NOVEMBER 04 , 2023

Perspectives of Informatics for

Human Genome Project

Ananya Arora (20BCB0031)
Bhavana Jain (20BCB0105)
Saloni Vyas (20BCB0064)
Sanjna Subramanian (20BCB0088)

Aim

A brief description of our work

The aim of the project is to leverage informatics and computational approaches to conduct a comprehensive comparative genomics analysis of primate and human genomes within the context of the Human Genome Project (HGP).

This comparative genomics initiative seeks to explore, understand, and elucidate the genetic similarities and differences between humans and primates, with a focus on a variety of aspects, including evolution, functional genomics, and potential implications for human health and biology.

Human Genome Project

The Human Genome Project (HGP) was an international scientific endeavor that aimed to map and sequence the entire human genome, which is the complete set of genes and genetic material in a human being. The HGP provided crucial insights into human genetics, paving the way for numerous advancements in medicine, genetics, and biotechnology. It remains a landmark achievement in the study of human biology and has had a profound impact on our understanding of human health and disease.

It had two primary goals:

1. Mapping the Human Genome: The first goal of the HGP was to create a comprehensive genetic map of the human genome. This involved identifying and locating the positions of specific genes and genetic markers on each of the 23 pairs of human chromosomes.

2. Sequencing the Human Genome: The second goal was to determine the precise order of the 3 billion base pairs in the human genome. This was a massive and technically challenging task, as it required determining the sequence of the DNA letters (A, T, C, and G) that make up human DNA.

Human Genome Project

The results of the HGP have had a profound impact on various fields of science and medicine. Some key findings and results include:

1. Gene Identification: The HGP identified and mapped approximately 20,000-25,000 protein-coding genes in the human genome. This information has been invaluable for understanding the genetic basis of human traits and diseases.

2. Non-Coding DNA: It revealed that a significant portion of the human genome consists of non-coding DNA, which was previously considered "junk DNA." While some non-coding regions have regulatory functions, others are still not fully understood.

3. Human Genetic Variation: The project also helped identify genetic variations among individuals. This information has been crucial in the study of human diversity, evolution, and susceptibility to diseases.

4. Medical Advances: The HGP has greatly accelerated research in genetics and genomics, leading to significant advances in personalized medicine, the diagnosis and treatment of genetic disorders, and our understanding of diseases like cancer, diabetes, and cardiovascular conditions.

Comparative Genomics

Comparative genomics is a branch of genomics that involves the study of similarities and differences in the genetic material (genomes) of different species. It's a powerful approach in genetics and biology that focuses on comparing the DNA sequences, gene structures, and functional elements of genomes to gain insights into evolutionary relationships, gene function, and the genetic basis of traits or diseases.

Here are some key aspects of comparative genomics:

- Genome Sequencing
- Evolutionary Relationships
- Gene Function and Conservation
- Gene Families
- Functional Annotations
- Identification of Regulatory Elements
- Structural Variations
- Disease Studies

Methodology – Species Selection

We will be choosing several primate species along with Homo sapiens. We will also add one seemingly distant mammal species into the mix to explore how much of the protein sequence is conserved and to explore some evolutionary relationships.

Main Control Species
Human - Homo sapiens

Primate Species
Chimpanzee - Pan troglodytes
Bonobo - Pan paniscus
Gorilla - Gorilla gorilla
Orangutan - Pongo
Rhesus monkey - Macaca mulatta

Distant Species
Mouse - Mus musculus

Methodology – Gene Selection

1. FOXP2 - forkhead box P2
2. BRCA 1 breast cancer early onset
3. CFTR (cystic fibrosis transmembrane conductance regulator)
4. APOE

Methodology – Tool Selection

UniProtKB

UniProtKB is the world's most comprehensive resource for protein sequence and annotation data. It contains over 200 million protein sequences from over 500,000 species, including humans, other animals, plants, and microbes. UniProtKB also contains a wealth of information about protein function, structure, and interactions.

Clustal Omega

Clustal is a multiple sequence alignment (MSA) program that is widely used in bioinformatics research. It is a progressive MSA algorithm, which means that it aligns sequences one at a time to a growing alignment. Clustal is known for its accuracy and speed, and it can be used to align a wide range of sequence types, including proteins, DNA, and RNA.

FOXP2 Gene

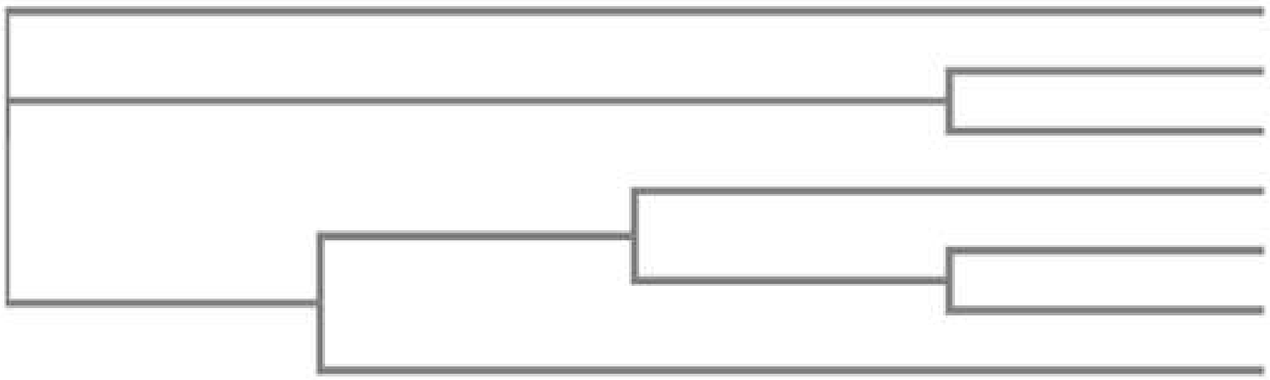
- The FOXP2 gene provides instructions for making a protein called forkhead box P2. This protein is a transcription factor.
- Researchers suspect that the forkhead box P2 protein may regulate hundreds of genes, although only some of its targets have been identified.
- Studies suggest that it plays important roles in brain development, including the growth of nerve cells (neurons) and the transmission of signals between them. It is also involved in synaptic plasticity, which is necessary for learning and memory. The forkhead box P2 protein appears to be essential for the normal development of speech and language.
- FOXP2 mutations cause a speech and language disorder, raising interest in potential roles of this gene in human evolution.

FOXP2 Gene

Phylogenetic Tree

This is a Neighbour-joining tree without distance corrections.

Branch length: ◉ Cladogram ○ Real



sp|P58463|FOXP2_MOUSE 0.0014
sp|O15409|FOXP2_HUMAN 0.0028
sp|Q8MJA0|FOXP2_PANTR 0
sp|Q8MJ98|FOXP2_PONPY 0.0014
sp|Q8MJ99|FOXP2_GORGO 0
sp|Q8MJ97|FOXP2_MACMU 0
sp|Q8HZ00|FOXP2_PANPA 0

FOXP2 Gene

```
#
#
#  Percent Identity  Matrix - created by Clustal2.1
#
#

     1: sp|P58463|FOXP2_MOUSE  100.00   99.58   99.72   99.86   99.86   99.86   99.86
     2: sp|O15409|FOXP2_HUMAN   99.58  100.00   99.58   99.72   99.72   99.72   99.72
     3: sp|Q8MJ98|FOXP2_PONPY   99.72   99.58  100.00   99.86   99.86   99.86   99.86
     4: sp|Q8MJ99|FOXP2_GORGO   99.86   99.72   99.86  100.00  100.00  100.00  100.00
     5: sp|Q8MJ97|FOXP2_MACMU   99.86   99.72   99.86  100.00  100.00  100.00  100.00
     6: sp|Q8MJA0|FOXP2_PANTR   99.86   99.72   99.86  100.00  100.00  100.00  100.00
     7: sp|Q8HZ00|FOXP2_PANPA   99.86   99.72   99.86  100.00  100.00  100.00  100.00
```

FOXP2 Gene

```
sp|P58463|FOXP2_MOUSE      MMQESATETISNSSMNQNGMSTLSSQLDAGSRDGRSSGDTSSEVSTVELLHLQQQQALQA 60
sp|O15409|FOXP2_HUMAN      MMQESATETISNSSMNQNGMSTLSSQLDAGSRDGRSSGDTSSEVSTVELLHLQQQQALQA 60
sp|Q8MJ98|FOXP2_PONPY      MMQESVTETISNSSMNQNGMSTLSSQLDAGSRDGRSSGDTSSEVSTVELLHLQQQQALQA 60
sp|Q8MJ99|FOXP2_GORGO      MMQESATETISNSSMNQNGMSTLSSQLDAGSRDGRSSGDTSSEVSTVELLHLQQQQALQA 60
sp|Q8MJ97|FOXP2_MACMU      MMQESATETISNSSMNQNGMSTLSSQLDAGSRDGRSSGDTSSEVSTVELLHLQQQQALQA 60
sp|Q8MJA0|FOXP2_PANTR      MMQESATETISNSSMNQNGMSTLSSQLDAGSRDGRSSGDTSSEVSTVELLHLQQQQALQA 60
sp|Q8HZ00|FOXP2_PANPA      MMQESATETISNSSMNQNGMSTLSSQLDAGSRDGRSSGDTSSEVSTVELLHLQQQQALQA 60
                           *****.************************************************

sp|P58463|FOXP2_MOUSE      ARQLLLQQQTSGLKSPKSSEKQRPLQVPVSVAMMTPQVITPQQMQQILQQQVLSPQQLQA 120
sp|O15409|FOXP2_HUMAN      ARQLLLQQQTSGLKSPKSSDKQRPLQVPVSVAMMTPQVITPQQMQQILQQQVLSPQQLQA 120
sp|Q8MJ98|FOXP2_PONPY      ARQLLLQQQTSGLKSPKSSDKQRPLQVPVSVAMMTPQVITPQQMQQILQQQVLSPQQLQA 120
sp|Q8MJ99|FOXP2_GORGO      ARQLLLQQQTSGLKSPKSSDKQRPLQVPVSVAMMTPQVITPQQMQQILQQQVLSPQQLQA 120
sp|Q8MJ97|FOXP2_MACMU      ARQLLLQQQTSGLKSPKSSDKQRPLQVPVSVAMMTPQVITPQQMQQILQQQVLSPQQLQA 120
sp|Q8MJA0|FOXP2_PANTR      ARQLLLQQQTSGLKSPKSSDKQRPLQVPVSVAMMTPQVITPQQMQQILQQQVLSPQQLQA 120
sp|Q8HZ00|FOXP2_PANPA      ARQLLLQQQTSGLKSPKSSDKQRPLQVPVSVAMMTPQVITPQQMQQILQQQVLSPQQLQA 120
                           ******************:*****************************************

sp|P58463|FOXP2_MOUSE      LLQQQQAVMLQQQQLQEFYKKQQEQLHLQLLQQQQQQQQQQQQQQQQQQQQ-QQQQQQQQ 179
sp|O15409|FOXP2_HUMAN      LLQQQQAVMLQQQQLQEFYKKQQEQLHLQLLQQQQQQQQQQQQQQQQQQQQ-QQQQQQQQ 179
sp|Q8MJ98|FOXP2_PONPY      LLQQQQAVMLQQQQLQEFYKKQQEQLHLQLLQQQQQQQQQQQQQQQQQQQQ--QQQQQQQQ 178
sp|Q8MJ99|FOXP2_GORGO      LLQQQQAVMLQQQQLQEFYKKQQEQLHLQLLQQQQQQQQQQQQQQQQQQQQ---QQQQQQQQ 177
sp|Q8MJ97|FOXP2_MACMU      LLQQQQAVMLQQQQLQEFYKKQQEQLHLQLLQQQQQQQQQQQQQQQQQQQQ--QQQQQQQQ 178
sp|Q8MJA0|FOXP2_PANTR      LLQQQQAVMLQQQQLQEFYKKQQEQLHLQLLQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQ 180
sp|Q8HZ00|FOXP2_PANPA      LLQQQQAVMLQQQQLQEFYKKQQEQLHLQLLQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQ 180
                           ************************************************   ********

sp|P58463|FOXP2_MOUSE      QQQQQQQQQQQQHPGKQAKEQQQQ-QQQQQLAAQQLVFQQQLLQMQQLQQQQHLLSLQRQ 238
sp|O15409|FOXP2_HUMAN      QQQQQQQQQQQQHPGKQAKEQQQQQQQQQQLAAQQLVFQQQLLQMQQLQQQQHLLSLQRQ 239
sp|Q8MJ98|FOXP2_PONPY      QQQQQQQQQQQQHPGKQAKEQQQQ-QQQQQLAAQQLVFQQQLLQMQQLQQQQHLLSLQRQ 237
sp|Q8MJ99|FOXP2_GORGO      QQQQQQQQQQQQHPGKQAKEQQQQQQQQQQLAAQQLVFQQQLLQMQQLQQQQHLLSLQRQ 237
sp|Q8MJ97|FOXP2_MACMU      QQQQQQQQQQQQHPGKQAKEQQQQQQQQQQLAAQQLVFQQQLLQMQQLQQQQHLLSLQRQ 238
sp|Q8MJA0|FOXP2_PANTR      QQQQQQQQQQQQHPGKQAKEQQQQQQQQQQLAAQQLVFQQQLLQMQQLQQQQHLLSLQRQ 240
sp|Q8HZ00|FOXP2_PANPA      QQQQQQQQQQQQHPGKQAKEQQQQQQQQQQLAAQQLVFQQQLLQMQQLQQQQHLLSLQRQ 240
                           ************************ ************************************
```

```
sp|P58463|FOXP2_MOUSE      GLISIPPGQAALPVQSLPQAGLSPAEIQQLWKEVTGVHSMEDNGIKHGGLDLTTNNSSST 298
sp|O15409|FOXP2_HUMAN      GLISIPPGQAALPVQSLPQAGLSPAEIQQLWKEVTGVHSMEDNGIKHGGLDLTTNNSSST 299
sp|Q8MJ98|FOXP2_PONPY      GLISIPPGQAALPVQSLPQAGLSPAEIQQLWKEVTGVHSMEDNGIKHGGLDLTTNNSSST 297
sp|Q8MJ99|FOXP2_GORGO      GLISIPPGQAALPVQSLPQAGLSPAEIQQLWKEVTGVHSMEDNGIKHGGLDLTTNNSSST 297
sp|Q8MJ97|FOXP2_MACMU      GLISIPPGQAALPVQSLPQAGLSPAEIQQLWKEVTGVHSMEDNGIKHGGLDLTTNNSSST 298
sp|Q8MJA0|FOXP2_PANTR      GLISIPPGQAALPVQSLPQAGLSPAEIQQLWKEVTGVHSMEDNGIKHGGLDLTTNNSSST 300
sp|Q8HZ00|FOXP2_PANPA      GLISIPPGQAALPVQSLPQAGLSPAEIQQLWKEVTGVHSMEDNGIKHGGLDLTTNNSSST 300
                           ************************************************************

sp|P58463|FOXP2_MOUSE      TSSTTSKASPPITHHSIVNGQSSVLNARRDSSSHEETGASHTLYGHGVCKWPGCESICED 358
sp|O15409|FOXP2_HUMAN      TSSNTSKASPPITHHSIVNGQSSVLSARRDSSSHEETGASHTLYGHGVCKWPGCESICED 359
sp|Q8MJ98|FOXP2_PONPY      TSSTTSKASPPITHHSIVNGQSSVLNARRDSSSHEETGASHTLYGHGVCKWPGCESICED 357
sp|Q8MJ99|FOXP2_GORGO      TSSTTSKASPPITHHSIVNGQSSVLNARRDSSSHEETGASHTLYGHGVCKWPGCESICED 357
sp|Q8MJ97|FOXP2_MACMU      TSSTTSKASPPITHHSIVNGQSSVLNARRDSSSHEETGASHTLYGHGVCKWPGCESICED 358
sp|Q8MJA0|FOXP2_PANTR      TSSTTSKASPPITHHSIVNGQSSVLNARRDSSSHEETGASHTLYGHGVCKWPGCESICED 360
sp|Q8HZ00|FOXP2_PANPA      TSSTTSKASPPITHHSIVNGQSSVLNARRDSSSHEETGASHTLYGHGVCKWPGCESICED 360
                           ***.*******************.********************************

sp|P58463|FOXP2_MOUSE      FGQFLKHLNNEHALDDRSTAQCRVQMQVVQQLEIQLSKERERLQAMMTHLHMRPSEPKPS 418
sp|O15409|FOXP2_HUMAN      FGQFLKHLNNEHALDDRSTAQCRVQMQVVQQLEIQLSKERERLQAMMTHLHMRPSEPKPS 419
sp|Q8MJ98|FOXP2_PONPY      FGQFLKHLNNEHALDDRSTAQCRVQMQVVQQLEIQLSKERERLQAMMTHLHMRPSEPKPS 417
sp|Q8MJ99|FOXP2_GORGO      FGQFLKHLNNEHALDDRSTAQCRVQMQVVQQLEIQLSKERERLQAMMTHLHMRPSEPKPS 417
sp|Q8MJ97|FOXP2_MACMU      FGQFLKHLNNEHALDDRSTAQCRVQMQVVQQLEIQLSKERERLQAMMTHLHMRPSEPKPS 418
sp|Q8MJA0|FOXP2_PANTR      FGQFLKHLNNEHALDDRSTAQCRVQMQVVQQLEIQLSKERERLQAMMTHLHMRPSEPKPS 420
sp|Q8HZ00|FOXP2_PANPA      FGQFLKHLNNEHALDDRSTAQCRVQMQVVQQLEIQLSKERERLQAMMTHLHMRPSEPKPS 420
                           ************************************************************

sp|P58463|FOXP2_MOUSE      PKPLNLVSSVTMSKNMLETSPQSLPQTPTTPTAPVTPITQGPSVITPASVPNVGAIRRRH 478
sp|O15409|FOXP2_HUMAN      PKPLNLVSSVTMSKNMLETSPQSLPQTPTTPTAPVTPITQGPSVITPASVPNVGAIRRRH 479
sp|Q8MJ98|FOXP2_PONPY      PKPLNLVSSVTMSKNMLETSPQSLPQTPTTPTAPVTPITQGPSVITPASVPNVGAIRRRH 477
sp|Q8MJ99|FOXP2_GORGO      PKPLNLVSSVTMSKNMLETSPQSLPQTPTTPTAPVTPITQGPSVITPASVPNVGAIRRRH 477
sp|Q8MJ97|FOXP2_MACMU      PKPLNLVSSVTMSKNMLETSPQSLPQTPTTPTAPVTPITQGPSVITPASVPNVGAIRRRH 478
sp|Q8MJA0|FOXP2_PANTR      PKPLNLVSSVTMSKNMLETSPQSLPQTPTTPTAPVTPITQGPSVITPASVPNVGAIRRRH 480
sp|Q8HZ00|FOXP2_PANPA      PKPLNLVSSVTMSKNMLETSPQSLPQTPTTPTAPVTPITQGPSVITPASVPNVGAIRRRH 480
                           ************************************************************
```

FOXP2 Gene

```
sp|P58463|FOXP2_MOUSE      SDKYNIPMSSEIAPNYEFYKNADVRPPFTYATLIRQAIMESSDRQLTLNEIYSWFTRTFA 538
sp|O15409|FOXP2_HUMAN      SDKYNIPMSSEIAPNYEFYKNADVRPPFTYATLIRQAIMESSDRQLTLNEIYSWFTRTFA 539
sp|Q8MJ98|FOXP2_PONPY      SDKYNIPMSSEIAPNYEFYKNADVRPPFTYATLIRQAIMESSDRQLTLNEIYSWFTRTFA 537
sp|Q8MJ99|FOXP2_GORGO      SDKYNIPMSSEIAPNYEFYKNADVRPPFTYATLIRQAIMESSDRQLTLNEIYSWFTRTFA 537
sp|Q8MJ97|FOXP2_MACMU      SDKYNIPMSSEIAPNYEFYKNADVRPPFTYATLIRQAIMESSDRQLTLNEIYSWFTRTFA 538
sp|Q8MJA0|FOXP2_PANTR      SDKYNIPMSSEIAPNYEFYKNADVRPPFTYATLIRQAIMESSDRQLTLNEIYSWFTRTFA 540
sp|Q8HZ00|FOXP2_PANPA      SDKYNIPMSSEIAPNYEFYKNADVRPPFTYATLIRQAIMESSDRQLTLNEIYSWFTRTFA 540
                           ************************************************************

sp|P58463|FOXP2_MOUSE      YFRRNAATWKNAVRHNLSLHKCFVRVENVKGAVWTVDEVEYQKRRSQKITGSPTLVKNIP 598
sp|O15409|FOXP2_HUMAN      YFRRNAATWKNAVRHNLSLHKCFVRVENVKGAVWTVDEVEYQKRRSQKITGSPTLVKNIP 599
sp|Q8MJ98|FOXP2_PONPY      YFRRNAATWKNAVRHNLSLHKCFVRVENVKGAVWTVDEVEYQKRRSQKITGSPTLVKNIP 597
sp|Q8MJ99|FOXP2_GORGO      YFRRNAATWKNAVRHNLSLHKCFVRVENVKGAVWTVDEVEYQKRRSQKITGSPTLVKNIP 597
sp|Q8MJ97|FOXP2_MACMU      YFRRNAATWKNAVRHNLSLHKCFVRVENVKGAVWTVDEVEYQKRRSQKITGSPTLVKNIP 598
sp|Q8MJA0|FOXP2_PANTR      YFRRNAATWKNAVRHNLSLHKCFVRVENVKGAVWTVDEVEYQKRRSQKITGSPTLVKNIP 600
sp|Q8HZ00|FOXP2_PANPA      YFRRNAATWKNAVRHNLSLHKCFVRVENVKGAVWTVDEVEYQKRRSQKITGSPTLVKNIP 600
                           ************************************************************

sp|P58463|FOXP2_MOUSE      TSLGYGAALNASLQAALAESSLPLLSNPGLINNASSGLLQAVHEDLNGSLDHIDSNGNSS 658
sp|O15409|FOXP2_HUMAN      TSLGYGAALNASLQAALAESSLPLLSNPGLINNASSGLLQAVHEDLNGSLDHIDSNGNSS 659
sp|Q8MJ98|FOXP2_PONPY      TSLGYGAALNASLQAALAESSLPLLSNPGLINNASSGLLQAVHEDLNGSLDHIDSNGNSS 657
sp|Q8MJ99|FOXP2_GORGO      TSLGYGAALNASLQAALAESSLPLLSNPGLINNASSGLLQAVHEDLNGSLDHIDSNGNSS 657
sp|Q8MJ97|FOXP2_MACMU      TSLGYGAALNASLQAALAESSLPLLSNPGLINNASSGLLQAVHEDLNGSLDHIDSNGNSS 658
sp|Q8MJA0|FOXP2_PANTR      TSLGYGAALNASLQAALAESSLPLLSNPGLINNASSGLLQAVHEDLNGSLDHIDSNGNSS 660
sp|Q8HZ00|FOXP2_PANPA      TSLGYGAALNASLQAALAESSLPLLSNPGLINNASSGLLQAVHEDLNGSLDHIDSNGNSS 660
                           ************************************************************

sp|P58463|FOXP2_MOUSE      PGCSPQPHIHSIHVKEEPVIAEDEDCPMSLVTTANHSPELEDDREIEEEPLSEDLE     714
sp|O15409|FOXP2_HUMAN      PGCSPQPHIHSIHVKEEPVIAEDEDCPMSLVTTANHSPELEDDREIEEEPLSEDLE     715
sp|Q8MJ98|FOXP2_PONPY      PGCSPQPHIHSIHVKEEPVIAEDEDCPMSLVTTANHSPELEDDREIEEEPLSEDLE     713
sp|Q8MJ99|FOXP2_GORGO      PGCSPQPHIHSIHVKEEPVIAEDEDCPMSLVTTANHSPELEDDREIEEEPLSEDLE     713
sp|Q8MJ97|FOXP2_MACMU      PGCSPQPHIHSIHVKEEPVIAEDEDCPMSLVTTANHSPELEDDREIEEEPLSEDLE     714
sp|Q8MJA0|FOXP2_PANTR      PGCSPQPHIHSIHVKEEPVIAEDEDCPMSLVTTANHSPELEDDREIEEEPLSEDLE     716
sp|Q8HZ00|FOXP2_PANPA      PGCSPQPHIHSIHVKEEPVIAEDEDCPMSLVTTANHSPELEDDREIEEEPLSEDLE     716
                           ************************************************************
```

BRCA1 Gene

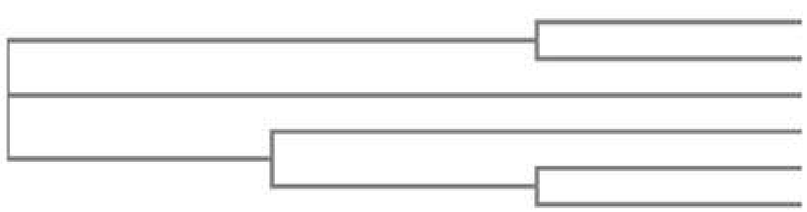
- Provides instructions for making a protein that acts as a tumor suppressor. Tumor suppressor proteins help prevent cells from growing and dividing too rapidly or in an uncontrolled way.
- The BRCA1 protein is involved in repairing damaged DNA. In the nucleus of many types of normal cells, the BRCA1 protein interacts with several other proteins to mend breaks in DNA.
- By helping to repair DNA, the BRCA1 protein plays a critical role in maintaining the stability of a cell's genetic information.

BRCA1 Gene

Phylogenetic Tree

This is a Neighbour-joining tree without distance corrections.

Branch length: ◉ Cladogram ○ Real

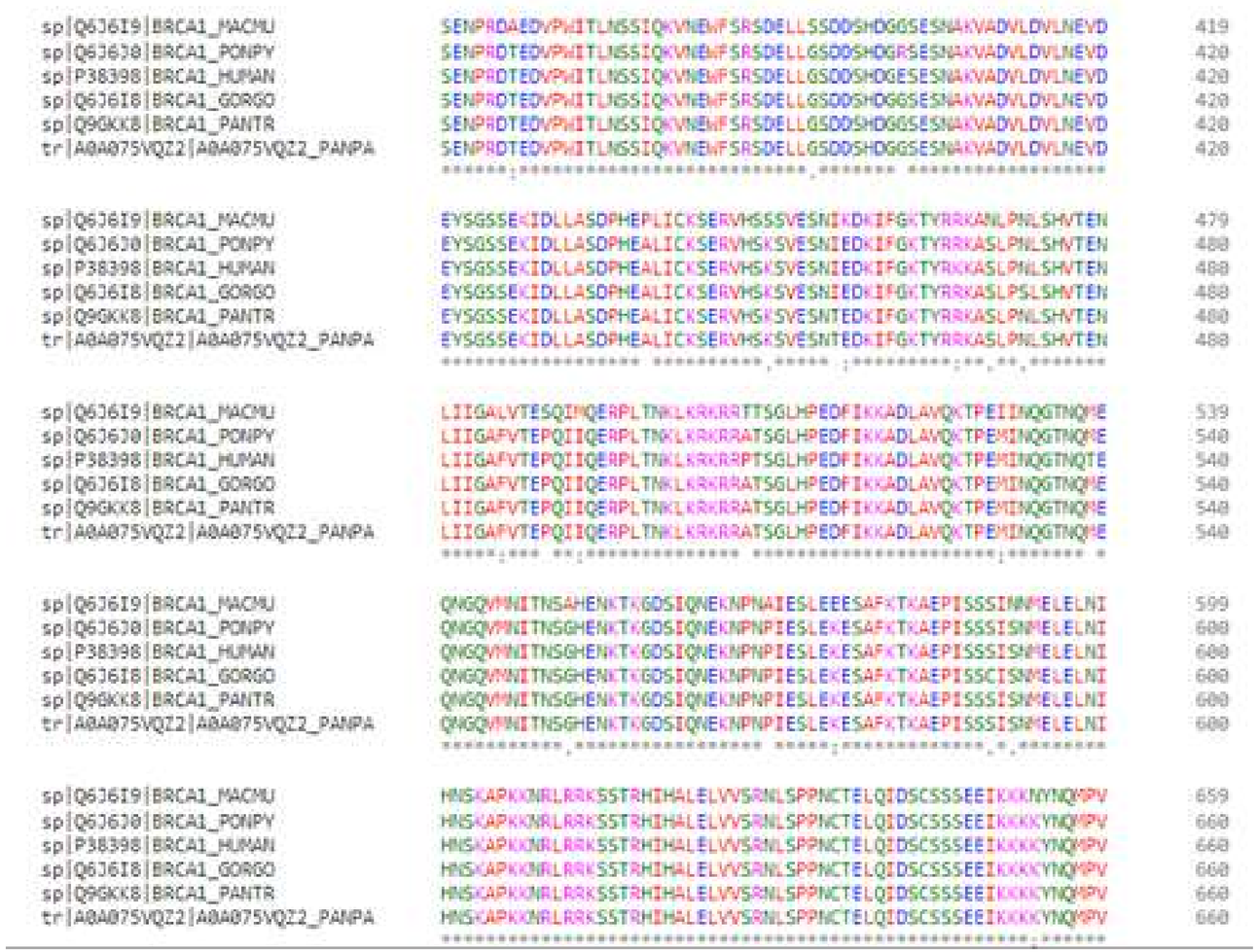


sp|Q6J6I9|BRCA1_MACMU 0.05105
sp|Q6J6J0|BRCA1_PONPY 0.0134
sp|P38398|BRCA1_HUMAN 0.01136
sp|Q6J6I8|BRCA1_GORGO 0.008
sp|Q9GKK8|BRCA1_PANTR 0
tr|A0A075VQZ2|A0A075VQZ2_PANPA 0.00108

BRCA1 Gene

```
#
#  Percent Identity  Matrix - created by Clustal2.1
#
#

     1: sp|Q6J6I9|BRCA1_MACMU          100.00   93.56   93.13   93.39   93.77   93.69
     2: sp|Q6J6J0|BRCA1_PONPY           93.56  100.00   96.83   97.16   97.58   97.47
     3: sp|P38398|BRCA1_HUMAN           93.13   96.83  100.00   98.01   98.44   98.33
     4: sp|Q6J6I8|BRCA1_GORGO           93.39   97.16   98.01  100.00   98.82   98.71
     5: sp|Q9GKK8|BRCA1_PANTR           93.77   97.58   98.44   98.82  100.00   99.89
     6: tr|A0A075VQZ2|A0A075VQZ2_PANPA  93.69   97.47   98.33   98.71   99.89  100.00
```

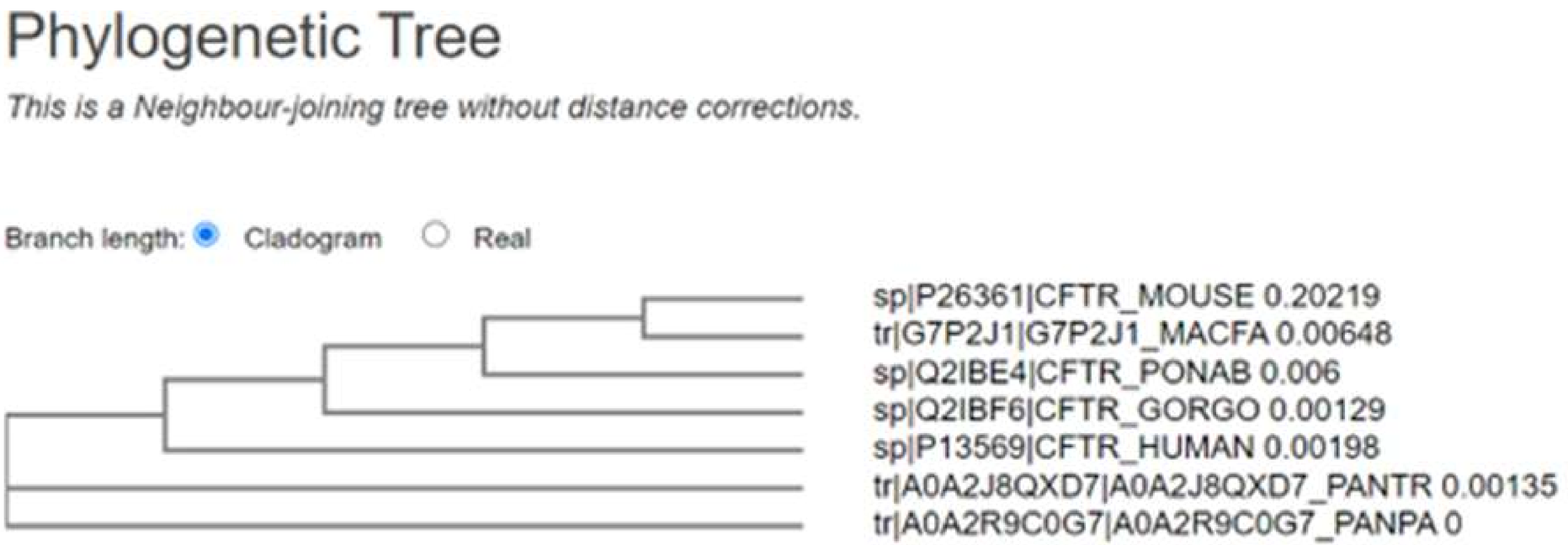
BRCA1 Gene


sp|Q6J6I9|BRCA1_MACMU SENPRDAEDVPWITLNSSIQKVNEWFSRSDELLSSDDSHDGGSESNAKVADVLDVLNEVD 419
sp|Q6J6J0|BRCA1_PONPY SENPRDTEDVPWITLNSSIQKVNEWFSRSDELLGSDDSHDGRSESNAKVADVLDVLNEVD 420
sp|P38398|BRCA1_HUMAN SENPRDTEDVPWITLNSSIQKVNEWFSRSDELLGSDDSHDGESESNAKVADVLDVLNEVD 420
sp|Q6J6I8|BRCA1_GORGO SENPRDTEDVPWITLNSSIQKVNEWFSRSDELLGSDDSHDGGSESNAKVADVLDVLNEVD 420
sp|Q9GKK8|BRCA1_PANTR SENPRDTEDVPWITLNSSIQKVNEWFSRSDELLGSDDSHDGGSESNAKVADVLDVLNEVD 420
tr|A0A075VQZ2|A0A075VQZ2_PANPA SENPRDTEDVPWITLNSSIQKVNEWFSRSDELLGSDDSHDGGSESNAKVADVLDVLNEVD 420
EYSGSSEKIDLLASDPHEPLICKSERVHSSSVESNIKDKIFGKTYRRKANLPNLSHVTEN 479
EYSGSSEKIDLLASDPHEALICKSERVHSKSVESNIEDKIFGKTYRRKASLPNLSHVTEN 480
EYSGSSEKIDLLASDPHEALICKSERVHSKSVESNIEDKIFGKTYRRKASLPSLSHVTEN 480
EYSGSSEKIDLLASDPHEALICKSERVHSKSVESNTEDKIFGKTYRRKASLPNLSHVTEN 480
LIIGALVTESQIMQERPLTNKLKRKRRTTSGLHPEDFIKKADLAVQKTPEIINQGTNQME 539
LIIGAFVTEPQIIQERPLTNKLKRKRRATSGLHPEDFIKKADLAVQKTPEMINQGTNQTE 540
LIIGAFVTEPQIIQERPLTNKLKRKRRPTSGLHPEDFIKKADLAVQKTPEMINQGTNQTE 540
LIIGAFVTEPQIIQERPLTNKLKRKRRATSGLHPEDFIKKADLAVQKTPEMINQGTNQME 540
LIIGAFVTEPQIIQERPLTNKLKRKRRATSGLHPEDFIKKADLAVQKTPEMINQGTNQME 540
QMGQVMNITNSAHENKTKGDSIQNEKNPNAIESLEEESAFKTKAEPISSSIMNMELELNI 599
QMGQVMNITNSGHENKTKGDSIQNEKNPNPIESLEKESAFKTKAEPISSSISNMELELNI 600
QMGQVMNITNSGHENKTKGDSIQNEKNPNPIESLEKESAFKTKAEPISSSISNMELELNI 600
QMGQVMNITNSGHENKTKGDSIQNEKNPNPIESLEKESAFKTKAEPISSCISNMELELNI 600
QMGQVMNITNSGHENKTKGDSIQNEKNPNPIESLEKESAFKTKAEPISSSISNMELELNI 600
QMGQVMNITNSGHENKTKGDSIQNEKNPNPIESLEKESAFKTKAEPISSSISNMELELNI 600
MNSKAPKKNRLRRKSSTRHIHALELVVSRNLSPPNCTELQIDSCSSSEEIKKKNYNQMPV 659
MNSKAPKKNRLRRKSSTRHIHALELVVSRNLSPPNCTELQIDSCSSSEEIKKKKYNQMPV 660
MNSKAPKKNRLRRKSSTRHIHALELVVSRNLSPPNCTELQIDSCSSSEEIKKKKYNQMPV 660
MNSKAPKKNRLRRKSSTRHIHALELVVSRNLSPPNCTELQIDSCSSSEEIKKKKYNQMPV 660
MNSKAPKKNRLRRKSSTRHIHALELVVSRNLSPPNCTELQIDSCSSSEEIKKKKYNQMPV 660

CFTR Gene

- It provides instructions for making a protein called the CF transmembrane conductance regulator (CFTR).
- This protein functions as a channel across the membrane of cells that produce mucus, sweat, saliva, tears, and digestive enzymes.
- The channel transports negatively charged particles called chloride ions into and out of cells. The transport of chloride ions helps control the movement of water in tissues, which is necessary for the freely flowing mucus.
- Mucus is a slippery substance that lubricates and protects the lining of the airways.

CFTR Gene

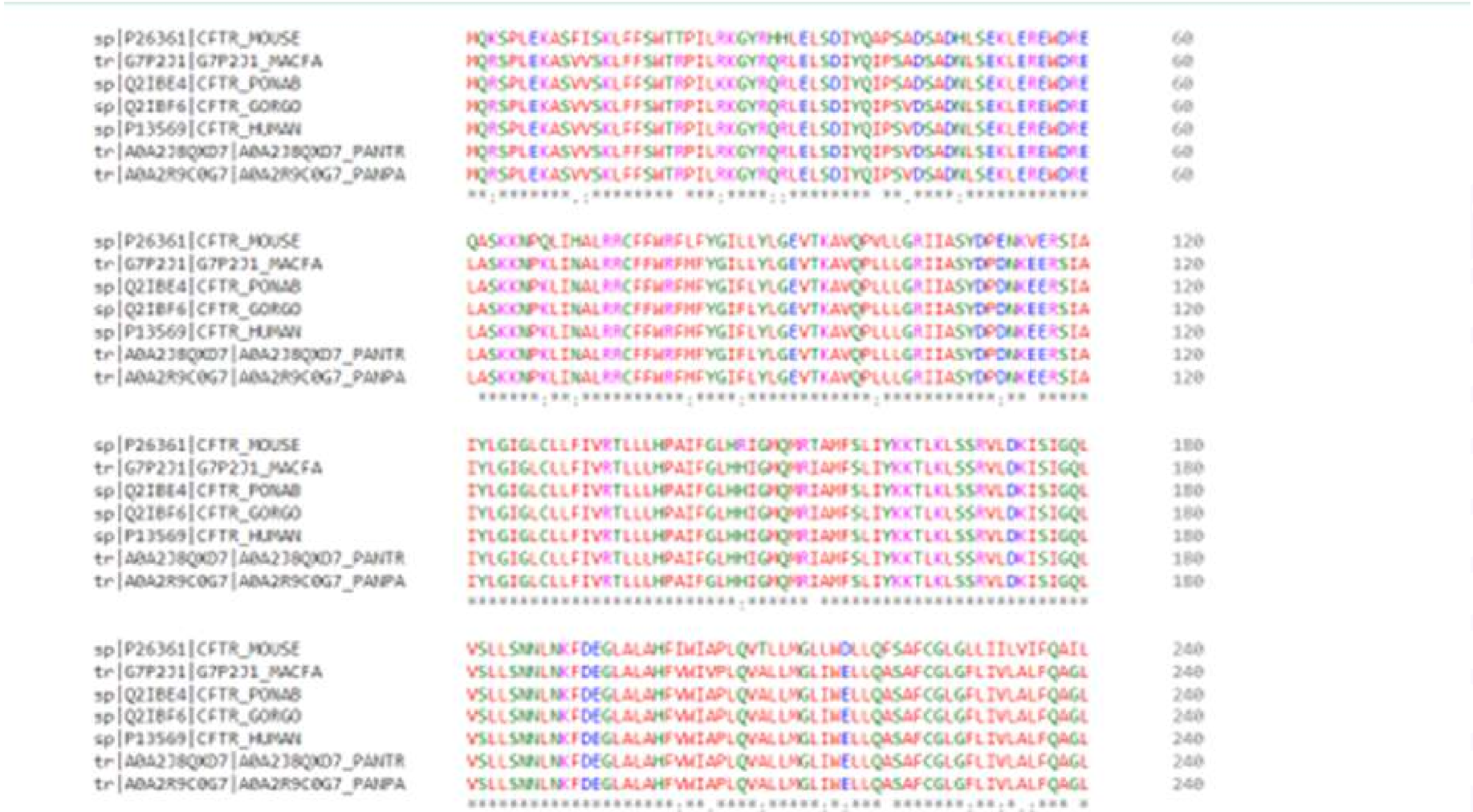

Phylogenetic Tree
This is a Neighbour-joining tree without distance corrections.
Branch length: Cladogram Real
sp|P26361|CFTR_MOUSE 0.20219
tr|G7P2J1|G7P2J1_MACFA 0.00648
sp|Q2IBE4|CFTR_PONAB 0.006
sp|Q2IBF6|CFTR_GORGO 0.00129
sp|P13569|CFTR_HUMAN 0.00198
tr|A0A2J8QXD7|A0A2J8QXD7_PANTR 0.00135
tr|A0A2R9C0G7|A0A2R9C0G7_PANPA 0

CFTR Gene

```
#
#
#  Percent Identity  Matrix - created by Clustal2.1
#
#
```

1: sp|P26361|CFTR_MOUSE	100.00	79.13	78.71	78.85	78.71	78.71	78.85
2: tr|G7P2J1|G7P2J1_MACFA	79.13	100.00	98.04	98.45	98.31	98.38	98.51
3: sp|Q2IBE4|CFTR_PONAB	78.71	98.04	100.00	99.05	98.92	98.99	99.12
4: sp|Q2IBF6|CFTR_GORGO	78.85	98.45	99.05	100.00	99.59	99.66	99.80
5: sp|P13569|CFTR_HUMAN	78.71	98.31	98.92	99.59	100.00	99.66	99.80
6: tr|A0A2J8QXD7|A0A2J8QXD7_PANTR	78.71	98.38	98.99	99.66	99.66	100.00	99.86
7: tr|A0A2R9C0G7|A0A2R9C0G7_PANPA	78.85	98.51	99.12	99.80	99.80	99.86	100.00

CFTR Gene


sp|P26361|CFTR_MOUSE
tr|G7P2J1|G7P2J1_MACFA
sp|Q2IBE4|CFTR_PONAB
sp|Q2IBF6|CFTR_GORGO
sp|P13569|CFTR_HUMAN
tr|A0A2J8QXD7|A0A2J8QXD7_PANTR
tr|A0A2R9C0G7|A0A2R9C0G7_PANPA
60
120
180
240

APOE GENE

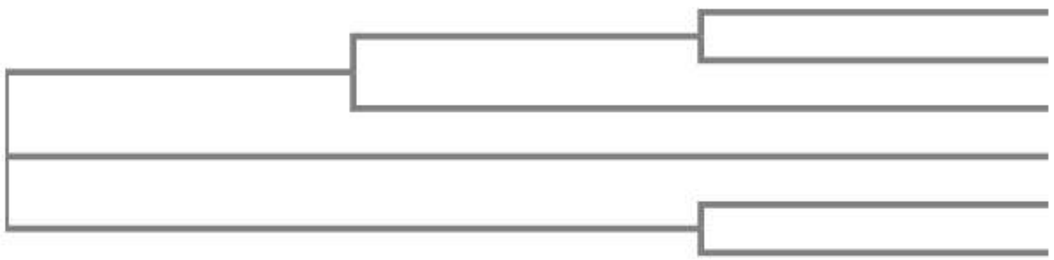
- The APOE gene, which stands for Apolipoprotein E, encodes a protein that plays a crucial role in the metabolism of lipids (fats) in the body.
- The APOE gene is involved in the transport and redistribution of lipids in the body, particularly cholesterol.
- The protein produced by the APOE gene is part of lipoproteins, which are responsible for carrying cholesterol and other fats in the bloodstream. The different APOE alleles affect the binding affinity of the protein for these lipoproteins, which, in turn, influences how cholesterol is transported and utilized in the body.
- It's important to note that the APOE gene's role in health and disease is complex and influenced by many factors, including interactions with other genes and environmental factors. Genetic testing for APOE status can provide information about an individual's potential risk factors for certain health conditions, but it cannot predict with certainty whether or not someone will develop a specific disease.

APOE GENE

Phylogenetic Tree

This is a Neighbour-joining tree without distance corrections.

Branch length: ◉ Cladogram ○ Real



sp|P08226|APOE_MOUSE 0.24472
sp|P10517|APOE_MACFA 0.0336
sp|Q9GLM7|APOE_PONPY 0.00765
sp|P02649|APOE_HUMAN 0.01718
sp|Q9GJU3|APOE_PANTR 0.005
sp|Q9GLM8|APOE_GORGO 0.00446

APOE GENE

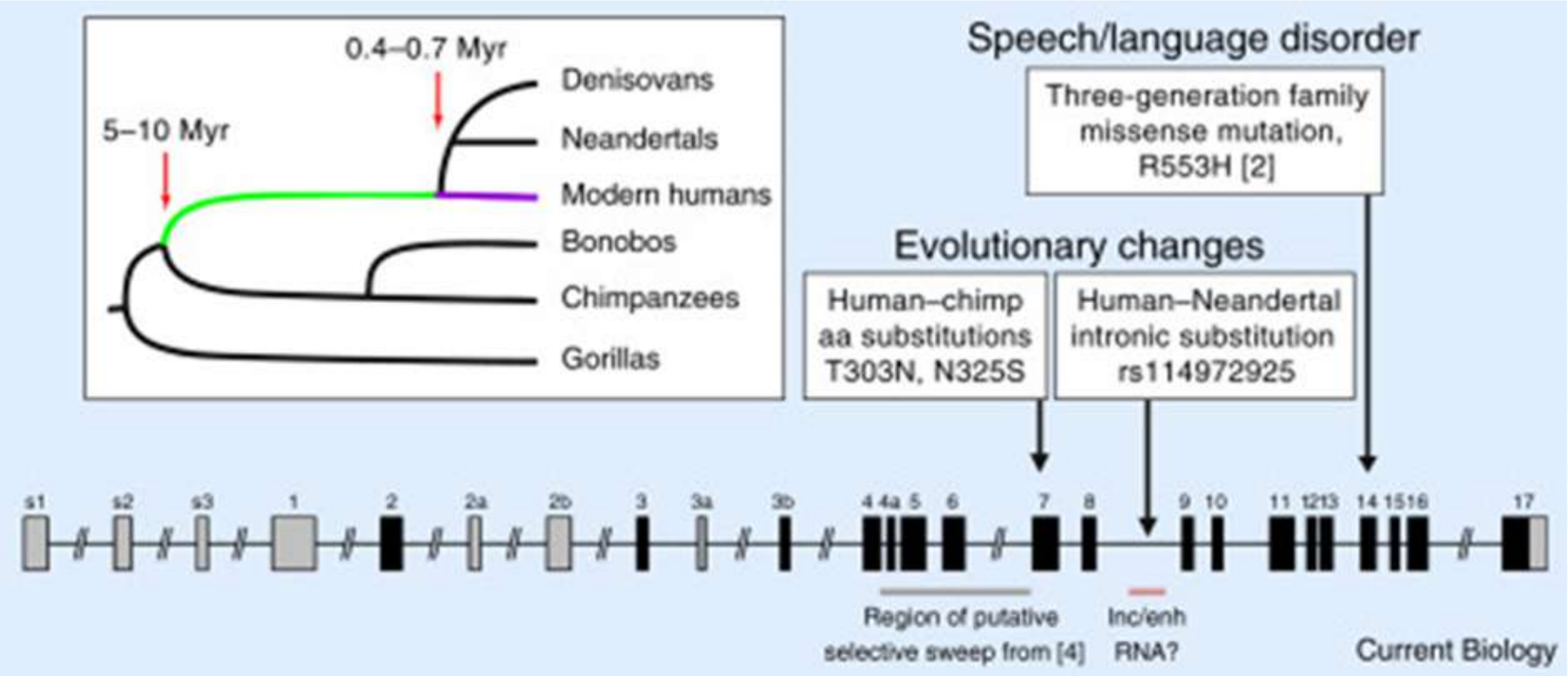
```
#
#
#  Percent Identity  Matrix - created by Clustal2.1
#
#

     1: sp|P08226|APOE_MOUSE  100.00   72.17   73.14   72.17   72.82   73.14
     2: sp|P10517|APOE_MACFA   72.17  100.00   94.32   93.38   94.01   94.01
     3: sp|Q9GLM7|APOE_PONPY   73.14   94.32  100.00   97.48   98.42   98.11
     4: sp|P02649|APOE_HUMAN   72.17   93.38   97.48  100.00   97.16   97.48
     5: sp|Q9GJU3|APOE_PANTR   72.82   94.01   98.42   97.16  100.00   99.05
     6: sp|Q9GLM8|APOE_GORGO   73.14   94.01   98.11   97.48   99.05  100.00
```

APOE GENE

```
sp|P08226|APOE_MOUSE      KAGAREGAERGVSAIRERLGPLVEQGRQRTANLGAGAAQPLRDRAQAFGDRIRGRLEEVG  232
sp|P10517|APOE_MACFA      QAGAREGAERGVSAIRERLGPLVEQGRVRAATVGSLASQPLQERAQALGERLRARMEEMG  240
sp|Q9GLM7|APOE_PONPY      QAGAREGAERGVSAIRERLGPLVEQGRVRAATVGSVAGKPLQERAQAWGERLRARMEEMG  240
sp|P02649|APOE_HUMAN      QAGAREGAERGLSAIRERLGPLVEQGRVRAATVGSLAGQPLQERAQAWGERLRARMEEMG  240
sp|Q9GJU3|APOE_PANTR      QAGAREGAERGVSAIRERLGPLVEQGRVRAATVGSLAGQPLQERAQAWGERLRARMEEMG  240
sp|Q9GLM8|APOE_GORGO      QAGAREGAERGVSAIRERLGPLVEQGRVRAATVGSLAGQPLQERAQAWGERLRARMEEMG  240
                          :**********:**************** *:*.:*: *.:**::**** *:*:*.*:**:*

sp|P08226|APOE_MOUSE      NQARDRLEEVREHMEEVRSKMEEQTQQIRLQAEIFQARLKGWFEPIVEDMHRQWANLMEK  292
sp|P10517|APOE_MACFA      SRTRDRLDEVKEQVAEVRAKLEEQAQQISLQAEAFQARLKSWFEPLVEDMQRQWAGLVEK  300
sp|Q9GLM7|APOE_PONPY      SRTRDRLDEVKEQVAEVRAKLEEQAQQIRLQAEAFQARLKSWFEPLVEDMQRQWAGLVEK  300
sp|P02649|APOE_HUMAN      SRTRDRLDEVKEQVAEVRAKLEEQAQQIRLQAEAFQARLKSWFEPLVEDMQRQWAGLVEK  300
sp|Q9GJU3|APOE_PANTR      SRTRDRLDEVKEQVAEVRAKLEEQAQQIRLQAEAFQARLKSWFEPLVEDMQRQWAGLVEK  300
sp|Q9GLM8|APOE_GORGO      SRTRDRLDEVKEQVAEVRAKLEEQAQQIRLQAEAFQARLKSWFEPLVEDMQRQWAGLVEK  300
                          .::****:**:*:: ***:*:***:*** **** ******.****:****:****.*:**

sp|P08226|APOE_MOUSE      IQASVATNPIITPVAQENQ   311
sp|P10517|APOE_MACFA      VQAAVGASTAPVPIDNH--   317
sp|Q9GLM7|APOE_PONPY      VQAAVGTSAAPVPSDNH--   317
sp|P02649|APOE_HUMAN      VQAAVGTSAAPVPSDNH--   317
sp|Q9GJU3|APOE_PANTR      VQAAMGTSAAPVPSDNH--   317
sp|Q9GLM8|APOE_GORGO      VQAAMGTSAAPVPSDNH--   317
                          :**::.:.   .*  :.
```

Conclusion

FOXP2



Conclusion

BRCA1

BRCA1, we analyzed complete BRCA1 gene sequences from 6 primate species. We show that specific amino acid sites have experienced repeated selection for amino acid replacement over primate evolution. This selection has been focused specifically on humans and our closest living relatives, chimpanzees (Pan troglodytes) and bonobos (Pan paniscus). After examining BRCA1 polymorphisms in bonobo, chimpanzee, and rhesus macaque (Macaca mulatta) individuals, we find considerable variation within each of these species and evidence for recent selection in chimpanzee populations. Finally, we also sequenced and analyzed BRCA2 from 6 primate species and find that this gene has also evolved under positive selection.

Conclusion

CFTR

- Mouse: The matrix data indicates that the mouse CFTR sequence has lower identity values with the primate species, reflecting a more distant genetic relationship.
- Chimpanzee, Bonobo, and Orangutan: These species are expected to have closer genetic relationships in the tree, with shorter branch lengths due to their shared common ancestry.
- Gorilla: Gorillas are expected to be closer to humans in the phylogenetic tree than to the more distant mouse.
- Rhesus Monkeys: The matrix data might show moderate to high identity values between rhesus monkeys and humans, reflecting the parallel findings in disease risk associations.

Conclusion

APOE

- Mouse: The matrix data indicates that the mouse CFTR sequence has lower identity values with the primate species, reflecting a more distant genetic relationship.
- Chimpanzee, Bonobo: These species are expected to have closer genetic relationships in the tree, with shorter branch lengths due to their shared common ancestry.
- Gorilla: Gorillas are expected to be closer to humans in the phylogenetic tree than to the more distant mouse.
- Some associations between APOE alleles and disease risk in rhesus monkeys parallel findings in humans, making them a useful model for human health research.


Thank you!

VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

Fall Semester 2023

Genomics and Proteomics

Review 3
November 4,2023

Perspectives of Informatics for Human Genome Project - Comparative Genomics of Primate and Human Genome

Professor: Dr.Kamalanathan A S

Submitted by
Ananya Arora (20BCB0031)
Bhavana Jain (20BCB0105)
Saloni Vyas (20BCB0064)
Sanjna Subramanian (20BCB0088)

Our Aim

The aim of the project is to leverage informatics and computational approaches to conduct a comprehensive comparative genomics analysis of the primate and human genomes within the context of the Human Genome Project (HGP). This comparative genomics initiative seeks to explore, understand, and elucidate the genetic similarities and differences between humans and primates, with a focus on a variety of aspects, including evolution, functional genomics, and potential implications for human health and biology.

Introduction to Human Genome Project

The Human Genome Project (HGP) was an international scientific research initiative that aimed to map and sequence the entire human genome, which is the complete set of an individual's DNA, including all genes and non-coding sequences. The project was officially launched in 1990 and involved a collaboration between multiple countries and research institutions. It had two primary goals:

1. **Mapping the Human Genome**: The first goal of the HGP was to create a comprehensive genetic map of the human genome. This involved identifying and locating the positions of specific genes and genetic markers on each of the 23 pairs of human chromosomes.

2. **Sequencing the Human Genome**: The second goal was to determine the precise order of the 3 billion base pairs in the human genome. This was a massive and technically challenging task, as it required determining the sequence of the DNA letters (A, T, C, and G) that make up human DNA.

The Human Genome Project was completed ahead of schedule and officially declared finished in April 2003. **The results of the HGP** have had a profound impact on various fields of science and medicine. Some key findings and results include:

1. **Gene Identification:** The HGP identified and mapped approximately 20,000-25,000 protein-coding genes in the human genome. This information has been invaluable for understanding the genetic basis of human traits and diseases.

2. **Non-Coding DNA**: It revealed that a significant portion of the human genome consists of non-coding DNA, which was previously considered "junk DNA." While some non-coding regions have regulatory functions, others are still not fully understood.

3. **Human Genetic Variation:** The project also helped identify genetic variations among individuals. This information has been crucial in the study of human diversity, evolution, and susceptibility to diseases.

4. **Medical Advances:** The HGP has greatly accelerated research in genetics and genomics, leading to significant advances in personalized medicine, the diagnosis and treatment of genetic disorders, and our understanding of diseases like cancer, diabetes, and cardiovascular conditions.

5. **Ethical, Legal, and Social Implications (ELSI)**: The project recognized the need to address ethical, legal, and social issues related to genetics and genomics. It established an ELSI program to explore and address these concerns, including privacy, discrimination, and informed consent.

In summary, the Human Genome Project was a groundbreaking scientific endeavor that provided a comprehensive map and sequence of the human genome. Its results have had far-reaching implications in genetics, medicine, and our understanding of human biology, while also raising important ethical and societal questions related to the use of genetic information.

Intro to Comparative Genomics

Comparative genomics is a field of biological research that involves comparing the genomes of different species to understand their similarities and differences. It is a powerful tool for studying the evolution, function, and organization of genes and other genomic elements. Here are some key aspects of comparative genomics:

- **Genome Sequencing:** Advances in DNA sequencing technology have made it possible to determine the complete DNA sequences of many organisms, ranging from bacteria to plants, animals, and humans. Comparative genomics relies on these complete genome sequences.
- **Evolutionary Relationships:** Comparative genomics helps scientists infer the evolutionary relationships between species by examining similarities and differences in their genomes. This can reveal common ancestry and divergence points.
- **Gene Function and Conservation:** By comparing genomes, researchers can identify genes that are conserved across different species. These conserved

genes often have essential functions that have been preserved throughout evolution.

- **Gene Families:** Comparative genomics helps in identifying gene families, which are groups of genes that share a common ancestor. These gene families can expand or contract in different species, and the variations can provide insights into the evolution of specific functions or traits.
- **Functional Annotations:** Comparative genomics can be used to annotate genes and non-coding regions in a genome. By comparing genes to known sequences and studying their arrangements, scientists can make educated guesses about their functions.
- **Identification of Regulatory Elements:** Comparative genomics can help identify regulatory elements in the genome, such as promoters and enhancers, by identifying conserved regions near genes that are likely to be involved in gene expression.
- **Structural Variations:** Researchers can also compare genomes to study structural variations, including duplications, deletions, inversions, and translocations of genomic regions. These structural variations can impact the evolution and function of organisms.
- **Disease Studies:** Comparative genomics is important in understanding the genetic basis of diseases. By comparing the genomes of healthy and affected individuals or species, researchers can identify genetic variations associated with diseases.
- **Genomic Adaptations**: It can reveal how species have adapted to different environments, lifestyles, and ecological niches by examining the genomic changes that underlie these adaptations.
- **Drug Discovery and Biotechnology:** Comparative genomics can be used in drug discovery and biotechnology to identify potential drug targets, study pathogen genomes, and engineer organisms for specific purposes.

Overall, comparative genomics provides valuable insights into the diversity of life on Earth, the mechanisms of evolution, and the functional aspects of genes and genomes across different species. It has numerous applications in biology, genetics, and medicine.

Literature Review

S. No.	Name	Author, Year	Learnings, Summary
1	The Human Genome Project—An Overview	David R. Bentley, 2000	The human genome sequence will underpin human biology and medicine in the next century, providing a single, essential reference to all genetic information. The availability of a reference sequence of the genome provides the basis for studying the nature of sequence variation. By comparing corresponding genomic sequences in different species, regions that have been highly conserved during evolution can be identified, many of which reflect conserved functions such as gene regulation. This paper gives an overview of the scale of the human genome project and the methodologies used. It further talks about the various implications and applications of the generated human genome, like
2	Comparative Primate Genomics	Wolfgang Enard and Svante Paabo, 2004	The comparison of the human genome to the genomes of species closely related to humans allows the identification of genomic features that set primates apart from other mammals and of features that set certain primates apart from other primates.
3	The Human Genome Project	Maynard V. Olson, 1993	The Human Genome Project in the United States, despite being only in its third year of substantial funding, has made substantial progress towards its central goals. The project's initial policy has proven to be forward-thinking, even amidst rapid technological changes. While the experimental methods employed have evolved from the project's inception, its core conceptual framework remains largely intact. The paper acknowledges that the Human Genome Project has led to various biological advancements, although establishing direct cause-and-effect relationships with specific advances is challenging. It's emphasized that this project is an internationally coordinated endeavor with contributions from many countries and diverse funding sources. Success in the project is seen when it aligns with existing trends in science, such as PCR, yeast genetics, and fluorescence microscopy, whereas attempts to introduce

			new trends have been less successful. In conclusion, the project has made significant progress in its mapping goals and has positively impacted biomedical research. Its future focus will be on sequencing the largely unexplored 99% of the human genome.
4	The Promise of Comparative Genomics in Mammals	Stephen J. O'Brien, Marilyn Menotti-Raymond, 1999	The conclusion of the research emphasizes the growing significance of comparative genomics in domesticated livestock and companion animals. In the past, this field played a secondary role to genetic advances in human and model organisms but is now gaining prominence due to improved technologies. The paper highlights several practical applications for these dense gene maps, including the creation of animal models for human genetic diseases based on gene homology. This enables the monitoring of disease progression and therapeutic testing. Comparative genomics also offers opportunities to identify polygenes affecting both human and veterinary diseases, assess multifactorial traits, and discover adaptations in mammals that could inform gene therapy. Additionally, it provides a bridge for translating human trial-based treatments to veterinary pathologies. Ultimately, comparative genomics opens the door to various advancements in understanding and treating diseases in both human and animal populations.
5	Comparative genomics approaches to study organism similarities and differences	Liping Wei, Yueyi Liu,Inna Dubchak,John Shon, and John Parka,2002	Comparative studies can be performed at different levels of the genomes to obtain multiple perspectives about the organisms. Discussed in detail the type of analyses that offer significant biological insights in the comparisons of (1) genome structure including overall genome statistics, repeats, genome rearrangement at both DNA and gene level, synteny, and breakpoints; (2) coding regions including gene content, protein content, orthologs, and paralogs; and (3) noncoding regions including the prediction of regulatory elements.Briefly review the currently available computational tools in comparative genomics such as algorithms for genome-scale sequence alignment, gene identification, and nonhomology-based function prediction.

6	Sequencing Primate Genomes: What Have We Learned?		We summarize the progress in whole-genome sequencing and analyses of primate genomes. This includes the characterization of genome structural variation, episodic changes in the repeat landscape, differences in gene expression, new models regarding speciation, and the ephemeral nature of the recombination landscape. The functional characterization of genomic differences important in primate speciation and adaptation remains a significant challenge.Next-generation sequencing technologies promise to greatly expand the number of available primate genome sequences; however, such draft genome sequences will likely miss critical genetic differences within complex genomic regions unless dedicated efforts are put forward to understand the full spectrum of genetic variation.
7	Comparative primate genomics: emerging patterns of genome content and dynamics	Jeffrey Rogers and Richard A. Gibbs	The paper discusses the state of comparative primate genomics, emphasizing its rapid growth and the valuable insights it has provided regarding the evolution of the human genome and the use of non-human primates in biomedical research. Here are some key learnings and conclusions from the passage: 1. Insights into Human Genome Evolution: One of the major impacts of comparative primate genomics is its contribution to our understanding of the history and mechanisms of human genome evolution. It has revealed evidence of complex genetic divergence and exchange among ancestral evolutionary lineages. 2. Role of Non-Human Primates in Biomedical Research: Non-human primate genomics is expanding the scope of biomedical research by providing innovative analyses using primate models of human diseases. This offers valuable insights into human health and diseases. 3. Indispensable Role of Non-Human Primates: Non-human primates are described as indispensable resources for comparative and experimental studies. They play a crucial role in understanding the origin of humans and elucidating the genetic basis of human diseases.

8	Global discovery of primate-specific genes in the human genome	Sen-Kwan Taya, Jason Blytheb, and Leonard Lipovicha,	The paper discusses the identification and characteristics of primate-specific genes (TUs) and their potential roles in primate evolution and biology. Here are some key learnings and conclusions from the passage: 1. Intronless Genes: Many primate-specific genes have only one exon. This suggests that these genes may have arisen through retroposition, a process that intensified in primates around 38-50 million years ago. Alternatively, the "introns-late" model, which suggests intron accumulation over evolutionary time, may explain why these young genes are unspliced or have fewer exons. 2. Location of Primate-Specific Genes: These genes are generally not located in segmental duplications, indicating that single-copy regions may play a role in the origin of primate-specific genes. 3. Alu Repeats: Alu repeats, which are characteristic of primates, overlap with exons of some primate-specific genes. This suggests a possible role of repeat-mediated recombination in gene genesis. 4. Expression in Reproductive Tissues: Primate-specific genes were found to be expressed in brain and neuronal tissues, as well as reproductive tissues. This suggests that these genes may play a role in phenotypic differences and speciation. Notably, they were enriched in reproductive, but not neuronal, expression, indicating their potential involvement in reproductive function and possibly disease.

Chosen Species

We will be choosing several primate species along with Homo sapiens. We will also add one seemingly distant mammal species into the mix to explore how much of the protein sequence is conserved, and to explore some evolutionary relationships.

Main Control Species

- Human - Homo sapiens

Primate Species

- Chimpanzee - Pan troglodytes
- Bonobo - Pan paniscus
- Gorilla - Gorilla gorilla
- Orangutan - Pongo
- Rhesus monkey - Macaca mulatta

Distant Species

- Mouse - Mus musculus

Tools used

UniprotKB

UniProtKB is the Universal Protein Resource Knowledgebase, a comprehensive resource for protein sequence and annotation data. It is freely accessible to the scientific community and is maintained by a consortium of organizations, including the European Bioinformatics Institute (EMBL-EBI), the Swiss Institute of Bioinformatics (SIB), and the Protein Information Resource (PIR).

UniProtKB is the world's most comprehensive resource for protein sequence and annotation data. It contains over 200 million protein sequences from over 500,000 species, including humans, other animals, plants, and microbes. UniProtKB also contains a wealth of information about protein function, structure, and interactions.

To use UniProtKB, you can search the database by protein name, sequence, identifier, or other criteria. The search results will provide you with detailed information about the protein of interest, including its sequence, function, structure, and cross-references to other databases.

ClustalW

ClustalW is a multiple sequence alignment (MSA) program that is widely used in bioinformatics research. It is a progressive MSA algorithm, which means that it aligns sequences one at a time to a growing alignment. ClustalW is known for its accuracy and speed, and it can be used to align a wide range of sequence types, including proteins, DNA, and RNA.

To use ClustalW, you will need to provide the program with a set of sequences to align. The sequences can be in any of the following formats: FASTA, PIR, EMBL, GDE, CLUSTAL, or GCG/MSF. ClustalW will then generate an MSA of the sequences, which can be saved in one of the supported output formats.
ClustalW has a number of parameters that can be adjusted to control the alignment process. These parameters include the gap penalty, the weight matrix, and the alignment method. The gap penalty is used to penalize the introduction of gaps into the alignment. The weight matrix is used to score the alignment of different amino acids or nucleic acids. The alignment method can be either fast/approximate or slow/accurate.

Some applications we will be using -

- **Identifying homologous sequences**: ClustalW can be used to identify homologous sequences, which are sequences that share a common ancestor. Homologous sequences are often similar in function and structure.
- **Studying protein evolution**: ClustalW can be used to study protein evolution by comparing the protein sequences of different organisms. This can help researchers to understand how proteins have changed over time and to identify the evolutionary relationships between different organisms.

Chosen Genes

1. FOXP2 gene

The FOXP2 gene provides instructions for making a protein called forkhead box P2. This protein is a transcription factor, which means that it controls the activity of other genes. It attaches (binds) to the DNA of these genes through a region known as a forkhead domain. Researchers suspect that the forkhead box P2 protein may regulate hundreds of genes, although only some of its targets have been identified. The forkhead box P2 protein is active in several tissues, including the brain, both before and after birth. Studies suggest that it plays important roles in brain development, including the growth of nerve cells (neurons) and the transmission of signals between them. It is also involved in synaptic plasticity, which is the ability of connections between neurons (synapses) to change and adapt to experience over time. Synaptic plasticity is necessary for learning and memory. The forkhead box P2 protein appears to be essential for the normal development of speech and language. Researchers are working to identify the genes regulated by forkhead box P2 that are critical for learning these skills.

In 2001, a study reported the first case of a gene mutated in a developmental speech and language disorder. The culprit, FOXP2, attracted the attention of researchers across

multiple disciplines. Given its link to acquisition of spoken language skills, one of the most distinctive capabilities of Homo sapiens, this gene was seen as an obvious candidate for evolutionary study.

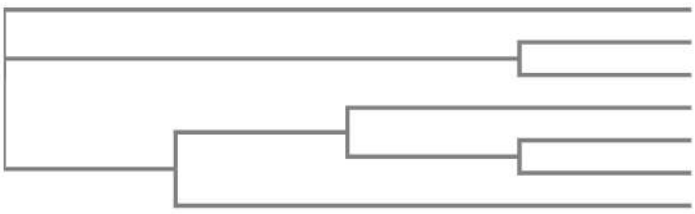
Here are the protein codes from UniProtKB for each species.

a. Human - O15409
b. Chimpanzee - Q8MJA0
c. Gorilla - Q8MJ99
d. Orangutan - Q8MJ98
e. Bonobo - Q8HZ00
f. Rhesus Monkey - Q8MJ97
g. Mouse - P58463

Phylogenetic Tree

This is a Neighbour-joining tree without distance corrections.

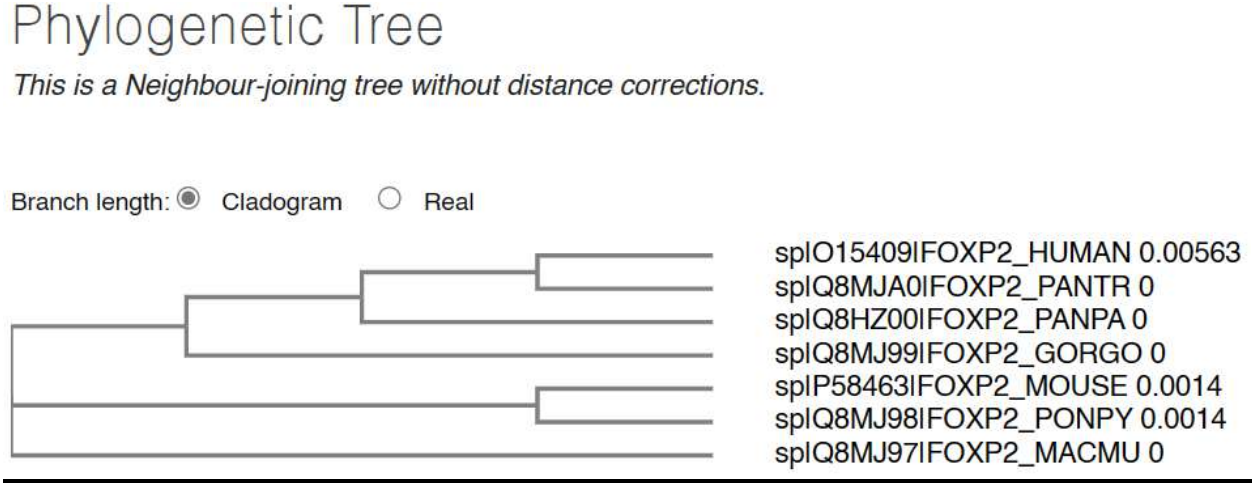
Branch length: ◉ Cladogram ○ Real




```
#
#
#  Percent Identity  Matrix - created by Clustal2.1
#
#

     1: sp|P58463|FOXP2_MOUSE  100.00   99.58   99.72   99.86   99.86   99.86   99.86
     2: sp|O15409|FOXP2_HUMAN   99.58  100.00   99.58   99.72   99.72   99.72   99.72
     3: sp|Q8MJ98|FOXP2_PONPY   99.72   99.58  100.00   99.86   99.86   99.86   99.86
     4: sp|Q8MJ99|FOXP2_GORGO   99.86   99.72   99.86  100.00  100.00  100.00  100.00
     5: sp|Q8MJ97|FOXP2_MACMU   99.86   99.72   99.86  100.00  100.00  100.00  100.00
     6: sp|Q8MJA0|FOXP2_PANTR   99.86   99.72   99.86  100.00  100.00  100.00  100.00
     7: sp|Q8HZ00|FOXP2_PANPA   99.86   99.72   99.86  100.00  100.00  100.00  100.00
```

```
sp|P58463|FOXP2_MOUSE      MMQESATETISNSSMNQNGMSTLSSQLDAGSRDGRSSGDTSSEVSTVELLHLQQQQALQA 60
sp|O15409|FOXP2_HUMAN      MMQESATETISNSSMNQNGMSTLSSQLDAGSRDGRSSGDTSSEVSTVELLHLQQQQALQA 60
sp|Q8MJ98|FOXP2_PONPY      MMQESVTETISNSSMNQNGMSTLSSQLDAGSRDGRSSGDTSSEVSTVELLHLQQQQALQA 60
sp|Q8MJ99|FOXP2_GORGO      MMQESATETISNSSMNQNGMSTLSSQLDAGSRDGRSSGDTSSEVSTVELLHLQQQQALQA 60
sp|Q8MJ97|FOXP2_MACMU      MMQESATETISNSSMNQNGMSTLSSQLDAGSRDGRSSGDTSSEVSTVELLHLQQQQALQA 60
sp|Q8MJA0|FOXP2_PANTR      MMQESATETISNSSMNQNGMSTLSSQLDAGSRDGRSSGDTSSEVSTVELLHLQQQQALQA 60
sp|Q8HZ00|FOXP2_PANPA      MMQESATETISNSSMNQNGMSTLSSQLDAGSRDGRSSGDTSSEVSTVELLHLQQQQALQA 60
                           *****.************************************************** 

sp|P58463|FOXP2_MOUSE      ARQLLLQQQTSGLKSPKSSEKQRPLQVPVSVAMMTPQVITPQQMQQILQQQVLSPQQLQA 120
sp|O15409|FOXP2_HUMAN      ARQLLLQQQTSGLKSPKSSDKQRPLQVPVSVAMMTPQVITPQQMQQILQQQVLSPQQLQA 120
sp|Q8MJ98|FOXP2_PONPY      ARQLLLQQQTSGLKSPKSSDKQRPLQVPVSVAMMTPQVITPQQMQQILQQQVLSPQQLQA 120
sp|Q8MJ99|FOXP2_GORGO      ARQLLLQQQTSGLKSPKSSDKQRPLQVPVSVAMMTPQVITPQQMQQILQQQVLSPQQLQA 120
sp|Q8MJ97|FOXP2_MACMU      ARQLLLQQQTSGLKSPKSSDKQRPLQVPVSVAMMTPQVITPQQMQQILQQQVLSPQQLQA 120
sp|Q8MJA0|FOXP2_PANTR      ARQLLLQQQTSGLKSPKSSDKQRPLQVPVSVAMMTPQVITPQQMQQILQQQVLSPQQLQA 120
sp|Q8HZ00|FOXP2_PANPA      ARQLLLQQQTSGLKSPKSSDKQRPLQVPVSVAMMTPQVITPQQMQQILQQQVLSPQQLQA 120
                           ******************:***************************************** 

sp|P58463|FOXP2_MOUSE      LLQQQQAVMLQQQQLQEFYKKQQEQLHLQLLQQQQQQQQQQQQQQQQQQQQ-QQQQQQQQ 179
sp|O15409|FOXP2_HUMAN      LLQQQQAVMLQQQQLQEFYKKQQEQLHLQLLQQQQQQQQQQQQQQQQQQQQ-QQQQQQQQ 179
sp|Q8MJ98|FOXP2_PONPY      LLQQQQAVMLQQQQLQEFYKKQQEQLHLQLLQQQQQQQQQQQQQQQQQQQQ--QQQQQQQQ 178
sp|Q8MJ99|FOXP2_GORGO      LLQQQQAVMLQQQQLQEFYKKQQEQLHLQLLQQQQQQQQQQQQQQQQQQQQ---QQQQQQQQ 177
sp|Q8MJ97|FOXP2_MACMU      LLQQQQAVMLQQQQLQEFYKKQQEQLHLQLLQQQQQQQQQQQQQQQQQQQQ--QQQQQQQQ 178
sp|Q8MJA0|FOXP2_PANTR      LLQQQQAVMLQQQQLQEFYKKQQEQLHLQLLQQQQQQQQQQQQQQQQQQQQQQQQQQQQQ 180
sp|Q8HZ00|FOXP2_PANPA      LLQQQQAVMLQQQQLQEFYKKQQEQLHLQLLQQQQQQQQQQQQQQQQQQQQQQQQQQQQQ 180
                           ************************************************   ******** 

sp|P58463|FOXP2_MOUSE      QQQQQQQQQQQQHPGKQAKEQQQQ-QQQQQLAAQQLVFQQQLLQMQQLQQQQHLLSLQRQ 238
sp|O15409|FOXP2_HUMAN      QQQQQQQQQQQQHPGKQAKEQQQQQQQQQQLAAQQLVFQQQLLQMQQLQQQQHLLSLQRQ 239
sp|Q8MJ98|FOXP2_PONPY      QQQQQQQQQQQQHPGKQAKEQQQQ-QQQQQLAAQQLVFQQQLLQMQQLQQQQHLLSLQRQ 237
sp|Q8MJ99|FOXP2_GORGO      QQQQQQQQQQQQHPGKQAKEQQQQQQQQQLAAQQLVFQQQLLQMQQLQQQQHLLSLQRQ 237
sp|Q8MJ97|FOXP2_MACMU      QQQQQQQQQQQQHPGKQAKEQQQQQQQQQLAAQQLVFQQQLLQMQQLQQQQHLLSLQRQ 238
sp|Q8MJA0|FOXP2_PANTR      QQQQQQQQQQQQHPGKQAKEQQQQQQQQQLAAQQLVFQQQLLQMQQLQQQQHLLSLQRQ 240
sp|Q8HZ00|FOXP2_PANPA      QQQQQQQQQQQQHPGKQAKEQQQQQQQQQLAAQQLVFQQQLLQMQQLQQQQHLLSLQRQ 240
                           ************************ ********************************** 
```

Upon knockout of 4 amino acids at different sites



```
#
#  Percent Identity  Matrix - created by Clustal2.1
#
#
```

1: sp|O15409|FOXP2_HUMAN	100.00	99.29	99.29	99.44	99.44	99.44	99.44
2: sp|P58463|FOXP2_MOUSE	99.29	100.00	99.72	99.86	99.86	99.86	99.86
3: sp|Q8MJ98|FOXP2_PONPY	99.29	99.72	100.00	99.86	99.86	99.86	99.86
4: sp|Q8MJ99|FOXP2_GORGO	99.44	99.86	99.86	100.00	100.00	100.00	100.00
5: sp|Q8MJ97|FOXP2_MACMU	99.44	99.86	99.86	100.00	100.00	100.00	100.00
6: sp|Q8MJA0|FOXP2_PANTR	99.44	99.86	99.86	100.00	100.00	100.00	100.00
7: sp|Q8HZ00|FOXP2_PANPA	99.44	99.86	99.86	100.00	100.00	100.00	100.00

```
sp|O15409|FOXP2_HUMAN      MMQESATETISNSSMNQNGMSTLSSQLDAGSRDGRSSGDTSSEVSTVELLHLQQQQALQA 60
sp|P58463|FOXP2_MOUSE      MMQESATETISNSSMNQNGMSTLSSQLDAGSRDGRSSGDTSSEVSTVELLHLQQQQALQA 60
sp|Q8MJ98|FOXP2_PONPY      MMQESVTETISNSSMNQNGMSTLSSQLDAGSRDGRSSGDTSSEVSTVELLHLQQQQALQA 60
sp|Q8MJ99|FOXP2_GORGO      MMQESATETISNSSMNQNGMSTLSSQLDAGSRDGRSSGDTSSEVSTVELLHLQQQQALQA 60
sp|Q8MJ97|FOXP2_MACMU      MMQESATETISNSSMNQNGMSTLSSQLDAGSRDGRSSGDTSSEVSTVELLHLQQQQALQA 60
sp|Q8MJA0|FOXP2_PANTR      MMQESATETISNSSMNQNGMSTLSSQLDAGSRDGRSSGDTSSEVSTVELLHLQQQQALQA 60
sp|Q8HZ00|FOXP2_PANPA      MMQESATETISNSSMNQNGMSTLSSQLDAGSRDGRSSGDTSSEVSTVELLHLQQQQALQA 60
                           *****.**************************************************

sp|O15409|FOXP2_HUMAN      ARQLLLQQQTSGLKSPKSSDKQRPLQVPVSVAMMTPQVITPQQMQQILQQQVLSPQQLQA 120
sp|P58463|FOXP2_MOUSE      ARQLLLQQQTSGLKSPKSSEKQRPLQVPVSVAMMTPQVITPQQMQQILQQQVLSPQQLQA 120
sp|Q8MJ98|FOXP2_PONPY      ARQLLLQQQTSGLKSPKSSDKQRPLQVPVSVAMMTPQVITPQQMQQILQQQVLSPQQLQA 120
sp|Q8MJ99|FOXP2_GORGO      ARQLLLQQQTSGLKSPKSSDKQRPLQVPVSVAMMTPQVITPQQMQQILQQQVLSPQQLQA 120
sp|Q8MJ97|FOXP2_MACMU      ARQLLLQQQTSGLKSPKSSDKQRPLQVPVSVAMMTPQVITPQQMQQILQQQVLSPQQLQA 120
sp|Q8MJA0|FOXP2_PANTR      ARQLLLQQQTSGLKSPKSSDKQRPLQVPVSVAMMTPQVITPQQMQQILQQQVLSPQQLQA 120
sp|Q8HZ00|FOXP2_PANPA      ARQLLLQQQTSGLKSPKSSDKQRPLQVPVSVAMMTPQVITPQQMQQILQQQVLSPQQLQA 120
                           ********************:***************************************

sp|O15409|FOXP2_HUMAN      LLQQQQAVMLQQQQLQEFYKKQQEQLHLQLQQQQQQQQQQQQQQQQQQQ--QQQQQQQQQ 178
sp|P58463|FOXP2_MOUSE      LLQQQQAVMLQQQQLQEFYKKQQEQLHLQLLQQQQQQQQQQQQQQQQQQQQ-QQQQQQQQ 179
sp|Q8MJ98|FOXP2_PONPY      LLQQQQAVMLQQQQLQEFYKKQQEQLHLQLLQQQQQQQQQQQQQQQQQQQQ--QQQQQQQQ 178
sp|Q8MJ99|FOXP2_GORGO      LLQQQQAVMLQQQQLQEFYKKQQEQLHLQLLQQQQQQQQQQQQQQQQQQQ---QQQQQQQQ 177
sp|Q8MJ97|FOXP2_MACMU      LLQQQQAVMLQQQQLQEFYKKQQEQLHLQLLQQQQQQQQQQQQQQQQQQQ--QQQQQQQQ 178
sp|Q8MJA0|FOXP2_PANTR      LLQQQQAVMLQQQQLQEFYKKQQEQLHLQLLQQQQQQQQQQQQQQQQQQQQQQQQQQQQ 180
sp|Q8HZ00|FOXP2_PANPA      LLQQQQAVMLQQQQLQEFYKKQQEQLHLQLLQQQQQQQQQQQQQQQQQQQQQQQQQQQQ 180
                           ******************************* ****************   ********

sp|O15409|FOXP2_HUMAN      QQQQQQQQQQQQHPGKQAKEQQQQQQQQQQLAAQQLVFQQQLLQMQQLQQQQHLLSLQRQ 238
sp|P58463|FOXP2_MOUSE      QQQQQQQQQQQQHPGKQAKEQQQQ-QQQQQLAAQQLVFQQQLLQMQQLQQQQHLLSLQRQ 238
sp|Q8MJ98|FOXP2_PONPY      QQQQQQQQQQQQHPGKQAKEQQQQ-QQQQQLAAQQLVFQQQLLQMQQLQQQQHLLSLQRQ 237
sp|Q8MJ99|FOXP2_GORGO      QQQQQQQQQQQQHPGKQAKEQQQQQQQQQQLAAQQLVFQQQLLQMQQLQQQQHLLSLQRQ 237
sp|Q8MJ97|FOXP2_MACMU      QQQQQQQQQQQQHPGKQAKEQQQQQQQQQQLAAQQLVFQQQLLQMQQLQQQQHLLSLQRQ 238
sp|Q8MJA0|FOXP2_PANTR      QQQQQQQQQQQQHPGKQAKEQQQQQQQQQQLAAQQLVFQQQLLQMQQLQQQQHLLSLQRQ 240
sp|Q8HZ00|FOXP2_PANPA      QQQQQQQQQQQQHPGKQAKEQQQQQQQQQQLAAQQLVFQQQLLQMQQLQQQQHLLSLQRQ 240
                           ************************ ************************************
```

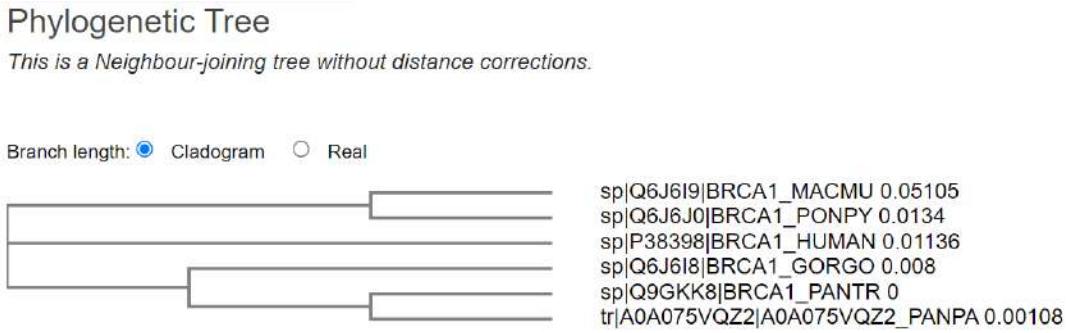
2. BRCA1 gene

The BRCA1 gene provides instructions for making a protein that acts as a tumor suppressor. Tumor suppressor proteins help prevent cells from growing and dividing too rapidly or in an uncontrolled way.

The BRCA1 protein is involved in repairing damaged DNA. In the nucleus of many types of normal cells, the BRCA1 protein interacts with several other proteins to mend breaks in DNA. These breaks can be caused by natural and medical radiation or other environmental exposures, and they also occur when chromosomes exchange genetic material in preparation for cell division. By helping to repair DNA, the BRCA1 protein plays a critical role in maintaining the stability of a cell's genetic information.

Here are the protein codes from UniProtKB for each species.

a. Human
b. Chimp
c. Gorilla
d. Orangutan
e. Bonobo
f. Rhesus Monkey
g. Mouse




```
#
#  Percent Identity  Matrix - created by Clustal2.1
#
#

     1: sp|Q6J6I9|BRCA1_MACMU          100.00   93.56   93.13   93.39   93.77   93.69
     2: sp|Q6J6J0|BRCA1_PONPY           93.56  100.00   96.83   97.16   97.58   97.47
     3: sp|P38398|BRCA1_HUMAN           93.13   96.83  100.00   98.01   98.44   98.33
     4: sp|Q6J6I8|BRCA1_GORGO           93.39   97.16   98.01  100.00   98.82   98.71
     5: sp|Q9GKK8|BRCA1_PANTR           93.77   97.58   98.44   98.82  100.00   99.89
     6: tr|A0A075VQZ2|A0A075VQZ2_PANPA  93.69   97.47   98.33   98.71   99.89  100.00
```

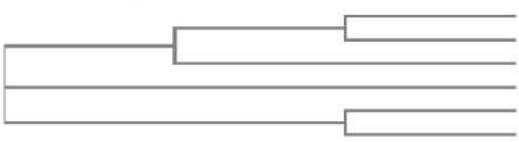
```
sp|Q6J6I9|BRCA1_MACMU           SENPRDAEDVPWITLNSSIQKVNEWFSRSDELLSSDDSHDGGSESNAKVADVLDVLNEVD   419
sp|Q6J6J0|BRCA1_PONPY           SENPRDTEDVPWITLNSSIQKVNEWFSRSDELLGSDDSHDGRSESNAKVADVLDVLNEVD   420
sp|P38398|BRCA1_HUMAN           SENPRDTEDVPWITLNSSIQKVNEWFSRSDELLGSDDSHDGESESNAKVADVLDVLNEVD   420
sp|Q6J6I8|BRCA1_GORGO           SENPRDTEDVPWITLNSSIQKVNEWFSRSDELLGSDDSHDGGSESNAKVADVLDVLNEVD   420
sp|Q9GKK8|BRCA1_PANTR           SENPRDTEDVPWITLNSSIQKVNEWFSRSDELLGSDDSHDGGSESNAKVADVLDVLNEVD   420
tr|A0A075VQZ2|A0A075VQZ2_PANPA  SENPRDTEDVPWITLNSSIQKVNEWFSRSDELLGSDDSHDGGSESNAKVADVLDVLNEVD   420
                                ******:************************.******* ******************

sp|Q6J6I9|BRCA1_MACMU           EYSGSSEKIDLLASDPHEPLICKSERVHSSSVESNIKDKIFGKTYRRKANLPNLSHVTEN   479
sp|Q6J6J0|BRCA1_PONPY           EYSGSSEKIDLLASDPHEALICKSERVHSKSVESNIEDKIFGKTYRRKASLPNLSHVTEN   480
sp|P38398|BRCA1_HUMAN           EYSGSSEKIDLLASDPHEALICKSERVHSKSVESNIEDKIFGKTYRKKASLPNLSHVTEN   480
sp|Q6J6I8|BRCA1_GORGO           EYSGSSEKIDLLASDPHEALICKSERVHSKSVESNIEDKIFGKTYRRKASLPSLSHVTEN   480
sp|Q9GKK8|BRCA1_PANTR           EYSGSSEKIDLLASDPHEALICKSERVHSKSVESNTEDKIFGKTYRRKASLPNLSHVTEN   480
tr|A0A075VQZ2|A0A075VQZ2_PANPA  EYSGSSEKIDLLASDPHEALICKSERVHSKSVESNTEDKIFGKTYRRKASLPNLSHVTEN   480
                                ***************** **********.***** :*********:**.**.*******

sp|Q6J6I9|BRCA1_MACMU           LIIGALVTESQIMQERPLTNKLKRKRRTTSGLHPEDFIKKADLAVQKTPEIINQGTNQME   539
sp|Q6J6J0|BRCA1_PONPY           LIIGAFVTEPQIIQERPLTNKLKRKRRATSGLHPEDFIKKADLAVQKTPEMINQGTNQME   540
sp|P38398|BRCA1_HUMAN           LIIGAFVTEPQIIQERPLTNKLKRKRRPTSGLHPEDFIKKADLAVQKTPEMINQGTNQTE   540
sp|Q6J6I8|BRCA1_GORGO           LIIGAFVTEPQIIQERPLTNKLKRKRRATSGLHPEDFIKKADLAVQKTPEMINQGTNQME   540
sp|Q9GKK8|BRCA1_PANTR           LIIGAFVTEPQIIQERPLTNKLKRKRRATSGLHPEDFIKKADLAVQKTPEMINQGTNQME   540
tr|A0A075VQZ2|A0A075VQZ2_PANPA  LIIGAFVTEPQIIQERPLTNKLKRKRRATSGLHPEDFIKKADLAVQKTPEMINQGTNQME   540
                                *****:*** **:************** ********************:******* *

sp|Q6J6I9|BRCA1_MACMU           QNGQVMNITNSAHENKTKGDSIQNEKNPNAIESLEEESAFKTKAEPISSSINNMELELNI   599
sp|Q6J6J0|BRCA1_PONPY           QNGQVMNITNSGHENKTKGDSIQNEKNPNPIESLEKESAFKTKAEPISSSISNMELELNI   600
sp|P38398|BRCA1_HUMAN           QNGQVMNITNSGHENKTKGDSIQNEKNPNPIESLEKESAFKTKAEPISSSISNMELELNI   600
sp|Q6J6I8|BRCA1_GORGO           QNGQVMNITNSGHENKTKGDSIQNEKNPNPIESLEKESAFKTKAEPISSCISNMELELNI   600
sp|Q9GKK8|BRCA1_PANTR           QNGQVMNITNSGHENKTKGDSIQNEKNPNPIESLEKESAFKTKAEPISSSISNMELELNI   600
tr|A0A075VQZ2|A0A075VQZ2_PANPA  QNGQVMNITNSGHENKTKGDSIQNEKNPNPIESLEKESAFKTKAEPISSSISNMELELNI   600
                                **********.**************** *****:************.*.********

sp|Q6J6I9|BRCA1_MACMU           HNSKAPKKNRLRRKSSTRHIHALELVVSRNLSPPNCTELQIDSCSSSEEIKKKNYNQMPV   659
sp|Q6J6J0|BRCA1_PONPY           HNSKAPKKNRLRRKSSTRHIHALELVVSRNLSPPNCTELQIDSCSSSEEIKKKKYNQMPV   660
sp|P38398|BRCA1_HUMAN           HNSKAPKKNRLRRKSSTRHIHALELVVSRNLSPPNCTELQIDSCSSSEEIKKKKYNQMPV   660
sp|Q6J6I8|BRCA1_GORGO           HNSKAPKKNRLRRKSSTRHIHALELVVSRNLSPPNCTELQIDSCSSSEEIKKKKYNQMPV   660
sp|Q9GKK8|BRCA1_PANTR           HNSKAPKKNRLRRKSSTRHIHALELVVSRNLSPPNCTELQIDSCSSSEEIKKKKYNQMPV   660
tr|A0A075VQZ2|A0A075VQZ2_PANPA  HNSKAPKKNRLRRKSSTRHIHALELVVSRNLSPPNCTELQIDSCSSSEEIKKKKYNQMPV   660
                                ************************************************.******
```

UPON KNOCKOUT

Phylogenetic Tree

This is a Neighbour-joining tree without distance corrections.

Branch length: ◉ Cladogram ○ Real



sp|Q6J6I9|BRCA1_MACMU 0.05104
sp|Q6J6J0|BRCA1_PONPY 0.0134
sp|P38398|BRCA1_HUMAN 0.01082
sp|Q6J6I8|BRCA1_GORGO 0.00816
sp|Q9GKK8|BRCA1_PANTR 0
tr|A0A075VQZ2|A0A075VQZ2_PANPA 0.00112

```
#
#
#  Percent Identity  Matrix - created by Clustal2.1
#
#

     1: sp|Q6J6I9|BRCA1_MACMU          100.00   93.56   93.22   93.39   93.77   93.69
     2: sp|Q6J6J0|BRCA1_PONPY           93.56  100.00   96.93   97.16   97.58   97.47
     3: sp|P38398|BRCA1_HUMAN           93.22   96.93  100.00   97.99   98.50   98.36
     4: sp|Q6J6I8|BRCA1_GORGO           93.39   97.16   97.99  100.00   98.82   98.71
     5: sp|Q9GKK8|BRCA1_PANTR           93.77   97.58   98.50   98.82  100.00   99.89
     6: tr|A0A075VQZ2|A0A075VQZ2_PANPA  93.69   97.47   98.36   98.71   99.89  100.00
```

,

```
sp|O15409|FOXP2_HUMAN      SDKYNIPMSSEIAPNYEFYKNADVRPPFTYATLIRQAIMESSDRQLTLNEIYSWFTRTFA 536
sp|P58463|FOXP2_MOUSE      SDKYNIPMSSEIAPNYEFYKNADVRPPFTYATLIRQAIMESSDRQLTLNEIYSWFTRTFA 538
sp|Q8MJ98|FOXP2_PONPY      SDKYNIPMSSEIAPNYEFYKNADVRPPFTYATLIRQAIMESSDRQLTLNEIYSWFTRTFA 537
sp|Q8MJ99|FOXP2_GORGO      SDKYNIPMSSEIAPNYEFYKNADVRPPFTYATLIRQAIMESSDRQLTLNEIYSWFTRTFA 537
sp|Q8MJ97|FOXP2_MACMU      SDKYNIPMSSEIAPNYEFYKNADVRPPFTYATLIRQAIMESSDRQLTLNEIYSWFTRTFA 538
sp|Q8MJA0|FOXP2_PANTR      SDKYNIPMSSEIAPNYEFYKNADVRPPFTYATLIRQAIMESSDRQLTLNEIYSWFTRTFA 540
sp|Q8HZ00|FOXP2_PANPA      SDKYNIPMSSEIAPNYEFYKNADVRPPFTYATLIRQAIMESSDRQLTLNEIYSWFTRTFA 540
                           ************************************************************

sp|O15409|FOXP2_HUMAN      YFRRNAATWKNAVRHNLSLHKCFVRVENVKGAV-TVDEVEYQKRRSQKITGSPTLVKNIP 595
sp|P58463|FOXP2_MOUSE      YFRRNAATWKNAVRHNLSLHKCFVRVENVKGAVWTVDEVEYQKRRSQKITGSPTLVKNIP 598
sp|Q8MJ98|FOXP2_PONPY      YFRRNAATWKNAVRHNLSLHKCFVRVENVKGAVWTVDEVEYQKRRSQKITGSPTLVKNIP 597
sp|Q8MJ99|FOXP2_GORGO      YFRRNAATWKNAVRHNLSLHKCFVRVENVKGAVWTVDEVEYQKRRSQKITGSPTLVKNIP 597
sp|Q8MJ97|FOXP2_MACMU      YFRRNAATWKNAVRHNLSLHKCFVRVENVKGAVWTVDEVEYQKRRSQKITGSPTLVKNIP 598
sp|Q8MJA0|FOXP2_PANTR      YFRRNAATWKNAVRHNLSLHKCFVRVENVKGAVWTVDEVEYQKRRSQKITGSPTLVKNIP 600
sp|Q8HZ00|FOXP2_PANPA      YFRRNAATWKNAVRHNLSLHKCFVRVENVKGAVWTVDEVEYQKRRSQKITGSPTLVKNIP 600
                           ********************************* ************************

sp|O15409|FOXP2_HUMAN      TSLGYGAALNASLQAALAESSLPLLSNPGLINNASSGLLQAVHEDLNGSLDHIDSNGNSS 655
sp|P58463|FOXP2_MOUSE      TSLGYGAALNASLQAALAESSLPLLSNPGLINNASSGLLQAVHEDLNGSLDHIDSNGNSS 658
sp|Q8MJ98|FOXP2_PONPY      TSLGYGAALNASLQAALAESSLPLLSNPGLINNASSGLLQAVHEDLNGSLDHIDSNGNSS 657
sp|Q8MJ99|FOXP2_GORGO      TSLGYGAALNASLQAALAESSLPLLSNPGLINNASSGLLQAVHEDLNGSLDHIDSNGNSS 657
sp|Q8MJ97|FOXP2_MACMU      TSLGYGAALNASLQAALAESSLPLLSNPGLINNASSGLLQAVHEDLNGSLDHIDSNGNSS 658
sp|Q8MJA0|FOXP2_PANTR      TSLGYGAALNASLQAALAESSLPLLSNPGLINNASSGLLQAVHEDLNGSLDHIDSNGNSS 660
sp|Q8HZ00|FOXP2_PANPA      TSLGYGAALNASLQAALAESSLPLLSNPGLINNASSGLLQAVHEDLNGSLDHIDSNGNSS 660
                           ************************************************************

sp|O15409|FOXP2_HUMAN      PGCSPQPHIHSIHVKEEPVIAEDEDCPMSLVTTANHSPELEDDREIEEEPLSEDLE     711
sp|P58463|FOXP2_MOUSE      PGCSPQPHIHSIHVKEEPVIAEDEDCPMSLVTTANHSPELEDDREIEEEPLSEDLE     714
sp|Q8MJ98|FOXP2_PONPY      PGCSPQPHIHSIHVKEEPVIAEDEDCPMSLVTTANHSPELEDDREIEEEPLSEDLE     713
sp|Q8MJ99|FOXP2_GORGO      PGCSPQPHIHSIHVKEEPVIAEDEDCPMSLVTTANHSPELEDDREIEEEPLSEDLE     713
sp|Q8MJ97|FOXP2_MACMU      PGCSPQPHIHSIHVKEEPVIAEDEDCPMSLVTTANHSPELEDDREIEEEPLSEDLE     714
sp|Q8MJA0|FOXP2_PANTR      PGCSPQPHIHSIHVKEEPVIAEDEDCPMSLVTTANHSPELEDDREIEEEPLSEDLE     716
sp|Q8HZ00|FOXP2_PANPA      PGCSPQPHIHSIHVKEEPVIAEDEDCPMSLVTTANHSPELEDDREIEEEPLSEDLE     716
                           ************************************************************
```

3. **<u>CFTR gene</u>**

Provides instructions for making a protein called the CF transmembrane conductance regulator (CFTR). This protein functions as a channel across the membrane of cells that produce mucus, sweat, saliva, tears, and digestive enzymes. The channel transports negatively charged particles called chloride ions into and out of cells. The transport of chloride ions helps control the movement of water in tissues, which is necessary for the production of thin, freely flowing mucus. Mucus is a slippery substance that lubricates and protects the lining of the airways, digestive system, reproductive system, and other organs and tissues.

Here are the protein codes from UniProtKB for each species.

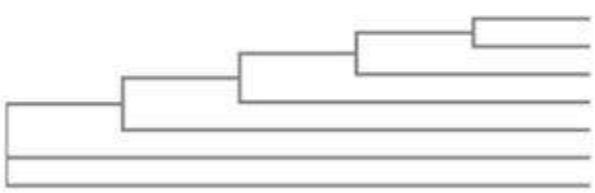
a. Human - P13569
b. Chimp - A0A2J8QXD7

c. Gorilla - Q2IBF6
d. Orangutan - Q2IBE4
e. Bonobo - A0A2R9C0G7
f. Rhesus Monkey(n/a) instead Cynomolgus monkey - G7P2J1
g. Mouse - P26361

Phylogenetic Tree

This is a Neighbour-joining tree without distance corrections.

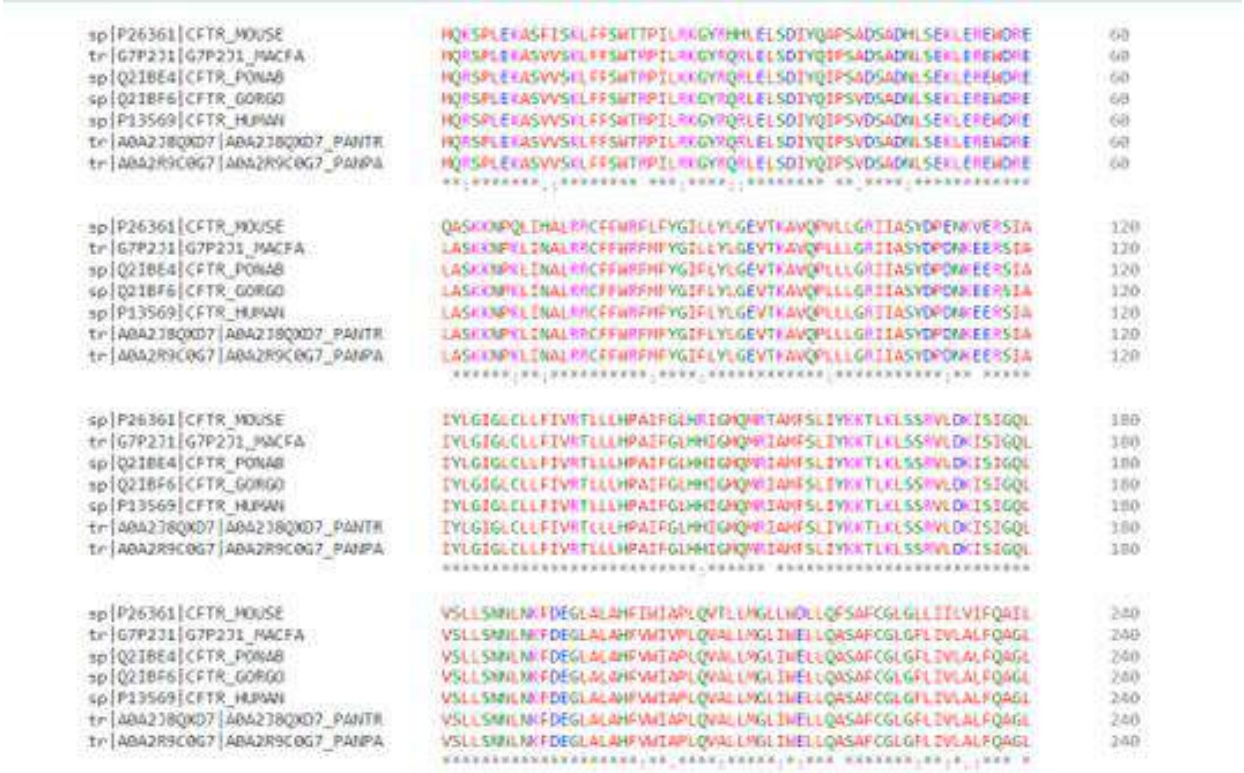
Branch length: Cladogram Real



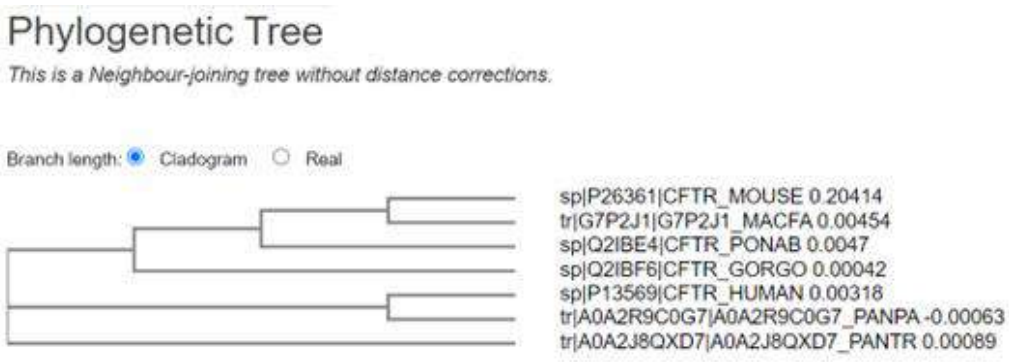
sp|P26361|CFTR_MOUSE 0.20219
tr|G7P2J1|G7P2J1_MACFA 0.00648
sp|Q2IBE4|CFTR_PONAB 0.006
sp|Q2IBF6|CFTR_GORGO 0.00129
sp|P13569|CFTR_HUMAN 0.00198
tr|A0A2J8QXD7|A0A2J8QXD7_PANTR 0.00135
tr|A0A2R9C0G7|A0A2R9C0G7_PANPA 0

```
#
#
#  Percent Identity  Matrix - created by Clustal2.1
#
#
```

1: sp|P26361|CFTR_MOUSE	100.00	79.13	78.71	78.85	78.71	78.71	78.85
2: tr|G7P2J1|G7P2J1_MACFA	79.13	100.00	98.04	98.45	98.31	98.38	98.51
3: sp|Q2IBE4|CFTR_PONAB	78.71	98.04	100.00	99.05	98.92	98.99	99.12
4: sp|Q2IBF6|CFTR_GORGO	78.85	98.45	99.05	100.00	99.59	99.66	99.80
5: sp|P13569|CFTR_HUMAN	78.71	98.31	98.92	99.59	100.00	99.66	99.80
6: tr|A0A2J8QXD7|A0A2J8QXD7_PANTR	78.71	98.38	98.99	99.66	99.66	100.00	99.86
7: tr|A0A2R9C0G7|A0A2R9C0G7_PANPA	78.85	98.51	99.12	99.80	99.80	99.86	100.00

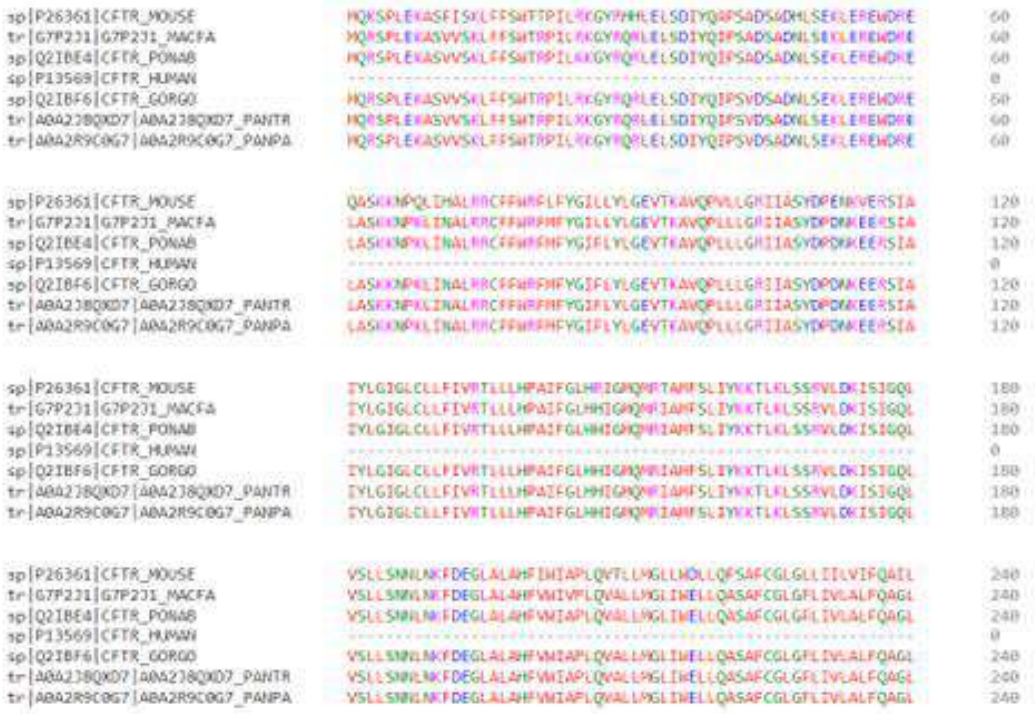


Upon Gene Knockout:



```
#
#
#  Percent Identity  Matrix - created by Clustal2.1
#
#
```

1: sp|P26361|CFTR_MOUSE	100.00	79.13	78.71	76.68	78.85	78.71	78.85
2: tr|G7P2J1|G7P2J1_MACFA	79.13	100.00	98.04	98.22	98.45	98.38	98.51
3: sp|Q2IBE4|CFTR_PONAB	78.71	98.04	100.00	98.90	99.05	98.99	99.12
4: sp|P13569|CFTR_HUMAN	76.68	98.22	98.90	100.00	99.58	99.58	99.75
5: sp|Q2IBF6|CFTR_GORGO	78.85	98.45	99.05	99.58	100.00	99.66	99.80
6: tr|A0A2J8QXD7|A0A2J8QXD7_PANTR	78.71	98.38	98.99	99.58	99.66	100.00	99.86
7: tr|A0A2R9C0G7|A0A2R9C0G7_PANPA	78.85	98.51	99.12	99.75	99.80	99.86	100.00



4.APOE gene

The **APOE** gene, short for Apolipoprotein E gene, is a human gene that encodes a protein involved in the metabolism of lipids (fats) and cholesterol in the body. This gene has several different alleles, with the two most common being APOE2, APOE3, and APOE4. These alleles vary in their structure and function and are associated with different health outcomes, particularly in relation to cholesterol and lipid metabolism.

Here are the protein codes from UniProtKB for each species.

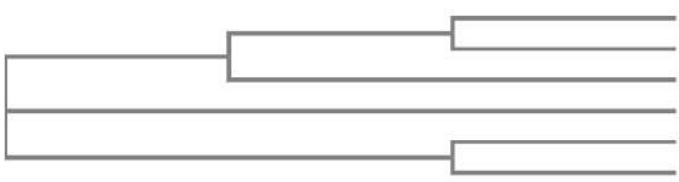
a. Human P02649
b. Chimp Q9GJU3
c. Gorilla Q9GLM8

d. Bornean Oragantun Q9GLM7
e. Rhesus Monkey P10517
f. Mouse P08226

Phylogenetic Tree

This is a Neighbour-joining tree without distance corrections.

Branch length: ◉ Cladogram ○ Real

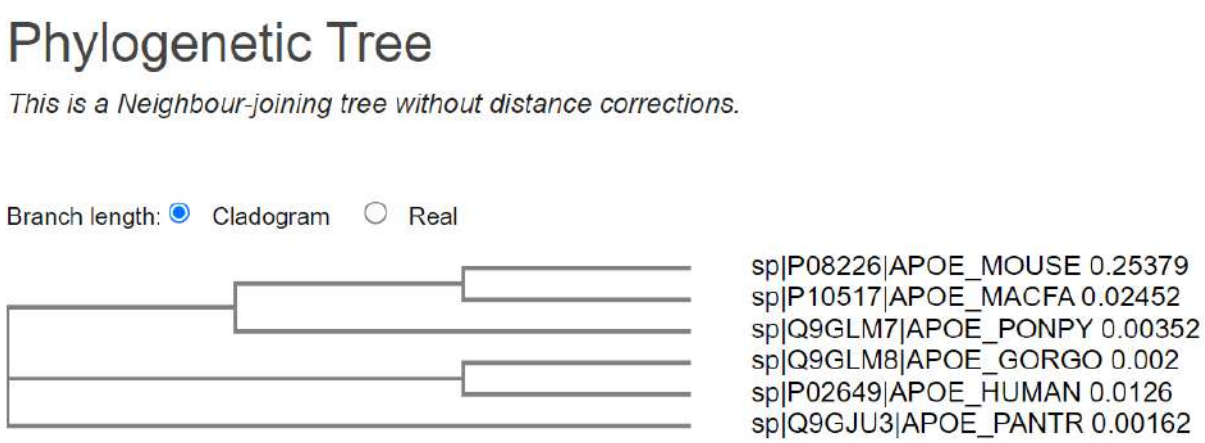


sp|P08226|APOE_MOUSE 0.24472
sp|P10517|APOE_MACFA 0.0336
sp|Q9GLM7|APOE_PONPY 0.00765
sp|P02649|APOE_HUMAN 0.01718
sp|Q9GJU3|APOE_PANTR 0.005
sp|Q9GLM8|APOE_GORGO 0.00446

```
#
#
#  Percent Identity  Matrix - created by Clustal2.1
#
#

     1: sp|P08226|APOE_MOUSE  100.00   72.17   73.14   72.17   72.82   73.14
     2: sp|P10517|APOE_MACFA   72.17  100.00   94.32   93.38   94.01   94.01
     3: sp|Q9GLM7|APOE_PONPY   73.14   94.32  100.00   97.48   98.42   98.11
     4: sp|P02649|APOE_HUMAN   72.17   93.38   97.48  100.00   97.16   97.48
     5: sp|Q9GJU3|APOE_PANTR   72.82   94.01   98.42   97.16  100.00   99.05
     6: sp|Q9GLM8|APOE_GORGO   73.14   94.01   98.11   97.48   99.05  100.00
```

```
sp|P08226|APOE_MOUSE      KAGAREGAERGVSAIRERLGPLVEQGRQRTANLGAGAAQPLRDRAQAFGDRIRGRLEEVG  232
sp|P10517|APOE_MACFA      QAGAREGAERGVSAIRERLGPLVEQGRVRAATVGSLASQPLQERAQALGERLRARMEEMG  240
sp|Q9GLM7|APOE_PONPY      QAGAREGAERGVSAIRERLGPLVEQGRVRAATVGSVAGKPLQERAQAWGERLRARMEEMG  240
sp|P02649|APOE_HUMAN      QAGAREGAERGLSAIRERLGPLVEQGRVRAATVGSLAGQPLQERAQAWGERLRARMEEMG  240
sp|Q9GJU3|APOE_PANTR      QAGAREGAERGVSAIRERLGPLVEQGRVRAATVGSLAGQPLQERAQAWGERLRARMEEMG  240
sp|Q9GLM8|APOE_GORGO      QAGAREGAERGVSAIRERLGPLVEQGRVRAATVGSLAGQPLQERAQAWGERLRARMEEMG  240
                          :**********:*************** *:*.:*: *.:**::**** *:*:*.*:**:*

sp|P08226|APOE_MOUSE      NQARDRLEEVREHMEEVRSKMEEQTQQIRLQAEIFQARLKGWFEPIVEDMHRQWANLMEK  292
sp|P10517|APOE_MACFA      SRTRDRLDEVKEQVAEVRAKLEEQAQQISLQAEAFQARLKSWFEPLVEDMQRQWAGLVEK  300
sp|Q9GLM7|APOE_PONPY      SRTRDRLDEVKEQVAEVRAKLEEQAQQIRLQAEAFQARLKSWFEPLVEDMQRQWAGLVEK  300
sp|P02649|APOE_HUMAN      SRTRDRLDEVKEQVAEVRAKLEEQAQQIRLQAEAFQARLKSWFEPLVEDMQRQWAGLVEK  300
sp|Q9GJU3|APOE_PANTR      SRTRDRLDEVKEQVAEVRAKLEEQAQQIRLQAEAFQARLKSWFEPLVEDMQRQWAGLVEK  300
sp|Q9GLM8|APOE_GORGO      SRTRDRLDEVKEQVAEVRAKLEEQAQQIRLQAEAFQARLKSWFEPLVEDMQRQWAGLVEK  300
                          .::****:**:*:: ***:*:***:*** **** ******.****:****:****.*:**

sp|P08226|APOE_MOUSE      IQASVATNPIITPVAQENQ   311
sp|P10517|APOE_MACFA      VQAAVGASTAPVPIDNH--   317
sp|Q9GLM7|APOE_PONPY      VQAAVGTSAAPVPSDNH--   317
sp|P02649|APOE_HUMAN      VQAAVGTSAAPVPSDNH--   317
sp|Q9GJU3|APOE_PANTR      VQAAMGTSAAPVPSDNH--   317
sp|Q9GLM8|APOE_GORGO      VQAAMGTSAAPVPSDNH--   317
                          :**::.:.   .*  :.
```

Upon Knockout

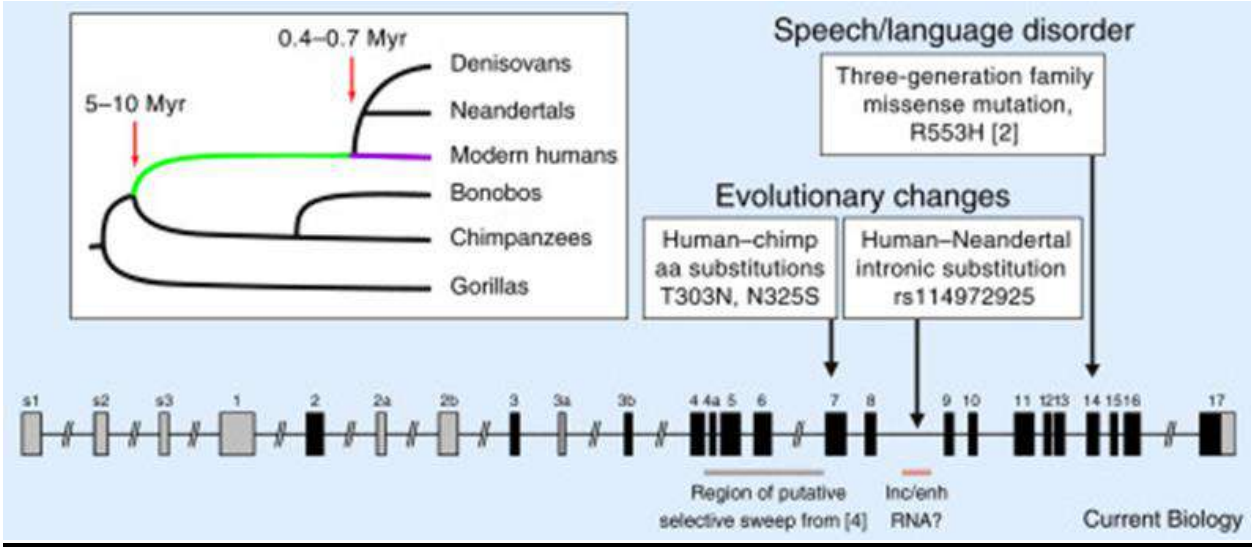



```
#
#
#  Percent Identity  Matrix - created by Clustal2.1
#
#

     1: sp|P08226|APOE_MOUSE  100.00   72.17   73.14   73.14   66.42   72.82
     2: sp|P10517|APOE_MACFA   72.17  100.00   94.32   94.01   94.89   94.01
     3: sp|Q9GLM7|APOE_PONPY   73.14   94.32  100.00   98.11   97.81   98.42
     4: sp|Q9GLM8|APOE_GORGO   73.14   94.01   98.11  100.00   98.54   99.05
     5: sp|P02649|APOE_HUMAN   66.42   94.89   97.81   98.54  100.00   98.54
     6: sp|Q9GJU3|APOE_PANTR   72.82   94.01   98.42   99.05   98.54  100.00
```

Results & Learnings

1. For FOXP2 Gene

- This led to two independent reports in 2002, which established that, despite high sequence conservation in primates, the human FOXP2 protein differed from its chimpanzee counterpart at two amino-acid sites.
- FOXP2 is ancient history; the gene is found in similar form in rodents, birds, reptiles and fish among others
- Knockdown of the avian FOXP2 orthologue in a key basal ganglia nucleus in brains of zebra finches affects the variability of the songs that they learn.
- Initial studies identified amino-acid substitutions in exon 7 that distinguish human FOXP2 from the chimpanzee protein and suggested that they were associated with a selective sweep within the last 100–200 thousand years.
- Researchers found that the protein-coding sequence of FOXP2 in Neandertals matched that of Homo sapiens. This result suggested that, rather than being subject to recent selection, the two amino-acid substitutions were more likely already fixed in the common ancestor of Neandertals and modern humans, which lived at least 400 thousand years ago.
- But there were mutation events that occurred before and after the neanderthal split.
- One report suggested that the target of the putative recent sweep may have been within an intronic regulatory element.



2. For BRCA1 Gene

BRCA1, we analyzed complete BRCA1 gene sequences from 6 primate species. We show that specific amino acid sites have experienced repeated selection for amino acid

replacement over primate evolution. This selection has been focused specifically on humans and our closest living relatives, chimpanzees (Pan troglodytes) and bonobos (Pan paniscus). After examining BRCA1 polymorphisms in bonobo, chimpanzee, and rhesus macaque (Macaca mulatta) individuals, we find considerable variation within each of these species and evidence for recent selection in chimpanzee populations. Finally, we also sequenced and analyzed BRCA2 from 6 primate species and find that this gene has also evolved under positive selection.

3. For CFTR Gene

- Mouse (sp P26361/CFTR_MOUSE): The matrix data indicates that the mouse CFTR sequence has lower identity values with the primate species, reflecting a more distant genetic relationship. This is likely supported by the tree data, where the mouse branch is expected to be relatively longer, indicating greater genetic divergence.
- Chimpanzee, Bonobo, and Orangutan: These species are expected to have closer genetic relationships in the tree, with shorter branch lengths due to their shared common ancestry. The matrix data should also show higher identity values among these species
- Gorilla: Gorillas are expected to be closer to humans in the phylogenetic tree than to the more distant mouse. This should be reflected in the matrix data with higher identity values between gorillas and humans compared to gorillas and mice.
- Rhesus Monkeys: The matrix data might show moderate to high identity values between rhesus monkeys and humans, reflecting the parallel findings in disease risk associations. The tree data would position rhesus monkeys closer to humans in the tree compared to more distantly related species like mice.

Incorporating both the matrix data and the phylogenetic tree data allows us to better understand the genetic relationships and evolutionary history of the CFTR gene in these species. The matrix data quantifies the sequence similarity, while the tree data visualizes the evolutionary connections among them. The alignment or mismatch between these two sets of data can provide insights into the conservation and divergence of the CFTR gene in different species.

4. For APOE Gene

The APOE gene, which encodes apolipoprotein E, is found in various primates, including humans. It plays a crucial role in lipid and cholesterol metabolism.Similar to humans, chimpanzees possess APOE alleles and share a common ancestor with humans.Gorillas also have APOE alleles, although less research has been conducted on this gene in gorillas compared to humans and chimpanzees.Orangutans, like other primates, likely have APOE genes.Research

on APOE in orangutans is limited, making it difficult to draw direct comparisons with humans.

Some associations between APOE alleles and disease risk in rhesus monkeys parallel findings in humans, making them a useful model for human health research.Mice have a functionally equivalent gene to human APOE called Apoe.Mouse studies have contributed significantly to understanding APOE's roles in lipid metabolism, neurobiology, and disease.While some findings are relevant to humans, mice differ significantly from humans in terms of physiology, diet, and disease susceptibility, so direct comparisons have limitations.
The APOE gene is conserved across primates, and various species have their own allelic variations.