Evolutionary and structural signatures of protein-coding function: synonymous acceleration, read-through, and structural impact of mutations.

by

Maxim Wolf

Submitted to the Department of Computational and Systems Biology
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Computational and Systems Biology

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2019

Author .....
Department of Computational and Systems Biology
August 27, 2019

Certified by .....
Manolis Kellis
Professor
Thesis Supervisor

Accepted by .....
Chris Burge
Chairman, Department Committee on Graduate Theses

Evolutionary and structural signatures of protein-coding function: synonymous acceleration, read-through, and structural impact of mutations.

by

Maxim Wolf

Submitted to the Department of Computational and Systems Biology
on August 27, 2019, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Computational and Systems Biology

Abstract

In this thesis I observe evolutionary signatures in coding regions to: (1) understand the sources of highly mutable coding regions in mammals; (2) to elucidate a new candidate function for a stop codon readthrough candidate gene, BRI3BP; and (3) to show how rapid sequence-based structure approximations can help predict the structural impact of amino-acid changes. (1) First, I searched for deviations from the evolutionary signatures of coding regions to recognize synonymous acceleration elements (SAEs) in protein coding genes. I showed that these are driven by an increased mutation rate, which persists in the human lineage, in otherwise evolutionarily-constrained protein-coding regions, providing an important resource to better characterize protein-coding constraint in mammals and within humans.(2) Second, I combined evolutionary signatures at the protein-coding and protein-folding level to characterize the functional implication of stop-codon readthrough in BRI3BP. I showed that this readthrough region has conserved spaced hydrophobic residues that pattern match to the c-terminal helix forming a coiled-coil-like domain. This change alters BRI3BP function from pro-growth to pro-apoptotic, similarly to VEGF-A. This suggests that readthrough-triggered apoptosis may represent a general mechanism for limiting growth of cells with aberrant ribosomal termination. (3) Third, I used rapid protein-structure approximation of burial of residues based on protein sequence to predict the structural impact of amino acid alterations. I show that the prediction can be improved over using exclusively the hydrophobicity change of the residue. Overall my work demonstrates how evolutionary and structural signatures can be used to predict highly mutational gene regions, readthrough function and structural impact of mutation.

Thesis Supervisor: Manolis Kellis
Title: Professor

Acknowledgments

Thanks to all of Kellis lab, past and present, for the support over the years.

Contents

List of Figures

List of Tables

Chapter 1

Introduction: Signatures of variation in coding regions

In the of the central dogma of biology, proteins are the ultimate step of gene expression. Variation in coding regions has the most direct and measurable impact on protein function. However, unlike transcriptional effects, which are relatively easy to measure experimentally, the functionality of proteins and their functional changes is much more difficult to quantify. In order to better understand the impact of coding region mutations on protein function it is essential to create models that allows us a way to estimate the effects of these changes. In this challenge, the lens of evolution gives us a unique perspective, because we can assess a set of information in order to observe the impact of mutations that have affected species over millions of years.

To effectively gain insight from evolutionary information, it is essential to focus on evolutionary distances that are representative of the problems that are being addressed. In this thesis I explore specifically how evolutionary signatures elucidate function of human genes. To this end, I consider the timescale of mammalian evolution combined with the evidence of variation in the human population. The timescale difference allows us to gain a broader view of the mutational landscape. The differences in mammals give us a way to observe sufficiently many mutations to easily identify regions of neutral selective pressure, but not so many mutations where we would expect a myriad of repeat or compensatory mutations. Comparing these

changes to variation within the human population allows us to see if the predictions that we are making based on evolutionary signatures can be relevant to the human genes.

For the cases where evolutionary signatures help us identify new patterns, we can use them to ascribe function to gene regions or to whole genes. Evolutionary signatures have been used to distinguish coding from non-coding regions using PhyloCSF [1], to group genes into functional networks by similarity of evolutionary signatures [2], and to study common regulation signatures in phenotypically linked genes [3]. In chapter 2 I discuss how these patterns in coding regions allow us to expand on the work identifying synonymous constraint regions [4] to identify synonymous acceleration elements which are closely related to regions of elevated local mutation rate. By taking mutation rate into account in evolutionary models, we can enhance functional predictions for specific variants which can improve experimental design for clarification of protein function.

In the ideal case, evolutionary signatures, when analyzed in greater detail within a particular gene, can provide for direct and experimentally testable hypotheses. Stop codon readthrough is a particularly interesting topic where the observation of evolutionary patterns can help identify functional elements. Generally, the process of ignoring the stop codon is aberrant in cells even though in many genes we can still observe ribosome footprints in the 3'UTR despite the fact that resulting proteins are likely non-functional [5]. However, we know that in Anopheles and Drosophila readthrough appears to be much more abundant and functional [6] in part using evolutionary evidence. The same types of signatures were used in identifying human readthrough candidates [7] among them is the BRI3BP gene. In chapter 3 I show how this prediction can be further supported by other computational methods and provide some preliminary experimental evidence in support of the readthrough region being instrumental in the reversal of BRI3BP's function.

Evolutionary information combined with structural data can be utilized to help identify functional regions in proteins [8]. However, even now some proteins remain difficult to impossible to crystallize and *de novo* structural prediction methods still

have many limitations [9],[10]. To reduce the the complexity of predicting secondary protein structure it is beneficial to be able to make approximations that are sufficient to still allow us to obtain biological insight. In chapter 4, I discuss the potential applications of one such model [11] to predict the structural impact of amino acid changes. The insights for how to use relevant information from structural approximations can help our understanding of protein function in cases where crystal structures are not yet available.

My hope is that in a field inundated by new data, my analysis and others like it can provide a means of directing the scope of future analysis in order to more effectively extract meaningful biological data.

Chapter 2

Analyzing Regions of Unusual Synonymous Constraint

2.1 Introduction

Evolutionary signatures in coding regions can provide important insights on gene function by distinguishing the type of selective constraints present in different sections of the gene. The highly-characteristic signatures of synonymous rates in protein-coding regions give us a unique background for recognizing overlapping functions and regions with unusual selective pressures on neutral base positions within protein-coding exons.

The presence of synonymous constraint can be detected in the coding regions of human genes using evolutionary information. However, there may be many potential causes of constraint. I compare synonymous constraint on the scale of mammals to the mutations on the scale of human population to observe how constraint has changed over evolutionary time. Depending on the kinds of features that influence constraint we may expect them to have an effect over different timescales.

In this chapter, I focus on unusually synonymously constrained regions(UCRs). While regions with unusually high synonymous constraint (Synonymous Constraint Elements (SCEs)) are relatively well studied, regions of unusually low synonymous constraint (Synonymous Acceleration Elements (SAEs)) are relatively less under-

stood. I demonstrate that these SAEs are driven by high local mutation rates which persist in the human lineage.

To this end, I describe the effects of codon bias on identifying UCRs as well as the presence of human RNA binding protein and transcription factor (TF) protein binding sites, their overlap with human single nucleotide variants (SNVs) and *De novo* mutations, and allele frequencies and genome wide association study enrichment in those variants. Additionally, I looked at the deleteriousness predictions of for the mutations that occur in the UCR regions.

For SCEs, I also analyzed their location within protein coding regions relative to position nucleosomes, and exon boundaries. I also studied the sequence content of the SCEs to determine whether I could identify new or old known sequence motifs or signatures. Finally, I worked with Rachel Sealfon to look at SCEs in viruses. In this project I helped create FRESCo which was used to identify overlapping genes and regulatory elements in viral genes. I looked at RNA secondary structure motifs that were identified and the pairing of nucleotides within constrained regions.

Our results suggest that local mutation rate variation within a gene can have a strong impact on the ability to detect functionally-important residues in human variation studies. We provide a useful proteome-wide benchmark for calibrating such studies in the detection of specific residues under selection.

2.2 Background

Algorithms to identify synonymous constraint elements (SCEs) have been used to identify fragments of coding regions that contain additional overlapping functions based on the fact that they have unusually low rates of synonymous substitution, implying that purifying selection has acted not only to conserve the protein sequence but also to prevent synonymous substitutions [4]. However, the local synonymous substitution rate data from these studies could also identify Synonymous Acceleration Elements (SAEs), which have not been analyzed previously.

SCEs were previously shown to contain functional elements that create additional

constraint to the constraints imposed by conservation of protein coding sequence. In a previous analysis, an evolutionary model based on codon substitution frequencies in a 29 mammal alignment was used to identify regions of overlapping constraint (SCEs) using a window size of 9 codons [4]. SCEs were shown to be enriched in cassette exons and ends of genes, regions that often contain regulatory elements, such as Exonic Splicing enhancers and RNA Binding protein motifs. They were used to identify an alternative transcription initiation site of a gene. It has also been suggested that SCEs tend to overlap disordered segments of proteins because the decreased protein-level constraint in these segments provides greater flexibility for overlapping functions [12].

We use our recently-developed model of codon evolution, as implemented in the FRESCo software [13], beginning with an input phylogenetic tree topology and global codon alignment, and computing maximum-likelihood tree branch lengths and a codon substitution matrix parametrized by synonymous and nonsynonymous substitution rates. Maximum-likelihood values for these two parameters are then computed on sliding windows.Finally, a likelihood ratio test is used to identify regions, SCEs, in which the synonymous rate of a window is significantly lower than the average synonymous rate across the full open reading frame (ORF).

In our analysis we apply FRESCo using the alignments for each gene and the tree of 29 mammals to determine regions with unusually high or low synonymous substitution rates within mammalian genes, to further characterize SCEs, and to investigate the cause of unusually high synonymous rates in SAE regions.

2.3 Identifying Unusually Constrained Regions(UCRs)

2.3.1 Computational methodology for UCR identification

I applied the FRESCo framework to CCDS regions from *Homo Sapiens* release 9 in the hg19 human reference genome assembly to obtain a new set of SCEs for humans. I adjusted the FRESCo algorithm to also identify SAEs – regions that had a significantly higher synonymous rate within the sliding window as compared to the average

for that particular transcript. We identified the regions using 9 codon windows and a 5% significance threshold. This process is graphically represented in Figure B-1.

The FRESCo Codon model is trained on the given alignments by looking at transition and transversion probabilities and the probabilities of mutation of each individual base given its codon position. This also takes into account the phylogenetic relationship between the different species in the alignment. An MG94 model is used to figure out the ratio between synonymous and non-synonymous substitutions. These parameters are assembled together to formulate a 64 by 64 matrix. For each window FRESCo learns local synonymous and nonsynonymous rates. Subsequently the window specific rate it compared to the gene-wide rate to see if the window specific rate is significantly better at describing the substitutions than the transcript-wide rate giving a likelihood ratio p-value that is used to determine the UCRs. The p-values are Bonferroni corrected for the length of the gene and windows with significant p-values are marked as UCRs.

2.3.2 Overall SAE and SCE statistics

Overall, we identified 79,261 non-overlapping SCEs in 22,615 transcripts and 79,275 non-overlapping SAEs in 22,785 transcripts with average lengths of 45.2 and 39.8 bases respectively. The SCE and SAE regions account for 8.2% and 7.2% of the analyzed CDS, respectively. We created a public track hub, "Human unusually constrained regions", that allows viewing the SCE and SAE regions in the UCSC genome browser.

2.3.3 Lack of codon bias and CpG enrichment in SAE and SCEs

Since we learn a different codon transition matrix for each individual transcript, we lack sufficient codon-to-codon substitutions to optimally learn the full matrix each time. In order to make sure that the codon model was picking up insightful evolutionary signatures despite this disadvantage, we analyzed UCRs for biased codon distributions. We inferred the mammalian ancestor using a Kimura 1980 evolution-

ary model [14]. Changes were computed to the human lineage with respect to the ancestor. The results showed that there is no bias toward the identity of codons that are enriched in SAE regions or depleted in SCEs. Synonymous substitutions in these regions follow a genomewide pattern with proportionally more and proportionally fewer substitutions in SAE and SCEs respectively (Figure B-2 B). This pattern remains irrespective of which species is chosen to be compared to the ancestor (Figure B-2 D).

We next investigated the extent to which SAE and SCE regions result from the high mutability of CpG dinucleotides, which would increase or decrease the synonymous substitution rates for regions in which codons with an ancestral CpG in the second and third position are enriched. Mutations here resulting in C to T on the forward strand would not be synonymous but a C to T on the opposite strand would result in a G to A transition. However, we found that guanines in the third position of such codons are not significantly more or less likely to mutated to adenine than to any of the other bases when compared with control regions (Figure B-2 C).

To test the possibility that SAE regions result from positive selection for more efficient codons, we analyzed the efficiency of codons gained in SAE and SCE regions. Previous studies have estimate codon efficiency using tRNA abundance and frequency of codons in highly expressed genes [15]. Using these estimates, we looked at the rates of mutation to or from the most efficient codons in our regions of interest. Overall, there is a genome-wide loss of the more efficient codons. In SAEs and SCEs, rates of this codon loss are as would be expected from their synonymous rates (Figure B-3) with SAE regions seeing more rapid loss while SCE regions see fewer changes. In the majority of species, the codons that are lost within SAE regions correlate weakly with codon efficiency suggesting that that positive selection due to maintaining codon efficiency is not the driver of the substitutions. It is possible that the general loss of efficient codons has to do with population size differences in human relative to the mammalian ancestor. We see a similar pattern of codon loss when comparing ancestral frequencies to those in some non-human species like dog and horse, but other species have a much weaker correlation between efficiency and gain/loss rate

(Figure B-4)

2.4 *De novo* mutations in humans suggest SAE regions have higher mutation rates

We subsequently tested the hypothesis that the excess of synonymous substitutions in SAEs are caused by higher mutation rates, using a database of *de novo* variants, denovo-db [16]. If the hypothesis were true, we would expect an enrichment of de novo variants in SAE regions. We found a mutation rate of $3.510 * 10^{-8}$ per base in SAE regions, a 1.24-fold enrichment relative to random subsets of coding regions of size similar to SAEs and SCEs, which had a mutation rate of $2.44 \pm 0.19 * 10^{-8}$ de novo mutations per base ($p \ll 0.01$). Crucially, the mutational enrichment is among both synonymous and nonsynonymous substitutions, suggesting that identifying SAE regions can help locate regions of high overall mutation rate and not just high synonymous substitution rate. Interestingly, SCEs have a slight depletion in both synonymous and nonsynonymous observed mutations, with an overall rate of $2.27 * 10^{-8}$ mutations per base ($p<0.01$), suggesting that enough of the mutations in SCEs are embryonic lethal to reduce observed local mutation rates and highlighting the functional importance of the overlapping elements identified by SCEs (Figure B-5(C))

2.5 Human variant distribution in UCRs

2.5.1 SAE regions have high frequency variants in the human population

In order to investigate the importance of UCR containing gene regions we looked at the polymorphisms in the overall human population by studying variants in the 0.3 release of the ExAC database. Among the full database there was little deviation from the expected number of Single Nucleotide Variants (SNVs) in SAE and SCE regions. However, stratification by allele frequency enables us to see some differences.

The SCE enrichment in private mutations in ExAC suggests that the mutations are relatively rare and when they do occur they are kept in relatively low frequencies in the human population. This further supports the hypothesis that mutations in SCE regions tend to be deleterious and there is selective pressure to remove those mutations from the population. In contrast SAE regions are depleted in private mutations but show minor enrichment in relatively low frequency alleles (<1 in 1000 individuals) in the ExAC population. This is consistent with SAE regions having higher mutation rates and possibly lower selection a the synonymous sites. There is also no evidence to suggest that they would be positively selected allowing them to rise to higher allele frequencies. With positive selection we would have expected a steady increase for variants at higher frequencies, instead we observe most of the mutations at allele counts of about 8 and then a stable non-increasing rate for higher allele counts. (Figure B-5 A,B,D)

2.5.2 Synonymous GWAS variants enrichment proportional to local synonymous rate

We then confined the analysis to alleles significantly associated with traits in genome-wide association studies (GWAS). While conserved regions would likely have more causal mutation, we expected to see fewer total GWAS hits in those regions relative to those that have less purifying selection. The mutations in UCR regions follow this pattern: SCEs have fewer GWAS-associated synonymous SNPs than similar-length non-UCR regions while the SAE regions have more . Interestingly, the median synonymous mutation in SAE is of slightly higher allele frequency than SCE in ExAC (Figure B-6). This increase, and a proportional decrease in allele frequency for SCE regions, could suggest that while SCEs have more functional GWAS implicated rare synonymous variants, SAE synonymous variants may be found in GWAS just by virtue of being more common.

2.5.3 Deleteriousness of SNVs in UCRs

In order to better predict the impact of the SNPs observed in the UCR regions we analyzed the predicted deleteriousness of the variation using Polyphen 2[17]. It is important to consider that the fact that synonymous and nonsynonymous rates would affect the Polyphen prediction since those rates are parameters in the model so we separated the regions by non-synonymous rates. The data shows that SAE regions have the highest frequency of mutations that are predicted to be deleterious by Polyphen (FIGURE B-7). Moreover, given increasing evidence that rare and private mutations can have major impacts on disease ([18], [19]) it is likely that some of these mutations are even more deleterious than predicted.

2.6 Genome Features in relation to UCRs

2.6.1 RNA binding protein and transcription Factor binding sites in UCRs

To verify the effects of higher mutation rate in SAE and increased purifying selection in SCEs we looked at known binding motifs present in the two types of regions. Using a database of Transcription factor binding sites (TFBS) and the motifs for RNA binding proteins identified by Ray et al we compared the incidence of these motifs in the UCR regions to identically sized regions randomly sampled from non-UCR containing gene regions among genes that had at least one UCR. For the TFBS we observed an enrichment in SCE regions and a depletion in SAE regions (Figure B-8 A). For the SCE regions, this is consistent with previous results suggesting that SCE regions have overlapping motifs. However it is important to note that these motifs in coding regions are very rare. SAE regions were depleted for these motifs which is not surprising since it would be difficult for the organism to maintain a functional motif in a highly mutable region, and, conversely, the presence of such a functional motif would decrease the synonymous rate. In the case of the RNA binding protein motifs we observed a different result, with no enrichment in SCEs and a small but significant

enrichment in SAE regions (Figure B-8 B). We hypothesize that the lack of enrichment in SCE regions is due to the fact that many of these motifs in coding regions are non-functional, so a small enrichment of functional motifs would be drowned out by the large number of non-functional ones that are neither enriched nor depleted. The enrichment of the RNABP motifs may be explained by them having less negative selection against having new motifs for translational control than for transcriptional control due to a further downstream function of TF regulation relative to RNABP which likely has a weaker impact on cell state. There may also be opportunities for positive selection due to a smaller and more localized effect on the specific gene.

The RNABP motifs were obtained from Ray *et al.* [20] and the TFBS motifs were obtained from the JASPAR database [21]. These motifs were input into FIMO [22] to find occurrences for the given motifs in the CCDS sequences. This was repeated for the SAE and SCE regions separately to ensure that we are not counting elements overlapping regions boundaries. The FIMO output was processed to quantify the number of bases covered by motifs in each sequence relative to the overall sequence size.

2.6.2 Identifying new motifs in SCEs

I used *de novo* motif finding algorithms [23] as well as looking for N-mer enrichment to analyze the combined SAE regions for new motifs. While it doesn't look like any particular motif is present in a large subset of SCEs certain 8-mers were extremely enriched (p-val $< 10^{-200}$). Some notable examples include the CTGCTGCT, CTACCTCA, and ACCAAAGA motifs. The former is a frame specific motif which was rather interesting to see given the lack of proportionally strong CTG codon enrichment. The frame specificity might suggest that there may be some specific function relating to the repeat use of the specific t-RNA for CTG (which also happens to be the most common codon for leucine). The CTACCTCA motif is the most enriched non-frame specific motif which makes it the ideal candidate for seeking some non-coding related function that generates synonymous constraint. Relative to the other two motifs the ACCAAAGA motif occurs just about as frequently on the reverse

strand suggesting that there may be some DNA related reason for constraint on the synonymous positions. Nevertheless, more research would be needed to ascertain the exact function of these and other enriched motifs.

2.6.3 SCE locations relative to nucleosome positions and miRNA binding sites

It was found that many SCEs span or are nearby exon-exon junctions. An unexpectedly large number of sites matching miRNA target sites were located within SCEs, however, those matches did not include many of the miRNA sites that were known to be in coding regions. Sequences consistent with exonic splicing enhancers were detected in many SCEs. Due to the low information content of the motifs it is difficult to remove false positives from this set. Relative to randomly picked coding regions, SCEs are more likely to occur near the nucleosome edges. With a median SCE size of about 90 there was a significant enrichment of SCEs with a 60 or fewer base overlap with the nucleosome covered regions. To identify relatively stable nucleosomes I used the data from from Valuev *et al* [24]. The real data was compared to simulations of shuffled SCEs to random positions within genes and a combination of shuffled SCEs and nucleosome positions. The results are shown in Figure B-9. The enrichment at nucleosome position boundary suggests that there may be some additional constraint required for precise nucleosome placement. This supports the theory that fixed nucleosome positions may be required for gene function rather than allowing the nucleosomes randomly distributed among the DNA.

2.7 Comparison to primate and vertebrate constrained elements

To see how resilient the identification of Synonymous constraint elements is, I identified UCRs using vertebrate and primate alignments. The primate alignments didn't have sufficient mutations to accurately identify UCRs and the UCRs were relatively

randomly distributed throughout the transcripts. The weakness of the signal relative to the noise of random mutations resulted in poor overlap of UCR identification relative to the mammalian regions.

In Contrast, the UCRs based on the tetrapod alignment were a lot more similar to the mammalian regions. Qualitatively the tetrapod UCRs are larger and many merge UCRs that were identified as separate regions in mammals. While some genes like BRI3BP (Figure B-10) were relatively consistent, others seemed to have highly fluctuating synonymous rates throughout the gene. Overall there was a 70% overlap of the SCE regions and a 57% overlap in the SAE regions identified from the mammalian alignment in the vertebrate UCRs, though the vast majority of the tetrapod regions overlap at least one mammalian UCR. The inconsistencies are likely related to the quality of the alignment and future analysis should include more stringent filtering for alignment quality focusing on the transcripts which span the full vertebrate tree.

2.8 Conclusion

FRESCo can be applied to mammalian lineages and still successfully identify multi-functional regions by locating regions of excess synonymous constraint. The output data from FRESCo can be further analyzed to identify regions with unusually high synonymous substitution rates. In identifying these regions, FRESCo successfully avoids codon bias in region identification and at our current window size the SAE regions were not affected by local increases in mutation rates due to CpG mutability. SAE regions tend to have higher mutation rates, a depletion in transcription factor binding sites and increased variation in the human population.

Evidence of high mutation rates in SAE regions combined with large quantities of mutations in humans in these regions suggest that we expect to encounter new deleterious non-synonymous mutations in regions that have had low non-synonymous rates throughout evolutionary history. This gives us some possible targets for future research since there is either a strong selection against preserving those mutations or the regions have become more mutable at non-synonymous sites in recent evolutionary

history. It would be an interesting topic for further research to see if there are any recent regulatory changes that may be associated with these regions' elevated mutation rates.

Additional insight about mutation rate variation could change the way we look for functional elements in genes. Combining knowledge of SAE regions with non-synonymous rate variations across the genes has the potential of identifying mis-annotated coding regions since SAE signatures combined with really high non-synonymous rates could be evidence of a non coding region. Alternatively, when an SAE regions is paired with a low non-synonymous rate, it could help identify essential protein components since the level of conservation is unexpectedly high given the region's mutability. Additionally, methods for identifying SAE regions could be used to improve evolutionary models that assume uniform mutation rates by providing a way to account for mutation rate variation across the gene and enhance the identification of functional elements.

Chapter 3

BRI3BP Readthrough Conservation and function

3.1 Introduction

Evolutionary signatures can be used to identify functional regions in genes. In the case of BRI3BP, signatures representing coding constraint were used to identify the region beyond the stop codon as coding-like, which raised the possibility of this gene containing a functional readthrough region. Subsequently, structure-level evolutionary signatures allowed for the identification of a conserved motif in the gene extension which would allow for the formation of a coiled-coil-like structure at the C-terminal end of the protein. This interaction, combined with wild type BRI3BP's role in the apoptosis pathway, allowed for the hypothesis that the readthrough variant of BRI3BP could cause a change in the gene's function by blocking the normal interaction partners of BRI3BP's tail end. In this chapter, I describe the structural evidence which was used to form this hypothesis, the CRISPR editing experiment I performed to test the function of the readthrough region in the wet lab, and the potential for readthrough-triggered apoptosis as a way for organisms to limit growth of dis-regulated cells.

3.2 Background

Functional stop codon readthrough is common in flies and other insects with over 280 candidates in *Drosophila* ([7]; [25]), there are substantially fewer cases of readthrough identified in humans. While there is still little known about the readthrough genes in general, they tend to fall into categories of regulated and constitutive readthrough. C-terminal regions in the constitutively readthrough peptides may to contain localization signals which allows for fractional distribution to sites with and without the signal proportional to the rate of readthrough. In the *Drosophila* set there is an overabundance of trans-membrane proteins in this set (P 0.01) supporting the theory that readthrough may contribute to organelle localization while the normal variant would localize to the cell membrane. This can be an efficient way to differentially localize a membrane protein from the same transcript and use rare readthrough to populate organelles with less membrane volume with the same proportion of protein. Out of the seven computationally identified human gene candidates (MAPK10, BRI3BP, APQ4, OPRL1, OPRK1, SACM1L, ACP [7]) six are known to be transmembrane proteins. However, these candidates do not possess any localization signals in the readthrough domain suggesting a different method of function for readthrough in humans.

One particularly interesting experimentally verified case of human readthrough is the VEGF-A gene. Under normal conditions this gene promotes cell growth but the programmed readthrough extension (VEGF-Ax) has been shown to have anti-angiogenic properties [26]. This example begs the question of whether stop codon readthrough is a good way of reversing a gene's function when the conditions for this reversal are favorable. From the list of known non-verified potential readthrough candidates, BRI3BP is the most promising to have a functional readthrough region.

The BRI3BP protein is localized to the ER and the mitochondria where it is involved in cell growth regulating pathways. BRI3BP is frequently found to be amplified or mutated in many types of cancer (Figure B-11). BRI3BP was shown to be an oncogene by transfection into mouse cells resulting in cancerous growths. BRI3BP overexpression resulted in matrigel invasion of cells and elevated p53 and PKC cas-

cade levels [27]. Non-readthrough BRI3BP is thought to function by interacting with p53 via a binding partner - BRI3. Under normal conditions BRI3 sequesters p53 in the mitochondria preventing apoptosis and encouraging cell growth. Unexpectedly, BRI3BP has also been implicated in having an inverse function in cells are treated with etoposide [28], where increased BRI3BP expression leads to cell death. This effect can be attributed to the gene having a different mode of function under drug induced stress. Additionally there is some weak evidence for ribosome footprints signal in the readthrough region (Figure B-12), but the functionality of this region has never been studied. Stop codon readthrough could be culprit behind the change in BRI3BP function under drug induced stress since there were no identified alternate forms or transcripts that could cause the phenotype change.

3.3 Computational evidence for readthrough in BRI3BP

Studying evolutionary signatures made it possible to identify BRI3BP as a candidate gene for programmed stop-codon readthrough.

3.3.1 Constraint signature in readthrough region

BRI3BP was first identified as a candidate for programmed readthrough by observing that this gene's post stop codon region had a positive PhyloCSF score combined with a well preserved secondary in-frame stop codon [7]. In this case it was observed that the region between the two stop codons has evolutionary properties that are more similar to a coding than a non-coding region. While BRI3BP did not have the highest conservation score compared to the other candidates it was especially of interest because it was implicated as an oncogene.

3.3.2 Preservation of hydrophobic residues

I investigated the roughly 30 codon length readthrough region for any large conserved RNA structures, RNA motifs, and Protein sequence motifs which could have explained

the need for additional constraint. There were no significant hits in the readthrough region. However there was is a pattern consisting of hydrophobic residues in the tail region across many species in the alignment. There are identifiable, spaced, and strongly conserved hydrophobic residues over a hydrophilic background. These residues could potentially interact with similarly spaced residues on the C-terminal helix of the non-readthrough, normal length, protein (Figure B-13). I used HHPred to predict the secondary structure of BRI3BP. While the majority of the protein is several transmembrane helices connected by short loops, the C-terminal helix is located in the cytoplasm. It also contains a coiled-coil-like domain which would be largely open for interactions. When the read-through region is added to the secondary structure prediction an additional conserved cytoplasmic helix becomes evident despite the lack of conservation of specific residues. The repeated motif of spaced hydrophobic residues followed by a conserved helix was suggestive of a potential structural interaction in the BRI3BP extension.

3.3.3 Protein structure prediction

Since most of the protein is transmembrane it is not possible to model in its entirety. However, modeling the C terminal cytoplasmic region by itself should be a sufficient approximation of the structure since only a small fraction of the protein is cytoplasmic and residues are likely to interact more in closer proximity where they have more degrees of freedom than in the membrane.

In order to model the interaction of the C-terminal region I used QUARK [29] for *de novo* structure prediction of the C-terminal region and SWISS-MODEL [30] for homology modeling of the C-terminal region. Both predicted structures involved close interaction between the readthrough region and the C-terminal helix. To assess the stability of the interaction I also predicted structures for shuffled amino-acids, shuffled nucleotides, and a frameshift causing mutations (Figure B-14 A)

Both structural prediction methods are in agreement with HHPred [31] about the structure of the C-terminal helix in the region proceeding the first stop. The homology based modeling predicts wider area interaction between this helix and the

readthrough regions. Both models show that it would be energetically favorable for the extension to interact with the adjacent helix and there is enough flexibility to form a turn for the interacting helix. The different structure of the shuffled region suggests that there is some constraint about the order of the amino acids that form those interactions. That constraint is likely relatively weak since the frameshift also demonstrates a similar interaction pattern with the main helix.

Nevertheless, despite the differences in amino acid composition, the predicted structures for other vertebrate species show strong conservation of the interacting structure. With the exception of the lineage where mouse and rat have a readthrough destroying frame-shift, the remainder of the species form a similar interacting structure with the two terminal helices (Figure B-14). While it is possible that the other species have lost the readthrough functionality, the general pattern of helix-helix interaction across the vertebrate tree adds to the evidence of evolutionary conservation that the readthrough region of BRI3BP is programmed to be functional.

3.4 Experimental manipulation of BRI3BP gene

I designed a series of wet-lab experiments with the goal to identify another example of functional stop codon readthrough in humans and show that regulation using readthrough can be involved in cell proliferation pathways similar to VEGF-Ax.

The experimental design (Figure B-15) had 3 primary objectives:

- Replicate previous experiments to reestablish BRI3BP function – decreased expression results in reduced growth and proliferation

- Edit the genome to alter the stop codon at the end of the gene – observe the effect of the readthrough variant

- Analyze mechanism of readthrough function if some effect is observed in the previous steps

3.4.1 Creating BRI3BPx with CRISPR

Previous experiments using mass spectrometry that looked for read-through signal were unsuccessful in identifying protein fragments representing the extended version of the BRI3BP gene (BRI3BPx) and it is likely that the readthrough only triggers under specific conditions. I used CRISPR protocol to change the first stop codon to a coding residue to test the effects of the readthrough version on cell function.

To create the optimal targets for the Cas guides I looked for conserved regions in the vicinity of the stop codon which would minimize the off-target effects which could result in modifications in wrong genes. I selected three guide sequences with the aid of Zhang lab tools [32] which were all used as part of 3 separate vector designs for the gene modification. The primary test sequence included a replacement of the TGA stop codon to TCA (serine). The aim of choosing this particular substitution was to minimize changes to the stop codon context by only making a single nucleotide substitution. The TCA change also avoids creating a large or hydrophobic residue which may interfere with the secondary structure in the cytoplasmic region of the protein. Once the experiment was conducted using each of the three guides we chose a the particular guide which had the highest edit efficiency for use with the controls.

To confirm that the phenotype changes to the cells that we were observing were directly related to BRI3BPx several controls were necessary. To test that the amino-acid sequence is essential and sufficient for the change in phenotype, 3 controls were created. These controls included a shuffling of the amino acids, dinucleotides and nucleotides of the readthrough region. The chosen sequences were also selected so that their predicted structures did not form an interaction with the C-terminal helix in the structure predicted with QUARK. Finally, to verify that the editing process itself was not causing a phenotype change, 3 additional controls with the same permutations were made which still included the original stop codon. Thus, the permuted sequences could be used to verify that the edit was successful and phenotype for the off-target effects could be observed.

The first set of experiments involved clonal isolation and expansion. In this experi-

ment we calculated the efficiency for editing by sequencing isolated clonal populations resulting in 68% editing and 89% targeting efficiency (19 total colonies tested) for the readthrough edit and 73% editing and 86% targeting efficiency (15 colonies tested). While the controls had good survival rates with 13 colonies ultimately surviving and 10 of those having the correct edit, the readthrough edit fared much worse. Only 5 colonies survived - none of which had the correct edit, but 13 of the 14 colonies that ultimately perished were edited successfully with the constitutive readthrough variant. While, we were not able to draw any specific conclusions from this, it showed that there is the potential for the edit to have pro-apoptotic properties that is incompatible with the stress of the clonal expansion from a single cell. Moreover, phenotypically in comparison to the control cells, the readthrough variant cells tended to clump and have altered physical appearance As a result, an unpublished double selection method developed by the lab was used instead so that we would be able to have a surviving population of edited cells. Using the double selection method we were able to obtain 80% edit efficiency as determined by sequencing *e.coli*-amplified and isolated clonal populations derived from the pooled cell population. Colonies from this population was used in further experiments.

3.4.2 siRNA knockdown of BRI3BP and BRI3BPx

An siRNA knockdown experiment was designed to verify the BRI3BP function and to see if we could rescue the BRI3BPx phenotype. A restoration of the wild-type phenotype by knockdown would imply that BRI3BPx is responsible for the observed changes. The construct was designed to target the second, most conserved exon of BRI3BP which is included in all known BRI3BP isoforms. A lentivirus was created for targeting. A fluorescence and qPCR readout was used to measure the siRNA efficiency. Observations of fluorescent vs. non-fluorescent cells estimated efficiency at around 81.5% while the qPCR measured it to be around 81.1%. This cell population was used in the apoptosis assay described below.

3.5 Etoposide Apoptosis assay

To test the cell status of a particular cell population a topoisomerase inhibitor (etoposide) was added to the cells for 40 hours. After the 40 hour mark the apoptotic state of the cells was measured by propidium iodide (PI) and Annexin V (AV). The former binds to DNA in cells that have burst open, while the latter recognizes the swapping of phosphotydilserine to the outer membrane. In combination, we can use fluorescence activated cell sorting (FACS) to count cells which are effected by one or both tags. The interpretation for every combination of fluorescence is shown in table A-1.

To interpret the remaining results I obtained a baseline for the effects of the etoposide (Figure B-16 KD A and B). Expectedly, the amount of apoptotic cells increased in that particular comparison. When the wild type cells are compared with BRI3BPx edited cells we notice an increase in apoptotic cells with a disproportionally large increase of cells in early apoptotic stages (B-16 C). This result is intriguing since the effect of BRI3BPx seems to be similar to what was observed by Yamazaki *et al.* with BRI3BP overexpression under drug treatment conditions. More experimental controls are necessary to determine that the observed effect is specifically due to the BRI3BP interaction and not due to the cells being generally sicker from the transformation procedure.

The BRI3BP siRNA knockdown cells would be expected to decrease apoptosis rates since higher expression of growth factors would results in cells being more vulnerable to drugs due to the stress of cell division. This prediction matches our observed results (Figure B-17).

These preliminary results suggest the BRI3BPx has the potential for increasing apoptosis rates in cells and that our designed siRNA is functional in reducing the normal effects of BRI3BP. Further experiments would be needed to see if the BRI3BPx phenotype could be rescued by use of the siRNA. Additionally, if the cells containing the shuffled BRI3BPx regions described in section 3.4.1 were tested with the same assay, then we would be able to confirm the necessity of the specific amino acid sequence in generating the phenotype.

3.6 Conclusion

The evolutionary information in conjunction with the results of the completed experiments suggest that the readthrough region of BRI3BP may have an evolutionarily preserved function which radically alters the performance of the protein. Considering the VEGF-Ax as another example of the mechanism it appears that the prevalence of use of translational read-through as an anti-proliferative measure may be more widespread than previously thought. There are clear benefits to a cell's ability to develop additional pathways for triggering apoptosis when environmental causes result in stop-codon readthrough errors.

Nevertheless, the completed experiments do not eliminate other possible hypotheses for mechanisms by which BRI3BPx causes the functional changes in cells. To provide more evidence that the interaction of the secondary structure elements in the C-terminal causes the changes in BRI3BP function and to analyze the mechanisms which cause BRI3BP stop codon readthrough more experiments would be necessary. For example, with the use of a v5 N-terminally tagged BRI3BP gene it should be possible to look at localization differences between BRI3BP and BRI3BPx. Subsequently crosslinking and IP for the tag would allow us to look for interaction parters of the two gene variants. If the readthrough region indeed blocks the coiled coil motif on the c-terminal end we would expect to see some differences between the two variants. With BRI3BP and BRI3BPx over-expression in WT cells it would be interesting to replicate the experiments presented by Yamazaki *et al* to observe whether the two overexpressed variants function in the same way. Also, ribosome footprinting in the drug treated cells within the readthrough region would inform us if the drug treatment is causal for the readthrough variant.

If future experiments successfully demonstrate BRI3BPx's pro-apoptotic properties, this natural anti-proliferative check could have clinical applications. For example, cancer cells with functional BRI3BP could be susceptible to drugs which increase readthrough rates while healthier cells may be able to resist the effects due to other non-apoptotic error correction measures.

Chapter 4

Analyzing Burial Traces of Human Genes

4.1 Introduction

There has been a push to develop models that can help us predict certain information about the structure of a protein based on its amino acid sequence. While *de novo* structural prediction is difficult and resource intensive, some other methods can be used to approximate some structural properties with fewer computational limitations. In this chapter, I describe how approximating the protein structure by predicting the burial of individual residues within a globular fold can be used to predict the magnitude of the impact on the mutation on protein structure. I discuss the relationship between these predicted structural changes and conservation and show that improving this prediction can further our ability to estimate the degree of impact of hypothetical mutations on protein function. I suggest that as rapid structural prediction improves it can be useful to consider as part of agglomerative models which make predictions about the functionality of new variants in patients.

4.2 Background

A model developed in the England lab [11] allows for the estimation of residue burial based on sequence. This model relies on modeling the amino acids as beads connected by springs and applying the forces on each residue to mainly come from the hydrophobicity: the extent to which they favor the cellular environment relative to the protein interior. The model then allows the polymer to fold and outputs the burial of each residue in the globular structure. In many cases structural stability of a protein is dependent on specific conditions and even single amino-acid substitutions can have a major impact on protein structure rendering the protein non-functional [33].

By comparing the fluctuation in the burial trace due to single amino acid substitutions I provide estimates for the amount of constraint on those positions and make a prediction about the impact of a mutation at these positions on protein structure. To evaluate the accuracy of these predictions I compared the structural fluctuation of simulated mutations to data sets of known human mutations that are known to have an effect on phenotype (as well as locations known not to have an effect) [17] and observing the fluctuations at these positions.

4.3 Defining and computing burial traces

Since the model for predicting the residue burial has the underlying assumption that the residues fall into a globular formation and do not interact with other cellular components to affect their structure, I focused on the domains listed in the conserved domain database (CDD). These domains are manually curated and generally fold into nearly identical structures when present in different kinds of proteins. Additionally, this choice allowed me to have an alignment centered on structural similarity which would allow us to sample more mutational space knowing that the protein secondary and tertiary structure remains generally the same.

Once the data set was selected, the first step towards predicting residue functionality was to observe the differences among burial traces of all the mutations present in

the alignment of proteins that share the domain structure. Three fluctuation metrics were used to compute the differences - the L^2 norm, L^∞ norm, and Pearson correlation. Finally, we can compute the burial traces for each single amino acid permutation of the CDD consensus sequence. This process is outlined in (Figure B-18). With this data we can measure the effects of conservation on changes in burial trace.

4.4 Checking that difference metrics accurately represent effects of mutation

A few verifications of the model performance were necessary to ensure that the burial trace changes that we are observing are proportional to the changes we qualitatively expect given the nature of the changed residue. Based on 100 randomly selected sequences from the CDD, I determined the magnitude of the metrics described in the previous section under every possible mutation. The mutations which represented the greatest changes in hydrophobicity resulted in the greatest changes to the burial trace. Since the model uses the hydrophobicity directly to compute the burial this is to be expected and was a promising first result. Since the nature of each amino acid, other than the hydrophobicity score, is not relevant for the model I wanted to make sure that similar changes to the hydrophobicity had the same effect on the burial trace no matter which amino acids were substituted. There was no significant difference observed when controlled for the hydrophobicity change. Additionally, I verified that the average hydrophobicity of the protein did not have an effect on the magnitude of the burial trace difference resulting from the simulated mutation. However, location in the protein sequence did appear to play a role with the effect on the trace. Simulated mutations in the edge 10% of the protein tended to cause larger mutational effects than those of a similar hydrophobicity change (Figure B-19). Due to the nature of the model representing beads on a string, the edges have more degrees of freedom and thus could more easily alter in burial. These observed changes are likely not biologically relevant since the end regions of the protein would likely interact with

many other forces externally to the protein itself which will likely have a major impact on the local structure. Further analysis suggested that the protein regions affected by these results are of uniform size among proteins of various lengths, implying that these edge effects are increasingly less relevant in the overall traces for larger sized proteins. For many mutations the effect of the burial trace was not local, meaning that the trace was not always most different in the vicinity of the mutated position. This suggests that despite the fact that the amino acids interact most locally, they could still have an impact on the folding of the global structure within this model.

4.5 Structural fluctuation as a predictor of conservation

One of the most basic tests for predicting importance of a particular residue to the domain structure is to look at the number of substitutions present in each alignment column. For every fluctuation metric, I compared the effect of mutating a given residue of the consensus sequence to the set of substitutions in the alignment (Figure B-20). Despite the high variance we see each metric approach the value of maximum similarity as the number of observed substitutions in the alignment increases. Even though the variance is relatively high across the full range of conservation the variance decreases as the number of substitutions goes up. This decrease could be due to the fact that some conservation is due to non-structural reasons like residues involved in PPI, active sites, etc, and potentially due to the fact that there wasn't necessarily time for evolution to try every allowable mutation.

Nevertheless, structural fluctuation is a poor predictor of conservation. For two conservation metrics - the number of alignment substitutions and the PAM matrix value of the mutation, the differences in burial trace do no better at predicting conservation than the difference in Kyte-Doolittle(KD) hydrophobicity values between the new and original residue (Figure B-21). For each metric the predictive value averages around R^2 0.075. Interestingly, for larger domains, structural fluctuation

is a better predictor of conservation than hydrophobicity(Figure B-22). This isn't surprising since for larger sized domains a higher fraction of the residues should be related to maintaining domain structure relative to the active site residues that will generally be of a similar quantity in domains of various size.

One goal of this analysis was to observe which mutations in proteins were consistent in maintaining the overall structure and which mutations would change the burial to the extent that the protein loses its structural stability. I generated burial traces for all the possible mutations from the consensus sequence for the domain. Then I compared the deviations in the burial traces for each mutation to the trace of the consensus. If you average the deviations for the mutations that are present in the alignment to those that are not, we see that the mutations not found in the alignment are causing greater changes to burial in the majority of the samples (Figure B-23). Nevertheless, there is relatively high variance in both categories of the mutations.

The combination of all the above results suggests that on average burial changes are indicative of the selection pressure against the possible mutation, but it difficult to make predictions about specific mutations. It seems that burial trace is a somewhat orthogonal measure of functional importance and could be used to separate the positions conserved for structural versus other functional reasons. Moreover, integrative methods would be able to use burial and conservation in addition to other parameters to improve functional prediction.

4.6 Predicting deleteriousness from burial traces

To further analyze the relationship between burial trace and residue functionality I look at specific mutations known to cause disease relative to likely neutral mutations. As the test set of classifying mutations into deleterious and non-deleterious, I chose the training data set used for Polyphen 2 which separated mutations into two groups. The first are SNPs implicated in Mendelian disease and the second containing SNPs predicted as neutral based on human-chimp divergence. Using this data set, I applied the metrics to measuring structural fluctuations in the two groups to see whether the

deleterious mutations could be predicted by their effects on burial trace.

Looking at the L^2 norm, L^∞ norm, and Pearson correlation for the differences in burial trace, was not performing well as a classifier of deleteriousness of mutations so I had developed two additional metrics that use information from the protein sequence and the KD hydrophobicity of the new residue. The first metric used was based on the concavity of the new trace relative to the wild-type protein's trace. Since it was suggested in the original paper that the model does a better job predicting peaks and valleys than it does predicting the absolute depth of each amino acid, it's reasonable to consider mutations that turn peaks into valleys and valleys into peaks as more substantial than those mutations which result in a peak being represented as having less absolute depth. I could use this to quantify the change in the concavity of the curve and consider mutations that result in greater concavity changes to be more deleterious. The second metric was introduced to account for the fact that some regions of the protein may be more flexible than others. Since we had previously computed the effects of each mutation onto the structural trace we had the data describing which protein regions tend to have a higher standard deviation in burial. It's reasonable to assume that if a new mutation introduces changes in a protein regions which is generally more stable, then it probably has a larger impact on protein function. For example, in a case study of dihydropteridine reductase the neutral mutations all cause fluctuations which fall in the flexible regions while most deleterious mutations create a change in burial around positions 70-100 where the structure generally remains stable (Figure B-24). For each mutation I calculate a "contribution curve" where for each amino acid the value is higher proportional to the square of the differences in burial traces at positions where the difference is greater than a standard deviation of all mutational effects described above. A visual example of how this score is computed is provided in figure B-25.

The two new metrics outperformed the correlation and the norms, but even when the metrics were combined with the magnitude in hydrophobicity change the classification was on par with performing as well as using the hydrophobicity alone (Figure B-26 A). However, I noticed that the two new metrics tend get things generally cor-

rect, but they mis-classify different kinds of mutations (Figure B-26 B). Using a linear combination of the two scores I was able to improve the classification of mutations into neutral and deleterious. The new metric performed somewhat better than hydrophobicity alone(AUC 0.68 vs AUC 0.65, respectively), but still significantly worse than Polyphen in classifying those mutations (Figure B-27).

4.7 Using homology modeling to benchmark model performance

To have the methodology of predicting deleteriousness be more versatile, it's essential that we do not limit ourselves to proteins with a known globular domain boundary. Since the method does not require creating the full protein structure it may be possible to make predictions based on the nearby protein region to the SNP. If small enough regions are considered then hopefully a globular approximation would be sufficient to identify cases where the mutations will cause major disruptions to tertiary structure. To test this hypothesis I considered nearby regions around the potential SNPs to identify the deleteriousness of those mutations. For each amino acid change the regions including 200 amino acids on both sides of the polymorphism was considered. Where this was not possible due to SNP position the terminal 400 residues containing the SNP were considered.

With the use of nearby regions the predictive ability of the fluctuation and concavity metrics reduced significantly. To see whether this decrease was caused due to the model limitations in predicting secondary structures of non-globular regions I used homology modeling to predict structures of the same regions. Interestingly, the correlations in burial for all the residues were relatively low between the homology models and Jeremy England's model (Figure B-28). However, there are a lot of challenges in making these predictions and at least the models were more correlated when regions within globular domains were used as reference.

To assess if the the model performs better when the structure is predicted more

accurately I considered the set of regions with the highest 50% of correlations between the two models ($R^2 > 0.2$). These regions performed significantly better for the set of nearby regions (Figure B-29 B). However, for the case when globular domains were used there was no improvement in deleteriousness prediction when the analysis was limited to regions highly correlated between the two models (Figure B-29 A).

Overall, these results highlight the difficulties in predicting structural properties from limited information. Nevertheless, they suggest that using a combination of structural models, even if those models aren't optimally suited to the particular sequence can improve our interpretation of their results. However, further refinement of the method is needed before making meaningful predictions with this information.

4.8 Variation in hemoglobin burial traces

To observe how the burial model changes among homologs of the proteins of the same structural class I analyzed the CDD structures of the hemoglobin gene in a variety of species. First, I compared the burial traces of each hemoglobin's CDS. To account for gaps in the sequences I only considered residues which aligned to the consensus hemoglobin sequence. The resulting traces had a substantial degree of variation (Figure B-30). In order to see if the variation was caused by non-aligned residues I also attempted to look at the trace of the whole sequence separately and subsequently compare the burials of the non-aligned positions. This approach did not improve the deviation in the results. The one promising aspect of the variation was that the regions connecting the peaks and troughs had less variation than the peaks themselves. This is consistent with my previous result which suggest the peak locations in the burial trace are being generally predicted consistently relative to their absolute height.

4.9 Conclusion

With a focus on human gene sequences, I looked at the impact of hydrophobicity altering Single Nucleotide polymorphisms (SNPs) on predicted protein structure. It was found that while highly conserved positions are more likely to break the structure when mutated a large fraction of constraint is NOT structurally caused. The model performed better on regions of confined structure (CDD domains) as compared with random coding regions of comparable lengths. Using several properties of the burial trace, I created a method to predict the deleteriousness of arbitrary mutations in protein coding regions. With a combination of burial trace concavity and fluctuations based on flexible regions, I was able to outperform using the raw input data as a classifier. However, as evidenced by the comparison of my method with Polyphen 2, my method performs substantially worse than other tools which use a large ensemble of data sources for classification. Additionally, I focused on a specific domain in greater detail to map out its structural flexibility. The multi-species alignment of hemoglobin was analyzed to obtain a spectrum of allowable mutations and compare it to mutations that are not present in the alignment. For a given structure of hemoglobin disallowed mutations were easily identifiable, but different hemoglobin homologs had relatively varied burial traces.

Appendix A

Tables

Table A.1: Interpreting presence of florescence in the apoptosis assay

	AV-	AV+
PI-	Healthy Cells	Early Stage Apopotosis
PI+	Necorsis/noise	Late stage apoptosis

Appendix B

Figures

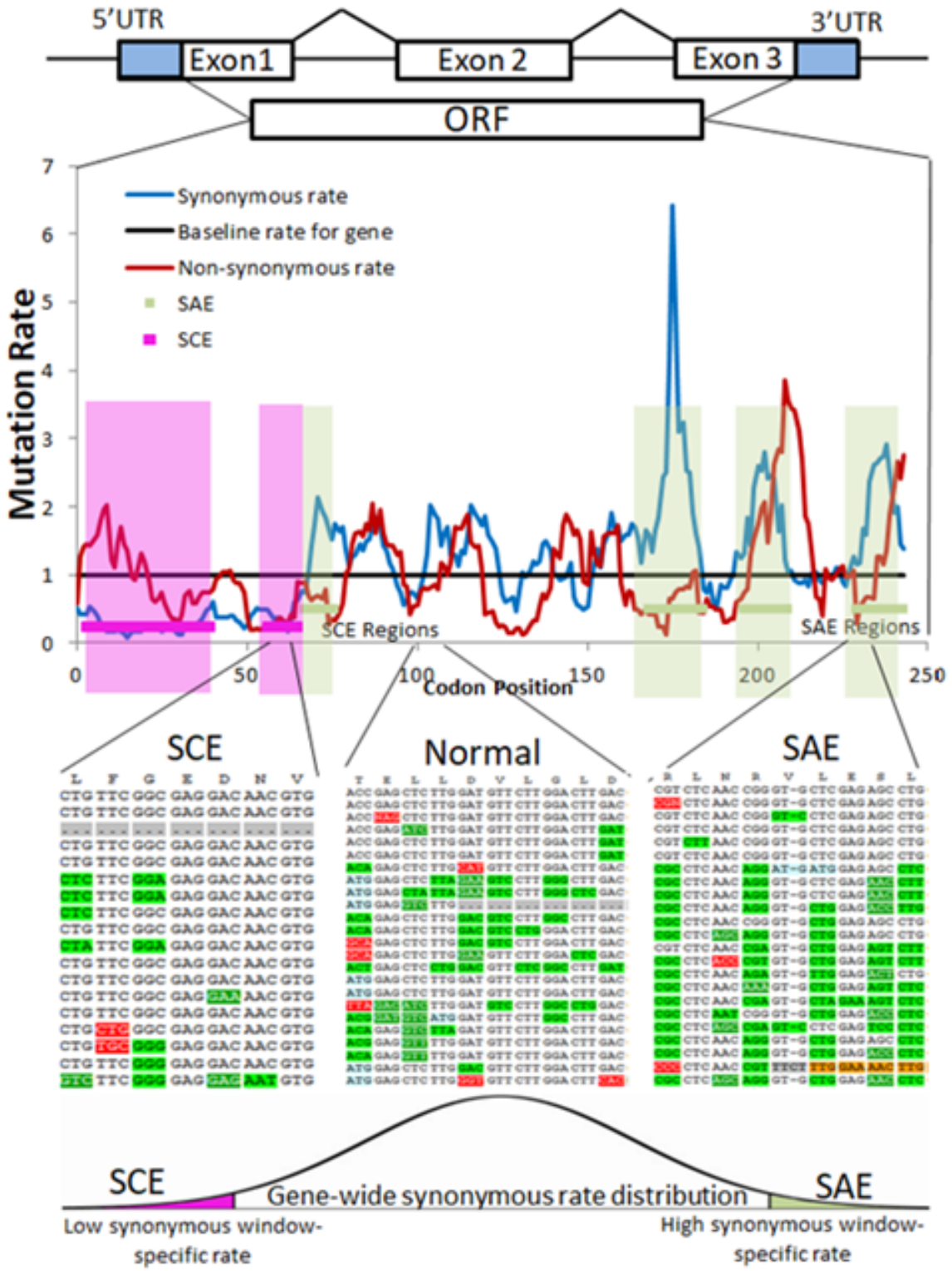


Figure B-1: To identify Synonymous Constraint Elements(SCEs) and regions of Synonymous Acceleration Elements (SAE) we take the CCDS ORF for every gene and analyze its window specific synonymous rates. With the average rate set to 1 we look at 9 codon windows where the local synonymous rate deviates significantly from from the gene average. The alignment at the bottom of the figure visualizes from let to right SCEs, normal regions and SAE regions respectively to emphasize the local change in synonymous rate.

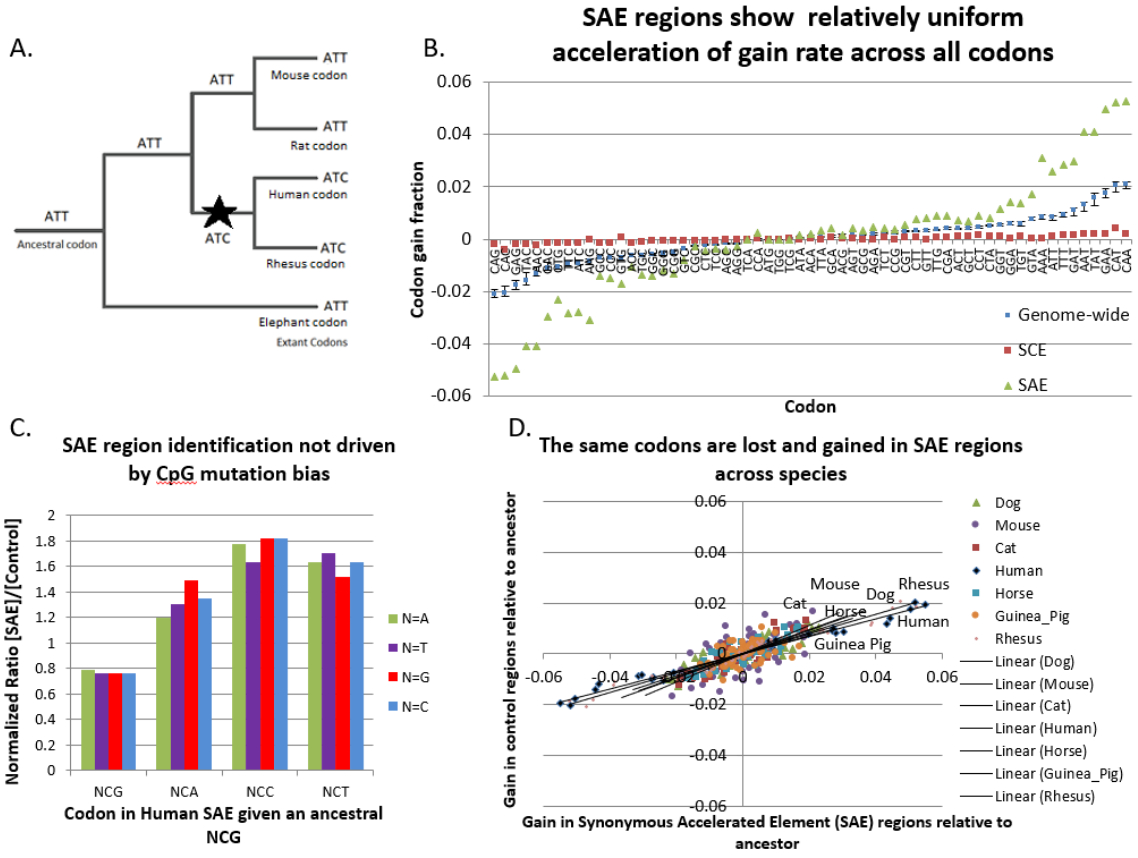


Figure B-2: (A) subset of the mammalian tree showing a sample mutation. In this case The human codon is e changed from ATT to ATC which is considered a loss of ATT and an ATC gain. (B) Gain and loss of individual codons in SAE, SCE, and control regions. SAE and SCE regions show the same patterns of gain and loss as the other coding regions, but at an accelerated or decelerated rate, respectively. (C) The first codon position matters little when it comes to predicting the mutability of the NCG codon. Moreover SAE regions show greater mutability of NCG codon, but fewer of those mutations go to the NCA codon than the NCC and the NCT codons. This suggests that the CpG-> CpA mutation is not the primary driving force of these mutations.(D) Comparing each species to the mammalian ancestor yields qualitatively similar results. For SAE regions - codons that are more frequent in once species than the ancestor tend to be more frequent in other species as well.

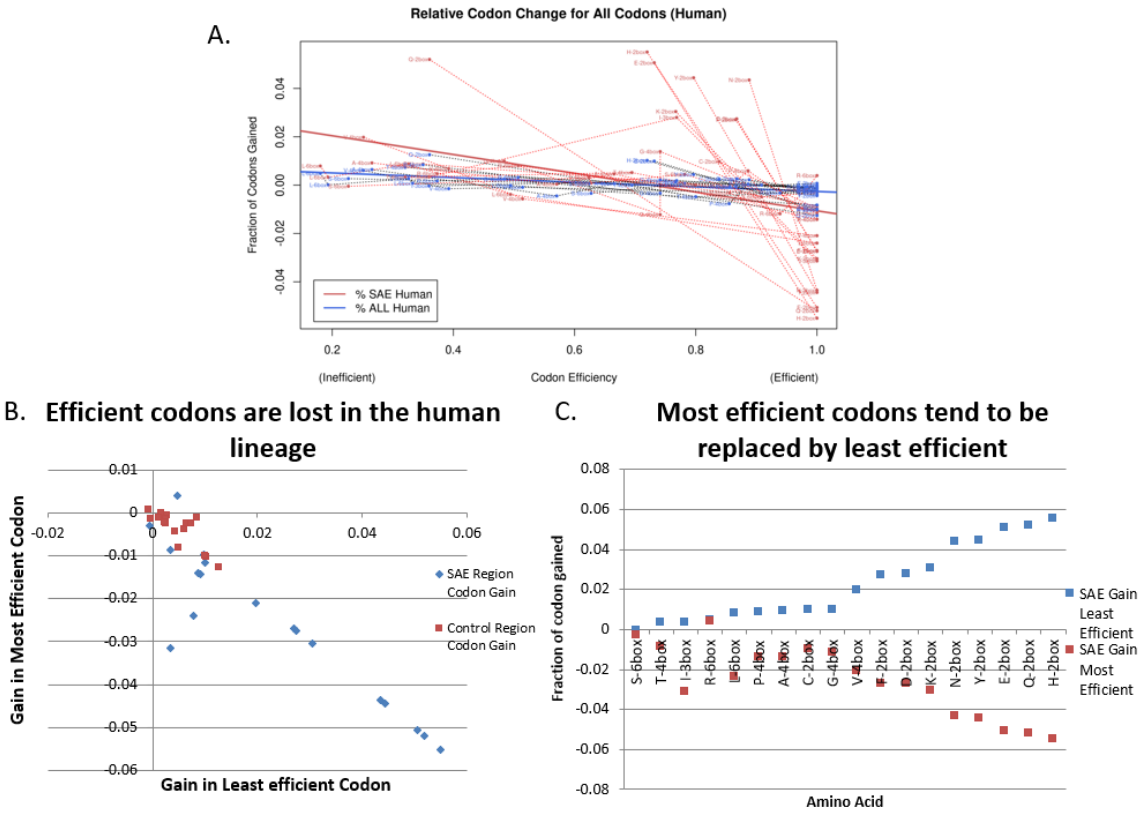


Figure B-3: More efficient codons have been lost in SAE regions at a greater rate than codons in control coding regions.

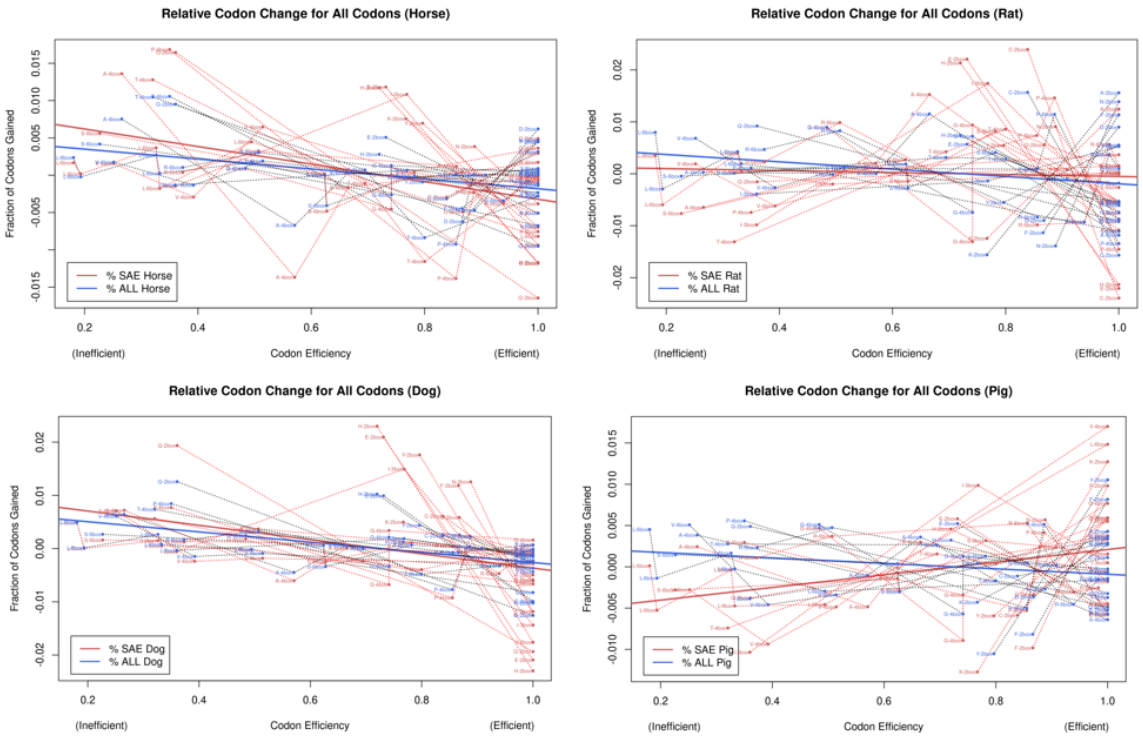


Figure B-4: The y axis shows the fractional increase of a specific codon and the x axis has the efficiency of the particular codon in humans. The red points show the gain in SAE regions while the blue points are the gain in control regions. The differences could potentially be due to poor correlations and species specific differences in codon efficiency

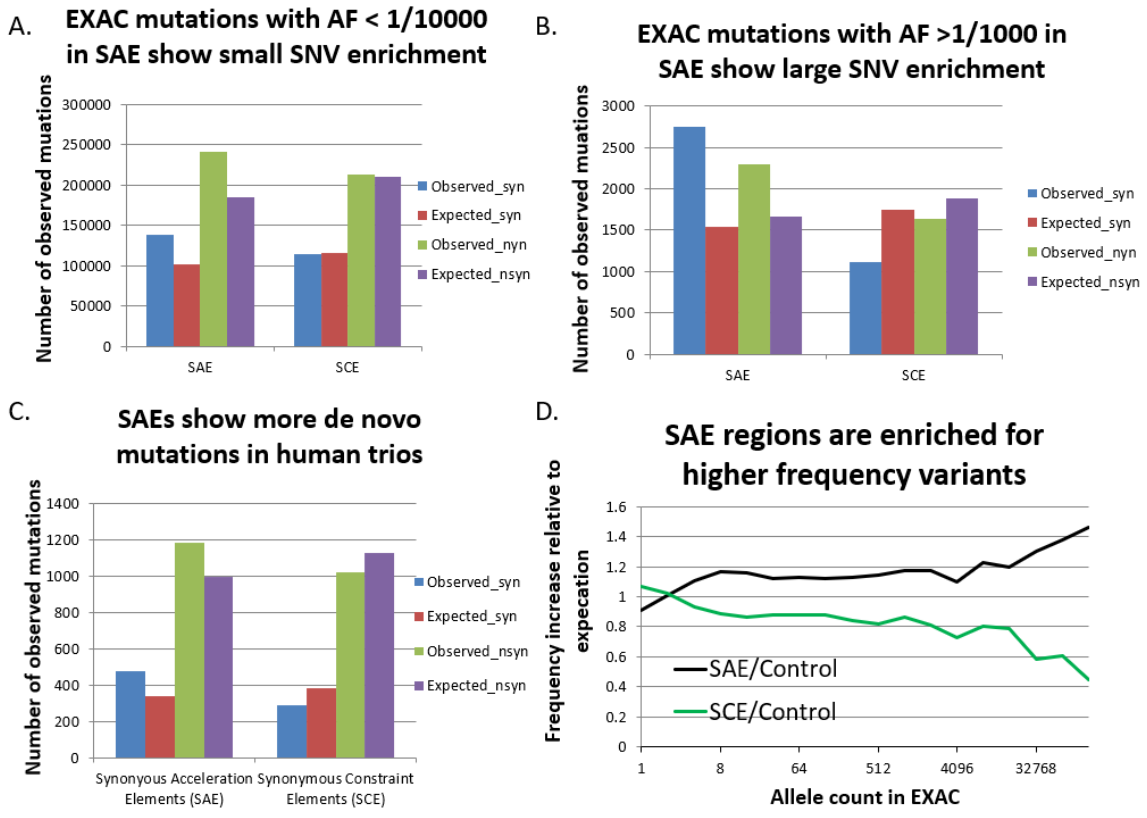


Figure B-5: (A-C) Enrichment in mutation rates affects both synonymous and non-synonymous rates among de novo mutations based on a database of trios. (D) Mutations in SAE are enriched for higher frequency mutations in the human population relative to SCE and control regions.

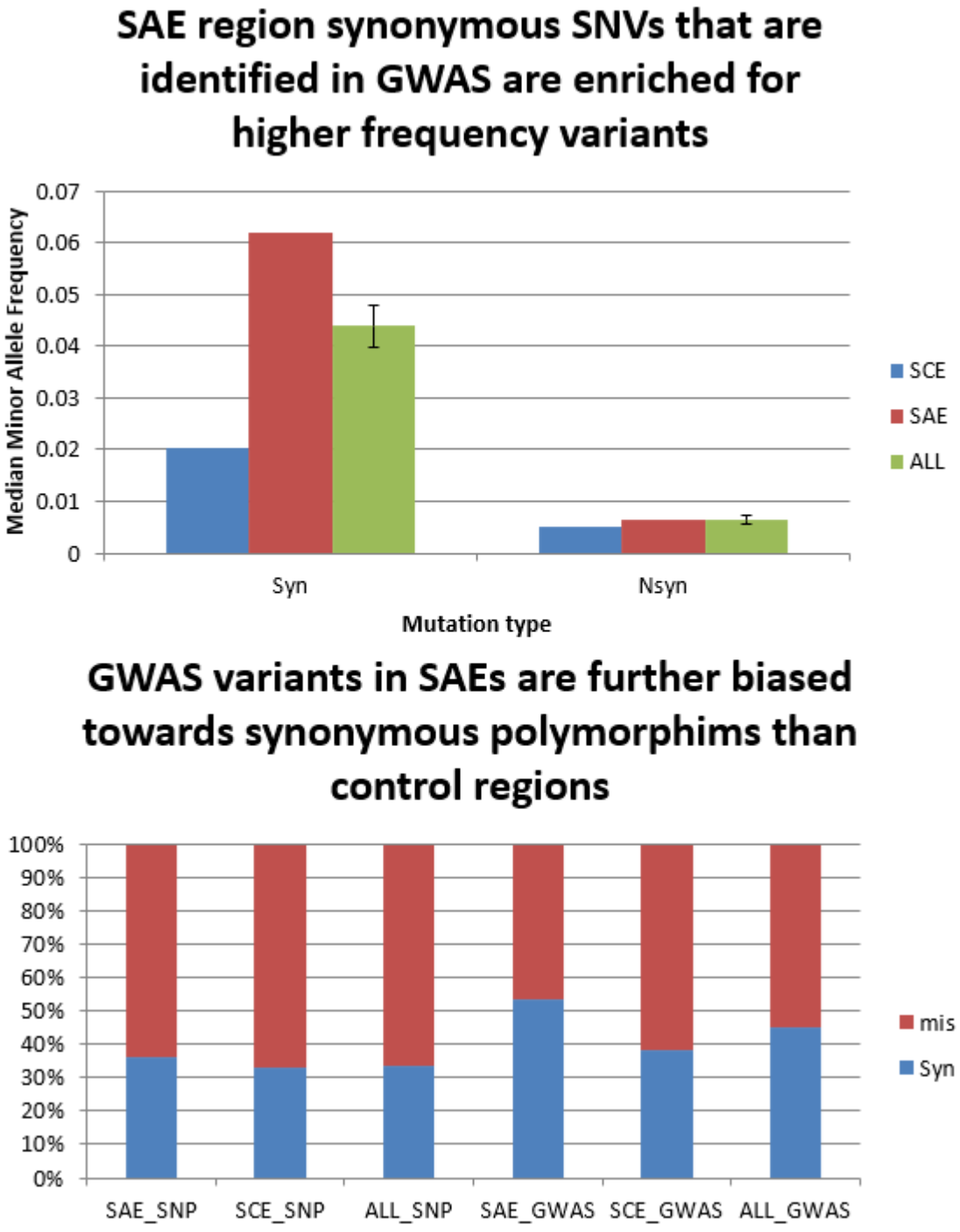


Figure B-6: GWAS hits in UCRs show higher preference for synonymous variation in SAE regions and pick up overall higher MAF polymorphisms in SAEs relative to SCEs.

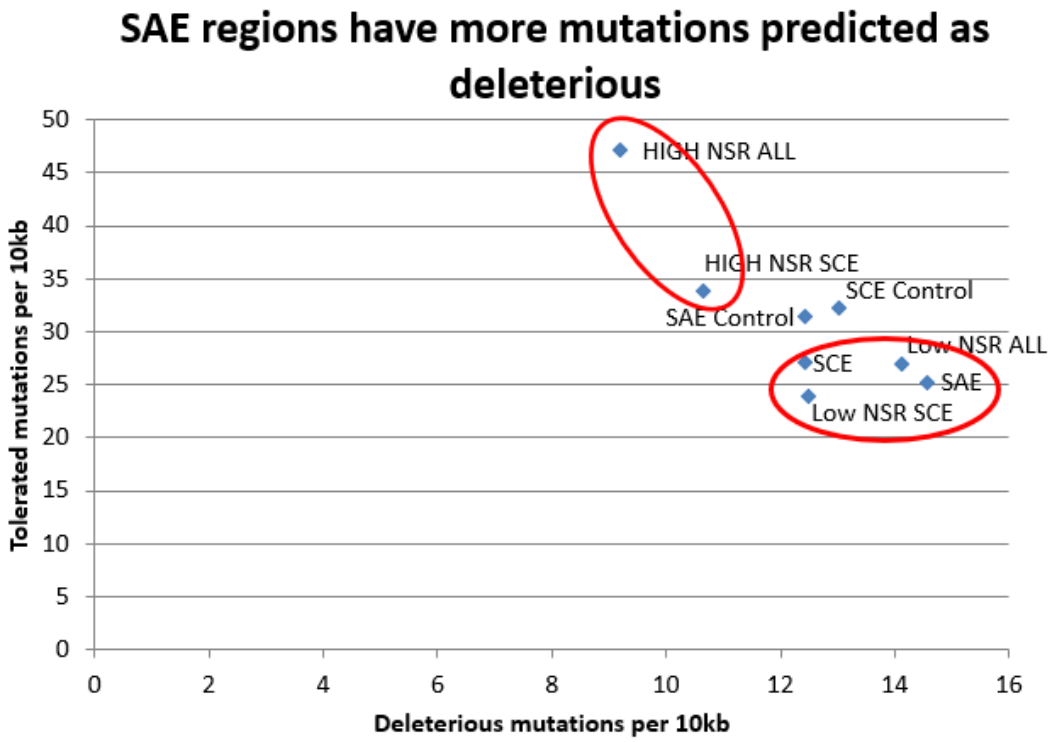


Figure B-7: The non-synonymous rate of coding regions does a good job separating CDS regions that contain mutations that are considered tolerated versus those that are considered deleterious. Interestingly, SAE regions fall into the category with relatively high prediction of deleterious mutations among the cluster of low non-synonymous rate sequences. This suggests that we're likely to see low non-synonymous rates in SAEs and that the mutations in these regions may be harmful suggesting a high likelihood of future *de novo* deleterious mutation in some individuals.

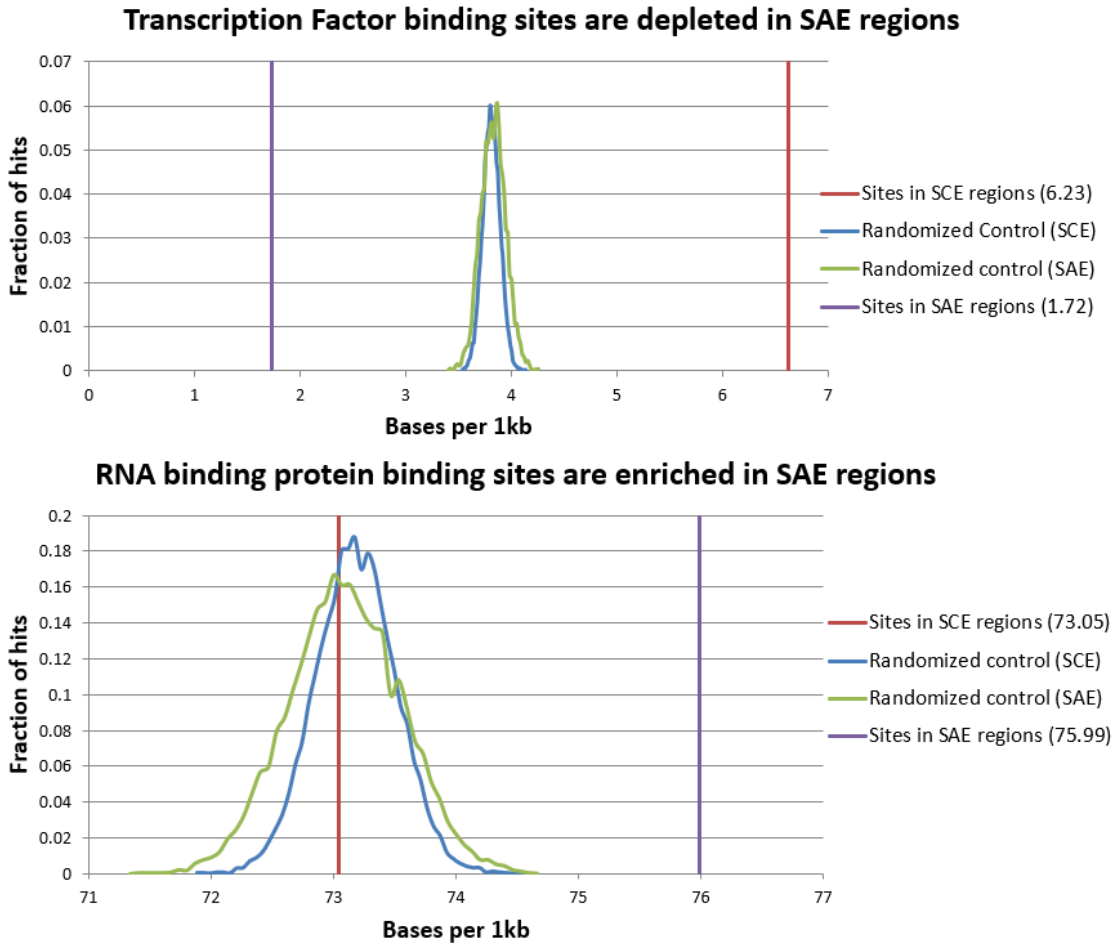


Figure B-8: Enrichment of SCE motifs in TFBS suggests that they contain several active sites that are maintained by selective pressure.

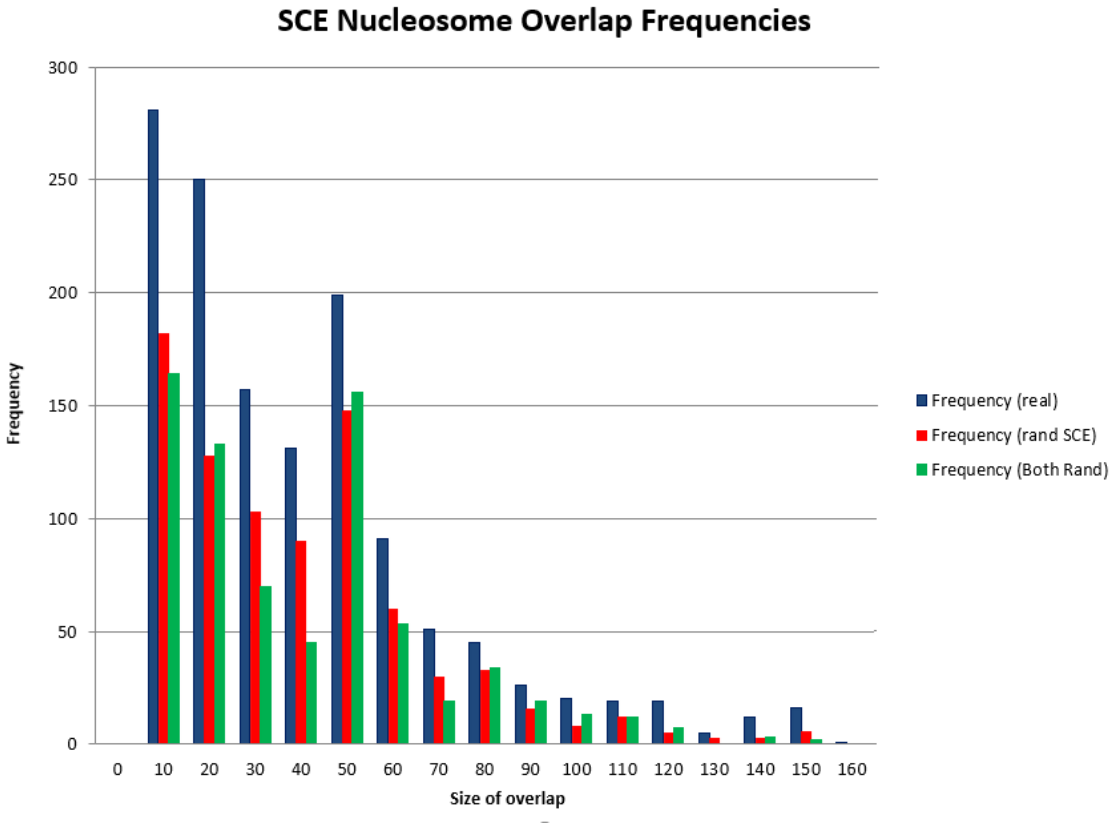


Figure B-9: Nucleosome overlap with SCE regions. The red and green bars show two different randomization approaches with randomly generated SCE and randomly generated nucleosome and SCE positions within the coding regions of genes. It seems that the significance in overlap is due to the SCEs being enriched for nucleosome boundaries.

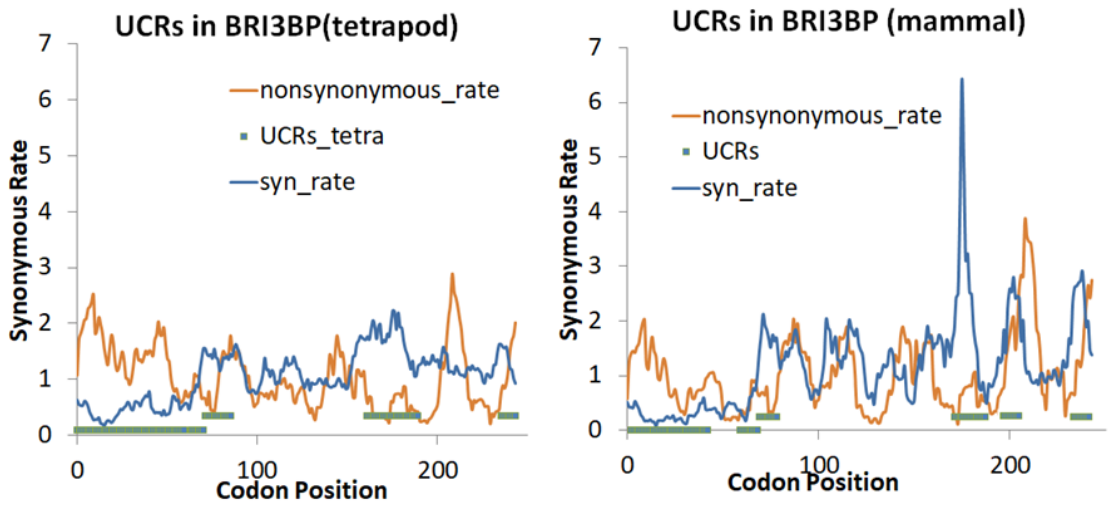


Figure B-10: Tetrapod UCRs are identified in the same positions, but expanded relative to mammalian UCR calls. Two mammalian SCEs and two SAEs are merged in the vertebrate data set.

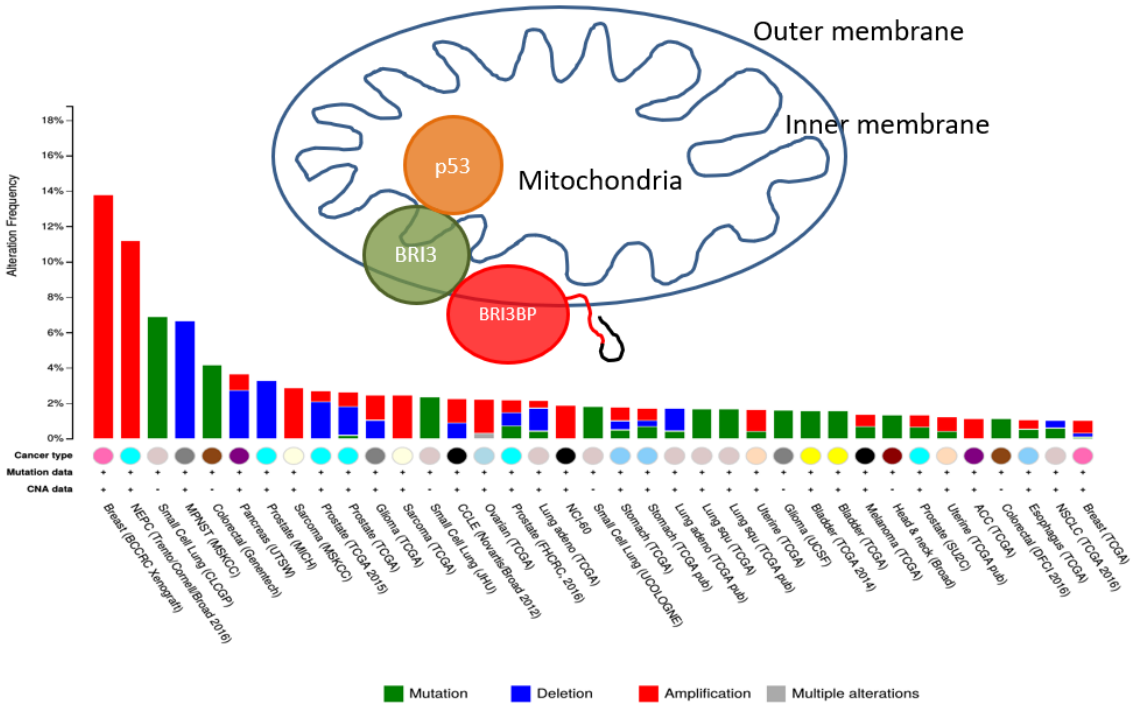


Figure B-11: BRI3BP is localized to the outer mitochondrial membrane where it normally functions by interacting with BRI3 which binds with and sequesters p53 in the mitochondria. The C-terminal end and the readthrough region (shown in red and black on the top diagram) extend to the cytoplasm where they should be open to protein-protein interactions. The bottom graph shows the fraction of cancers of each type which appear to have alterations in BRI3BP. Most commonly, there is an amplification of BRI3BP expression in breast and prostate cancers.

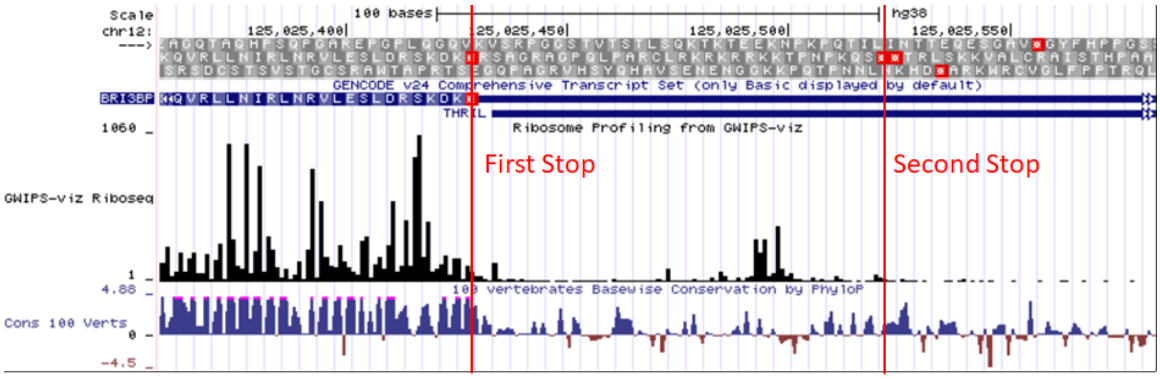


Figure B-12: C-terminal region of the BRI3BP gene showing mild conservation and some ribosome footprinting signal beyond the primary stop codon.

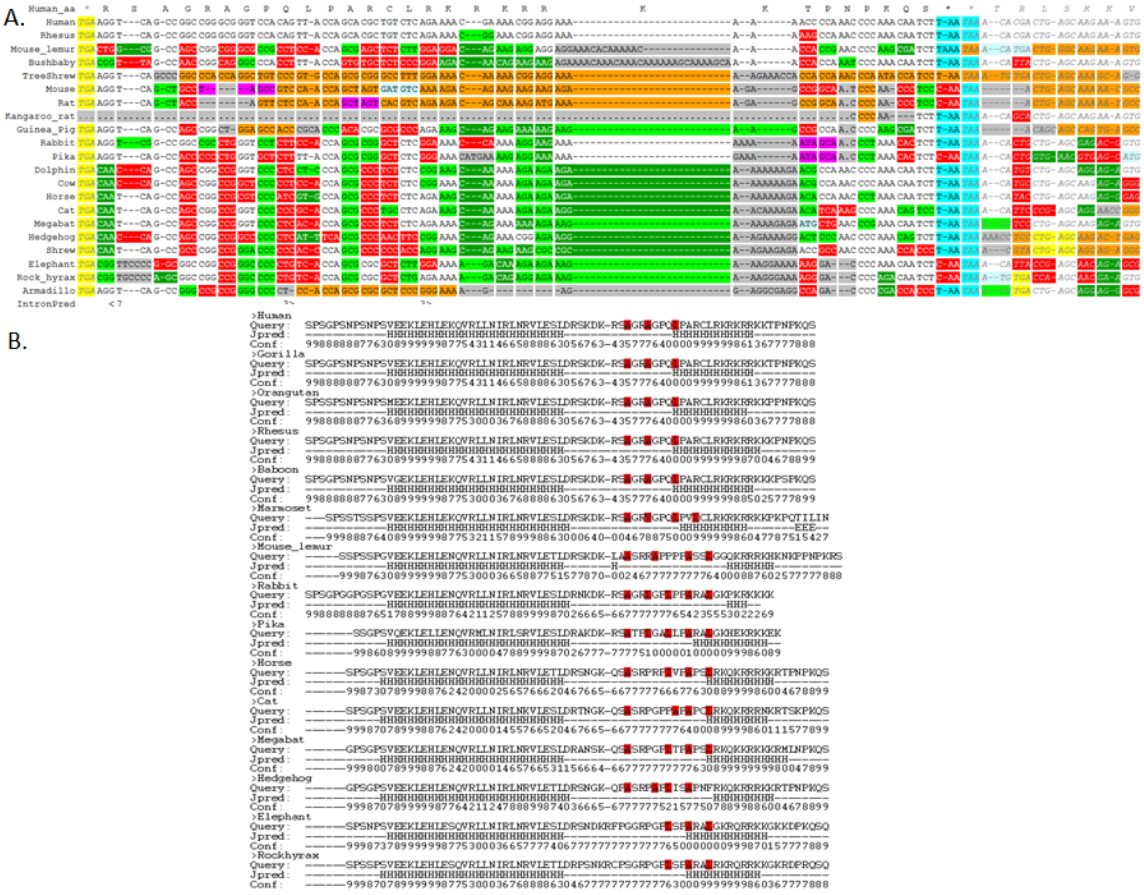


Figure B-13: (A) the alignment of the region between the two stop codons as seen using CodAlignView [34]. The large number of non-synonymous substitutions is readily apparent. (B) The conservation of spaced hydrophobic residues indicative of a possible coiled-coil structural motif.

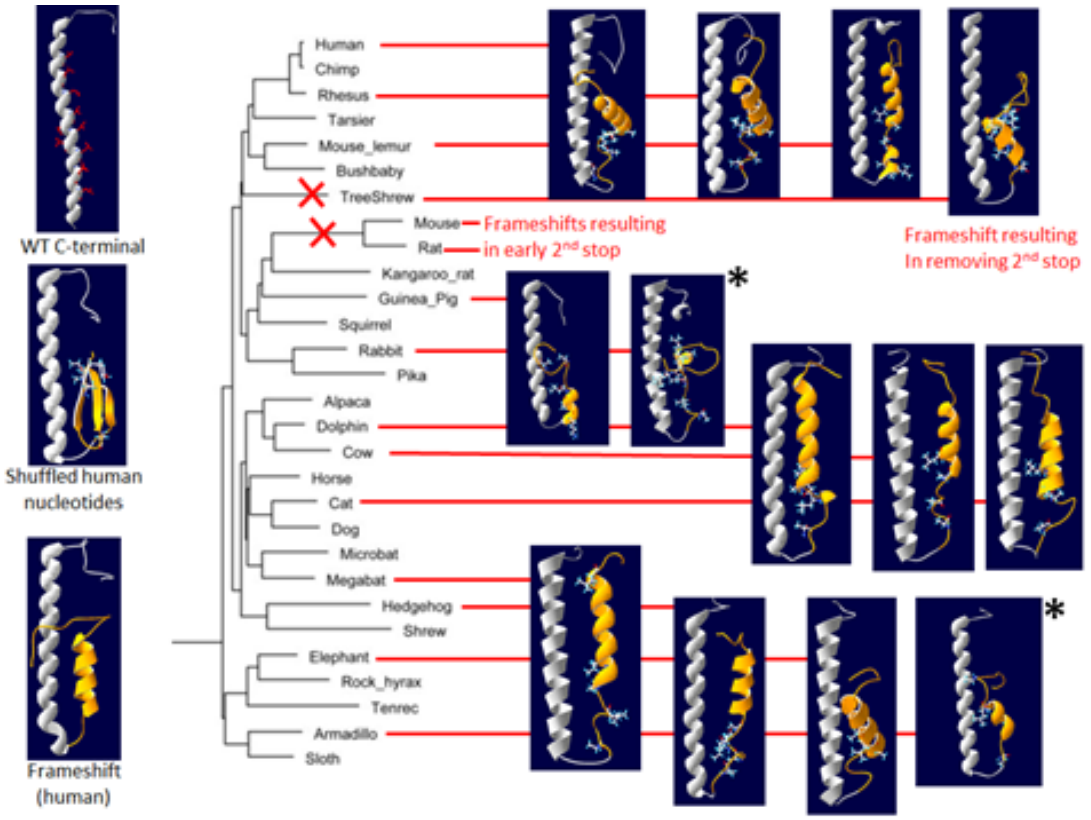


Figure B-14: Predicted secondary structures of the amino-acid sequences of mammals. The yellow region represents the readthrough segment and the gray region includes all C-terminal cytoplasmic residues of BRI3BP. The left side shows the effects of predicting the structure without the readthrough region, with shuffled nucleotides and with frame-shifted nucleotides.

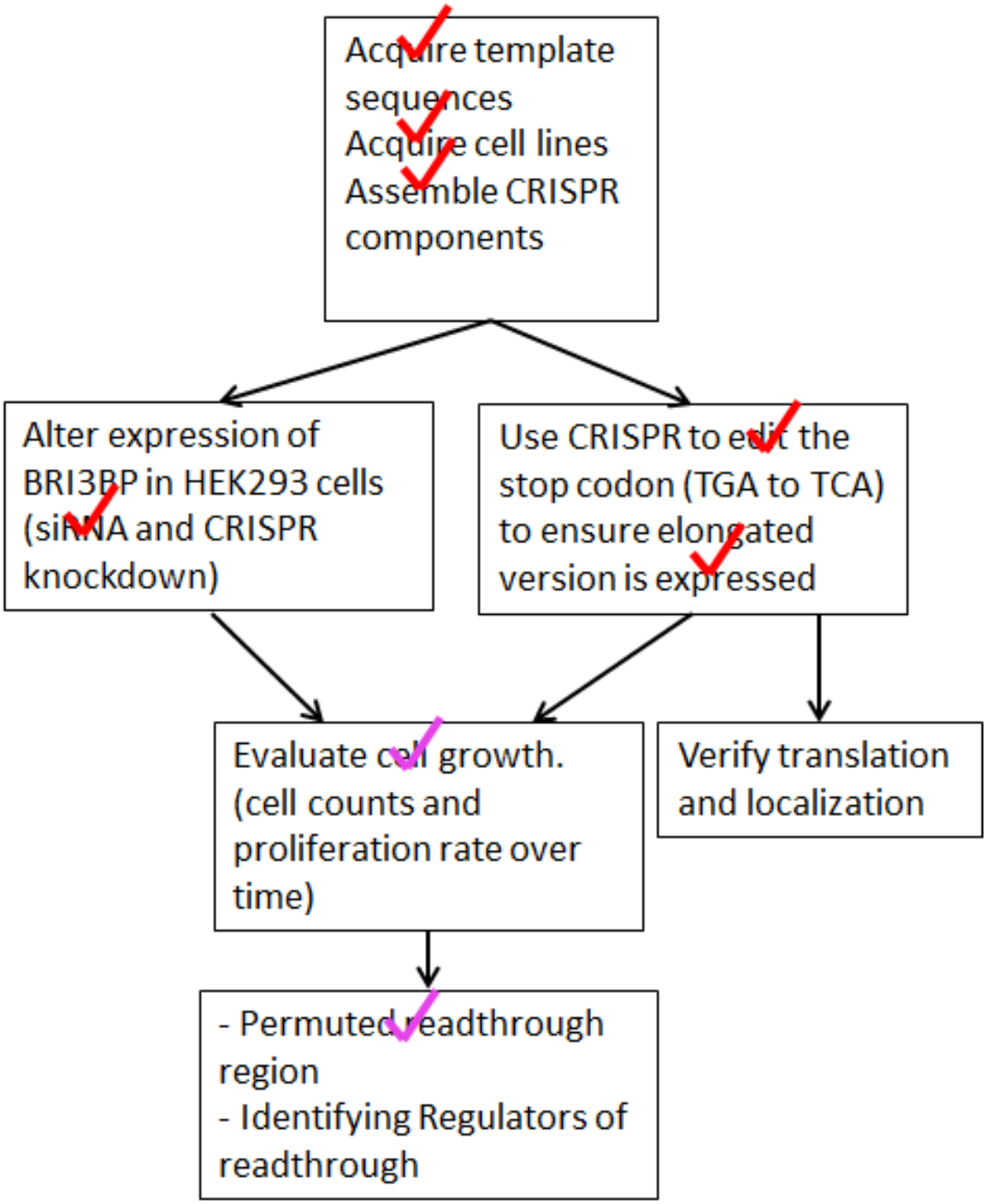


Figure B-15: Work-flow for the planned experiments with experimental modification of BRI3BP. The red and pink check-marks indicate the completed and partially completed objectives respectively. The remainder are planned, but unfinished experiments.

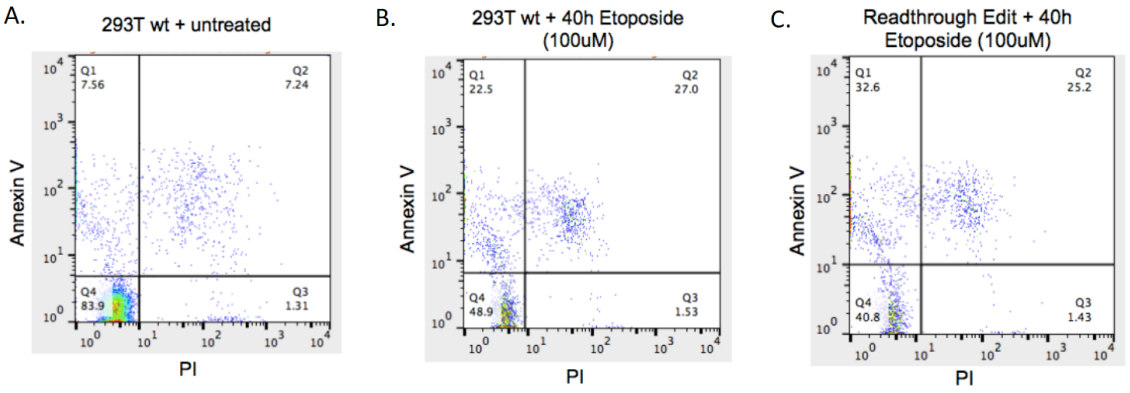


Figure B-16: FACS assay measuring the response to Etoposide drug treatment by observing PI and Annexin V markers in the cell population: elucidating the effects of BRI3BP knockdown on apoptosis rates in drug treated cells. (A) Distribution of cell Wild type cells subjected to etoposide has medium rates of apoptosis with 49.8% healthy cells. BRI3BP knockdown reduces apoptotic rates as is consistent with previous studies.

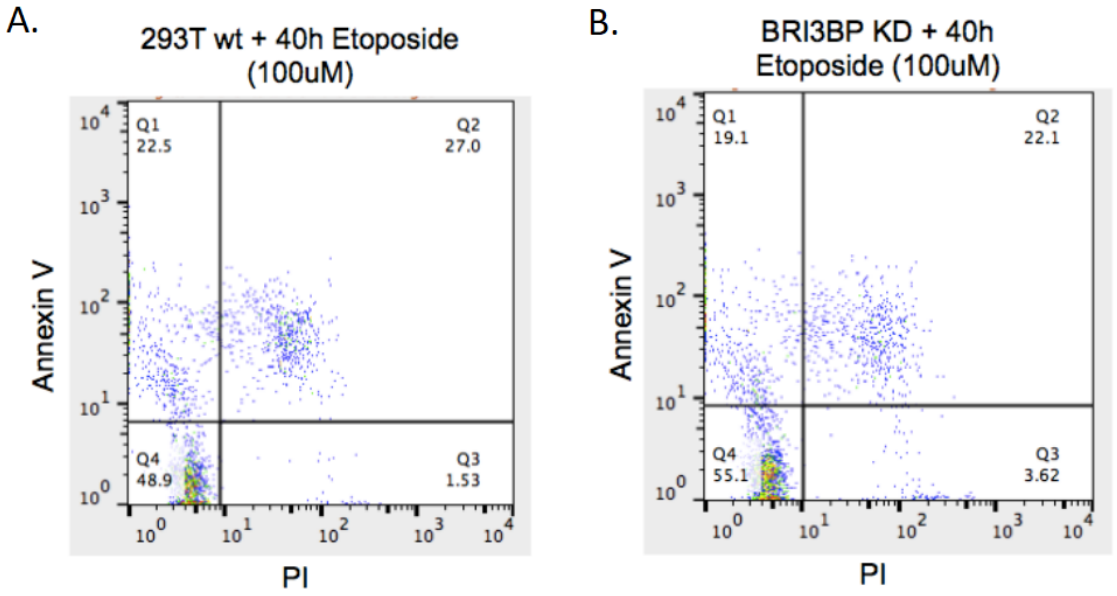


Figure B-17: FACS assay measuring the response to Etoposide drug treatment by observing PI and Annexin V markers in the cell population. (A) Untreated cells exhibit little apoptosis. (B) Apoptosis levels increase post drug treatment. (C) Further amounts of apoptosis in edited BRI3BP cells suggesting the edit may be pro-apoptotic.

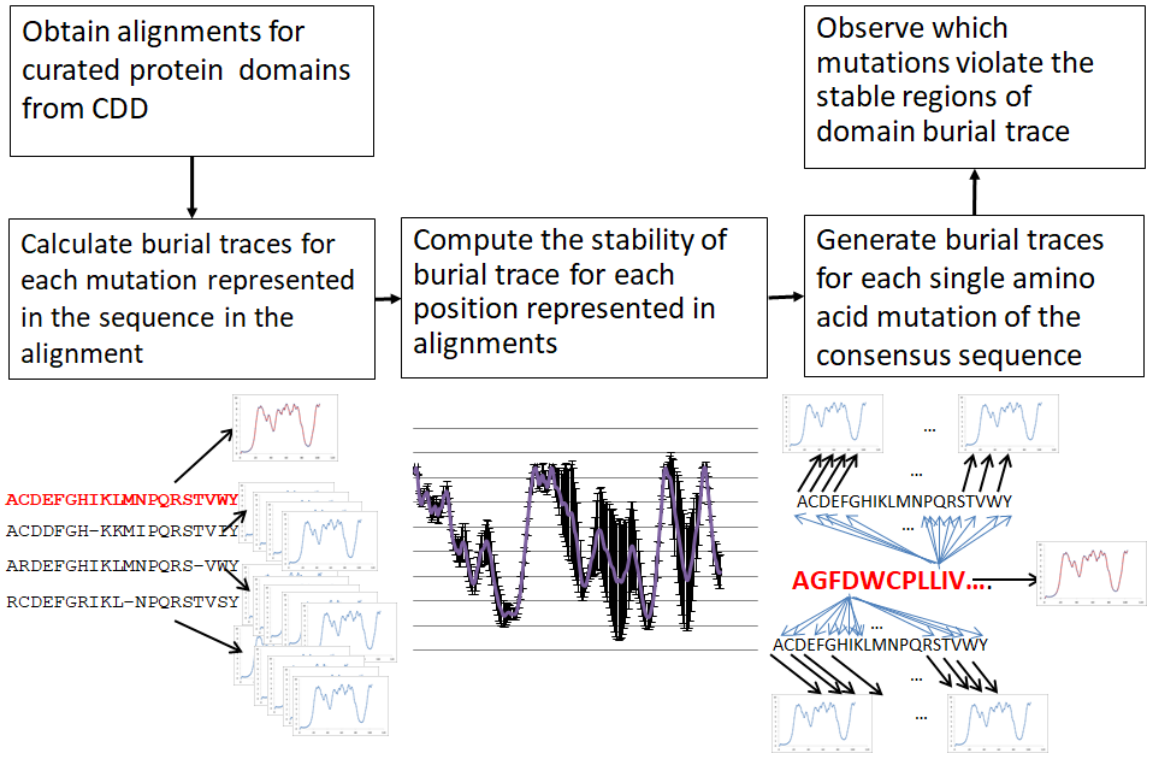


Figure B-18: A workflow of predicting the mutational effects of single amino-acid substitutions. The left panel shows that we generate a trace from each individual mutation from the consensus in every position in the alignment. The middle panel shows an example of the resulting set of fluctuations While some areas are flexible, showing high error bars for the burial of a particular residue in the previously computed mutational ensemble, other positions maintain burial depth under most conditions. The right panel shows how we get a burial trace from the consensus sequence by mutating each amino acid to the remaining 19 possibilities and generating a trace from each mutation.

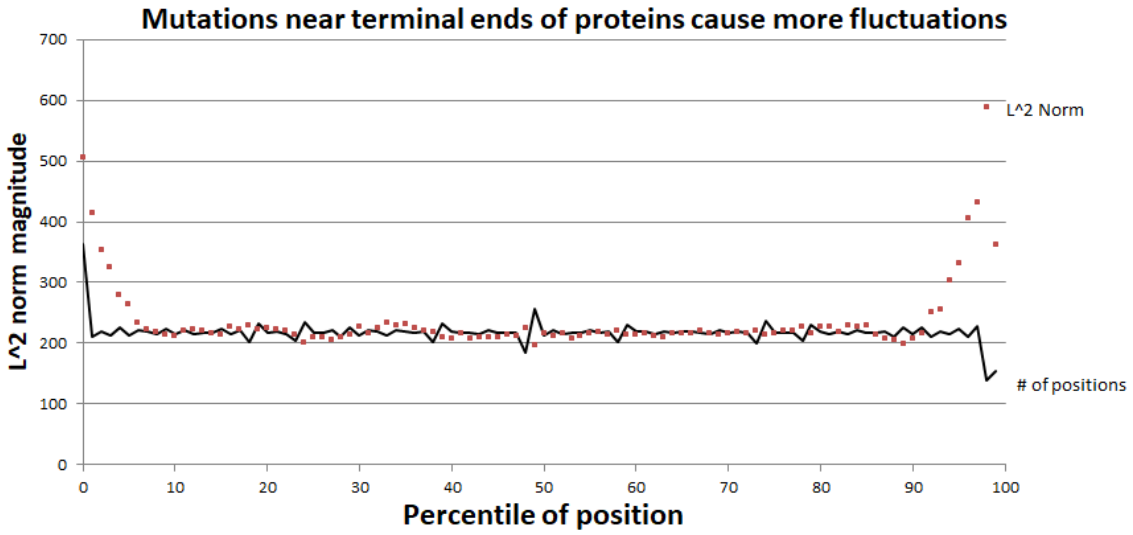


Figure B-19: Mutations in the tail ends of proteins appear to cause greater fluctuations in the protein according to the model. The number of residues in each bin is also plotted to make sure that the deviations in fluctuation are not as a result of biased binning.

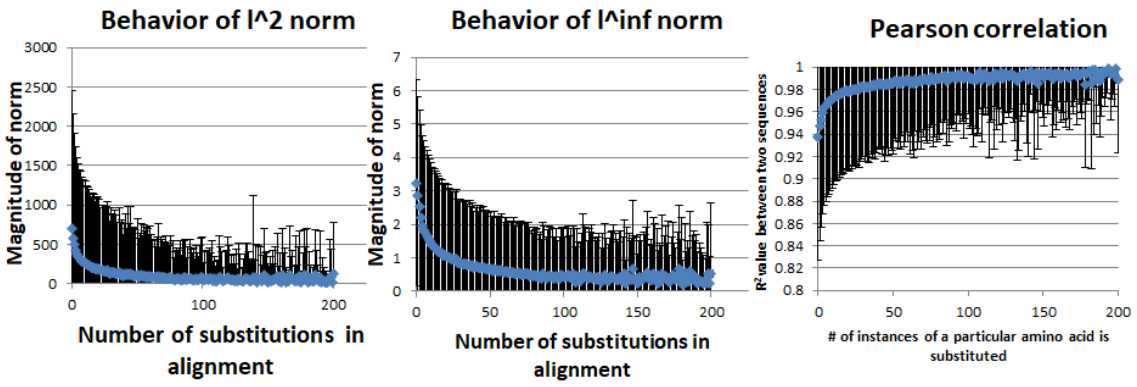


Figure B-20: The L^2 norm, L^∞ norm, and Pearson correlation were used to compare the burial traces for all mutations in each position. For each CDD, and at every position, the average value of the comparison metric was computed and binned by the number of substitution at that alignment position. For each metric it was observed that mutating residues which have much amino acid variance in similar structural domains has a smaller effect on changing the burial trace of the protein.

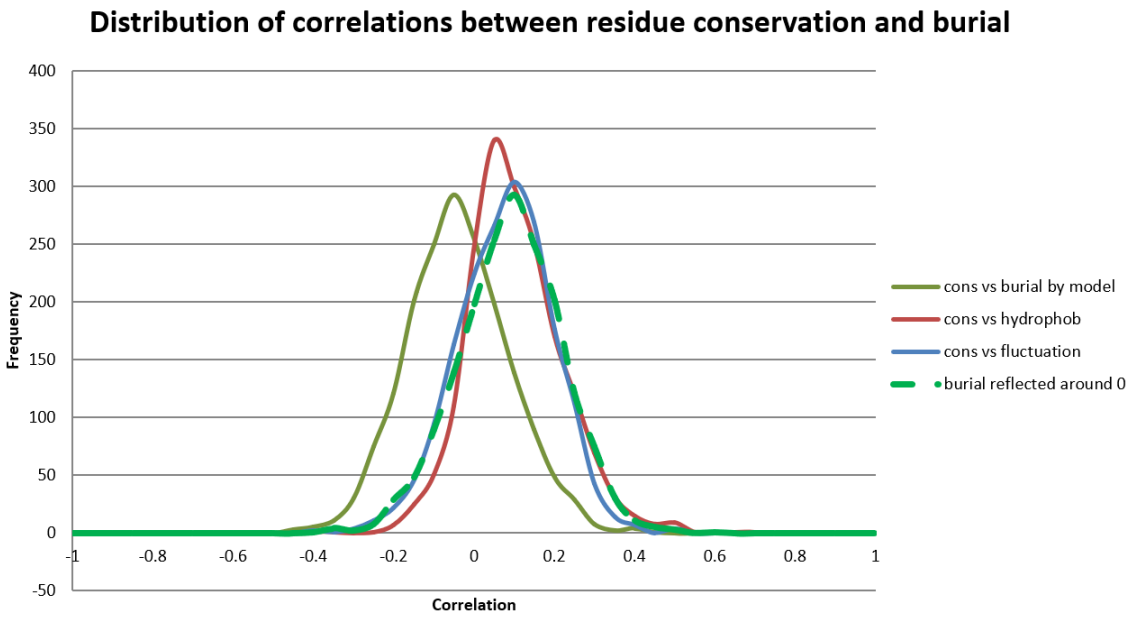


Figure B-21: PAM matrix value of mutation is correlated with the differences in burial trace of the two sequences (fluctuation), the burial of the changed residue (burial) and the difference in hydrophobicity (hydrophobicity) for all observed mutations within a subset of CDD domains(N=1700). Despite the high variance, the fluctuation metric has barely any correlation with conservation and performs no better than Kyte-Doolittle(KD) hydrophobicity. The burial is inversely correlated so the reflected burial correlations are shown to improve comparison.

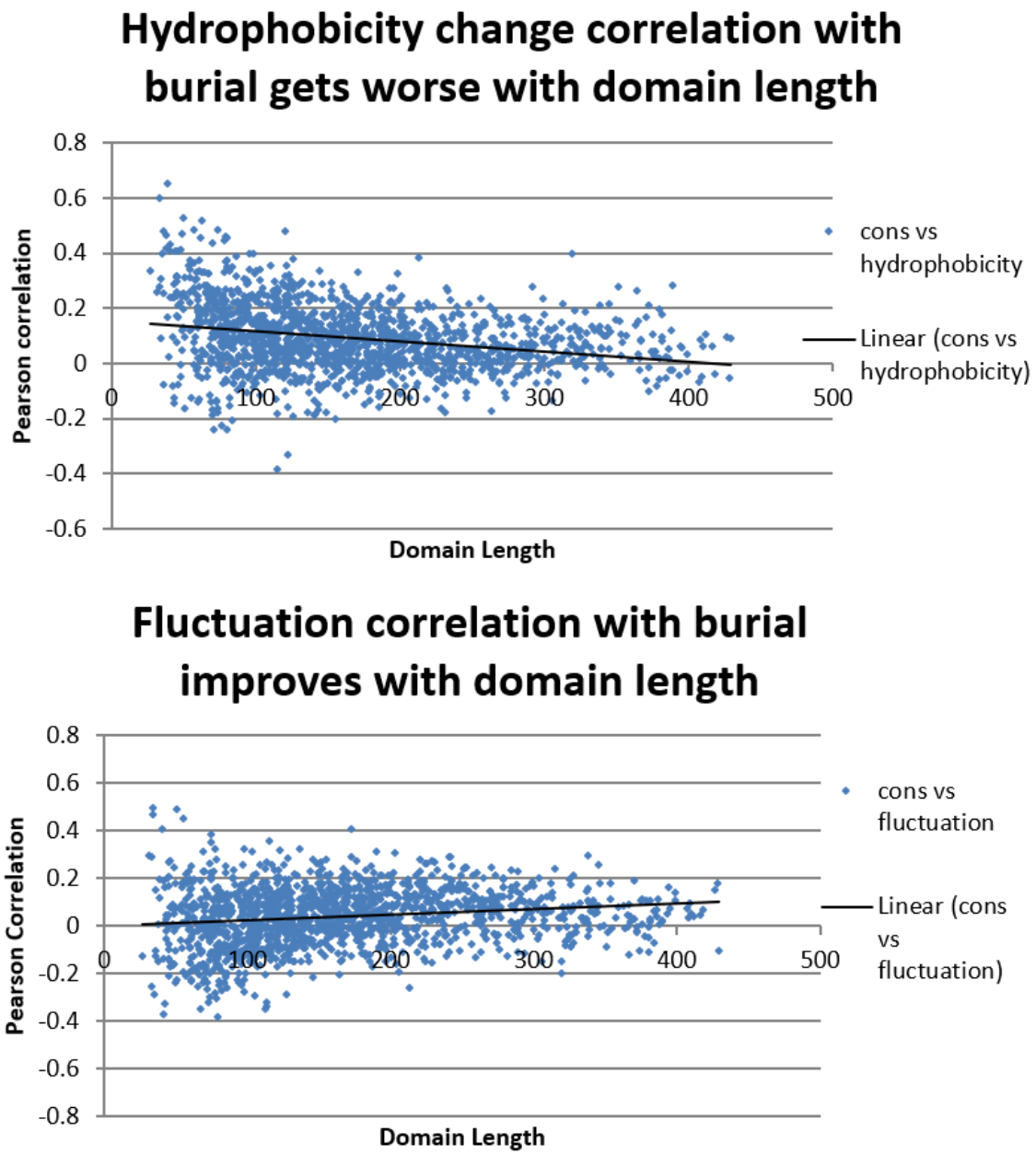


Figure B-22: The top graph shows the trend in correlation between the hydrophobicity change conservation as the length of the domain increases. The bottom graph shows the correlation of conservation with the magnitude of fluctuation associated with the mutation. As the length of the domain increases fluctuation becomes better correlated with conservation while the hydrophobicity becomes a worse predictor.

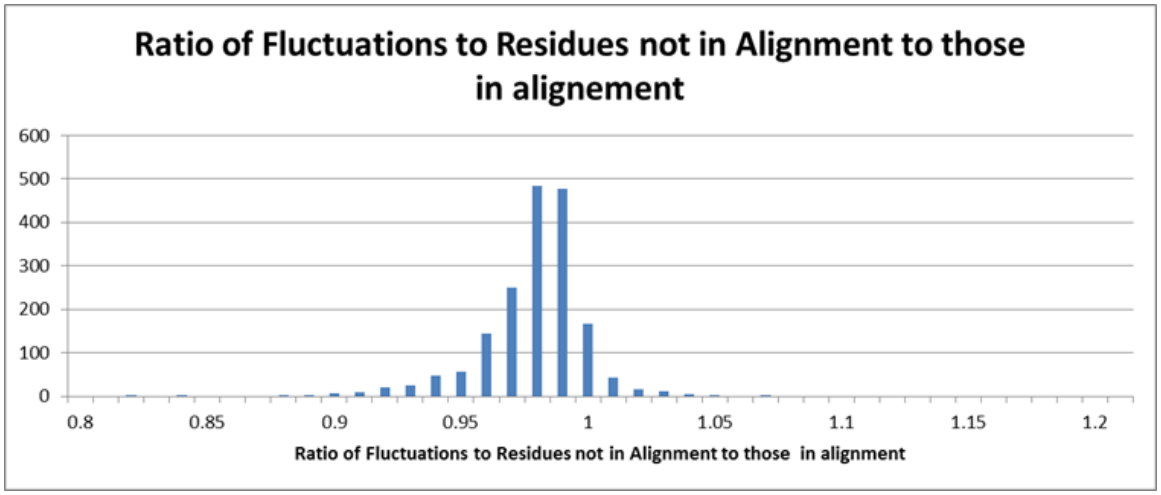


Figure B-23: For each CDD domain, this plot shows the ratio of the average deviation from consensus for every mutation in the CDD alignment to the average deviation for all mutations not present in the alignment. For the majority of structural domains the ratio falls bellow one suggesting that single amino acid substitutions not present in the alignment result in greater structural changes.

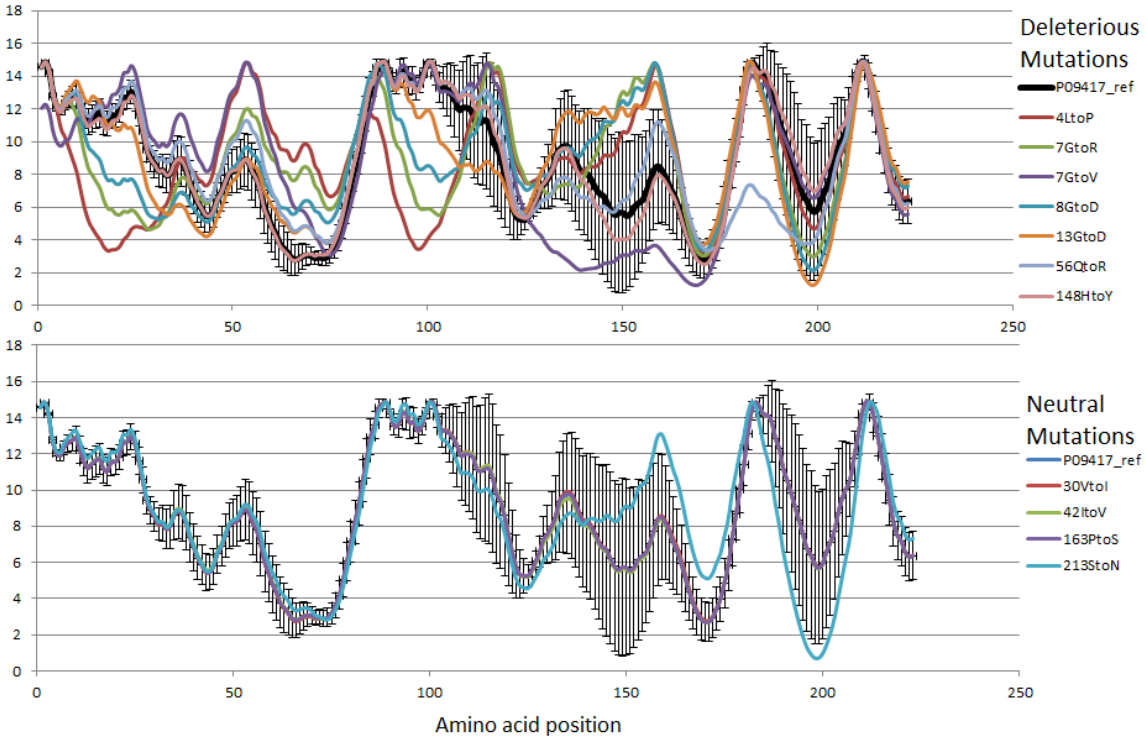


Figure B-24: Deleterious and neutral mutations in the dihydropteridine reductase protein. It is of note that the neutral mutations - even those which have an effect of burial mostly affect the structure in flexible regions while the deleterious mutations have a more global effect on burial.

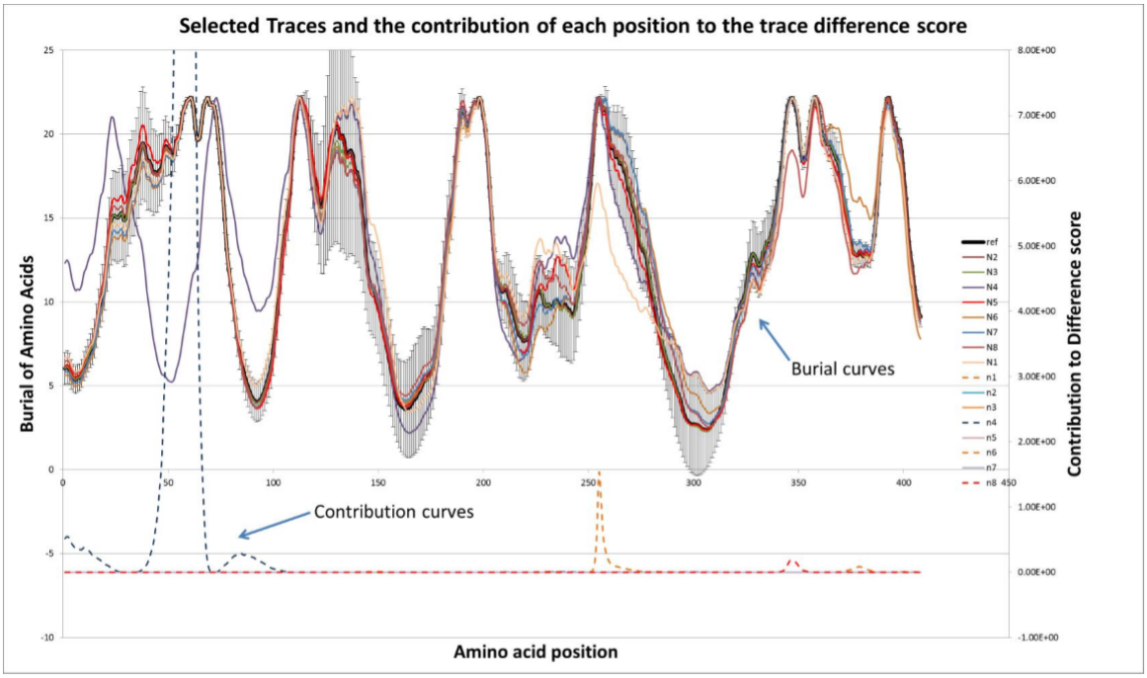
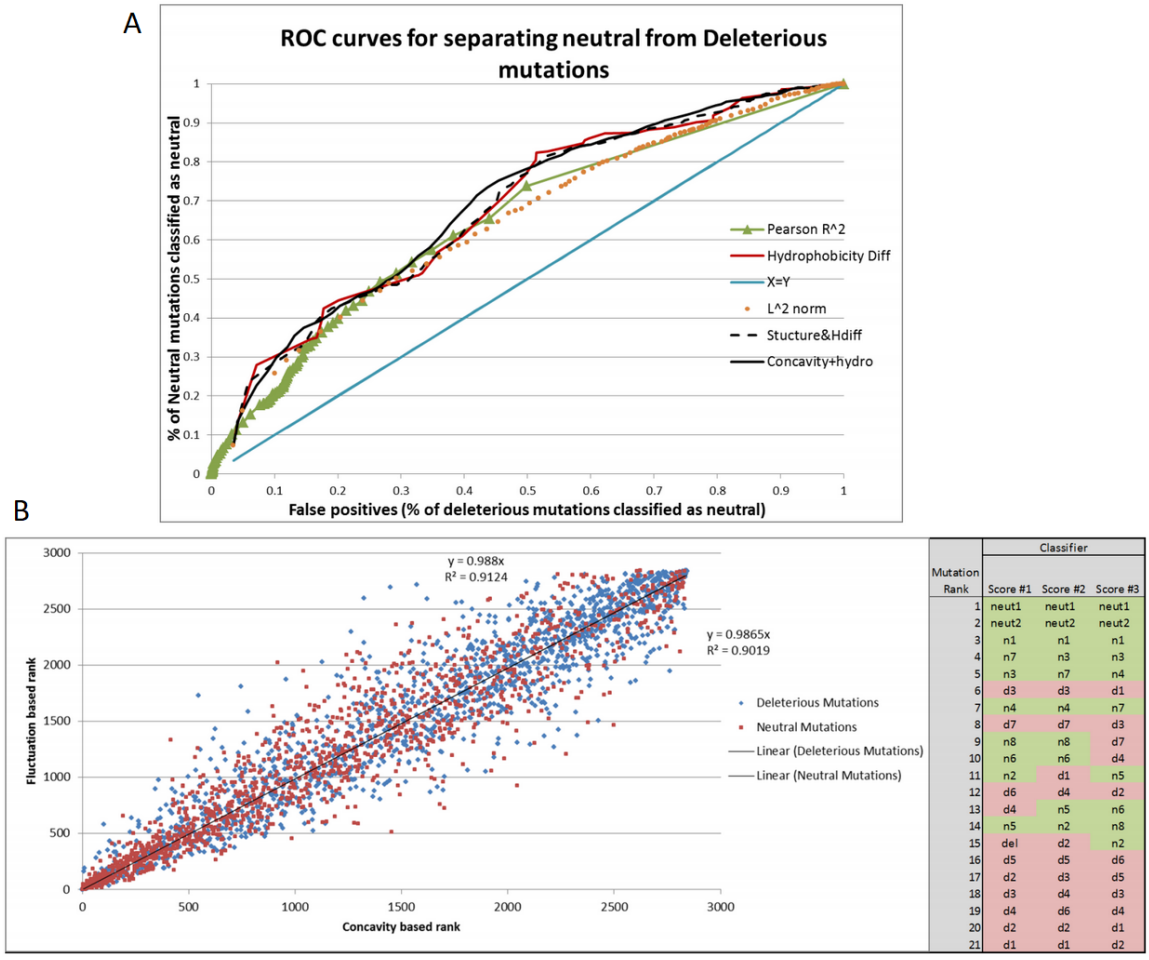


Figure B-25: Example traces for a set of 8 mutations are shown on the left axis. The right y axis plots the contribution of the deviation of the burial trace from the reference at that position. The final score for this fluctuation based metric is the area under the contribution curve. Large deviations where the protein is generally stable have large effects on the score (purple curve residues 55-65) while similarly large deviations in a more flexible region have less contribution (purple curve residues 90-95).



Mutation Rank	Classifier		
	Score #1	Score #2	Score #3
1	neut1	neut1	neut1
2	neut2	neut2	neut2
3	n1	n1	n1
4	n7	n3	n3
5	n3	n7	n4
6	d3	d3	d1
7	n4	n4	n7
8	d7	d7	d3
9	n8	n8	d7
10	n6	n6	d4
11	n2	d1	n5
12	d6	d4	d2
13	d4	n5	n6
14	n5	n2	n8
15	del	d2	n2
16	d5	d5	d6
17	d2	d3	d5
18	d3	d4	d3
19	d4	d6	d4
20	d2	d2	d1
21	d1	d1	d2

Figure B-26: A. The classification performance of various metrics to predict whether a mutations is neutral or deleterious from burial trace. There is no significant improvement of performance relative to just using the hydrophobicity difference. B. The performance of each mutations using the concavity and the fluctuation metric. Despite the fact that there is decent correlation between the two, each metric prioritizes a different aspect of the trace resulting in an alternate ranked list of mutation deleteriousness as shown in the example on the right. The three scores are the fluctuation metric, concavity metric, and the raw hydrophobicity difference. SNPs considered neutral in the training data are green and the deleterious SNPs are red.

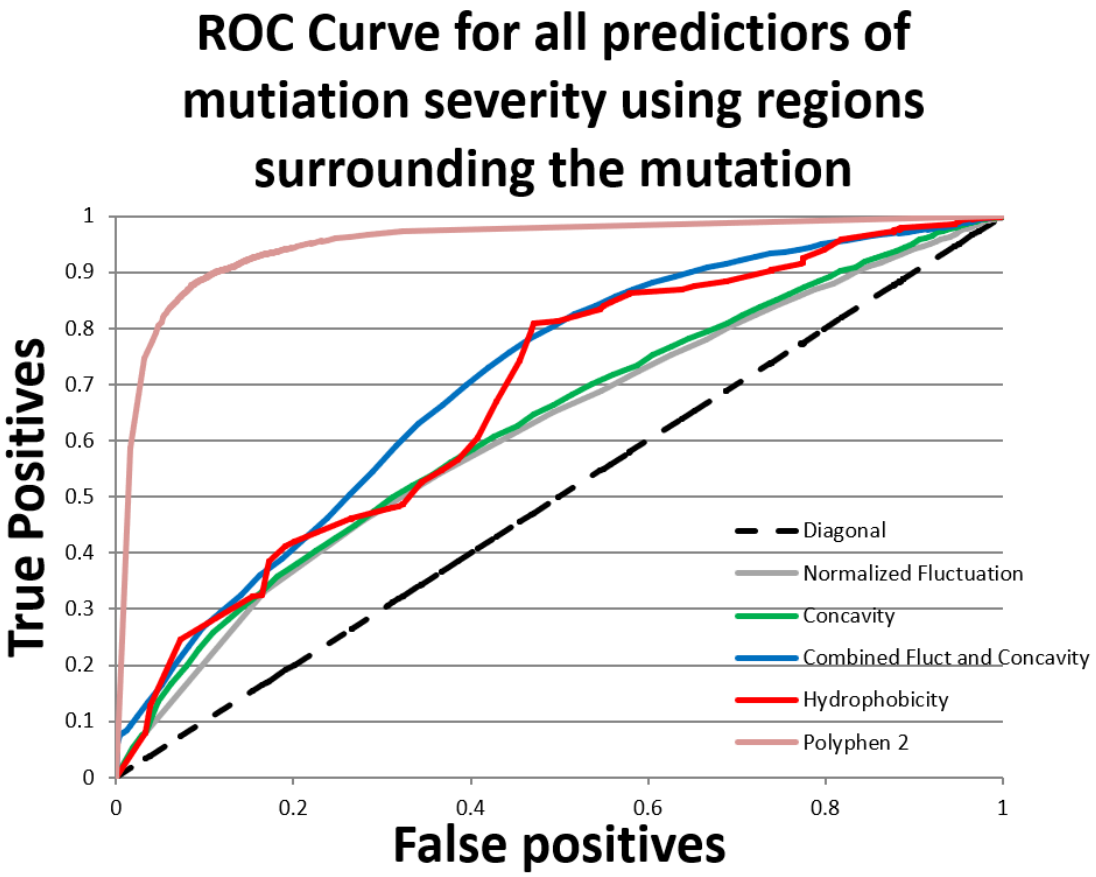


Figure B-27: ROC curves showing how well each metric predicts how deleterious each mutation is based on the effects of the mutation on sequence and predicted structural properties. The Polyphen prediction is shown in pink in comparison to the remaining metrics which use only the sequence information. The best performing curve is a linear combination of concavity and fluctuation metrics which barely outperform the classification based solely on the hydrophobicity change of the mutated residue.

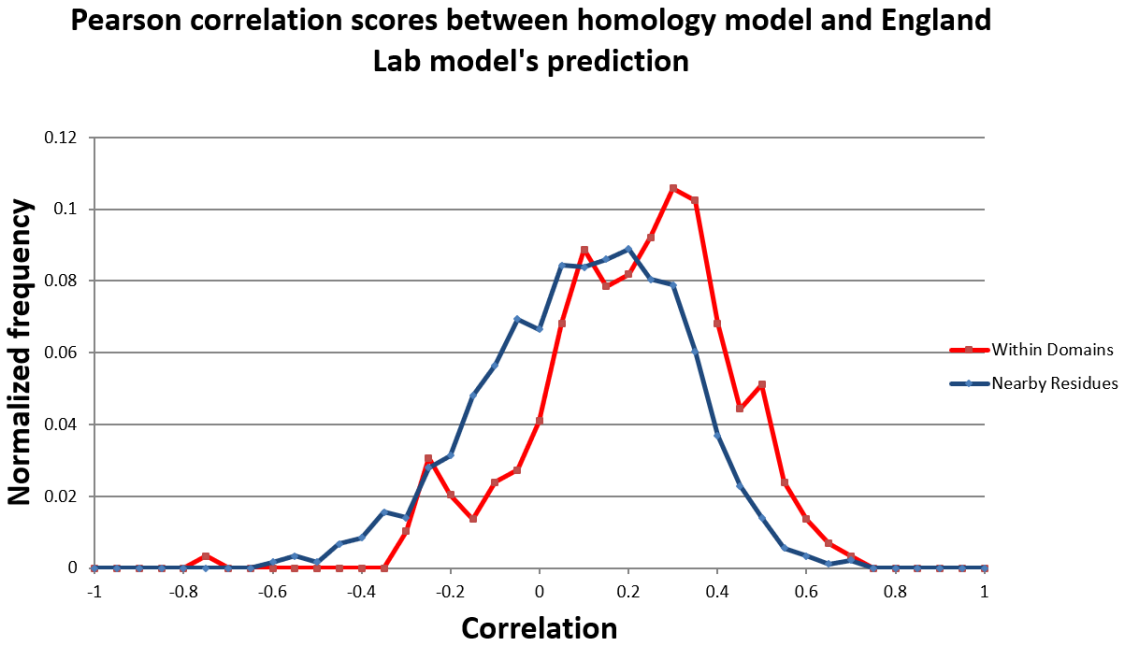


Figure B-28: Two curves are shown that compare the relation between burial predictions generated by the England lab model and homology modeling. When the sequence which us used as input for the models is part of a Conserved domain database(CDD) domain the two models tend to agree more with each other. It's expected that in the CDD domain case the structures are more globular which likely makes both kinds of predictions more accurate.

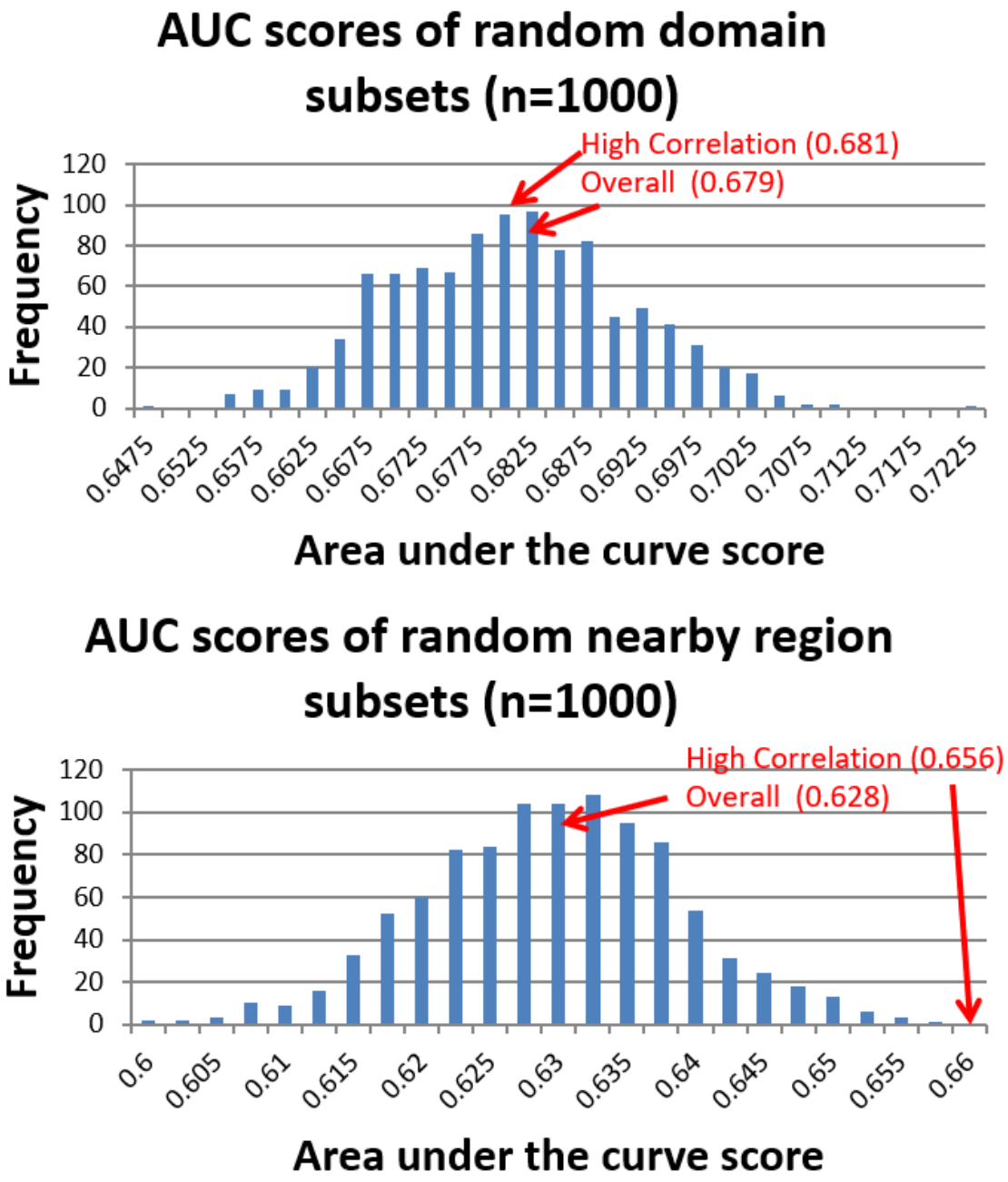


Figure B-29: The two graphs show the correlation for random region subsets of the the same length (400 residues) with the top graph showing the results for 1000 permutations in each of the CDD domains while the bottom graph shows the scores for non CDD proteins. In each case the value of the true AUC for the gene region around the mutation is shown as well as the AUC of the top 10% of genes where the correlation between the homology model and the burial trace is the highest. It's evident that in the case of non-domain limited regions a higher correlation between the two models improves the ability to predict mutation severity.

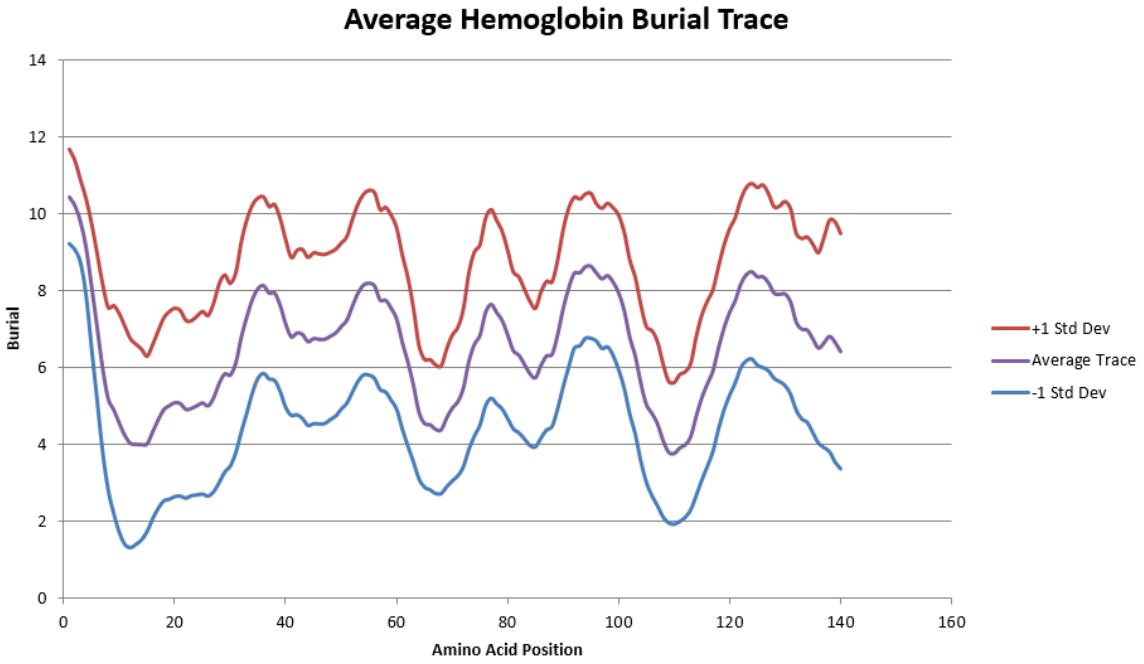


Figure B-30: The average burial trace of all the hemoglobins. The red and blue curves show one standard deviation of the burial trace in the positive and negative direction from the average respectively.

Bibliography

[1] Lin MF, Jungreis I, Kellis M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. Bioinformatics. 2011 Jul 1;27(13):i275-82. doi: 10.1093/bioinformatics/btr209.

[2] Priedigkeit N, Wolfe N, Clark NL. Evolutionary signatures amongst disease genes permit novel methods for gene prioritization and construction of informative gene-based networks. PLoS Genet. 2015 Feb 13;11(2):e1004967. doi: 10.1371/journal.pgen.1004967. eCollection 2015 Feb.

[3] Shen JJ, Wang TY, Yang W. Regulatory and evolutionary signatures of sex-biased genes on both the X chromosome and the autosomes. Biol Sex Differ. 2017 Nov 2;8(1):35. doi: 10.1186/s13293-017-0156-4.

[4] Lin MF, Kheradpour P, Washietl S, Parker BJ, Pedersen JS, Kellis M. Locating protein-coding sequences under selection for additional, overlapping functions in 29 mammalian genomes. Genome Res. 2011 Nov;21(11):1916-28. doi: 10.1101/gr.108753.110. Epub 2011 Oct 12.

[5] Li C, Zhang J. Stop-codon read-through arises largely from molecular errors and is generally nonadaptive. PLoS Genet. 2019;15(5):e1008141. Published 2019 May 23. doi:10.1371/journal.pgen.1008141

[6] Jungreis I, Chan CS, Waterhouse RM, Fields G, Lin MF, Kellis M. Evolutionary Dynamics of Abundant Stop Codon Readthrough. Mol Biol Evol. 2016;33(12):3108–3132. doi:10.1093/molbev/msw189

[7] Jungreis I, Lin MF, Spokony R, Chan CS, Negre N, Victorsen A, White KP, Kellis M. Evidence of abundant stop codon readthrough in Drosophila and other metazoa. Genome Res. 2011 Dec;21(12):2096-113. doi: 10.1101/gr.119974.110. Epub 2011 Oct 12.

[8] Nemoto W, Saito A, Oikawa H. Recent advances in functional region prediction by using structural and evolutionary information - Remaining problems and future extensions. Comput Struct Biotechnol J. 2013 Dec 5;8:e201308007. doi: 10.5936/csbj.201308007. eCollection 2013.

[9] Yang Y, Gao J, Wang J, et al. Sixty-five years of the long march in protein secondary structure prediction: the final stretch?. Brief Bioinform. 2018;19(3):482–494. doi:10.1093/bib/bbw129

[10] Cheng J, Tegge AN, Baldi P. Machine learning methods for protein structure prediction. IEEE Rev Biomed Eng. 2008;1:41-9. doi: 10.1109/RBME.2008.2008239.

[11] England JL. Allostery in protein domains reflects a balance of steric and hydrophobic effects. Structure. 2011 Jul 13;19(7):967-75. doi: 10.1016/j.str.2011.04.009.

[12] Macossay-Castillo M, Kosol S, Tompa P, Pancsa R. Synonymous Constraint Elements Show a Tendency to Encode Intrinsically Disordered Protein Segments. PLOS Computational Biology. May 8, 2014. https://doi.org/10.1371/journal.pcbi.1003607

[13] Sealfon RS, Lin MF, Jungreis I, Wolf MY, Kellis M, Sabeti PC. FRESCo: finding regions of excess synonymous constraint in diverse viruses. Genome Biol. 2015 Feb 17;16:38. doi: 10.1186/s13059-015-0603-7.

[14] Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol. 1980 Dec;16(2):111-20.

[15] Komar AA. The Yin and Yang of codon usage. Hum Mol Genet. 2016 Oct 1;25(R2):R77-R85. Epub 2016 Jun 27.

[16] Turner TN, Yi Q, Krumm N, Huddleston J, Hoekzema K, F Stessman HA, Doebley AL, Bernier RA, Nickerson DA, Eichler EE. denovo-db: a compendium of human de novo variants. Nucleic Acids Res. 2017 Jan 4;45(D1):D804-D811. doi: 10.1093/nar/gkw865. Epub 2016 Oct 5.

[17] Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nat Methods. 2010 Apr;7(4):248-9. doi: 10.1038/nmeth0410-248.

[18] McClellan J, King M. Genetic Heterogeneity in Human Disease. Cell. 2010 April 16; Volume 141, Issue 2: Pages 210-217. https://doi.org/10.1016/j.cell.2010.03.032

[19] Gao F, Keinan A. High burden of private mutations due to explosive human population growth and purifying selection. BMC Genomics. 2014; 15(Suppl 4):S3. https://dx.doi.org/10.1186

[20] Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, Gueroussov S, Albu M, Zheng H, Yang A, Na H, Irimia M, Matzat LH, Dale RK, Smith SA, Yarosh CA, Kelly SM, Nabet B, Mecenas D, Li W, Laishram RS, Qiao M, Lipshitz HD, Piano F, Corbett AH, Carstens RP, Frey BJ, Anderson RA, Lynch KW, Penalva LO, Lei EP, Fraser AG, Blencowe BJ, Morris QD, Hughes TR. A compendium of RNA-binding motifs for decoding gene regulation. Nature. 2013 Jul 11;499(7457):172-7. doi: 10.1038/nature12311.

[21] Khan, A. et al. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Res. 2018; 46:D260–D266, doi: 10.1093/nar/gkx1126

[22] Grant C, Bailey T, Noble, W. "FIMO: Scanning for occurrences of a given motif", Bioinformatics 27(7):1017–1018, 2011.

[23] Timothy L. Bailey and Charles Elkan, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers", Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, pp. 28-36, AAAI Press, Menlo Park, California, 1994.

[24] Valouev A, Johnson SM, Boyd SD, Smith CL, Fire AZ, Sidow A. Determinants of nucleosome organization in primary human cells. Nature. 2011 May 22;474(7352):516-20. doi: 10.1038/nature10002.

[25] Jungreis I, Chan CS, Waterhouse RM, Fields G, Lin MF, Kellis M. Evolutionary Dynamics of Abundant Stop Codon Readthrough. Mol Biol Evol. 2016 Dec;33(12):3108-3132. Epub 2016 Sep 7.

[26] Eswarappa SM, Potdar AA, Koch WJ, Fan Y, Vasu K, Lindner D, Willard B, Graham LM, DiCorleto PE, Fox PL. Programmed translational readthrough generates antiangiogenic VEGF-Ax.Cell. 2014 Jun 19;157(7):1605-18. doi: 10.1016/j.cell.2014.04.033.

[27] Ha SA, Shin SM, Lee YJ, Kim S, Kim HK, Namkoong H, Lee H, Lee YS, Cho YS, Park YG, Jeon HM, Oh C, Kim JW. HCCRBP-1 directly interacting with HCCR-1 induces tumorigenesis through P53 stabilization. Int J Cancer. 2008 Feb 1;122(3):501-8.

[28] Yamazaki T, Sasaki N, Nishi M, Yamazaki D, Ikeda A, Okuno Y, Komazaki S, Takeshima H. Augmentation of drug-induced cell death by ER protein BRI3BP. Biochem Biophys Res Commun. 2007 Nov 3;362(4):971-5. Epub 2007 Aug 27.

[29] Xu D, Zhang Y. (2012) Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins, 2012, 80, 1715-1735.

[30] Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F.T., de Beer, T.A.P., Rempfer, C., Bordoli, L., Lepore, R., Schwede, T. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. 46(W1), W296-W303 (2018).

[31] Zimmermann L, Stephens A, Nam SZ, Rau D, Kübler J, Lozajic M, Gabler F, Söding J, Lupas AN, Alva V. A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. J Mol Biol. 2018 Jul 20. S0022-2836(17)30587-9.

[32] Bae S., Park J., Kim J.-S. Cas-OFFinder: A fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. Bioinformatics 30, 1473-1475 (2014).

[33] Studer RA, Dessailly BH, Orengo CA. Residue mutations and their impact on protein structure and function: detecting beneficial and pathogenic changes. Biochem J. 2013 Feb 1;449(3):581-94. doi: 10.1042/BJ20121221.

[34] Jungreis I, Lin M, Chan C, Kellis M. CodAlignView: The Codon Alignment Viewer, available from http://data.broadinstitute.org/compbio1/cav.php