# Analysis of Mnx1 protein, its homologous sequences, phylogenetic tree and mutations

Authors:
G. Görkem Köse - 25359
Emir Can Ülkü - 26740
Ahmet Ölçüm - 25915
Sayed Damon Sadraije Najafi - 26260

## Introduction

Proteins are molecules with certain shapes and sizes with a task, these tasks may range from building other proteins to helping with other functions in living things. Mutant proteins may have altered functions, this may cause them to not bind to certain sites or bind too good to some sites. The focus of the project is the Mnx1 protein whose mutations may cause the Currarino Syndrome, which is a disease that is autosomal dominant (Merello, 2014), this is a peculiar situation as autosomal dominant illnesses cannot be carried and only be expressed in the phenotype, sick organisms with such diseases usually do not survive therefore the sickness will not be transported to the next generation. The gene coding Mnx1 protein is called mnx1, which was previously called the hlxb9 gene (Merello, 2014). The protein is a homeobox protein which is a developmental factor that is responsible for pancreatic cells' differentiation, neuronal differentiation and probably many more unknown differentiations. The mutant protein may cause serious anorectal malformations or may not have any symptoms present. Mnx1 is differentially expressed, which means the proteins expressed under certain conditions are met, this causes the protein to be expressed at the area that is supposed to differentiate to certain cells. The project has the aim to uncover certain aspects of the protein with a computational approach, this is done by checking conserved regions of similar protein codes from different species and cross checking the single point substitution mutations found in clinical papers and databases.

## Materials and Methods

The project has started with searching for a rare genetic disease which is inherited in a Mendelian type fashion. Currarino Triad Syndrome (CS) was chosen among a wide variety of diseases, due to its autosomal *dominant* inheritance. The mutations that cause Currarino Syndrome occur on Mnx1 protein, a homeobox protein that is responsible for the morphogenesis during embryonic development. Therefore, the sequence for Mnx1 protein was retrieved from the UniProt database. The sequence was then Blasted against 100, 250, 500, 1000, and 5000 sequences using BlastP in order to get its homologous sequences.

However, the analysis for 100 and 250-sequence files were decided to be left out since there was only a single Homo sapiens sequence and there was no other isoform or a different homologous protein resulting from a gene duplication event. Then, using other sequence files, the homologous sequences were aligned by the MEGA11, Muscle algorithm. However, the alignment for 5000 sequences could not be completed due to its complexity and the time shortage. Hence, the analysis was performed using the sequence files of 500 and 1000 sequences, which contains two different isoform sequences for the Mnx1 protein. The

resulting alignments for 500 and 1000 sequences were employed to construct the phylogenetic tree using the Neighbor-Joining method. Then, midpoint rerooting was applied to the two phylogenetic trees of 500 and 1000 sequences in FigTree. But in the final analysis, only the sequence file of 500 sequences is employed since in the tree of 1000 sequences, it is observed that the clade that contains a single Homo sapiens sequence of isoform 1, consists of 974 sequences, therefore leaving the other clade out does not create a drastic difference.

The aligned sequences were used to calculate the conservation scores for each amino acid site in the sequence. For each position, the amino acid which has the highest number of occurrences among all aligned sequences was retrieved in order to obtain the consensus sequence. The frequency of the amino acids in the consensus sequence is computed, which would be treated as the conservation scores for the respective position.

```
filename = "../aligned/fas/aligned-500.fas"
#filename = "./aligned/fas/aligned-pruned-500.fas"
#filename = "./aligned/fas/realigned-pruned-500.fas"
seqDict = fastareader(filename)
sequences = list(seqDict.values())
aaList = ["A", "R", "N" , "D" , "C" , "Q" , "E" , "G" , "H" , "I" , "L" , "K" , "M" , "F" , "P" , "S" , "T" , "W" ,"Y" ,"V"]

consensus = ""
consensus_aminoacid_score = {}
for pos in range(len(sequences[0])):
    aa_percent_dict = dict.fromkeys(aaList, 0)
    aminoacids_onsamepos = ""
    for seq in sequences:
        aminoacids_onsamepos += seq[pos]
    for aa in aaList:
        count_aa = aminoacids_onsamepos.count(aa)
        aa_percent_dict[aa] = count_aa / len(sequences)
    aa_percent_dict = dict(sorted(aa_percent_dict.items(), key=lambda x: x[1], reverse=True))
    consensus_aa = list(aa_percent_dict.keys())[0]
    max_score = list(aa_percent_dict.values())[0]
    consensus += consensus_aa
    consensus_aminoacid_score[pos] = {consensus_aa:max_score}
```

Figure 1: Code snippet to calculate the conservation scores and consensus sequence

Then, the phylogenetic tree is pruned to construct the subset tree which contains only the clade of a single Homo sapiens sequence and the clade that contains the original protein sequence is taken into consideration for the next step. The nearest common ancestor of both Homo sapiens sequences are retrieved and the clade that contains the original sequence was selected. Another conservation score calculation was performed for only the members of the clade, and for further analysis upon the conservations, a specific alignment session was conducted for the members from scratch.

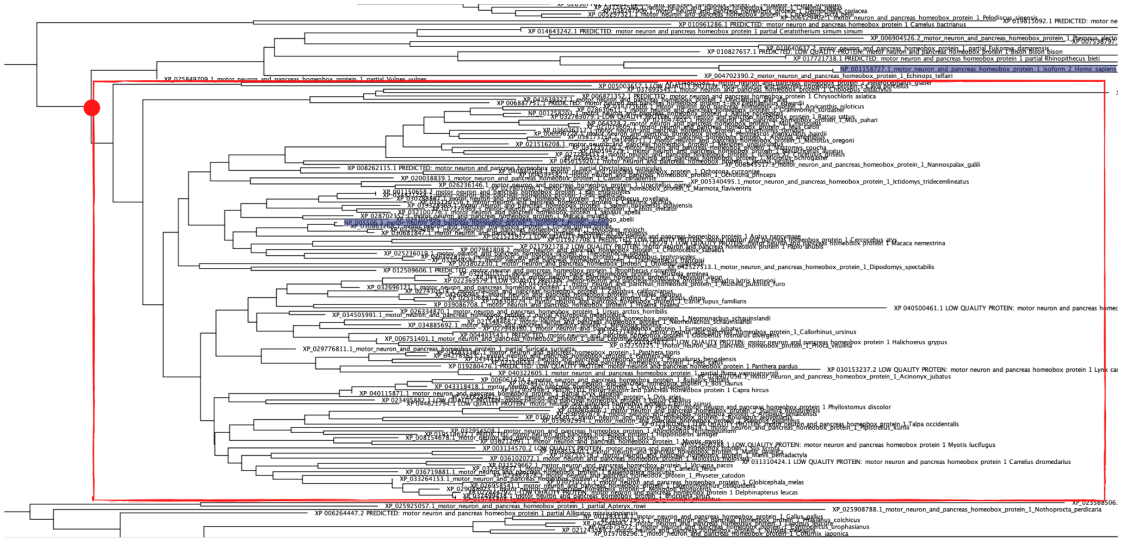

Figure 2: Red point indicates the common ancestor of two Homo sapiens sequence and the frame contains the selected subtree.

```
from ete3 import Tree

t = Tree("../trees/500_midpoint_rooted.nwk", format=1, quoted_node_names=True)

homo_sapiens = ['NP_001158727.1_motor_neuron_and_pancreas_homeobox_protein_1_isoform_2_Homo_sapiens', \
                'NP_005506.3_motor_neuron_and_pancreas_homeobox_protein_1_isoform_1_Homo_sapiens']

nodes_to_keep = t.get_common_ancestor(homo_sapiens).get_leaves()
t.prune(nodes_to_keep)
single_hs = t.get_children()[1].get_leaves()
other_hs = t.get_children()[0].get_leaves()
if len(other_hs) > len(single_hs):
    single_hs = other_hs
t.prune(single_hs)

t.write(format=1, outfile="pruned_500.nwk")

inFile = open("../trees/clade_sequences.txt", "W")
for sequence in list(single_hs):
    inFile.write(sequence.name + "\n")

inFile.close()
```

Figure 3: Code snippet to retrieve the subtree that is indicated inside the frame in figure 9.

Next, the single-point mutations are retrieved from three different sources. First source is the clinical papers about Currarino disease which identifies novel mutations on Mnx1 protein that are definitely pathogenic. Second source is ClinVar, which is a publicly accessible database that reports the relationships among variations and the phenotypes. The last source is gnomAD, a database of reporting structural variants for populations.
For the mutations whose clinical significance was known, the three conservation scores - one that is builded by all set of aligned sequences, second that is builded by only the aligned sequences on the respective clade, and third that is builded by the realignment session for the sequences on the respective clade - on the mutation occurring site was reported by our

project. For the mutations whose clinical significance are not known, the classification is performed using all three methods for comparison purposes. The findings are reported in the Results section.

Lastly, in order to see if there is a relationship between the mutation type - neutral or pathogenic- with the frequency of the mutation among the population; an independent t-test has been performed. For each classification that results from different conservation score computation approaches, the mutations are separated into two groups: pathogenic and non-pathogenic. Then, it is examined whether there is a significant difference between these two samples in terms of the allele frequency. The findings are presented in the Results section as well.

## Results

In this project, a homeobox protein, Mnx1, is examined with respect to its homologous sequences, conserved areas, and mutations. As it is further discussed in the Materials and Methods section, the protein sequence is used to retrieve the homologous sequences using BlastP and the file that contains 500 homologous sequences is chosen in order to perform the analysis steps.

The homologous sequences are aligned using the MEGA11 tool, Muscle algorithm. Below, you can see a part of the alignment, the whole alignment file is uploaded to Github.

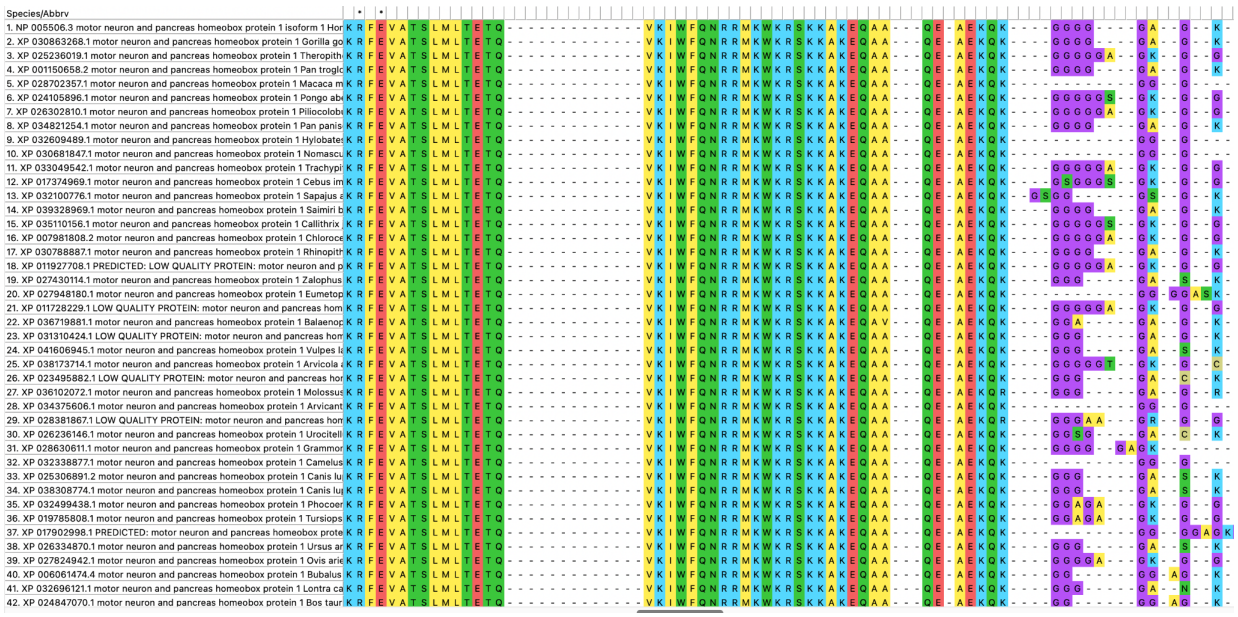


Figure 4: Part of the alignment of 500 homologous sequences

For each position for the aligned sequences, a conservation score calculation is carried out by finding the most frequent amino acid for that position and its frequency. Then, using the protein sequence alignments, a phylogenetic tree is constructed using the Neighbor-joining method. In the phylogenetic tree, there are two different sequences that belong to Homo sapiens. In order to focus the analysis on the original sequence, the clade that contains the original protein sequence is selected. The sequences that are at the leaves of the resulting pruned tree are subsetted and the conservation score computation is conducted using the aligned sequences of the pruned tree. For further analysis, a specific alignment session is performed using MEGA11 for the sequences that are part of our focus clade. Then, the conservation scores for the new specific alignment are calculated. Below, you can examine the overall value distribution for the conservation scores that are employing three different alignment outputs.

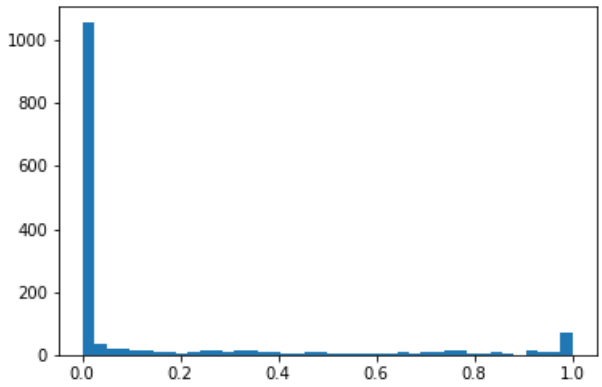


Figure 5: Distribution of conservation scores for the alignment of 500 sequences

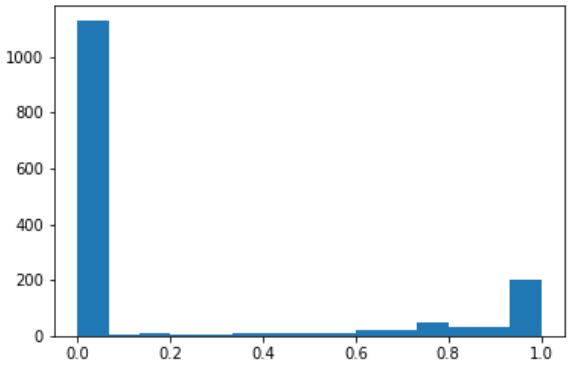


Figure 6: Distribution of conservation scores for the already aligned sequences that are part of the focus clade

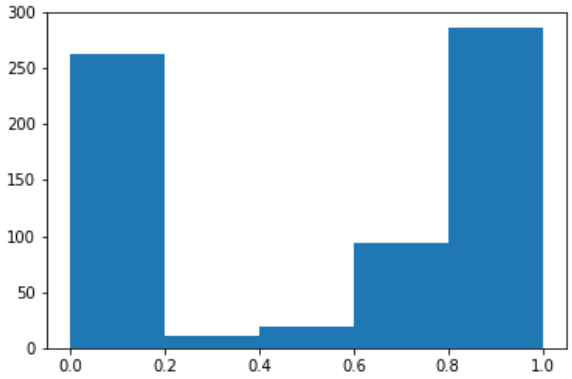


Figure 7: Distribution of conservation scores for the specific alignment of the sequences that are part of the focus clade

As the sequences become closer to the original sequence in the phylogenetic three, the conservation scores increase drastically. For the second strategy, it can be observed that the number of high conservation scores are increased while the number of low conservation scores remain the same. The reason behind that the conservation scores are calculated using the original and prior alignment results, therefore there may be gap insertions at some points that are not necessary to align the present sequences but it was necessary to align all 500 sequences before the examination of the phylogenetic tree. Yet, if one prefers a new alignment session only for the sequences in the focus clade, then the number of high conservation scores increases as the number of low conservation scores decreases. In order to observe the most conserved areas among the sequence, another plot series is generated. Below, you can see the plot for conservation scores by positions for the three different alignment strategies.

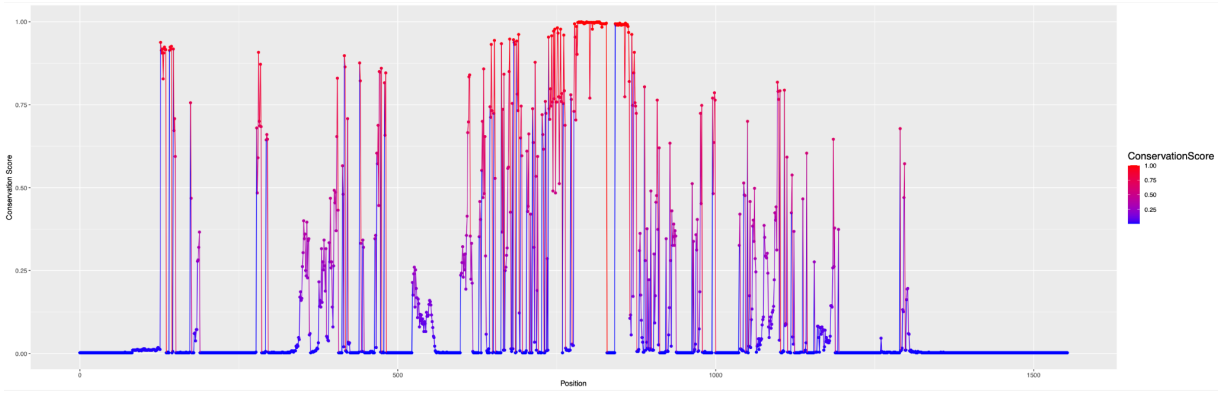


Figure 8: Conservation scores for the positions of the 500 aligned protein sequences

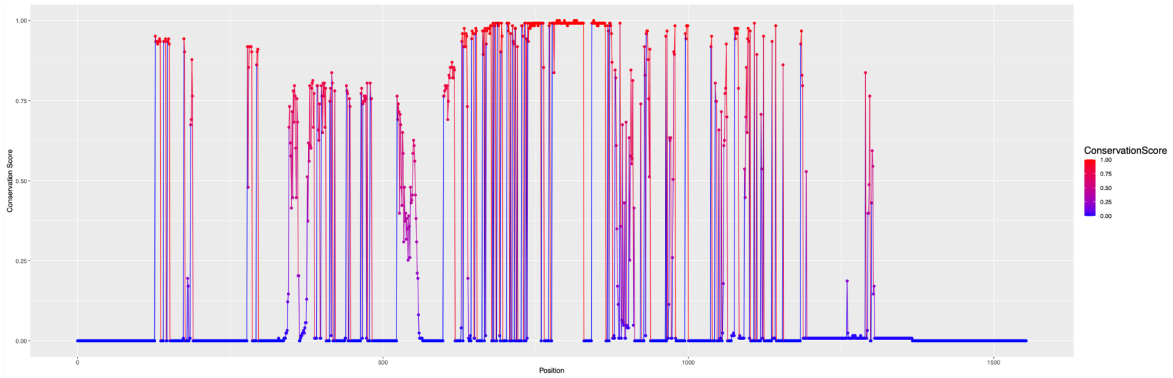


Figure 9: Conservation scores for the positions of the aligned sequences that are part of the focus clade

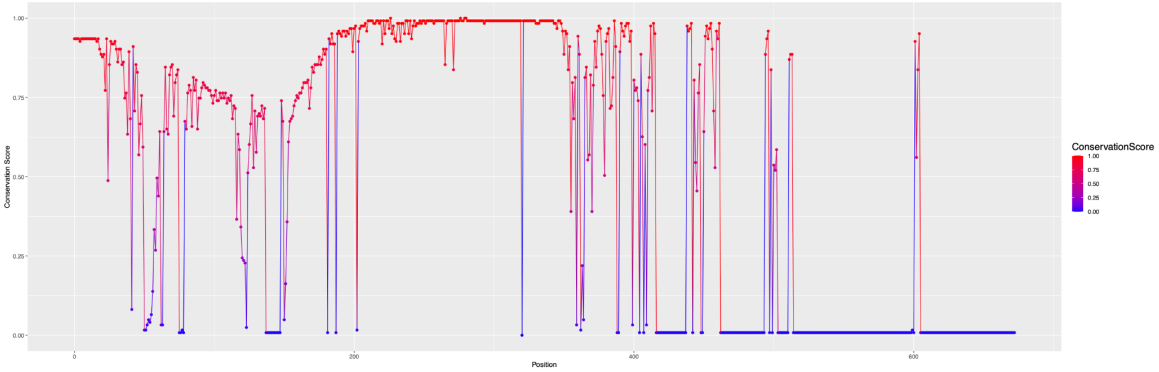


Figure 10: Conservation scores for the positions of the *realigned* sequences that are part of the focus clade

The main reason behind developing three different conservation score strategies is to compare how well they perform predicting the clinical significance of mutations. In order to classify the mutations as pathogenic or neutral, one has to have a threshold for the conservation scores. Yet, the ClinVar database only reports 6 missense mutations, however, there are no mutations that have clinical significance. Therefore, in order to develop a threshold, the clinical papers which identify novel mutations on Mnx1 protein that cause Currarino syndrome are examined and such mutations are collected. Since the number of such mutations is 17, a complex prediction cannot be carried out by fitting a Machine Learning model. Therefore, for each pathogenic mutation, the conservation scores for the mutation occurring sites are gathered in a list for each of the conservation score strategies. This means that, for the pathogenic mutation set; there are three different conservation score lists. The classification threshold is set to be the mean of the respective conservation score list for each of the classification sessions that would be carried out using a different conservation score calculation strategy. The threshold is set to 0.680875 for the first conservation score set, that is calculated using all of the 500 aligned homologous sequences; 0.894817 for the second conservation score set, that is calculated using the already aligned

sequences of the focus clade; and 0.912601 for the last conservation score set, that is calculated using the new alignments for the sequences of the focus clade. This observation shows that for a general alignment which has a lot of gap insertions; if a mutation position, by chance, coincides with a position that has a lot of gaps, the score becomes low. This is why the mean of the conservation scores of the pathogenic mutation occurring sites are low in the most general alignment of 500 sequences as compared to the most specific alignment.

Using gnomAD, all missense mutations are retrieved. For each such single-point mutation, the conservation score is calculated on the occurring site of the mutation. If the conservation score is greater than or equal to the threshold value, which is the mean of the conservation scores at the pathogenic mutation occurring sites; the corresponding mutation is classified as pathogenic. This process is repeated for the three different conservation score strategies.

In total, there are 173 mutations to be classified as neutral or pathogenic. 77 of the mutations are identified as pathogenic by the first conservation score set, 104 of the mutations are identified as pathogenic by the second conservation score set, and 102 of the mutations are identified as pathogenic by the last conservation score set.

Furthermore, in order to investigate a relationship between the frequency of the mutation among the population reported in gnomAD and the type of mutation (pathogenic or neutral), independent t-test experiments are conducted. For each conservation score calculation approach, the classified mutations are separated into two distinct samples with respect to the mutation type; and it is measured whether there is a significant statistical difference between the frequency of the allele or not. The p-values, indicators of the significance, are all greater than 0.05 with values 0.18, 0.08, 0.35; for the first, second, and third conservation score sets respectively. Below, you can see the scatter-plots for conservation scores and allele frequencies in order to determine whether a negative correlation exists or not.

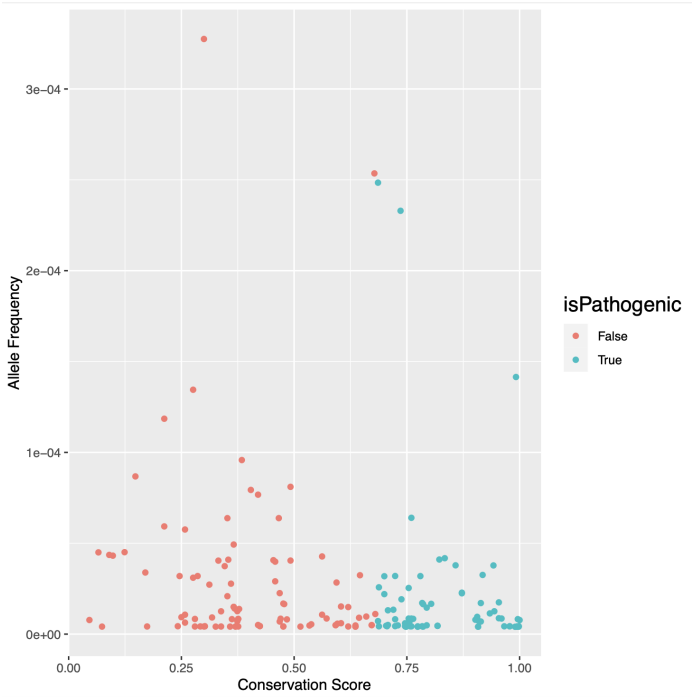


Figure 11: Conservation Score and Allele Frequency scatter plot of mutations classified using 500 aligned protein sequences

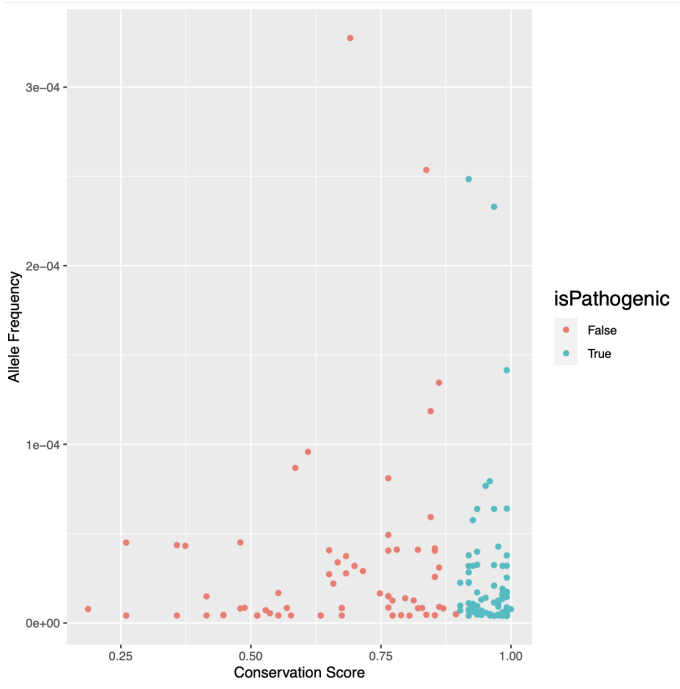


Figure 12: Conservation Score and Allele Frequency scatter plot of mutations classified using the aligned sequences that are part of the focus clade

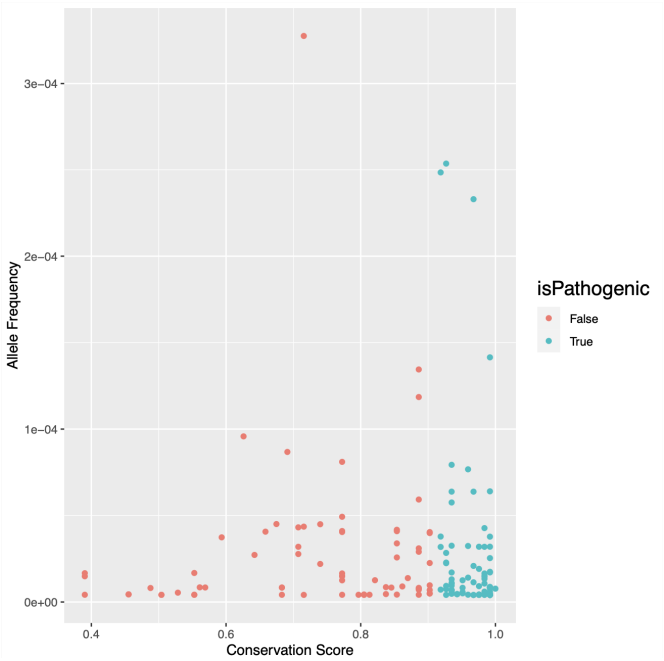


Figure 13: Conservation Score and Allele Frequency scatter plot of mutations classified using the *realigned* sequences that are part of the focus clade

## Discussion

The original idea is examining as many sequences as possible to observe the orthologs/paralogs and pruning the trees where the gene duplication events occurred. Yet, only the hits file which contains 5000 sequences has multiple sequences that belong to Homo sapiens. However, due to complexity issues, aligning 5000 sequences is not possible with the current resources. Hence, the analysis is performed using 500 sequences.

Since a lot of gaps are inserted in the most general alignment of 500 sequences, the conservation scores tend to be low, therefore not even half of the mutations are identified as pathogenic in the first approach. But for the second and third strategies, since the sequences that are probably causing the 'unnecessary' gap insertions are removed, overall conservation score trend is observed to be high. It seems that carrying out another session for realigning the sequences from the focus clade only causes the threshold to be increased while not contributing to the conservation scores of the mutations that the clinical significance is not known for. Therefore, it seems there is not a dramatic difference between the second and third approach of computing conservation scores. Yet, since the second strategy lacks the additional step of realigning, it may be preferable over the third strategy due to complexity

reasons. But in order to make a final decision, one needs more known pathogenic mutations that are identified.

Moreover, one can claim that as the classification threshold increases given the conservation score calculation strategy remains the same, the number of mutations that are classified as pathogenic decreases, hence the number of false negatives increases. Therefore, if one chooses a low threshold, sensitivity would get hurt. On the other hand, as the classification threshold decreases given the conservation score calculation strategy remains the same, the number of mutations that are identified as pathogenic increases which results in false positives. Therefore the specificity of the classification would suffer.

Lastly, it seems like for our classification results, there is no relationship between the mutation type and the allele frequency of the mutation as the p-values indicate that the statistical difference between the pathogenic and neutral mutations with respect to the allele frequencies are not significant. If one examines the scatter-plots of conservation scores and allele frequencies, it would be seen that there is no negative correlation between the conservation scores and allele frequencies. So we can't conclude that the less frequent alleles tend to be on more conserved sites.

## References


1) Medical Publications to find more Mutations: https://ojrd.biomedcentral.com/articles/10.1186/s13023-021-01799-0 https://pubmed.ncbi.nlm.nih.gov/24095820/

2) Merello, E., De Marco, P., Ravegnani, M., Riccipetitoni, G., Cama, A., & Capra, V. (2013). Novel MNX1 mutations and clinical analysis of familial and sporadic Currarino cases. European journal of medical genetics, 56(12), 648–654. https://doi.org/10.1016/j.ejmg.2013.09.011

3) *The Genome Aggregation Database (gnomAD)*. (2021). Https://Www.Nature.Com/Immersive/D42859-020-00002-x/Index.Html. Retrieved 2021, from https://www.nature.com/immersive/d42859-020-00002-x/index.html?error=cookies_not_supported&code=3a68239a-d991-4765-ac68-b7b392c2bbeb

4) *ClinVar Database*. (2021). Https://Www.Ncbi.Nlm.Nih.Gov/Clinvar/Intro/#:~:Text=ClinVar%20is%20a%20freely%20accessible,And%20phenotypes%2C%20with%20supporting%20evidence. Retrieved 2021, from https://www.ncbi.nlm.nih.gov/clinvar/intro/#:~:text=ClinVar%20is%20a%20freely%20accessible,and%20phenotypes%2C%20with%20supporting%20evidence.