Structural and functional atlas of frameshift variation capacity in human genome

by

Nan Hu

A Thesis

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

in partial fulfillment of the requirements for the

Degree of Master of Science

in

Bioinformatics and computational biology

May 2018

APPROVED:

_____

Professor Dmitry Korkin, Thesis Advisor

_____

Professor Elizabeth F. Ryder, Thesis Reader

_____

Professor Dmitry Korkin, Head of Department

Abstract

Currently, it is widely accepted that frameshift mutations yield truncated and dysfunctional proteins. Frameshift mutation products are mainly non-functional, abnormal, polypeptides and therefore have gained little attention from the point of view of their structural and functional analyses. However, recent studies have shown that frameshift proteins do have structures and can be functional. While most studies about frameshift mutation focus on the nucleotide sequence level, here we simulate and directly analyze frameshift mutation on the protein domain level. We focus on the protein domain, because it is the smallest structural and functional protein unit. By using protein blast tool to analyze the protein domain yield from all coding gene sequences in the human genome (45,139 mRNAs), we found out that 11,313 polypeptide sequences resulting from +1 frameshift mutation and 10,278 sequences resulting from +2 frameshift mutation are homologous to the existing proteins and could potentially carry out a function. Moreover, for 464 and 448 frameshift products in each type, respectively, we detected at least one protein domain by using Interproscan tool. We also compared the genes where we found the frameshift-produced protein domains with the genes associated with frameshift mutations reported by Clinvar database. The result shows that 47 genes from our set were also found to carry clinically-relevant frameshift mutations. This work provides the first whole-genome view of the frameshift effects on the protein domain structure and function, which would shed new insights about this variation mechanism with applications in a wide range of areas from evolutionary biology to precision medicine.

Acknowledgements

I would like to express my gratitude to all those who helped me during the writing of this thesis.

My deepest gratitude goes first and foremost to my advisor, Professor Dmitry Korkin, for his constant encouragement and guidance. Then, I would like to thank my thesis reader, Professor Elizabeth F. Ryder, for her advice and support of completing of this thesis.

I would also like to appreciate all the members in Korkin Lab who light me up when we had discussions. In the meantime, I have to thank my best friend and classmate Xiaojun Wang who helps me debug the programs.

Last but not the least, I am so thankful for my families that support me during my graduate study.

All in all, thank everyone who helped me during my graduate time.

Table of Contents

List of Figures

List of Tables

1 Background

1.1 Introduction

The research we performed in this thesis belongs to the area of molecular biology; this is a critical area of research since mutations are the contributor of evolution, as well as the contributor of genetic disease.

Unraveling the consequence of frameshift mutation has a lot benefits. First of all, it is well understood that beneficial point mutations accumulated and finally contribute to species evolution [1], but whether an organism can ever benefit from a frameshift mutation is still mysterious. Besides, frameshift mutations lead cancers and only gene therapy could be used to treat disease nowadays [2]. Studying the consequence of frameshift mutation could discover potential drug target and help researchers to design medicines. For this reason, the researcher`s role in expanding the knowledge and understanding of frameshift mutation is critical and valuable.

Due to recent surge in research on frameshift mutation, the consequences of it either partially or completely change the DNA sequence after the spot of frameshift mutation happens [3]. Even so, researchers still need to study how the mutation will bring to change to its transcripts or even protein products. In this research, we use computational methods to study the changes in mRNA sequence resulting from frameshift mutation. We attempted to use bioinformatics tools such as mutation simulation to tackle this problem.

Furthermore, recent research has shown that protein domain centric approach is better than genetic centric approach. That is analyzing the frameshift mutation from a protein domain perspective instead of looking its mRNA nucleotide sequences [4]. The benefits of doing protein domain centric approach is obvious because protein domain is the functional unit of protein. Clearly,

frameshift mutation will cause protein losing its function or gain protein a new function. In this paper, we are trying to identify those frameshift mutations that lead to a protein that perform the same or a new function.

The main challenge usually faced when using these computational tools is to simulate frameshift mutation in mRNA, then translate the frameshifted mRNAs to protein sequences, and finally blast against protein in-real-world and to detect their domain structures. These works are computationally expensive, especially for novel domain structure conformations.

Our results could contribute to a future research of evolution and also drug design to produce a precise medicine against genetic disease.

1.2 Frameshift mutation

Frameshift mutation is a genetic mutation caused by a number of base indels (insertion or deletion) in DNA which cannot be divided by three. Because gene expression count on codon which consists of three nucleotides, frameshift mutation will lead to a different reading frame, and finally translate into a whole new peptide contrast with the original peptide sequences. Everything after the spot of indels will partially or completely change [5].

Based on the sequence shift position against the original sequence, frameshift mutation can be categorized into two types: +1, +2. They are also written as +1, -1 in some literature. (Figure 1).

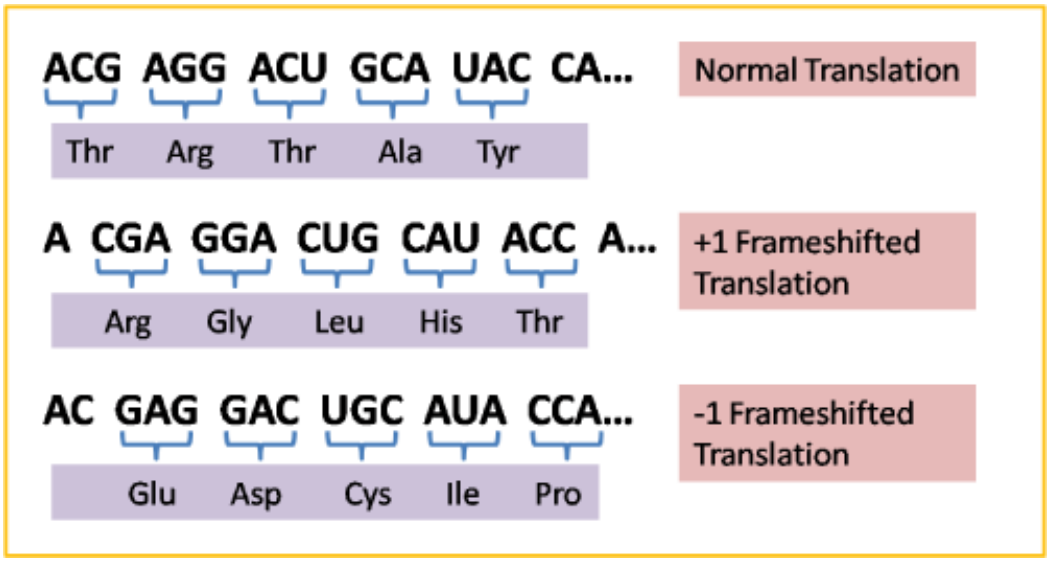

Figure 1. Schematic representation of frame-shift events with their +1 and -1 versions. [6]

In this research, we will analyze the protein sequence at domain level because the domain considered as the function unit in a protein. The domain themselves usually play a role of a particular function or are responsible for interaction, contributing to the overall function of a protein.

Three types of peptides will generate due to frameshift mutation. (1). Domain disappears and comes out of a plain protein sequence that is a sequence that has no homology to any identified protein domain; (2) The original domain will be the same functional domain or a new functional domain. (3) The plain protein sequence becomes to a domain. In this paper, we are trying to analyze the option (2), and interpreting our findings which could be a great deal of evolution biology and oncology studies.

1.3 Evolution factors

Currently, evolution is known to be caused by natural selection, gene flow, environmental factors, and mutations such as missense, nonsense, and duplications [7]. It is unknown whether frameshift mutations contribute to evolution.

Researchers identified and studied these mutations by analyzing tissue samples collected from patients affected by pathogenic frameshift mutations. Therefore, many frameshift mutations, along with the evolutionary information they may contain, remain unidentified. However, previous studies showed that frameshift coding genes can be expressed, and frameshift proteins can be functional by themselves [8].

Recent study has shown that the universal genetic code, protein coding genes and genomes of all species were optimized for frameshift tolerance [8]. This work points out that frameshift homologs are defined as a set of frameshifted but yet functional coding genes/proteins that were evolved from a common ancestor gene via frameshift mutation.

Another study shows that a frameshift mutation in CCR5 genes will give resistance ability to infect with HIV [9]. CCR5 is a co-factor which is responsible for HIV entering the cell [10]. A 32-base pair deletion in CCR5 has been identified as a mutation that negates the likelihood of an HIV infection [11]. This region on the open reading frame contains a frameshift mutation which introduce a premature stop codon [12]. This mutation leads to the loss of function of biding HIV. CCR5-1 is considered as the wild type and CCR5-2 is regarded as the mutant allele. People with a heterozygous CCR5 mutation were less sensitive to the infection with HIV [13]. In a study, even through one exposure to a high concentration level of HIV virus, no one homozygous for the CCR5-2 mutation was reported as positive for HIV [14]. This kind of frameshift mutation could be considered as a beneficial for individual`s life, and thus this mutation could be considered as an evolution to some degree. Besides, researchers can mimic this molecular behavior and decide to knock out protein domain in order to prevent from infection by pathogen.

For these reasons, this paper will help identify a protein domain evolution road map, and provide information to researchers to navigate these maps. This

method could also expand to other species to detect the protein domain evolution route since protein domains exist in all species. Moreover, this method could apply to other form of RNAs such as 'NR_', which is for RNA not coding, to help decipher the regulatory regions.

1.4 Genetic disorder factors

A genetic disorder is a disease caused by an abnormality in DNA. These abnormalities can range from a single nucleotide mutation to a deletion or insertion of an entire chromosome [15].

Frameshift mutations are mutations caused by insertions or deletions of one or two nucleotides from a DNA sequence. Because tRNA translates codons, groups of three mRNA nucleotides, to amino acids [16], frameshift mutations lead to a shift in the tRNA reading frame and thus a perturbed protein [17]. These mutations generally occur in hot spots, repeated sequences of one or two nucleotides. This is due to a ‘slip’ of the DNA polymerase followed by the realignment of the DNA template and nascent strand during replication [18]. However, frameshift mutations can also occur elsewhere in a DNA sequence. They lead to either an inactive protein or a protein with an altered structure and function. Both of these cases are very dangerous and can result in many severe diseases such as Crohn’s disease [19], Cystic Fibrosis [20], Tay-Sachs disease [21] and several types of cancer.

For this reason, this paper will generate potential cancer development candidates in the level of protein domain, and provide information to researchers to identify drug target and design new medicines.

2 Methods

The purpose of this research is trying to find all possible protein domains introduced by frameshift mutation across human genome. All the works are in silico. Data collection and analysis are mostly using Python. Software such as InterproScan requires Linux environment. The SMART which used to detect domain architecture is a web-based tool.

In general, the pipeline of this research is collecting all mRNAs of homo sapiens. Then apply simulation frameshift mutation at the translation position without considering any stop codons, including the stop codons that are created by the frameshift and its original stop codon, and translate the nucleotide sequences into peptides. Third, by blasting the protein sequences we generated will filter out the sequences which meet the e-value of 0.0001. These sequences are considered as potential functional proteins. These potential functional proteins will put into Interproscan to identify their domains and homologue superfamily. The output of Interproscan would be a bunch of sequences that annotated with domain and superfamily information. These sequences will be used to build up a domain architecture and finally construct a frameshift domain structural atlas. The gene we found will be compare with genes that are known to be associated with frameshift disease. This practice is regarded as an evaluation of our results. See the pipeline workflow below(Figure2).

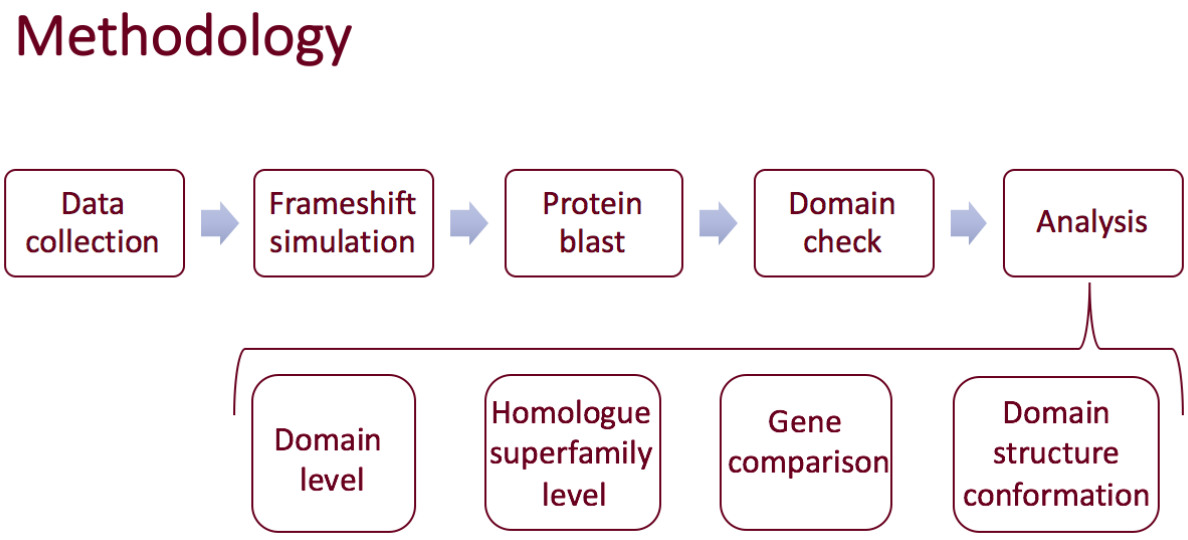

Figure 2. Methodology workflow

2.1 Data collection

The mRNA accession numbers were retrieved from the file RefSeq transcripts of GRCh38 downloaded from NCBI. Then we developed a script to connect to the NCBI Entrez API and fetch mRNAs from the NCBI nucleotide database [22]. 45,139 mRNAs were retrieved from the database in genebank format. This format allows us to identify the start position of translation on mRNA. This is the position that we simulate our frameshift mutation.

2.2 Frameshift simulation

After we get the mRNA sequence, we started to simulate frameshift mutation in the open reading frame. In this step, we use biopython packages [23] to perform +1/-1 simulation in mRNA, and then translate them into protein sequence (Figure 3).

The frameshift start point at the translation start point in mRNA (marked as a dark red arrow), all the rest residues will be translated. We labeled it with a pink grill

pattern. Because we ignore all the stop codons, the poly-A tail in mRNA will also be translate into proteins.

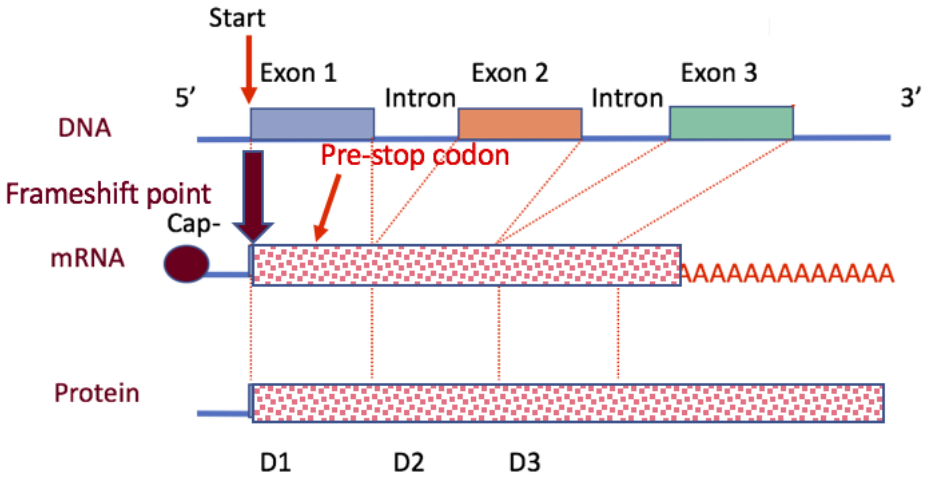

Figure 3. Simulation design.

Since frameshift mutation could happen in any spot of the sequence in real scenario, we set up a simulation design without considering any inner stop codon among the whole sequence. In other words, there are multiple stop codons within the frameshifted sequence we get. In this way, it allows us to include all possible frameshift cases and all possible incoming protein domains. We give an example in figure 4 The nucleotide sequence frameshifted and resulting in 5 inner stop codons. In order to include all possible domains, we set up a "stop=False" in our program and obtained the peptides.

The packages we use is SeqIO which allows us to extract nucleotide sequence from genebank format file. Imbedded function "translate" will directly give us frameshift protein products. "Stop=False" is a parameter in the "translate" function.

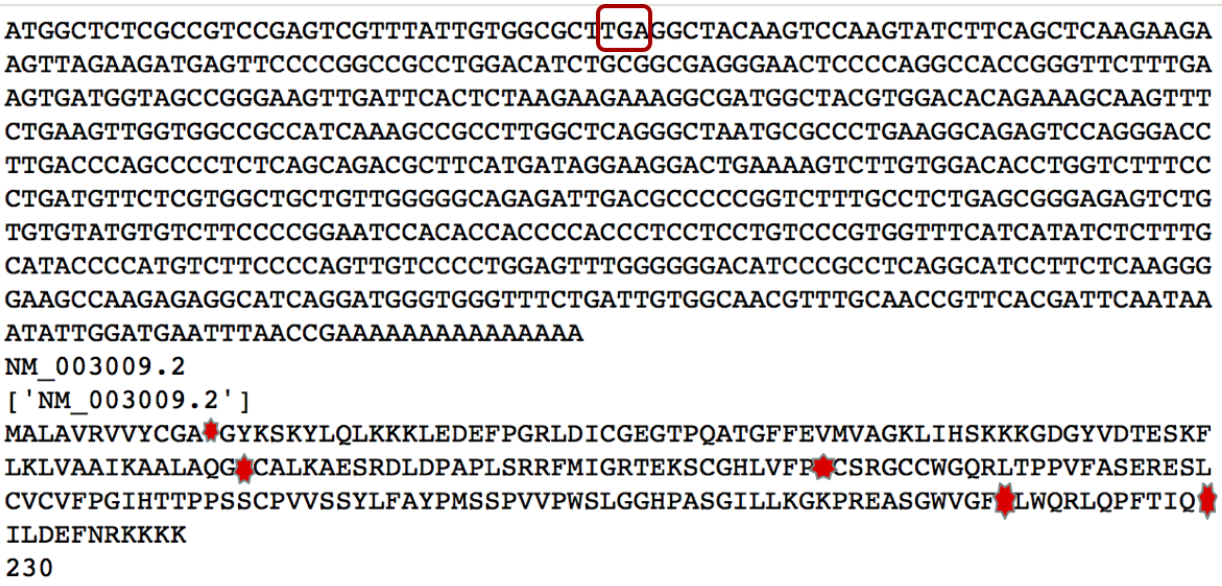

Figure 4. An example of frameshift Simulation design.

There is another reason that the stop codons need to be ignored. We found evidence that stop codons can code amino acids. The transcript accession number is NM_003009. The protein accession number is NP_003000.1. It is reported that UGA stop codon recoded as selenocysteine. This is rare case, but it exists. Therefore, the simulation will keep stop codons in the simulation frameshifted sequences.

2.3 Protein Blast

The Basic Local Alignment Search Tool (BLAST) helps identify regions of local similarity between sequences. This program compares nucleotide or amino acid sequence databases and calculates a bit score which is the statistical significance of matches [24]. The BLAST can be used to infer functional and evolutionary relationships between sequences as well as find members of gene families [25].

We wrote a script and used the NCBI 'BLASTP' command line in order to submit a large number of BLAST jobs. This step is extremely computationally expensive, so we downloaded the protein database which we BLASTed against

within a local computer, and did the protein BLAST locally. The protein reference sequence is 'GRCh38_latest_protein.faa'.

In this step, we keep * in the sequences as a stop codon for the following reasons. First, the BLSAT program recognize the * mark as a stop codon. Second, the * will not match to any amino acids if it includes in a query sequence, and this practice will lower the total score of that alignment. Therefore, we keep * in sequences and BLAST them.

The aim of doing blast is select frameshift proteins comparing with the well documented protein database. So, in this step, the frameshifted proteins are blast against human RefSeq proteins (GRCh38_latest_protein.faa, downloaded from NCBI on Jan/15/2018). 'GRCh38_latest_protein.faa' include NP_ labled protein sequence as well as XP_ labled protien sequences. NP_ labled protein sequences are those with biochemical evidence, while XP_ are those predicted proteins. The threshold is a 0.0001 e-value, and the BLOSUM62 matrix was used. The Expect Value was reduced to 0.0001 from the default of 10 in order to increase the speed of the BLAST processing time and ensure meaningful results. BLOSUM62 is most effective in finding all potential similarities including 30-40%. Using BLOSUM62 will cover a broader range of potential functional frameshift proteins into our candidates.

We retain those sequences which have blast results and discard other sequences. Now, these retained proteins are called potential functional protein since they exist in the real-world to some degree. We report the BLAST criteria we performed below. The e-value and matrix are selected; all others are default values.

database	GRCh38_latest_protein.faa
e-value	0.0001
word size	2
gap open	11
gap extend	1
matrix	BLOSUM62 matrix
comp based stats	0

Table 1. BLAST parameters

2.4 Domain detection by Interproscan

InterProScan is the software package that scan the sequences (protein and nucleic) against InterPro's signatures and return annotations of the input sequences. By classifying sequences into families and predicting the presence of domains and important sites, InterPro uses as a tool that analyze the structure and function of protein sequences [26]. To reach this goal, InterPro uses signatures which is the predictive models. These models are provided by several different databases (referred to as member databases) which make up the InterPro resources.

This step is also computationally expensive. In order to save time and submit a large number of jobs at the same time, we download the Interproscan tool and do protein domain detection locally. The version of the software is interproscan-5.27-66.0 which downloaded on Feb 24th,2018. This software requires Linux environment. We write a bash script to automatically submit thousands of jobs orderly to Interproscan.

In this step, we aim to identify if these potential functional proteins have domains. We detect +1/-1 protein domain first and then select those candidates who have the domain found. If the frameshifted ones turn up positive, we will annotate its original sequences by InterproScan. This is a trick to shorten the time of running Interproscan.

Problem is that the InterproScan will not allow a sequence with a * inside. Stops are converted to X's in our sequences to allow comparison of the entire frameshifted sequence to the database.

2.5 Domain analysis

In this step, we aim to classify potential functional proteins into two categories: no domain detected group and domain found group. The domain found group will retain for future analysis.

The tool we use is a simply XML parser written in python. This step was performed in Jupyter notebook [27]. We retrieve those domains who meet the requirement of entry.attrib['type']=="DOMAIN". Domains are distinct structural, functional or sequence units that may exist in a rich and variety of biological contexts. A match to an InterPro entry of this type indicates the presence of a domain [28]. We extract information including accessing number; domain name, domain position, domain function of both original and frameshifted sequence. We then compare the position of the original sequence and the frameshift sequence. Once there is a position overlap between two domains, we claim that they are evolution connected because of frameshift mutation (fig.5).

A domain evolution map caused by frameshift mutation was made in order to come up with the pattern that human can easily recognize. Using JavaScript, we were able to build an interactive protein domain evolutionary map. Also, we pay special attentions to domains which only become to a new domain, and generate patterns based on their mutation strength and domain frequency. This work was performed in Cystoscope.

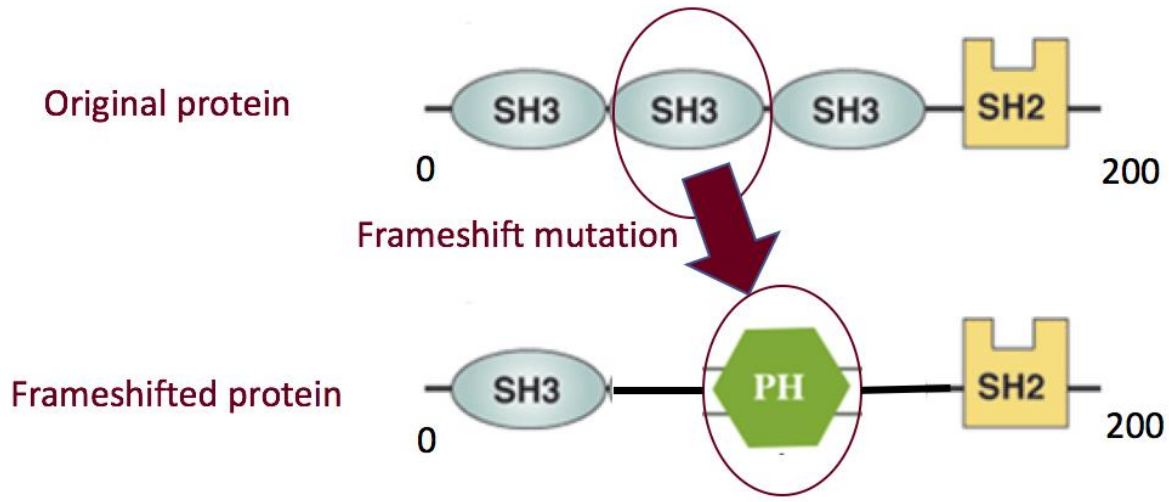

Figure 5. Position overlap between the original SH3 domain and frameshifted PH domain.

2.6 Homologue superfamily analysis

A homologous superfamily is a group of proteins that share a common evolutionary origin, reflected by the similarity in their structure [29]. Since superfamily members often display very low similarity at the sequence level, this type of InterPro entry is usually based on a collection of underlying hidden Markov models, rather than a single signature [30].

The tool we use is a simple XML parser [31] written in python. This step was performed in Jupyter notebook. We retrieve those domains who meet the requirement of entry.attrib['type']=="HOMOLOGUE_SUPERFAMILY".

We extract information, including access number; homologue superfamily name; homologue superfamily position of both original and frameshifted sequence.

We then compare the position of original sequence and frameshift sequence. Once there is an overlap between two homologue superfamily, we claim that they are evolution connected because of frameshift mutation. A homologue superfamily

evolution map caused by frameshift mutation was made in order to come up with the pattern that human can easily recognize.

An interactive protein domain evolutionary map was generated by JavaScript. Also, we filtered out homologue superfamily which only become to a new homologue superfamily, and generate patterns based on their mutation strength and homologue superfamily frequency. This work performs in Cystoscope.

2.7 Gene comparison

We also compare genes, which have frameshifted results according to our analysis, to the ClinVar database. ClinVar is a freely accessible, public archive of reports of the relationships among human variations and phenotypes, with supporting evidence [32]. We retrieve all genes (4059) which have been annotated by Clinvar as frameshift mutation disease.

We prepare our candidate diseases associated genes in the following ways. First, we identify the mRNAs which have frameshifted domain found. Then, we extract gene names, including gene name synonyms and turn these into the candidate data. There are 863 mRNAs coding frameshifted proteins which are domains found by InterproScan. 1931 gene names including synonyms are extracted from these mRNAs.

This step could be considered as an evaluation of the whole research work. Because all we do is in silico experiment, and if the evidence found in the real world, we could say that partially our experiment design is right. Besides, other potential functional proteins might exist, but not reported yet. These results will give a new guidance of research in cancer biology.

2.8 Domain structure conformation

The Simple Modular Architecture Research Tool (SMART) [33] was used to create protein domain architecture from the mutated, functional protein products returned by InterPro. SMART was accessed manually because there were very few sequences that needed to be analyzed. Instead, we can do a batch search.

Firstly, the original protein sequence of match genes according to the last step will be collected and formatted in to FASTA format. Then we run them in SMART. The genes which reported containing a protein domain will be collected. We also collect the frameshifted protein sequences of these genes and run them in SMART again in order to compare the original and frameshifted domains.

Problem is that the batch search in SMART will only return results that the input sequences are in their database [34]. In other words, if the input sequence is not in their database, there will return nothing. So, Therefore, we analyzed domain architecture manually, one gene at a time.

3 Results and analysis

3.1 Protein blast results

After simulating frameshift mutation in both +1 and +2 types, we blast these frameshifted protein sequences against Refseq Protein (GRCh38). 11,313 of 45,139 mRNAs as +1 frameshift mutation sequences and 10,278 of 45,139 as +2 frameshift mutation sequences are reported with blast results (fig.6). In other words, these frameshift products would exist in the real world according to the human protein atlas. Thus, they are considered as potential functional proteins.

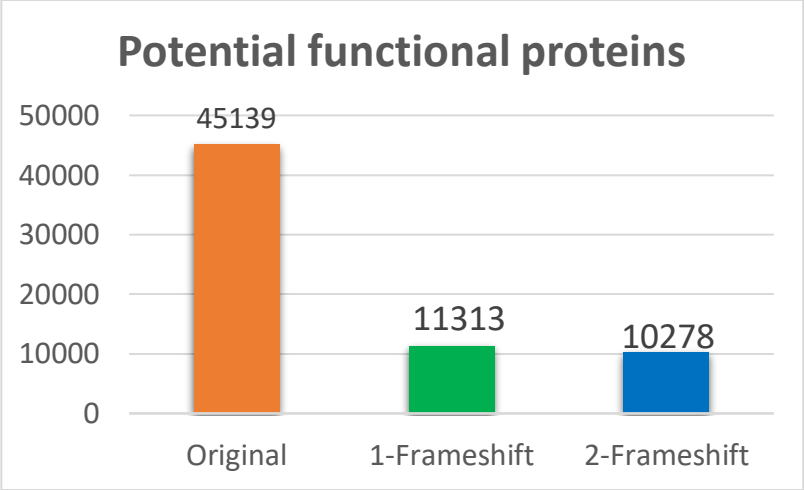

Figure 6. Quantity of frameshifted products with blast results and original ones

The sequence identity is a good estimate value for blast alignment results. The extent to which two amino acid sequences have the same residues at the same positions in an alignment, often expressed as a percentage. We include all aligned sequences and calculate their sequence identities. For example, NM_000020.2 matches 14 sequences through BLAST. We include all the identities to see the quality of the frameshifted proteins. (Table2)

match_id	identity	%	decimal
NM_000020.2>NP_001265116.1	(28, 35)	80%	0.8
NM_000020.2>NP_001268361.1	(32, 49)	65%	0.65
NM_000020.2>NP_001287848.1	(22, 26)	85%	0.85
NM_000020.2>NP_001317315.1	(30, 41)	73%	0.73
NM_000020.2>NP_001317316.1	(30, 41)	73%	0.73
NM_000020.2>XP_006712090.1	(30, 41)	73%	0.73
NM_000020.2>XP_011527398.1	(22, 36)	61%	0.61
NM_000020.2>XP_011531179.1	(30, 41)	73%	0.73
NM_000020.2>XP_011531180.1	(30, 41)	73%	0.73
NM_000020.2>XP_016856993.1	(28, 43)	65%	0.65
NM_000020.2>XP_016856994.1	(28, 43)	65%	0.65
NM_000020.2>XP_016859721.1	(30, 41)	73%	0.73
NM_000020.2>XP_016859722.1	(30, 41)	73%	0.73
NM_000020.2>XP_016861836.1	(25, 35)	71%	0.71

Table 2. NM_000020.2 matches and their sequence identity.

By analyzing the sequence identity in either 1-frameshift or 2-framshift mutation, we found that most of the scores are located in range 40% to 70% (Figure 7 and Figure 8). This indicates that most of the sequences are identical to the extent of 40%--70%.

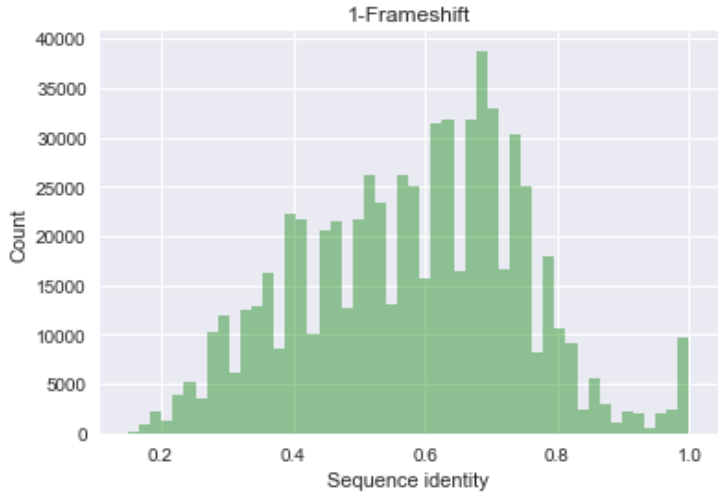

Figure 7. 1-Frameshift sequence identity

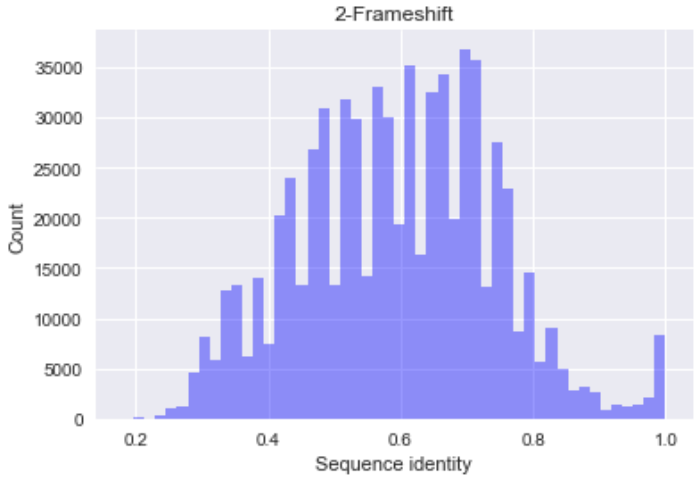

Figure 8. 2-Frameshift sequence identity

3.2 Domain and homologue superfamily detection

The InterproScan returns results in XML format (Figure 9). In tag "entry", "type=DOMAIN" is the filter threshold. We filter out results that contain information "type= DOMAIN ", which suggest that this is a domain part within the blast sequence. Domain name, function, and location of the domain are also shown in the results.

This practice is the same to filter out homologue superfamily sequences. The threshold is "type=HOMOLOGUE_SUPERFAMILY".

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<protein-matches xmlns="http://www.ebi.ac.uk/interpro/resources/schemas/interproscan5" interproscan-version="5.27-66.0">
    <protein>
        <sequence md5="746042b812b157e96d8143864ad2df77">MVPPVWTLLLLVGAALFRKEKPPDQKLVVRSSRDNYVLTQCDFEDDAKPLCDWSQVSADDEDWV
        <xref id="NM_003386.2_original"/>
        <matches>
            <hmmer2-match evalue="6.0E-85" score="298.2">
                <signature ac="SM00137" name="MAM_2">
                    <entry ac="IPR000998" desc="MAM domain" name="MAM_dom" type="DOMAIN">
                        <go-xref category="CELLULAR_COMPONENT" db="GO" id="GO:0016020" name="membrane"/>
                    </entry>
                    <models>
                        <model ac="SM00137" name="MAM_2"/>
                    </models>
                    <signature-library-release library="SMART" version="7.1"/>
                </signature>
                <locations>
                    <hmmer2-location score="160.2" evalue="2.0E-43" hmm-start="1" hmm-end="192" start="369" end="536"/>
                    <hmmer2-location score="100.6" evalue="1.9E-25" hmm-start="1" hmm-end="192" start="36" end="204"/>
                    <hmmer2-location score="37.4" evalue="2.8E-7" hmm-start="1" hmm-end="192" start="206" end="368"/>
                </locations>
```

Figure 9. An example of Interproscan results.

3.3 Summary of simulation, blast and domain detection

The total mRNAs we get was 45,139. For each mRNA, we applied 1-frameshft and 2-frameshift mutations. Through BLAST, we found 76% of them did not return any blast results, 24% returned with results which indicated these frameshifted proteins. The absolute number was 11,313 and 10,279 for 1-frameshift and 2-frameshift, respectively. According to InterproScan results, only 1% of frameshifted proteins contain domains. The absolute number was 464 and 448 for 1-frameshift and 2-frameshift, respectively.

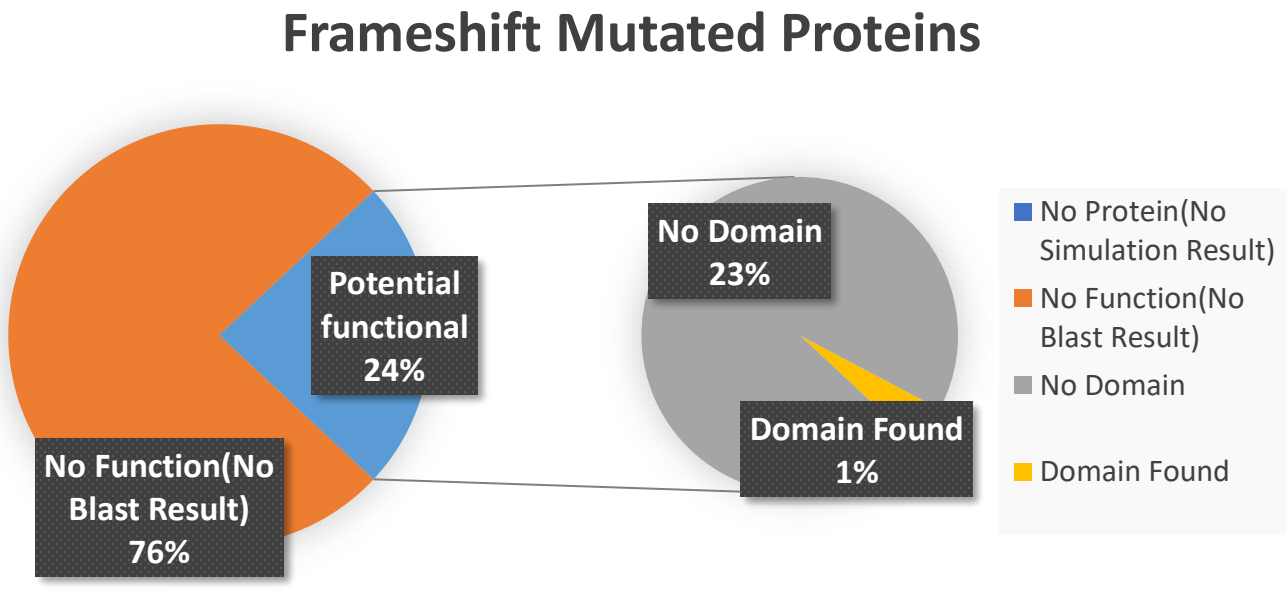

Figure 10. A summary of frameshift mutated proteins

3.4 Domain conformation

The structural protein domain architecture of the CDS, 1-frameshift, and 2-frameshift products are shown below in respective order.

Three types of peptides generated due to frameshift mutation. (1). Domain disappears and comes out of a plain peptide. As shown in Figure 11-Type1, domain CARD frameshifted to plain peptide; (2) The original domain will be the same domain or will become to a new domain. As shown in Figure 11-Type2-1, domain GVQW frameshifted to TOP2c which is itself; and Figure 12-Type2-2. (3) The plain protein sequence becomes to a domain. As shown in Figure 12-Type3, plain peptides frameshifted to GVQW domain reported by Pfam database.

In this research, we are trying to analyze option (2), and interpreting our findings which could be a great deal of evolution biology and oncology studies.

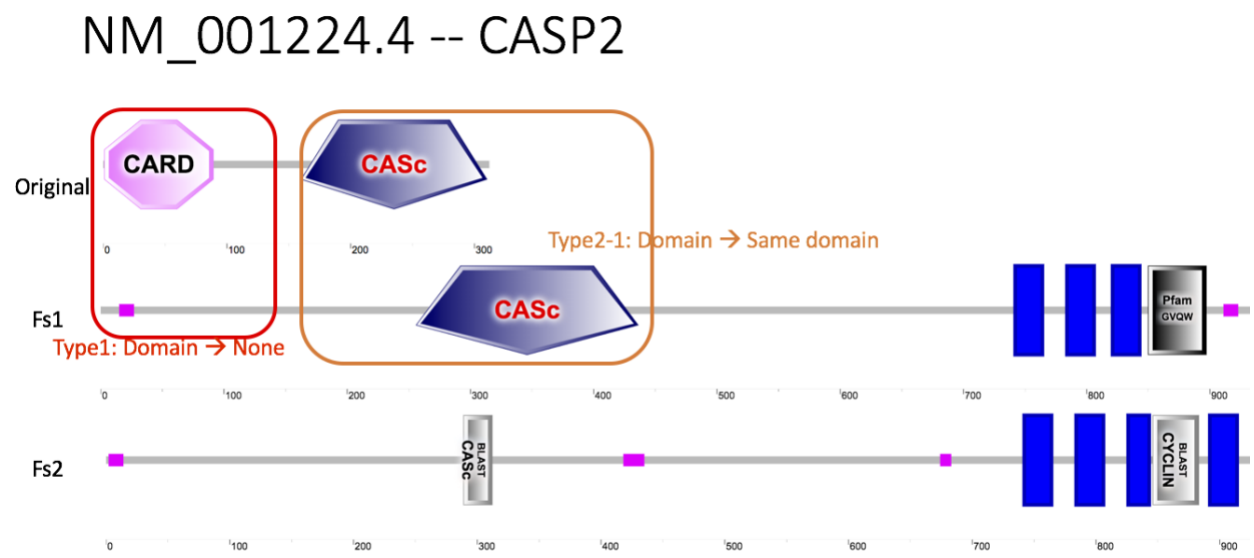

Figure 11. Domain conformation example NM_001224.4, the structural protein domain architecture of the original, 1-Frameshift, and 2-Frameshift products are shown above in order.

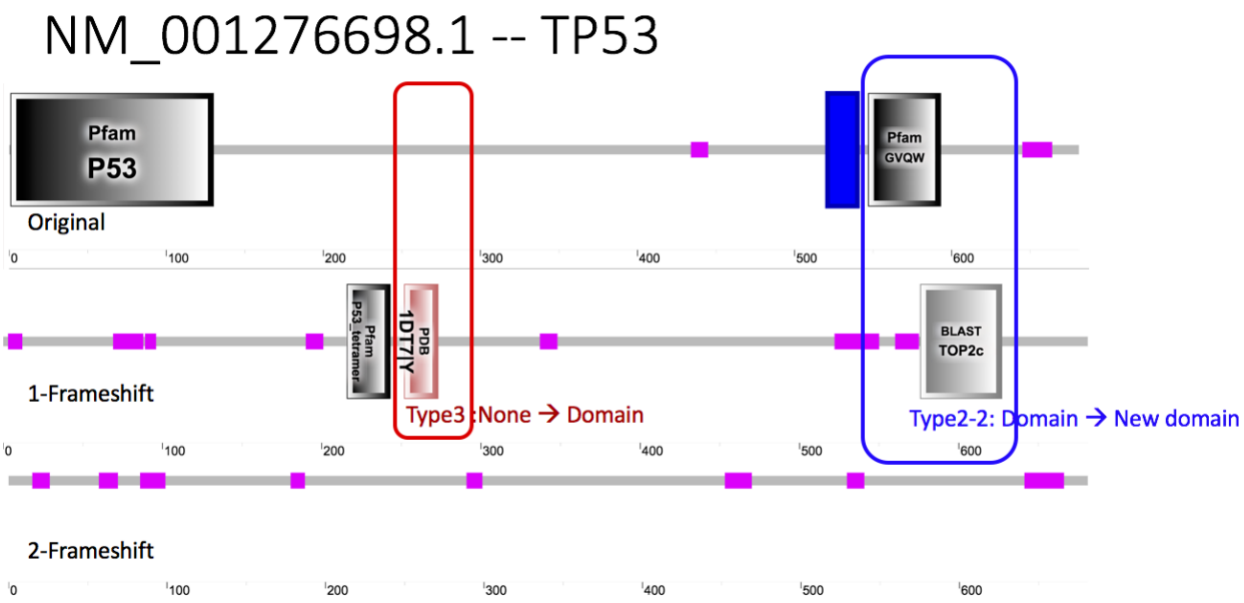

Figure 12.Domain conformation example NM_001276698.1

3.5 Evolution route map -- Domain level

Via frameshift mutation, original domain will become itself or a whole new domain detected by InterproScan. Here, we only consider the case that the original domain will frameshift to a new one. This is because it shows an evolution route map caused by frameshift mutation.

We found 124 of type2 relations, 66 of them are domain frameshift to same domain, and 58 of them are evolution relation between different domains. And they can build up a lot small networks. Most of the network are one to one relation, but there still have a small network that include 3 to 4 domains. We give three examples below (fig13).

The first route map in fig13 shows that domain Sacchrp_dh_NADP frameshift to Znf_C2H2_type domain, the strength between them is 1 which means there is only one case we found among our analysis. 112 represents the node size of Znf_C2H2_type, which means its appearance frequency in all the frameshift scenarios. This indicate that Znf_C2H2_type has a functional robust to frameshift mutations. The domain KRAB is another source to frameshift to Znf_C2H2_type with a strength of 3. Besides, domain Znf_C2H2_type is not only a target role in the route map, but also could be a source to domain Hlx-hairpin-Hlx_DNA-bd_motif with a strength 20, which means a Zinc finger domain frame-shifted to a helix-hairpin-helix domain 20 times. The full domain relation will report in appendix A. We only include frameshift mutation happens in a domain and resulting into a new domain.

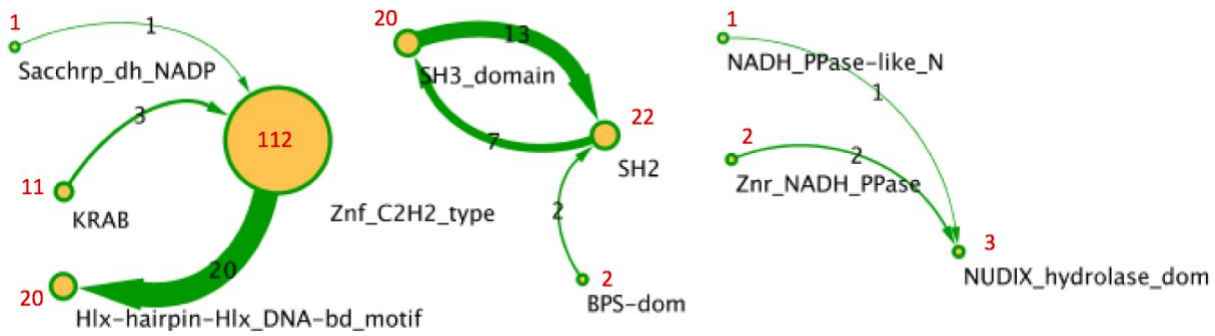

Figure 13. An example of evolution route map in domain level.

3.6 Sequence identity within the same named domains

Since multiple peptide will classified to a domain, it is not saying that the residue of these peptides is 100% identical. In order to see the similarity and identity within each node as described in the route map. We collect the domain sequences and alignment them in T-coffee. T-Coffee is a multiple sequence alignment program. The main characteristic of T-Coffee is that it will allow users to combine results obtained with several alignment methods [35].

Here we give an example of domain Znf_C2H2_type as a source and domain Hlx-hairpin-Hlx_DNA-bd_motif as a target. So, each of them investigate with five sequences. The result shows they are in good alignment (fig14). The hue with label "BAD AVG GOOD" represents their alignment situations. In red, it means the alignment is good, and vice versa. In our result, they all colored with red which means they are well alignment. The last line label with * : . or space indicate that the residue alignment is identical, vary similar, less similar or not relevant [36].

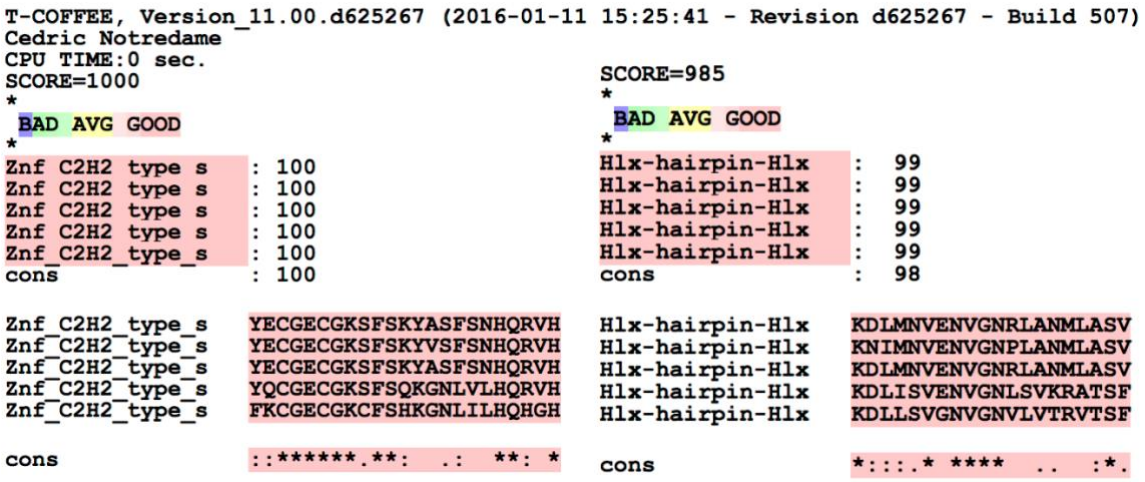

Figure 14. Sequence identity within the same name domains

3.7 Evolution route map -- Homologue superfamily level

A homologous superfamily is a group of proteins that share a common evolutionary origin, reflected by similarity in their structure. We classify the domains to its superfamily based on the position they located in the same peptides. Then the evolution route map in homologue superfamily level are built in Cytoscape (fig15). This practice allows to find a potential larger network within our data results.

The results show that most of the route still in a size of 2-3 nodes. The largest route map with node equals to 5 which is a slightly larger than the evolution route map in domain level. The node size represents domain quantities of that homologue superfamily.

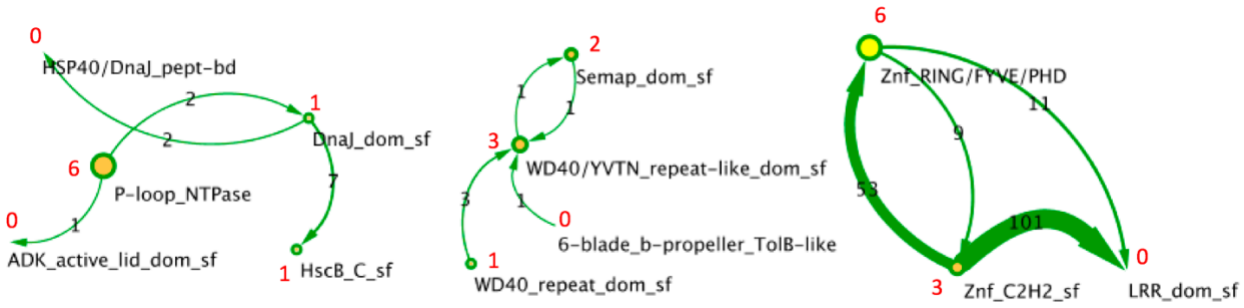

Figure 15. An example of evolution route map in homologue superfamily level.

The full homologue superfamily relation will report in appendix A. We only include frameshift mutation happens in a homologue superfamily and resulting into a new homologue superfamily.

3.8 Gene detection

For those sequences, which have frameshifted protein domain found, we filtered out them and extract their gene names, including gene name synonyms. We compare with the genes associated with frameshift disease reported by Clinvar. The result shows that 47 of our candidate genes have been reported and recorded (fig12).

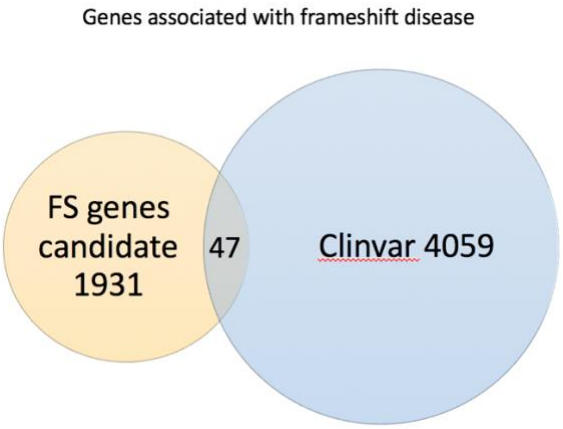

Figure 16. Candidate genes compare with known genes associated with frameshift disease

These 47 genes include AASS, MYCN, TK2, GLYCTK, GLB1, RNF135, RAD51, PAX2, RBCK1, MEOX1, FMR1, DENND5A, GRIN2A, GRM1, AK2, CDKN2A, PDE11A, IGLL1, TMEM237, TP53, EDA2R, DSC2, NGLY1, MYD88, GNRHR, RBMX, GHR, CDH3, BRCA1, RBBP8, FHL1, FAS, CASP8, NKX2-5, EMX2, POLH, CFC1, SON, KMT2B, GSS, LDHA, MEFV, G6PC, DICER1, ORC4, ARMC4, ARMC5.

3.9 Protein domain structural atlas of known genes associated with frameshift disease

Those 47 genes match 71 mRNA transcripts. These are the evidence that frameshift mutation can generate functional peptides. We will give BRCA1 as an example to illustrate their original sequence and frameshifted sequences architectures. Some of others will list right after.

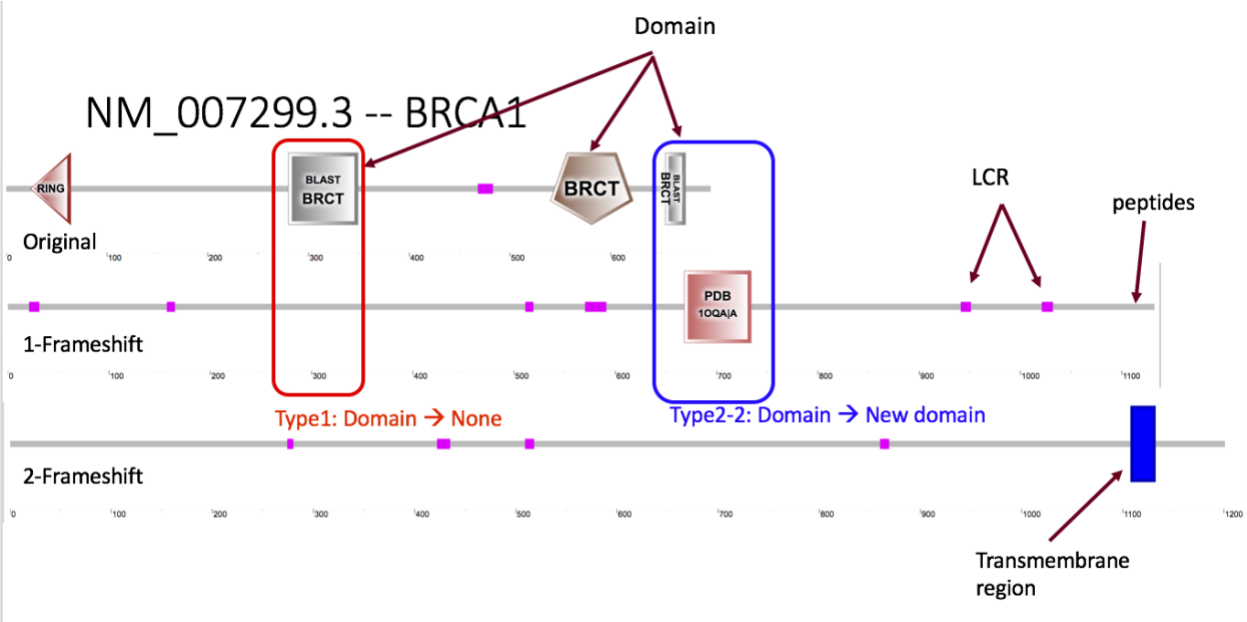

Figure 17 Protein domain architecture of NM_007299.3 and its frameshift products

The structural architectures of transcript NM_007299.3 of gene BRCA1 and its +1/+2 frameshift products are shown in figure 17. From above to bottom, they are original sequence, 1-frameshifted peptides and 2-frameshifted peptides respectively. The horizontal grey bar represents the peptides with a position annotation at the bottom. Polygons on the bar are domains found by SMART. The pink area is low complexity region and the blue rectangular is transmembrane region. In original sequence, there are four domains and one low complexity region. Domain RING plays a role in protein binding (GO:0005515), zinc ion binding (GO:0008270). Domain "blast BRCT" is a SMART BRCT domain. The domain was found using schnipsel database, and is an outlier domain. The BRCT domain is found predominantly in proteins involved in cell cycle checkpoint functions responsive to DNA damage [37], for example, as found in the breast cancer DNA-repair protein BRCA1. The domain is an approximately 100 amino acid tandem repeat, which appears to act as a phospho-protein binding domain [37]. Domain BRCT is a domain found in SMART database. The fourth one in the original sequence is also a "Blast BRCT" domain found in schnipsel database, reported by Smart. The pink region is the low complexity region (LCR). This region is abundant in the protein universe. LCR-containing proteins tend to have more binding partners across different PPI networks than proteins that have no LCRs. In 1-frameshift peptides, one domain and six LCRs were found. The domain annotated as "PDB 1OQA|A, meaning that this domain was found in PBD and its PBD id is 1OQA|A. In 2-frameshift peptide, one blue rectangular is a transmembrane helix region; and four pink regions are LCRs.

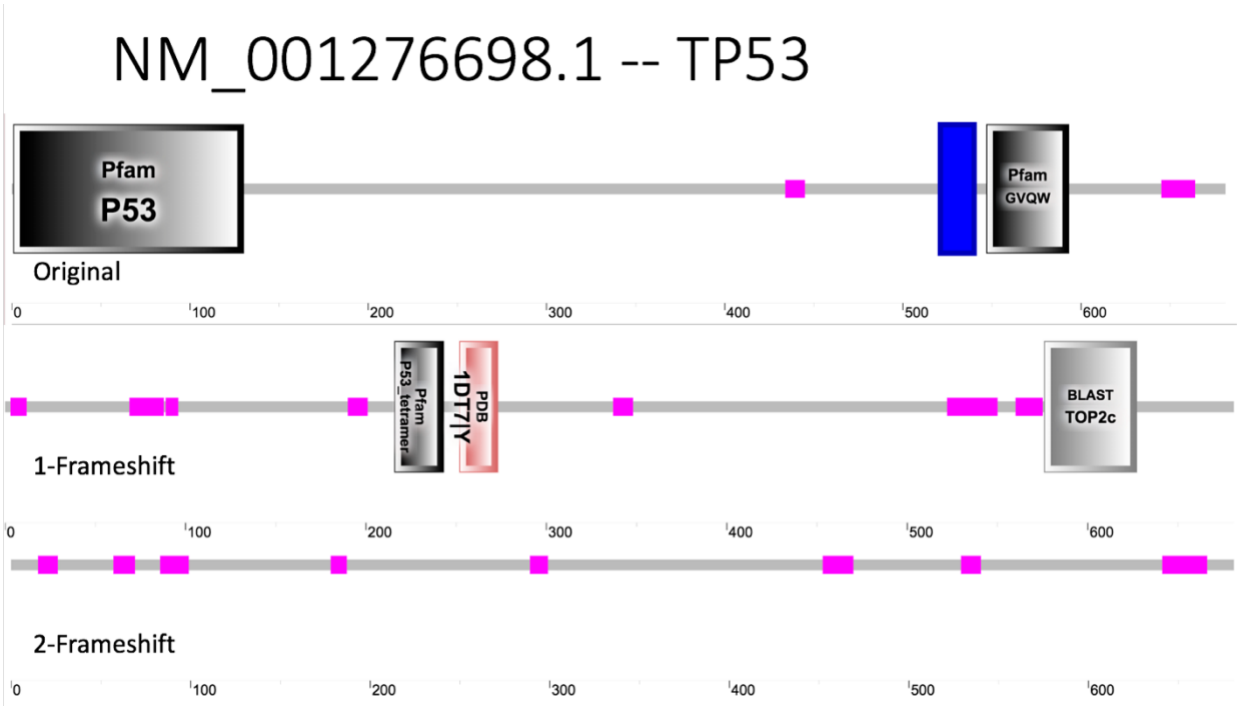

Figure 18. Protein domain architecture of NM_001276698.1 and its frameshift products

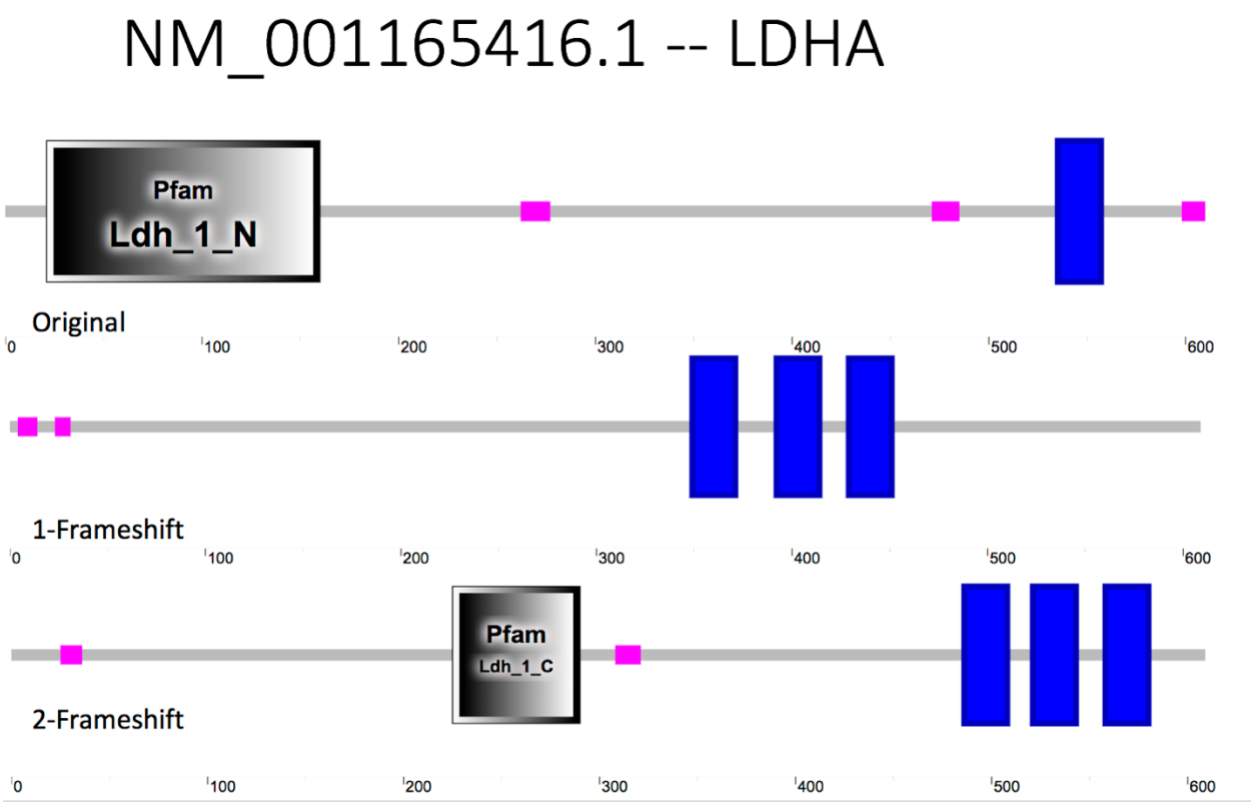

Figure 19. Protein domain architecture of NM_0011654146.1 and its frameshift products

4 Conclusion

The results of this simulation suggest that frameshift mutations can produce functional proteins that are the same or a whole new functional protein domain. Specifically, 464 and 448 polypeptides that are products of the 1-frameshift and 2-frameshift mutations, respectively, are found with InterproScan to carry protein domains. These are 1% of the whole transcripts that we studied, which is all mRNA transcripts in human genome. Besides, 23% of these transcripts are reported with a match through BLAST, even though they do not found protein domains according to Interproscan. This does not mean that there is no protein domain within these transcripts, because they might not be recorded in the database. So, these 24% are named with potential functional proteins.

At the domain level, 124 of domain frameshift to domain relations have been found, 66 of them are domain frameshift to the same functional domain, and 58 of them are relations between different domains.

At the homologue superfamily level, 83 relations were found. 43 of them are homologue superfamily frameshift to the same homologue superfamily, and 40 of them are relations between different homologue superfamily.

At the gene level, 863 genes are responsible for 464 and 448 polypeptides that are products of the 1-frameshift and 2-frameshift mutations. Each transcript has a gene annotation in the raw Genebank format data we collected. By considering synonymous names, the list of genes was increased to 1,931. Out of these list, we found 47 genes that have frameshift mutations in the Clinvar database.

5 Discussion

By simulation frameshift at the beginning of the open reading frame(ORF) in all human mRNAs and investigate the structure and function of these frameshift products, we are able to build a structural and functional atlas of frameshift mutation in human genome.

Investigation of the whole frameshift sequences without considering the introduced stop codons will help including all the possible frameshift mutation spot cases. The biopython packages are used to generate the frameshifted peptides, and the stop codon which they return as a * value was replaced by "X". This is because Interproscan will not allow match sequences including a *. Instead, the * replaced as "X" which means an unknown residue.

This design is the most important process in this research. If the frameshift mutation position is not in the beginning of the ORF, but in the middle of the ORF, the consequence of the frameshift mutation would be a conjunction of a former part of original peptide and a frameshift peptide. Through our results, we can combine the domain we found in order to study their structure and functions. This could be a good suggestion to find new drug target and help to design medicines.

Furthermore, this design is a good way to explain that frameshift mutation is a contributor to evolution of the protein domains. The direct position overlap between original peptides and frameshift peptides are considered as a linkage of evolution route. This is hard to study in DNA sequence level, because the universal genetic code is not one to one coding relationship. One amino acid can be code by multiple codons. Besides, domains can be represented by many different amino acid sequences.

We only studied the human mRNAs. This method could also be applied to other type of RNAs, such as ncRNA, to investigate the relationship between

regulatory RNAs. In reference transcripts, the RNA transcripts labeled as "XM" means predicted mRNA model. Our methods can also be applied to these sequences in order to build a larger structural and functional atlas in human genome. Moreover, our method can be used to study evolution in domain level in other species, such as cow, mouse, pig, rat, frog and zebrafish.

Despite such benefits of this design, we still recognize that there are limitations in it. Although we narrow down the frameshift cases into we would never see the conjunction of all the frameshift peptides that we produce, but we do narrow down the number of cases from 45,139 to 464 and 448 for 1-Frameshift and 2-Frameshift, respectively. Besides, we still need to check the frameshift domain residues whether they contain a "X" as a stop codon, to check whether the "X" can code amino acid. However, as a bioinformatics project which conduct in silico, we do reach our goal. That is to find possible biomarkers for biologist and save their time and funding. Hopefully we can provide valuable and interesting findings and trigger their thoughts.

References

[1] S.-A. R. E. B. &. W. P. C. S. Andreas Crameri, "DNA shuffling of a family of genes from diverse species accelerates directed evolution," *Nature,* 15 Jan 1998.

[2] H. Hu and R. A. Gatti, "New approaches to treatment of primary immunodeficiencies: fixing mutations with chemicals," *Current Opinion in Allergy and Clinical Immunology,* vol. 8, no. 6, pp. 540-546, 1 Dec 2008.

[3] Y. O. J. E. J. N. A. T. E. T. a. M. I. George Streisinger, Frameshift Mutations and the Genetic Code, Cold Spring Harb Symp Quant Biol, 1966.

[4] P. T. P. D. K. M. Nehrt NL, "Domain landscapes of somatic mutations in cancer," *BMC Genomics,* vol. 13, no. 9, p. Suppl 4, 18 Jun 2012.

[5] J. D.Watson, Molecular biology of the gene, 6th ed., San Francisco: Pearson/Benjamin Cummings, 2008.

[6] S. TR, "Singh TR (2013) Mitochondrial Genomes and Frameshift Mutations: Hidden Stop Codons, their Functional Consequences and Disease Associations.," *Journal of Clinical & Medical Genomics,* 8 July 2013.

[7] D. Q. Charles Darwin, On the Origin of Species, New York : Sterling , 2008.

[8] H. P. C. L. X. W. Y. W. G. C. J. Z. Xiaolong Wang, "Premature termination codons signaled targeted repair of frameshift mutation by nonsense-mediated gene editing.," 5 April 2017. [Online]. Available: https://doi.org/10.1101/069971.

[9] L. F. V. G. P. M. Blanpain C, "Mechanism of transdominant inhibition of CCR5-mediated HIV-1 infection by ccr5delta32.," *JBC,* vol. 272, no. 49, pp. 30603-6, Jan 1998.

[10] F. L. G. V. &. M. P. Cédric Blanpain, "CCR5 and HIV infection," *Receptors and Channels,* vol. 8, no. 1, pp. 19-13, 2011.

[11] K. T. G. B. I. S. O. S. D. M. Carrington M, "Novel Alleles of the Chemokine-Receptor Gene CCR5," *AJHC,* vol. 61, no. 6, pp. 1261-1267, Dec 1997.

[12] B. L. M. T. B. P. A. B. F. L. M. S. V. W. G. V. R. W. D. M. P. Cédric Blanpain, "Multiple nonfunctional alleles of CCR5 are frequent in various human populations.," *Blood* , vol. 96, no. 5, pp. 1638-1645, 1 Sep 2000.

[13] K. H. S. M. T. P. N. H. a. M. V. M. Marmor, "Resistance to HIV Infection," *J Urban Health,* vol. 83, no. 1, pp. 5-17, Jan 2006.

[14] P. A. B.-W. A. A. G. S. T. K. J. C. C. …. M. P. M. Zimmerman, "Inherited resistance to HIV-1 conferred by an inactivating mutation in CC chemokine receptor 5: studies in populations with contrasting clinical phenotypes, defined racial background, and quantified risk.," *Molecular medicine,* vol. 3, no. 1, pp. 23-36, Jan 1997.

[15] J. R. Lupski, "Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits," *Trend in genetics,* vol. 14, no. 10, pp. 4147-422, 1 Oct 1998.

[16] S. &. B. W. Clancy, "Translation: DNA to mRNA to Protein," *Nature,* 2008.

[17] D. J. M. S. D. C. Maehigashi T, "Structural insights into +1 frameshifting promoted by expanded or modification-deficient anticodon stem loops.," *PANS,* vol. 111, no. 35, pp.

12740-12745, 2 Sep 2014.

[18] †. E. T. G. S. J. E. M. I. a. A. T. Y. Okada, "A FRAME-SHIFT MUTATION INVOLVING THE ADDITION OF TWO BASE PAIRS IN THE LYSOZYME GENE OF PHAGE T4," *PANS,* vol. 56, no. 6, p. 1692–1698, Dec 1966.

[19] B. D. I. N. N. D. C. F. R. R. B. H. M. T. K. R. D. R. A. J. B. S. B. T. K. B. H. S. N. G. C. J. Ogura Y1, "A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease.," *Nature,* vol. 411, no. 6837, pp. 603-6, 31 May 2001.

[20] S. R. C. F. H. C. H. N. S. T. B. L. D. M. W. M. G. B. e. a. Iannuzzi MC1, "Two frameshift mutations in the cystic fibrosis gene.," *AJHG,* vol. 48, no. 2, p. 227–231, Feb 1991.

[21] M. M. H. L. a. E. F. Neufeld, "A frameshift mutation in a patient with Tay-Sachs disease causes premature termination and defective intracellular transport of the alpha-subunit of beta-hexosaminidase.," *THE JOURNAL OF BIOLOGICAL CHEMISTRY* , vol. 264, no. 35, pp. 21376-21380, 15 Dec 1989.

[22] N. R. Coordinators, "Database resources of the National Center for Biotechnology Information," *Nucleic Acids Research,* vol. 44, no. D1, p. D7–D19, 4 Jan 2015.

[23] T. A. J. T. C. B. A. C. C. J. C. A. D. I. F. T. H. F. K. B. W. a. M. J. L. d. H. Peter J. A. Cock, "Biopython: freely available Python tools for computational molecular biology and bioinformatics," *Bioinformatics,* vol. 25, no. 11, p. 1422–1423., 1 Jun 2009.

[24] G. W. States DJ, "Combined use of sequence similarity and codon bias for coding region identification.," *J Comput Biol.,* vol. 1, no. 1, pp. 39-50, 1994.

[25] T. L. M. A. A. S. J. Z. Z. Z. W. M. D. J. L. Stephen F. Altschul, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.," *Nucleic Acids Research,* vol. 25, no. 17, p. 3389–3402, 1 Sep 1997.

[26] D. B. H.-Y. C. M. F. W. L. C. M. H. M. J. M. A. M. G. N. S. P. A. F. Q. A. S.-V. M. S. S.-Y. Y. R. L. a. S. H. Philip Jones, "InterProScan 5: genome-scale protein function classification," *Bioinformatics,* vol. 30, no. 9, p. 1236–1240, 1 May 2014.

[27] B. E. G. Fernando Perez, "IPython: A System for Interactive Scientific Computing," *Computing in Science & Engineering,* vol. 9, no. 3, May-Jun 2007.

[28] M. A. Sangrador-Vegas A, "Protein classification: An introduction to EMBL-EBI resources," [Online]. Available: http://europepmc.org/abstract/CTX/C7836.

[29] K. K. H. R. C. C. Gough J, "Assignment of homology to genome sequences using a library of Hidden Markov Models that represent all proteins of known structure," *Journal of molecular biology,* vol. 313, no. 4, pp. 903-919, 2 Nov 2001.

[30] R. A. T. K. A. A. B. A. B. D. B. P. B. V. B. L. C. R. C. E. C. U. D. L. D. M. D. R. F. W. F. J. G. D. H. N. H. S. H. D. K. A. K. A. K. A. L. P. S. L.-G. D. L. R. L. I. L. M. M. J. M. C. M. J. M. J. M. A. M. A. N. N. S. O. C. O. R. P. J. D. S. C. J. A. S. P. D. T. F. V. D. W. C. H. W. a. C. Y. Nicola J. Mulder, "New developments in the InterPro database," *Nucleic Acids Research,* vol. 35, no. D224–D228, Jan 2007.

[31] G. v. Rossum, "Python tutorial," Technical Report CS-R9526,, Amsterdam, 1995.

[32] J. M. L. G. R. R. W. J. W. S. R. D. M. C. a. D. R. M. Melissa J. Landrum, "ClinVar: public archive of relationships among sequence variation and human phenotype," *Nucleic Acids Research,* vol. 42, no. D980–D985, 1 Jan 2014.

[33] M. F. B. P. P. C. Schultz J, "SMART, a simple modular architecture research tool:

Identification of signaling domains," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 95, no. 11, pp. 5857-5864, 1998.

[34] "SMART: Batch access," [Online]. Available: http://smart.embl-heidelberg.de/smart/batch.pl.

[35] H. Notredame, "T-Coffee: A novel method for multiple sequence alignments.," *JMB,* vol. 302, pp. 205-217, 2000.

[36] S. M. I. X. M. O. A. M. J.-M. C. J.-F. T. a. C. N. Paolo Di Tommaso, "T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension," *Bioinformatics,* no. W13–W17, 1 Jul 2011.

[37] K. H. P. B. A. F. N. S. F. A. E. V. K. P. Bork, "A superfamily of conserved domains in DNA damage-responsive cell cycle checkpoint proteins," *FASEB J.,* no. 1, pp. 68-76, 11 Jan 1997.

Appendix A：Domain frameshift to a new domain

Source	mutation	target	strength
Znf_C2H2_type	fs	Hlx-hairpin-Hlx_DNA-bd_motif	20
SH3_domain	fs	SH2	13
FN3_dom	fs	Cullin_homology	11
F-box_dom	fs	F-box-assoc_dom	9
SH2	fs	SH3_domain	7
Prot_kinase_dom	fs	AGC-kinase_C	6
Death_domain	fs	TIR_dom	5
TLC-dom	fs	RWD-domain	5
GAF	fs	Cyt_c-like_dom	4
Small_GTP-bd_dom	fs	DnaJ_domain	4
DML1/Misato_tubulin	fs	SRCR	3
DUF4749	fs	Znf_LIM	3
DnaJ_domain	fs	DnaJ_C	3
EGF-like_dom	fs	Trypsin_dom	3
Importin-beta_N	fs	4Fe4S_Fe-S-bd	3
KRAB	fs	Znf_C2H2_type	3
RRM_dom	fs	RBM1CTR	3
VWF_A	fs	Sushi_SCR_CCP_dom	3
Apple	fs	Trypsin_dom	2
BPS-dom	fs	SH2	2
Dynamin_central	fs	GED	2
LisH	fs	CRA_dom	2
LisH	fs	CTLH/CRA	2
Neur_chan_lig-bd	fs	Neurotrans-gated_channel_TM	2
PAS	fs	bHLH_dom	2
PH_domain	fs	IRS_PTB	2
Pan_app	fs	Trypsin_dom	2
Pept_C14_p20	fs	Pept_C14A	2
RA_dom	fs	SARAH_dom	2
Znr_NADH_PPase	fs	NUDIX_hydrolase_dom	2
Acoase/IPM_deHydtase_lsu_aba	fs	Ricin_B_lectin	1
Acyl_transferase	fs	PKS_acyl_transferase	1
DNA_recomb/repair_Rad51_C	fs	RecA_monomer-monomer_interface	1
Dynamin_central	fs	GED_dom	1
EGF-like_Ca-bd_dom	fs	Trypsin_dom	1

FN3_dom	fs	Interferon/interleukin_rcp_dom	1
FXR_C1	fs	FXR_C3	1
Flavoprot_Pyr_Nucl_cyt_Rdtase	fs	OxRdtase_FAD/NAD-bd	1
Ig_sub	fs	Ig-like_dom	1
Ig_sub2	fs	Ig-like_dom	1
Integrin_bsu_VWA	fs	Integrin_beta_N	1
Integrin_bsu_VWA	fs	PSI	1
Interferon_reg_fact_DNA-bd_dom	fs	Interferon_reg_factor-3	1
Interferon_reg_factor-3	fs	Interferon_reg_fact_DNA-bd_dom	1
Kinase_OSR1/WNK_CCT	fs	Prot_kinase_dom	1
MyD88_Death	fs	TIR_dom	1
NADH_PPase-like_N	fs	NUDIX_hydrolase_dom	1
Neurotrans-gated_channel_TM	fs	Neur_chan_lig-bd	1
PSI	fs	Semap_dom	1
Pept_C14A	fs	Pept_C14_p10	1
Prot_kinase_dom	fs	Ser-Thr/Tyr_kinase_cat_dom	1
Rad51_DMC1_RadA	fs	DNA_recomb/repair_Rad51_C	1
Rad51_DMC1_RadA	fs	RecA_monomer-monomer_interface	1
RecA_ATP-bd	fs	DNA_recomb/repair_Rad51_C	1
RecA_ATP-bd	fs	RecA_monomer-monomer_interface	1
Sacchrp_dh_NADP	fs	Znf_C2H2_type	1
UmuC	fs	Ricin_B_lectin	1
Unchr_dom_Cys-rich	fs	TIL_dom	1

Appendix B：Homologue superfamily frameshift to a new Homologue superfamily

Source	mutation	target	strength
Znf_C2H2_sf	fs	LRR_dom_sf	101
Znf_C2H2_sf	fs	Znf_RING/FYVE/PHD	53
Znf_RING/FYVE/PHD	fs	LRR_dom_sf	11
Znf_RING/FYVE/PHD	fs	Znf_C2H2_sf	9
Neur_chan_lig-bd_sf	fs	Neuro-gated_channel_TM_sf	8
DnaJ_dom_sf	fs	HscB_C_sf	7
Neuro-gated_channel_TM_sf	fs	Neur_chan_lig-bd_sf	6
GAF-like_dom_sf	fs	Cyt_c-like_dom_sf	4
SH3-like_dom_sf	fs	SH2_dom_sf	4
CTDL_fold	fs	C-type_lectin-like/link_sf	3
Gln_synt_N	fs	Gln_synth/guanido_kin_cat_dom	3
WD40_repeat_dom_sf	fs	WD40/YVTN_repeat-like_dom_sf	3
Znf_CCCH_sf	fs	Znf_CCHC_sf	3
C-type_lectin-like/link_sf	fs	CTDL_fold	2
DEATH-like_dom_sf	fs	Toll_tir_struct_dom_sf	2
DnaJ_dom_sf	fs	HSP40/DnaJ_pept-bd	2
Elafin-like_sf	fs	Kunitz_BPTI_sf	2
Nucleotide-diphossugar_trans	fs	Ricin_B-like_lectins	2
P-loop_NTPase	fs	DnaJ_dom_sf	2
6-blade_b-propeller_TolB-like	fs	WD40/YVTN_repeat-like_dom_sf	1
Arg_repress-like_C	fs	RRF_sf	1
DEATH-like_dom_sf	fs	Caspase-like_dom_sf	1
F-box-like_dom_sf	fs	Galactose-bd-like_sf	1
FN3_sf	fs	Ig-like_fold	1
Fibrinogen-like_C	fs	Fibrinogen_a/b/g_C_2	1
Fibrinogen_a/b/g_C_1	fs	Fibrinogen-like_C	1
Fibrinogen_a/b/g_C_1	fs	Fibrinogen_a/b/g_C_2	1
NA-bd_OB-fold	fs	KH_dom_type_1_sf	1
P-loop_NTPase	fs	ADK_active_lid_dom_sf	1
RGS_subdom1	fs	RGS_sf	1
Rib_L2_dom2	fs	Translation_prot_SH3-like_sf	1
SH2_dom_sf	fs	SH3-like_dom_sf	1
SMAD-like_dom_sf	fs	SMAD_FHA_dom_sf	1
SMAD_FHA_dom_sf	fs	SMAD-like_dom_sf	1

Semap_dom_sf	fs	WD40/YVTN_repeat-like_dom_sf	1
Transglutaminase_C_sf	fs	Ig-like_fold	1
VHL_beta_dom_sf	fs	VHL_sf	1
VHL_sf	fs	VHL_alpha_dom_sf	1
WD40/YVTN_repeat-like_dom_sf	fs	Semap_dom_sf	1
vWFA_dom_sf	fs	Sushi/SCR/CCP_sf	1