Insertion-deletions are depleted in protein regions with predicted secondary structure

Yi Yang, Matthew Braga, and Matthew D. Dean

Molecular and Computational Biology
University of Southern California
1050 Childs Way
Los Angeles, CA 90089

matthew.dean@usc.edu

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 events (indels) 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, 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 under-represented 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.

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 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 Hedrick 2005; Hartl and Clark 2007; Nielsen and Slatkin 2013). Understanding the selective impact of insertion-deletion events (indels), which can extend many nucleotides, has proven to be much more complicated.

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; Gavrilov et al. 2015; Grocholski et al. 2015; Liu et al. 2015; Liu et al. 2016; Jackson et al. 2017; Gavrilov et al. 2018; 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 frame-shifts (Iengar 2012; Chong et al. 2013; Montgomery 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; Gavrilov et al. 2015; Gavrilov et al. 2018; Zhang et al. 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 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: 1) 97,382 indels identified from 11,444 five-species alignments in the tree (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, 2) 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, 3) skews in observed vs. 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 4) 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 overall structure and therefore the function of a protein.

Materials and Methods

Interspecific insertion-deletion (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. 11,444 genes had one-to-one orthologs across all five species.

We aligned proteins using GUIDANCE (Penn, Privman, Landan, et al. 2010; Penn, Privman, Ashkenazy, et al. 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: 1) 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, 2) 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 overlapping 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 to insertion-deletion events (Snir and Pachter 2006; Salari et al. 2008; Jilani et al. 2022), 3) 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 4) MERGED: INTERNAL indels after merging coordinates that overlapped, so that sites in an alignment that were in different overlapping regions only contributed once. We present the results from these four subsets as supplementary files, 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 8 different AlphaFold2 predictions that contain 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 50-52 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, 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 re-calculated 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 a more appropriate background for randomization.

Gene Ontology enrichment. For the MERGED indels only, we identified relative outliers by counting the number of sites in the alignment overlapping NONE vs. not, versus sites overlapping indels vs. not. We excluded alignments that had fewer than 5 positions in any of these four cells of this 2x2 table, then applied a X^2 test and corrected resulting p-values (Benjamini and Hochberg 1995). Genes with a -log10 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 Panther Classification system (Mi et al. 2013; Mi et al. 2017; Mi et al. 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 Gene Ontology analyses for genes which had no indels across the five species analyzed.

Accessibility and plDDT 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 (Slodkowicz 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 FreeSASA (Mitternacht 2016) with the “--format=rsa” option, using the AlphaFold2 structure as input. We also compared plDDT 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 (ROC) 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 – glm(indel~secondary_structure) or glm(indel~rsa), respectively. A third model included both as independent variables – glm(indel~secondary_structure + 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 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 5-species alignments above, we reverse-translated each protein to its transcript, downloaded from Ensembl version 107. We counted the proportion of synonymous vs. nonsynonymous variants occurring over the different secondary structures, compared to invariant sites. We only quantified synonymous vs. 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) (The Genomes Project 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 frame-shifted 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). Indel positions were most under-represented 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 in NONE, meaning indels were much more likely 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 Figure 1, Supplementary Table 1), with one exception: indels over TURN and BEND are not under-represented in the very stringent subset GU94_PA100_GD40 (Supplementary Figure 1, Supplementary Table 1).

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 parsimony. In other words, if amino acid sequences existed for mouse and rat, but not for the other species, that indel was mapped as an insertion on the branch leading to rodents.

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 (Figure 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.

For the four secondary structures, O:E values were consistently lower in rodent lineages compared to primate lineages. There are four secondary structure 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 Figure 2. 46 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 tend 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 Figure 2).

GO analysis. We identified 797 alignments (genes) where the enrichment of indels over NONE was especially high. 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 Cellular Component. In sum, there were no striking or consistent patterns of Gene Ontology 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 2). This result suggests that genes involved in neurotransmission may be especially intolerant of indel mutations. Interestingly, genes involved in immune response appeared to be under-represented 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 plDDT scores. Accessibility and plDDT scores varied according to secondary structure. STRAND had low accessibility and high plDDT 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 plDDT scores, indicating external and unstable regions of proteins (Fig. 4).

Importantly, sites that overlapped indels consistently showed higher accessibility and lower plDDT 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 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 plDDT 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 plDDT scores. This pattern held across all four subsets of indels described above (Supplementary Figure 3).

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 dataset, 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 1). 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 1). In other words, we still observed the general pattern that nonsynonynmous variants were under-represented 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 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 6 geographic regions, indel sites spanning NONE occurred at nearly twice the frequency than secondary structures. 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 X^2= 37.8, df = 2, p-value < 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 non-mutually exclusive models – a mutational bias model versus a selection model – that could explain the non-random 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 1975; 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 they are the same secondary structures in different parts of the same protein. Second, the observed vs. expected ratios (Table 1) are stronger in rodents compared to primates (Figure 3). A mutational bias hypothesis cannot account for this interspecific variation unless different species also experience different mutational biases. 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 (Kimura 1983; Lynch 2007; 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; Geraldes et al. 2008; Geraldes et al. 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 in mutational process 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 vs. expected across the entire genome. It remains unknown how much the strength of selection varies across individual indels, although our Gene Ontology 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, are 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 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.

Data and resource 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). [for reviewers only: that link is not yet public; this link provides access: https://datadryad.org/stash/share/5NLwY6IUt75oIgY16DgFRkywjBUx2eoela6RYDHFHdg]

Acknowledgements

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 data; 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 Center for Advanced Research Computing (CARC) at the University of Southern California – Tomek Osinski provided many useful discussions. Funding was provided by National Science Foundation grant #2027373.

Figure Legends

Figure 1. Schematic of main methodology. Shown is a hypothetical protein alignment between five species, which identified two unique indel events (positions 50-52 and positions 530-534). 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 SHEET. During randomization, we would permute the starting locations of these two indel events, then extend them by their observed length. Intraspecific analyses of human genomes proceeded in almost the same manner, except that indels were already called in their corresponding .vcf files.

Figure 2. Comparison of observed vs. 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 at top originate at the mean expectation for each group, and terminate at the observed value. Indel sites overlap NONE 132% more than expected, and overlap the four secondary structures less than expected (ranging from 62% expectation in STRAND to 84% expectation in TURN). Also see Table 1.

Figure 3. Observed:Expected ratios of indels, polarized into insertions (above branch) versus deletions (below branch), using Dog as outgroup. There is no consistent

difference in O:E in insertions and deletions, but the branches leading to rodent species generally show stronger skews than branches leading to primates.

Figure 4. Weighted means of relative solvent accessibility (red, left axis) and pIDDT scores (blue, 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 |).

Figure 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.

Supplementary Figure 1. A repeat of Figure 2, but for each of the four different subsets of indels.

Supplementary Figure 3. A repeat of Figure 3, but for each of the four different subsets of indels.

Supplementary Figure 3. A repeat of Figure 4, but for each of the four different subsets of indels.

References

Arpino JAJ, Reddington SC, Halliwell LM, Rizkallah PJ, Jones DD. 2014. Random Single Amino Acid Deletion Sampling Unveils Structural Tolerance and the Benefits of Helical Registry Shift on GFP Folding and Structure. *Structure* 22:889–898.

Banerjee A, Levy Y, Mitra P. 2019. Analyzing Change in Protein Stability Associated with Single Point Deletions in a Newly Defined Protein Structure Database. *J. Proteome Res.* 18:1402–1410.

Barton HJ, Zeng K. 2019. The Impact of Natural Selection on Short Insertion and Deletion Variation in the Great Tit Genome. *Genome Biology and Evolution* 11:1514–1524.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)* 57:289–300.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The Protein Data Bank. *Nucleic Acids Research* 28:235–242.

Bermejo-Das-Neves C, Nguyen H-N, Poch O, Thompson JD. 2014. 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* 15:111.

Bustamante CD, Townsend JP, Hartl DL. 2000. Solvent Accessibility and Purifying Selection Within Proteins of Escherichia coli and Salmonella enterica. *Molecular Biology and Evolution* 17:301–308.

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* 185:3426-3440.e19.

Charlesworth B. 2009. Effective population size and patterns of molecular evolution and variation. *Nat Rev Genet* 10:195–205.

de la Chaux N, Messer PW, Arndt PF. 2007. DNA indels in coding regions reveal selective constraints on protein evolution in the human lineage. *BMC Evol Biol* 7:191.

Chen J, Guo J. 2021. Structural and functional analysis of somatic coding and UTR indels in breast and lung cancer genomes. *Sci Rep* 11:21178.

Chong Z, Zhai W, Li C, Gao M, Gong Q, Ruan J, Li J, Jiang L, Lv X, Hungate E, et al. 2013. The Evolution of Small Insertions and Deletions in the Coding Genes of Drosophila melanogaster. *Molecular Biology and Evolution* 30:2699–2708.

Chou PY, Fasman GD. 1975. Conformational parameters for amino acids in helical, β-sheet, and random coil regions calculated from proteins. *ACS Publications* [Internet]. Available from: https://pubs.acs.org/doi/pdf/10.1021/bi00699a001

Chowdhury B, Garai G. 2017. A review on multiple sequence alignment from the perspective of genetic algorithm. *Genomics* 109:419–431.

Dean AM, Neuhauser C, Grenier E, Golding GB. 2002. The Pattern of Amino Acid Replacements in α/β-Barrels. *Molecular Biology and Evolution* 19:1846–1864.

Echave J, Spielman SJ, Wilke CO. 2016. Causes of evolutionary rate variation among protein sites. *Nat Rev Genet* 17:109–121.

Fitch WM, Smith TF. 1983. Optimal sequence alignments. *Proceedings of the National Academy of Sciences* 80:1382–1386.

Franzosa EA, Xia Y. 2009. Structural Determinants of Protein Evolution Are Context-Sensitive at the Residue Level. *Molecular Biology and Evolution* 26:2387–2395.

Fujiwara K, Toda H, Ikeguchi M. 2012. Dependence of α-helical and β-sheet amino acid propensities on the overall protein fold type. *BMC Structural Biology* 12:18.

Gavrilov Y, Dagan S, Levy Y. 2015. Shortening a loop can increase protein native state entropy. *Proteins: Structure, Function, and Bioinformatics* 83:2137–2146.

Gavrilov Y, Dagan S, Reich Z, Scherf T, Levy Y. 2018. An NMR Confirmation for Increased Folded State Entropy Following Loop Truncation. *J. Phys. Chem. B* 122:10855–10860.

Geraldes A, Basset P, Gibson B, Smith KL, Harr B, Yu HT, Bulatova N, Ziv Y, Nachman MW. 2008. Inferring the history of speciation in house mice from autosomal, X-linked, Y-linked and mitochondrial genes. *Mol Ecol* 17:5349–5363.

Geraldes A, Basset P, Smith KL, Nachman MW. 2011. Higher differentiation among subspecies of the house mouse (*Mus musculus*) in genomic regions with low recombination. *Mol Ecol* 20:4722–4736.

Goldman N, Thorne JL, Jones DT. 1998. Assessing the Impact of Secondary Structure and Solvent Accessibility on Protein Evolution. *Genetics* 149:445–458.

Gonzalez CE, Roberts P, Ostermeier M. 2019. Fitness Effects of Single Amino Acid Insertions and Deletions in TEM-1 β-Lactamase. *Journal of Molecular Biology* 431:2320–2330.

Grantham R. 1974. Amino acid difference formula to help explain protein evolution. *Science* 185:862–864.

Grocholski T, Dinis P, Niiranen L, Niemi J, Metsä-Ketelä M. 2015. Divergent evolution of an atypical S-adenosyl-l-methionine–dependent monooxygenase involved in anthracycline biosynthesis. *Proceedings of the National Academy of Sciences* 112:9866–9871.

Guo B, Zou M, Wagner A. 2012. Pervasive Indels and Their Evolutionary Dynamics after the Fish-Specific Genome Duplication. *Molecular Biology and Evolution* 29:3005–3022.

Halliwell LM, Jathoul AP, Bate JP, Worthy HL, Anderson JC, Jones DD, Murray JAH. 2018. ΔFlucs: Brighter Photinus pyralis firefly luciferases identified by surveying consecutive single amino acid deletion mutations in a thermostable variant. *Biotechnology and Bioengineering* 115:50–59.

Hartl DL, Clark AG. 2007. Principles of population genetics. 4th ed. Sunderland, MA: Sinauer

Hedrick PW. 2005. Genetics of populations. 3rd ed. Boston: Jones and Bartlett

Hormozdiari F, Salari R, Hsing M, Schönhuth A, Chan SK, Sahinalp SC, Cherkasov A. 2009. The Effect of Insertions and Deletions on Wirings in Protein-Protein Interaction Networks: A Large-Scale Study. *Journal of Computational Biology* 16:159–167.

Hsing M, Cherkasov A. 2008. Indel PDB: A database of structural insertions and deletions derived from sequence alignments of closely related proteins. *BMC Bioinformatics* 9:293.

Iengar P. 2012. An analysis of substitution, deletion and insertion mutations in cancer genes. *Nucleic Acids Research* 40:6401–6413.

Jackson EL, Spielman SJ, Wilke CO. 2017. Computational prediction of the tolerance to amino-acid deletion in green-fluorescent protein. *PLOS ONE* 12:e0164905.

Jayaraman V, Toledo-Patiño S, Noda-García L, Laurino P. 2022. Mechanisms of protein evolution. *Protein Science* 31:e4362.

Jilani M, Haspel N, Jagodzinski F. 2022. 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. p. 89–99. Available from: 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, Žídek A, Potapenko A. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* 596:583–589.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research* 30:3059–3066.

Khan T, Douglas GM, Patel P, Nguyen Ba AN, Moses AM. 2015. Polymorphism Analysis Reveals Reduced Negative Selection and Elevated Rate of Insertions and Deletions in Intrinsically Disordered Protein Regions. *Genome Biology and Evolution* 7:1815–1826.

Kim PM, Lu LJ, Xia Y, Gerstein MB. 2006. Relating Three-Dimensional Structures to Protein Networks Provides Evolutionary Insights. *Science* 314:1938–1941.

Kim R, Guo J. 2010. Systematic analysis of short internal indels and their impact on protein folding. *BMC Struct Biol* 10:24.

Kimura M. 1983. The neutral theory of molecular evolution. Cambridge: Cambridge University Press

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, et al. 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536:285–291.

Levy Karin E, Susko E, Pupko T. 2014. Alignment Errors Strongly Impact Likelihood-Based Tests for Comparing Topologies. *Molecular Biology and Evolution* 31:3057–3067.

Light S, Sagit R, Ekman D, Elofsson A. 2013. Long indels are disordered: A study of disorder and indels in homologous eukaryotic proteins. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* 1834:890–897.

Light S, Sagit R, Sachenkova O, Ekman D, Elofsson A. 2013. Protein Expansion Is Primarily due to Indels in Intrinsically Disordered Regions. *Molecular Biology and Evolution* 30:2645–2653.

Lin M, Whitmire S, Chen J, Farrel A, Shi X, Guo J. 2017. Effects of short indels on protein structure and function in human genomes. *Sci Rep* 7:9313.

Liu S, Wei X, Dong X, Xu L, Liu J, Jiang B. 2015. Structural plasticity of green fluorescent protein to amino acid deletions and fluorescence rescue by folding-enhancing mutations. *BMC Biochemistry* 16:17.

Liu S, Wei X, Ji Q, Xin X, Jiang B, Liu J. 2016. A facile and efficient transposon mutagenesis method for generation of multi-codon deletions in protein sequences. *Journal of Biotechnology* 227:27–34.

Lynch M. 2007. The origins of genome architecture. Available from: https://repository.library.georgetown.edu/handle/10822/548280

Mariani V, Biasini M, Barbato A, Schwede T. 2013. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* 29:2722–2728.

Marsh JA, Teichmann SA. 2014. Parallel dynamics and evolution: Protein conformational fluctuations and assembly reflect evolutionary changes in sequence and structure. *BioEssays* 36:209–218.

Mi H, Huang X, Muruganujan A, Tang H, Mills C, Kang D, Thomas PD. 2017. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic acids research* 45:D183–D189.

Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD. 2019. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic acids research* 47:D419–D426.

Mi H, Muruganujan A, Thomas PD. 2013. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res* 41:D377-386.

Mills RE, Pittard WS, Mullaney JM, Farooq U, Creasy TH, Mahurkar AA, Kemeza DM, Strassler DS, Ponting CP, Webber C, et al. 2011. Natural genetic variation caused by small insertions and deletions in the human genome. *Genome Res.* 21:830–839.

Mintseris J, Weng Z. 2005. Structure, function, and evolution of transient and obligate protein–protein interactions. *Proceedings of the National Academy of Sciences* 102:10930–10935.

Mitternacht S. 2016. FreeSASA: An open source C library for solvent accessible surface area calculations. *F1000Res* 5:189.

Montgomery SB, Goode DL, Kvikstad E, Albers CA, Zhang ZD, Mu XJ, Ananda G, Howie B, Karczewski KJ, Smith KS, et al. 2013. The origin, evolution, and functional impact of short insertion–deletion variants identified in 179 human genomes. *Genome Res.* 23:749–761.

Nielsen R, Slatkin M. 2013. An introduction to population genetics: theory and applications. Sinauer Associates Sunderland, MA

Nishizawa M, Nishizawa K. 2002. 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* 55:706–717.

Ohta T. 1972. Evolutionary rate of cistrons and DNA divergence. *J Mol Evol* 1:150–157.

Pascarella S, Argos P. 1992. Analysis of insertions/deletions in protein structures. *Journal of Molecular Biology* 224:461–471.

Penn O, Privman E, Ashkenazy H, Landan G, Graur D, Pupko T. 2010. GUIDANCE: a web server for assessing alignment confidence scores. *Nucleic acids research* 38:W23–W28.

Penn O, Privman E, Landan G, Graur D, Pupko T. 2010. An alignment confidence score capturing robustness to guide tree uncertainty. *Mol Biol Evol* 27:1759–1767.

Privman E, Penn O, Pupko T. 2012. Improving the performance of positive selection inference by filtering unreliable alignment regions. *Molecular Biology and Evolution* 29:1–5.

Rockah-Shmuel L, Tóth-Petróczy Á, Sela A, Wurtzel O, Sorek R, Tawfik DS. 2013. Correlated Occurrence and Bypass of Frame-Shifting Insertion-Deletions (InDels) to Give Functional Proteins. *PLOS Genetics* 9:e1003882.

Salari R, Schönhuth A, Hormozdiari F, Cherkasov A, Sahinalp SC. 2008. 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. p. 330–341.

Savino S, Desmet T, Franceus J. 2022. Insertions and deletions in protein evolution and engineering. *Biotechnology Advances* 60:108010.

Scherrer MP, Meyer AG, Wilke CO. 2012. Modeling coding-sequence evolution within the context of residue solvent accessibility. *BMC Evolutionary Biology* 12:179.

Shahmoradi A, Sydykova DK, Spielman SJ, Jackson EL, Dawson ET, Meyer AG, Wilke CO. 2014. Predicting Evolutionary Site Variability from Structure in Viral Proteins: Buriedness, Packing, Flexibility, and Design. *J Mol Evol* 79:130–142.

Shih C-H, Chang C-M, Lin Y-S, Lo W-C, Hwang J-K. 2012. Evolutionary information hidden in a single protein structure. *Proteins: Structure, Function, and Bioinformatics* 80:1647–1657.

Simm AM, Baldwin AJ, Busse K, Jones DD. 2007. Investigating protein structural plasticity by surveying the consequence of an amino acid deletion from TEM-1 β-lactamase. *FEBS Letters* 581:3904–3908.

Sinden RR, Potaman VN, Oussatcheva EA, Pearson CE, Lyubchenko YL, Shlyakhtenko LS. 2002. Triplet repeat DNA structures and human genetic disease: dynamic mutations from dynamic DNA. *J Biosci* 27:53–65.

Slodkowicz G, Goldman N. 2020. Integrated structural and evolutionary analysis reveals common mechanisms underlying adaptive evolution in mammals. *Proc Natl Acad Sci USA* 117:5977–5986.

Snir S, Pachter L. 2006. 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. p. 265–280.

Tao S, Fan Y, Wang W, Ma G, Liang L, Shi Q. 2007. Patterns of Insertion and Deletion in Mammalian Genomes. *Current Genomics* 8:370–378.

Taylor MS, Ponting CP, Copley RR. 2004. Occurrence and Consequences of Coding Sequence Insertions and Deletions in Mammalian Genomes. *Genome Res.* 14:555–566.

The Genomes Project C. 2015. A global reference for human genetic variation. *Nature* 526:68–74.

Thomas PD, Ebert D, Muruganujan A, Mushayahama T, Albou L-P, Mi H. 2022. PANTHER: Making genome-scale phylogenetics accessible to all. *Protein Science* 31:8–22.

Tien MZ, Meyer AG, Sydykova DK, Spielman SJ, Wilke CO. 2013. Maximum Allowed Solvent Accessibilites of Residues in Proteins. *PLoS One* 8:e80635.

Tóth-Petróczy Á, Tawfik DS. 2011. Slow protein evolutionary rates are dictated by surface–core association. *Proceedings of the National Academy of Sciences* 108:11151–11156.

Tóth-Petróczy Á, Tawfik DS. 2014. Hopeful (Protein InDel) Monsters? *Structure* 22:803–804.

Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, Yuan D, Stroe O, Wood G, Laydon A. 2022. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research* 50:D439–D444.

Won YJ, Hey J. 2005. Divergence population genetics of chimpanzees. *Mol Biol Evol* 22:297–307.

Woods H, Schiano DL, Aguirre JI, Ledwitch KV, McDonald EF, Voehler M, Meiler J, Schoeder CT. 2023. Computational modeling and prediction of deletion mutants. *Structure* 31:713-723.e3.

Yeh S-W, Liu J-W, Yu S-H, Shih C-H, Hwang J-K, Echave J. 2014. Site-Specific Structural Constraints on Protein Sequence Evolutionary Divergence: Local Packing Density versus Solvent Exposure. *Molecular Biology and Evolution* 31:135–139.

Zhang Z, Huang J, Wang Z, Wang L, Gao P. 2011. Impact of Indels on the Flanking Regions in Structural Domains. *Molecular Biology and Evolution* 28:291–301.

Zhang Z, Wang J, Gong Y, Li Y. 2018. Contributions of substitutions and indels to the structural variations in ancient protein superfamilies. *BMC Genomics* 19:771.

Zhang Z, Wang Y, Wang L, Gao P. 2010. The Combined Effects of Amino Acid Substitutions and Indels on the Evolution of Structure within Protein Families. *PLOS ONE* 5:e14316.

Zhao Z, Jin L, Fu YX, Ramsay M, Jenkins T, Leskinen E, Pamilo P, Trexler M, Patthy L, Jorde LB, et al. 2000. Worldwide DNA sequence variation in a 10-kilobase noncoding region on human chromosome 22. *Proc Natl Acad Sci USA* 97:11354–11358.

Figure 1

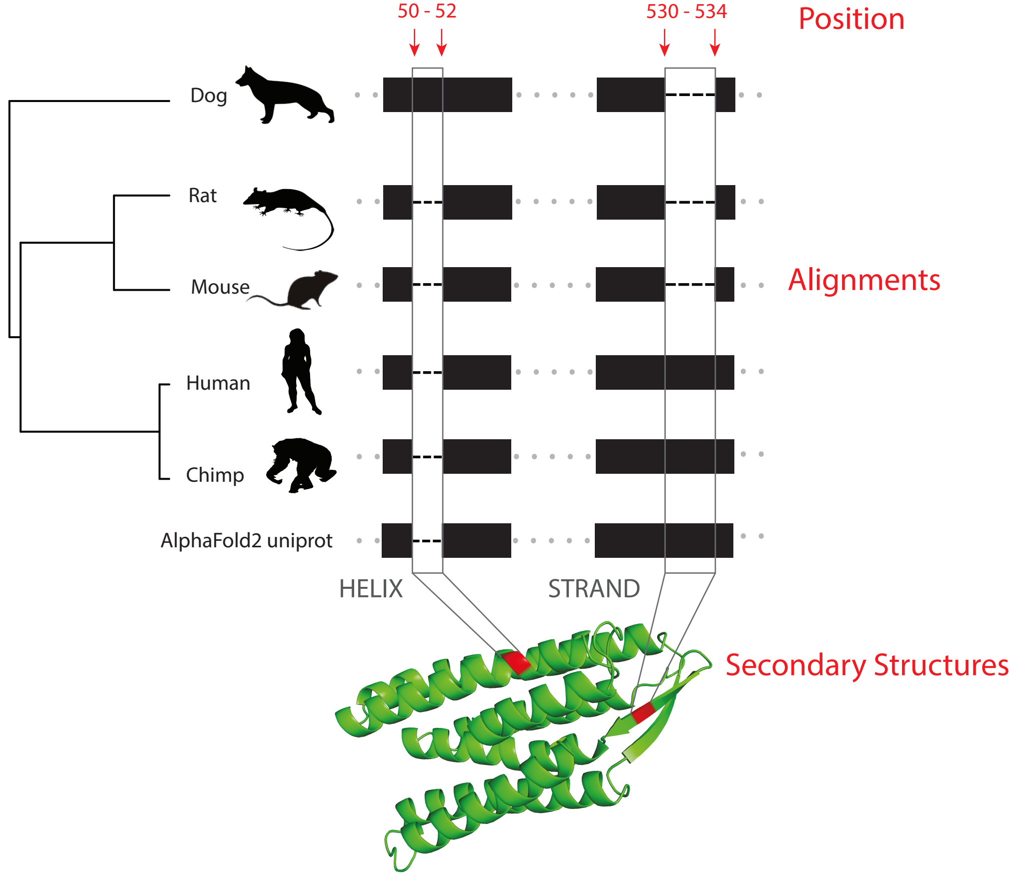
50 - 52
530 - 534
Position
Dog
Rat
Mouse
Human
Chimp
AlphaFold2 uniprot
Alignments
HELIX
STRAND
Secondary Structures

Figure 2

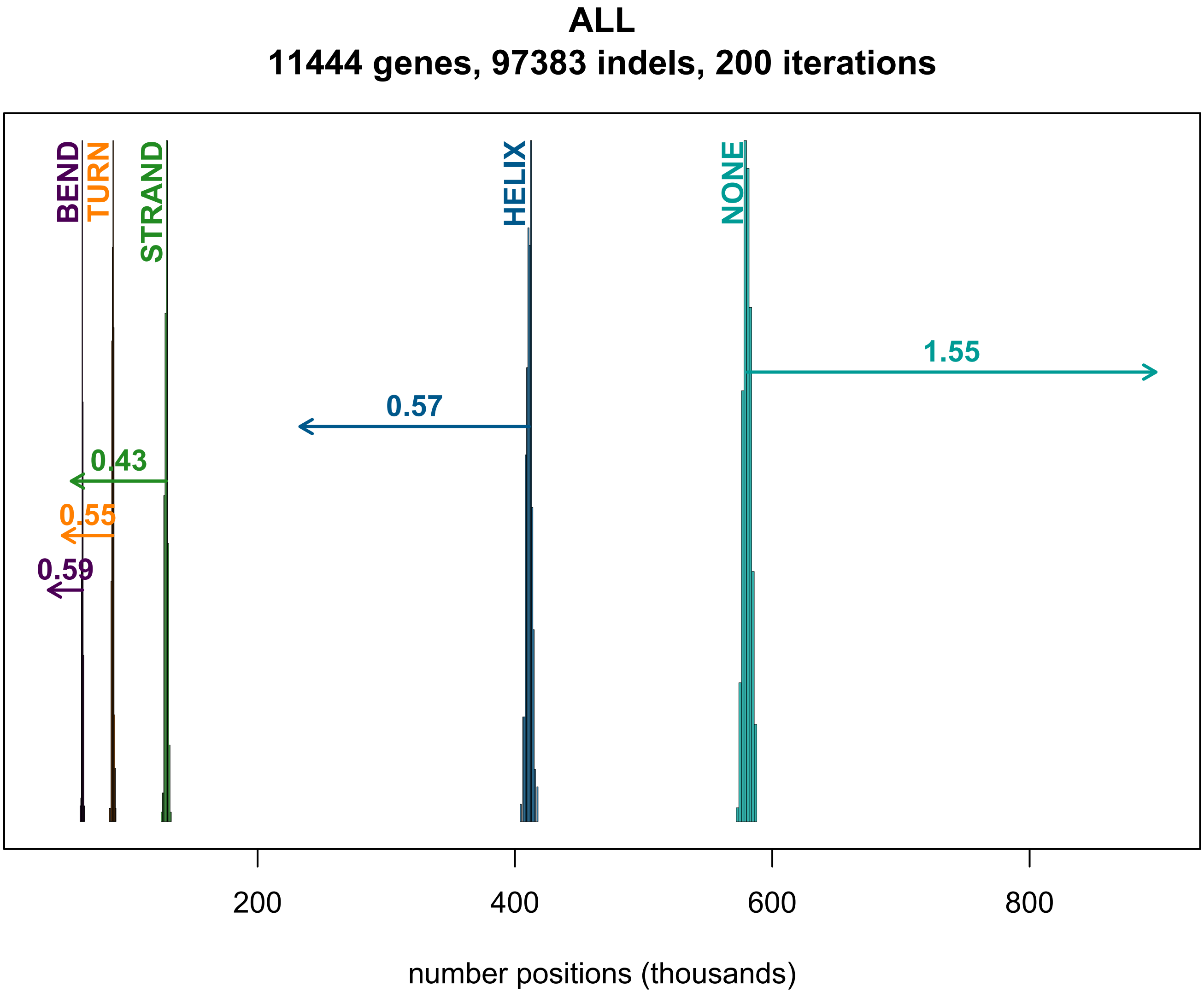
ALL
11444 genes, 97383 indels, 200 iterations
BEND
TURN
STRAND
HELIX
NONE
1.55
0.57
0.43
0.55
0.59
200
400
600
800
number positions (thousands)

Figure 3

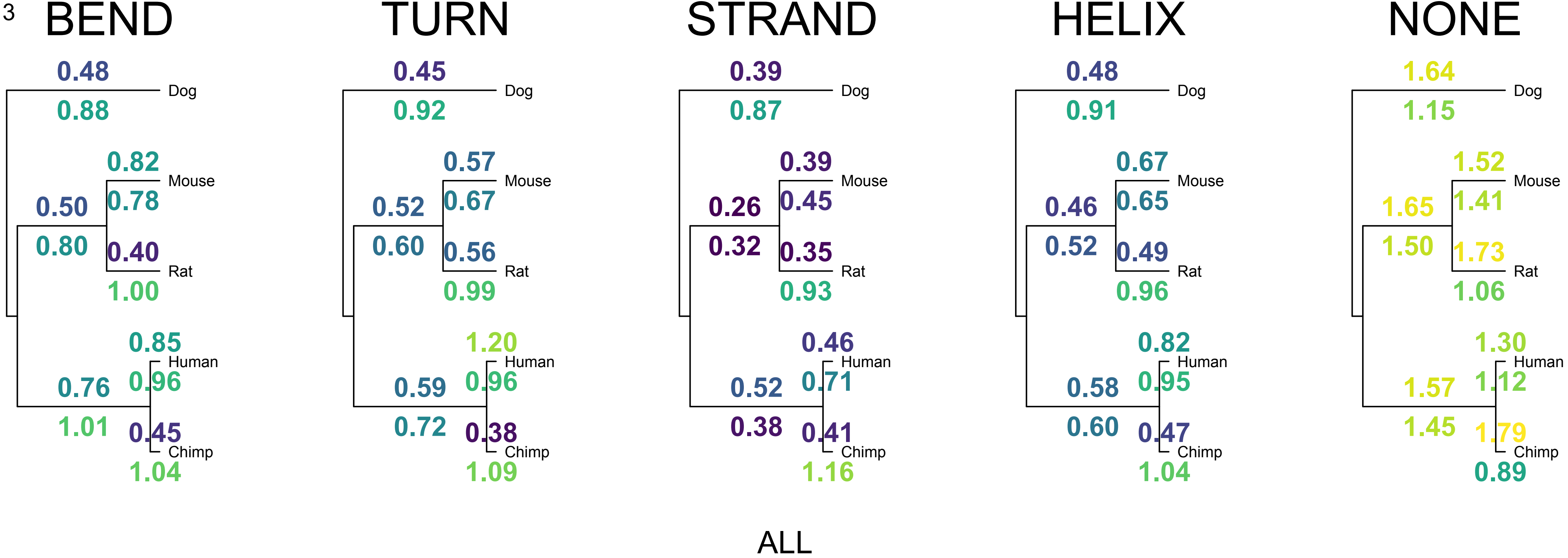
BEND
0.48
Dog
0.88
0.82
Mouse
0.50
0.78
0.80
0.40
Rat
1.00
0.85
Human
0.76
0.96
1.01
0.45
Chimp
1.04
TURN
0.45
Dog
0.92
0.57
Mouse
0.52
0.67
0.60
0.56
Rat
0.99
1.20
Human
0.59
0.96
0.72
0.38
Chimp
1.09
STRAND
0.39
Dog
0.87
0.39
Mouse
0.26
0.45
0.32
0.35
Rat
0.93
0.46
Human
0.52
0.71
0.38
0.41
Chimp
1.16
HELIX
0.48
Dog
0.91
0.67
Mouse
0.46
0.65
0.52
0.49
Rat
0.96
0.82
Human
0.58
0.95
0.60
0.47
Chimp
1.04
NONE
1.64
Dog
1.15
1.52
Mouse
1.65
1.41
1.50
1.73
Rat
1.06
1.30
Human
1.57
1.12
1.45
1.79
Chimp
0.89
ALL

Figure 4

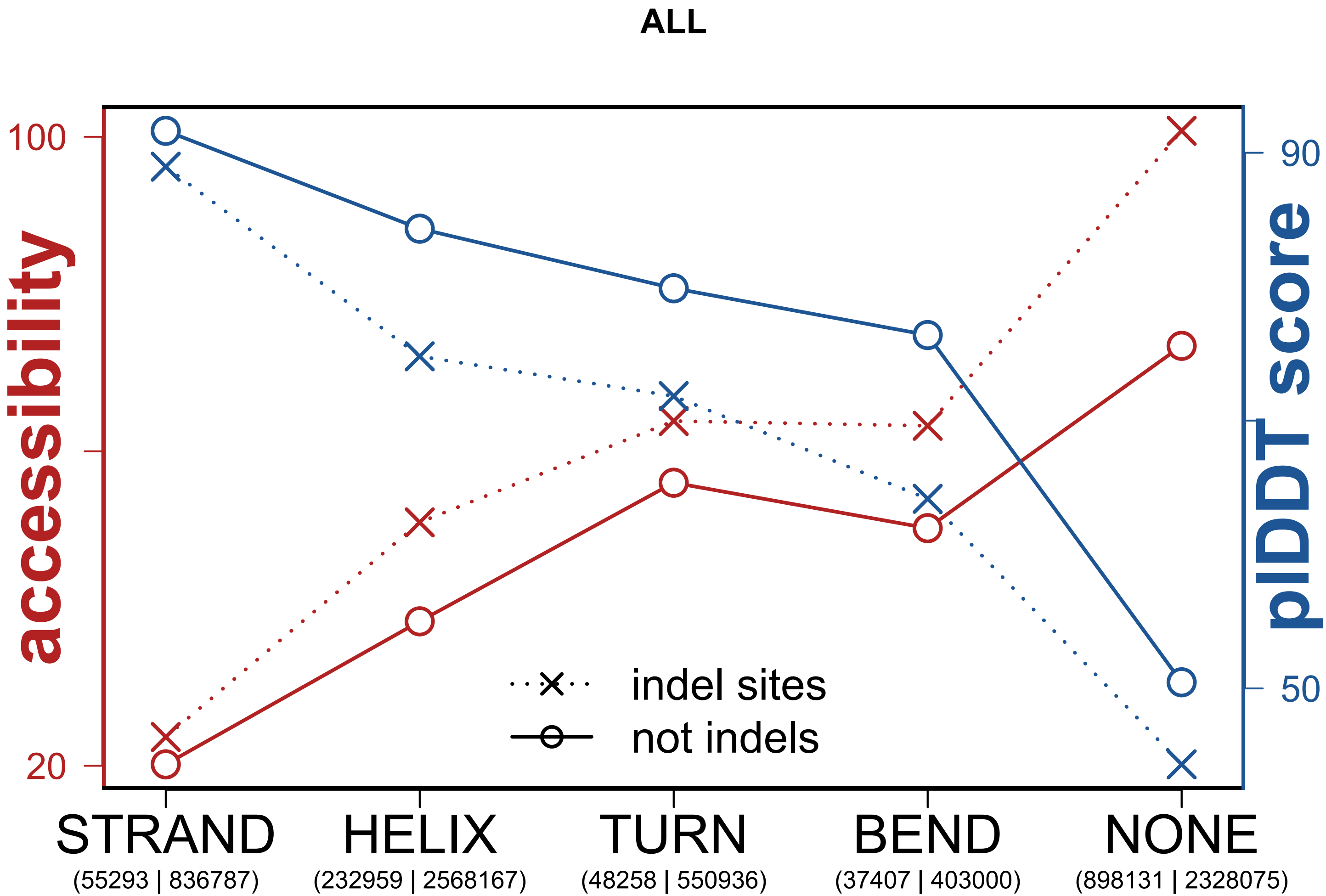
ALL
accessibility
100
20
plDDT score
90
50
indel sites
not indels
STRAND
(55293 | 836787)
HELIX
(232959 | 2568167)
TURN
(48258 | 550936)
BEND
(37407 | 403000)
NONE
(898131 | 2328075)

Figure 5

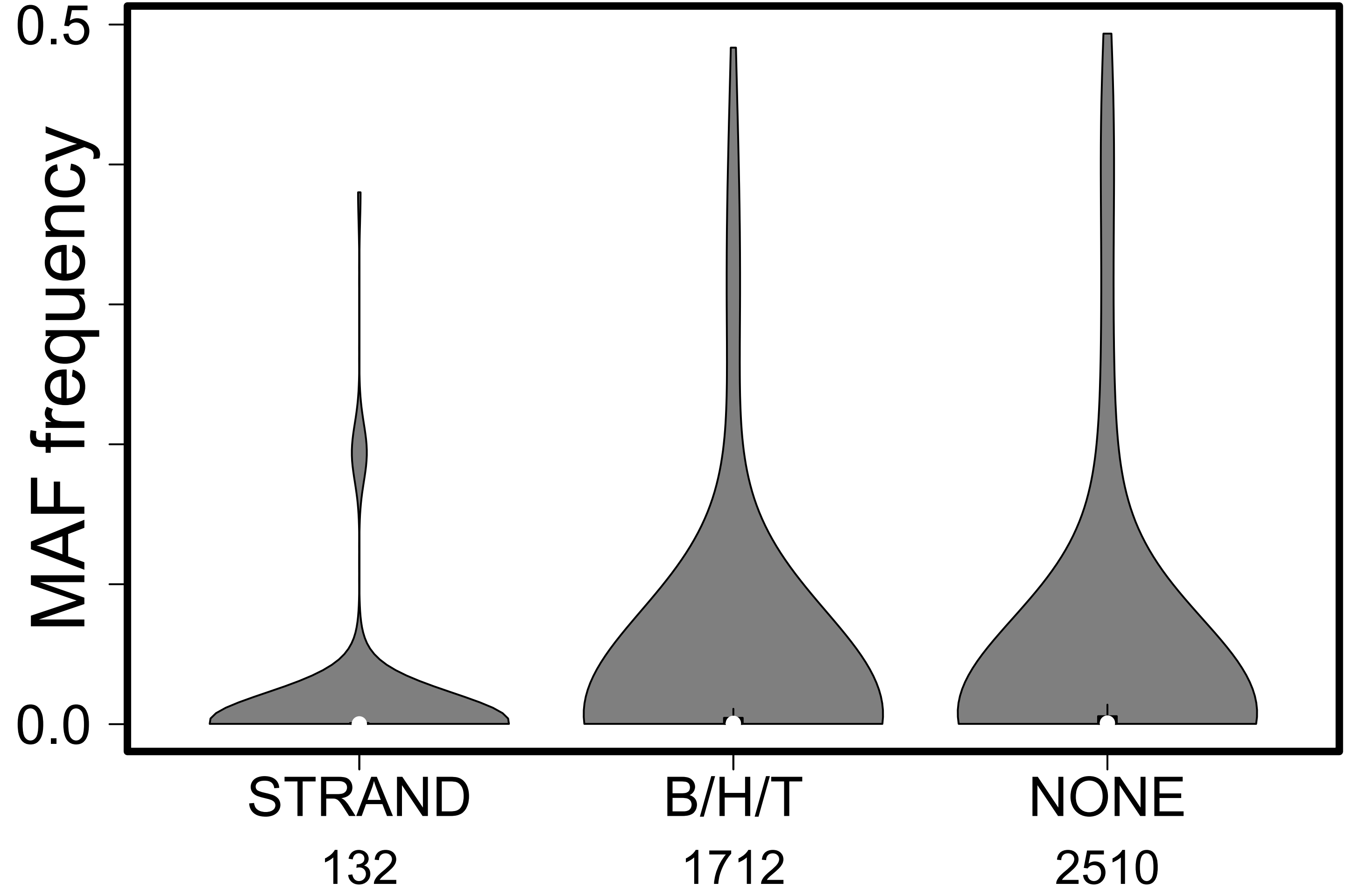

Table 1. Number of indels or codon mutations that overlap secondary structures. Observed=number of positions in alignments that map over each category. Expected=Number expected based on randomization. Codons are classified as invariant (Invariant), synonymous (Syn.) or nonsynonsymous (Non.). p=proportion of sites within their respective columns that fall within each category. This table is repeated as Supplementary Tables 2, after employing four different subsetting strategies.

	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

Table 2. AUC metrics for three Generalized Linear Models. Mean (standard deviation) AUC from 5 iterations of randomly sampling sites across alignments.

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)