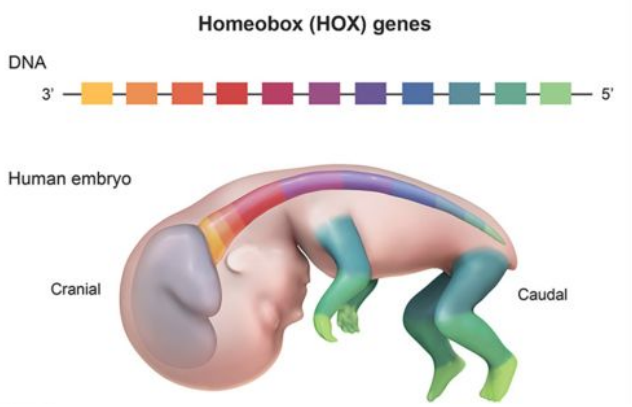
# Mnx1 protein

Analysis using its homologous sequences, phylogenetic tree, and mutations

# Background

- Currarino triad syndrome is an autosomal dominant inherited disease, caused by mutations in the human MNX1 gene, where mutants may have anorectal malformations. [1]
- Mutations in the MNX1 gene influences the translation of Mnx1 protein, a homeobox protein that is responsible for the morphogenesis during embryonic development.
- The mutations on the genes which are translated into homeobox proteins cause severe developmental diseases, since they contain a highly conserved DNA-binding domain known as the homeodomain. [2]



# Workflow

Retrieving homologous sequences for number of 100, 250, 500, 1000, 5000


9606

Choosing a hits fasta file that contains two Homo sapiens sequences

Building MUSCLE alignment and neighbor-joining tree

Computing conservation scores for the aligned sequences

Applying midpoint rooting to the phylogenetic tree

Pruning the tree from the common ancestor of two Homo sapiens sequences and subsetting a single clade, using ete3

Computing conservation scores for the post-alignment sequences on the selected clade

Subsetting the sequences from pre-alignment hits

Building new alignment using MEGA, Muscle

Computing conservation scores for newly aligned subsetted sequences

Comparing the conservation scores for three approaches by plotting histograms

Gathering the neutral mutations from GNOMAD, disease-causing mutations from Clinvar and novel mutations identified in scientific papers for Currarino syndrome

Mapping the mutation positions given for the original protein sequence to the positions on the aligned sequence

Observing any relationship between the mutation type and the conservation score on the mutation occurring site

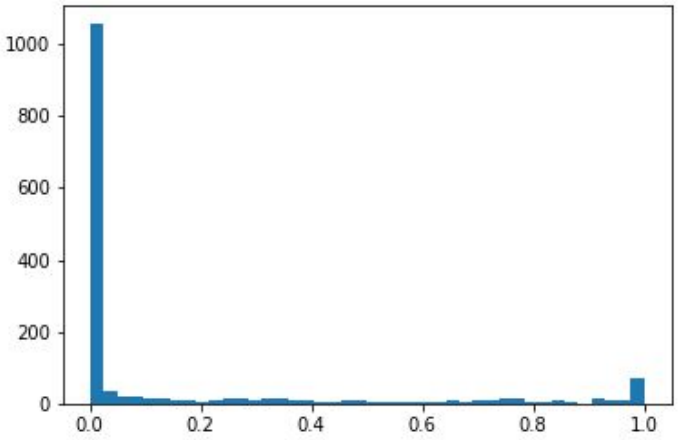



Figure 1: Conservation scores for 500 aligned homologous sequences

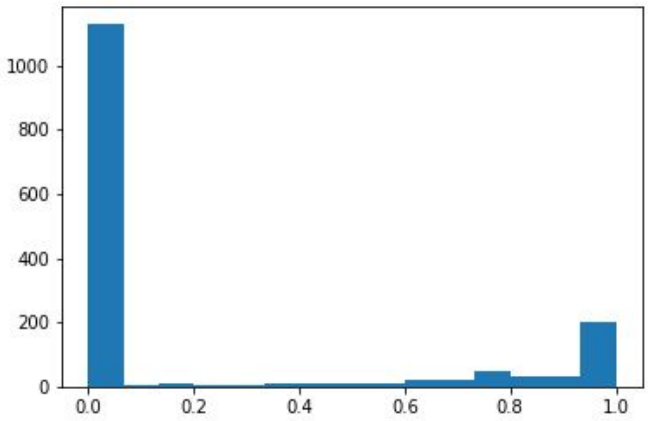



Figure 2: Conservation scores for the aligned sequences in the clade of interest

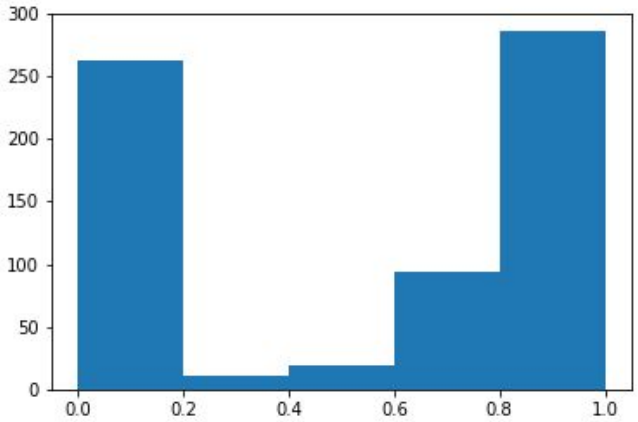



Figure 3: Conservation scores for the sequences in the clade of interest after a new alignment session

# Encountered Issues

- MSA for 5000 sequences has not progressed since the 3rd iteration, and the first three iterations took around 8-9 hours to complete.
- The format of the variants given in ClinVar database cannot be well understood.

**GRCh38/hg38 7p22.3-q36.3(chr7:53985-159282531)x1**

| | |
|---|---|
| **Allele ID:** | 73088 |
| **Variant type:** | copy number loss |
| **Variant length:** | 159,228,547 bp |
| **Cytogenetic location:** | 7p22.3-q36.3 |
| **Genomic location:** | 7: 53985-159282531 (GRCh38) GRCh38 UCSC<br>7: 53985-159075220 (GRCh37) GRCh37 UCSC<br>7: 149068-158767981 (NCBI36) NCBI36 UCSC |

**HGVS:**

| Nucleotide | Protein | Molecular consequence |
|---|---|---|
| NC_000007.14:g.(?_53985)_(159282531_?)del | | |
| NC_000007.13:g.(?_53985)_(159075220_?)del | | |
| NC_000007.12:g.(?_149068)_(158767981_?)del | | |

| | |
|---|---|
| **Protein change:** | - |
| **Other names:** | - |
| **Canonical SPDI:** | - |
| **Functional consequence:** | - |
| **Global minor allele frequency (GMAF):** | - |

# References


1. https://www.sciencedirect.com/topics/neuroscience/homeodomain-protein
2. https://omim.org/entry/176450?search=%22currarino%20syndrome%22&highlight=%22currarino%20%28syndromic%7Csyndrome%29%22