independently while maintaining the preference towards IDRs (Fig. 12C). Gamma distribution has a good continuous fit to log(S+1) for randomly sampled STY amino acids located in IDRs. Given its universality and low number of parameters (Friedman, Cai, and Xie 2006; Reiss, Facciotti, and Baliga 2008; Mendoza-Parra et al. 2013), we have selected gamma distribution as a reasonable model for log(S+1) (Fig. 12C). Assuming that the distribution of log(S+1) values for phosphosites located in phospho-islands should belong to the same family and fixing the parameters of the previously obtained distribution, we decomposed the distribution of log(S+1) values into the weighted sum of two gamma distributions, one of which corresponds to STYs located in phospho-islands and the other one, to remaining STYs in IDRs (Fig 12A, red and grey curves, respectively). From these two gamma distributions we obtained parameters for a hidden Markov model, which, in turn, was used to map phosphorylation islands. The distributions of S values for phosphosites in identified islands and the distribution for other phosphosites yielded a good match to the expected ones (Fig. 12BD).

Both for the HMR and mouse datasets, more than half of phosphosites are located in phospho-islands (61% and 56%, respectively) (Fig. 12E, Suppl. Fig. S22AB). For human phosphosites, however, we see a larger proportion of sites (53%) located outside phospho-islands. In the latter case the distributions in the decomposition differ less, compared to the former two cases (Fig. 12A, Suppl. Fig. S22). It could be caused by a larger density of phosphosites in IDRs of the human proteome, resulting from higher experimental coverage; that would lead to generally lower S values, which, in turn, could cause the right peak in the log(S+1) distribution to merge with the left peak, rendering the underlying gamma-distributions less distinguishable. To validate this explanation, we randomly sampled 40% of human phosphosites, so that the sample size matched the one for mouse phosphosites;

however, the results on this rarefied dataset did not change (Fig. 12E, Suppl. Fig. S22C) indicating that our procedure is robust with respect to phosphosite sample sizes. Hence, phospho-islands for the human dataset are identified with a lower accuracy than those for the HMR and mouse datasets. This could be caused by different experimental technique applied to the human phosphosites, compared to the one used for mouse and rat phosphosites, and by a possibly large number of false-positive phosphosites in the former case (Huttlin et al. 2010; Bekker-Jensen et al. 2017; Xu et al. 2017) (see 5.4. Discussion).

In phospho-islands, the overall pSTY-amino acid composition differs from that of individual phosphosites, mainly because the fraction of threonines is significantly higher in phospho-islands at the expense of the lower fraction of tyrosines (Fig. 12F). Also, the conservation of residues in phospho-islands is larger than that of the individual sites (Fig. 12G). Overall, the general properties of clustered phosphosites seem to differ from those of individual phosphosites.

A similar attempt to decompose the S distribution for phosphosites located in ordered regions yielded the distribution of $\log(S+1)$ values highly skewed to the left (small distances), even relative to the distribution of $\log(S+1)$ values in phospho-islands in IDRs (Suppl. Fig. S22E). This precluded decomposition of the S distribution into a weighted sum of two distributions. A more complex model possibly incorporating features of the tertiary protein structure might be required to infer and analyze phospho-islands located in ORs, which is beyond the scope of the present study.

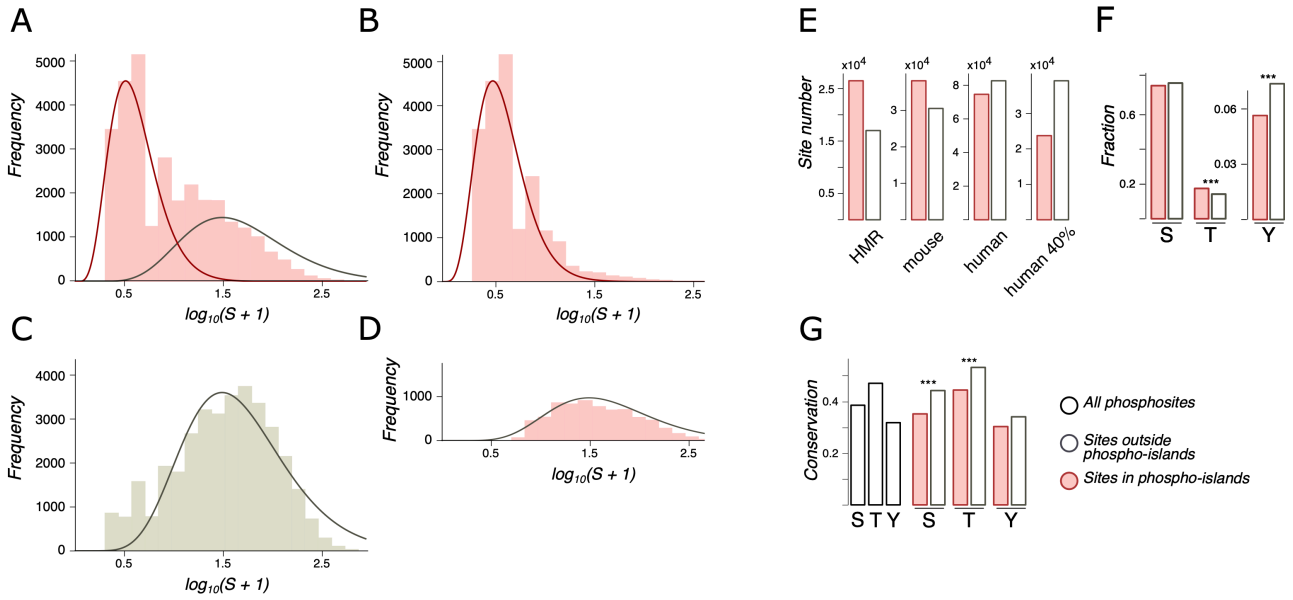


Figure 12 | Phospho-islands for the HMR phosphosite dataset. (A) The distribution of log10(S+1) values (pink histogram) and its decomposition in two gamma distributions: the one for phospho-islands (red curve) and for individual phosphosites (red curve). **(B)** The distribution of log10(S+1) values for phosphosites predicted to be in phospho-islands. **(C)** log10(S+1) values for non-phosphorylated STY amino acids randomly sampled from DRs with the same sample size and amino acid content as in the HMR dataset. **(D)** log10(S+1) values for predicted individual phosphosites. **(E)** Numbers of individual phosphosites and sites in phospho-islands for four datasets. **(F)** Amino acid content of phospho-islands and individual phosphosites. **(G)** Frequency of mutations for phosphosites and individual amino acids. Asterisks depict significantly different values ($p < 0.001$, $\chi 2$ test).

5.3.3. Mutational patterns of phosphorylated amino acids

Next, we have reconstructed the ancestral states for all mammalian orthologous protein groups not containing paralogs and calculated the proportions of mutations $P(X_1 \rightarrow X_2)$, where X_1 and X_2 are different amino acids. We treated phosphorylated and non-phosphorylated states of STY amino acids as distinct states. We then introduced a measure of difference in mutation rates for phosphorylated STY and their non-phosphorylated counterparts. For a mutation of an STY amino acid X to a non-STY amino acid Z we define $R(X, Z) = P(pX \rightarrow Z) / P(X \rightarrow Z)$. If X^* is another STY amino acid,

$R(X, X^*) = P(pX \rightarrow pX^*) / P(X \rightarrow X^*)$. Thus, the *R* value for a given type of mutations is the proportion of the considered mutation of a phosphorylated STY amino acid among other mutations normalized by the fraction of respective mutations of the non-phosphorylated STY counterpart. The *R* values are thus not affected by differences in the mutation rates between phosphorylated and non-phosphorylated amino acids, as all probabilities are implicitly normalized by the mutation rates of *pX* and *X*.

We firstly consider phosphosites located in IDRs. For phosphorserines from the HMR dataset we confirm earlier observations: phosphoserines mutate to NCA more frequently than non-phosphorylated serines (Fig. 13A). The *R* values for serine mutation to aspartate, *R*(S,D), and glutamate, *R*(S,E), are both significantly larger than 1 (1.2, $p<0.01$ and 1.7, $p<0.001$, respectively; χ^2 test) and, interestingly, they differ substantially ($p<0.001$, multiple random Poisson sampling test). Similarly, asparagine and glutamine *R* values differ, with *R*(S,N)=0.9 ($p<0.001$, χ^2 test), significantly lower than 1, and *R*(S,Q)=1.4 ($p<0.001$, χ^2 test), significantly higher than 1. The rate of mutation to lysine significantly differs for phosphorylated and non-phosphorylated serines ($p<0.001$, χ^2 test). Interestingly, the mutation rate to another positively charged amino acid, arginine, is significantly lower than expected ($p<0.01$, χ^2 test). For non-polar amino acids generally no significant differences in the *R* values between phosphorylated and non-phosporylated serines are observed, but for methionine and proline, the calculated values are significant: *R*(S,M)>1 ($p<0.001$, χ^2 test) and *R*(S,P)<1 ($p<0.01$, χ^2 test).

In earlier studies, only mutations of serines or to serines had been considered, as the available data did not allow for statistically significant results for threonine and tyrosine (Kurmangaliyev, Goland, and Gelfand 2011; Miao et al. 2018). Here, we see that phosphorylated threonines from the HMR dataset tend to mutate to serines (Fig. 13B). At that, phosphorylated serines mutate to threonines more frequently than their non-phosphorylated counterparts for all considered samples, i.e. for the human, mouse and HMR sets (Fig. 13B, Suppl. Figs. S23-S28). Phosphorylated tyrosines tend to avoid mutations to isoleucine ($p<0.05$, χ^2 test) and, for human samples, to arginine ($p<0.05$, χ^2 test) and

glycine ($p<0.001$, χ^2 test) (Fig. 13B, Suppl. Figs. S23-S28). Phospho-tyrosines in the mouse dataset show a weaker tendency for the avoidance of the mutations to aspartate than the non-phosphorylated ones ($p<0.05$, χ^2 test) while the rate of pY-to-I mutations is higher (Fig. 13B).

Separate analysis of mutations in phospho-islands and in individual phosphosites yields three observations. Firstly, alterations of mutation patterns of phosphoserines and phosphothreonines (pST) in IDRs relative to non-phosphorylated ST in IDRs are similar to the patterns observed for the clustered pST and, to a lesser extent, to those observed for individual pSTs (Fig. 13B). This is mostly due to the fact that the mutational patterns of clustered pSTs generally differ from those of their non-phosphorylated counterparts to a greater extent than the mutational patterns of individual phosphoserines do (Fig. 13B, Suppl. Fig. S21). Secondly, for phosphotyrosines, alterations in their mutational patterns brought about by phosphorylation are mostly explained by individual phosphotyrosines. The mutational patterns of individual sites deviate from the ones observed for non-phosphorylated tyrosines more than those of clustered phosphotyrosines (Fig. 13B, Suppl. Figs. S23-S28). Also, if we compare the *R* values calculated for all possible mutations in clustered *vs*. individual phosphosites, the *R* value corresponding to the S-to-E mutation will be significantly higher for the set of clustered phosphosites ($p=0.009$, χ^2 test, Suppl. Fig. S30). Hence, we posit that the general phosphosite mutational pattern alterations can be explained mostly by mutations in clustered phosphosites for phosphoserines and phosphothreonines and by individual sites when phosphotyrosines are considered.

We also studied mutation patterns in ordered regions (ORs), and observed that phosphothreonines located in ORs demonstrate higher T-to-S mutation rates (Fig. 13B) relative to those of non-phosphorylated threonines located in ORs. Also, sites located in ORs demonstrate enhanced S-to-T and Y-to-T mutation rates relative to non-phosphorylated serines and threonines in ORs, respectively (Fig. 13B).

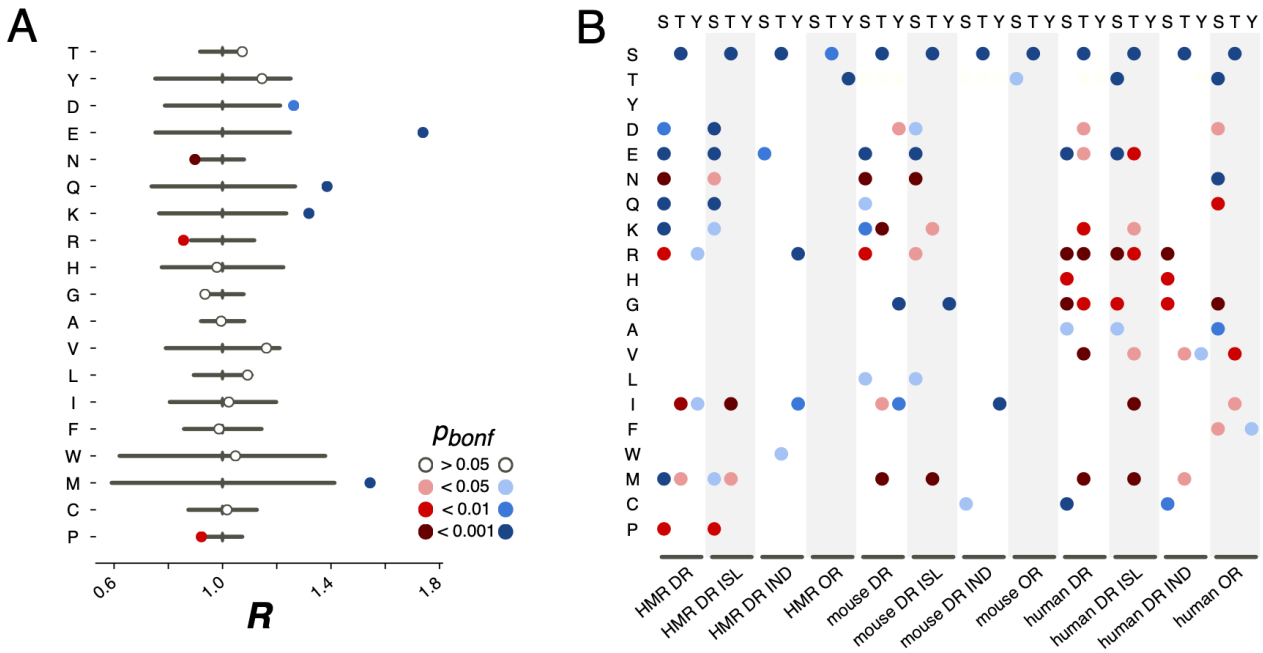


Figure 13 | pX0→X1 substitution vectors. (A) R values of the pS→X substitutions for serines from the HMR dataset located in DRs. **(B)** Substitution probabilities for phosphorylated STY amino acids significantly different from those for non-phosphorylated STY amino acids for several datasets. The significance levels are shown with the colors introduced in the panel in (A). Abbreviations on the horizontal axis: ISL, phosphosites located in phospho-islands; IND, individual phosphosites; DR, phosphosites from disordered regions; OR, phosphosites from ordered regions. On both panels, blue and red dots represent statistically significant over- and underrepresentation of pSTY-to-X mutations relative to non-phosphorylated STY, respectively. Darker shading represents higher statistical significance.

5.3.4. Phosphosite contexts

Sequence contexts of phosphosites generally fall into three categories: acidic (A), basic (B), and proline (P) motifs, with tyrosine phosphosites comprising a special class (Y) (Villén et al. 2007; Huttlin et al. 2010). For each phosphosite from each dataset we have identified its context. As in previous studies (Villén et al. 2007; Huttlin et al. 2010), phosphosites not assigned with any of these

context classes were considered as having “other” (O) motif. We studied the distribution of these motifs for all classes of phosphosites.

In IDRs, relative to ORs, we observed a higher percentage of phosphosites with assigned contexts (Fig. 14A). P-phosphosites demonstrate the highest overrepresentation in IDRs, with 25% of IDR phosphosites having the proline motif. Phospho-islands, compared to individual phosphosites, contain more phosphosites with assigned motifs relative to individual phosphosites. In IDRs, there are more B- and P-phosphosites and fewer A-phosphosites among clustered sites than among individual ones. Notably, the fraction of phospho-tyrosines is substantially higher in ordered regions. However, this effect could be at least partially explained by the general tendency of aromatic residues, including tyrosine, to occur in ordered protein regions (Receveur-Bréchot et al. 2005).

5.3.5. Phosphorylation breadth

An important feature of a phosphosite is its “phosphorylation breadth”, that is, the number of tissues where it is phosphorylated. In this study, the maximal phosphorylation breadth is nine, as the phosphorylation data for nine mouse tissues are available (Huttlin et al. 2010). Among broadly expressed phosphosites (present in all nine tissues), compared to tissue-specific ones (present in only one tissue), very few sites have unassigned contexts (O) and almost none are tyrosine phosphosites. The fraction of acidic phosphosites (24%) is substantially lower among tissue-specific sites relative to broadly phosphorylated ones (37%) ($p<0.001$, χ^2 test) (Fig. 14A).

As mentioned above, the pS-to-E mutation yields the highest value, R(S,E) (Fig. 13A) and represents the only mutation with significantly different R values in phospho-islands and individual sites ($p=0.009$, χ^2 test, Suppl. Fig. S21). At that, R(S,E) significantly increase with increasing breadth of expression (Fig. 14B), from R(S,E)=1.14 for tissue-specific phosphosites to R(S,E)=6.64 for broadly expressed phosphosites ($p=0.016$, t-test).

Finally, we compared percentages of phosphosites with different breadths in ORs *vs.* IDRs and in phospho-islands *vs.* individual phosphosites (Fig. 14CD). As the phosphorylation breadth increases,

so does the fraction of clustered phosphosites, reaching 85% for sites phosphorylated in nine tissues; the fraction of phosphosites in IDRs also increases, reaching 95.4%.

Hence, broadly expressed phosphosites have well-defined motifs, tend towards disordered regions and to phospho-islands, have mostly acidic context, and mutate to NCA more frequently than tissue-specific phosphosites.

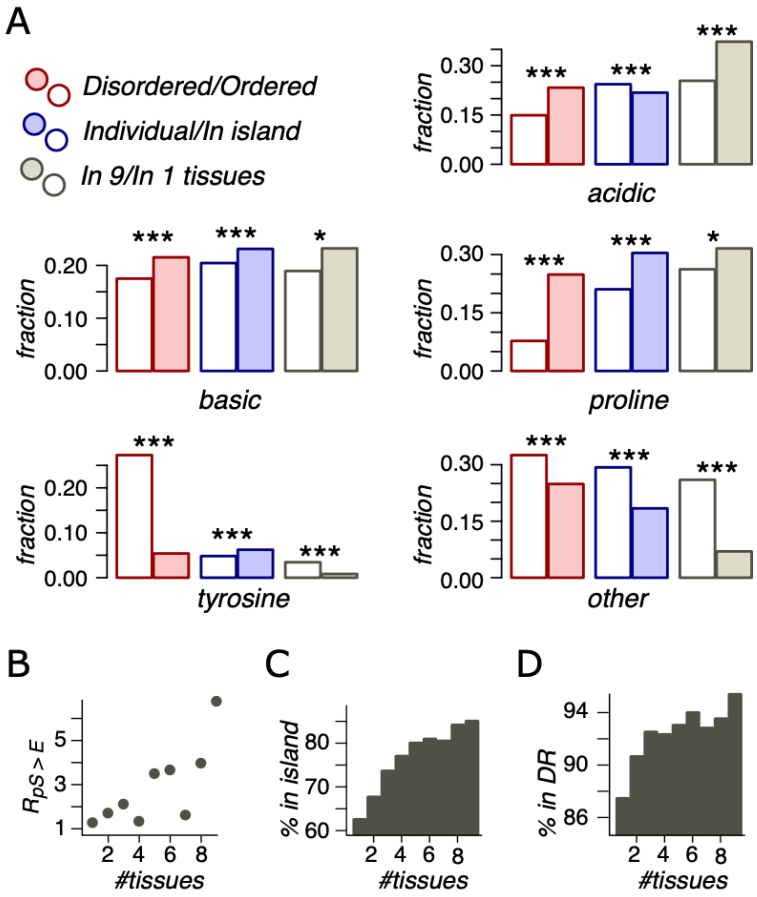


Figure 14 | Phosphosite contexts and phosphorylation breadth. (A) Overrepresentation of phosphosite contexts in ordered vs. disoredred regions, in phospho-islands vs. individual phosphosites and for broadly vs. narrowly distributed phosphorylated amino acids. One asterisk and three asterisks indicate statistical significance at the levels of 0.05 and 0.001 respectively (χ2 test). **(B)** The dependance of R(pS,E) on the phosphosite breadth. Pearson's r2 is equal to 0.53 with the t-test p = 0.016. **(C)** The dependance of phosphosite fraction in phospho-islands on the phosphorylation breadth

(p = 9*10−41, χ2 test). **(D)** Percent of phosphosites in disordered regions vs. phosphosite breadth (p = 4.1*10−10, χ2 test).

5.3.6. Mutation patterns in the proximity of phosphosites

We now show that not only phoshosites require special motifs (Huttlin et al. 2010), but the mutational context of clustered phosphosites differs from that of individual sites. To assess evolutionary dynamics associated with phosphosite motifs, we analyzed mutational patterns in ±3 amino acid windows of HMR ST phosphosites located in IDRs and compared them with those of non-phosphorylated ST amino acids from IDRs. The ±3 window was selected, as it yielded the strongest effect in terms of the number of mutations with rates statistically distinct from the expected ones (Suppl. Fig. S31AB). We did not consider phosphotyrosines, as they have not been shown to possess any discernible general motif apart from the phosphorylated tyrosine itself (Huttlin et al. 2010).

We introduce the measure Q defined as $Q\left(X_1^p \rightarrow X_2^p\right) = P(X_1^p \rightarrow X_2^p)/P(X_1^n \rightarrow X_2^n)$, where X_1^p and X_2^p are amino acids near phosphorylated serines and threonines and X_1^n and X_2^n are amino acids near non-phosphorylated serines and threonines. Q measures overrepresentation of a given mutation in the proximity of pST amino acids relative to ST amino acids. We also considered sites located in phospho-islands and individual phosphosites separately (Fig. 15, Suppl. Fig. S31CD).

In the whole HMR dataset, 22 types of non-phosphorylated amino acid substitutions out of the total of 289 have Q values statistically different from the expected value 1 (p<0.05, χ^2 test with the Bonferroni correction), among them three pairs of mutually reverse mutations (Fig. 15). As expected from the conservation of the phosphosite contexts, mutations between positively and negatively charged amino acids, potentially changing acidic to basic contexts and *vice versa*, are underrepresented, whereas E-to-D, D-to-E and K-to-R, not changing the context type, are overrepresented. The P-to-A substitution is overrepresented, thus indicating the instability of proline contexts. Interestingly, all three mutations with Q values exceeding 2.5 involve lysine, two of them being reverse mutations F-to-K and K-to-F. The fourth most overrepresented mutation, Y-to-G with

Q(Y→G)=2.5, could explain the lack of tyrosine phosphosites in IDRs, as a large fraction of IDR phosphosites are clustered with the distances between sites not exceeding three amino acids. Thus, a large Q(Y→G) value would lead to general underrepresentation of tyrosines in IDRs.

Types of mutations with significant Q values generally differ near clustered and individual phosphosites (Suppl. Fig. S31CD). E-to-D, not changing the local acidic context type (Huttlin et al. 2010), is overrepresented and E-to-K, disrupting the acidic context (Huttlin et al. 2010), is underrepresented in both cases. On the other hand, around individual phosphosites, Q(F→K)=3.4 and Q(P→A)=1.12, indicating an enhanced birth rate of the basic context and disruption of the proline context, respectively. The R-to-D mutation, disrupting the local basic context, also is overrepresented near individual phosphosites. In general, among seven overrepresented mutations near clustered phosphosites, only the K-to-P mutation disrupts the local basic context in favor of the proline context and among seven overrepresented mutations near individual phosphosites, three mutations (E-to-F, R-to-D, and P-to-A) could be regarded as context-disrupting. Hence, the individual phosphosite contexts are somewhat less evolutionary stable and thus the lower percentage of individual phosphosites with identifiable contexts might be due to specific local context-disrupting mutation patterns for these phosphosites.

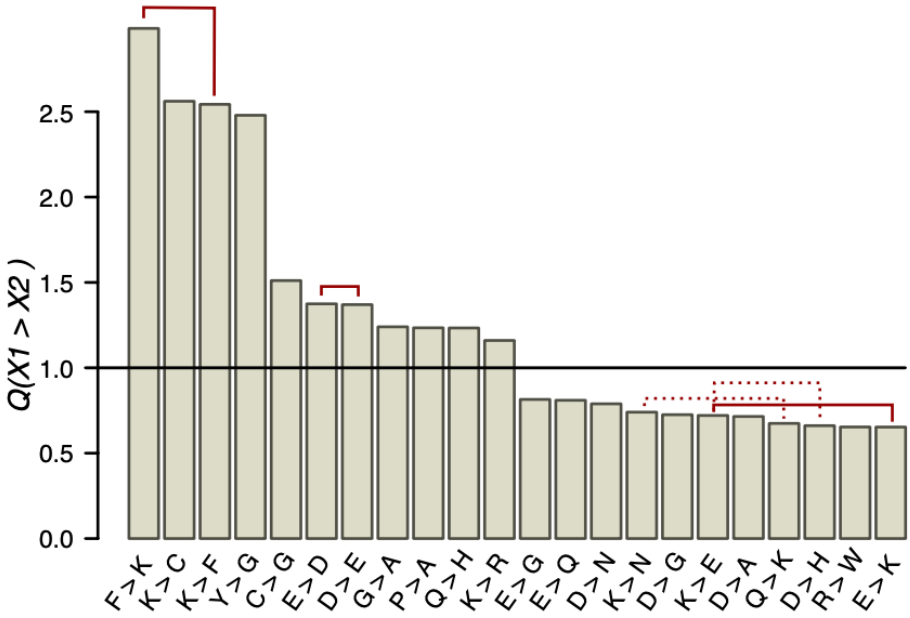


Figure 15 | Q values of mutations near ST phosphosites with probabilities significantly different from the expected ones. Solid red lines connect mutually reverse mutations. Dashed lines indicate quazy-reverse mutations of amino acids with common chemical properties.

5.4. Discussion

5.4.1. Clustered vs. individual phosphosites

We have demonstrated that clustered phosphosites differ from non-clustered ones in a number of aspects: (i) overrepresentation of phosphothreonines and underrepresentation of phosphotyrosines in phospho-islands (Fig. 12F); (ii) stronger conservation of clustered phosphoserines and phosphothreonines (Fig. 12G); (iii) larger proportion of sites phosphorylated in many tissues (Fig. 14C); (iv) significantly larger probability of mutations to glutamate for clustered relative to the individual phosphoserines; (v) larger fraction of sites with specific motifs in phospho-islands (Fig. 14A); (vi) mutational patterns in the proximity of phosphosites consistent with the context-retention hypothesis (Fig. 15). What are possible explanations for the observed effects?

Underrepresentation of phosphotyrosines in phospho-islands could be explained by phosphorylation of clustered phosphosites being co-operative. As serines and threonines are more

similar to each other in their tendency to being phosphorylated by similar enzymes than they are to tyrosine (Villén et al. 2007; Huttlin et al. 2010; Landry et al. 2014; Studer et al. 2016), one would expect phospho-tyrosines to disrupt co-operative phosphorylation of adjacent ST amino acids by being phosphorylated independently, thus introducing a negative charge which would affect phosphorylation probabilities of the neighbouring amino acids (Landry et al. 2014). Hence phospho-tyrosines could have been purged by selection from pST clusters.

Secondly, phosphosites located in phospho-islands are more conserved than individual ones (Fig. 12G), as opposed to an earlier hypothesis that individual phosphosites are more conserved than their clustered counterparts (Landry et al. 2014). Our result seems to contradict the notion that the cellular function of phosphosites in an island depends on the number of phosphorylated residues rather than specific phosphorylated sites, whereas individual phosphosites operate as single-site switches and hence should be more conserved (Landry et al. 2014). However, this argument implies that phosphorylation of most individual phosphosites is important for the organism's fitness, which may be not true (Landry et al. 2014; Miao et al. 2018) and hence our results do not contradict the model of evolution of functionally important phosphosites.

Overrepresentation of phosphosites with defined motifs among the clustered ones (Fig. 14A) and reduced numbers of mutations disrupting the local contexts of the clustered sites (Suppl. Fig. S30CD) may indicate enhanced selective pressure on clustered phosphosites and their contexts. An indirect support of this claim comes from the overrepresentation of ubiquitously phosphorylated sites among the clustered ones (Fig. 14C). Indeed, broad phosphorylation requires a stronger local context and indicates the reduced probability of a phosphosite being detected simply due to the noise inherent to the phosphorylation machinery (Landry et al. 2014).

Mutations of phosphoserines located in IDRs to NCA are generally overrepresented among all mutations of the type pS-to-X relative to the corresponding mutations of non-phosphorylated serines (Fig. 13B). This effect is stronger for clustered phosphosites and for ubiquitously phosphorylated sites.

Together with the observation about clustered phosphosites being on average more broadly phosphorylated than the individual ones, this suggests that a large fraction of phosphosite clusters might be phosphorylated (nearly) constitutively, and thus changes of individual phospho-serines to NCAs could experience lesser degrees of negative selection acting upon the corresponding mutations, as these mutations introduce smaller degrees of local electric charge shifts on the protein globule than the mutations of non-phosphorylated serines to NCAs do.

5.4.2. Two types of mutations

In all considered phosphosite datasets, we have observed two types of pSTY-to-X mutations overrepresented relative to STY-to-X mutations (Fig. 13B): (i) pSTY-to-pSTY, especially pT-to-pS mutation and (ii) pSTY-to-NCA, especially pS-to-E mutations. The former effect could be explained by the relaxed selection against pST-to-pST mutations due to the phosphorylation machinery often not distinguishing between serines and threonines (Huttlin et al. 2010; Miao et al. 2018). The overrepresentation of pT-to-pS mutation for all datasets, including sites located in ORs, could stem from the higher probability of phosphosite retention following a pT-to-pS mutation relative to the probability of phosphorylated threonine retention when no mutations have occurred (Fig. 11C). Thus, the observed enhanced pT-to-pS mutation rate could be due to the enhanced evolutionary stability of serine phosphorylation relative to the threonine phosphorylation.

The enhanced serine-to-NCA mutation rates could stem from the physico-chemical similarity of phosphorylated serines and negatively charged amino acids: both types of residues introduce negatively charged groups of similar size to the protein globule. Thus, if phosphorylation is (almost) constitutive, i.e. happens very frequently in a large number of tissues, we would expect the serine-to-NCA mutation rate to be enhanced. Indeed, ubiquitous phosphorylated serines have the pS-to-E mutation rate more than six-fold larger than the S-to-E mutation rate (Fig. 14B). However, the same pattern does not hold for phospho-threonines (Fig. 13B).

The differences in the mutation rates observed for phosphosites are stronger when clustered phosphosites are considered. Although this might be explained by individual phosphosites likely resulting from noise generated by the phosphorylation machinery (Landry et al. 2014), this could also indicate a general pattern of phosphosites constantly arising at random points of the proteome due to a constant evolutionary process. If phosphorylation at a focal site turns out to be advantageous, its individual context could be reinforced yielding broader phosphorylation pattern of this site or, alternatively, other phosphosites could emerge in the vicinity of this phosphosite, thus forming phospho-islands. As the vast majority of broadly phosphorylated sites are clustered, and clustered phosphosites demonstrate stronger phosphosite-specific features than individual phosphosites do, we suggest that formation of phosphorylation clusters around beneficial phosphosites is the prevalent process compared to context reinforcement of just one site. However, this hypothesis requires further verification.

5.4.3. Human phosphosites

The results obtained for the human set of phosphosites differ somewhat from those for the mouse and HMR sets, like in cases with different STY amino acids representation among phosphorylated amino acids (Fig. 11D), proportion of phosphosites located in phospho-islands (Fig. 12E) or some mutational patterns of phosphorylated STY amino acids (Fig. 13B). This could be explained by differences in experimental procedures used to obtain phosphosite lists for human and for mouse and rat. Whereas for classic laboratory organisms, phosphosites are obtained directly from the analysis of an organism or an analysis of its live organ (Huttlin et al. 2010), for human phosphosite inference immortalized cell lines, such as HeLa, are used (Bekker-Jensen et al. 2017; Xu et al. 2017), with conditions differing from those *in vivo*, and hence one could expect different patterns of phosphorylation. In particular, the lower rate of mutations to NCA could be explained by overrepresentation of sites with noisy phosphorylation manifesting only in cell lines under the conditions of experiments. The mutation of such a residue to NCA would most likely result in the

deleterious effect of an average non-phosphorylated serine mutation to NCA (Jin and Pawson 2012). Thus, we propose that phosphosites conserved between human and rodent lineages, called here HMR sites, are more robust with respect to experimental techniques, and hence are better suited for phosphosite evolutionary studies.

5.4.4. Mutations to lysine

As shown on Fig. 13, pS-to-K mutations are overrepresented in mouse and HMR datasets along with mutations to NCA. We see two possible explanations here: 1) This pattern may be due to the dynamics of basic contexts of clustered phosphosites. Indeed, if a phosphosite cluster possesses a basic context, the substitution of one of the phosphosites to a positively charged lysine should reinforce the context of neighbouring sites. This hypothesis is indirectly supported by pS-to-K mutations being overrepresented for clustered, but not for individual sites in the HMR dataset (Fig. 13B). 2) The most overrepresented mutation of phosphoserines relative to non-phosphorylated serines is pS-to-E. Serine is coded by TCN and AGY codon families, and glutamate – by GAR codons. N represents any of the four nucleotides, Y – cytosine or thymine and R – adenine or guanine. Thus, S-to-K a mutation would require at least two transversions: TCR (S) → GCR (A) → GAR (E)/TCR (S) → TAR (Stop) → GAR (E), or, alternatively, at least two transitions and one transversion: AGY (S) → AGR (R) → GGR (G) → GAR (E)/AGY (S) → AGR (R) → AAR (K) → GAR (E)/AGY (S) → AAY (N) → AAR (K) → GAR (E)/AGY (S) → AAY (N) → GAY (D) → GAR (E)/AGY (S) → GGY (G) → GAY (D) → GAR (E)/AGY (S) → GGY (G) → GGR (G) → GAR (E). Note that lysine is present in two of the paths, hence the overrepresentation of pS-to-K mutations may be due to the presence of intermediate stages of pS-to-E substitutions. Both explanations cannot currently be proven due to the lack of data, however this may be a subject for future investigations.

5.4.5. Evolution of non-studied phosphosite groups

Previous studies dedicated to the evolution of phosphosites have focused on phosphoserines located in IDRs. The large datasets employed in the present study enabled us to assess the patterns of phosphothreonines, phosphotyrosines and sites located in ORs. Apart from the largely enhanced pT-to-pS mutation proportions relative to T-to-S ones (Fig. 13B) no patterns with straightforward biological explanation were observed in these cases. However, an interesting observation here is the consistent, significantly enhanced rate of pY-to-I mutations relative to the Y-to-I mutations in the mouse and HMR datasets (Fig. 13B).

Chapter 6. Conclusions

Both A-to-I RNA editing and protein phosphorylation discussed in the present thesis heritably change the encoded information beyond the genomic blueprint. These modifications, while changing the expressed genomic information in an organism, are also associated with divergent mutational patterns in populations in respective sites or, as is the case with coleoid A-to-I editing, with apparent positive selection at heavily modified sites. Both of these modifications also tend to cluster along the transcripts of proteins, thus adding to the expressed variability they generate.

Briefly, the main results reported here are:

1. Editing in coleoid cephalopods sometime masks beneficial A-to-G substitutions.
2. A-to-I mRNA editing in coleoids largely depends on the local secondary RNA structure. These structures are both the main factor contributing to heredity of editing sites and to clusterization of editing sites.
3. Non-synonimous clustered editing in coleoids contributes almost a half to the total proteome variability and clustered editing in general contributes almost half to the transcriptome variability.
4. Phosphorylation sites in mammalian proteomes are prone to mutations to negatively charged amino acids.
5. Clustered phosphosites have more acidic contexts and are substituted to negatively charged amino acids more frequently than individual ones.

RNA editing sites are much more numerous in soft-bodied cephalopods (coleoids) than in any other studied group (Liscovitch-Brauer et al., 2017). In our study (Moldovan et al., 2020), we show the capacity of numerous coleoid RNA editing sites to function as surrogates for beneficial A-to-G substitutions. This effect is more pronounced for heavily edited sites. At that, the latter are surrounded by stronger local RNA secondary structure (expectedly) and feature different sequence context (unexpectedly). The RNA structure is even stronger around edited adenines homologous to guanines

in sister species. Edited adenines tend to be substituted to guanines, and this tendency is supported by positive selection at highly edited sites.

These observations may be explained by the beneficial effect of increased phenotypic diversity in a low-polymorphic population, enhancing adaptation and facilitating the evolutionary process. Besides, A-to-I editing at sites where G would be preferred provides a larger (than a single nucleotide position) target for mutations increasing the editing level. Together with similar recent observations on *Drosophila* and human editing sites (Popitsch et al. 2020), this points at a general role of RNA editing in the molecular evolution of metazoans.

RNA editing sites tend to cluster. In the follow-up study (Moldovan et al., 2022), we study this property in detail and show that the clustering of A-to-I editing sites in coleoid cephalopods can be largely explained by three types of RNA structural features. The clustering itself contributes about a half to the proteome variance generated by editing.

In our study of protein phosphorylation (Moldovan and Gelfand, 2020), we propose a simple yet accurate homology-based approach for the ancestral phosphosite inference yielding in our case the set of HMR phosphosites. As the predicted fractions of phosphorylation labels falsely assigned to internal tree nodes are much smaller than the ones for other phosphosite datasets, HMR set poses a valuable source of data for evolutionary studies. A practical extension of our homology-based approach could be a phosphosite prediction procedure incorporating additional pieces of information such as the tendency of phosphosites to cluster, the local phosphosite contexts, and the tree structure into the probabilistic model, which would predict phosphosites with a high degree of accuracy. On the other hand, it would be interesting to infer the interplay between phosphorylation and selection using population-genetics data.

In general, our studies provide important pieces to the puzzle that is the evolution of information alterations in biological populations. Namely the concordance between RNA editing and

positive selection in coleoid cephalopods and dependence between protein phosphorylation and selective constraint on local amino acid substitutions.

Bibliography

Albertin, Caroline B., Oleg Simakov, Therese Mitros, Z. Yan Wang, Judit R. Pungor, Eric Edsinger-Gonzales, Sydney Brenner, Clifton W. Ragsdale, and Daniel S. Rokhsar. 2015. 'The Octopus Genome and the Evolution of Cephalopod Neural and Morphological Novelties'. *Nature* 524 (7564): 220–24. https://doi.org/10.1038/nature14668.

Al-Khouri, Anna Maria, Yuliang Ma, Summanuna H. Togo, Scott Williams, and Tomas Mustelin. 2005. 'Cooperative Phosphorylation of the Tumor Suppressor Phosphatase and Tensin Homologue (PTEN) by Casein Kinases and Glycogen Synthase Kinase 3β'. *Journal of Biological Chemistry* 280 (42): 35195–202. https://doi.org/10.1074/jbc.M503045200.

Allison, A. C. 1954. 'Protection Afforded by Sickle-Cell Trait against Subtertian Malareal Infection'. *British Medical Journal* 1 (4857): 290–94. https://doi.org/10.1136/bmj.1.4857.290.

Alon, Shahar, Sandra C Garrett, Erez Y Levanon, Sara Olson, Brenton R Graveley, Joshua J C Rosenthal, and Eli Eisenberg. 2015. 'The Majority of Transcripts in the Squid Nervous System Are Extensively Recoded by A-to-I RNA Editing'. *ELife* 4 (January): e05198. https://doi.org/10.7554/eLife.05198.

Alon, Shahar, Eyal Mor, Francois Vigneault, George M. Church, Franco Locatelli, Federica Galeano, Angela Gallo, Noam Shomron, and Eli Eisenberg. 2012. 'Systematic Identification of Edited MicroRNAs in the Human Brain'. *Genome Research* 22 (8): 1533–40. https://doi.org/10.1101/gr.131573.111.

Altenhoff, Adrian M, Natasha M Glover, Clément-Marie Train, Klara Kaleb, Alex Warwick Vesztrocy, David Dylus, Tarcisio M de Farias, et al. 2018. 'The OMA Orthology Database in 2018: Retrieving Evolutionary Relationships among All Domains of Life through Richer Web and Programmatic Interfaces'. *Nucleic Acids Research* 46 (D1): D477–85. https://doi.org/10.1093/nar/gkx1019.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. 'Basic Local Alignment Search Tool'. *Journal of Molecular Biology* 215 (3): 403–10. https://doi.org/10.1016/S0022-2836(05)80360-2.

Ancel, L. W. 2000. 'Undermining the Baldwin Expediting Effect: Does Phenotypic Plasticity Accelerate Evolution?' *Theoretical Population Biology* 58 (4): 307–19. https://doi.org/10.1006/tpbi.2000.1484.

Ardila, Alfredo. 2016. 'The Evolutionary Concept of "Preadaptation" Applied to Cognitive Neurosciences'. *Frontiers in Neuroscience* 10 (March). https://doi.org/10.3389/fnins.2016.00103.

Ardito, Fatima, Michele Giuliani, Donatella Perrone, Giuseppe Troiano, and Lorenzo Lo Muzio. 2017. 'The Crucial Role of Protein Phosphorylation in Cell Signaling and Its Use as Targeted Therapy (Review)'. *International Journal of Molecular Medicine* 40 (2): 271–80. https://doi.org/10.3892/ijmm.2017.3036.

Athanasiadis, Alekos, Alexander Rich, and Stefan Maas. 2004. 'Widespread A-to-I RNA Editing of Alu-Containing MRNAs in the Human Transcriptome'. Edited by Marv Wickens. *PLoS Biology* 2 (12): e391. https://doi.org/10.1371/journal.pbio.0020391.

Atkins, John F., Gary Loughran, Pramod R. Bhatt, Andrew E. Firth, and Pavel V. Baranov. 2016. 'Ribosomal Frameshifting and Transcriptional Slippage: From Genetic Steganography and Cryptography to Adventitious Use'. *Nucleic Acids Research*, July, gkw530. https://doi.org/10.1093/nar/gkw530.

Avery, P. J., and W. G. Hill. 1977. 'Variability in Genetic Parameters among Small Populations'. *Genetical Research* 29 (3): 193–213. https://doi.org/10.1017/S0016672300017286.

Baeza, Josue, Michael J. Smallegan, and John M. Denu. 2016. 'Mechanisms and Dynamics of Protein Acetylation in Mitochondria'. *Trends in Biochemical Sciences* 41 (3): 231–44. https://doi.org/10.1016/j.tibs.2015.12.006.

Bahn, Jae Hoon, Jae-Hyung Lee, Gang Li, Christopher Greer, Guangdun Peng, and Xinshu Xiao. 2012. 'Accurate Identification of A-to-I RNA Editing in Human by Transcriptome Sequencing'. *Genome Research* 22 (1): 142–50. https://doi.org/10.1101/gr.124107.111.

Barak, Michal, Erez Y. Levanon, Eli Eisenberg, Nurit Paz, Gideon Rechavi, George M. Church, and Ramit Mehr. 2009. 'Evidence for Large Diversity in the Human Transcriptome Created by Alu RNA Editing'. *Nucleic Acids Research* 37 (20): 6905–15. https://doi.org/10.1093/nar/gkp729.

Barton, Nick, and Linda Partridge. 2000. 'Limits to Natural Selection'. *BioEssays* 22 (12): 1075–84. https://doi.org/10.1002/1521-1878(200012)22:12<1075::AID-BIES5>3.0.CO;2-M.

Bass, Brenda L., and Harold Weintraub. 1988. 'An Unwinding Activity That Covalently Modifies Its Double-Stranded RNA Substrate'. *Cell* 55 (6): 1089–98. https://doi.org/10.1016/0092-8674(88)90253-X.

Bekker-Jensen, Dorte B., Christian D. Kelstrup, Tanveer S. Batth, Sara C. Larsen, Christa Haldrup, Jesper B. Bramsen, Karina D. Sørensen, et al. 2017. 'An Optimized Shotgun Strategy for the Rapid Generation of Comprehensive Human Proteomes'. *Cell Systems* 4 (6): 587-599.e4. https://doi.org/10.1016/j.cels.2017.05.009.

Bell, Graham. 2013. 'Evolutionary Rescue and the Limits of Adaptation'. *Philosophical Transactions of the Royal Society B: Biological Sciences* 368 (1610): 20120080. https://doi.org/10.1098/rstb.2012.0080.

Bendixsen, Devin P., Bjørn Østman, and Eric J. Hayden. 2017. 'Negative Epistasis in Experimental RNA Fitness Landscapes'. *Journal of Molecular Evolution* 85 (5–6): 159–68. https://doi.org/10.1007/s00239-017-9817-5.

Berezikov, Eugene. 2011. 'Evolution of MicroRNA Diversity and Regulation in Animals'. *Nature Reviews Genetics* 12 (12): 846–60. https://doi.org/10.1038/nrg3079.

Berg, Johannes, Stana Willmann, and Michael Lässig. 2004. 'Adaptive Evolution of Transcription Factor Binding Sites'. *BMC Evolutionary Biology* 4 (October): 42. https://doi.org/10.1186/1471-2148-4-42.

Berg, Otto G., and Peter H. von Hippel. 1988. 'Selection of DNA Binding Sites by Regulatory Proteins'. *Trends in Biochemical Sciences* 13 (6): 207–11. https://doi.org/10.1016/0968-0004(88)90085-0.

Bertram, Jason, and Joanna Masel. 2019. 'Different Mechanisms Drive the Maintenance of Polymorphism at Loci Subject to Strong versus Weak Fluctuating Selection'. *Evolution* 73 (5): 883–96. https://doi.org/10.1111/evo.13719.

Bisswanger, Hans. 2014. 'Enzyme Assays'. *Perspectives in Science* 1 (1–6): 41–55. https://doi.org/10.1016/j.pisc.2014.02.005.

Bonferroni. 1936. *Teoria Statistica Delle Classi e Calcolo Delle Probabilità*. Firenze: Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze.

Booker, Tom R., Benjamin C. Jackson, and Peter D. Keightley. 2017. 'Detecting Positive Selection in the Genome'. *BMC Biology* 15 (1): 98. https://doi.org/10.1186/s12915-017-0434-y.

Bromham, Lindell, and David Penny. 2003. 'The Modern Molecular Clock'. *Nature Reviews Genetics* 4 (3): 216–24. https://doi.org/10.1038/nrg1020.

Brusa, Rossella, Frank Zimmermann, Duk-Su Koh, Dirk Feldmeyer, Peter Gass, Peter H. Seeburg, and Rolf Sprengel. 1995. 'Early-Onset Epilepsy and Postnatal Lethality Associated with an Editing-Deficient *GluR-B* Allele in Mice'. *Science* 270 (5242): 1677–80. https://doi.org/10.1126/science.270.5242.1677.

Bubb, K L, D Bovee, D Buckley, E Haugen, M Kibukawa, M Paddock, A Palmieri, et al. 2006. 'Scan of Human Genome Reveals No New Loci Under Ancient Balancing Selection'. *Genetics* 173 (4): 2165–77. https://doi.org/10.1534/genetics.106.055715.

Buchumenski, Ilana, Osnat Bartok, Reut Ashwal-Fluss, Varun Pandey, Hagit T. Porath, Erez Y. Levanon, and Sebastian Kadener. 2017. 'Dynamic Hyper-Editing Underlies Temperature Adaptation in Drosophila'. Edited by Jin Billy Li. *PLOS Genetics* 13 (7): e1006931. https://doi.org/10.1371/journal.pgen.1006931.

Cadotte, Marc W., Sara E. Campbell, Shao-peng Li, Darwin S. Sodhi, and Nicholas E. Mandrak. 2018. 'Preadaptation and Naturalization of Nonnative Species: Darwin's Two Fundamental Insights into Species Invasion'. *Annual Review of Plant Biology* 69 (1): 661–84. https://doi.org/10.1146/annurev-arplant-042817-040339.

Carter, Ashley J.R., Joachim Hermisson, and Thomas F. Hansen. 2005. 'The Role of Epistatic Gene Interactions in the Response to Selection and the Evolution of Evolvability'. *Theoretical Population Biology* 68 (3): 179–96. https://doi.org/10.1016/j.tpb.2005.05.002.

Casinos, Adrià. 2017. 'From Cuénot's Préadaptation to Gould and Vrba's Exaptation: A Review'. *Biological Journal of the Linnean Society* 121 (2): 239–47. https://doi.org/10.1093/biolinnean/blw038.

Charlesworth, Brian. 1979. 'Evidence against Fisher's Theory of Dominance'. *Nature* 278 (5707): 848–49. https://doi.org/10.1038/278848a0.

Charlesworth, Deborah. 2006. 'Balancing Selection and Its Effects on Sequences in Nearby Genome Regions'. *PLoS Genetics* 2 (4): e64. https://doi.org/10.1371/journal.pgen.0020064.

Charlesworth, Deborah, Nicholas H. Barton, and Brian Charlesworth. 2017. 'The Sources of Adaptive Variation'. *Proceedings of the Royal Society B: Biological Sciences* 284 (1855): 20162864. https://doi.org/10.1098/rspb.2016.2864.

Chen, Jessica Walton, Pedro Romero, Vladimir N. Uversky, and A. Keith Dunker. 2006. 'Conservation of Intrinsic Disorder in Protein Domains and Families: II. Functions of Conserved Disorder'. *Journal of Proteome Research* 5 (4): 888–98. https://doi.org/10.1021/pr060049p.

Chen, Liang. 2013. 'Characterization and Comparison of Human Nuclear and Cytosolic Editomes'. *Proceedings of the National Academy of Sciences* 110 (29). https://doi.org/10.1073/pnas.1218884110.

Conover, David O., Tara A. Duffy, and Lyndie A. Hice. 2009. 'The Covariance between Genetic and Environmental Influences across Ecological Gradients: Reassessing the Evolutionary Significance of Countergradient and Cogradient Variation'. *Annals of the New York Academy of Sciences* 1168 (1): 100–129. https://doi.org/10.1111/j.1749-6632.2009.04575.x.

Crick, F. H. 1958. 'On Protein Synthesis'. *Symposia of the Society for Experimental Biology* 12: 138–63.

Crick, F. H., L. Barnett, S. Brenner, and R. J. Watts-Tobin. 1961. 'General Nature of the Genetic Code for Proteins'. *Nature* 192 (December): 1227–32. https://doi.org/10.1038/1921227a0.

Crispo, Erika. 2007. 'THE BALDWIN EFFECT AND GENETIC ASSIMILATION: REVISITING TWO MECHANISMS OF EVOLUTIONARY CHANGE MEDIATED BY PHENOTYPIC PLASTICITY'. *Evolution* 61 (11): 2469–79. https://doi.org/10.1111/j.1558-5646.2007.00203.x.

Crooks, Gavin E., Gary Hon, John-Marc Chandonia, and Steven E. Brenner. 2004. 'WebLogo: A Sequence Logo Generator: Figure 1'. *Genome Research* 14 (6): 1188–90. https://doi.org/10.1101/gr.849004.

Crow, J. F. 1958. 'Some Possibilities for Measuring Selection Intensities in Man'. *Human Biology* 30 (1): 1–13.

Darwin, Charles. 1872. *The Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life, 6th Edition; with Additions and Corrections*. London: John Murray.

Dobzhansky, Theodosius, Max K Hecht, and William C Steere. 1968. *Evolutionary Biology: Volume 2*. Boston, MA: Springer US. https://doi.org/10.1007/978-1-4684-8094-8.

Duan, Yuange, Shengqian Dou, Shiqi Luo, Hong Zhang, and Jian Lu. 2017. 'Adaptation of A-to-I RNA Editing in Drosophila'. Edited by Jianzhi Zhang. *PLOS Genetics* 13 (3): e1006648. https://doi.org/10.1371/journal.pgen.1006648.

Duan, Yuange, Shengqian Dou, Hong Zhang, Changcheng Wu, Mingming Wu, and Jian Lu. 2018. 'Linkage of A-to-I RNA Editing in Metazoans and the Impact on Genome Evolution'. *Molecular Biology and Evolution* 35 (1): 132–48. https://doi.org/10.1093/molbev/msx274.

Duclos, Kevin K., Jesse L. Hendrikse, and Heather A. Jamniczky. 2019. 'Investigating the Evolution and Development of Biological Complexity under the Framework of Epigenetics'. *Evolution & Development* 21 (5): 276–93. https://doi.org/10.1111/ede.12301.

Durrett, Rick, and Deena Schmidt. 2008. 'Waiting for Two Mutations: With Applications to Regulatory Sequence Evolution and the Limits of Darwinian Evolution'. *Genetics* 180 (3): 1501–9. https://doi.org/10.1534/genetics.107.082610.

Earl, David J., and Michael W. Deem. 2004. 'Evolvability Is a Selectable Trait'. *Proceedings of the National Academy of Sciences of the United States of America* 101 (32): 11531–36. https://doi.org/10.1073/pnas.0404656101.

Edgar, Robert C. 2004. 'MUSCLE: A Multiple Sequence Alignment Method with Reduced Time and Space Complexity'. *BMC Bioinformatics* 5 (1): 113. https://doi.org/10.1186/1471-2105-5-113.

Eggington, Julie M., Tom Greene, and Brenda L. Bass. 2011. 'Predicting Sites of ADAR Editing in Double-Stranded RNA'. *Nature Communications* 2 (1): 319. https://doi.org/10.1038/ncomms1324.

Eisenberg, Eli, and Erez Y. Levanon. 2018. 'A-to-I RNA Editing — Immune Protector and Transcriptome Diversifier'. *Nature Reviews Genetics* 19 (8): 473–90. https://doi.org/10.1038/s41576-018-0006-1.

Ejsmond, Maciej, Wiesław Babik, and Jacek Radwan. 2010. 'MHC Allele Frequency Distributions under Parasite-Driven Selection: A Simulation Model'. *BMC Evolutionary Biology* 10 (1): 332. https://doi.org/10.1186/1471-2148-10-332.

Eldredge, N., and S. J. Gould. 1997. 'On Punctuated Equilibria'. *Science (New York, N.Y.)* 276 (5311): 338–41. https://doi.org/10.1126/science.276.5311.337c.

Ensterö, Mats, Chammiran Daniel, Helene Wahlstedt, François Major, and Marie Öhman. 2009. 'Recognition and Coupling of A-to-I Edited Sites Are Determined by the Tertiary Structure of the RNA'. *Nucleic Acids Research* 37 (20): 6916–26. https://doi.org/10.1093/nar/gkp731.

Farajollahi, Sanaz, and Stefan Maas. 2010. 'Molecular Diversity through RNA Editing: A Balancing Act'. *Trends in Genetics* 26 (5): 221–30. https://doi.org/10.1016/j.tig.2010.02.001.

Feketová, Zuzana, Tomáš Mašek, Václav Vopálenský, and Martin Pospíšek. 2010. 'Ambiguous Decoding of the CUG Codon Alters the Functionality of the Candida Albicans Translation Initiation Factor 4E: Candida Albicans EIF4E'. *FEMS Yeast Research*, April, no-no. https://doi.org/10.1111/j.1567-1364.2010.00629.x.

Feldmeyer, Dirk, Kalev Kask, Rossella Brusa, Hans–Christian Kornau, Rohini Kolhekar, Andrei Rozov, Nail Burnashev, et al. 1999. 'Neurological Dysfunctions in Mice Expressing Different Levels of the Q/R Site–Unedited AMPAR Subunit GluR–B'. *Nature Neuroscience* 2 (1): 57–64. https://doi.org/10.1038/4561.

Fisher, R. A. 1928. 'Two Further Notes on the Origin of Dominance'. *The American Naturalist* 62 (683): 571–74. https://doi.org/10.1086/280234.

Fitzpatrick, Benjamin M. 2012. 'Underappreciated Consequences of Phenotypic Plasticity for Ecological Speciation'. *International Journal of Ecology* 2012: 1–12. https://doi.org/10.1155/2012/256017.

Ford, Edmund Briscoe. 1971. *Ecological Genetics*. 3rd ed. London: Chapman and Hall.

Friedman, Nir, Long Cai, and X. Sunney Xie. 2006. 'Linking Stochastic Dynamics to Population Distribution: An Analytical Framework of Gene Expression'. *Physical Review Letters* 97 (16): 168302. https://doi.org/10.1103/PhysRevLett.97.168302.

Fuhs, Stephen Rush, and Tony Hunter. 2017. ‘PHisphorylation: The Emergence of Histidine Phosphorylation as a Reversible Regulatory Modification’. *Current Opinion in Cell Biology* 45 (April): 8–16. https://doi.org/10.1016/j.ceb.2016.12.010.

Garland, Theodore, and Scott A. Kelly. 2006. ‘Phenotypic Plasticity and Experimental Evolution’. *Journal of Experimental Biology* 209 (12): 2344–61. https://doi.org/10.1242/jeb.02244.

Garrett, Sandra, and Joshua J. C. Rosenthal. 2012. ‘RNA Editing Underlies Temperature Adaptation in K $^{+}$ Channels from Polar Octopuses’. *Science* 335 (6070): 848–51. https://doi.org/10.1126/science.1212795.

Ghalambor, C. K., J. K. McKAY, S. P. Carroll, and D. N. Reznick. 2007. ‘Adaptive versus Non-Adaptive Phenotypic Plasticity and the Potential for Contemporary Adaptation in New Environments’. *Functional Ecology* 21 (3): 394–407. https://doi.org/10.1111/j.1365-2435.2007.01283.x.

Ghalambor, Cameron K., Kim L. Hoke, Emily W. Ruell, Eva K. Fischer, David N. Reznick, and Kimberly A. Hughes. 2015. ‘Non-Adaptive Plasticity Potentiates Rapid Adaptive Evolution of Gene Expression in Nature’. *Nature* 525 (7569): 372–75. https://doi.org/10.1038/nature15256.

Gommans, Willemijn M., Sean P. Mullen, and Stefan Maas. 2009. ‘RNA Editing: A Driving Force for Adaptive Evolution?’ *BioEssays* 31 (10): 1137–45. https://doi.org/10.1002/bies.200900045.

Gould, S. J., and R. C. Lewontin. 1979. ‘The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme’. *Proceedings of the Royal Society of London. Series B, Biological Sciences* 205 (1161): 581–98. https://doi.org/10.1098/rspb.1979.0086.

Gould, Stephen Jay, and Elisabeth S. Vrba. 1982a. ‘Exaptation—a Missing Term in the Science of Form’. *Paleobiology* 8 (1): 4–15. https://doi.org/10.1017/S0094837300004310.

———. 1982b. ‘Exaptation—a Missing Term in the Science of Form’. *Paleobiology* 8 (1): 4–15. https://doi.org/10.1017/S0094837300004310.

Grether, Gregory F. 2005. ‘Environmental Change, Phenotypic Plasticity, and Genetic Compensation’. *The American Naturalist* 166 (4): E115–23. https://doi.org/10.1086/432023.

GTEx Consortium, Meng How Tan, Qin Li, Raghuvaran Shanmugam, Robert Piskol, Jennefer Kohler, Amy N. Young, et al. 2017. ‘Dynamic Landscape and Regulation of RNA Editing in Mammals’. *Nature* 550 (7675): 249–54. https://doi.org/10.1038/nature24041.

Harjanto, Dewi, Theodore Papamarkou, Chris J. Oates, Violeta Rayon-Estrada, F. Nina Papavasiliou, and Anastasia Papavasiliou. 2016a. ‘RNA Editing Generates Cellular Subsets with Diverse Sequence within Populations’. *Nature Communications* 7 (1): 12145. https://doi.org/10.1038/ncomms12145.

———. 2016b. ‘RNA Editing Generates Cellular Subsets with Diverse Sequence within Populations’. *Nature Communications* 7 (1): 12145. https://doi.org/10.1038/ncomms12145.

Hedges, S. B., J. Dudley, and S. Kumar. 2006. ‘TimeTree: A Public Knowledge-Base of Divergence Times among Organisms’. *Bioinformatics* 22 (23): 2971–72. https://doi.org/10.1093/bioinformatics/btl505.

Hiller, Michael, and Matthias Platzer. 2008. ‘Widespread and Subtle: Alternative Splicing at Short-Distance Tandem Sites’. *Trends in Genetics* 24 (5): 246–55. https://doi.org/10.1016/j.tig.2008.03.003.

Ho, Wei-Chin, and Jianzhi Zhang. 2018. 'Evolutionary Adaptations to New Environments Generally Reverse Plastic Phenotypic Changes'. *Nature Communications* 9 (1): 350. https://doi.org/10.1038/s41467-017-02724-5.

Hoekstra, Hopi E., and Jerry A. Coyne. 2007. 'THE LOCUS OF EVOLUTION: EVO DEVO AND THE GENETICS OF ADAPTATION: THE LOCUS OF EVOLUTION'. *Evolution* 61 (5): 995–1016. https://doi.org/10.1111/j.1558-5646.2007.00105.x.

Huang, Hongzhan, Cecilia N Arighi, Karen E Ross, Jia Ren, Gang Li, Sheng-Chih Chen, Qinghua Wang, Julie Cowart, K Vijay-Shanker, and Cathy H Wu. 2018. 'IPTMnet: An Integrated Resource for Protein Post-Translational Modification Network Discovery'. *Nucleic Acids Research* 46 (D1): D542–50. https://doi.org/10.1093/nar/gkx1104.

Hughes, Kimberly A., and Jeff Leips. 2017. 'Pleiotropy, Constraint, and Modularity in the Evolution of Life Histories: Insights from Genomic Analyses'. *Annals of the New York Academy of Sciences* 1389 (1): 76–91. https://doi.org/10.1111/nyas.13256.

Huttlin, Edward L., Mark P. Jedrychowski, Joshua E. Elias, Tapasree Goswami, Ramin Rad, Sean A. Beausoleil, Judit Villén, Wilhelm Haas, Mathew E. Sowa, and Steven P. Gygi. 2010. 'A Tissue-Specific Atlas of Mouse Protein Phosphorylation and Expression'. *Cell* 143 (7): 1174–89. https://doi.org/10.1016/j.cell.2010.12.001.

Iakoucheva, L. M. 2004. 'The Importance of Intrinsic Disorder for Protein Phosphorylation'. *Nucleic Acids Research* 32 (3): 1037–49. https://doi.org/10.1093/nar/gkh253.

Jiang, Daohan, and Jianzhi Zhang. 2019. 'The Preponderance of Nonsynonymous A-to-I RNA Editing in Coleoids Is Nonadaptive'. *Nature Communications* 10 (1): 5411. https://doi.org/10.1038/s41467-019-13275-2.

Jin, Jing, and Tony Pawson. 2012. 'Modular Evolution of Phosphorylation-Based Signalling Systems'. *Philosophical Transactions of the Royal Society B: Biological Sciences* 367 (1602): 2540–55. https://doi.org/10.1098/rstb.2012.0106.

Jin, Yongfeng, Wenjing Zhang, and Qi Li. 2009. 'Origins and Evolution of ADAR-Mediated RNA Editing'. *IUBMB Life* 61 (6): 572–78. https://doi.org/10.1002/iub.207.

Kallman, A. M. 2003. 'ADAR2 A->I Editing: Site Selectivity and Editing Efficiency Are Separate Events'. *Nucleic Acids Research* 31 (16): 4874–81. https://doi.org/10.1093/nar/gkg681.

Karczewski, Konrad J., Laurent C. Francioli, Grace Tiao, Beryl B. Cummings, Jessica Alföldi, Qingbo Wang, Ryan L. Collins, et al. 2020. 'The Mutational Constraint Spectrum Quantified from Variation in 141,456 Humans'. *Nature* 581 (7809): 434–43. https://doi.org/10.1038/s41586-020-2308-7.

Kim, Dennis D.Y., Thomas T.Y. Kim, Thomas Walsh, Yoshifumi Kobayashi, Tara C. Matise, Steven Buyske, and Abram Gabriel. 2004. 'Widespread RNA Editing of Embedded *Alu* Elements in the Human Transcriptome'. *Genome Research* 14 (9): 1719–25. https://doi.org/10.1101/gr.2855504.

Kimura, Motoo. 1960. 'Genetic Load of a Population and Its Significance in Evolution'. *The Japanese Journal of Genetics* 35 (1): 7–33. https://doi.org/10.1266/jjg.35.7.

———. 1983. *The Neutral Theory of Molecular Evolution*. Cambridge [Cambridgeshire] ; New York: Cambridge University Press.

Klironomos, Filippos D., Johannes Berg, and Sinéad Collins. 2013. 'How Epigenetic Mutations Can Affect Genetic Evolution: Model and Mechanism: Problems & Paradigms'. *BioEssays* 35 (6): 571–78. https://doi.org/10.1002/bies.201200169.

Kolmogorov, Andrey N. 1933. 'Sulla Determinazione Empirica Di Una Legge Di Distribuzione', no. 4: 83–91.

Kondrashov, Dmitry A., and Fyodor A. Kondrashov. 2015. 'Topological Features of Rugged Fitness Landscapes in Sequence Space'. *Trends in Genetics* 31 (1): 24–33. https://doi.org/10.1016/j.tig.2014.09.009.

Koonin, Eugene V. 2006. 'The Origin of Introns and Their Role in Eukaryogenesis: A Compromise Solution to the Introns-Early versus Introns-Late Debate?' *Biology Direct* 1 (August): 22. https://doi.org/10.1186/1745-6150-1-22.

Koonin, Eugene V. 2012. 'Does the Central Dogma Still Stand?' *Biology Direct* 7 (1): 27. https://doi.org/10.1186/1745-6150-7-27.

Koonin, Eugene V. 2016. 'The Meaning of Biological Information'. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374 (2063): 20150065. https://doi.org/10.1098/rsta.2015.0065.

Koonin, Eugene V., Miklos Csuros, and Igor B. Rogozin. 2013. 'Whence Genes in Pieces: Reconstruction of the Exon-Intron Gene Structures of the Last Eukaryotic Common Ancestor and Other Ancestral Eukaryotes: Whence Genes in Pieces'. *Wiley Interdisciplinary Reviews: RNA* 4 (1): 93–105. https://doi.org/10.1002/wrna.1143.

Koonin, Eugene V, and Yuri I Wolf. 2009. 'Is Evolution Darwinian or/and Lamarckian?' *Biology Direct* 4 (1): 42. https://doi.org/10.1186/1745-6150-4-42.

Koshi, Jeffrey M., and Richard A. Goldstein. 1996. 'Probabilistic Reconstruction of Ancestral Protein Sequences'. *Journal of Molecular Evolution* 42 (2): 313–20. https://doi.org/10.1007/BF02198858.

Kronholm, Ilkka, and Sinéad Collins. 2016. 'Epigenetic Mutations Can Both Help and Hinder Adaptive Evolution'. *Molecular Ecology* 25 (8): 1856–68. https://doi.org/10.1111/mec.13296.

Kumar, Sudhir, Glen Stecher, Michael Suleski, and S. Blair Hedges. 2017. 'TimeTree: A Resource for Timelines, Timetrees, and Divergence Times'. *Molecular Biology and Evolution* 34 (7): 1812–19. https://doi.org/10.1093/molbev/msx116.

Kurmangaliyev, Yerbol Z, Sammi Ali, and Sergey V Nuzhdin. 2016. 'Genetic Determinants of RNA Editing Levels of ADAR Targets in *Drosophila Melanogaster*'. *G3 Genes|Genomes|Genetics* 6 (2): 391–96. https://doi.org/10.1534/g3.115.024471.

Kurmangaliyev, Yerbol Z, Alexander Goland, and Mikhail S Gelfand. 2011. 'Evolutionary Patterns of Phosphorylated Serines'. *Biology Direct* 6 (1): 8. https://doi.org/10.1186/1745-6150-6-8.

Landry, Christian R., Luca Freschi, Taraneh Zarin, and Alan M. Moses. 2014. 'Turnover of Protein Phosphorylation Evolving under Stabilizing Selection'. *Frontiers in Genetics* 5 (July). https://doi.org/10.3389/fgene.2014.00245.

Lanfear, Robert, Hanna Kokko, and Adam Eyre-Walker. 2014. 'Population Size and the Rate of Evolution'. *Trends in Ecology & Evolution* 29 (1): 33–41. https://doi.org/10.1016/j.tree.2013.09.009.

Langmead, Ben, and Steven L Salzberg. 2012. 'Fast Gapped-Read Alignment with Bowtie 2'. *Nature Methods* 9 (4): 357–59. https://doi.org/10.1038/nmeth.1923.

Lee, Robin van der, Marija Buljan, Benjamin Lang, Robert J. Weatheritt, Gary W. Daughdrill, A. Keith Dunker, Monika Fuxreiter, et al. 2014. 'Classification of Intrinsically Disordered Regions and Proteins'. *Chemical Reviews* 114 (13): 6589–6631. https://doi.org/10.1021/cr400525m.

Leffler, E. M., Z. Gao, S. Pfeifer, L. Segurel, A. Auton, O. Venn, R. Bowden, et al. 2013. 'Multiple Instances of Ancient Balancing Selection Shared Between Humans and Chimpanzees'. *Science* 339 (6127): 1578–82. https://doi.org/10.1126/science.1234070.

Levanon, Erez Y., and Eli Eisenberg. 2015. 'Does RNA Editing Compensate for Alu Invasion of the Primate Genome?: Prospects & Overviews'. *BioEssays* 37 (2): 175–81. https://doi.org/10.1002/bies.201400163.

Levis, Nicholas A., and David W. Pfennig. 2016. 'Evaluating "Plasticity-First" Evolution in Nature: Key Criteria and Empirical Approaches'. *Trends in Ecology & Evolution* 31 (7): 563–74. https://doi.org/10.1016/j.tree.2016.03.012.

———. 2019. 'Phenotypic Plasticity, Canalization, and the Origins of Novelty: Evidence and Mechanisms from Amphibians'. *Seminars in Cell & Developmental Biology* 88 (April): 80–90. https://doi.org/10.1016/j.semcdb.2018.01.012.

Lewontin, R. C. 1958. 'A General Method for Investigating the Equilibrium of Gene Frequency in a Population'. *Genetics* 43 (3): 419–34. https://doi.org/10.1093/genetics/43.3.419.

Lewontin, R C. 1964. 'THE INTERACTION OF SELECTION AND LINKAGE. I. GENERAL CONSIDERATIONS; HETEROTIC MODELS'. *Genetics* 49 (1): 49–67. https://doi.org/10.1093/genetics/49.1.49.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. 'The Sequence Alignment/Map Format and SAMtools'. *Bioinformatics* 25 (16): 2078–79. https://doi.org/10.1093/bioinformatics/btp352.

Li, Qiye, Zongji Wang, Jinmin Lian, Morten Schiøtt, Lijun Jin, Pei Zhang, Yanyan Zhang, et al. 2014. 'Caste-Specific RNA Editomes in the Leaf-Cutting Ant Acromyrmex Echinatior'. *Nature Communications* 5 (1): 4943. https://doi.org/10.1038/ncomms5943.

Libby, R. T., and J. A. Gallant. 1991. 'The Role of RNA Polymerase in Transcriptional Fidelity'. *Molecular Microbiology* 5 (5): 999–1004. https://doi.org/10.1111/j.1365-2958.1991.tb01872.x.

Limoges, C. 1976. 'Natural Selection, Phagocytosis, and Preadaptation: Lucien Cuénot, 1886-1901'. *Journal of the History of Medicine and Allied Sciences* 31 (2): 176–214. https://doi.org/10.1093/jhmas/xxxi.2.176.

Liscovitch-Brauer, Noa, Shahar Alon, Hagit T. Porath, Boaz Elstein, Ron Unger, Tamar Ziv, Arie Admon, Erez Y. Levanon, Joshua J.C. Rosenthal, and Eli Eisenberg. 2017a. 'Trade-off between Transcriptome Plasticity and Genome Evolution in Cephalopods'. *Cell* 169 (2): 191-202.e11. https://doi.org/10.1016/j.cell.2017.03.025.

———. 2017b. ‘Trade-off between Transcriptome Plasticity and Genome Evolution in Cephalopods’. *Cell* 169 (2): 191-202.e11. https://doi.org/10.1016/j.cell.2017.03.025.

Liu, Xiao-Xiao, Qian-Huan Guo, Wei-Bo Xu, Peng Liu, and Kang Yan. 2022. ‘Rapid Regulation of Alternative Splicing in Response to Environmental Stresses’. *Frontiers in Plant Science* 13 (March): 832177. https://doi.org/10.3389/fpls.2022.832177.

Lorenz, Ronny, Stephan H Bernhart, Christian Höner zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. 2011. ‘ViennaRNA Package 2.0’. *Algorithms for Molecular Biology* 6 (1): 26. https://doi.org/10.1186/1748-7188-6-26.

Lush, Jay Laurence. 1937. *Animal Breeding Plans*. Ames, Iowa: Collegiate Press, Inc.

Lynch, Michael. 2020. ‘The Evolutionary Scaling of Cellular Traits Imposed by the Drift Barrier’. *Proceedings of the National Academy of Sciences* 117 (19): 10435–44. https://doi.org/10.1073/pnas.2000446117.

Lynch, Michael, and Kyle Hagner. 2015. ‘Evolutionary Meandering of Intermolecular Interactions along the Drift Barrier’. *Proceedings of the National Academy of Sciences* 112 (1): E30–38. https://doi.org/10.1073/pnas.1421641112.

Lynch, Michael, and Bruce Walsh. 1998. *Genetics and Analysis of Quantitative Traits*. Sunderland, Mass: Sinauer.

Maas, Stefan, Yukio Kawahara, Kristen M. Tamburro, and Kazuko Nishikura. 2006. ‘A-to-I RNA Editing and Human Disease’. *RNA Biology* 3 (1): 1–9. https://doi.org/10.4161/rna.3.1.2495.

Macek, Boris, Florian Gnad, Boumediene Soufi, Chanchal Kumar, Jesper V. Olsen, Ivan Mijakovic, and Matthias Mann. 2008. ‘Phosphoproteome Analysis of E. Coli Reveals Evolutionary Conservation of Bacterial Ser/Thr/Tyr Phosphorylation’. *Molecular & Cellular Proteomics* 7 (2): 299–307. https://doi.org/10.1074/mcp.M700311-MCP200.

Mann, H. B., and D. R. Whitney. 1947. ‘On a Test of Whether One of Two Random Variables Is Stochastically Larger than the Other’. *The Annals of Mathematical Statistics* 18 (1): 50–60. https://doi.org/10.1214/aoms/1177730491.

Markov, A. V., and S. B. Ivnitsky. 2016. ‘Evolutionary Role of Phenotypic Plasticity’. *Moscow University Biological Sciences Bulletin* 71 (4): 185–92. https://doi.org/10.3103/S0096392516040076.

Markov, A. V., S. B. Ivnitsky, M. B. Kornilova, E. B. Naimark, N. G. Shirokova, and K. S. Perfilieva. 2016. ‘Maternal Effect Obscures Adaptation to Adverse Environments and Hinders Divergence in Drosophila Melanogaster’. *Biology Bulletin Reviews* 6 (5): 429–35. https://doi.org/10.1134/S2079086416050054.

McCandlish, David M., and Arlin Stoltzfus. 2014. ‘Modeling Evolution Using the Probability of Fixation: History and Implications’. *The Quarterly Review of Biology* 89 (3): 225–52. https://doi.org/10.1086/677571.

McLennan, Deborah A. 2008. ‘The Concept of Co-Option: Why Evolution Often Looks Miraculous’. *Evolution: Education and Outreach* 1 (3): 247–58. https://doi.org/10.1007/s12052-008-0053-8.

Mendoza-Parra, Marco-Antonio, Malgorzata Nowicka, Wouter Van Gool, and Hinrich Gronemeyer. 2013. 'Characterising ChIP-Seq Binding Patterns by Model-Based Peak Shape Deconvolution'. *BMC Genomics* 14 (1): 834. https://doi.org/10.1186/1471-2164-14-834.

Merkin, Jason, Caitlin Russell, Ping Chen, and Christopher B. Burge. 2012. 'Evolutionary Dynamics of Gene and Isoform Regulation in Mammalian Tissues'. *Science* 338 (6114): 1593–99. https://doi.org/10.1126/science.1228186.

Miao, Benpeng, Qingyu Xiao, Weiran Chen, Yixue Li, and Zhen Wang. 2018. 'Evaluation of Functionality for Serine and Threonine Phosphorylation with Different Evolutionary Ages in Human and Mouse'. *BMC Genomics* 19 (1): 431. https://doi.org/10.1186/s12864-018-4661-6.

Michaelis, L., and M. Menten. 1913. 'Die Kinetik Der Invertinwirkung', no. 49: 333–69.

Minelli, Alessandro, and Giuseppe Fusco. 2010. 'Developmental Plasticity and the Evolution of Animal Complex Life Cycles'. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365 (1540): 631–40. https://doi.org/10.1098/rstb.2009.0268.

Mironov, Aleksei, Stepan Denisov, Alexander Gress, Olga V. Kalinina, and Dmitri D. Pervouchine. 2021. 'An Extended Catalogue of Tandem Alternative Splice Sites in Human Tissue Transcriptomes'. Edited by Ilya Ioshikhes. *PLOS Computational Biology* 17 (4): e1008329. https://doi.org/10.1371/journal.pcbi.1008329.

Mironov, Andrey A., James Wildon Fickett, and Mikhail S. Gelfand. 1999. 'Frequent Alternative Splicing of Human Genes'. *Genome Research* 9 (12): 1288–93. https://doi.org/10.1101/gr.9.12.1288.

Moldovan, Mikhail A., Zoe S. Chervontseva, Daria S. Nogina, and Mikhail S. Gelfand. 2022. 'A Hierarchy in Clusters of Cephalopod MRNA Editing Sites'. *Scientific Reports* 12 (1): 3447. https://doi.org/10.1038/s41598-022-07460-5.

Moldovan, Mikhail, Zoe Chervontseva, Georgii Bazykin, and Mikhail S. Gelfand. 2020. 'Adaptive Evolution at MRNA Editing Sites in Soft-Bodied Cephalopods'. *PeerJ* 8 (November): e10456. https://doi.org/10.7717/peerj.10456.

Moldovan, Mikhail, and Mikhail S. Gelfand. 2020. 'Phospho-Islands and the Evolution of Phosphorylated Amino Acids in Mammals'. *PeerJ* 8: e10436. https://doi.org/10.7717/peerj.10436.

Morse, Daniel P., P. Joseph Aruscavage, and Brenda L. Bass. 2002. 'RNA Hairpins in Noncoding Regions of Human Brain and *Caenorhabditis Elegans* MRNA Are Edited by Adenosine Deaminases That Act on RNA'. *Proceedings of the National Academy of Sciences* 99 (12): 7906–11. https://doi.org/10.1073/pnas.112704299.

Morse, Daniel P., and Brenda L. Bass. 1999. 'Long RNA Hairpins That Contain Inosine Are Present in *Caenorhabditis Elegans* Poly(A) $^{+}$ RNA'. *Proceedings of the National Academy of Sciences* 96 (11): 6048–53. https://doi.org/10.1073/pnas.96.11.6048.

Moses, Alan M., and Christian R. Landry. 2010. 'Moving from Transcriptional to Phospho-Evolution: Generalizing Regulatory Evolution?' *Trends in Genetics* 26 (11): 462–67. https://doi.org/10.1016/j.tig.2010.08.002.

Mustonen, V., J. Kinney, C. G. Callan, and M. Lassig. 2008. 'Energy-Dependent Fitness: A Quantitative Model for the Evolution of Yeast Transcription Factor Binding Sites'. *Proceedings of the National Academy of Sciences* 105 (34): 12376–81. https://doi.org/10.1073/pnas.0805909105.

Mustonen, V., and M. Lassig. 2005. 'Evolutionary Population Genetics of Promoters: Predicting Binding Sites and Functional Phylogenies'. *Proceedings of the National Academy of Sciences* 102 (44): 15936–41. https://doi.org/10.1073/pnas.0505537102.

———. 2010. 'Fitness Flux and Ubiquity of Adaptive Evolution'. *Proceedings of the National Academy of Sciences* 107 (9): 4248–53. https://doi.org/10.1073/pnas.0907953107.

Nam, Kiwoong, Kasper Munch, Thomas Mailund, Alexander Nater, Maja Patricia Greminger, Michael Krützen, Tomàs Marquès-Bonet, and Mikkel Heide Schierup. 2017. 'Evidence That the Rate of Strong Selective Sweeps Increases with Population Size in the Great Apes'. *Proceedings of the National Academy of Sciences* 114 (7): 1613–18. https://doi.org/10.1073/pnas.1605660114.

Narasimhan, Vagheesh, Petr Danecek, Aylwyn Scally, Yali Xue, Chris Tyler-Smith, and Richard Durbin. 2016. 'BCFtools/RoH: A Hidden Markov Model Approach for Detecting Autozygosity from next-Generation Sequencing Data'. *Bioinformatics* 32 (11): 1749–51. https://doi.org/10.1093/bioinformatics/btw044.

Navarro, Arcadio, and Nick H Barton. 2002. 'The Effects of Multilocus Balancing Selection on Neutral Variability'. *Genetics* 161 (2): 849–63. https://doi.org/10.1093/genetics/161.2.849.

Nguyen, Lam-Tung, Heiko A. Schmidt, Arndt von Haeseler, and Bui Quang Minh. 2015. 'IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies'. *Molecular Biology and Evolution* 32 (1): 268–74. https://doi.org/10.1093/molbev/msu300.

Nishi, Hafumi, Alexey Shaytan, and Anna R. Panchenko. 2014. 'Physicochemical Mechanisms of Protein Regulation by Phosphorylation'. *Frontiers in Genetics* 5 (August). https://doi.org/10.3389/fgene.2014.00270.

Nishikura, K., C. Yoo, U. Kim, J. M. Murray, P. A. Estes, F. E. Cash, and S. A. Liebhaber. 1991. 'Substrate Specificity of the DsRNA Unwinding/Modifying Activity'. *The EMBO Journal* 10 (11): 3523–32.

Nishikura, Kazuko. 2006. 'Editor Meets Silencer: Crosstalk between RNA Editing and RNA Interference'. *Nature Reviews Molecular Cell Biology* 7 (12): 919–31. https://doi.org/10.1038/nrm2061.

———. 2010. 'Functions and Regulation of RNA Editing by ADAR Deaminases'. *Annual Review of Biochemistry* 79 (1): 321–49. https://doi.org/10.1146/annurev-biochem-060208-105251.

———. 2016. 'A-to-I Editing of Coding and Non-Coding RNAs by ADARs'. *Nature Reviews Molecular Cell Biology* 17 (2): 83–96. https://doi.org/10.1038/nrm.2015.4.

Ostrom, J. H. 1979. 'Bird Flight: How Did It Begin?' *American Scientist* 67 (1): 46–56.

Ostrom, John H. 1974. 'Archaeopteryx and the Origin of Flight'. *The Quarterly Review of Biology* 49 (1): 27–47. https://doi.org/10.1086/407902.

Ou, Xumin, Jingyu Cao, Anchun Cheng, Maikel P. Peppelenbosch, and Qiuwei Pan. 2019. 'Errors in Translational Decoding: TRNA Wobbling or Misincorporation?' Edited by Ronald B Gartenhaus. *PLOS Genetics* 15 (3): e1008017. https://doi.org/10.1371/journal.pgen.1008017.

Paenke, Ingo, Bernhard Sendhoff, and Tadeusz J. Kawecki. 2007. 'Influence of Plasticity and Learning on Evolution under Directional Selection'. *The American Naturalist* 170 (2): E47–58. https://doi.org/10.1086/518952.

Partridge, Linda, and Nicholas H. Barton. 2000. 'Evolving Evolvability'. *Nature* 407 (6803): 457–58. https://doi.org/10.1038/35035173.

Paz-Yaacov, Nurit, Erez Y. Levanon, Eviatar Nevo, Yaron Kinar, Alon Harmelin, Jasmine Jacob-Hirsch, Ninette Amariglio, Eli Eisenberg, and Gideon Rechavi. 2010. 'Adenosine-to-Inosine RNA Editing Shapes Transcriptome Diversity in Primates'. *Proceedings of the National Academy of Sciences* 107 (27): 12174–79. https://doi.org/10.1073/pnas.1006183107.

Pearlman, Samuel M., Zach Serber, and James E. Ferrell. 2011. 'A Mechanism for the Evolution of Phosphorylation Sites'. *Cell* 147 (4): 934–46. https://doi.org/10.1016/j.cell.2011.08.052.

Pearson, Karl. 1895. 'VII. Note on Regression and Inheritance in the Case of Two Parents'. *Proceedings of the Royal Society of London* 58 (347–352): 240–42. https://doi.org/10.1098/rspl.1895.0041.

Pecori, Riccardo, Salvatore Di Giorgio, J. Paulo Lorenzo, and F. Nina Papavasiliou. 2022. 'Functions and Consequences of AID/APOBEC-Mediated DNA and RNA Deamination'. *Nature Reviews Genetics*, March. https://doi.org/10.1038/s41576-022-00459-8.

Peng, Zhen-Ling, and Lukasz Kurgan. 2012. 'Comprehensive Comparative Assessment of In-Silico Predictors of Disordered Regions'. *Current Protein & Peptide Science* 13 (1): 6–18. https://doi.org/10.2174/138920312799277938.

Penn, Dustin J., Kristy Damjanovich, and Wayne K. Potts. 2002. 'MHC Heterozygosity Confers a Selective Advantage against Multiple-Strain Infections'. *Proceedings of the National Academy of Sciences* 99 (17): 11260–64. https://doi.org/10.1073/pnas.162006499.

Pfennig, David W., Matthew A. Wund, Emilie C. Snell-Rood, Tami Cruickshank, Carl D. Schlichting, and Armin P. Moczek. 2010. 'Phenotypic Plasticity's Impacts on Diversification and Speciation'. *Trends in Ecology & Evolution* 25 (8): 459–67. https://doi.org/10.1016/j.tree.2010.05.006.

Pickrell, Joseph K., Athma A. Pai, Yoav Gilad, and Jonathan K. Pritchard. 2010. 'Noisy Splicing Drives MRNA Isoform Diversity in Human Cells'. Edited by Emmanouil T. Dermitzakis. *PLoS Genetics* 6 (12): e1001236. https://doi.org/10.1371/journal.pgen.1001236.

Pigliucci, Massimo. 2008. 'Is Evolvability Evolvable?' *Nature Reviews Genetics* 9 (1): 75–82. https://doi.org/10.1038/nrg2278.

Pigliucci, Massimo, Courtney J. Murren, and Carl D. Schlichting. 2006. 'Phenotypic Plasticity and Evolution by Genetic Assimilation'. *Journal of Experimental Biology* 209 (12): 2362–67. https://doi.org/10.1242/jeb.02070.

Pinto, Yishay, Haim Y Cohen, and Erez Y Levanon. 2014. 'Mammalian Conserved ADAR Targets Comprise Only a Small Fragment of the Human Editosome'. *Genome Biology* 15 (1): R5. https://doi.org/10.1186/gb-2014-15-1-r5.

Polson, A. G., and B. L. Bass. 1994. 'Preferential Selection of Adenosines for Modification by Double-Stranded RNA Adenosine Deaminase'. *The EMBO Journal* 13 (23): 5701–11.

Popitsch, Niko, Christian D Huber, Ilana Buchumenski, Eli Eisenberg, Michael Jantsch, Arndt von Haeseler, and Miguel Gallach. 2020. 'A-to-I RNA Editing Uncovers Hidden Signals of Adaptive Genome Evolution in Animals'. Edited by Adam Eyre-Walker. *Genome Biology and Evolution* 12 (4): 345–57. https://doi.org/10.1093/gbe/evaa046.

Prasanth, Kannanganattu V., Supriya G. Prasanth, Zhenyu Xuan, Stephen Hearn, Susan M. Freier, C. Frank Bennett, Michael Q. Zhang, and David L. Spector. 2005. 'Regulating Gene Expression through RNA Nuclear Retention'. *Cell* 123 (2): 249–63. https://doi.org/10.1016/j.cell.2005.08.033.

Price, Trevor D., Anna Qvarnström, and Darren E. Irwin. 2003. 'The Role of Phenotypic Plasticity in Driving Genetic Evolution'. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 270 (1523): 1433–40. https://doi.org/10.1098/rspb.2003.2372.

Ptacek, J, and M Snyder. 2006. 'Charging It up: Global Analysis of Protein Phosphorylation'. *Trends in Genetics* 22 (10): 545–54. https://doi.org/10.1016/j.tig.2006.08.005.

Raj, Arjun, and Alexander van Oudenaarden. 2008. 'Nature, Nurture, or Chance: Stochastic Gene Expression and Its Consequences'. *Cell* 135 (2): 216–26. https://doi.org/10.1016/j.cell.2008.09.050.

Ram, Yoav, and Lilach Hadany. 2014. 'Stress-Induced Mutagenesis and Complex Adaptation'. *Proceedings of the Royal Society B: Biological Sciences* 281 (1792): 20141025. https://doi.org/10.1098/rspb.2014.1025.

Ramaswami, Gokul, Wei Lin, Robert Piskol, Meng How Tan, Carrie Davis, and Jin Billy Li. 2012. 'Accurate Identification of Human Alu and Non-Alu RNA Editing Sites'. *Nature Methods* 9 (6): 579–81. https://doi.org/10.1038/nmeth.1982.

Receveur-Bréchot, Véronique, Jean-Marie Bourhis, Vladimir N. Uversky, Bruno Canard, and Sonia Longhi. 2005. 'Assessing Protein Disorder and Induced Folding'. *Proteins: Structure, Function, and Bioinformatics* 62 (1): 24–45. https://doi.org/10.1002/prot.20750.

Reenan, Robert A. 2005. 'Molecular Determinants and Guided Evolution of Species-Specific RNA Editing'. *Nature* 434 (7031): 409–13. https://doi.org/10.1038/nature03364.

Reiss, David J., Marc T. Facciotti, and Nitin S. Baliga. 2008. 'Model-Based Deconvolution of Genome-Wide DNA Binding'. *Bioinformatics* 24 (3): 396–403. https://doi.org/10.1093/bioinformatics/btm592.

Rieder, Leila E., Cynthia J. Staber, Barry Hoopengardner, and Robert A. Reenan. 2013. 'Tertiary Structural Elements Determine the Extent and Specificity of Messenger RNA Editing'. *Nature Communications* 4 (1): 2232. https://doi.org/10.1038/ncomms3232.

Riedl, Rupert. 1977. 'A Systems-Analytical Approach to Macro-Evolutionary Phenomena'. *The Quarterly Review of Biology* 52 (4): 351–70. https://doi.org/10.1086/410123.

———. 2000. *Strukturen der Komplexit??t: Eine Morphologie des Erkennens und Erkl??rens*. http://link.springer.com/openurl?genre=book&isbn=978-3-642-63111-5.

Riedl, Rupert, and Daniela Auer. 1975. *Die Ordnung Des Lebendigen: Systembedingungen d. Evolution*. Hamburg ; Berlin: Parey.

Rousselle, Marjolaine, Paul Simion, Marie-Ka Tilak, Emeric Figuet, Benoit Nabholz, and Nicolas Galtier. 2020. 'Is Adaptation Limited by Mutation? A Timescale-Dependent Effect of Genetic Diversity on the Adaptive Substitution Rate in Animals'. Edited by Jianzhi Zhang. *PLOS Genetics* 16 (4): e1008668. https://doi.org/10.1371/journal.pgen.1008668.

Sadowski, Martin, and Boris Sarcevic. 2010. 'Mechanisms of Mono- and Poly-Ubiquitination: Ubiquitination Specificity Depends on Compatibility between the E2 Catalytic Core and Amino Acid Residues Proximal to the Lysine'. *Cell Division* 5 (1): 19. https://doi.org/10.1186/1747-1028-5-19.

Sagulenko, Pavel, Vadim Puller, and Richard A Neher. 2018. 'TreeTime: Maximum-Likelihood Phylodynamic Analysis'. *Virus Evolution* 4 (1). https://doi.org/10.1093/ve/vex042.

Samhita, Laasya. 2021. 'The Boggarts of Biology: How Non-Genetic Changes Influence the Genotype'. *Current Genetics* 67 (1): 65–77. https://doi.org/10.1007/s00294-020-01108-5.

Savva, Yiannis A, Leila E Rieder, and Robert A Reenan. 2012. 'The ADAR Protein Family'. *Genome Biology* 13 (12): 252. https://doi.org/10.1186/gb-2012-13-12-252.

Schmalhausen, I.I. 1949. *Factors of Evolution: The Theory of Stabilizing Selection*. Philadelphia: The Blakiston Company.

Schweiger, Regev, and Michal Linial. 2010a. 'Cooperativity within Proximal Phosphorylation Sites Is Revealed from Large-Scale Proteomics Data'. *Biology Direct* 5 (1): 6. https://doi.org/10.1186/1745-6150-5-6.

———. 2010b. 'Cooperativity within Proximal Phosphorylation Sites Is Revealed from Large-Scale Proteomics Data'. *Biology Direct* 5 (1): 6. https://doi.org/10.1186/1745-6150-5-6.

Seplyarskiy, Vladimir B., Ruslan A. Soldatov, Evan Koch, Ryan J. McGinty, Jakob M. Goldmann, Ryan D. Hernandez, Kathleen Barnes, et al. 2021. 'Population Sequencing Data Reveal a Compendium of Mutational Processes in the Human Germ Line'. *Science* 373 (6558): 1030–35. https://doi.org/10.1126/science.aba7408.

Shoshan, Yoav, Noa Liscovitch-Brauer, Joshua J C Rosenthal, and Eli Eisenberg. 2021. 'Adaptive Proteome Diversification by Nonsynonymous A-to-I RNA Editing in Coleoid Cephalopods'. Edited by Mary O'Connell. *Molecular Biology and Evolution* 38 (9): 3775–88. https://doi.org/10.1093/molbev/msab154.

Sievers, Fabian, Andreas Wilm, David Dineen, Toby J Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, et al. 2011. 'Fast, Scalable Generation of High-quality Protein Multiple Sequence Alignments Using Clustal Omega'. *Molecular Systems Biology* 7 (1): 539. https://doi.org/10.1038/msb.2011.75.

Singh, Pooja, Christine Börger, Heather More, and Christian Sturmbauer. 2017. 'The Role of Alternative Splicing and Differential Gene Expression in Cichlid Adaptive Radiation'. *Genome Biology and Evolution* 9 (10): 2764–81. https://doi.org/10.1093/gbe/evx204.

Slijper, E. J. 1942. ‘Biologic Anatomical Investigations on the Bipedal Gait and Upright Posture in Mammals-With Special Reference to a Little Goat Born without Forelegs’, no. 45: 407–15.

Smith, J. M. 1970. ‘Natural Selection and the Concept of a Protein Space’. *Nature* 225 (5232): 563–64. https://doi.org/10.1038/225563a0.

Smith, J. Maynard. 1976. ‘What Determines the Rate of Evolution?’ *The American Naturalist* 110 (973): 331–38. https://doi.org/10.1086/283071.

Sniegowski, Paul D., Philip J. Gerrish, and Richard E. Lenski. 1997. ‘Evolution of High Mutation Rates in Experimental Populations of E. Coli’. *Nature* 387 (6634): 703–5. https://doi.org/10.1038/42701.

Soldatov, Ruslan A., Svetlana V. Vinogradova, and Andrey A. Mironov. 2014. ‘RNASurface: Fast and Accurate Detection of Locally Optimal Potentially Structured RNA Segments’. *Bioinformatics* 30 (4): 457–63. https://doi.org/10.1093/bioinformatics/btt701.

Sommer, Bernd, Martin Köhler, Rolf Sprengel, and Peter H. Seeburg. 1991. ‘RNA Editing in Brain Controls a Determinant of Ion Flow in Glutamate-Gated Channels’. *Cell* 67 (1): 11–19. https://doi.org/10.1016/0092-8674(91)90568-J.

Srinivasan, Bharath. 2021. ‘A Guide to the Michaelis–Menten Equation: Steady State and Beyond’. *The FEBS Journal*, July, febs.16124. https://doi.org/10.1111/febs.16124.

Stefl, Richard, Florian C. Oberstrass, Jennifer L. Hood, Muriel Jourdan, Michal Zimmermann, Lenka Skrisovska, Christophe Maris, et al. 2010. ‘The Solution Structure of the ADAR2 DsRBM-RNA Complex Reveals a Sequence-Specific Readout of the Minor Groove’. *Cell* 143 (2): 225–37. https://doi.org/10.1016/j.cell.2010.09.026.

Studer, Romain A., Ricard A. Rodriguez-Mias, Kelsey M. Haas, Joanne I. Hsu, Cristina Viéitez, Carme Solé, Danielle L. Swaney, et al. 2016. ‘Evolution of Protein Phosphorylation across 18 Fungal Species’. *Science* 354 (6309): 229–32. https://doi.org/10.1126/science.aaf2144.

Su, Chun-Hao, Dhananjaya D, and Woan-Yuh Tarn. 2018. ‘Alternative Splicing in Neurogenesis and Brain Development’. *Frontiers in Molecular Biosciences* 5 (February): 12. https://doi.org/10.3389/fmolb.2018.00012.

The UniProt Consortium. 2019. ‘UniProt: A Worldwide Hub of Protein Knowledge’. *Nucleic Acids Research* 47 (D1): D506–15. https://doi.org/10.1093/nar/gky1049.

Tian, B. 2005. ‘A Large-Scale Analysis of MRNA Polyadenylation of Human and Mouse Genes’. *Nucleic Acids Research* 33 (1): 201–12. https://doi.org/10.1093/nar/gki158.

Tian, Bin, and James L. Manley. 2017. ‘Alternative Polyadenylation of MRNA Precursors’. *Nature Reviews Molecular Cell Biology* 18 (1): 18–30. https://doi.org/10.1038/nrm.2016.116.

Toth-Petroczy, Agnes, Balint Meszaros, Istvan Simon, A. Keith Dunker, Vladimir N. Uversky, and Monika Fuxreiter. 2008. ‘Assessing Conservation of Disordered Regions in Proteins’. *The Open Proteomics Journal* 1 (1): 46–53. https://doi.org/10.2174/1875039700801010046.

True, H. L., and S. L. Lindquist. 2000. ‘A Yeast Prion Provides a Mechanism for Genetic Variation and Phenotypic Diversity’. *Nature* 407 (6803): 477–83. https://doi.org/10.1038/35035005.

Usmanova, Dinara R., Luca Ferretti, Inna S. Povolotskaya, Peter K. Vlasov, and Fyodor A. Kondrashov. 2015. ‘A Model of Substitution Trajectories in Sequence Space and Long-Term Protein Evolution’. *Molecular Biology and Evolution* 32 (2): 542–54. https://doi.org/10.1093/molbev/msu318.

Valente, Louis, and Kazuko Nishikura. 2007. ‘RNA Binding-Independent Dimerization of Adenosine Deaminases Acting on RNA and Dominant Negative Effects of Nonfunctional Subunits on Dimer Functions’. *Journal of Biological Chemistry* 282 (22): 16054–61. https://doi.org/10.1074/jbc.M611392200.

Villén, Judit, Sean A. Beausoleil, Scott A. Gerber, and Steven P. Gygi. 2007. ‘Large-Scale Phosphorylation Analysis of Mouse Liver’. *Proceedings of the National Academy of Sciences* 104 (5): 1488–93. https://doi.org/10.1073/pnas.0609836104.

Visser, J. Arjan G. M. de, and Joachim Krug. 2014. ‘Empirical Fitness Landscapes and the Predictability of Evolution’. *Nature Reviews. Genetics* 15 (7): 480–90. https://doi.org/10.1038/nrg3744.

Viterbi, A. 1967. ‘Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm’. *IEEE Transactions on Information Theory* 13 (2): 260–69. https://doi.org/10.1109/TIT.1967.1054010.

Waddington, C. H. 1942. ‘CANALIZATION OF DEVELOPMENT AND THE INHERITANCE OF ACQUIRED CHARACTERS’. *Nature* 150 (3811): 563–65. https://doi.org/10.1038/150563a0.

———. 1953a. ‘GENETIC ASSIMILATION OF AN ACQUIRED CHARACTER’. *Evolution* 7 (2): 118–26. https://doi.org/10.1111/j.1558-5646.1953.tb00070.x.

———. 1953b. ‘THE “BALDWIN EFFECT,” “GENETIC ASSIMILATION” AND “HOMEOSTASIS”’. *Evolution* 7 (4): 386–87. https://doi.org/10.1111/j.1558-5646.1953.tb00099.x.

———. 1959. ‘Canalization of Development and Genetic Assimilation of Acquired Characters’. *Nature* 183 (4676): 1654–55. https://doi.org/10.1038/1831654a0.

Waddington, C.H. 1961. ‘Genetic Assimilation’. In *Advances in Genetics*, 10:257–93. Elsevier. https://doi.org/10.1016/S0065-2660(08)60119-4.

Wagner, G. P. 1981. ‘Feedback Selection and the Evolution of Modifiers’. *Acta Biotheoretica* 30 (2): 79–102. https://doi.org/10.1007/BF00047674.

Wagner, Gunter P., and Lee Altenberg. 1996. ‘Perspective: Complex Adaptations and the Evolution of Evolvability’. *Evolution* 50 (3): 967. https://doi.org/10.2307/2410639.

Wagner, G□nter P., and Manfred D. Laubichler. 2004. ‘Rupert Riedl and the Re-Synthesis of Evolutionary and Developmental Biology: Body Plans and Evolvability’. *Journal of Experimental Zoology* 302B (1): 92–102. https://doi.org/10.1002/jez.b.20005.

Wahba, Albert J., Robert S. Gardner, Carlos Basilio, Robert S. Miller, Joseph F. Speyer, and Peter Lengyel. 1963. ‘SYNTHETIC POLYNUCLEOTIDES AND THE AMINO ACID CODE, VIII’. *Proceedings of the National Academy of Sciences* 49 (1): 116–22. https://doi.org/10.1073/pnas.49.1.116.

Walsh, Bruce, Michael Lynch, and Michael Lynch. 2018. *Evolution and Selection of Quantitative Traits*. New York, NY: Oxford University Press.

Wang, Andrew H.-J., Toshio Hakoshima, Gijs van der Marel, Jacques H. van Boom, and Alexander Rich. 1984. 'AT Base Pairs Are Less Stable than GC Base Pairs in Z-DNA: The Crystal Structure of d(M5CGTAm5CG)'. *Cell* 37 (1): 321–31. https://doi.org/10.1016/0092-8674(84)90328-3.

Wang, Guey-Shin, and Thomas A. Cooper. 2007. 'Splicing in Disease: Disruption of the Splicing Code and the Decoding Machinery'. *Nature Reviews Genetics* 8 (10): 749–61. https://doi.org/10.1038/nrg2164.

Wang, Yan, Jing Liu, Bo Huang, Yan-Mei Xu, Jing Li, Lin-Feng Huang, Jin Lin, et al. 2015. 'Mechanism of Alternative Splicing and Its Regulation'. *Biomedical Reports* 3 (2): 152–58. https://doi.org/10.3892/br.2014.407.

Wang, Zefeng, and Christopher B. Burge. 2008. 'Splicing Regulation: From a Parts List of Regulatory Elements to an Integrated Splicing Code'. *RNA* 14 (5): 802–13. https://doi.org/10.1261/rna.876308.

West-Eberhard, Mary Jane. 2003. *Developmental Plasticity and Evolution*. Oxford ; New York: Oxford University Press.

———. 2005. 'Developmental Plasticity and the Origin of Species Differences'. *Proceedings of the National Academy of Sciences* 102 (suppl_1): 6543–49. https://doi.org/10.1073/pnas.0501844102.

Wilcoxon, Frank. 1945. 'Individual Comparisons by Ranking Methods'. *Biometrics Bulletin* 1 (6): 80. https://doi.org/10.2307/3001968.

Wong, Swee Kee, Shuji Sato, and David W. Lazinski. 2001. 'Substrate Recognition by ADAR1 and ADAR2'. *RNA* 7 (6): 846–58. https://doi.org/10.1017/S135583820101007X.

Wright, Daniel J, Christopher R Force, and Brent M Znosko. 2018. 'Stability of RNA Duplexes Containing Inosine·cytosine Pairs'. *Nucleic Acids Research* 46 (22): 12099–108. https://doi.org/10.1093/nar/gky907.

Wright, Sewall. 1929. 'Fisher's Theory of Dominance'. *The American Naturalist* 63 (686): 274–79. https://doi.org/10.1086/280260.

———. 1934. 'Physiological and Evolutionary Theories of Dominance'. *The American Naturalist* 68 (714): 24–53. https://doi.org/10.1086/280521.

Xu, Guixia, and Jianzhi Zhang. 2014. 'Human Coding RNA Editing Is Generally Nonadaptive'. *Proceedings of the National Academy of Sciences* 111 (10): 3769–74. https://doi.org/10.1073/pnas.1321745111.

Xu, Hong, Xuanyi Chen, Nanjiao Ying, Meixia Wang, Xiaoli Xu, Rongyi Shi, and Yuejin Hua. 2017. 'Mass Spectrometry-Based Quantification of the Cellular Response to Ultraviolet Radiation in HeLa Cells'. Edited by Michael Shing-Yan Huen. *PLOS ONE* 12 (11): e0186806. https://doi.org/10.1371/journal.pone.0186806.

Xue, Bin, Roland L. Dunbrack, Robert W. Williams, A. Keith Dunker, and Vladimir N. Uversky. 2010. 'PONDR-FIT: A Meta-Predictor of Intrinsically Disordered Amino Acids'. *Biochimica et*

Biophysica Acta (BBA) - Proteins and Proteomics 1804 (4): 996–1010. https://doi.org/10.1016/j.bbapap.2010.01.011.

Yablonovitch, Arielle L., Patricia Deng, Dionna Jacobson, and Jin Billy Li. 2017. 'The Evolution and Adaptation of A-to-I RNA Editing'. Edited by Jianzhi Zhang. *PLOS Genetics* 13 (11): e1007064. https://doi.org/10.1371/journal.pgen.1007064.

Yampolsky, Lev Y., and Arlin Stoltzfus. 2001. 'Bias in the Introduction of Variation as an Orienting Factor in Evolution'. *Evolution and Development* 3 (2): 73–83. https://doi.org/10.1046/j.1525-142x.2001.003002073.x.

Yang, Yun, Jianning Lv, Bin Gui, Heng Yin, Xiaojie Wu, Yaozhou Zhang, and Yongfeng Jin. 2008. 'A-to-I RNA Editing Alters Less-Conserved Residues of Highly Conserved Coding Regions: Implications for Dual Functions in Evolution'. *RNA* 14 (8): 1516–25. https://doi.org/10.1261/rna.1063708.

Yang, Ziheng. 2007. 'PAML 4: Phylogenetic Analysis by Maximum Likelihood'. *Molecular Biology and Evolution* 24 (8): 1586–91. https://doi.org/10.1093/molbev/msm088.

Yu, Yao, Hongxia Zhou, Yimeng Kong, Bohu Pan, Longxian Chen, Hongbing Wang, Pei Hao, and Xuan Li. 2016. 'The Landscape of A-to-I RNA Editome Is Shaped by Both Positive and Purifying Selection'. Edited by Mikkel H. Schierup. *PLOS Genetics* 12 (7): e1006191. https://doi.org/10.1371/journal.pgen.1006191.

Zhang, Hong, Yirong Wang, Xinkai Wu, Xiaolu Tang, Changcheng Wu, and Jian Lu. 2021. 'Determinants of Genome-Wide Distribution and Evolution of UORFs in Eukaryotes'. *Nature Communications* 12 (1): 1076. https://doi.org/10.1038/s41467-021-21394-y.

Zhang, Junmei, Robert Sprung, Jimin Pei, Xiaohong Tan, Sungchan Kim, Heng Zhu, Chuan-Fa Liu, Nick V. Grishin, and Yingming Zhao. 2009. 'Lysine Acetylation Is a Highly Abundant and Evolutionarily Conserved Modification in Escherichia Coli'. *Molecular & Cellular Proteomics* 8 (2): 215–25. https://doi.org/10.1074/mcp.M800187-MCP200.

Zhang, Rui, Patricia Deng, Dionna Jacobson, and Jin Billy Li. 2017. 'Evolutionary Analysis Reveals Regulatory and Functional Landscape of Coding and Non-Coding RNA Editing'. Edited by Jianzhi Zhang. *PLOS Genetics* 13 (2): e1006563. https://doi.org/10.1371/journal.pgen.1006563.

Zhang, Wentao, Bojing Du, Di Liu, and Xiaoting Qi. 2014. 'Splicing Factor SR34b Mutation Reduces Cadmium Tolerance in Arabidopsis by Regulating Iron-Regulated Transporter 1 Gene'. *Biochemical and Biophysical Research Communications* 455 (3–4): 312–17. https://doi.org/10.1016/j.bbrc.2014.11.017.

Zhang, Yi, Lian Liu, Qiongzi Qiu, Qing Zhou, Jinwang Ding, Yan Lu, and Pengyuan Liu. 2021. 'Alternative Polyadenylation: Methods, Mechanism, Function, and Role in Cancer'. *Journal of Experimental & Clinical Cancer Research* 40 (1): 51. https://doi.org/10.1186/s13046-021-01852-7.

Zhang, Zuo, and Gordon G. Carmichael. 2001. 'The Fate of DsRNA in the Nucleus'. *Cell* 106 (4): 465–76. https://doi.org/10.1016/S0092-8674(01)00466-4.

Zhou, Jianhong, Christopher J. Oldfield, Wenying Yan, Bairong Shen, and A.Keith Dunker. 2020. 'Identification of Intrinsic Disorder in Complexes from the Protein Data Bank'. *ACS Omega* 5 (29): 17883–91. https://doi.org/10.1021/acsomega.9b03927.

Zuckerkandl, Emile, and Linus Pauling. 1965. ‘Molecules as Documents of Evolutionary History’. *Journal of Theoretical Biology* 8 (2): 357–66. https://doi.org/10.1016/0022-5193(65)90083-4.

Supplementary materials

Supplementary materials for chapter 3: Adaptive evolution at mRNA editing sites in soft-bodied cephalopods.

Extrapolation

Unpolarized *R* values strongly depend on the EL threshold. Due to insufficient data, analogous, statistically significant observations for polarized *Q* values at EL thresholds higher than 0 could not be obtained; however, we can indirectly demonstrate that *Q* must vary at different EL levels. Consider two closely related organisms, "1" and "2", and their common ancestor, "anc". By definition given in the main text:

$$Q^{1\to 2} = \frac{R^{1\to 2}_{\to \mathrm{G}}}{R^{1\to 2}_{\to \mathrm{Y}}} = \frac{\dfrac{p(\mathrm{E}^1 \to \mathrm{G}^2)}{p(\mathrm{A}^1 \to \mathrm{G}^2)}}{\dfrac{p(\mathrm{E}^1 \to \mathrm{Y}^2)}{p(\mathrm{A}^1 \to \mathrm{Y}^2)}} =$$

$$= \frac{\left(p(\mathrm{E}^{\mathrm{anc}} \to \mathrm{G}^2) + p(\mathrm{G}^{\mathrm{anc}} \to \mathrm{E}^1)\right) \times \left(p(\mathrm{A}^{\mathrm{anc}} \to \mathrm{Y}^2) + p(\mathrm{Y}^{\mathrm{anc}} \to \mathrm{A}^1)\right)}{\left(p(\mathrm{A}^{\mathrm{anc}} \to \mathrm{G}^2) + p(\mathrm{G}^{\mathrm{anc}} \to \mathrm{A}^1)\right) \times \left(p(\mathrm{E}^{\mathrm{anc}} \to \mathrm{Y}^2) + p(\mathrm{Y}^{\mathrm{anc}} \to \mathrm{E}^1)\right)}$$

Thus, by definition, $Q^{1\to 2}$ is monotonic on both $Q^2_{\to *}$ and $Q^1_{*\to}$. Similarly, $Q^{2\to 1}$ is monotonic on both $Q^1_{\to *}$ and $Q^2_{*\to}$

As no other terms are present in the ratio of *R* values, we conclude that at least one of the directed *Q* values should increase at the increase of the *R* ratio, the latter being observed in the case of its dependence on the EL threshold. Hence at higher EL values, at least one directed *Q* value should increase.

Caveats

Underpredicted editing sites.

The procedure to identify editing sites employed by Liscovitch-Brauer et al. is based on the alignment of RNA and genomic reads to constructed transcripts and the analysis of mismatches (Liscovitch-Brauer et al. 2017). Editing sites in transcripts with low read coverage are likely to be missed, as the average editing level does not exceed 10% in all studied species. As we calculate the substitution matrices using substitutions that happened between the ancestral and descendant states, the former inferred from at least two species, this could result in underprediction of ancestral editing sites, which, in turn, would inflate the number of A-to-E transitions. However, for NES we observe the same behavior as for the majority of sites (Fig. 3), but for SES the effect is much weaker, even considering the fact that they are generally less conserved than NES, while if the underprediction had determined our results, we would observe the same picture for NES and SES.

Heterozygous editing sites.

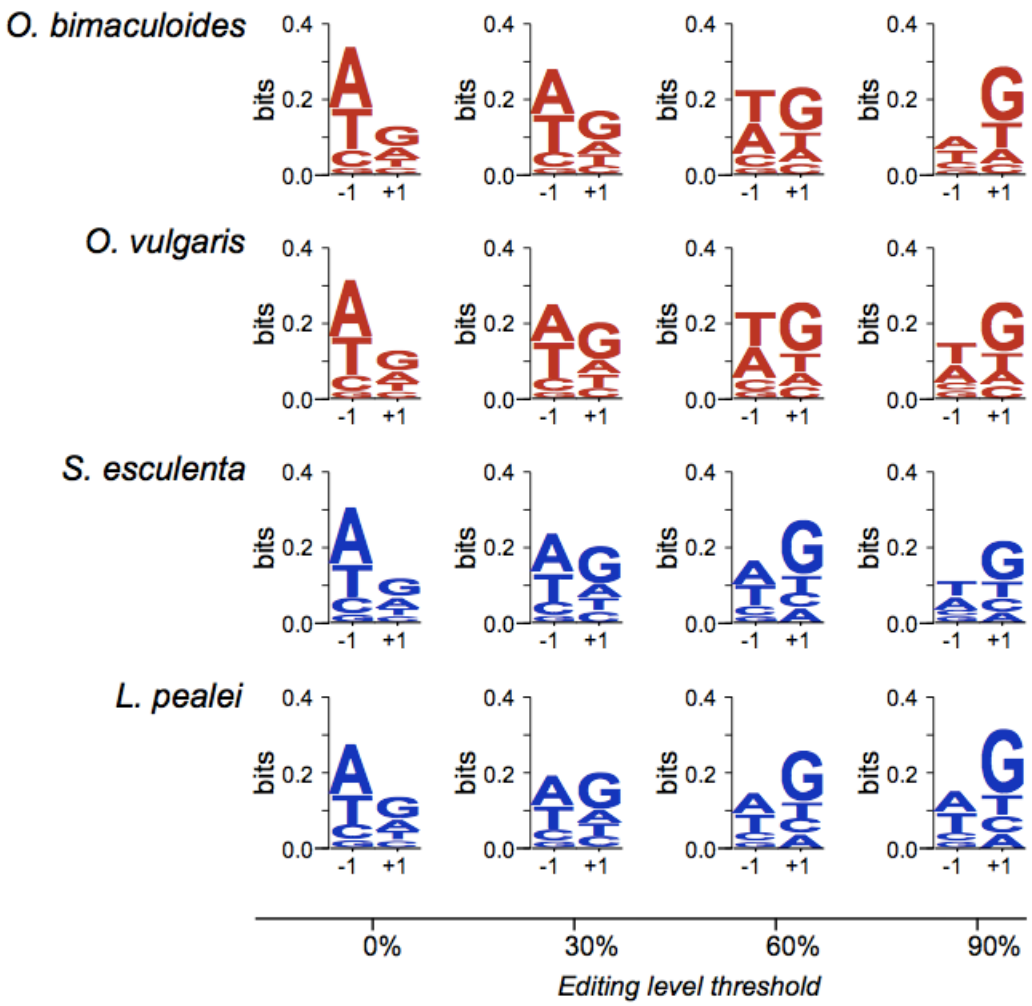
During construction of the editing-site list, heterozygous sites A-G were not considered (Liscovitch-Brauer et al. 2017), and hence some editing sites could not be predicted. However, heterozygous sites influenced the constructed transcriptomes, which by definition contained the allele prevailing in the reads (Liscovitch-Brauer et al. 2017). This could influence the results in two ways.

Firstly, consider a heterozygous A-G site in one species and a homozygous G-G site in another species. Clearly, there has been a G-to-A transition; at that, the procedures for the construction of the transcriptome and edited site list would generate the former transcript strictly with either G or unedited A. In the former case the substitution would not be counted, and in the latter case we would count a G-to-(unedited)A transition. If the adenine in the heterozygous state is edited, the above procedure would report either no transition of this adenine, or a transition involving an unedited adenine, hence

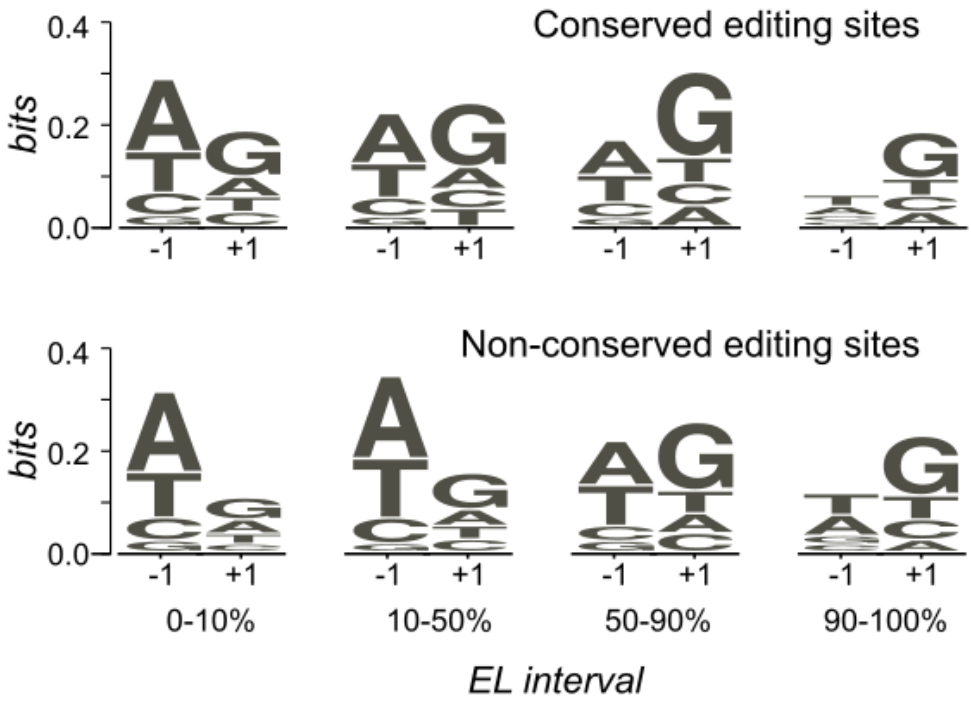
decreasing the real number of E-to-G. Similarly, in the case of A-G and A-A sites, the G-to-E substitutions would also be undercounted.

Secondly, the lack of data about heterozygous editing sites would result in a general underprediction of A-to-I editing sites discussed above. Thus, this underprediction of E-to-G and G-to-E transitions could influence the calculated R_G values, but it would act against the reported effects, as it lowers the $p(E \rightarrow \mathrm{G})$ values.

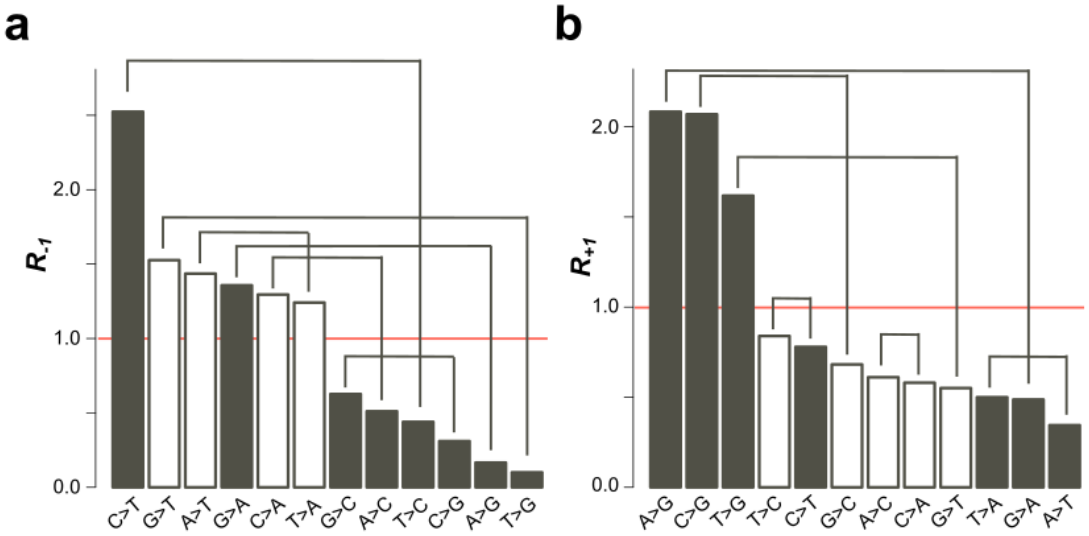
Supplementary Figures



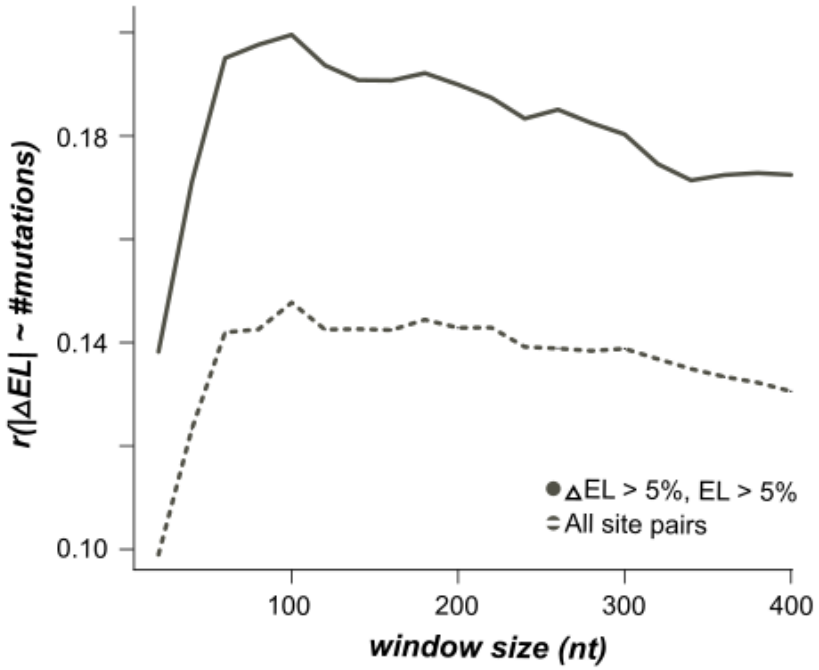
Supplementary Figure S1 | LOGOs of nucleotides adjacent to editing sites in all studied organisms at different editing levels.



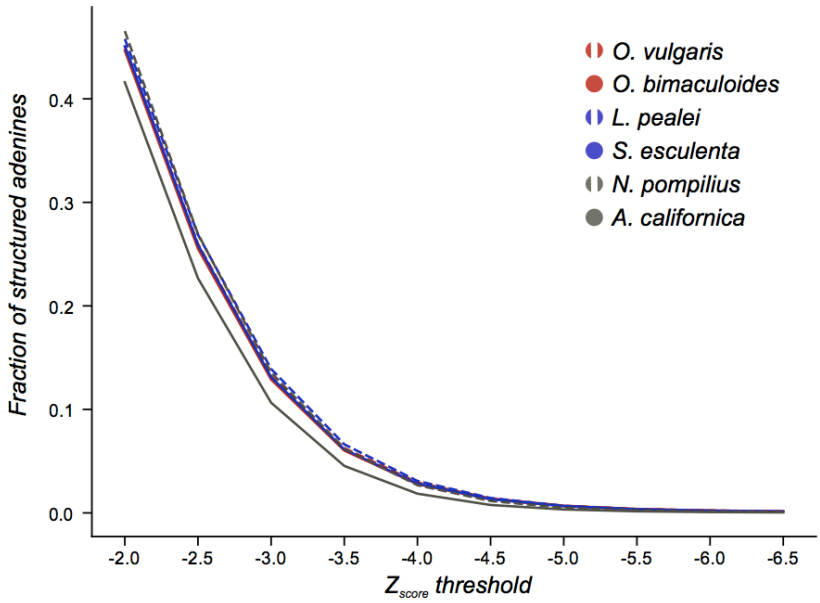
Supplementary Figure S2 | LOGOs of nucleotides adjacent to conserved and non-conserved editing sites in the squid-cuttlefish pair.



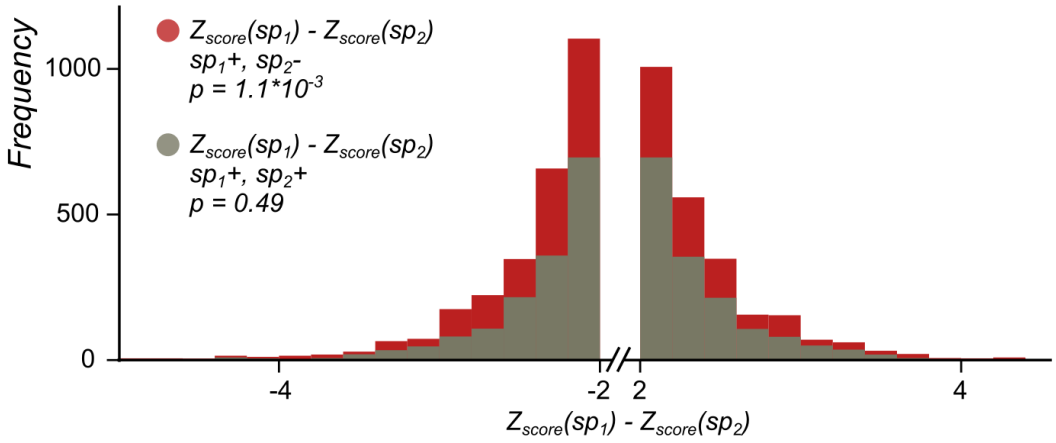
Supplementary Figure S3 | Over- and underrepresented mismatch *R* values in the local context of non-conserved editing sites in the pair of *Octopus* species. *R* values are defined for the ±1 positions relative to transcriptomic adenines as the ratio of the probability of a given substitution near the edited adenine and the respecteve probability for the non-edited adenine. **(a)** In position –1 relative to NCES sites. **(b)** In position +1 relative to NCES sites. Lines represent mutually reversed substitutions. White bars represent *R* values which do not significantly differ from 1 ($p > 0.05$) and grey bars represent statistically significant *R* values.



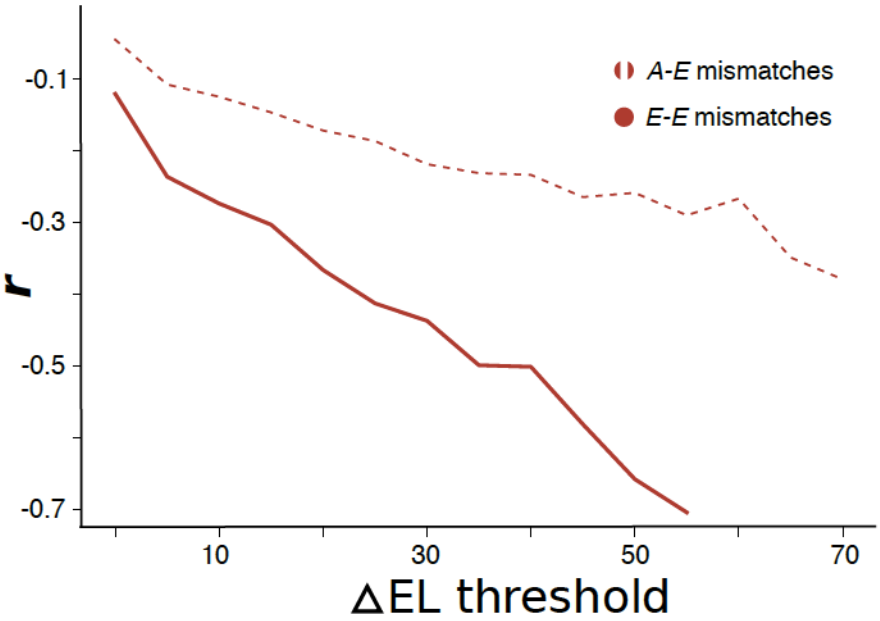
Supplementary Figure S4 | Pearson's correlation between the absolute value of difference of ELs in homologous sites and the number of mismatches in a window of a given size for different window sizes. The dashed line represents values obtained for all homologous site pairs; the solid line represents values obtained for pairs of homologous edited adenines with EL above 5% and with the absolute value of EL difference greater than 5%.



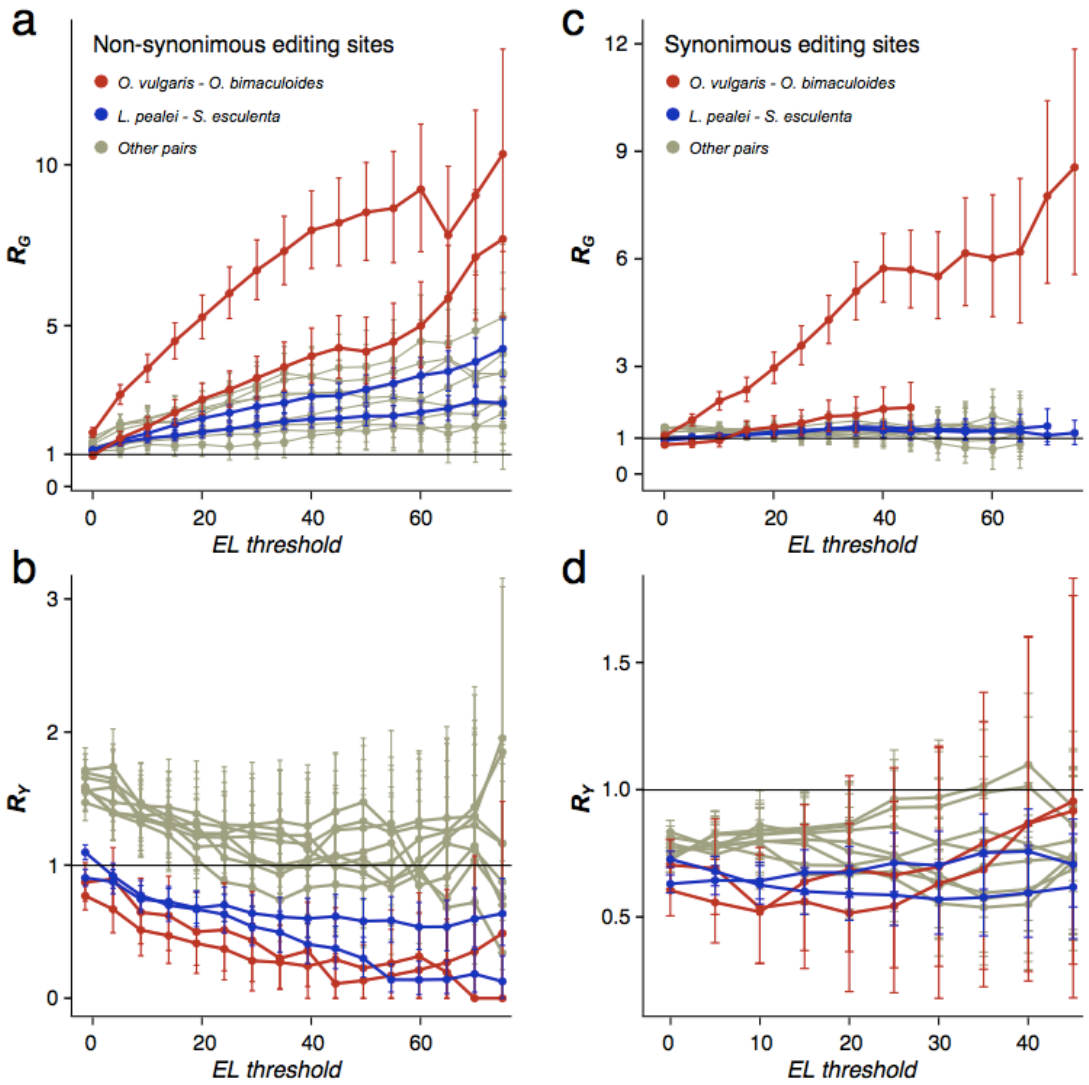
Supplementary Figure S5 | *A. californica* has less adenines in structured regions then other species. The classification of structured and unstructured segments was performed with various thresholds (see Methods). The grey line corresponding to *A. californica* is considerably lower than cephalopod lines.



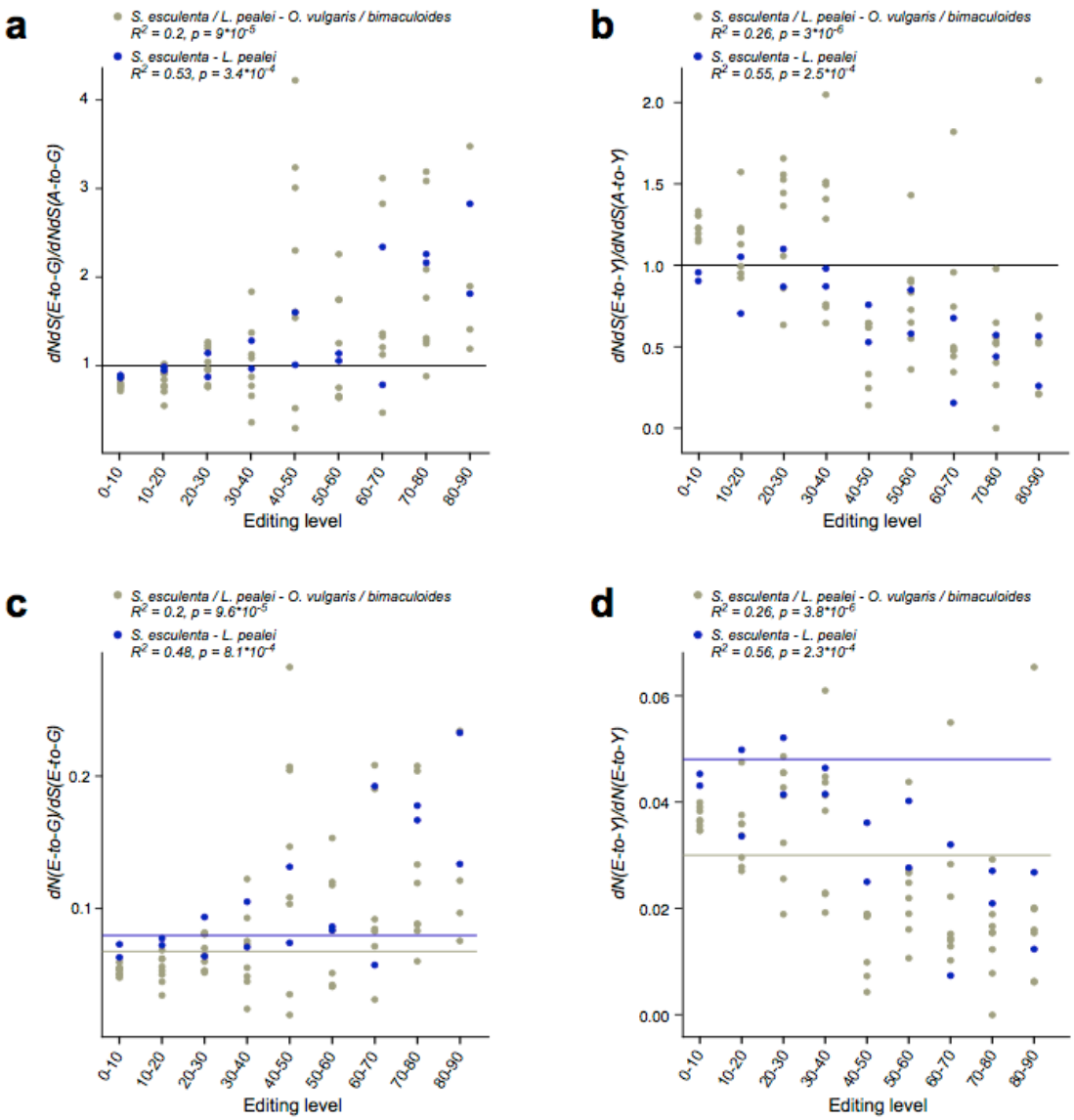
Supplementary Figure S6 | The local secondary structure is more stable at edited adenines than at homologous, non-edited adenines. The distribution of the difference of the minimal free energy Z-score between homologous sites in two octopuses, *O. vulgaris* and *O. bimaculoides* is shown in red when two homologous sites have different editing status (edited minus unedited) and in grey when both sites in a pair are edited. The left tail of the red histogram is heavier than the right one ($p = 6.03\times10^{-17}$ versus 0.57 for the grey histogram), showing that the editing sites tend to regions with higher secondary structure stability.



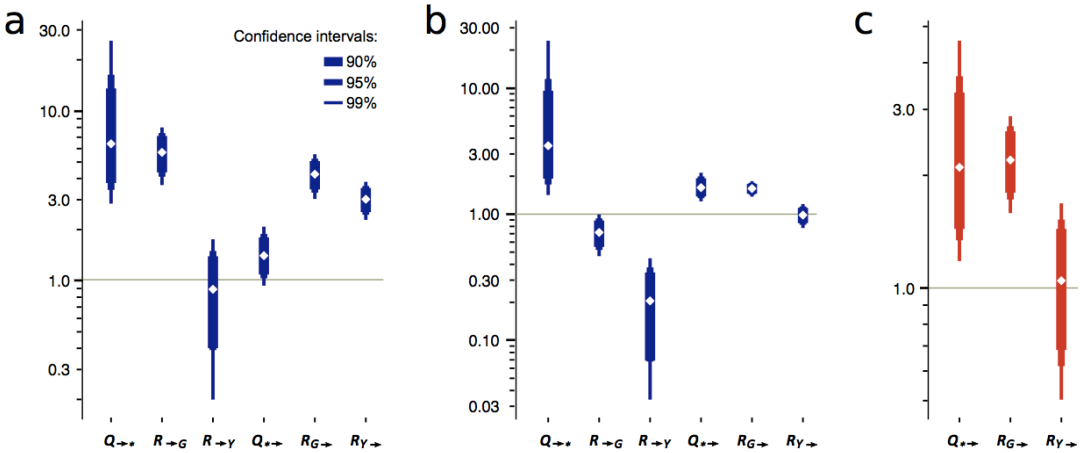
Supplementary Figure S7 | Dependence of Pearson's correlation between the difference of structural Z-scores and the difference in ELs of homologous sites on the minimal considered difference in ELs. The dashed line represents values obtained for pairs on homologous adenines, such that one adenine in a pair is edited, and the other is not. The EL of all unedited adenines was set to 0. The solid line represents values obtained for pairs of homologous edited adenines. Only correlation coefficients with p-values below 0.05 are shown.



Supplementary Figure S8 | Dependence of R_G and R_Y on the editing level considered separately for non-synonymous (ab) and synonymous (cd) sites. Two curves for each pair are given, since R_N is a reference-based measure. The red curves correspond to the *Octopus* pair, the blue curves, to the pair cuttlefish–squid, the grey curves, to distant pairs.



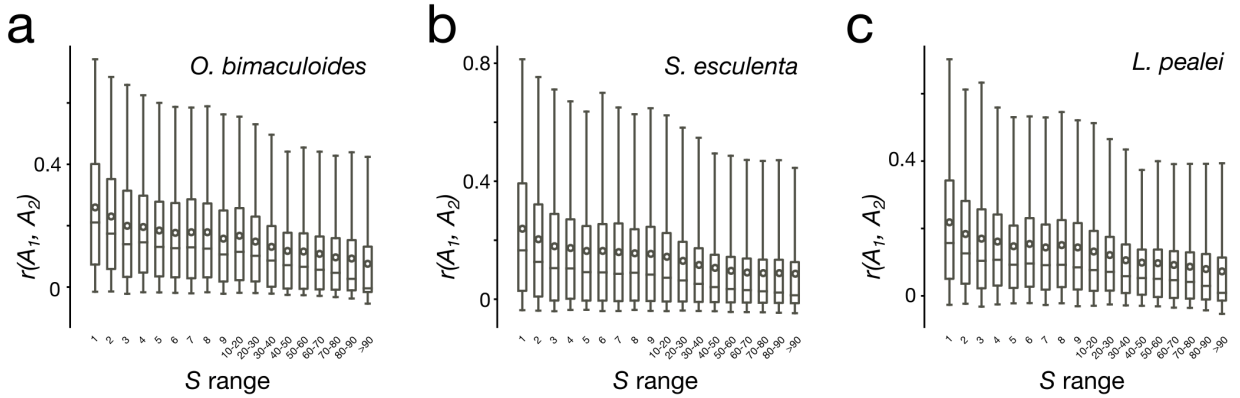
Supplementary Figure S9 | Normalized *dN/dS* of editing site substitutions. *dN/dS* of editing site substitutions normalized by respective *dN/dS* of unedited adenines to G **(a)** and to pyrimidines **(b)**. *dN/dS* of editing sites normalized by the respective ξ^{non}/ξ^{syn} value ratio (Supplementary Table 1) for substitutions to G **(c)** and to pyrimidines **(d)**. Horizontal lines represent average *dN/dS* values for unedited adenines. R^2 values and *p*-values calculated from the *F* statistic are shown.



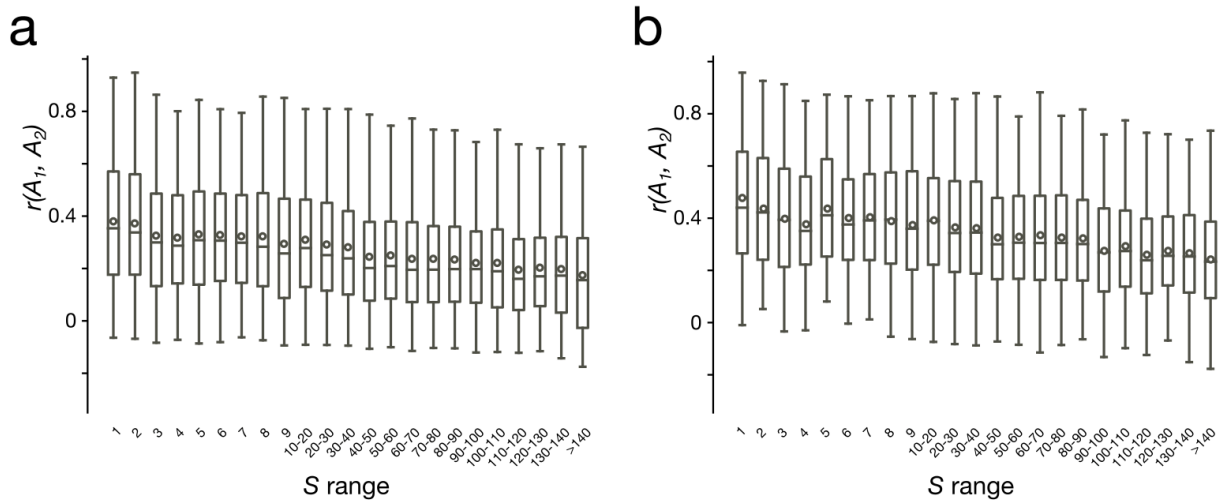
Supplementary Figure S10 | Mutational characteristics of editing sites. (a,b) Mutational characteristics of edited sites for the squid–cuttlefish pair separately for non-synonymous (NES) **(a)** and synonymous (SES) **(b)** sites. **(c)** $Q_{*\rightarrow}$ and $R_{*\rightarrow}$ values calculated for the octopus total substitution matrix; there are insufficient data for the separate analysis of NES and SES.

Supplementary materials for chapter 4: A hierarchy in clusters of cephalopod mRNA editing sites.

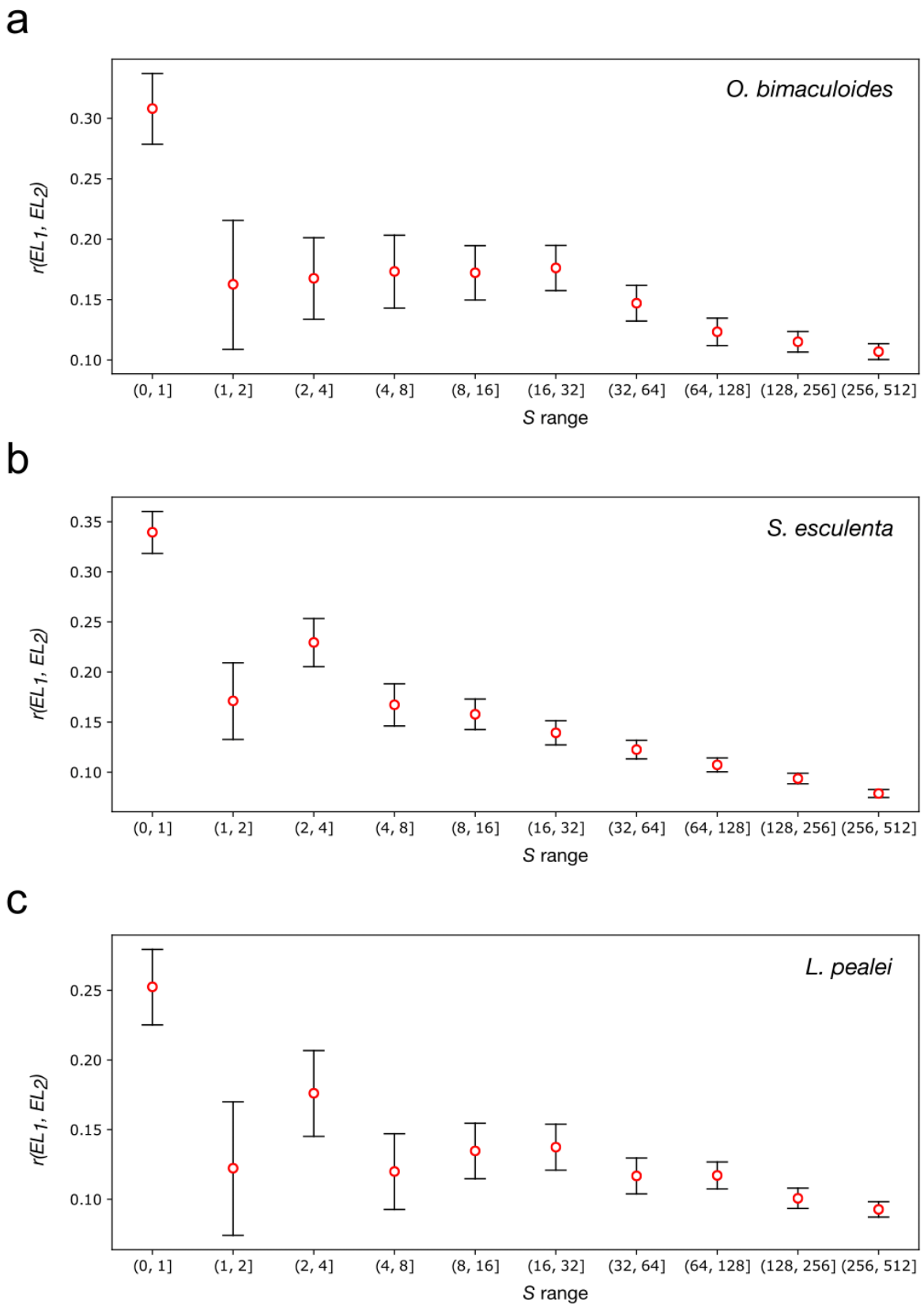
Supplementary Figures



Supplementary Figure S11 | The distributions of correlation coefficients of coleoid editing at two sites with respect to the distances between sites. Boxes represent the quartile borders; red circles represent the means; the grey lines indicate the 95% two-sided confidence intervals of the distributions. **(a)** *O. bimaculoides* **(b)** *S. esculenta* **(c)** *L. pealei*. Notation as in Fig. 2a.

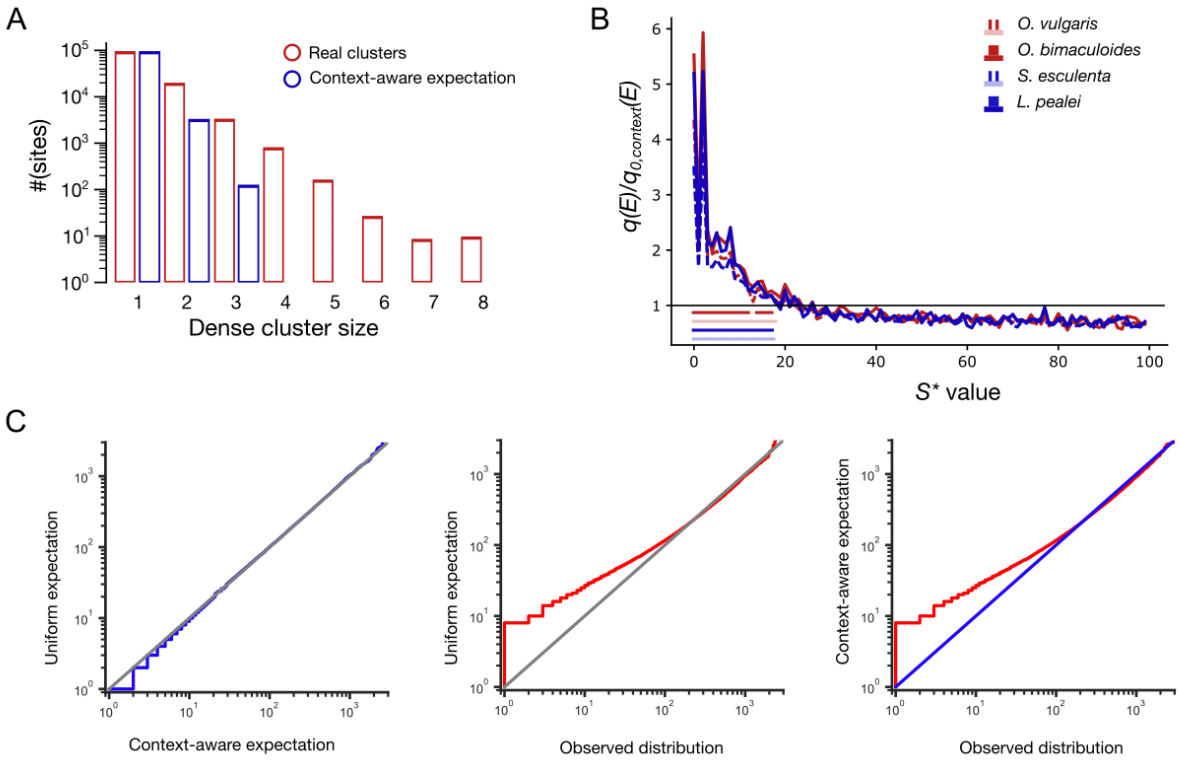


Supplementary Figure S12 | The distributions of correlation coefficients of *O. vulgaris* editing at two sites with respect to the distances between sites for different minimal editing level threshold values. Boxes represent the quartile borders; red circles represent the means; and the grey lines indicate the 95% two-sided confidence intervals of the distributions. **(a)** Threshold set to 5% **(b)** Threshold set 10%. Notation as in Fig. 2a, Suppl. Fig. S1.

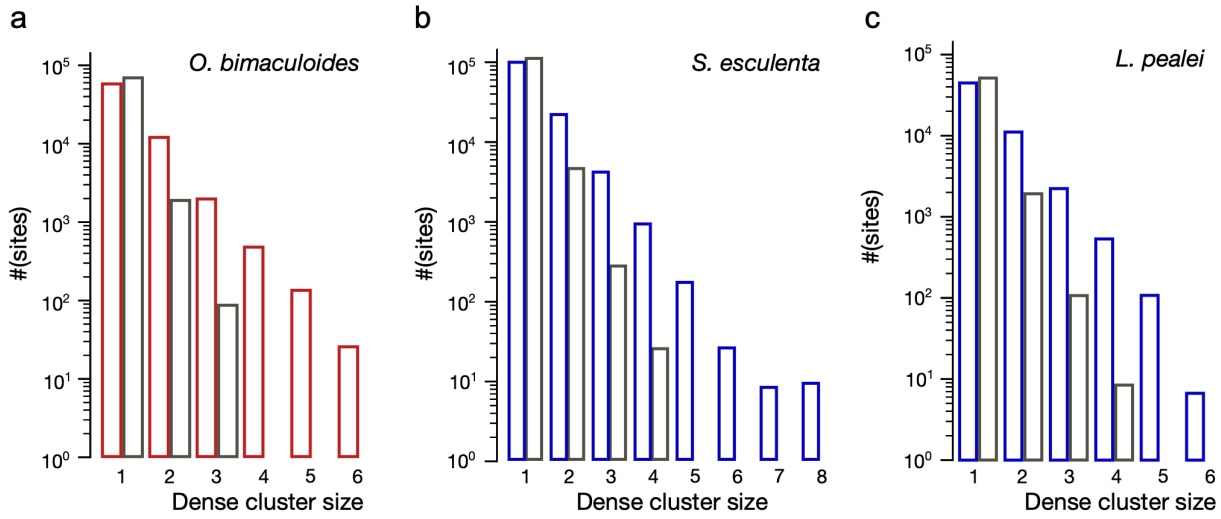


Supplementary Figure S13 | The dependence of the correlations of ELs on the *S* values for the considered coleoid A-to-I editing site datasets. The red circles mark the values of the correlation coefficients, and the grey lines represent the Bonferroni corrected 95% two-sided confidence intervals

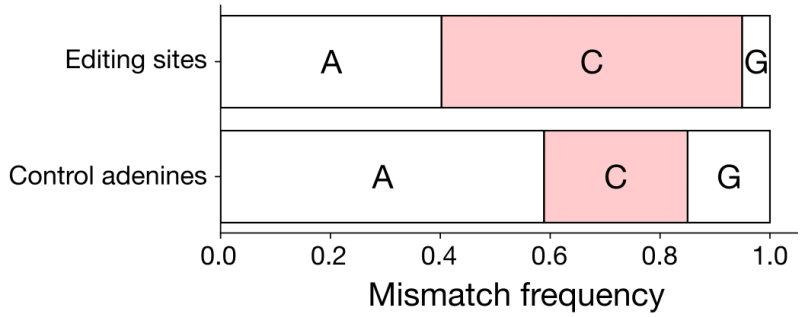
obtained from the *t*-distribution. **(a)** *O. bimaculoides* **(b)** *S. esculenta* **(c)** *L. pealei*. Notation as in Fig. 2b.



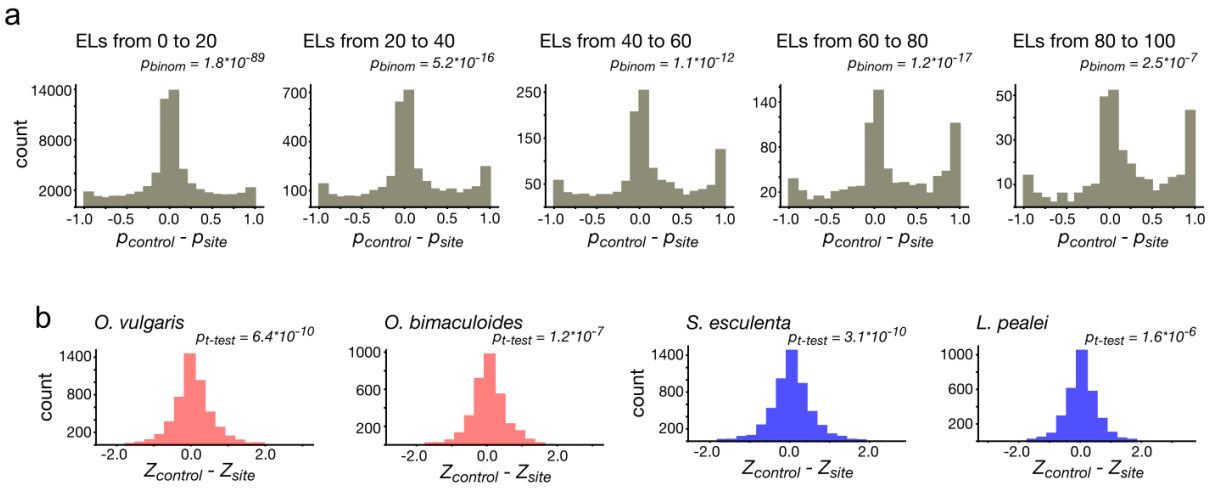
Supplementary Figure S14 | Results obtained with the context-aware expectation of the distribution of edited adenine. (A) Histogram of dense cluster sizes (nt) for the real *O. vulgaris* editing site dataset (red) and the matching random context-aware dataset (blue). **(B).** Deviation of the editing probabilities of adenines located near editing sites ($q(E)$) from the respective expected probabilities ($q_0(E)$) as dependent on the S^* values. Notation as in Fig. 5A. **(C)** Comparisons of all three distributions.



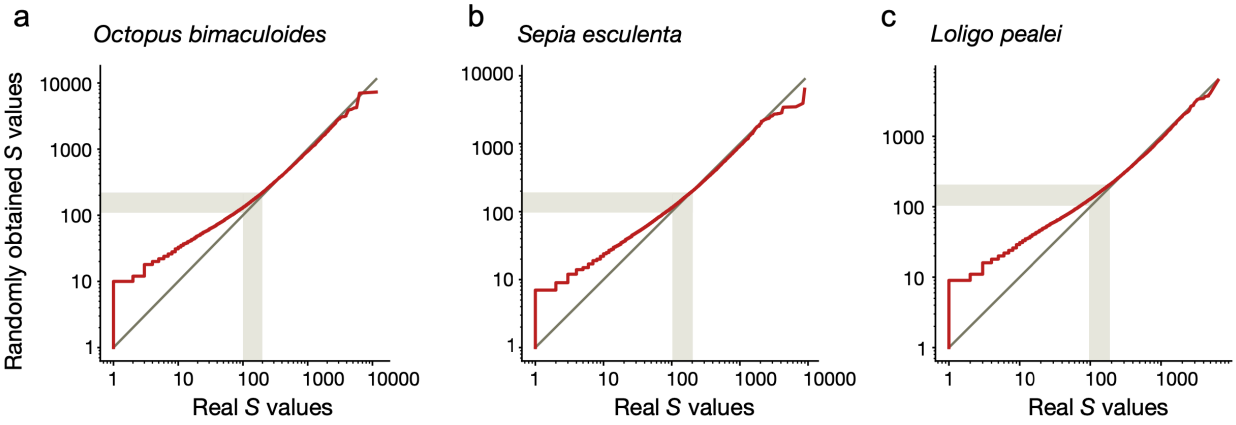
Supplementary Figure S15 | Histograms of dense cluster sizes (nt) for the real coleoid editing site datasets (red and blue) and the corresponding randomly obtained ones (grey). (a) *O. bimaculoides* **(b)** *S. esculenta* **(c)** *L. pealei*. Notation as in Fig. 3A.



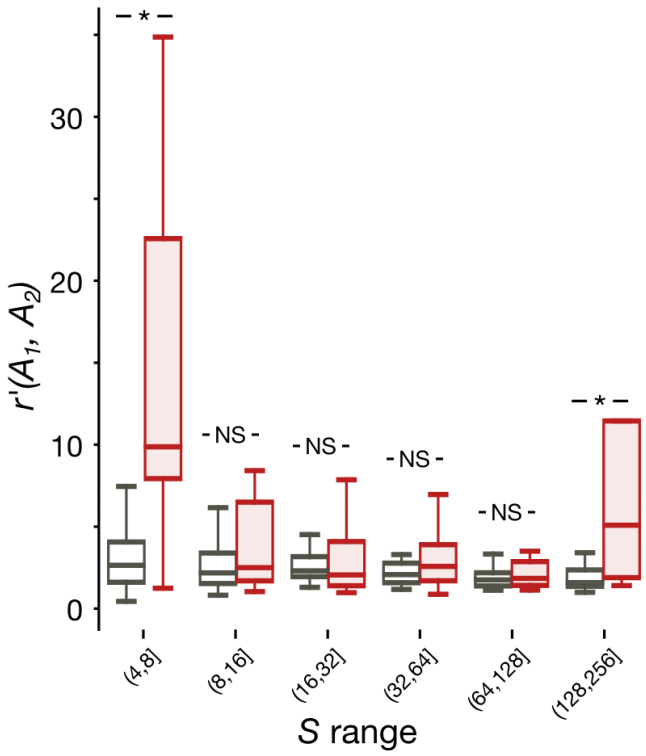
Supplementary Figure S16 | Distributions of mismatches of adenines in double RNA helices. See Supplementary Methods for details.



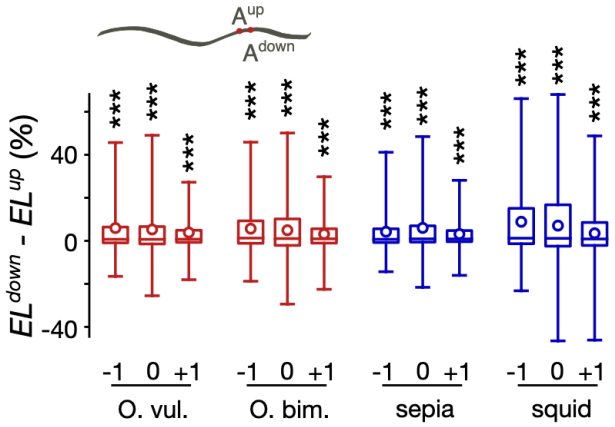
Supplementary Figure S17| Structural properties of coleoid A-to-I editing sites. (a) Differences in base pairing probabilities between A-to-I editing sites and nearest unedited adenines used as controls for different editing level (EL) intervals in *O. vulgaris*. The significance (*p*-values, the binomial test) of the site–control differences is shown. **(b)** Differences in RNA secondary structure free energy between editing sites and control adenines in four coleoid species. The significance (*p*-values, the *t*-test) of the control sites structural potentials being larger than those of editing sites is shown.



Supplementary Figure S18 | Clustering of A-to-I editing sites in coleoid transcriptomes. Red lines show the dependencies between the sorted real and the randomly obtained *S* value sets. Grey lines represent the expected dependence of the form $y=x$. Grey stripes represent the predicted borders of the regions affecting editing sites. **(a)** *O. bimaculoides* **(b)** *S. esculenta* **(c)** *L. pealei*. Notation as in Fig. 5A.



Supplementary Figure S19 | Distributions of the $r'(A_i, A_j)$ values calculated for the structurally close editing sites (red boxes) and for the control site pairs with no predicted secondary RNA structure between the sites in a pair (grey boxes), where both sites A_i and A_j are located in the same exon of *O. bimaculoides*. Notation as in Fig. 5B.



Supplementary Figure S20 | Distributions of differences in ELs between down- and upstream editing site in two-adenine dense clusters obtained for three reading frames. Three asterisks mark statistical significance of the differences in means ($p < 0.001$, Chi-squared contingency test). The values -1, 0 and +1 below indicate the coordinate of the dense cluster reading frame relative to the predicted protein-coding frames in the coleoid transcriptomes.

Supplementary Tables

Supplementary table S1 | SRA identifiers of RNAseq libraries employed in the present study

SRA ID	Organism	Bioproject	#G bases
SRR2045866	*O. bimaculoides*	PRJNA285380	5.6
SRR2045870	*O. bimaculoides*	PRJNA285380	7.4
SRR2047107	*O. bimaculoides*	PRJNA285380	7.2
SRR2047109	*O. bimaculoides*	PRJNA285380	7.9
SRR2047111	*O. bimaculoides*	PRJNA285380	7.1
SRR2047114	*O. bimaculoides*	PRJNA285380	6.8
SRR2047116	*O. bimaculoides*	PRJNA285380	6.4
SRR2047118	*O. bimaculoides*	PRJNA285380	6.1
SRR2047120	*O. bimaculoides*	PRJNA285380	6.3
SRR2047122	*O. bimaculoides*	PRJNA285380	7.2
SRR2048495	*O. bimaculoides*	PRJNA285380	3.2
SRR2048496	*O. bimaculoides*	PRJNA285380	3.2
SRR2048497	*O. bimaculoides*	PRJNA285380	3.2
SRR2048498	*O. bimaculoides*	PRJNA285380	3.2
SRR2048521	*O. bimaculoides*	PRJNA285380	3.2
SRR2048522	*O. bimaculoides*	PRJNA285380	3.2
SRR2048523	*O. bimaculoides*	PRJNA285380	3.2
SRR2048524	*O. bimaculoides*	PRJNA285380	3.2
SRR2048525	*O. bimaculoides*	PRJNA285380	3.2
SRR2857272	*O. vulgaris*	PRJNA299756	38.2
SRR2857274	*O. vulgaris*	PRJNA299756	38.4
SRR2855904	*S. esculenta*	PRJNA299756	39.6
SRR2856422	*S. esculenta*	PRJNA299756	48.3
SRR1522987	*L. pealei*	PRJNA255916	17.5
SRR1522988	*L. pealei*	PRJNA255916	17.5
SRR1725163	*L. pealei*	PRJNA255916	6.9
SRR1725164	*L. pealei*	PRJNA255916	6.9
SRR1725167	*L. pealei*	PRJNA255916	5.6
SRR1725169	*L. pealei*	PRJNA255916	7.7
SRR1725171	*L. pealei*	PRJNA255916	5.6

SRR1725172	*L. pealei*	PRJNA255916	4.1
SRR1725213	*L. pealei*	PRJNA255916	10.8
SRR1725235	*L. pealei*	PRJNA255916	12.4
SRR1725236	*L. pealei*	PRJNA255916	8.3

Supplementary Table S2 | Variance in the transcriptome and proteome explained by editing and by correlations in editing events

	O. vul.	O. bim.	sepia	squid
nucleotide SD due to editing	97.3027	82.2	110.8405	94.5532
% nucleotide variance explained by clustering	40.34	27.69	46.28	31.59
amino acid SD due to editing	80.6424	68.39	92.008	78.88
% amino acid variance explained by clustering	40.69	28.46	46.5	32.34
% nucleotide variance explained by DCs	4.3	3.6	4.038	4.057
% amino acid variance explained by DCs	4.48	3.7	4.168	4.175

Supplementary Table S3 | Effect sizes and confidence intervals for the data presented on Fig. 10C. Red color of a letter indicates the nucleotide in a dinucleotide, for which base-pairing probabilities are considered. $p_1 - p_2$ represents the mean difference in base pairing probabilities.

Comparison	p_1 - p_2	Wilcoxon p (Bonferroni corrected)
EE vs AA	-0.0268	1.0×10^{-6}
EE vs AA	0.11917	2.8×10^{-72}
EE vs AA	-0.1399	3.8×10^{-109}
EE vs AA	0.00603	0.8475
EE vs AE	-0.0776	4.0×10^{-16}
EE vs AE	0.03575	0.0011
EE vs AE	-0.1826	1.1×10^{-70}
EE vs AE	-0.0693	2.7×10^{-13}
EE vs EA	0.07293	8.4×10^{-6}
EE vs EA	0.17896	4.5×10^{-24}
EE vs EA	-0.0519	0.0022
EE vs EA	0.05410	0.0017

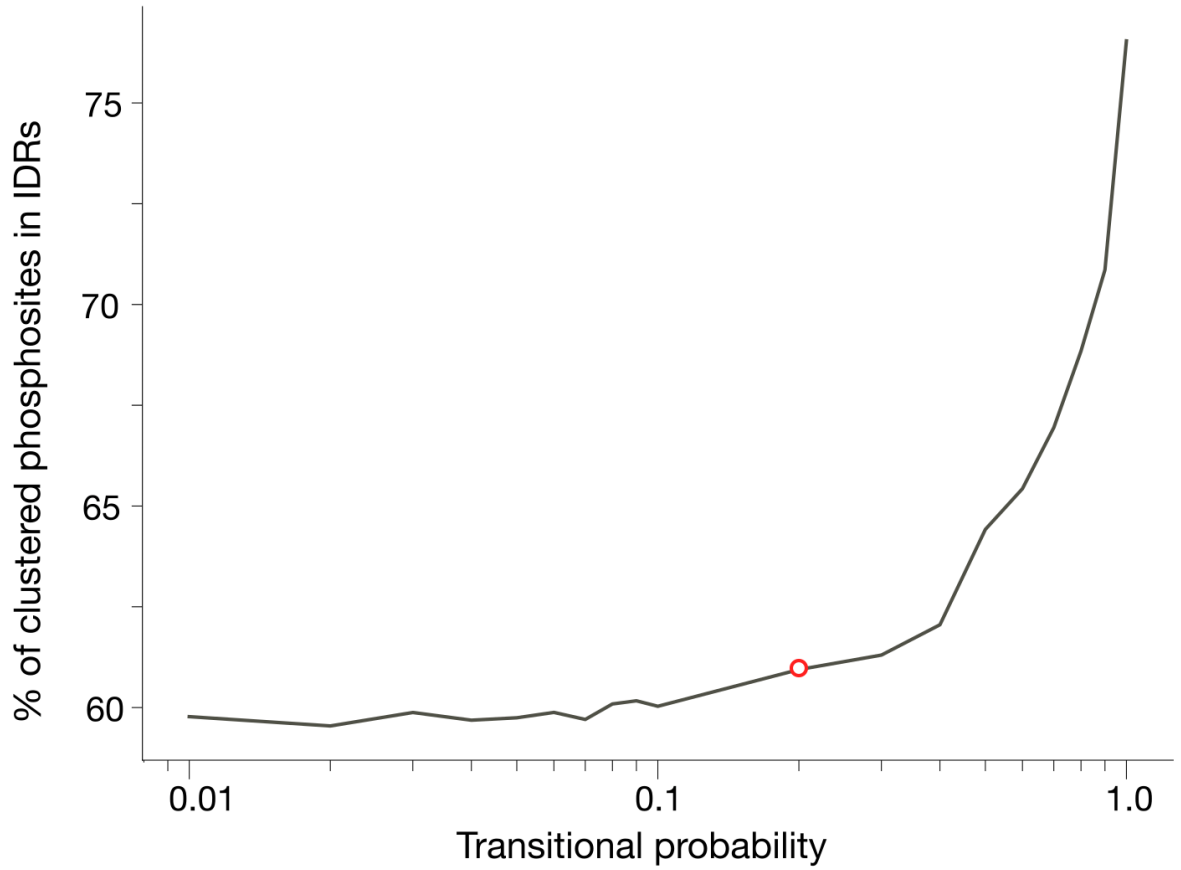
Supplementary Table S4 | 95% Confidence intervals of differences in base-pairing probabilities between paired editing sites (EE) and three types of control AA-dinucleotides obtained by random sampling. The color code is as in Table S3.

position \ site pair	AA - EE	AE - EE	EA - EE
1	0.0139 to 0.0396	0.0573 to 0.0968	-0.1057 to -0.0399

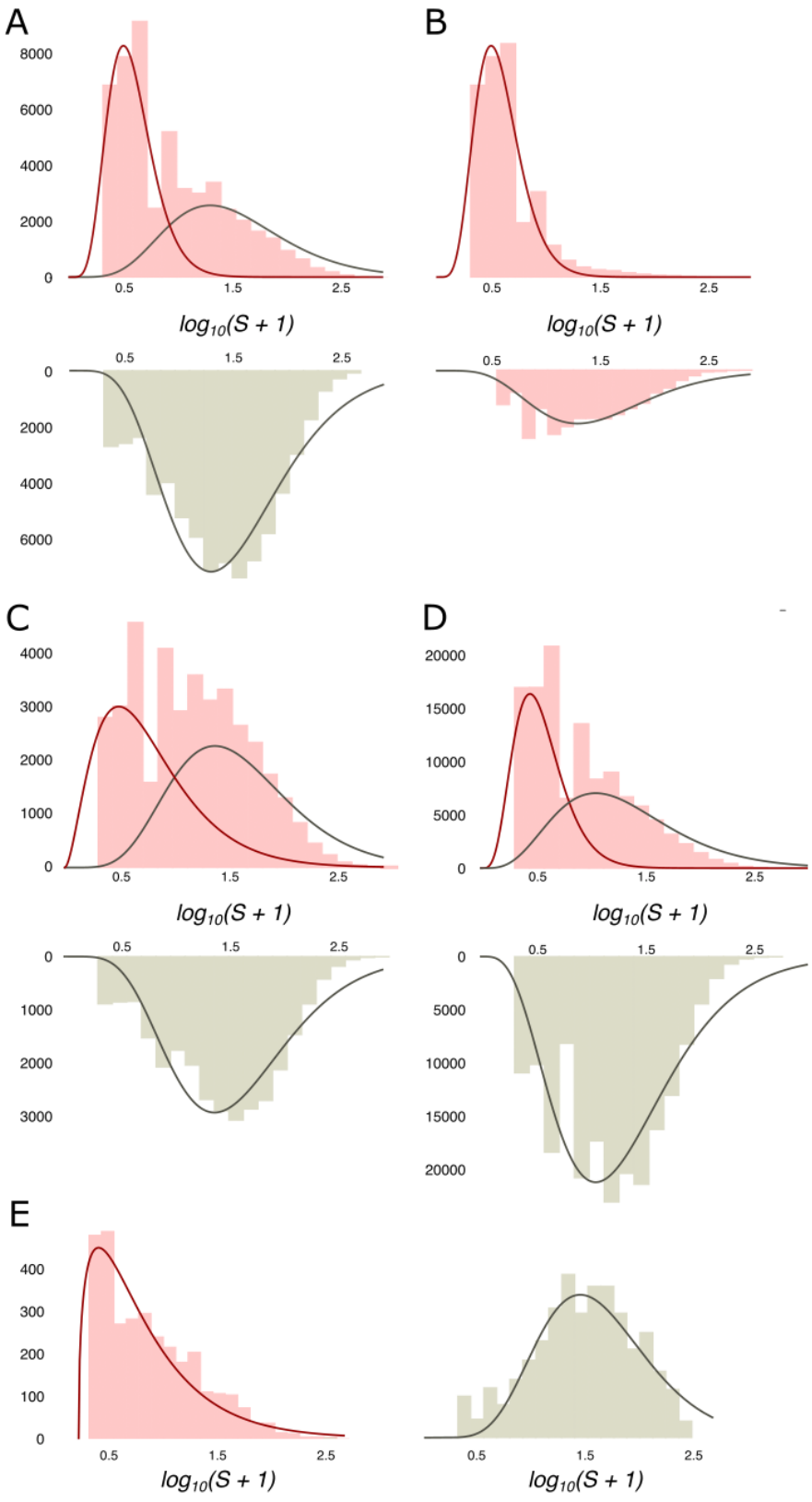
2	-0.0187 to 0.0072	0.0494 to 0.0896	-0.0881 to -0.02

Supplementary materials for chapter 5: Phospho-islands and the evolution of phosphorylated amino acids in mammals.

Supplementary Figures

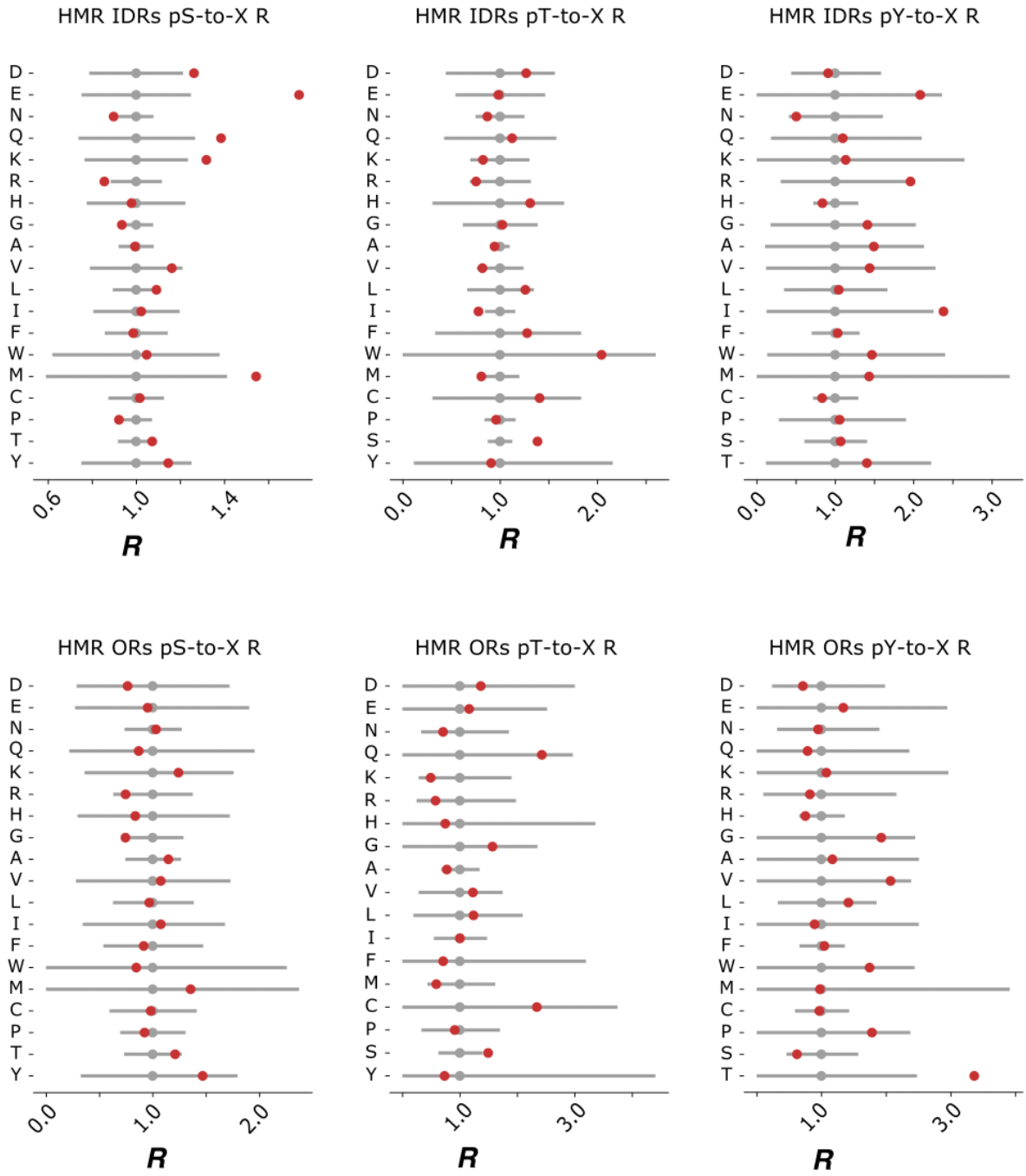


Supplementary Figure S21 | Stability of hidden Markov model clustered editing site predictions with respect to transitional probability values. Red circle represents the chosen value 0.2.



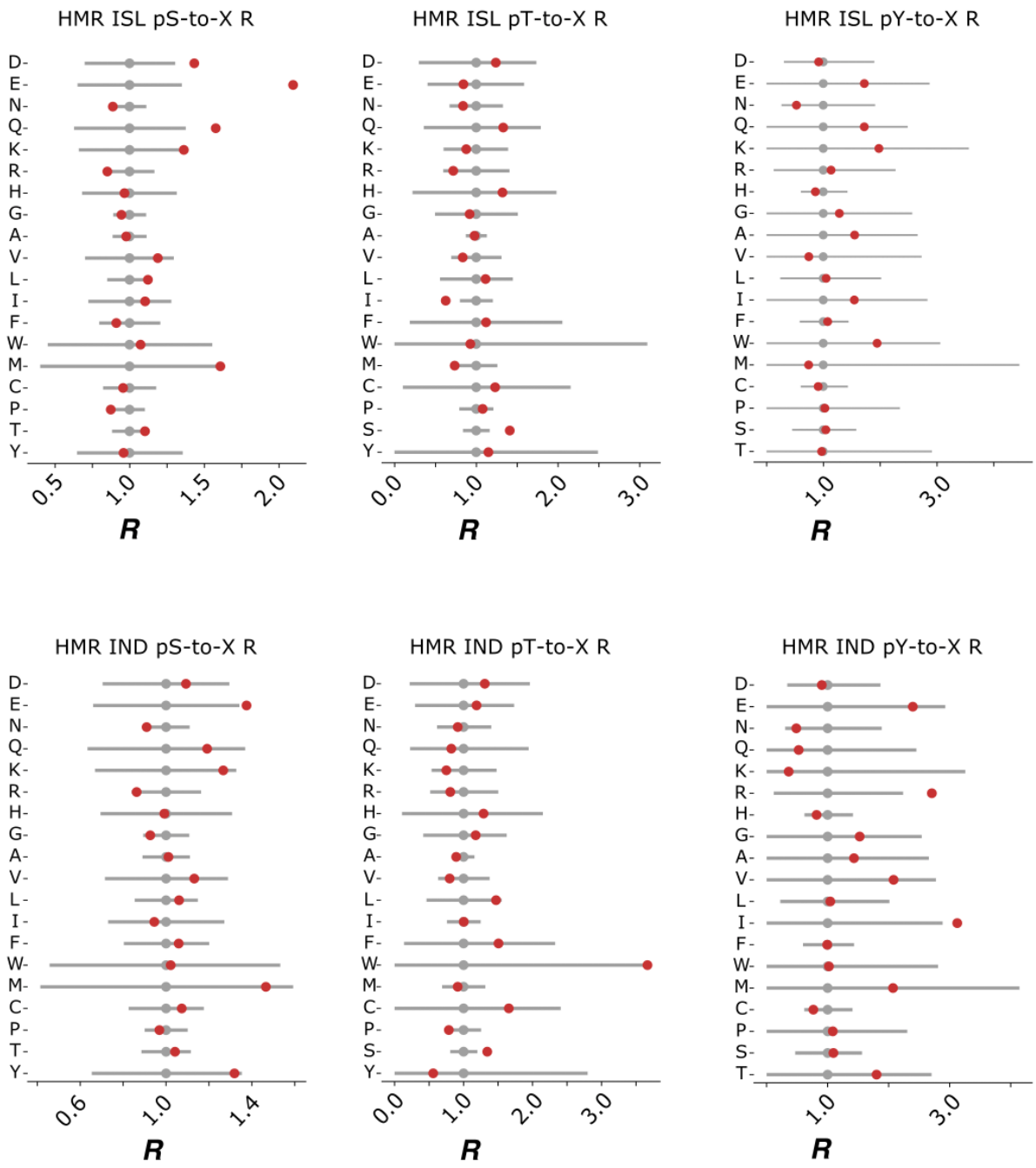
Supplementary Figure S22 | Phospho-island analysis in various datasets. (A, B) Prediction of phospho-islands for the mouse dataset. **(A)** (Above) The distribution of $\log_{10}(S + 1)$ values (pink histogram; see the main text for definitions) and its decomposition in two gamma distributions, for phospho-islands (red curve) and for individual phosphosites (red curve). (Below) $\log_{10}(S + 1)$ values for non-phosphorylated STY amino acids randomly sampled from IDRs with the same sample size and amino acid content as in the HMR dataset. **(B)** (Above) The distribution of $\log_{10}(S + 1)$ values for

phosphosites predicted to be in phospho-islands. (Below) $\log_{10}(S + 1)$ values for predicted individual phosphosites. **(C)** Histograms of $\log_{10}(S + 1)$ values for real (pink) and randomly assigned (grey) phosphosites with respective gamma-approximations for the rarefied human dataset. **(D)** Histograms of S values for real (pink) and randomly assigned (grey) phosphosites with respective gamma-approximations for the human dataset. **(E)** Distributions of the $\log_{10}(S + 1)$ values of phosphosites located in ORs (pink) and the distribution of $\log_{10}(S + 1)$ for STY amino acids randomly sampled from ORs (grey) with respective gamma-approximation. Note the good fit of a single gamma-distribution for the $\log_{10}(S + 1)$ values observed in ORs.

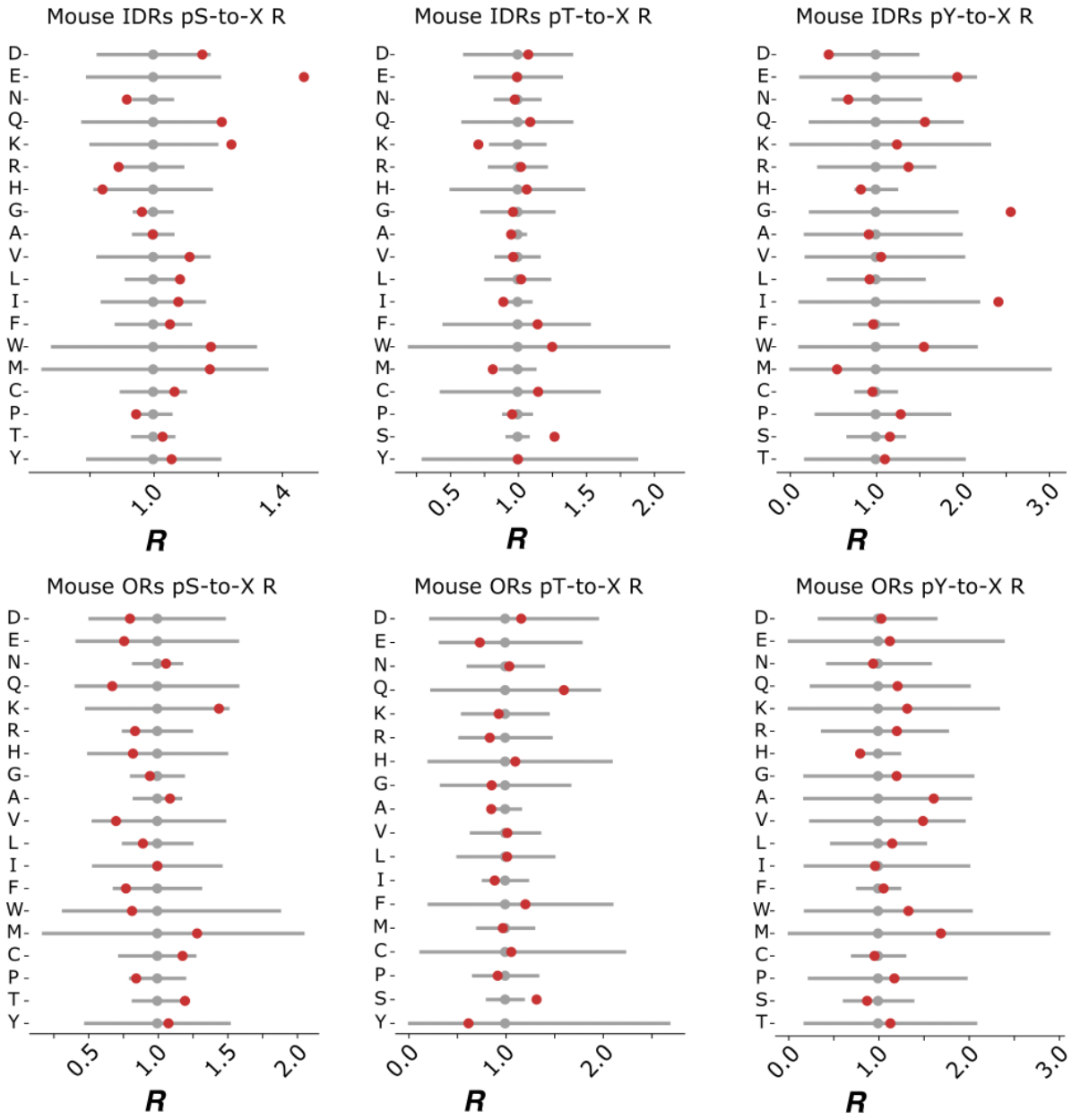


Supplementary Figure S23 | Comparison of the mutational patterns observed in HMR phosphosites located in intrinsically disordered and ordered regions. *R* values are defined in the

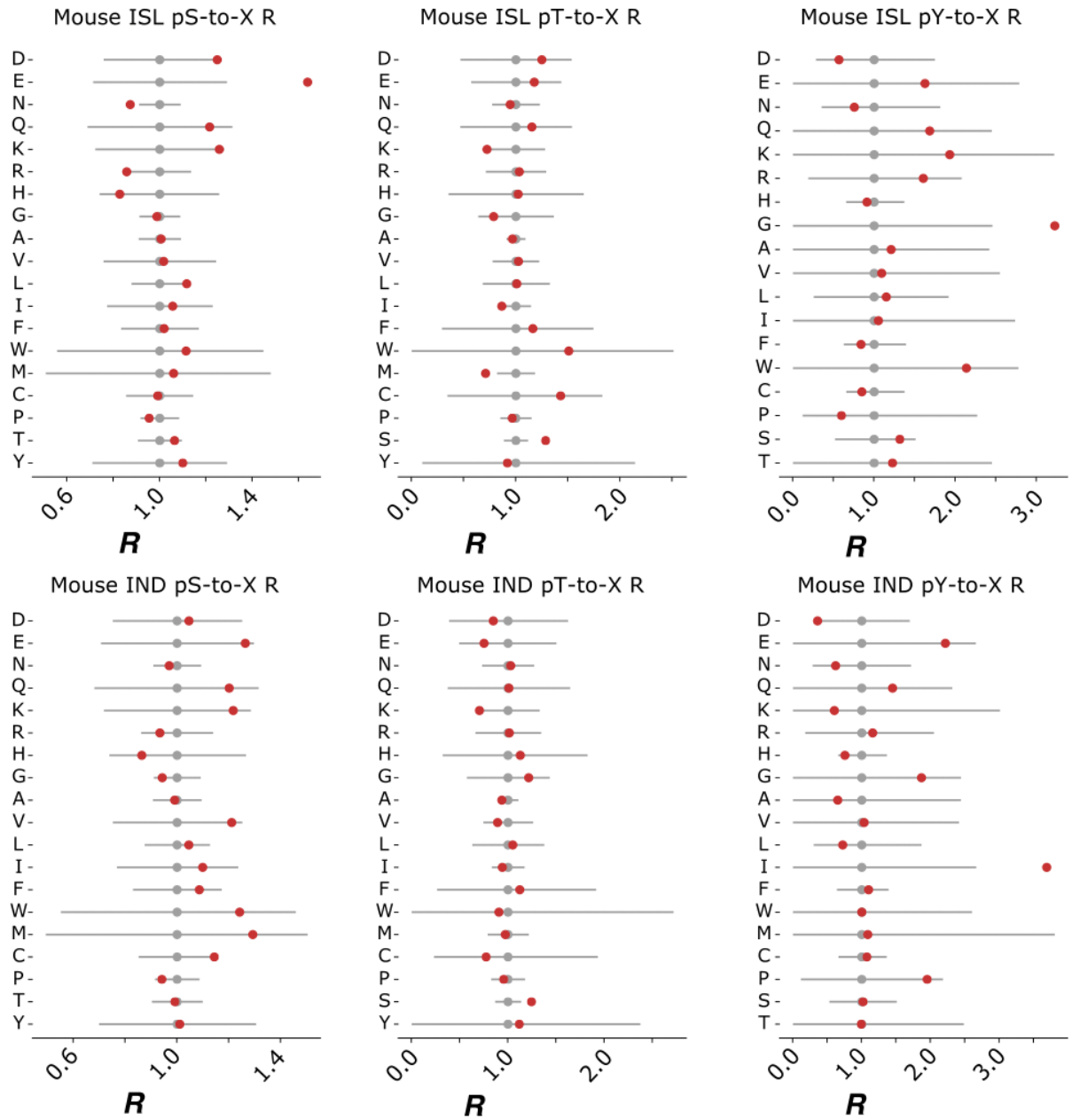
main text. 95% two-tail confidence intervals are generated from the chi-squared distribution with the Bonferroni correction. Red dots mark the actual *R* values. Grey dots represent the respective *R* values of non-phosphorylated STY amino acids.



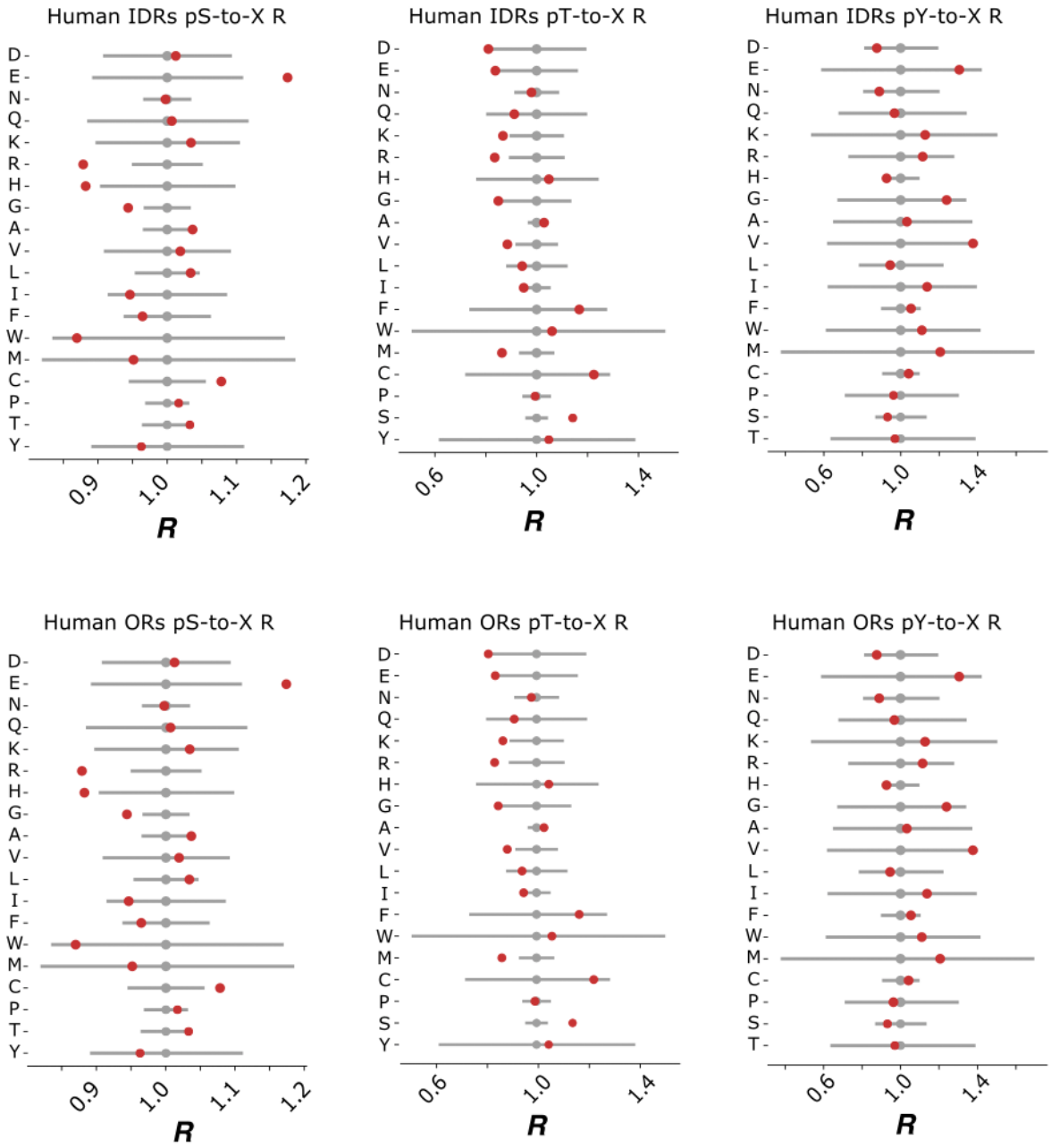
Supplementary Figure S24 | Comparison of the mutational patterns observed in HMR phosphosites located in phospho-islands *vs*. individual ones. Notation as in Suppl. Fig. S2.



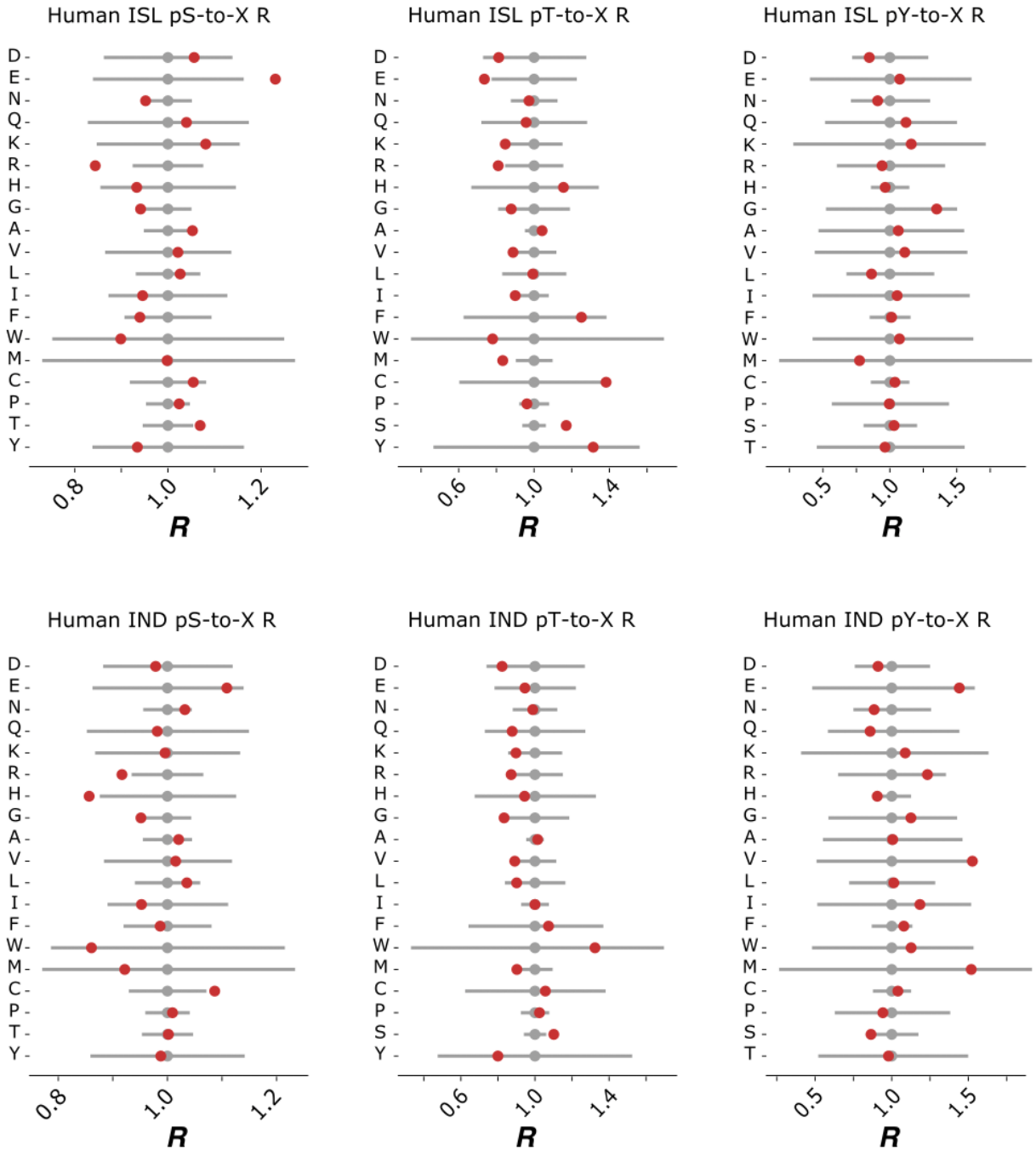
Supplementary Figure S25 | Comparison of the mutational patterns observed in mouse phosphosites located in intrinsically disordered and ordered regions. Notation as in Suppl. Fig. S2.



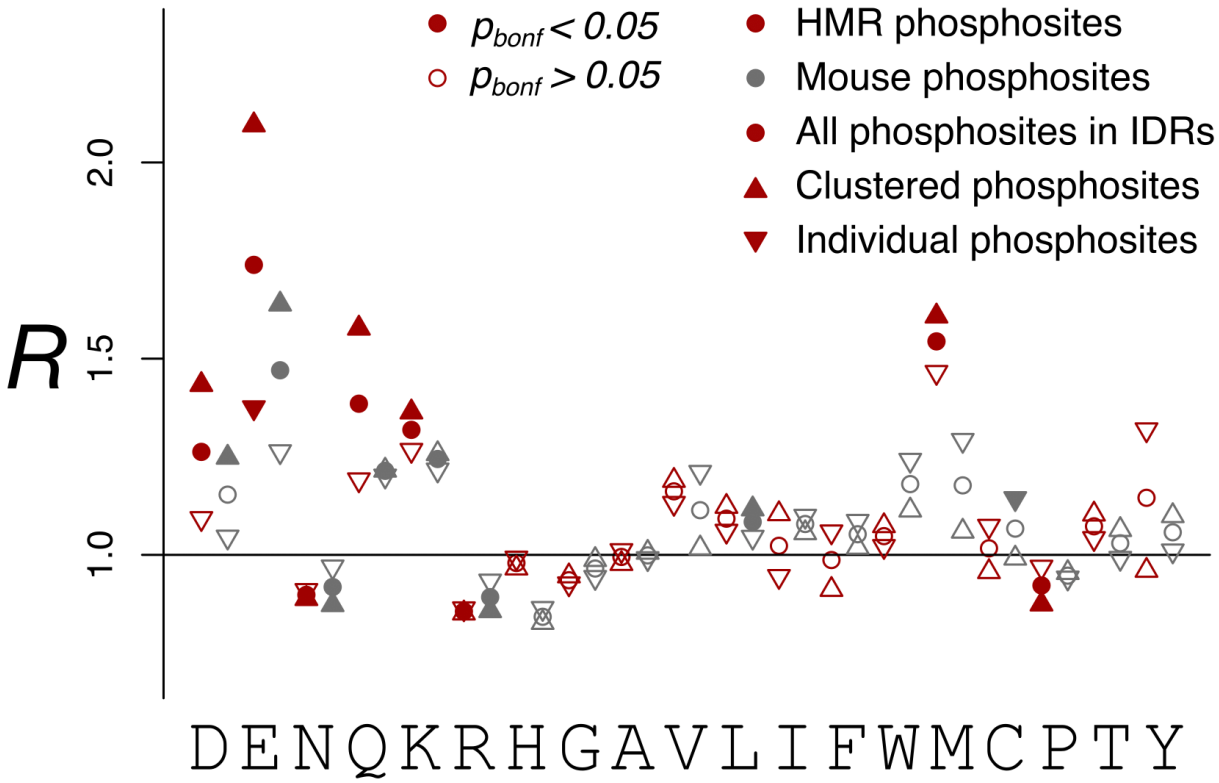
Supplementary Figure S26 | Comparison of the mutational patterns observed in mouse phosphosites located in phospho-islands *vs*. individual ones. Notation as in Suppl. Fig. S2.



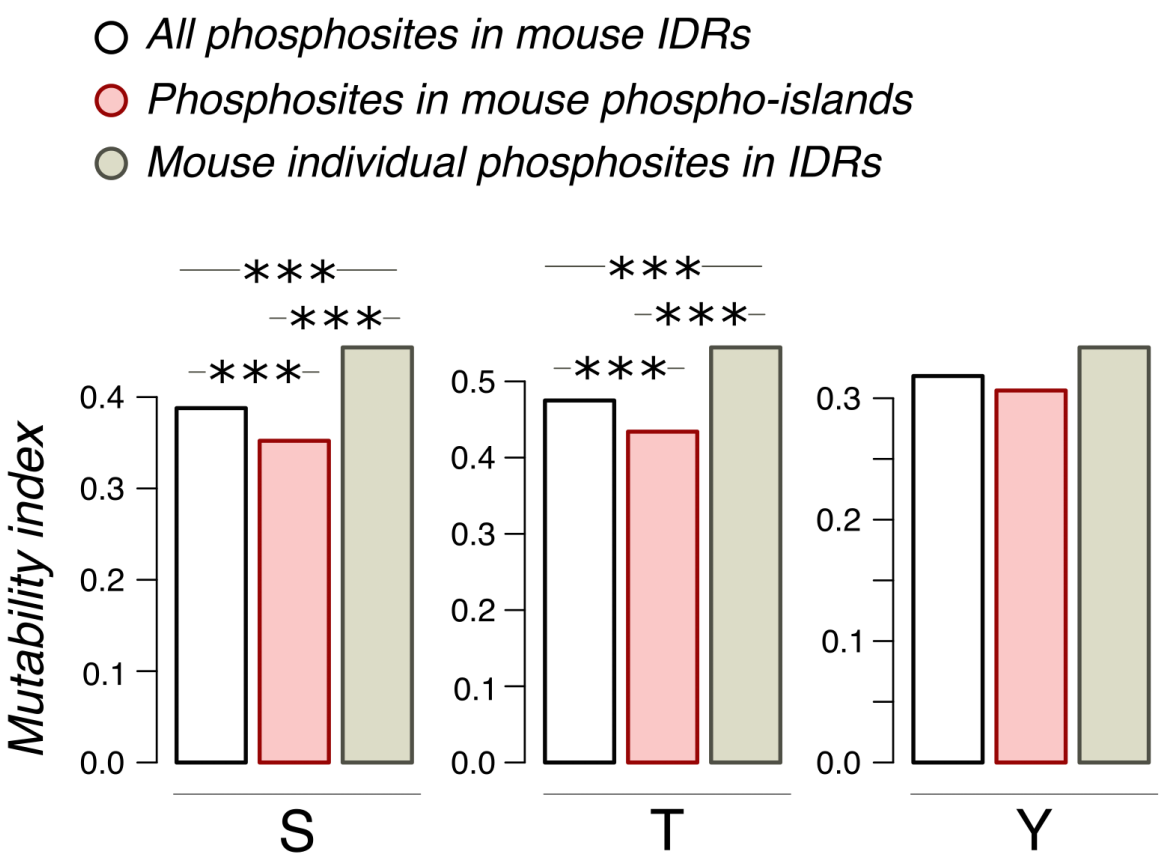
Supplementary Figure S27 | Comparison of the mutational patterns observed in human phosphosites located in intrinsically disordered and ordered regions. Notation as in Suppl. Fig. S2.



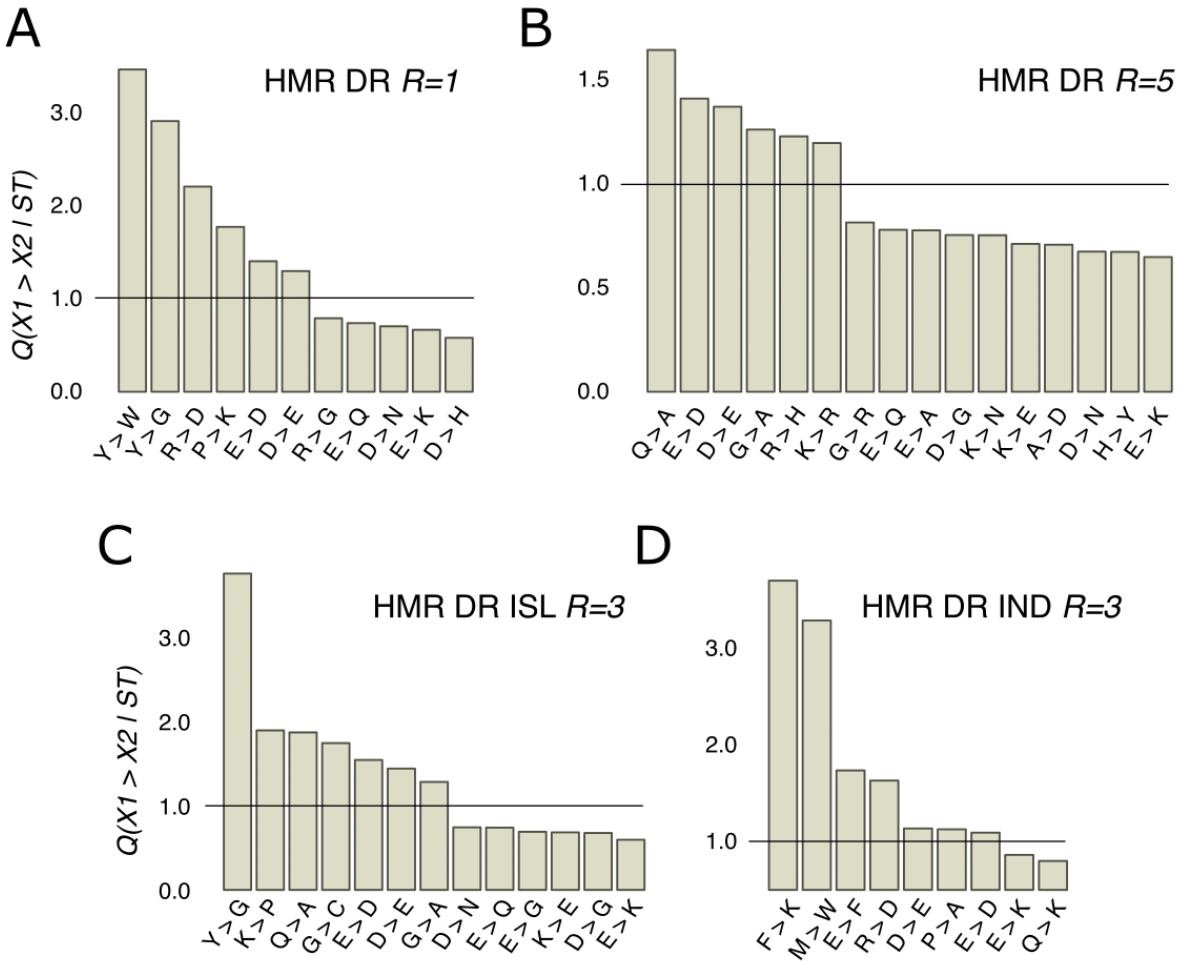
Supplementary Figure S28 | Comparison of the mutational patterns observed in human phosphosites located in phospho-islands *vs.* individual ones. Notation as in Suppl. Fig. S2.



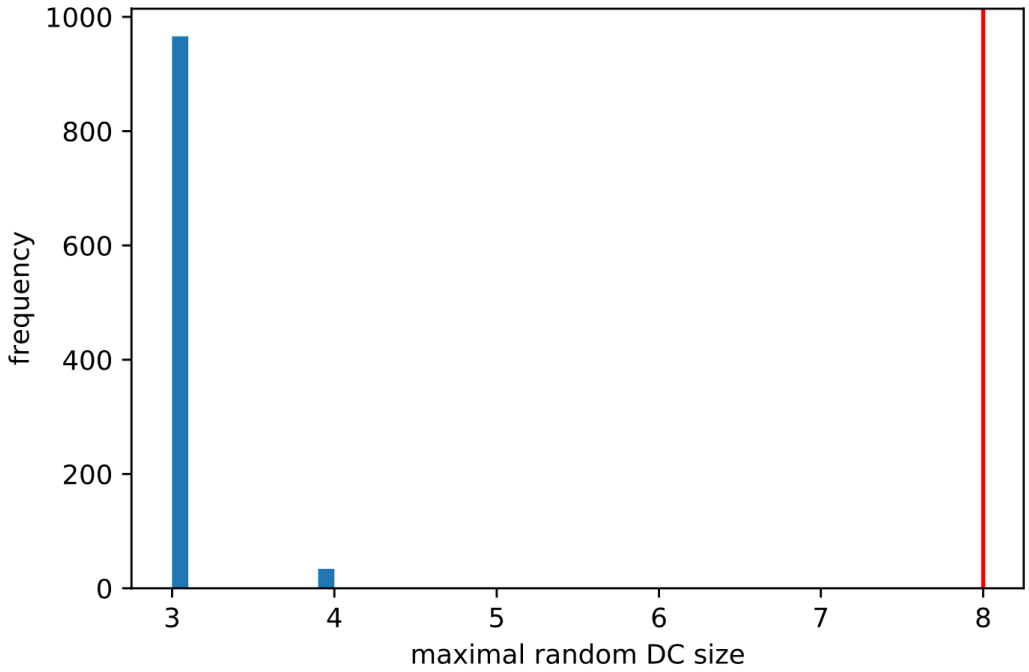
Supplementary Figure S29 | *R* values of the pS-to-X mutations for different phosphosite sets. Note the stronger effects observed for the HMR dataset compared to the mouse dataset and for clustered phosphoserines with respect to individual ones.



Supplementary Figure S30 | Conservation of phosphorylated amino acids for various mouse datasets. The mutability index is calculated for a phosphorylated amino acid with respect to its non-phosphorylated counterpart as the sum of probabilities of STY-to-X mutations calculated for all tree branches for the phosphorylated STY amino acid divided by the same sum for the respective non-phosphorylated amino acid. Three asterisks indicate statistically significant differences ($p < 0.001$, chi-squared contingency test).



Supplementary Figure S31 | *R* values of mutations near ST phosphosites with probabilities significantly different from the expected ones for various HMR subsets. (A) HMR sites located in intrinsically disordered regions with the window radius 1. **(B)** HMR sites located in disordered regions with the window radius 3. **(C)** Clustered HMR sites with the window radius 3. **(D)** Individual HMR sites with the window radius 3.



Supplementary Figure S32 | Maximal DC sizes in randomly computed uniform expectations (blue histogram) and the actual maximal editing DC size in *O. vulgaris* (red line).

Supplementary Tables

Supplementary tables are provided in the additional XLSX file published online at: https://peerj.com/articles/10436/.

List of supplementary tables:
/hmr_sites_list: List of all HMR-sites considered in the present study. Each site is marked as present in ordered/disordered region, having one of the five motif types, and being a part of a phosphosite cluster or individual. Phosphosites located in IDRs are considered as being neither clustered nor individual.
/mouse_site_list: List of mouse dataset phosphosites.
/hmr_psite_islands: List of HMR phosphor-islands.
/mouse_psite_islands: List of mouse phosphor-islands.
/R table: Table with *R* values and the respective corrected *p*-values for all considered phosphosite datasets. Fig. 13 and Suppl. Figs. S22-S26 correspond to this table.
/local_mutations: Q values of contextual mutations for different HMR phosphosite datasets and different window radius. Corresponds to Fig. 15 and Suppl. Fig. S29.
/phosphosite conservation: Mutability statistics for the mouse dataset. Corresponds to Suppl. Fig. S28.