


Insertion–Deletion Events Are Depleted in Protein Regions with Predicted Secondary Structure

Yi Yang¹, Matthew V. Braga¹, and Matthew D. Dean ^{1,*}

¹Molecular and Computational Biology, University of Southern California, Los Angeles, CA 90089, USA

*Corresponding author: E-mail: matthew.dean@usc.edu.

Accepted: April 21, 2024

Abstract

A fundamental goal in evolutionary biology and population genetics is to understand how selection shapes the fate of new mutations. Here, we test the null hypothesis that insertion–deletion (indel) events in protein-coding regions occur randomly with respect to secondary structures. We identified indels across 11,444 sequence alignments in mouse, rat, human, chimp, and dog genomes and then quantified their overlap with four different types of secondary structure—alpha helices, beta strands, protein bends, and protein turns—predicted by deep-learning methods of AlphaFold2. Indels overlapped secondary structures 54% as much as expected and were especially underrepresented over beta strands, which tend to form internal, stable regions of proteins. In contrast, indels were enriched by 155% over regions without any predicted secondary structures. These skews were stronger in the rodent lineages compared to the primate lineages, consistent with population genetic theory predicting that natural selection will be more efficient in species with larger effective population sizes. Nonsynonymous substitutions were also less common in regions of protein secondary structure, although not as strongly reduced as in indels. In a complementary analysis of thousands of human genomes, we showed that indels overlapping secondary structure segregated at significantly lower frequency than indels outside of secondary structure. Taken together, our study shows that indels are selected against if they overlap secondary structure, presumably because they disrupt the tertiary structure and function of a protein.

Key words: insertion, deletion, indels, evolution, selection.

Significance

How do insertion–deletion mutations, which occur when short stretches of amino acids are either added or deleted from a protein, accumulate in genomes? Here, we show that insertion–deletion events are less common in regions of proteins that are predicted to form secondary structures. We present multiple lines of evidence to show that this is most likely caused by selection against insertion–deletion events that disrupt the secondary structure and, therefore, the overall function of a protein.

Introduction

Understanding the fate of new mutations is critical to defining the evolutionary processes that shape biological diversity. At the level of single nucleotides, a rich body of theory has been developed to infer whether mutations

are neutral, deleterious, or beneficial (reviewed by Nielsen and Slatkin 2013; Hedrick 2005; Hartl and Clark 2007). Understanding the selective impact of insertion–deletion (indel) events, which can extend many nucleotides, has proven to be much more complicated.

© The Author(s) 2024. Published by Oxford University Press on behalf of Society for Molecular Biology and Evolution.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

Previous studies investigating the functional impact of indels generally fall into two categories (Savino et al. 2022). First, protein engineering studies have shown that indels can impact a protein's function, especially if they overlap important secondary structures (Simm et al. 2007; Arpino et al. 2014; Tóth-Petróczy and Tawfik 2014; Gavrilo et al. 2015, 2018; Grocholski et al. 2015; Liu et al. 2015, 2016; Jackson et al. 2017; Halliwell et al. 2018; Gonzalez et al. 2019; Woods et al. 2023). For example, Liu et al. (2016) found that experimentally deleting amino acids in beta strands and alpha helices of green fluorescent protein tended to reduce fluorescence, while deletions outside such regions were relatively neutral.

Second, evolutionary and population genetic studies have suggested that indels are relatively deleterious if they are long (Pascarella and Argos 1992; Taylor et al. 2004; Tao et al. 2007; Hsing and Cherkasov 2008; Kim and Guo 2010; Mills et al. 2011; Rockah-Shmuel et al. 2013; Lek et al. 2016; Zhang et al. 2018), cause frameshifts (lengar 2012; Montgomery et al. 2013; Chong et al. 2013; Bermejo-Das-Neves et al. 2014; Chen and Guo 2021), occur internally in the protein (Lin et al. 2017), alter flanking amino acids (Zhang et al. 2011), or fall outside of disordered regions (Taylor et al. 2004; Light, Sagit, Ekman, et al. 2013; Light, Sagit, Sachenkova, et al. 2013; Bermejo-Das-Neves et al. 2014; Khan et al. 2015). Protein families with indels tend to diverge in their structure and function relative to protein families without indels (Salari et al. 2008; Hormozdiari et al. 2009; Zhang et al. 2010, 2018; Gavrilo et al. 2015, 2018; Banerjee et al. 2019; Jayaraman et al. 2022), suggesting indels can be an important source of evolutionary novelty. Indeed, one study estimated that >70% of indels that have reached fixation have done so through positive selection (Barton and Zeng 2019).

Two important evolutionary studies identified orthologs across species and then overlapped inferred indels with experimentally determined protein structures in the Protein Data Bank (PDB; Berman et al. 2000). Following the publication of the human, mouse, and rat genomes, Taylor et al. (2004) identified 52 orthologous protein-coding genes that had an indel *and* a protein structure. Of these 52 indels, 31.5% of their sequence overlapped secondary structure of any kind, compared to 52.5% expected. A few years later, de la Chaux et al. (2007) analyzed the distribution of 343 protein-coding indels identified from human–chimp–rhesus orthologs that also occurred in the PDB. They found a deficiency of indels that overlapped alpha helices, but no difference in indels that overlapped beta strands.

As impactful as these studies were, they may not paint a full picture of the functional consequences of indel variation. The set of genes that could be studied was small, mostly limited by structural protein data or annotated

Pfam domains. Pfam domains do not necessarily correlate with 3D structure and the PDB represents a biased set of proteins (or protein regions) that are amenable to the experimental approaches required for structural proteomics, such as their ability to be crystallized. The relatively biased set of proteins for which we have structural data thus limits a systematic analysis across full genomes. For example, one study of duplicated genes could not analyze full-length proteins because of the divergence between aligned gene sequences and proteins represented in the PDB (Guo et al. 2012). However, the recent release of AlphaFold2—a deep-learning project that accurately predicts the 3D structure of a protein from its amino acid sequence (Jumper et al. 2021; Varadi et al. 2022)—provides a unique opportunity to systematically study indels across full proteins and whole genomes.

Here, we combine genome-wide predictions of AlphaFold2 with evolutionary and population genetic methods to ask whether indels occur randomly with respect to secondary structure, providing the most comprehensive evolutionary investigation into the fate of indels in protein-coding regions. We report four main results: (i) 97,382 indels identified from 11,444 five-species alignments in the phylogeny (dog, [mouse, rat], [human, chimp]) overlapped secondary structures 54% as often as expected but were 155% more common than expected in regions with no predicted secondary structures; (ii) indels that overlapped beta strands and occurred internally in a protein were especially rare, consistent with the known importance of these regions in overall protein structure; (iii) skews in observed versus expected were stronger in the rodent lineages compared to the primate lineages, consistent with theory predicting more efficient selection in rodents given their larger effective population sizes; and (iv) within human populations, indels that overlapped secondary structures occurred at significantly lower frequency compared to indels outside of secondary structures. Taken together, our results indicate selection acts against indels when they arise over structurally important regions of proteins, presumably because they can disrupt the overall structure and therefore the function of a protein.

Materials and Methods

Interspecific Indel Events

We downloaded protein sequences from all protein-coding genes identified as one-to-one orthologs between mouse, rat, human, chimp, and dog from Ensembl version 107 (ensembl.org). In the case of alternative transcripts, we chose the longest translated transcript to represent the gene. A total of 11,444 genes had one-to-one orthologs across all five species.

We aligned proteins using GUIDANCE (Penn, Privman, Ashkenazy, et al. 2010; Penn, Privman, Landan, et al.

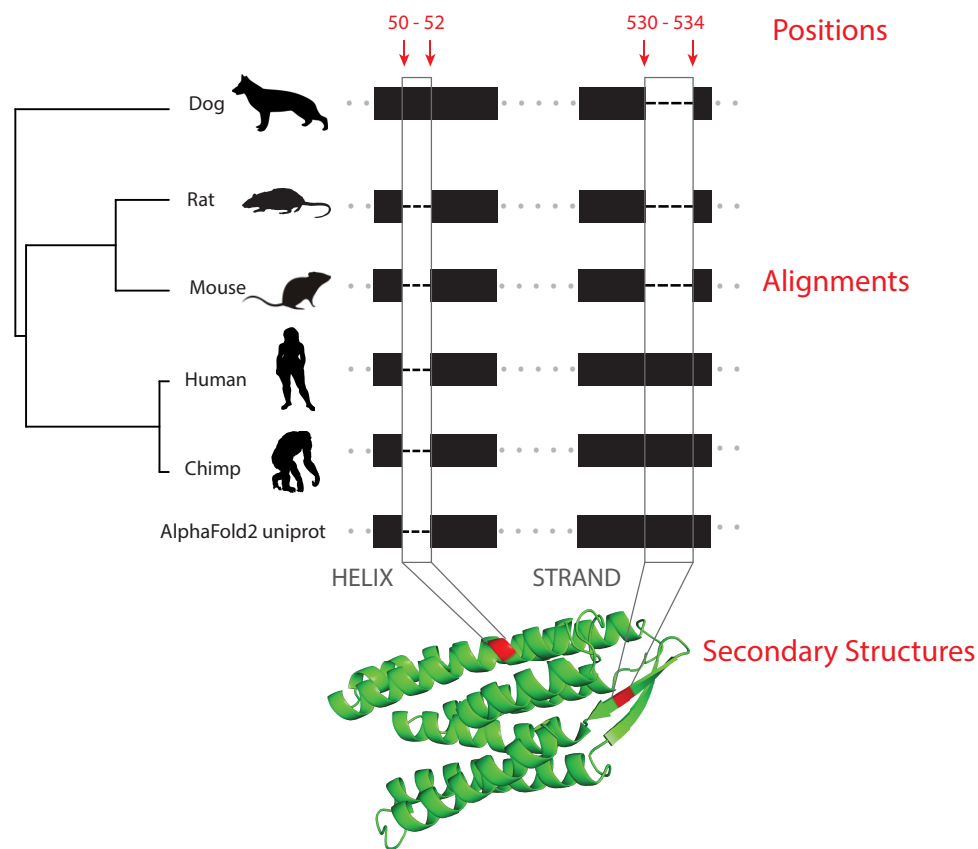


Fig. 1.—Schematic of main methodology. Shown is a hypothetical protein alignment between five species, which identified two unique indel events (positions 50 to 52 and positions 530 to 534). By including the UniProt sequence from AlphaFold2, we mapped from indel coordinates into predicted secondary structures. In this example, three positions fell over HELIX and five positions fell over STRAND. During randomization, we would permute the starting locations of these two indel events and then extend them by their observed lengths. Intraspecific analyses of human genomes proceeded in almost the same manner, except that indels were already called in their corresponding .vcf files. Silhouettes downloaded from phylocip.org: d149744f-8330-46df-8683-fcd4b385aa07 (Dog); a460430b-472b-4018-ba03-6b8eeb57fa5c (Rat); 570c7d9e-e6d1-46f5-b165-988981bfc5f6 (Mouse); 9fae30cd-fb59-4a81-a39c-e1826a35f612 (Human); 7133ab33-cc79-4d7c-9656-48717359abb4 (Chimp). All images are "dedicated to the public domain" or "free of known restrictions".

2010; Privman et al. 2012; Levy Karin et al. 2014). This approach estimates per-site alignment confidence by calculating its consistency across different starting guide trees, allowing us to incorporate a measure of confidence in downstream analyses. Importantly, we could use GUIDANCE scores to estimate error in indel placement and identify indels that were confidently placed. In each GUIDANCE iteration, we aligned protein sequences with MAFFT (Katoh et al. 2002). We ran MAFFT under the recommended default parameters; in the case of indels, the most important default parameters were the gap opening penalty (default = 1.53) and gap offset value (similar to gap extension penalty, default = 0.123). We then identified all indels as gaps from all 11,444 alignments (Fig. 1).

Our analyses could be impacted by sequencing errors or annotation errors that result in spurious inclusion or exclusion of amino acids from certain genes or by alignment errors (Fitch and Smith 1983; Chowdhury and Garai 2017).

Therefore, we repeated all downstream analyses after subsetting indels in four different ways: (i) *INTERNAL*: any indels that reached the beginning or ends of alignments were excluded, as visual inspection indicated these were noisy regions of alignment that could be related to incomplete annotation of full length genes; (ii) *GU94_PA100_GD40*: *INTERNAL* indels whose flanking five positions on both 5' and 3' ends (10 flanking positions total) had an average GUIDANCE confidence score of at least 0.94 (median observed), contained no indels, and had an average Grantham distance (Grantham 1974) of less than 40 (median observed), where Grantham distance was calculated using the R package AGVGD (<https://CRAN.R-project.org/package=agvgd>); this subset was meant to enrich for well-anchored indels and avoid problems distinguishing gaps in alignment due to protein divergence, versus gaps in alignment due to indel events (Snir and Pachter 2006; Salari et al. 2008; Jilani et al. 2022); (iii) *LENGTH_LTE20*: *INTERNAL* indels that were

less than or equal to 20 amino acids long in length, minimizing the impact of large indels that sometimes appeared to be spurious; and (iv) *MERGED*: INTERNAL indels after merging coordinates that overlapped, so that sites in an alignment that were in different overlapping indels only contributed once. We present the results from these four subsets as **Supplementary Material**, but they all produced essentially identical results as analyzing ALL indels.

AlphaFold2

AlphaFold2 is a deep learning approach developed by DeepMind to predict the 3D structure of proteins from only their amino acid sequence (Jumper et al. 2021; Varadi et al. 2022). Comparison to empirical data indicates these computational predictions are over 90% accurate.

AlphaFold2 assigns 43 different secondary structures to different regions of a protein, which we collapsed into five main categories. There were 32 different AlphaFold2 predictions that contained the phrase HELX, which are predictions of different helices; we collapsed these into the single term HELIX. There were eight different AlphaFold2 predictions that contained the phrase TURN, which are regions where the polypeptide is predicted to reverse direction in 3D space; we collapsed these into the single term TURN. We included the single AlphaFold2 prediction STRAND as-is, which are regions predicted to contain beta strands (also referred to as beta sheets). We included the single AlphaFold2 prediction BEND as-is, which are regions where the polypeptide is predicted to change direction but not fully reverse. There was one last AlphaFold2 prediction OTHER, but we did not observe any instances of this prediction in any of the proteins analyzed in this study so ignored that term. Each residue in the UniProt protein used by AlphaFold2 was assigned to one of these four categories or assigned the term NONE if they occurred outside any predicted secondary structure.

To link AlphaFold2 predictions to our five-species alignments above, we included the UniProt sequence in the alignment (Fig. 1). In rare cases, the AlphaFold2-downloaded UniProt sequence did not match the Ensembl-downloaded UniProt sequence, in which case we discarded the alignment from all analyses. Each position in each indel was then assigned HELIX, STRAND, TURN, BEND, or NONE (Fig. 1). In cases where the UniProt sequence was “deleted” (for example, indel 50 to 52 in Fig. 1), we assigned one-half of the deleted positions to whatever was assigned to its 5'-flanking residue and one-half to whatever was assigned to its 3'-flanking residue.

Randomization of Indel Positions

We generated null expectations through a randomization procedure. For each alignment, we randomly shuffled the

starting position of each indel and then extended each randomized indel by its observed length. In cases where a randomized indel extended past the end of an alignment, we wrapped the randomized indel to the front of the alignment. After shuffling the unique indels within each alignment, we recalculated the number of residues falling in each secondary structure, exactly as described above. We repeated this process 200 times to generate null expectations. We repeated this entire process for the four different subsets described above. For these four subsets, the relevant alignments were first truncated to match included regions and provide the appropriate background for randomization.

Gene Ontology (GO) Enrichment

For the MERGED indels only, we identified relative outliers by counting the number of sites in the alignment overlapping NONE versus not, versus sites overlapping indels versus not. We excluded alignments that had fewer than five positions in any of these four cells of this 2 × 2 table and then applied a χ^2 test and corrected the resulting *P*-values (Benjamini and Hochberg 1995). Genes with a $-\log_{10}$ *P*-value of at least 10 and at least a 1.5 fold change in expectation were taken as relative outliers. We tested whether these relative outlier genes were enriched for any Biological Process, Molecular Function, or Cellular Component using the Panther Classification system (Mi et al. 2013, 2017, 2019; Thomas et al. 2022), run from PantherDB (<https://pantherdb.org/>), with the settings “Test Type = Fisher’s Exact Test” and “Correction = Calculate False Discovery Rate”. We also performed enrichment analyses for genes that had no indels across the five species analyzed.

Accessibility and pIIDD Scores

Sites that are relatively internal on a 3D protein evolve more slowly than external sites, both at the level of nonsynonymous mutations (Goldman et al. 1998; Bustamante et al. 2000; Dean et al. 2002; Franzosa and Xia 2009; Tóth-Petróczy and Tawfik 2011; Scherrer et al. 2012; Shih et al. 2012; Marsh and Teichmann 2014; Shahmoradi et al. 2014; Yeh et al. 2014) and indel variation (Hsing and Cherkasov 2008; Guo et al. 2012). This correlation is complicated by whether or not external residues interact with other proteins (Mintseris and Weng 2005; Kim et al. 2006) or if externally oriented residues form active sites of proteins (Słodkiewicz and Goldman 2020).

For each site in each alignment, we calculated relative solvent accessibility, which is the degree to which a residue occurs on the outside of a folded protein (Tien et al. 2013), using FREEASA (Mitternacht 2016) with the “--format = rsa” option, using the AlphaFold2 structure as input. We also compared pIIDD scores (Mariani et al. 2013) across

an alignment. pLDDT scores are computational measures of confidence included in AlphaFold2 predictions. According to AlphaFold2, pLDDT scores < 50 likely represent intrinsically disordered or unstructured regions. As above, any “deletions” in the UniProt sequence were divided, and one-half of their sites were assigned the accessibility and pLDDT scores of their 5′-flanking residue and the other half to the scores of their 3′-flanking residue.

As will be shown below, secondary structure and relative solvent accessibility are strongly correlated. In an attempt to separate the effects of these two features on the probability of observing an indel, we compared receiver operating characteristic curves and area under the curve (AUC) values from three generalized linear models and then compared their likelihoods. Two models tested whether the probability of observing an indel was a function of secondary structure or relative solvent accessibility alone— $\text{glm}(\text{indel} \sim \text{secondary_structure})$ or $\text{glm}(\text{indel} \sim \text{rsa})$, respectively. A third model included both as independent variables— $\text{glm}(\text{indel} \sim \text{secondary_structure} + \text{rsa})$. We quantified the gain in likelihood when we included both independent variables, versus each one separately. For all three models, we included the “family = binomial” argument to model logistic variance. Our approach closely followed that of Jackson et al. (2017), modifying their scripts to suit our approach.

Because sites in a protein are not independent from each other, before applying generalized linear models, we randomly sampled a single site from each alignment. However, we did not sample sites with equal probability. Instead, we downweighted the probability of sampling by the inverse of the grand total of the five secondary structures (HELIX, STRAND, TURN, BEND, or NONE). By including this weighting scheme, we ensured even sampling of secondary structures, increasing the power of all three generalized linear models.

Comparison to Synonymous and Nonsynonymous Mutations

To provide additional context with which to interpret the distribution of indels, we tested three different nucleotide-based sites. First, we quantified the distribution of invariant sites across secondary structure as a kind of null distribution. Then, we quantified the same with respect to synonymous and nonsynonymous sites. We predicted that synonymous sites should distribute similarly to invariant sites, because they do not alter the protein sequence and thus probably have relatively minor effect on secondary structure. Conversely, we predicted that nonsynonymous sites would occur less frequently over secondary structure because, all else equal, their resulting amino acid changes could alter secondary structure.

Using the same five-species alignments above, we associated each protein to its transcript, downloaded from

Ensembl version 107, preserving the overall protein-based alignment. We counted the proportion of synonymous versus nonsynonymous variants occurring over the different secondary structures, compared to invariant sites. We only quantified synonymous versus nonsynonymous variants from the same alignments and sites that were used in our indel analyses.

Intraspecific Indel Events

As a complementary analysis to the interspecific analyses described above, we analyzed intraspecific variation from Phase 3 of the 1000 Human Genomes project (<https://www.internationalgenome.org/data-portal/data-collection/30x-grch38>) (1000 Genomes Project Consortium 2015; Byrska-Bishop et al. 2022). This database contains haplotype-phased indel calls (files named like ALL.chr1.shapeit2_integrated_snvindels_v2a_27022019.GRCh38.phased.INDELS.vcf) from 2,504 unrelated samples from 26 populations, with sample size ranging from 61 to 113 per population. These 26 populations derive from five large geographic areas: Africa, East Asia, South Asia, South America, and Europe.

Indel coordinates were truncated to match exon coordinates downloaded from UCSC Table Browser (table name = unipAliSwissprot from GRCh38). For any protein-coding genes that contained at least one indel, we assembled the reference and alternative alleles from the human genome, computationally placed indels, and then translated both alleles. Any indels that resulted in a frameshift in the first 95% of the protein-coding transcript (counted from 5′ translation start site) were excluded, because it is unclear whether reference and alternative alleles share 3D structure if they are dramatically frameshifted with respect to each other.

We only analyzed genes that were part of the five-species interspecific analyses described above. Otherwise, we would have included recent human-specific duplicates, where predictions might become noisy because of uncertainty about the exact timing of duplication along the lineage to modern humans.

Results

Indels Were Depleted in Regions with Secondary Structure

There were 11,444 genes that had one-to-one orthologs between dog, mouse, rat, chimp, and human genomes. Across these 11,444 alignments, we identified 97,382 indels spanning 1,272,048 positions. Indel sizes ranged from 1 to 2,870 residues long, but most were small: the 25%, 50%, and 75% quantiles were 1, 3, and 10 residues, respectively. Indel positions overlapped secondary structures significantly less than expected (Fig. 2; Table 1).

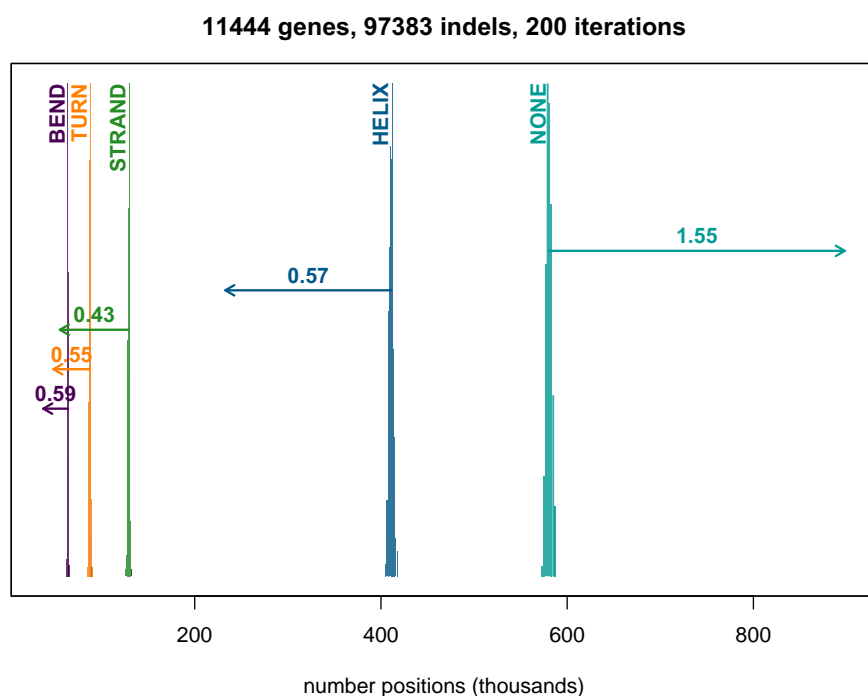


Fig. 2.—Comparison of observed versus expected number of alignment positions that overlap indels in the 11,444 alignments, stratified by secondary structure. Histograms built from randomizing indel positions across the alignments. Arrows originate at the mean expectation for each group, and terminate at the observed value. Indel sites overlap NONE 155% more than expected and overlap the four secondary structures less than expected (ranging from 43% expectation in STRAND to 59% expectation in BEND). Also see [Table 1](#).

Table 1

Number of indels or codon mutations that overlap secondary structures

	Indels			Codon-based							
	Observed	Expected	O/E	Invariant	p Inv.	Syn.	p Syn.	Syn./Inv.	Non.	p Non.	Non./Inv.
STRAND	55,293	129,070	0.43	455,059	0.120	278,936	0.130	1.09	143,454	0.086	0.72
TURN	48,258	87,473	0.55	287,034	0.076	189,311	0.088	1.17	110,149	0.066	0.87
HELIX	232,959	411,110	0.57	1,381,189	0.364	827,926	0.386	1.06	532,917	0.320	0.88
BEND	37,407	63,890	0.59	209,328	0.055	137,150	0.064	1.16	84,632	0.051	0.92
NONE	898,131	580,490	1.55	1,464,044	0.386	709,265	0.331	0.86	796,815	0.478	1.24

Observed = number of positions in alignments that map over each category. Expected = number expected based on randomization. Codons are classified as invariant (Invariant), synonymous (Syn.), or nonsynonymous (Non.). p = proportion of sites within their respective columns that fall within each category. This table is repeated as [supplementary table S2, Supplementary Material online](#), after employing four different subsetting strategies.

Indel positions were most underrepresented in STRAND, occurring at 43% expectations (calculated as 55,293 indel sites that overlapped STRAND, compared to 129,070 averaged across 200 randomizations), followed by indel positions occurring in TURN (55%), HELIX (57%), and BEND (59%) ([Table 1](#)). In contrast, indel positions occurred at 155% expectation over NONE, meaning indels were much more likely to occur in protein regions with no predicted secondary structure ([Table 1](#)). All observed values fell far outside the distributions from randomization ([Fig. 2](#)), translating into a *P*-value of essentially 0. We reached nearly identical conclusions after subsetting indels

in four ways described above ([supplementary fig. S1 and table S1, Supplementary Material online](#)), with one exception: indels over TURN and BEND are not underrepresented in the very stringent subset *GU94_PA100_GD40* ([supplementary fig. S1 and table S1, Supplementary Material online](#)).

Skews in Indel Distribution Were Stronger in Rodents

By using dog as an outgroup, we polarized all indels into either an insertion or deletion and placed each indel event on a specific branch in the phylogenetic tree, using simple

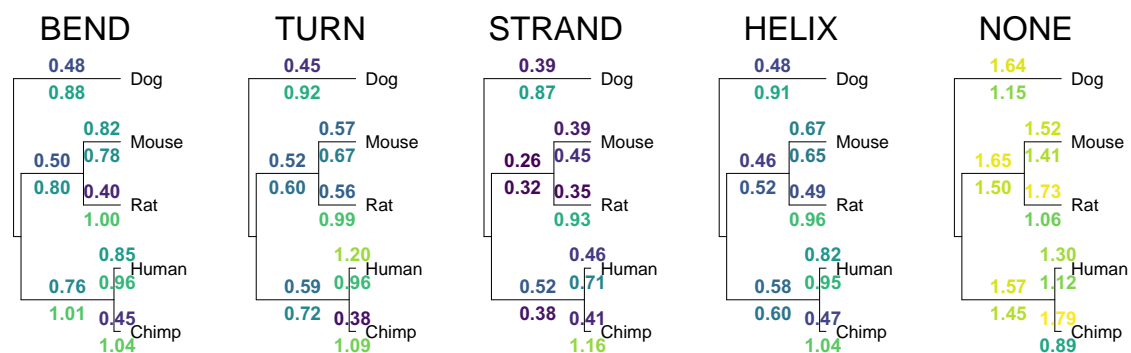


Fig. 3.—Observed:Expected ratios of indels, polarized into insertions (above branch) versus deletions (below branch), using Dog as outgroup. O:E ratios of insertions are more skewed than deletions, and the branches leading to rodent species generally show stronger skews than branches leading to primates.

parsimony. In other words, if amino acid sequences existed for human and chimp, but not for the other species (for example, indel 540-534 in Fig. 1), that indel was mapped as an insertion on the branch leading to primates.

There are seven branches on the phylogenetic tree analyzed here. Across the four secondary structures (BEND, TURN, STRAND, and HELIX), 24 of 28 O:E values were lower for insertions compared to deletions (Fig. 3). Conversely, across NONE sites all branches showed higher O:E for insertions compared to deletions. Taken together, these results suggest that insertions over secondary structure are more deleterious than deletions and therefore more likely to occur over NONE.

For the four secondary structures, O:E values were consistently lower in rodent lineages compared to primate lineages. There are four secondary structures that can be mapped to three rodent branches and three primate branches, where each branch contains insertions and deletions, for a total of 48 O:E values in Fig. 3. Forty-six of these 48 O:E values were lower in the rodent lineages compared to primate lineages. For example, O:E values for insertions over STRAND in the three rodent lineages = 0.26, 0.39, and 0.35, while in primates, the three values = 0.52, 0.46, and 0.41. Conversely, O:E values for NONE sites tended to be higher in rodents compared to primates. In sum, indels were especially unlikely to overlap secondary structures in rodents. All patterns described held after analyzing the four different subsets of indels described above (supplementary fig. S2, Supplementary Material online).

GO Analysis

We identified 797 alignments (genes) where the enrichment of indels over NONE was especially high (-log₁₀ P-value of at least 10 and at least a 1.5 fold difference in expectation). Compared to the rest of the 4,995 alignments, these 797 genes showed no statistical enrichment of Biological Process, but under the Cellular Component and

Molecular Function, ontologies showed enrichment of terms associated with cilia and ubiquitination. This enrichment lacks an obvious explanation.

We identified 88 alignments (genes) whose indels overlapped NONE much less than expected. None of these 88 genes showed enrichment of Biological Process or Molecular Function but showed enrichment of gene products localized to the nucleus under the Cellular Component. In sum, there were no striking or consistent patterns of GO enrichment associated with outlier genes in either direction.

We also analyzed the 904 genes which had no indels across any of the five species in the alignment. GO analysis uncovered many functional terms associated with neurotransmission, including synapse localization and synaptic transmission (supplementary table S2, Supplementary Material online). This result suggests that genes involved in neurotransmission may be especially intolerant of indel mutations. Interestingly, genes involved in immune response appeared to be underrepresented among genes with no indels. This result may indicate that immune genes undergo indel mutations more often than expected.

Indels Were Enriched in Regions with High Accessibility and Low pIIDD Scores

Accessibility and pIIDD scores varied according to secondary structure. STRAND had low accessibility and high pIIDD scores, indicating these secondary structures tend to fall on the inside of proteins and are relatively stable (Fig. 4). On the other end of the spectrum, NONE sites were much more accessible, with lower pIIDD scores, indicating external and unstable regions of proteins (Fig. 4).

Importantly, sites that overlapped indels consistently showed higher accessibility and lower pIIDD scores (compare X vs. O within each group; Fig. 4). In other words, within each secondary structure, indels were more commonly observed at sites that were relatively external and

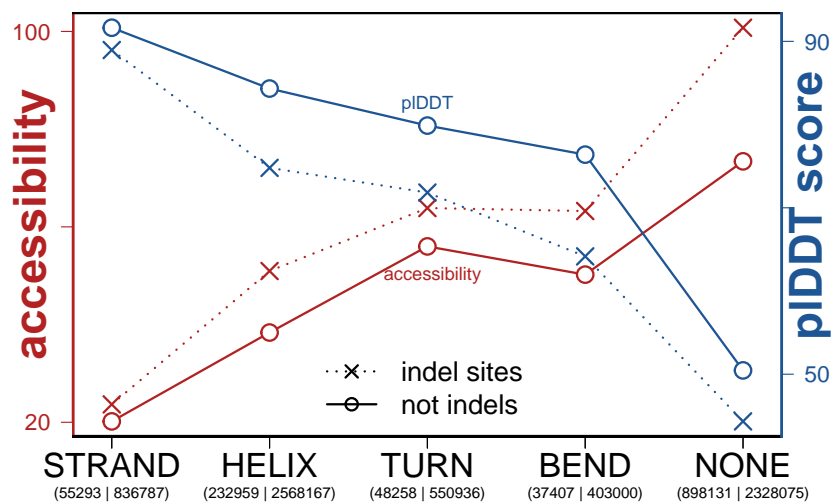


Fig. 4.—Weighted means of relative solvent accessibility (left axis) and pIDDT scores (right axis) across secondary structures, stratified by sites occurring over indels (X) versus sites not overlapping indels (O). Numbers on x axis indicate the number of sites that overlap an indel versus not (separated by |).

Table 2

AUC metrics for three generalized linear models

analysis_type	indel~SS	indel~RSA	indel~SS + RSA
ALL	0.684 (0.004)	0.707 (0.008)	0.720 (0.009)
INTERNAL	0.614 (0.010)	0.604 (0.007)	0.612 (0.005)
GU94_PA100_GD40	0.622 (0.006)	0.597 (0.015)	0.610 (0.013)
LENGTH_LTE20	0.618 (0.009)	0.610 (0.010)	0.621 (0.011)
MERGED	0.618 (0.006)	0.610 (0.014)	0.618 (0.011)

Mean (standard deviation) AUC from five iterations of randomly sampling sites across alignments.

in relatively unstable regions, compared to sites that did not overlap indels. Woods et al. (2023) found that experimentally deleting amino acids that reside in regions of high pIDDT were most likely to have a deleterious effect on protein function, providing an explanation for why we observe indels more frequently in regions with low pIDDT scores. This pattern held across all four subsets of indels described above (supplementary fig. S3, Supplementary Material online).

Comparing three different generalized linear models demonstrated that the effects of secondary structure were indistinguishable from the effects of relative solvent accessibility (Table 2). In the ALL data set, secondary structure performed about as well as relative solvent accessibility (AUC = 0.684 vs. 0.707, respectively), and including both as independent variables had only minor improvement to AUC (0.720) compared to single regressions. Similar results were obtained across the four subsets of data described above (Table 2). This shows that secondary structure and relative solvent accessibility are so correlated with each other that their effects cannot be meaningfully separated.

Nonsynonymous Variants Were Also Depleted in Protein Regions with Secondary Structure

Among the 11,444 alignments, we analyzed 3.8, 2.14, and 1.67 million codons that were invariant, synonymous, or nonsynonymous, respectively (Table 1). Synonymous codons overlapped secondary structures as often as invariant codons (synonymous-to-invariant ratios ranging from 0.86 to 1.17; Table 1). In contrast, nonsynonymous codons occurred far less frequently across the four secondary structures (nonsynonymous-to-invariant ratios ranging from 0.71 to 0.92) and more over NONE (nonsynonymous-to-invariant ratio of 1.24) (Table 1). These nonsynonymous-to-invariant ratios were generally smaller in magnitude than the *O:E* ratios estimated from indel distribution (Table 1). For example, indels occurred at 43% expectation over STRAND, while nonsynonymous codons occurred at 71% “expectation” (Table 1).

Similar patterns emerged after analyzing the four subsets of indels (supplementary table S1, Supplementary Material online). The main exception was that nonsynonymous-to-invariant ratios ranged from 0.91 to 0.98 across the four secondary structures and from 1.05 to 1.09 for NONE (supplementary table S1, Supplementary Material online). In other words, we still observed the general pattern that nonsynonymous variants were underrepresented across the four secondary structures and enriched over NONE, although at a smaller magnitude compared to the overall analysis.

Human Intraspecific Variation

We identified 1,921 indels from 1,436 unique genes, comprising a total of 4,354 positions. Most of these occurred at

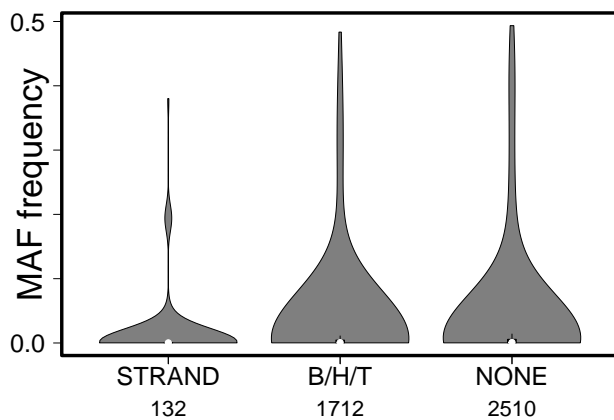


Fig. 5.—Violin plot of the minor allele frequency of indels in protein-coding regions, segregating within humans, stratified by secondary structure. B/H/T = pooled BEND + HELIX + TURN. Numbers on x axis indicate number of positions observed. Figure includes all human populations pooled; results remain qualitatively the same if we analyze populations separately.

a frequency of 1 allele observed among 5,008 phased alleles in the 1000 Genomes Project. We did not exclude these; even if they are due to sequencing or mapping errors, there is no reason to believe they would inflate our overall false positive rate as such errors should occur blindly with respect to secondary structure of proteins. In addition, an indel at a frequency of 1 allele could be especially deleterious, so we included them.

Across all six geographic regions, indel sites spanning NONE occurred at nearly twice the frequency than secondary structures (Fig. 5). NONE indels reached a mean frequency of 4 alleles out of 5,008 phased alleles, compared to BEND/HELIX/TURN indels (3 alleles) and STRAND indels (1 allele) (Kruskal–Wallis $\chi^2 = 37.8$, $df = 2$, $P < 10^{-8}$). If we use a minor allele frequency cutoff of 1%, 3%, or 5%, these patterns disappear, indicating that the majority of signal comes from the fact that a large proportion of STRAND indels occur as singletons.

Discussion

Our study combined the recent revolution in protein structure, ushered in by the AlphaFold2 project (Jumper et al. 2021), with evolutionary, population genetic, and permutation-based analyses to demonstrate that indels were depleted in regions of predicted secondary structure. This skew is especially strong for STRAND, which is consistent with these structures being internal and stable regions that are important for the overall 3D structure of a protein (Echave et al. 2016).

There are two nonmutually exclusive models—a mutational bias model versus a selection model—that could explain the nonrandom distribution of indels that we observe

here. Under a mutational bias model, the four secondary structures experience fundamentally different rates of indel mutation. The four different secondary structures tested here display systematic differences in amino acid composition (Chou and Fasman 1974; Fujiwara et al. 2012), which predicts different base composition and/or repetitive elements in the underlying DNA, which in turn could influence mutation rate.

However, three patterns in our data argue against the mutational bias hypothesis and instead provide support for a model where selection acts against indels that are more likely to disrupt protein function. First, *within* each secondary structure, positions with indels tend to occur in externally oriented and high pLDDT regions of proteins (Fig. 4). A mutational bias hypothesis cannot account for this discrepancy because we are comparing sites with or without indels within secondary structures. Second, the observed versus expected ratios (Table 1) are stronger in rodents compared to primates (Fig. 3). A mutational bias hypothesis cannot account for this interspecific variation unless different species also experience different mutational biases in a way that is related to secondary structure. In contrast, this pattern is predicted by a model of selection, because natural selection will operate more efficiently in species with large effective population size (Lynch 2007; Kimura 1983; Charlesworth 2009). Rodents have an effective population size that is roughly 10-fold larger than primates (Ohta 1972; Zhao et al. 2000; Won and Hey 2005; Geraldine et al. 2008, 2011). Finally, we showed that nonsynonymous variants were also depleted in regions of secondary structure, although not to the same degree (Table 1). A mutational bias hypothesis cannot explain the depletion of both indels and nonsynonymous variants over secondary structure, because these two classes differ in their mutational process.

To be sure, it is unlikely that indel mutations arise randomly. For example, G + C content often correlates with a genomic region's susceptibility to insertions or deletions (Sinden et al. 2002; Taylor et al. 2004), as well as features suggestive of a slippage mechanism (Nishizawa and Nishizawa 2002). However, a model of selection does not require indel mutation to be completely random. A selection model only requires any non-randomness to be equally distributed across the five categories of secondary structure tested here. It should also be pointed out that our study reports average deviations in observed versus expected across the entire genome. It remains unknown how much the strength of selection varies across individual indels, although our GO results did not uncover any functional similarity among the most highly skewed genes.

It is noteworthy that even within humans, we observed proportionately fewest indels over STRAND—exactly the secondary structure where indels were depleted in our five-species analyses. The low historical effective population size

of humans, coupled with multiple bottlenecks, is expected to reduce the efficiency of selection, yet we still observe skews in indel locations.

In conclusion, our analyses indicate that any change in amino acid sequence is likely to be deleterious for secondary structure, especially if that change is not a single nonsynonymous mutation, but the insertion or deletion of multiple amino acids. Indels that overlap STRAND and/or buried regions of the protein appear to be the most deleterious, while indels over NONE are the least. By analyzing the AlphaFold2 predictions, we have quantified these effects over whole genomes and full-length proteins, revealing a role for protein structure on the evolution of its primary sequence.

Supplementary Material

Supplementary material is available at *Genome Biology and Evolution* online.

Acknowledgments

Cornelius Gati provided many useful discussions on protein structure. Mark Chaisson provided many useful suggestions regarding 1000 genomes data. Jeff Jensen discussed population genetic inference. We thank Jackson et al. (2017) for publishing their Supplementary Material; our generalized linear modeling largely borrowed their code. Two anonymous reviewers and an associate editor greatly improved the quality of the manuscript. This project was born in BISC444; we thank the students of that class for their input. Computation was supported by the Center for Advanced Research Computing (CARC) at the University of Southern California—Tomek Osinski provided many useful discussions. Funding was provided by the National Science Foundation grant #2027373.

Data Availability

All data, code, and intermediate files required to reproduce the results here, as well as a README file, are available on Dryad (<https://doi.org/10.5061/dryad.bk3j9kdk9>) as a single protein_structure.tar.gz file (8.5 Gb) (<https://datadryad.org/stash/share/5NLwY6IUt75olgY16DgFRkyw jBUx2eola6RYDHFHdg>).

Literature Cited

1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*. 2015;526(7571):68–74. <https://doi.org/10.1038/nature15393>.

Arpino JAJ, Reddington SC, Halliwell LM, Rizkallah PJ, Jones DD. Random single amino acid deletion sampling unveils structural tolerance and the benefits of helical registry shift on GFP folding and structure. *Structure*. 2014;22(6):889–898. <https://doi.org/10.1016/j.str.2014.03.014>.

Banerjee A, Levy Y, Mitra P. Analyzing change in protein stability associated with single point deletions in a newly defined protein structure database. *J Proteome Res*. 2019;18(3):1402–1410. <https://doi.org/10.1021/acs.jproteome.9b00048>.

Barton HJ, Zeng K. The impact of natural selection on short insertion and deletion variation in the great tit genome. *Genome Biol Evol*. 2019;11(6):1514–1524. <https://doi.org/10.1093/gbe/evz068>.

Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc*. 1995;57(1):289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. *Nucleic Acids Res*. 2000;28(1):235–242. <https://doi.org/10.1093/nar/28.1.235>.

Bermejo-Das-Neves C, Nguyen H-N, Poch O, Thompson JD. A comprehensive study of small non-frameshift insertions/deletions in proteins and prediction of their phenotypic effects by a machine learning method (KD4i). *BMC Bioinformatics*. 2014;15:111. <https://doi.org/10.1186/1471-2105-15-111>.

Bustamante CD, Townsend JP, Hartl DL. Solvent accessibility and purifying selection within proteins of *Escherichia coli* and *Salmonella enterica*. *Mol Biol Evol*. 2000;17(2):301–308. <https://doi.org/10.1093/oxfordjournals.molbev.a026310>.

Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, Corvelo A, Clarke WE, Musunuri R, Nagulapalli K, et al. 2022. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell*. 185(18):3426–3440.e19. doi: 10.1016/j.cell.2022.08.004.

Charlesworth B. Effective population size and patterns of molecular evolution and variation. *Nat Rev Genet*. 2009;10(3):195–205. <https://doi.org/10.1038/nrg2526>.

Chen J, Guo J-T. Structural and functional analysis of somatic coding and UTR indels in breast and lung cancer genomes. *Sci Rep*. 2021;11(1):21178. <https://doi.org/10.1038/s41598-021-00583-1>.

Chong Z, Zhai W, Li C, Gao M, Gong Q, Ruan J, Li J, Jiang L, Lv X, Hungate E, et al. The evolution of small insertions and deletions in the coding genes of *Drosophila melanogaster*. *Mol Biol Evol*. 2013;30(12):2699–2708. <https://doi.org/10.1093/molbev/mst167>.

Chou PY, Fasman GD. Conformational parameters for amino acids in helical, β -sheet, and random coil regions calculated from proteins. *Biochemistry*. 1974;13(2):211–222. <https://doi.org/10.1021/bi00699a001>.

Chowdhury B, Garai G. A review on multiple sequence alignment from the perspective of genetic algorithm. *Genomics*. 2017;109(5-6):419–431. <https://doi.org/10.1016/j.ygeno.2017.06.007>.

Dean AM, Neuhauser C, Grenier E, Golding GB. The pattern of amino acid replacements in α - β -barrels. *Mol Biol Evol*. 2002;19(11):1846–1864. <https://doi.org/10.1093/oxfordjournals.molbev.a004009>.

de la Chaux N, Messer PW, Arndt PF. DNA indels in coding regions reveal selective constraints on protein evolution in the human lineage. *BMC Evol Biol*. 2007;7:191. <https://doi.org/10.1186/1471-2148-7-191>.

Echave J, Spielman SJ, Wilke CO. Causes of evolutionary rate variation among protein sites. *Nat Rev Genet*. 2016;17(2):109–121. <https://doi.org/10.1038/nrg.2015.18>.

Fitch WM, Smith TF. Optimal sequence alignments. *Proc Natl Acad Sci U S A*. 1983;80(5):1382–1386. <https://doi.org/10.1073/pnas.80.5.1382>.

Franzosa EA, Xia Y. Structural determinants of protein evolution are context-sensitive at the residue level. *Mol Biol Evol*. 2009;26(10):2387–2395. <https://doi.org/10.1093/molbev/msp146>.

- Fujiwara K, Toda H, Ikeguchi M. Dependence of α -helical and β -sheet amino acid propensities on the overall protein fold type. *BMC Struct Biol.* 2012;12(1):18. <https://doi.org/10.1186/1472-6807-12-18>.
- Gavrilov Y, Dagan S, Levy Y. Shortening a loop can increase protein native state entropy. *Proteins.* 2015;83(12):2137–2146. <https://doi.org/10.1002/prot.24926>.
- Gavrilov Y, Dagan S, Reich Z, Scherf T, Levy Y. An NMR confirmation for increased folded state entropy following loop truncation. *J Phys Chem B.* 2018;122(48):10855–10860. <https://doi.org/10.1021/acs.jpcc.8b09658>.
- Geraldes A, Basset P, Gibson B, Smith KL, Harr B, Yu HT, Bulatova N, Ziv Y, Nachman MW. Inferring the history of speciation in house mice from autosomal, X-linked, Y-linked and mitochondrial genes. *Mol Ecol.* 2008;17(24):5349–5363. <https://doi.org/10.1111/j.1365-294X.2008.04005.x>.
- Geraldes A, Basset P, Smith KL, Nachman MW. Higher differentiation among subspecies of the house mouse (*Mus musculus*) in genomic regions with low recombination. *Mol Ecol.* 2011;20(22):4722–4736. <https://doi.org/10.1111/j.1365-294X.2011.05285.x>.
- Goldman N, Thorne JL, Jones DT. Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics.* 1998;149(1):445–458. <https://doi.org/10.1093/genetics/149.1.445>.
- Gonzalez CE, Roberts P, Ostermeier M. Fitness effects of single amino acid insertions and deletions in TEM-1 β -lactamase. *J Mol Biol.* 2019;431(12):2320–2330. <https://doi.org/10.1016/j.jmb.2019.04.030>.
- Grantham R. Amino acid difference formula to help explain protein evolution. *Science.* 1974;185(4154):862–864. <https://doi.org/10.1126/science.185.4154.862>.
- Grocholski T, Dinis P, Niiranen L, Niemi J, Metsä-Ketelä M. Divergent evolution of an atypical S-adenosyl-L-methionine-dependent monooxygenase involved in anthracycline biosynthesis. *Proc Natl Acad Sci U S A.* 2015;112(32):9866–9871. <https://doi.org/10.1073/pnas.1501765112>.
- Guo B, Zou M, Wagner A. Pervasive indels and their evolutionary dynamics after the fish-specific genome duplication. *Mol Biol Evol.* 2012;29(10):3005–3022. <https://doi.org/10.1093/molbev/mss108>.
- Halliwell LM, Jathoul AP, Bate JP, Worthy HL, Anderson JC, Jones DD, Murray JAH. Δ fluc: brighter *Photinus pyralis* firefly luciferases identified by surveying consecutive single amino acid deletion mutations in a thermostable variant. *Biotechnol Bioeng.* 2018;115(1):50–59. <https://doi.org/10.1002/bit.26451>.
- Hartl DL, Clark AG. Principles of population genetics. 4th ed. Sunderland (MA): Sinauer; 2007.
- Hedrick PW. Genetics of populations. 3rd ed. Boston: Jones and Bartlett; 2005.
- Hormozdiari F, Salari R, Hsing M, Schönhuth A, Chan SK, Sahinalp SC, Cherkasov A. The effect of insertions and deletions on wirings in protein–protein interaction networks: a large-scale study. *J Comput Biol.* 2009;16(2):159–167. <https://doi.org/10.1089/cmb.2008.03TT>.
- Hsing M, Cherkasov A. Indel PDB: a database of structural insertions and deletions derived from sequence alignments of closely related proteins. *BMC Bioinformatics.* 2008;9(1):293. <https://doi.org/10.1186/1471-2105-9-293>.
- lengar P. An analysis of substitution, deletion and insertion mutations in cancer genes. *Nucleic Acids Res.* 2012;40(14):6401–6413. <https://doi.org/10.1093/nar/gks290>.
- Jackson EL, Spielman SJ, Wilke CO. Computational prediction of the tolerance to amino-acid deletion in green-fluorescent protein. *PLoS One.* 2017;12(4):e0164905. <https://doi.org/10.1371/journal.pone.0164905>.
- Jayaraman V, Toledo-Patiño S, Noda-García L, Laurino P. Mechanisms of protein evolution. *Protein Sci.* 2022;31(7):e4362. <https://doi.org/10.1002/pro.4362>.
- Jilani M, Haspel N, Jagodzinski F. Detection and analysis of amino acid insertions and deletions. In: Haspel N Jagodzinski F, Molloy K, editors. Algorithms and methods in structural bioinformatics. Computational Biology. Cham: Springer International Publishing; 2022. p. 89–99. https://doi.org/10.1007/978-3-031-05914-8_5.
- Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Židek A, Potapenko A et al. Highly accurate protein structure prediction with AlphaFold. *Nature.* 2021;596(7873):583–589. <https://doi.org/10.1038/s41586-021-03819-2>.
- Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 2002;30(14):3059–3066. <https://doi.org/10.1093/nar/gkf436>.
- Khan T, Douglas GM, Patel P, Nguyen Ba AN, Moses AM. Polymorphism analysis reveals reduced negative selection and elevated rate of insertions and deletions in intrinsically disordered protein regions. *Genome Biol Evol.* 2015;7(6):1815–1826. <https://doi.org/10.1093/gbe/ewi105>.
- Kim PM, Lu LJ, Xia Y, Gerstein MB. Relating three-dimensional structures to protein networks provides evolutionary insights. *Science.* 2006;314(5807):1938–1941. <https://doi.org/10.1126/science.1136174>.
- Kim R, Guo J. Systematic analysis of short internal indels and their impact on protein folding. *BMC Struct Biol.* 2010;10(1):24. <https://doi.org/10.1186/1472-6807-10-24>.
- Kimura M. The neutral theory of molecular evolution. Cambridge: Cambridge University Press; 1983.
- Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature.* 2016;536(7616):285–291. <https://doi.org/10.1038/nature19057>.
- Levy Karin E, Susko E, Pupko T. Alignment errors strongly impact likelihood-based tests for comparing topologies. *Mol Biol Evol.* 2014;31(11):3057–3067. <https://doi.org/10.1093/molbev/msu231>.
- Light S, Sagit R, Ekman D, Elofsson A. Long indels are disordered: a study of disorder and indels in homologous eukaryotic proteins. *Biochim Biophys Acta.* 2013;1834(5):890–897. <https://doi.org/10.1016/j.bbapap.2013.01.002>.
- Light S, Sagit R, Sachenkova O, Ekman D, Elofsson A. Protein expansion is primarily due to indels in intrinsically disordered regions. *Mol Biol Evol.* 2013;30(12):2645–2653. <https://doi.org/10.1093/molbev/mst157>.
- Lin M, Whitmire S, Chen J, Farrel A, Shi X, Guo JT. Effects of short indels on protein structure and function in human genomes. *Sci Rep.* 2017;7(1):9313. <https://doi.org/10.1038/s41598-017-09287-x>.
- Liu S, Wei X, Dong Xue, Xu Liang, Liu Jia, Jiang Biao. Structural plasticity of green fluorescent protein to amino acid deletions and fluorescence rescue by folding-enhancing mutations. *BMC Biochem.* 2015;16(1):17. <https://doi.org/10.1186/s12858-015-0046-5>.
- Liu S, Wei X, Ji Q, Xin X, Jiang B, Liu J. A facile and efficient transposon mutagenesis method for generation of multi-codon deletions in protein sequences. *J Biotechnol.* 2016;227:27–34. <https://doi.org/10.1016/j.jbiotec.2016.03.038>.
- Lynch M. The origins of genome architecture; 2007 [accessed 2023 August 11]. <https://repository.library.georgetown.edu/handle/10822/548280>.
- Mariani V, Biasini M, Barbato A, Schwede T. IDDT: a local superposition-free score for comparing protein structures and

- models using distance difference tests. *Bioinformatics*. 2013;29(21):2722–2728. <https://doi.org/10.1093/bioinformatics/btt473>.
- Marsh JA, Teichmann SA. Parallel dynamics and evolution: protein conformational fluctuations and assembly reflect evolutionary changes in sequence and structure. *BioEssays*. 2014;36(2):209–218. <https://doi.org/10.1002/bies.201300134>.
- Mi H, Huang X, Muruganujan A, Tang H, Mills C, Kang D, Thomas PD. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res*. 2017;45(D1):D183–D189. <https://doi.org/10.1093/nar/gkw1138>.
- Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res*. 2019;47(D1):D419–D426. <https://doi.org/10.1093/nar/gky1038>.
- Mi H, Muruganujan A, Thomas PD. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res*. 2013;41(D1):D377–D386. <https://doi.org/10.1093/nar/gks1118>.
- Mills RE, Pittard WS, Mullaney JM, Farooq U, Creasy TH, Mahurkar AA, Kemeza DM, Strassler DS, Ponting CP, Webber C, et al. Natural genetic variation caused by small insertions and deletions in the human genome. *Genome Res*. 2011;21(6):830–839. <https://doi.org/10.1101/gr.115907.110>.
- Mintseris J, Weng Z. Structure, function, and evolution of transient and obligate protein–protein interactions. *Proc Natl Acad Sci U S A*. 2005;102(31):10930–10935. <https://doi.org/10.1073/pnas.0502667102>.
- Mitternacht S. FreeSASA: an open source C library for solvent accessible surface area calculations. *F1000Res*. 2016;5:189. <https://doi.org/10.12688/f1000research.7931.1>.
- Montgomery SB, Goode DL, Kvikstad E, Albers CA, Zhang ZD, Mu XJ, Ananda G, Howie B, Karczewski KJ, Smith KS, et al. The origin, evolution, and functional impact of short insertion–deletion variants identified in 179 human genomes. *Genome Res*. 2013;23(5):749–761. <https://doi.org/10.1101/gr.148718.112>.
- Nielsen R, Slatkin M. *An introduction to population genetics: theory and applications*. Sunderland (MA): Sinauer Associates; 2013.
- Nishizawa M, Nishizawa K. A DNA sequence evolution analysis generalized by simulation and the Markov chain Monte Carlo method implicates strand slippage in a majority of insertions and deletions. *J Mol Evol*. 2002;55(6):706–717. <https://doi.org/10.1007/s00239-002-2366-5>.
- Ohta T. Evolutionary rate of cistrons and DNA divergence. *J Mol Evol*. 1972;1(2):150–157. <https://doi.org/10.1007/BF01659161>.
- Pascarella S, Argos P. Analysis of insertions/deletions in protein structures. *J Mol Biol*. 1992;224(2):461–471. [https://doi.org/10.1016/0022-2836\(92\)91008-D](https://doi.org/10.1016/0022-2836(92)91008-D).
- Penn O, Privman E, Ashkenazy H, Landan G, Graur D, Pupko T. GUIDANCE: a web server for assessing alignment confidence scores. *Nucleic Acids Res*. 2010;38(Web Server):W23–W28. <https://doi.org/10.1093/nar/gkq443>.
- Penn O, Privman E, Landan G, Graur D, Pupko T. An alignment confidence score capturing robustness to guide tree uncertainty. *Mol Biol Evol*. 2010;27(8):1759–1767. <https://doi.org/10.1093/molbev/msq066>.
- Privman E, Penn O, Pupko T. Improving the performance of positive selection inference by filtering unreliable alignment regions. *Mol Biol Evol*. 2012;29(1):1–5. <https://doi.org/10.1093/molbev/msr177>.
- Rockah-Shmuel L, Tóth-Petróczy Á, Sela A, Wurtzel O, Sorek R, Tawfik DS. Correlated occurrence and bypass of frame-shifting insertion–deletions (InDels) to give functional proteins. *PLoS Genet*. 2013;9(10):e1003882. <https://doi.org/10.1371/journal.pgen.1003882>.
- Salari R, Schönhuth A, Hormozdiari F, Cherkasov A, Sahinalp SC. The relation between indel length and functional divergence: a formal study. In: Crandall KA, Lagergren J, editors. *Algorithms in bioinformatics*. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer; 2008. p. 330–341. https://doi.org/10.1007/978-3-540-87361-7_28.
- Savino S, Desmet T, Franceus J. Insertions and deletions in protein evolution and engineering. *Biotechnol Adv*. 2022;60:108010. <https://doi.org/10.1016/j.biotechadv.2022.108010>.
- Scherrer MP, Meyer AG, Wilke CO. Modeling coding-sequence evolution within the context of residue solvent accessibility. *BMC Evol Biol*. 2012;12(1):179. <https://doi.org/10.1186/1471-2148-12-179>.
- Shahmoradi A, Sydykova DK, Spielman SJ, Jackson EL, Dawson ET, Meyer AG, Wilke CO. Predicting evolutionary site variability from structure in viral proteins: buriedness, packing, flexibility, and design. *J Mol Evol*. 2014;79(3-4):130–142. <https://doi.org/10.1007/s00239-014-9644-x>.
- Shih C-H, Chang C-M, Lin Y-S, Lo W-C, Hwang J-K. Evolutionary information hidden in a single protein structure. *Proteins*. 2012;80(6):1647–1657. <https://doi.org/10.1002/prot.24058>.
- Simm AM, Baldwin AJ, Busse K, Jones DD. Investigating protein structural plasticity by surveying the consequence of an amino acid deletion from TEM-1 β -lactamase. *FEBS Lett*. 2007;581(21):3904–3908. <https://doi.org/10.1016/j.febslet.2007.07.018>.
- Sinden RR, Potaman VN, Oussatcheva EA, Pearson CE, Lyubchenko YL, Shlyakhtenko LS. Triplet repeat DNA structures and human genetic disease: dynamic mutations from dynamic DNA. *J Biosci*. 2002;27(1):53–65. <https://doi.org/10.1007/BF02703683>.
- Slodkowitz G, Goldman N. Integrated structural and evolutionary analysis reveals common mechanisms underlying adaptive evolution in mammals. *Proc Natl Acad Sci U S A*. 2020;117(11):5977–5986. <https://doi.org/10.1073/pnas.1916786117>.
- Snir S, Pachter L. Phylogenetic profiling of insertions and deletions in vertebrate genomes. In: Apostolico A, Guerra C, Istrail S, Pevzner PA, Waterman M, editors. *Research in computational molecular biology*. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer; 2006. p. 265–280. https://doi.org/10.1007/11732990_23.
- Tao S, Fan Y, Wang W, Ma G, Liang L, Shi Q. Patterns of insertion and deletion in mammalian genomes. *Curr Genomics*. 2007;8(6):370–378. <https://doi.org/10.2174/138920207783406479>.
- Taylor MS, Ponting CP, Copley RR. Occurrence and consequences of coding sequence insertions and deletions in mammalian genomes. *Genome Res*. 2004;14(4):555–566. <https://doi.org/10.1101/gr.1977804>.
- Thomas PD, Ebert D, Muruganujan A, Mushayahama T, Albou L-P, Mi H. PANTHER: making genome-scale phylogenetics accessible to all. *Protein Sci*. 2022;31(1):8–22. <https://doi.org/10.1002/pro.4218>.
- Tien MZ, Meyer AG, Sydykova DK, Spielman SJ, Wilke CO. Maximum allowed solvent accessibilities of residues in proteins. *PLoS One*. 2013;8(11):e80635. <https://doi.org/10.1371/journal.pone.0080635>.
- Tóth-Petróczy Á, Tawfik DS. Slow protein evolutionary rates are dictated by surface–core association. *Proc Natl Acad Sci U S A*. 2011;108(27):11151–11156. <https://doi.org/10.1073/pnas.1015994108>.
- Tóth-Petróczy Á, Tawfik DS. Hopeful (protein InDel) monsters? *Structure*. 2014;22(6):803–804. <https://doi.org/10.1016/j.str.2014.05.013>.
- Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, Yuan D, Stroe O, Wood G, Laydon A, et al. AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic*

- Acids Res. 2022;50(D1):D439–D444. <https://doi.org/10.1093/nar/gkab1061>.
- Won Y-J, Hey J. Divergence population genetics of chimpanzees. *Mol Biol Evol.* 2005;22(2):297–307. <https://doi.org/10.1093/molbev/msi017>.
- Woods H, Schiano DL, Aguirre JI, Ledwitch KV, McDonald EF, Voehler M, Meiler J, Schoeder CT. Computational modeling and prediction of deletion mutants. *Structure.* 2023;31(6):713–723.e3. <https://doi.org/10.1016/j.str.2023.04.005>.
- Yeh S-W, Liu J-W, Yu S-H, Shih C-H, Hwang J-K, Echave J. Site-specific structural constraints on protein sequence evolutionary divergence: local packing density versus solvent exposure. *Mol Biol Evol.* 2014;31(1):135–139. <https://doi.org/10.1093/molbev/mst178>.
- Zhang Z, Huang J, Wang Z, Wang L, Gao P. Impact of indels on the flanking regions in structural domains. *Mol Biol Evol.* 2011;28(1):291–301. <https://doi.org/10.1093/molbev/msq196>.
- Zhang Z, Wang J, Gong Y, Li Y. Contributions of substitutions and indels to the structural variations in ancient protein superfamilies. *BMC Genomics.* 2018;19(1):771. <https://doi.org/10.1186/s12864-018-5178-8>.
- Zhang Z, Wang Y, Wang L, Gao P. The combined effects of amino acid substitutions and indels on the evolution of structure within protein families. *PLoS One.* 2010;5(12):e14316. <https://doi.org/10.1371/journal.pone.0014316>.
- Zhao Z, Jin L, Fu Y-X, Ramsay M, Jenkins T, Leskinen E, Pamilo P, Trexler M, Patthy L, Jorde LB, et al. Worldwide DNA sequence variation in a 10-kilobase noncoding region on human chromosome 22. *Proc Natl Acad Sci U S A.* 2000;97(21):11354–11358. <https://doi.org/10.1073/pnas.200348197>.

Associate editor: Brian Golding