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

G. Görkem Köse - 25359

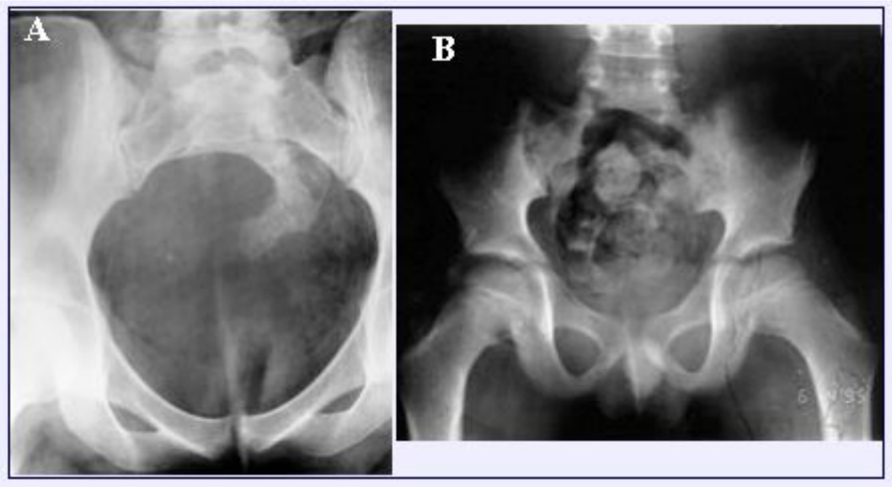
Ahmet Ölçüm - 25915

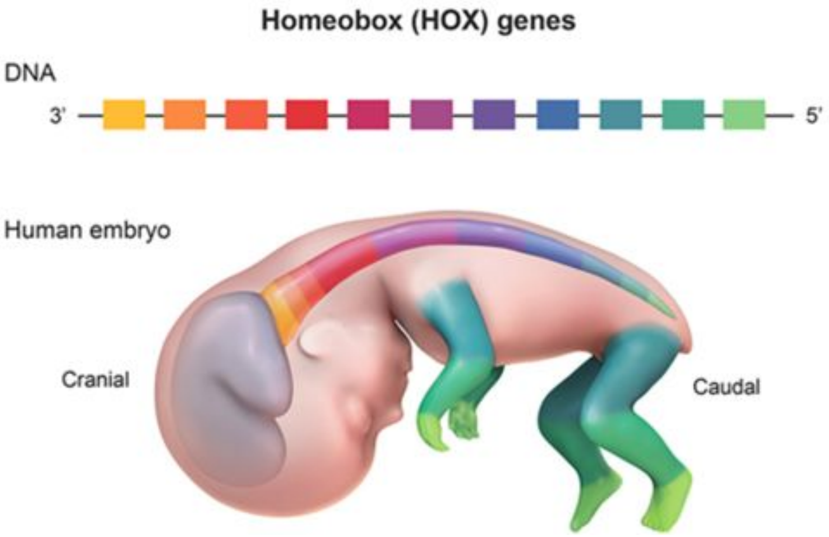
Sayed Damon Sadraije Najafi - 26260

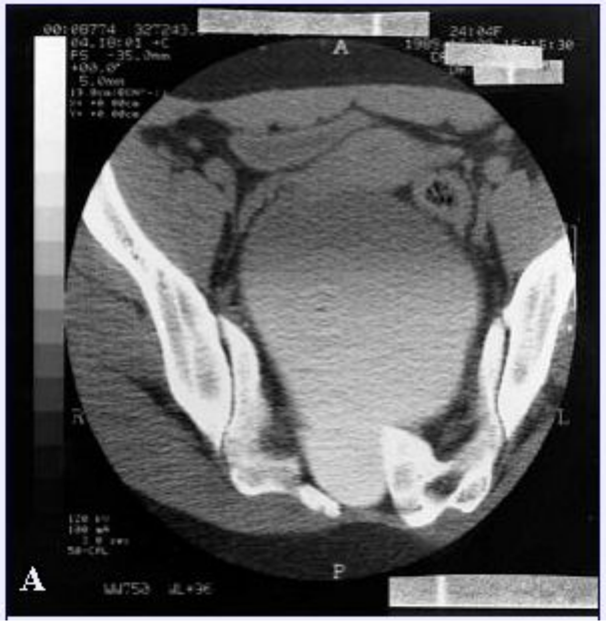
Emir Can Ülkü - 26740

Currarino Syndrome

- Teratoma
- Hamartoma
- Neurenteric cyst
- Anterior meningocele

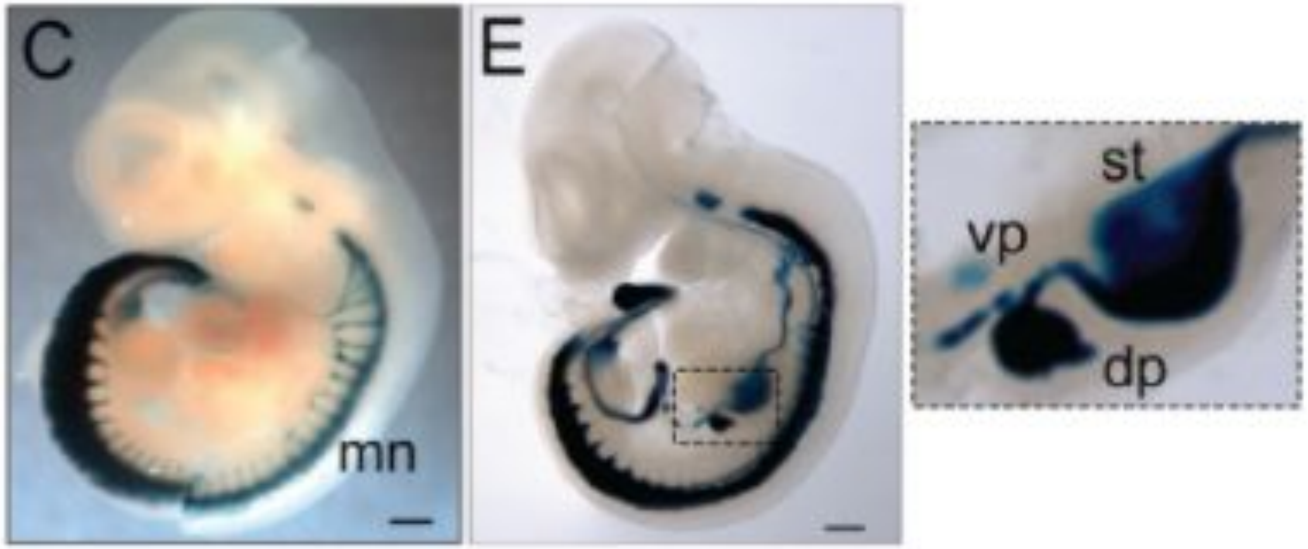


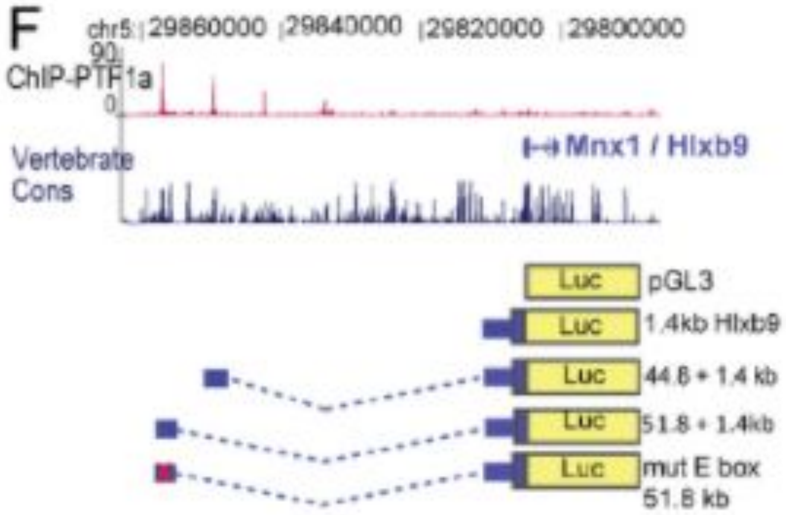




Mnx1 Protein

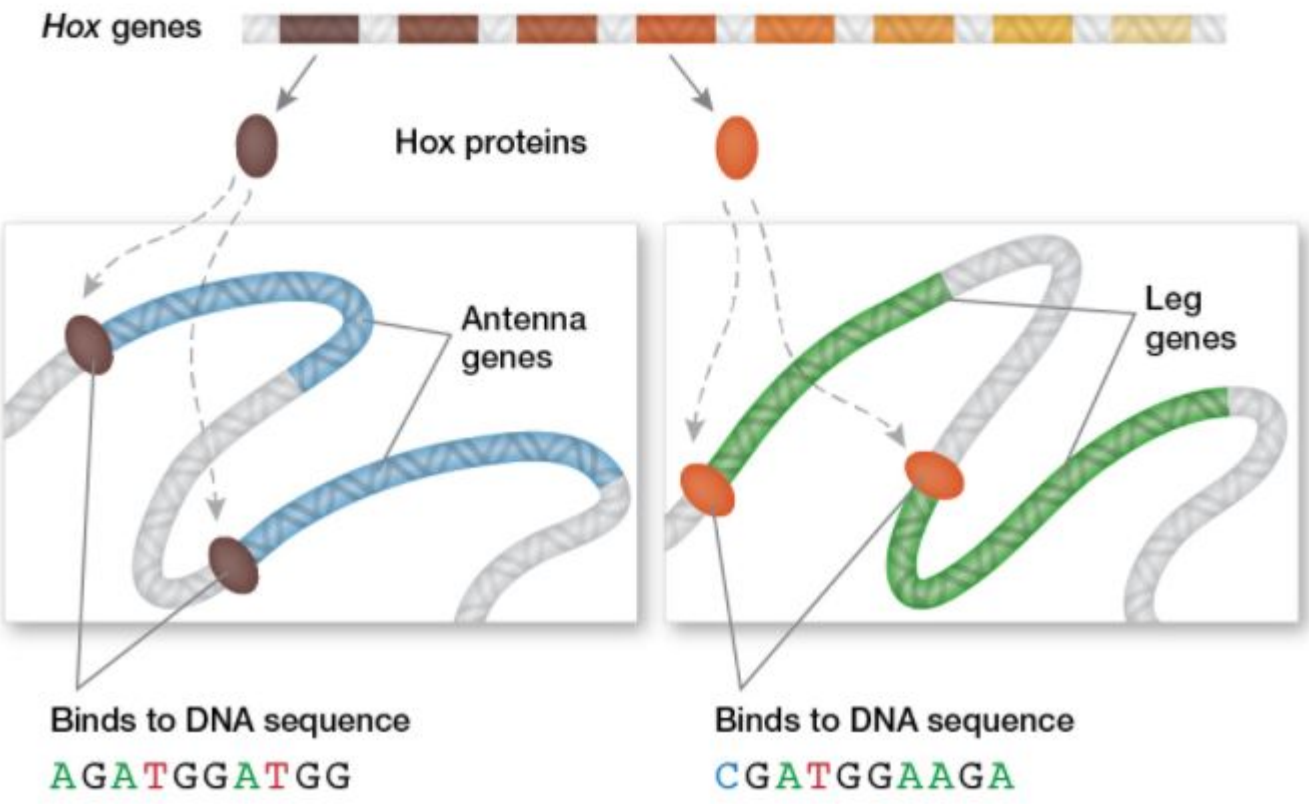
- Nuclear Protein
- Homeodomain
- Hox Protein Like
- Pancreas
- Motor-neuron

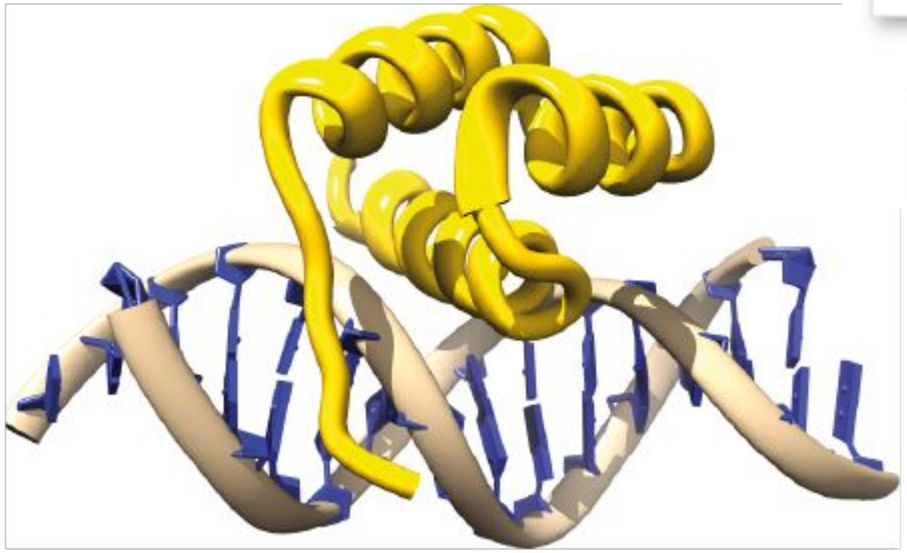




Homeodomain Proteins

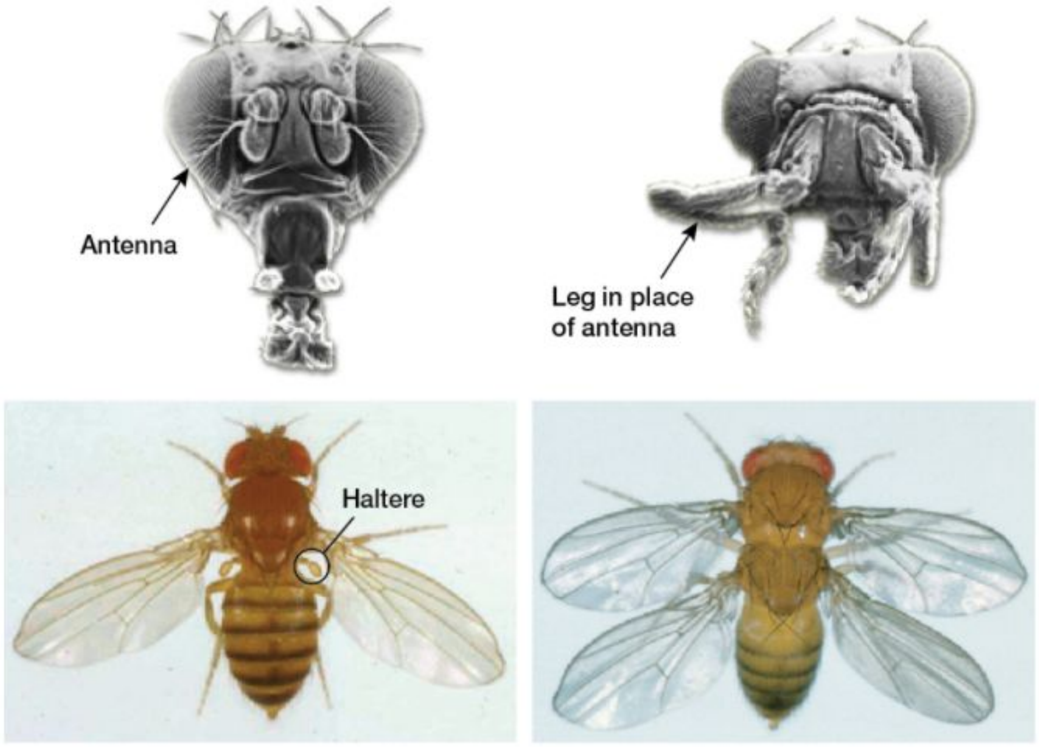
- Gene expression regulation
- Developmental Proteins



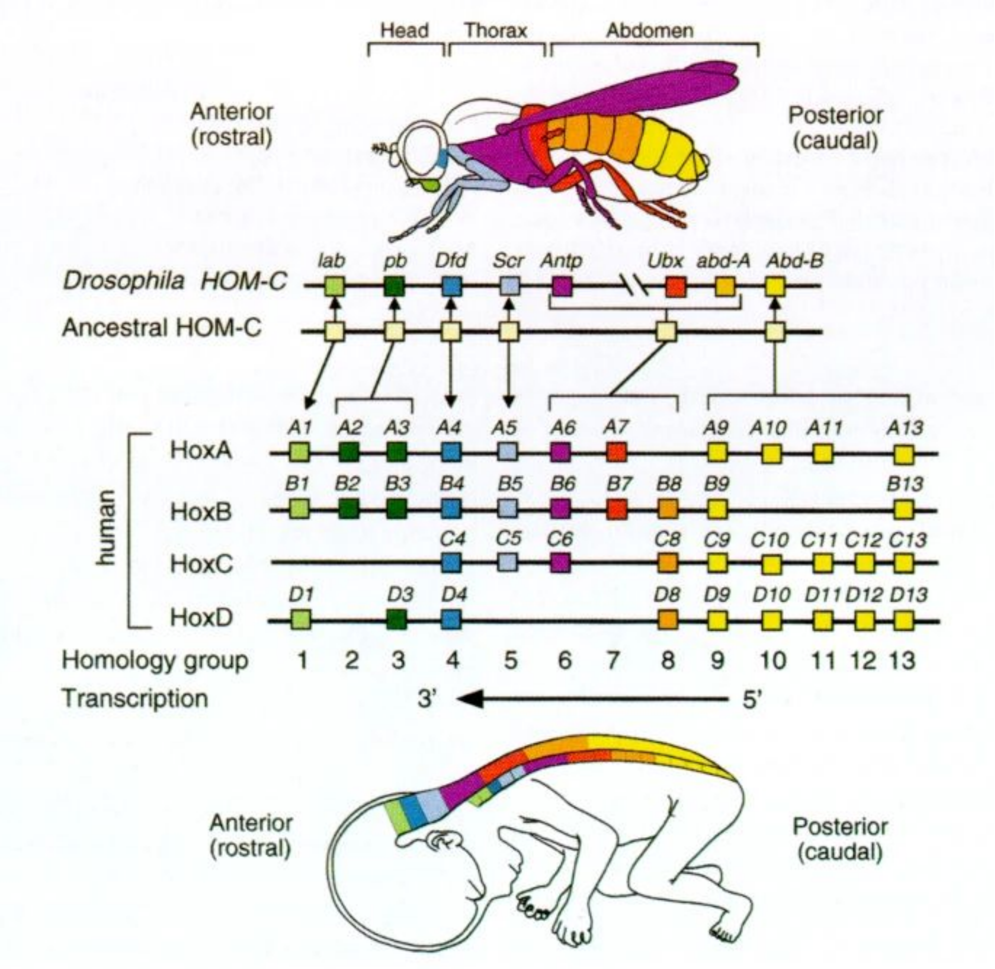


Homeodomain Proteins

- Gene expression regulation
- Developmental Proteins



Top: (Left) Normal fruitfly; (Right) Fruitfly with mutation in antennapedia gene Bottom: (Left) Normal fruitfly; (Right) Fruitfly with a homeotic mutation that gives it two thoraxes. Bottom images courtesy of the Archives, California Institute of Technology.



Materials & Methods

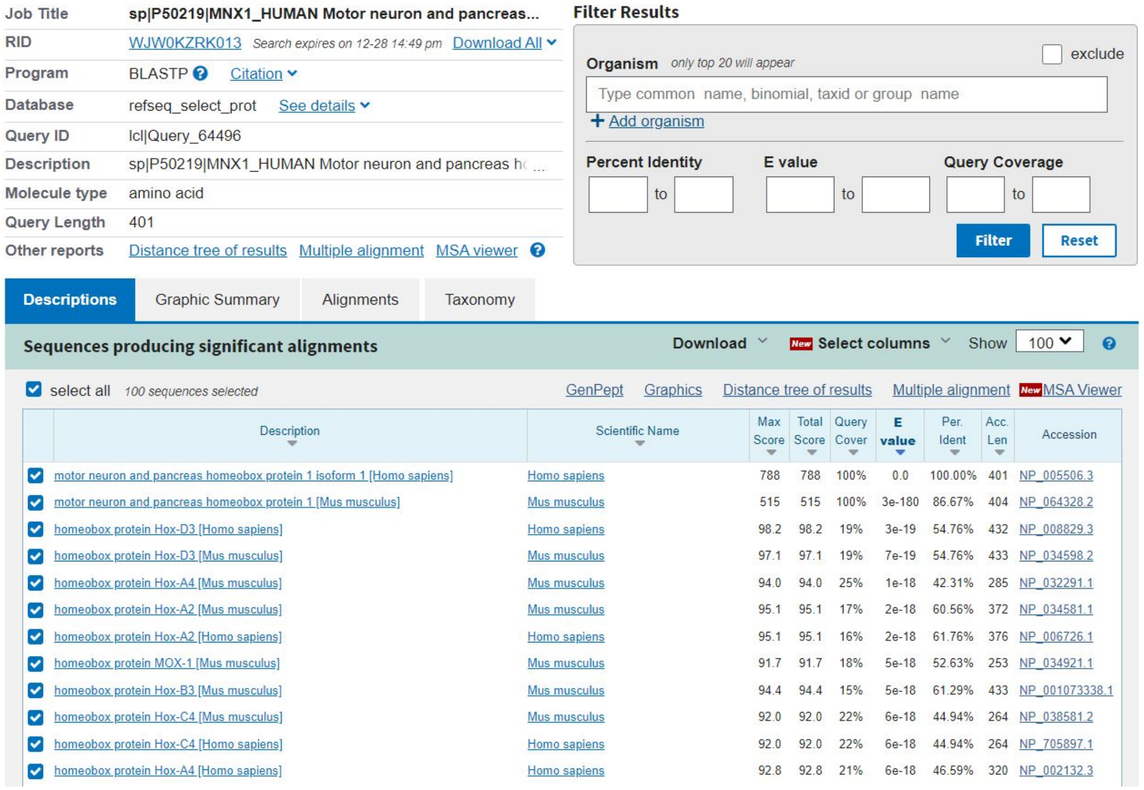
1. Research
2. Getting the protein sequence for Mnx1 (UniProt)
3. Finding homologous proteins (BlastP)
4. Alignment of the homologous proteins (MEGA11)
5. Building Trees (MEGA11)
6. Rerooting the Phylogenetic Tree (FigTree)
7. Pruning the Clade (Python, ete3 library)
8. Calculating conservation scores (Python)
9. Retrieving mutations (Clinical papers & gnomAD)
10. Mapping mutation occurring sites to the aligned sequences (Python)
11. Classification
12. Statistical tests to assess the effect of allele frequencies on mutation type (Python, scipy library)

Research

- From rarediseases.info.nih.gov website, Currarino Triad Syndrome is chosen.
- The protein which cause Currarino Triad Syndrome when a mutation occurs is identified.
- The protein sequence for Mnx1 is retrieved from UniProt.

Finding homologous sequences

- BlastP



Job Title	sp|P50219|MNX1_HUMAN Motor neuron and pancreas...
RID	WJW0KZRK013 *Search expires on 12-28 14:49 pm* Download All
Program	BLASTP Citation
Database	refseq_select_prot See details
Query ID	lcl|Query_64496
Description	sp|P50219|MNX1_HUMAN Motor neuron and pancreas h ...
Molecule type	amino acid
Query Length	401
Other reports	Distance tree of results Multiple alignment MSA viewer

Filter Results

Organism *only top 20 will appear* □ exclude

Type common name, binomial, taxid or group name

+ Add organism

Percent Identity ___ to ___ E value ___ to ___ Query Coverage ___ to ___

Filter Reset

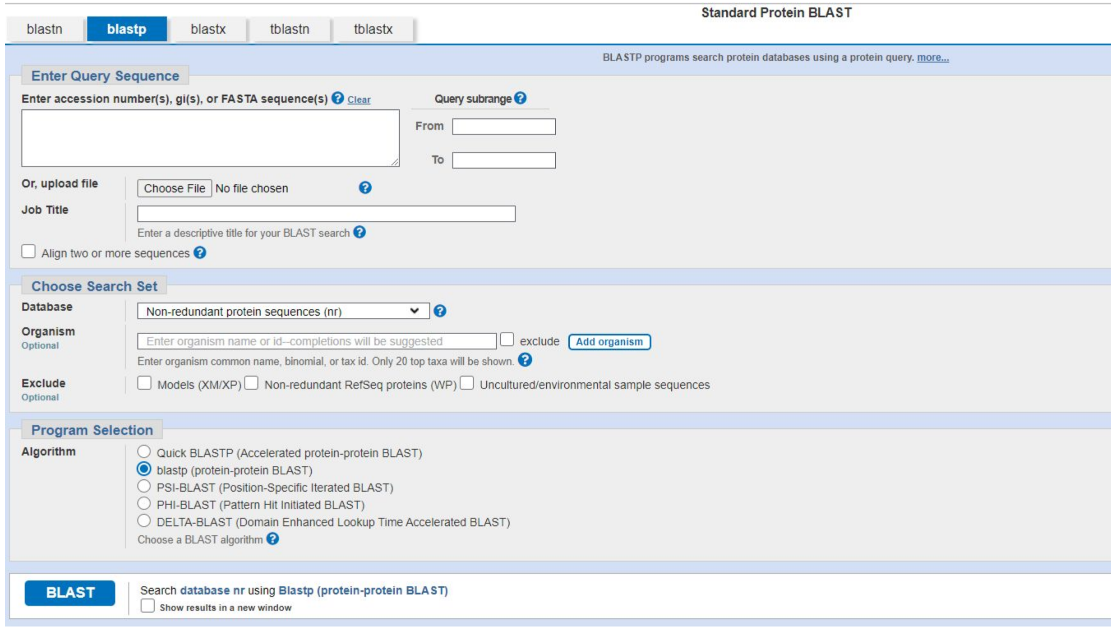
Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments

Download New Select columns Show 100

☑ select all *100 sequences selected* GenPept Graphics Distance tree of results Multiple alignment New MSA Viewer

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	motor neuron and pancreas homeobox protein 1 isoform 1 [Homo sapiens]	Homo sapiens	788	788	100%	0.0	100.00%	401	NP_005506.3
☑	motor neuron and pancreas homeobox protein 1 [Mus musculus]	Mus musculus	515	515	100%	3e-180	86.67%	404	NP_064328.2
☑	homeobox protein Hox-D3 [Homo sapiens]	Homo sapiens	98.2	98.2	19%	3e-19	54.76%	432	NP_008829.3
☑	homeobox protein Hox-D3 [Mus musculus]	Mus musculus	97.1	97.1	19%	7e-19	54.76%	433	NP_034598.2
☑	homeobox protein Hox-A4 [Mus musculus]	Mus musculus	94.0	94.0	25%	1e-18	42.31%	285	NP_032291.1
☑	homeobox protein Hox-A2 [Mus musculus]	Mus musculus	95.1	95.1	17%	2e-18	60.56%	372	NP_034581.1
☑	homeobox protein Hox-A2 [Homo sapiens]	Homo sapiens	95.1	95.1	16%	2e-18	61.76%	376	NP_006726.1
☑	homeobox protein MOX-1 [Mus musculus]	Mus musculus	91.7	91.7	18%	5e-18	52.63%	253	NP_034921.1
☑	homeobox protein Hox-B3 [Mus musculus]	Mus musculus	94.4	94.4	15%	5e-18	61.29%	433	NP_001073338.1
☑	homeobox protein Hox-C4 [Mus musculus]	Mus musculus	92.0	92.0	22%	6e-18	44.94%	264	NP_038581.2
☑	homeobox protein Hox-C4 [Homo sapiens]	Homo sapiens	92.0	92.0	22%	6e-18	44.94%	264	NP_705897.1
☑	homeobox protein Hox-A4 [Homo sapiens]	Homo sapiens	92.8	92.8	21%	6e-18	46.59%	320	NP_002132.3



Standard Protein BLAST

blastn | **blastp** | blastx | tblastn | tblastx

BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

Query subrange
From
To

Or, upload file: Choose File No file chosen

Job Title:
Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: Non-redundant protein sequences (nr)

Organism (Optional): Enter organism name or id--completions will be suggested □ exclude Add organism
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude (Optional): □ Models (XM/XP) □ Non-redundant RefSeq proteins (WP) □ Uncultured/environmental sample sequences

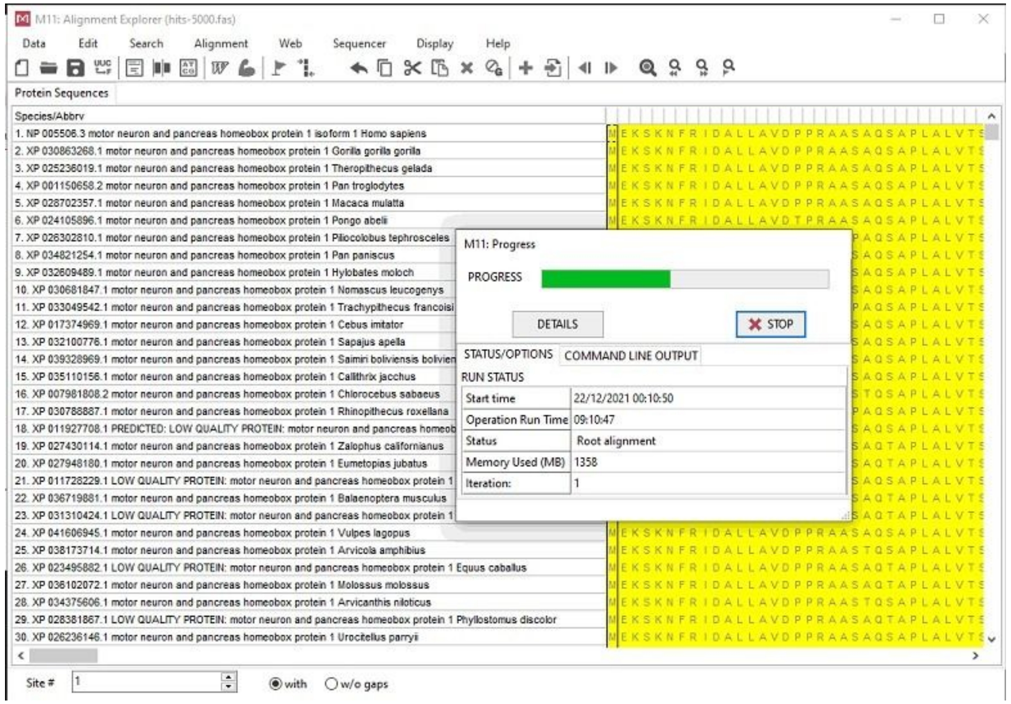
Program Selection

Algorithm:
○ Quick BLASTP (Accelerated protein-protein BLAST)
● blastp (protein-protein BLAST)
○ PSI-BLAST (Position-Specific Iterated BLAST)
○ PHI-BLAST (Pattern Hit Initiated BLAST)
○ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)
Choose a BLAST algorithm

BLAST | Search database nr using Blastp (protein-protein BLAST)
□ Show results in a new window

Multiple sequence alignment

- MEGA11, Muscle (MUltiple Sequence Comparison by Log-Expectation)



Species/Abbrv
1. NP 005506.3 motor neuron and pancreas homeobox protein 1 isoform 1 Homo sapiens
2. XP 030863268.1 motor neuron and pancreas homeobox protein 1 Gorilla gorilla gorilla
3. XP 025236019.1 motor neuron and pancreas homeobox protein 1 Theropithecus gelada
4. XP 001150658.2 motor neuron and pancreas homeobox protein 1 Pan troglodytes
5. XP 028702357.1 motor neuron and pancreas homeobox protein 1 Macaca mulatta
6. XP 024105896.1 motor neuron and pancreas homeobox protein 1 Pongo abelii
7. XP 026302810.1 motor neuron and pancreas homeobox protein 1 Piliocolobus tephrosceles
8. XP 034821254.1 motor neuron and pancreas homeobox protein 1 Pan paniscus
9. XP 032609489.1 motor neuron and pancreas homeobox protein 1 Hylobates moloch
10. XP 030681847.1 motor neuron and pancreas homeobox protein 1 Nomascus leucogenys
11. XP 033049542.1 motor neuron and pancreas homeobox protein 1 Trachypithecus francoisi
12. XP 017374969.1 motor neuron and pancreas homeobox protein 1 Cebus imitator
13. XP 032100776.1 motor neuron and pancreas homeobox protein 1 Sapajus apella
14. XP 039328969.1 motor neuron and pancreas homeobox protein 1 Saimiri boliviensis bolvien
15. XP 035110156.1 motor neuron and pancreas homeobox protein 1 Callithrix jacchus
16. XP 007981808.2 motor neuron and pancreas homeobox protein 1 Chlorocebus sabaeus
17. XP 030788887.1 motor neuron and pancreas homeobox protein 1 Rhinopithecus roxellana
18. XP 011927708.1 PREDICTED: LOW QUALITY PROTEIN: motor neuron and pancreas homeob
19. XP 027430114.1 motor neuron and pancreas homeobox protein 1 Zalophus californianus
20. XP 027948180.1 motor neuron and pancreas homeobox protein 1 Eumetopias jubatus
21. XP 011728229.1 LOW QUALITY PROTEIN: motor neuron and pancreas homeobox protein 1
22. XP 036719881.1 motor neuron and pancreas homeobox protein 1 Balaenoptera musculus
23. XP 031310424.1 LOW QUALITY PROTEIN: motor neuron and pancreas homeobox protein 1
24. XP 041606945.1 motor neuron and pancreas homeobox protein 1 Vulpes lagopus
25. XP 038173714.1 motor neuron and pancreas homeobox protein 1 Arvicola amphibius
26. XP 023495882.1 LOW QUALITY PROTEIN: motor neuron and pancreas homeobox protein 1 Equus caballus
27. XP 036102072.1 motor neuron and pancreas homeobox protein 1 Molossus molossus
28. XP 034375606.1 motor neuron and pancreas homeobox protein 1 Arvicanthis niloticus
29. XP 028381867.1 LOW QUALITY PROTEIN: motor neuron and pancreas homeobox protein 1 Phyllostomus discolor
30. XP 026236146.1 motor neuron and pancreas homeobox protein 1 Urocitellus parryii

Site # 1 ● with ○ w/o gaps

M11: Progress

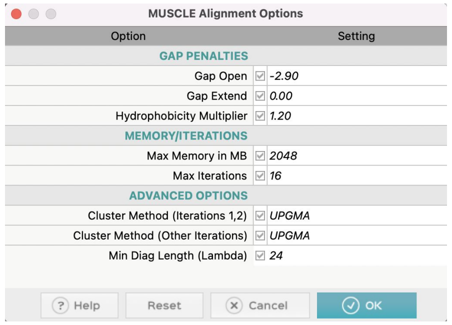
PROGRESS

DETAILS ✖ STOP

STATUS/OPTIONS | COMMAND LINE OUTPUT

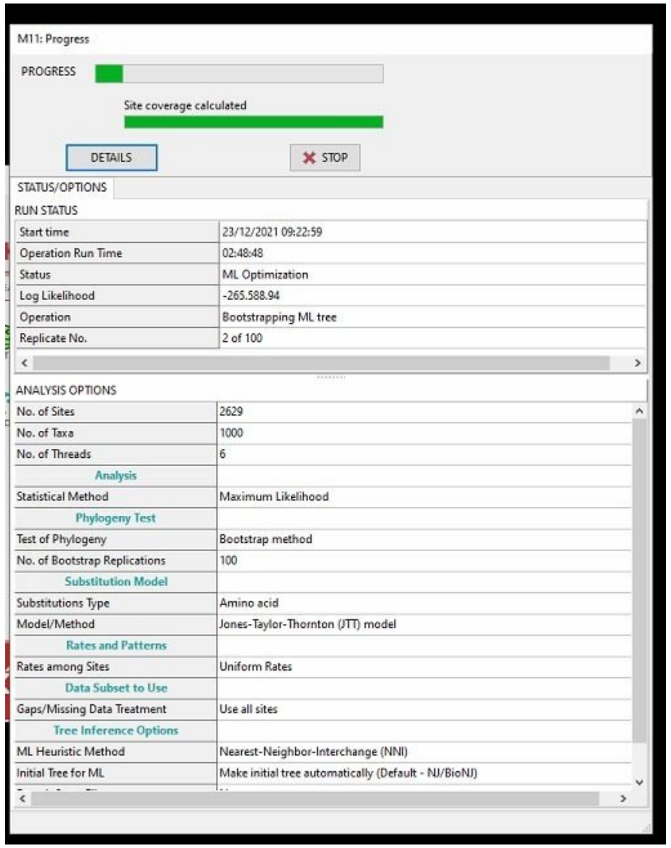
RUN STATUS

Start time	22/12/2021 00:10:50
Operation Run Time	09:10:47
Status	Root alignment
Memory Used (MB)	1358
Iteration:	1



MUSCLE Alignment Options

Option	Setting
GAP PENALTIES	
Gap Open	☑ -2.90
Gap Extend	☑ 0.00
Hydrophobicity Multiplier	☑ 1.20
MEMORY/ITERATIONS	
Max Memory in MB	☑ 2048
Max Iterations	☑ 16
ADVANCED OPTIONS	
Cluster Method (Iterations 1,2)	☑ UPGMA
Cluster Method (Other Iterations)	☑ UPGMA
Min Diag Length (Lambda)	☑ 24



M11: Progress

PROGRESS

Site coverage calculated

DETAILS ✖ STOP

STATUS/OPTIONS

RUN STATUS

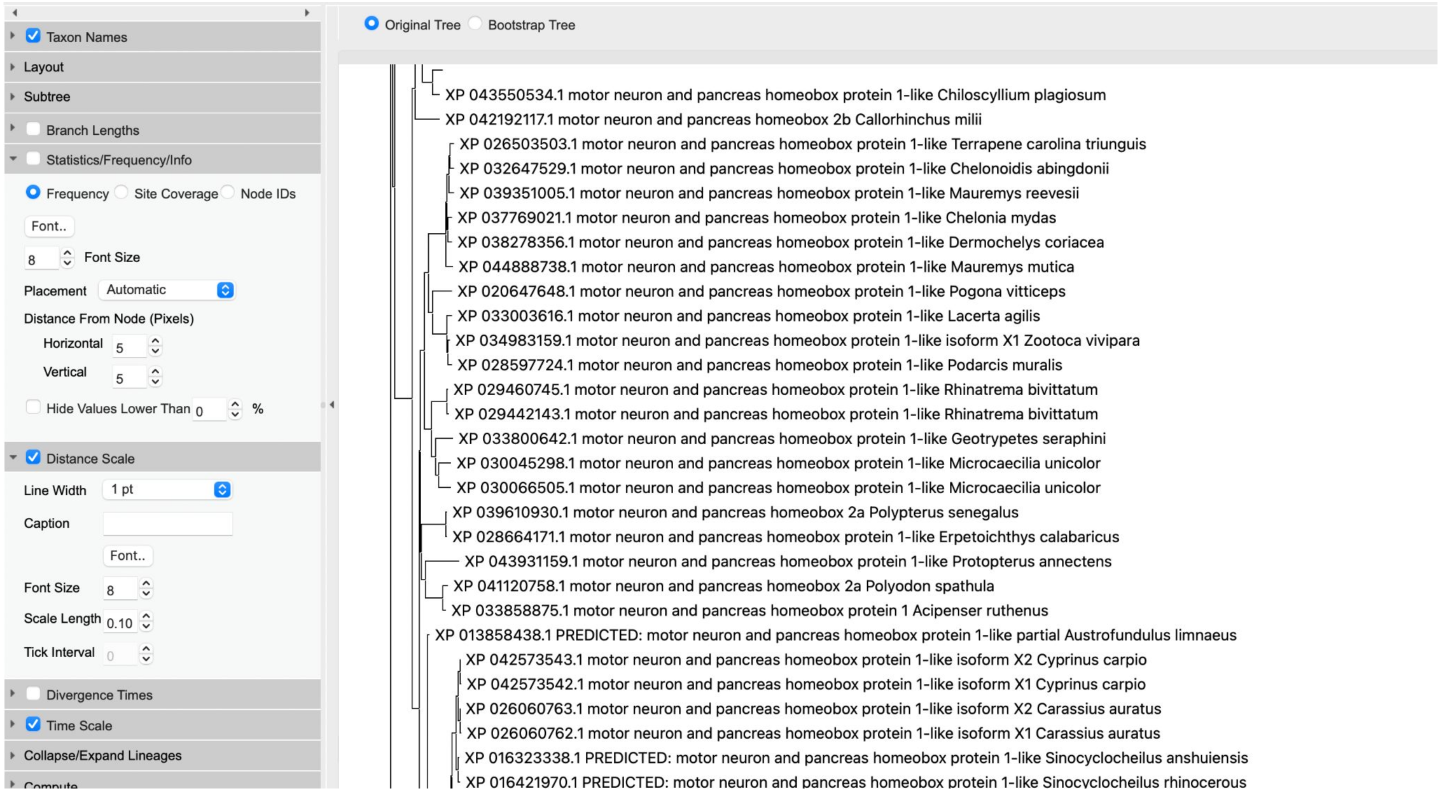
Start time	23/12/2021 09:22:59
Operation Run Time	02:48:48
Status	ML Optimization
Log Likelihood	-265.588.94
Operation	Bootstrapping ML tree
Replicate No.	2 of 100

ANALYSIS OPTIONS

No. of Sites	2629
No. of Taxa	1000
No. of Threads	6
Analysis	
Statistical Method	Maximum Likelihood
Phylogeny Test	
Test of Phylogeny	Bootstrap method
No. of Bootstrap Replications	100
Substitution Model	
Substitutions Type	Amino acid
Model/Method	Jones-Taylor-Thornton (JTT) model
Rates and Patterns	
Rates among Sites	Uniform Rates
Data Subset to Use	
Gaps/Missing Data Treatment	Use all sites
Tree Inference Options	
ML Heuristic Method	Nearest-Neighbor-Interchange (NNI)
Initial Tree for ML	Make initial tree automatically (Default - NJ/BioNJ)

Phylogenetic tree construction

- MEGA11, Neighbor-Joining method



● Frequency ○ Site Coverage ○ Node IDs

Font Size: 8

Placement: Automatic

Distance From Node (Pixels)

Horizontal: 5

Vertical: 5

□ Hide Values Lower Than 0 %

☑ Distance Scale

Line Width: 1 pt

Caption:

Font Size: 8

Scale Length: 0.10

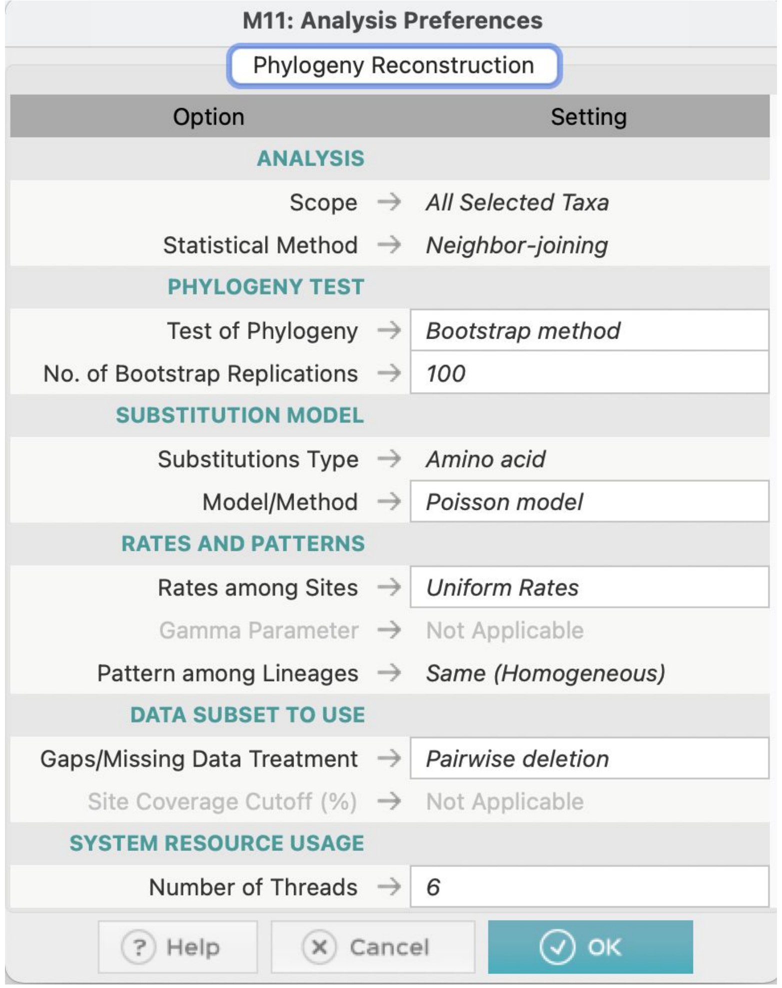
Tick Interval: 0

□ Divergence Times

☑ Time Scale

● Original Tree ○ Bootstrap Tree

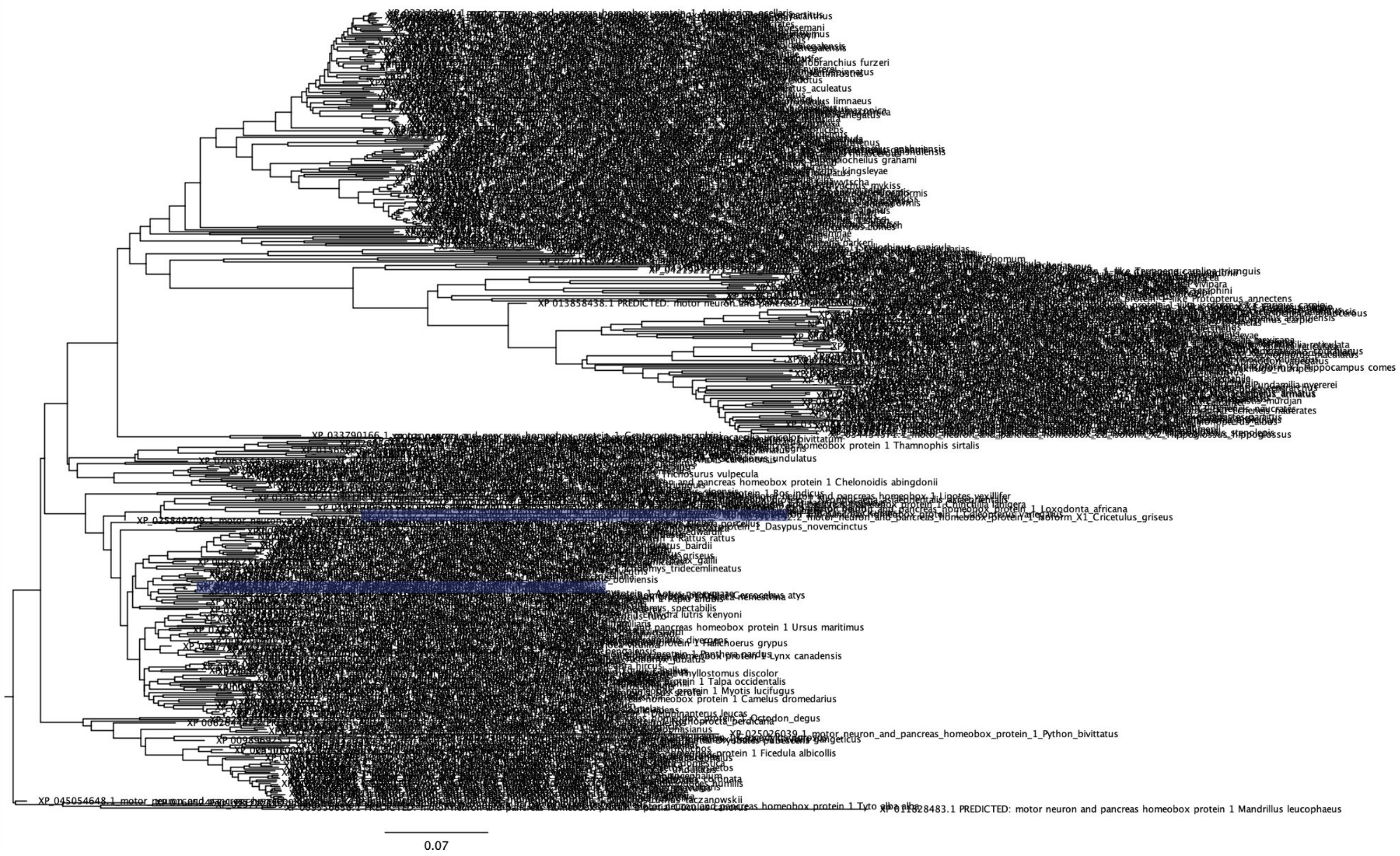
```
XP 043550534.1 motor neuron and pancreas homeobox protein 1-like Chiloscyllium plagiosum
XP 042192117.1 motor neuron and pancreas homeobox 2b Callorhinchus milii
XP 026503503.1 motor neuron and pancreas homeobox protein 1-like Terrapene carolina triunguis
XP 032647529.1 motor neuron and pancreas homeobox protein 1-like Chelonoidis abingdonii
XP 039351005.1 motor neuron and pancreas homeobox protein 1-like Mauremys reevesii
XP 037769021.1 motor neuron and pancreas homeobox protein 1-like Chelonia mydas
XP 038278356.1 motor neuron and pancreas homeobox protein 1-like Dermochelys coriacea
XP 044888738.1 motor neuron and pancreas homeobox protein 1-like Mauremys mutica
XP 020647648.1 motor neuron and pancreas homeobox protein 1-like Pogona vitticeps
XP 033003616.1 motor neuron and pancreas homeobox protein 1-like Lacerta agilis
XP 034983159.1 motor neuron and pancreas homeobox protein 1-like isoform X1 Zootoca vivipara
XP 028597724.1 motor neuron and pancreas homeobox protein 1-like Podarcis muralis
XP 029460745.1 motor neuron and pancreas homeobox protein 1-like Rhinatrema bivittatum
XP 029442143.1 motor neuron and pancreas homeobox protein 1-like Rhinatrema bivittatum
XP 033800642.1 motor neuron and pancreas homeobox protein 1-like Geotrypetes seraphini
XP 030045298.1 motor neuron and pancreas homeobox protein 1-like Microcaecilia unicolor
XP 030066505.1 motor neuron and pancreas homeobox protein 1-like Microcaecilia unicolor
XP 039610930.1 motor neuron and pancreas homeobox 2a Polypterus senegalus
XP 028664171.1 motor neuron and pancreas homeobox protein 1-like Erpetoichthys calabaricus
XP 043931159.1 motor neuron and pancreas homeobox protein 1-like Protopterus annectens
XP 041120758.1 motor neuron and pancreas homeobox 2a Polyodon spathula
XP 033858875.1 motor neuron and pancreas homeobox protein 1 Acipenser ruthenus
XP 013858438.1 PREDICTED: motor neuron and pancreas homeobox protein 1-like partial Austrofundulus limnaeus
XP 042573543.1 motor neuron and pancreas homeobox protein 1-like isoform X2 Cyprinus carpio
XP 042573542.1 motor neuron and pancreas homeobox protein 1-like isoform X1 Cyprinus carpio
XP 026060763.1 motor neuron and pancreas homeobox protein 1-like isoform X2 Carassius auratus
XP 026060762.1 motor neuron and pancreas homeobox protein 1-like isoform X1 Carassius auratus
XP 016323338.1 PREDICTED: motor neuron and pancreas homeobox protein 1-like Sinocyclocheilus anshuiensis
XP 016421970.1 PREDICTED: motor neuron and pancreas homeobox protein 1-like Sinocyclocheilus rhinocerous
```



Option	Setting
ANALYSIS	
Scope →	*All Selected Taxa*
Statistical Method →	*Neighbor-joining*
PHYLOGENY TEST	
Test of Phylogeny →	*Bootstrap method*
No. of Bootstrap Replications →	*100*
SUBSTITUTION MODEL	
Substitutions Type →	*Amino acid*
Model/Method →	*Poisson model*
RATES AND PATTERNS	
Rates among Sites →	*Uniform Rates*
Gamma Parameter →	Not Applicable
Pattern among Lineages →	*Same (Homogeneous)*
DATA SUBSET TO USE	
Gaps/Missing Data Treatment →	*Pairwise deletion*
Site Coverage Cutoff (%) →	Not Applicable
SYSTEM RESOURCE USAGE	
Number of Threads →	*6*

Rerooting the phylogenetic tree

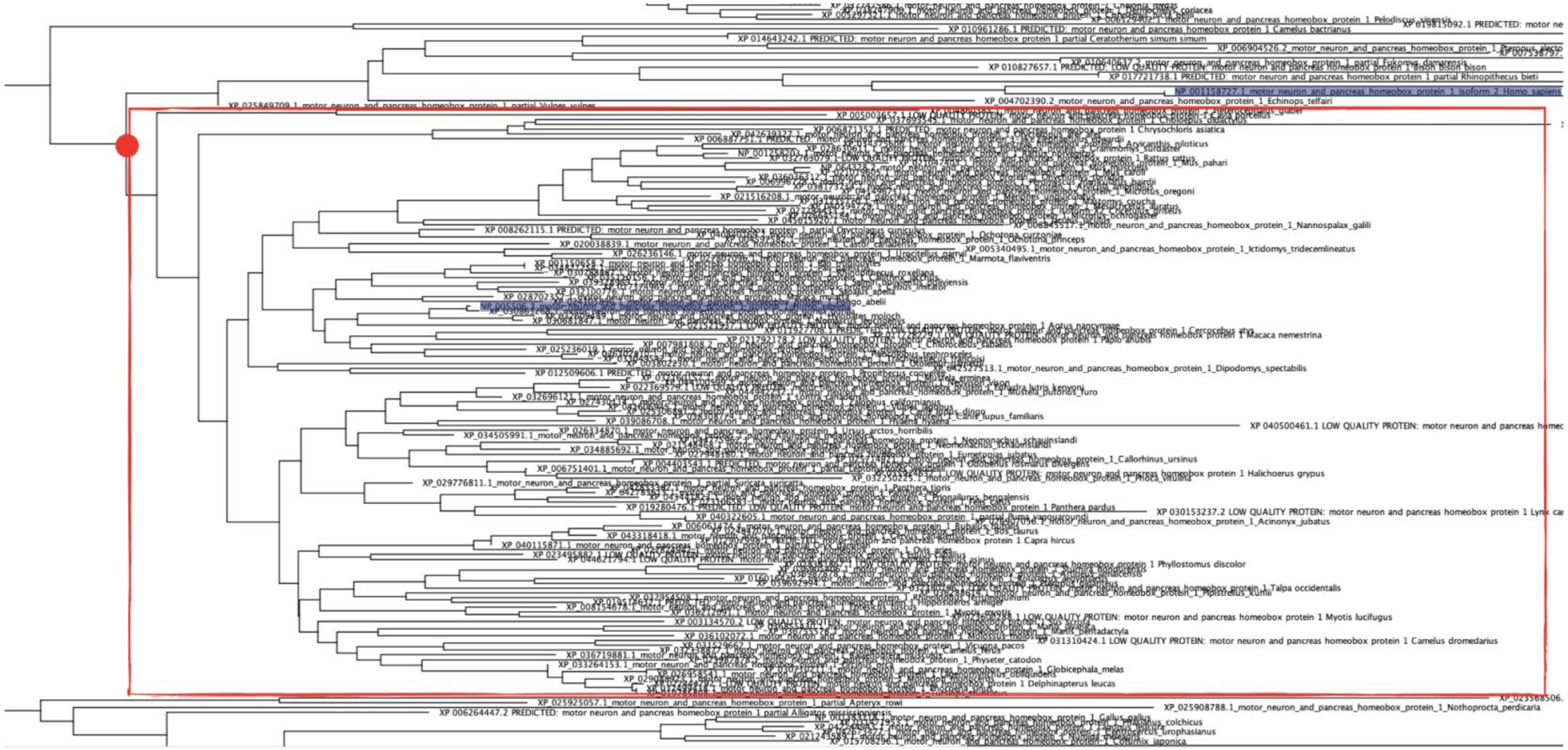
- FigTree



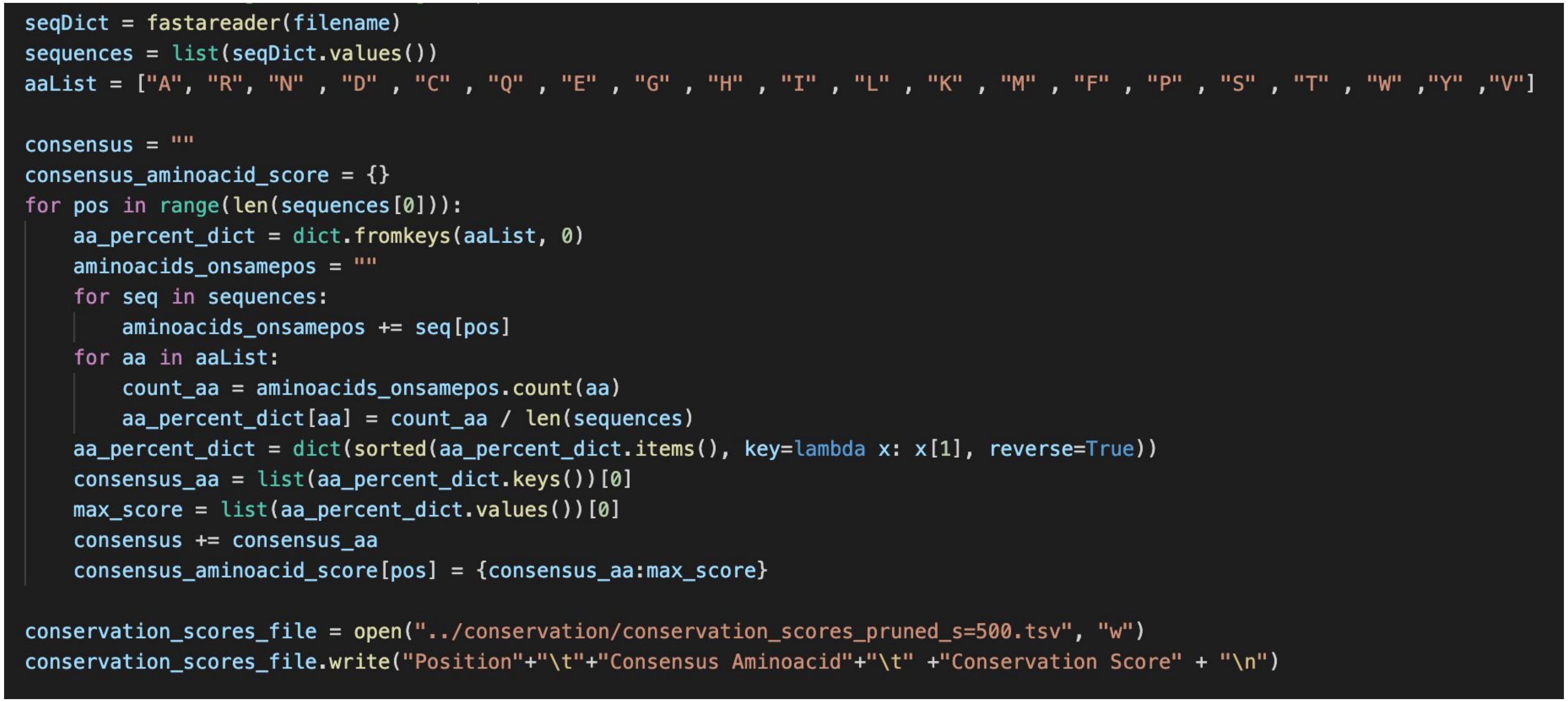
XP_013858438.1 PREDICTED: motor neuron and pancreas
homeobox protein 1 Thamnophis sirtalis
Trichosurus vulpecula
pancreas homeobox protein 1 Chelonoidis abingdonii
Loxodonta_africana
_Dasypus_novemcinctus
mys_tridecemlineatus
lutris kenyoni
pancreas homeobox protein 1 Ursus maritimus
Lynx canadensis
Phyllostomus discolor
Talpa occidentalis
Myotis lucifugus
homeobox protein 1 Camelus dromedarius
Octodon_degus
XP_025026039.1_motor_neuron_and_pancreas_homeobox_protein_1_Python_bivittatus
protein 1 Ficedula albicollis
XP_045054648.1_motor
XP_011828483.1 PREDICTED: motor neuron and pancreas homeobox protein 1 Mandrillus leucophaeus
Protopterus annectens
Hippocampus comes
0.07

Pruning the clade

- Python3, Ete3 library


XP 019815092.1 PREDICTED: motor ne
XP 010961286.1 PREDICTED: motor neuron and pancreas homeobox protein 1 Camelus bactrianus
XP 014643242.1 PREDICTED: motor neuron and pancreas homeobox protein 1 partial Ceratotherium simum simum
XP_006904526.2_motor_neuron_and_pancreas_homeobox_protein_1
XP 010827657.1 PREDICTED: LOW QUALITY PROTEIN: motor neuron and pancreas homeobox protein 1 Bison bison bison
XP 017721738.1 PREDICTED: motor neuron and pancreas homeobox protein 1 partial Rhinopithecus bieti
NP 001158727.1 motor neuron and pancreas homeobox protein 1 isoform 2 Homo sapiens
XP_004702390.2_motor_neuron_and_pancreas_homeobox_protein_1_Echinops_telfairi
XP 025849709.1 motor neuron and pancreas homeobox protein 1 partial Vulpes vulpes
XP 006871352.1 PREDICTED: motor neuron and pancreas homeobox protein 1 Chrysochloris asiatica
XP 008262115.1 PREDICTED: motor neuron and pancreas homeobox protein 1 partial Oryctolagus cuniculus
XP_006845517.1_motor_neuron_and_pancreas_homeobox_protein_1_Nannospalax_galili
XP_020038839.1_motor_neuron_and_pancreas_homeobox_protein_1_Castor_canadensis
XP_005340495.1_motor_neuron_and_pancreas_homeobox_protein_1_Ictidomys_tridecemlineatus
Marmota_flaviventris
Macaca nemestrina
XP_042527513.1_motor_neuron_and_pancreas_homeobox_protein_1_Dipodomys_spectabilis
XP 012509606.1 PREDICTED: motor neuron and pancreas homeobox protein 1 Propithecus coquereli
Canis_lupus_familiaris
XP 040500461.1 LOW QUALITY PROTEIN: motor neuron and pancreas homec
XP_034505991.1_motor_neuron_and_pancreas_homeobox_protein_1_partial_Ailuropoda_melanoleuca
XP_034885692.1_motor_neuron_and_pancreas_homeobox_protein_1_Neomonachus_schauinslandi
homeobox_protein_1_Callorhinus_ursinus
XP_006751401.1_motor_neuron_and_pancreas_homeobox_protein_1_partial_Leptonychotes_weddellii
LOW QUALITY PROTEIN: motor neuron and pancreas homeobox protein 1 Halichoerus grypus
XP_029776811.1_motor_neuron_and_pancreas_homeobox_protein_1_partial_Suricata_suricatta
XP 019280476.1 PREDICTED: LOW QUALITY PROTEIN: motor neuron and pancreas homeobox protein 1 Panthera pardus
XP 030153237.2 LOW QUALITY PROTEIN: motor neuron and pancreas homeobox protein 1 Lynx ca
XP 040322605.1 motor neuron and pancreas homeobox protein 1 partial Puma yagouaroundi
motor_neuron_and_pancreas_homeobox_protein_1_Acinonyx_jubatus
motor neuron and pancreas homeobox protein 1 Capra hircus
motor neuron and pancreas homeobox protein 1 Phyllostomus discolor
motor neuron and pancreas homeobox protein 1 Talpa occidentalis
LOW QUALITY PROTEIN: motor neuron and pancreas homeobox protein 1 Myotis lucifugus
XP 031310424.1 LOW QUALITY PROTEIN: motor neuron and pancreas homeobox protein 1 Camelus dromedarius
XP_036102072.1_motor_neuron_and_pancreas_homeobox_protein_1_Molossus_molossus
XP_033264153.1_motor_neuron_and_pancreas_homeobox_protein_1_Physeter_catodon
Globicephala_melas
protein 1 Delphinapterus leucas
XP_025925057.1_motor_neuron_and_pancreas_homeobox_protein_1_partial_Apteryx_rowi
XP_025908788.1_motor_neuron_and_pancreas_homeobox_protein_1_Nothoprocta_perdicaria
XP 006264447.2 PREDICTED: motor neuron and pancreas homeobox protein 1 partial Alligator mississippiensis
Phasianus_colchicus
Tympanuchus_urophasianus
XP_015708296.1_motor_neuron_and_pancreas_homeobox_protein_1_Coturnix_japonica

Calculating conservation scores

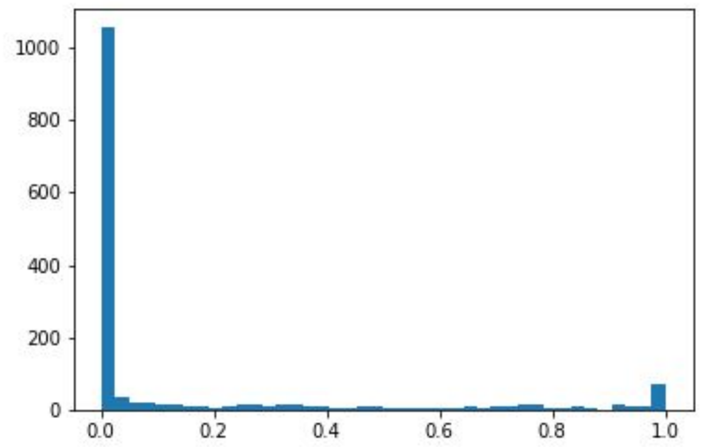


```
seqDict = fastareader(filename)
sequences = list(seqDict.values())
aaList = ["A", "R", "N" , "D" , "C" , "Q" , "E" , "G" , "H" , "I" , "L" , "K" , "M" , "F" , "P" , "S" , "T" , "W" ,"Y" ,"V"]

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}

conservation_scores_file = open("../conservation/conservation_scores_pruned_s=500.tsv", "w")
conservation_scores_file.write("Position"+"\t"+"Consensus Aminoacid"+"\t" +"Conservation Score" + "\n")
```

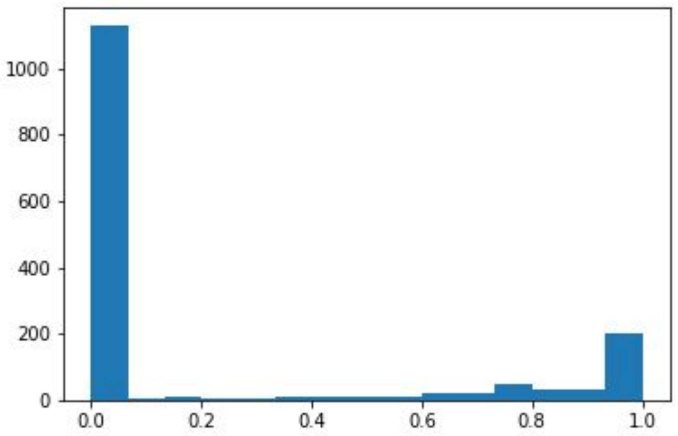
3 different conservation score calculation sets

- First method:
  - 500 aligned sequences
- Second method:
  - Already aligned sequences that are in the focus clade
- Third method:
  - Realigned sequences that are in the focus clade

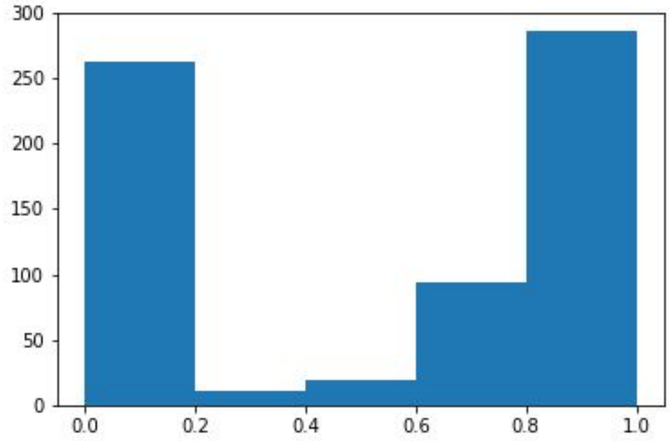
Conservation score histograms for different methods



First method



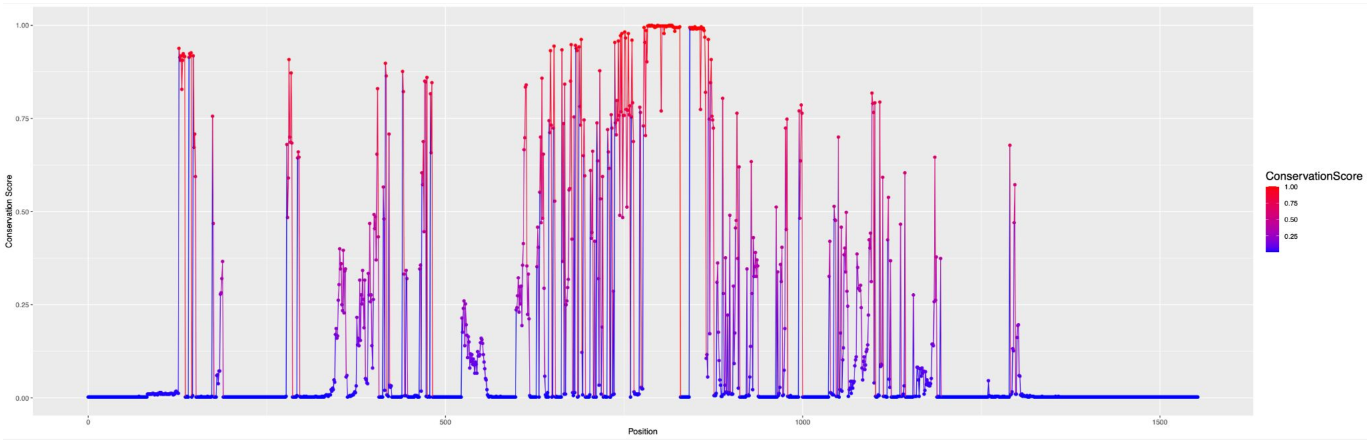
Second method



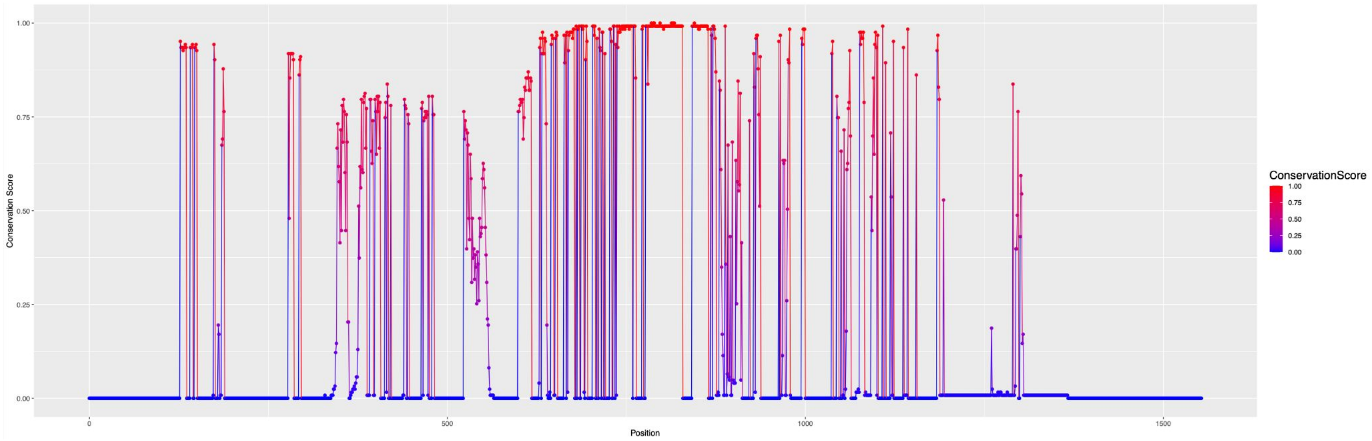
Third method

Conservation scores per position

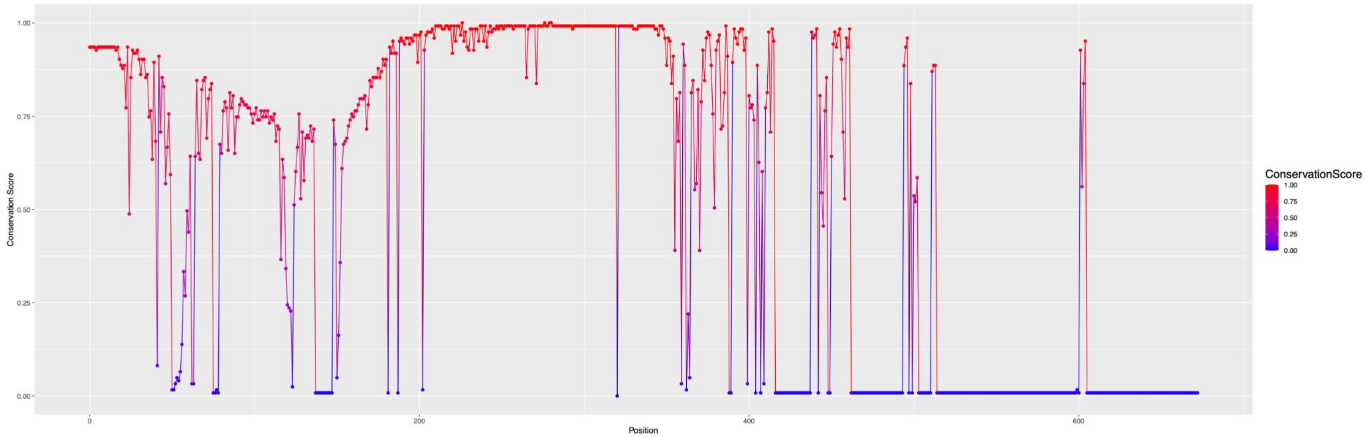
First method



Second method

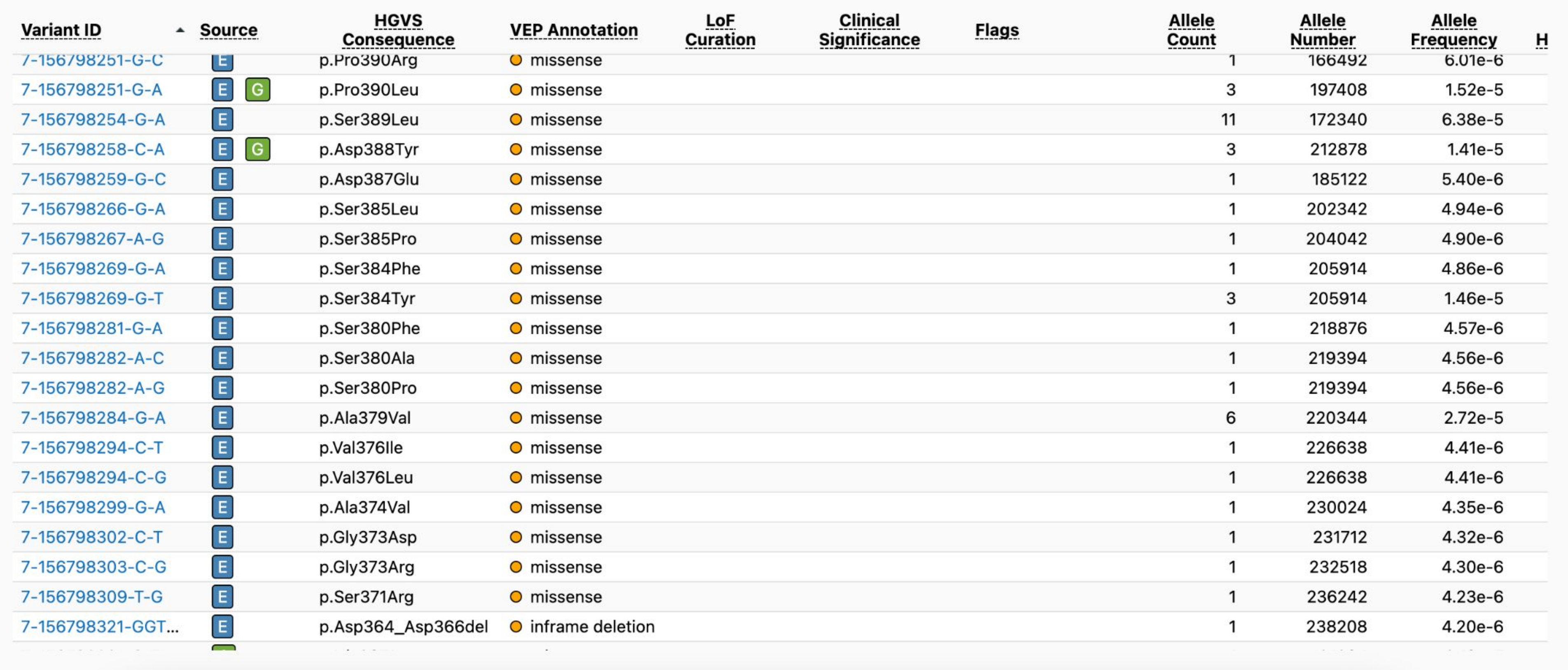


Third method



Retrieving mutations

- Known pathogenic mutations are retrieved from clinical papers about Currarino Syndrome. (17 such mutations)
- The mutations for which the clinical significance is not know are retrieved from gnomAD. (173 such mutations)



Variant ID	Source	HGVS Consequence	VEP Annotation	LoF Curation	Clinical Significance	Flags	Allele Count	Allele Number	Allele Frequency	H
7-156798251-G-C	E	p.Pro390Arg	● missense				1	166492	6.01e-6	
7-156798251-G-A	E G	p.Pro390Leu	● missense				3	197408	1.52e-5	
7-156798254-G-A	E	p.Ser389Leu	● missense				11	172340	6.38e-5	
7-156798258-C-A	E G	p.Asp388Tyr	● missense				3	212878	1.41e-5	
7-156798259-G-C	E	p.Asp387Glu	● missense				1	185122	5.40e-6	
7-156798266-G-A	E	p.Ser385Leu	● missense				1	202342	4.94e-6	
7-156798267-A-G	E	p.Ser385Pro	● missense				1	204042	4.90e-6	
7-156798269-G-A	E	p.Ser384Phe	● missense				1	205914	4.86e-6	
7-156798269-G-T	E	p.Ser384Tyr	● missense				3	205914	1.46e-5	
7-156798281-G-A	E	p.Ser380Phe	● missense				1	218876	4.57e-6	
7-156798282-A-C	E	p.Ser380Ala	● missense				1	219394	4.56e-6	
7-156798282-A-G	E	p.Ser380Pro	● missense				1	219394	4.56e-6	
7-156798284-G-A	E	p.Ala379Val	● missense				6	220344	2.72e-5	
7-156798294-C-T	E	p.Val376Ile	● missense				1	226638	4.41e-6	
7-156798294-C-G	E	p.Val376Leu	● missense				1	226638	4.41e-6	
7-156798299-G-A	E	p.Ala374Val	● missense				1	230024	4.35e-6	
7-156798302-C-T	E	p.Gly373Asp	● missense				1	231712	4.32e-6	
7-156798303-C-G	E	p.Gly373Arg	● missense				1	232518	4.30e-6	
7-156798309-T-G	E	p.Ser371Arg	● missense				1	236242	4.23e-6	
7-156798321-GGT...	E	p.Asp364_Asp366del	● inframe deletion				1	238208	4.20e-6	

Mapping mutation occurring sites to aligned sequences

- Original sequence:
  - MEKSKNFRIDALLAVDP...

- Aligned sequence:
  - ...--------------MEKSKNFRI-----DA...

- For a mutation such as M1A, the position is not 1 in the aligned sequence. Therefore, the actual positions and the aligned positions are mapped.

Classification

- For each conservation score table, the conservation scores at the positions for the known pathogenic mutations are averaged and set as a threshold for classification.

-
  ```
  For each unknown-type mutation:
      if the conservation score is above the threshold:
       classify as pathogenic
     else
       classify as neutral
  ```

Classification results

First method:

- Classification threshold: 0.68
- Number of pathogenic mutations: 77
- Number of neutral mutations: 96

Second method:

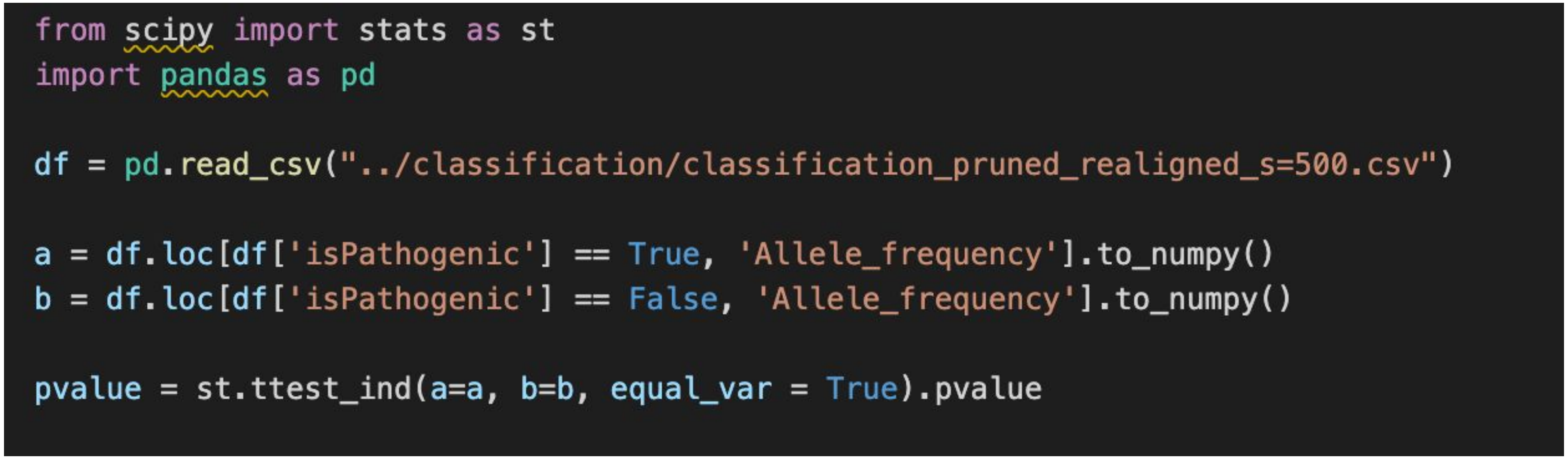
- Classification threshold: 0.89
- Number of pathogenic mutations: 104
- Number of neutral mutations: 69

Third method:

- Classification threshold: 0.91
- Number of pathogenic mutations: 102
- Number of neutral mutations: 71

Statistical tests to assess the effect of allele frequencies on mutation type

- Pathogenic and neutral mutations are separated into two samples, and an independent t-test is conducted on the allele frequencies.

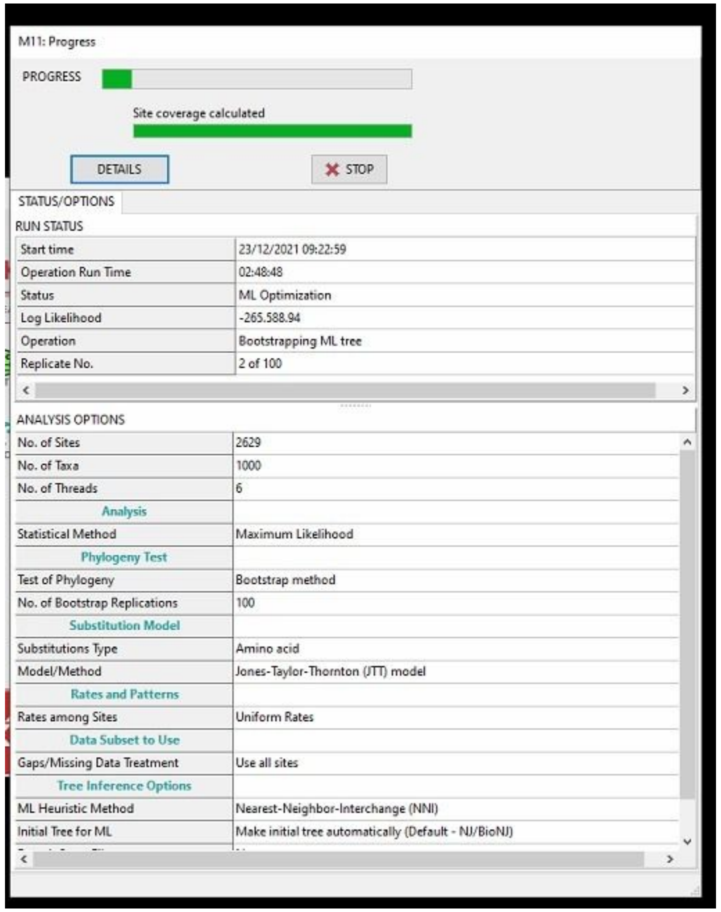


```python
from scipy import stats as st
import pandas as pd

df = pd.read_csv("../classification/classification_pruned_realigned_s=500.csv")

a = df.loc[df['isPathogenic'] == True, 'Allele_frequency'].to_numpy()
b = df.loc[df['isPathogenic'] == False, 'Allele_frequency'].to_numpy()

pvalue = st.ttest_ind(a=a, b=b, equal_var = True).pvalue
```

- p-values for:
  - First method → 0.18
  - Second method → 0.08
  - Third method → 0.35

Issues

- Why we did not use the sequence files of 100 and 250?
- Why we did not use the sequence files of 1000 and 5000?



Site coverage calculated

RUN STATUS

Start time	23/12/2021 09:22:59
Operation Run Time	02:48:48
Status	ML Optimization
Log Likelihood	-265.588.94
Operation	Bootstrapping ML tree
Replicate No.	2 of 100

ANALYSIS OPTIONS

No. of Sites	2629
No. of Taxa	1000
No. of Threads	6
Analysis	
Statistical Method	Maximum Likelihood
Phylogeny Test	
Test of Phylogeny	Bootstrap method
No. of Bootstrap Replications	100
Substitution Model	
Substitutions Type	Amino acid
Model/Method	Jones-Taylor-Thornton (JTT) model
Rates and Patterns	
Rates among Sites	Uniform Rates
Data Subset to Use	
Gaps/Missing Data Treatment	Use all sites
Tree Inference Options	
ML Heuristic Method	Nearest-Neighbor-Interchange (NNI)
Initial Tree for ML	Make initial tree automatically (Default - NJ/BioNJ)

Discussion

- Number of known pathogenic mutations should increase in order to classify the unknown mutations accurately.
- As the genetic relevance of the sequences increases in a set of sequences, conservation scores also increase.
- Given the same conservation score calculation method:
    - as classification threshold increases → sensitivity decreases.
    - as classification threshold decreases → specificity decreases.
- Number of mutations which identified as pathogenic and the classification thresholds are very close for the second and third method.
- With the observation of p-values, there is no statistical significance of the allele frequencies.

References

- Thompson, Nancy & Gésina, Emilie & Scheinert, Peter & Bucher, Philipp & Grapin-Botton, Anne. (2012). RNA Profiling and Chromatin Immunoprecipitation-Sequencing Reveal that PTF1a Stabilizes Pancreas Progenitor Identity via the Control of MNX1/HLXB9 and a Network of Other Transcription Factors. Molecular and cellular biology. 32. 1189-99. 10.1128/MCB.06318-11.
- Truscott M., Nepveu A. (2005) Homeodomains. In: Encyclopedic Reference of Genomics and Proteomics in Molecular Medicine. Springer, Berlin, Heidelberg . https://doi.org/10.1007/3-540-29623-9_2740
- Currarino syndrome Lynch, SA Review on Currarino syndrome, with data on clinics, and the genes involved. Jean-Loup Huret (Editor-in-Chief) INIST-CNRS (Publisher) 2008
- Mark, M., Rijli, F. & Chambon, P. Homeobox Genes in Embryogenesis and Pathogenesis. *Pediatr Res* 42, 421–429 (1997). https://doi.org/10.1203/00006450-199710000-00001
- Genetic Science Learning Center. (2016, March 1) Homeotic Genes and Body Patterns. Retrieved December 25, 2021, from https://learn.genetics.utah.edu/content/basics/hoxgenes/
- Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, Chen X, Taipale J, Hughes TR, Weirauch MT.(2018) The Human Transcription Factors. Cell. 172(4):650-665. doi: 10.1016/j.cell.2018.01.029. Review.