Phylogenetic Analysis of BRCA1: Exon 11 in Certain Primates

Kaitlyn Barnes, Rachel Rompala, & Cameron Vannoy

Department of Ecology, Evolution and Organismal Biology, Iowa State University

EEOB 563: Molecular Phylogenetics

Dr. Dennis Lavrov

April 15, 2021

I. GitHub

Link: https://github.com/vannoycm/EEOB563_FinalProject

II. Introduction

The BRCA1 (BReast CAncer gene 1) and BRCA2 (BReast CAncer gene 2) genes are responsible for producing proteins that repair damaged DNA which leads to them being referred to as tumor suppressor genes (National Cancer Institute at the National Institutes of Health, 2020). In humans, each parent passes on one copy of each gene to their offspring. Unfortunately, parents sometimes carry harmful variants of the BRCA genes that increases the chances of the offspring developing certain types of cancer, specifically breast and ovarian cancer. In certain populations the likelihood an individual will inherit a mutation of BRCA1 or BRCA2 gene can vary. For the general population, 0.2-0.3% (or 1 in 400 people) will carry a harmful variant, while people of Ashkenazi Jewish descent will have a 2.0% chance of having one of the harmful variants (Nelson et al., 2013).

Variants appear in different frequencies depending on the racial/ethnic group and geographic location of the population, and understanding these differences is important for public health reasons. For example in the United States of America, certain BRCA1 variants are more common than others while a different variant of BRCA1 is the most common when looking at a study done on populations in Iceland (Hall et al., 2009). As a woman ages, the likelihood of developing cancer increases, especially if she inherited a harmful variant of BRCA1 or BRCA2. When looking at cases of breast cancer and the general population, a woman's chances of developing breast cancer increased from 13% to 55-72% if she inherited a harmful BRCA1 variant (Kuchenbaecker et al., 2017). A similar correlation is seen with ovarian cancer as a woman's chances of developing the cancer increase from 1.2% to 39-44% with the presence of a harmful BRCA1 variant.

In 2000, a study was published comparing the BRCA 1 gene in humans to the BRCA gene 1 in other primates to show the adaptive evolution of the gene (Hutley et al., 2000). Understanding the action of BRCA 1 in non-human primates could improve understanding of the gene in humans. As a result of research being poured into breast cancer these past few decades, the BRCA1 gene has become most commonly known for the role that the gene's mutated version plays in the development of breast cancer amongst females. BRCA1 sequences have been historically collected from breast cancer patients, ovarian cancer patients, and normal tissue samples. BRCA1 is composed of 24 exons with roughly 60% of proteins embedded within exon 11, however there seems to be several mRNA splicing isoforms (Hutley et al., 2000, 1). Although exon 11 is not incorporated into all of the common mRNA splicing isoforms, its alternative splicing presents a difference between the BRCA1 levels in cancer and normal tissues (Tommaro et al., 2012, 2).

Figure 1

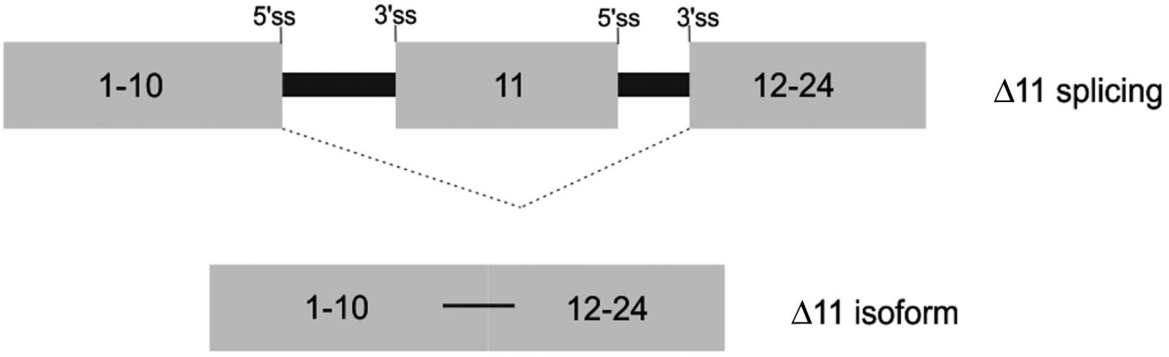

Figure 1 demonstrates the complete mRNA splicing of exon 11 (**Δ**11 isoform). This isoform of BRCA1 has not demonstrated any relevance in cancer development. **Δ**11 has been involved with cell proliferation and cell death (Tommaro et al., 2012, 2).

Figure 2

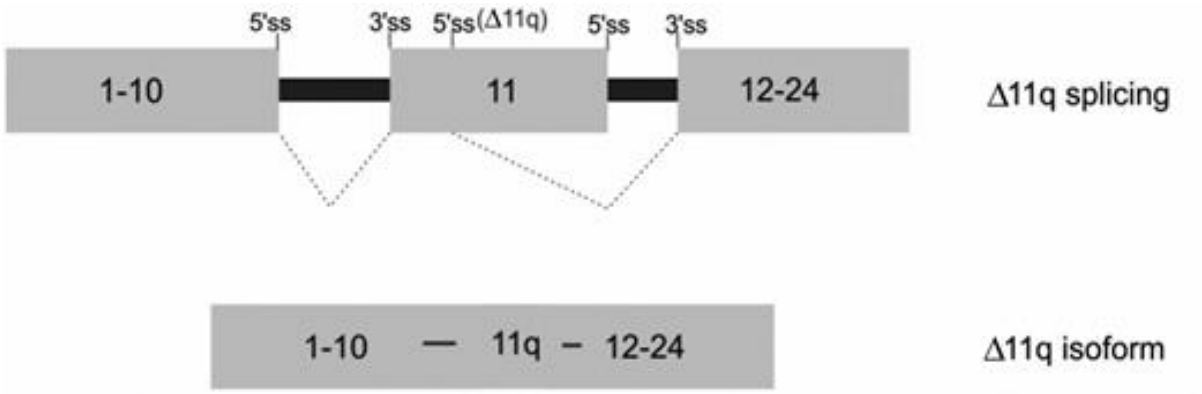

Figure 2 exhibits the partial mRNA splicing of exon 11 (**Δ**11q isoform). Nucleotides 905-4215 are not included within exon 11. The overexpression of BRCA1 **Δ**11q possesses the ability to suppress breast cancer cell growth. It is important to note that cytoplasmic **Δ**11q has been correlated with tumor formation (Tommaro et al., 2012, 2).

The question we would like to analyze is which primate is most closely related to humans when looking at exon 11 in BRCA1 tumor suppressor gene? We hypothesize that chimpanzees will be most closely related to humans when examining exon 11 in the BRCA1 tumor suppressor gene, because chimps are the primates that are widely known to be the closet relative of homosapiens.

To meet this goal we will be performing a distance analysis using Neighbor Joining and UPGMA and a Bayesian Analysis of the BRCA1: Exon 11 protein sequence collected from six different primates.

III. Methods

A. Sequence Selection

After a literature review, we determined that for evolutionary analysis of BRCA1 across primates we would be using the protein sequence of exon 11 of the BRCA1

gene. Sequences for exon 11 protein products were downloaded from the NCBI GenBank as follows:

Organism	**Accession Code**
Human (*Homo sapiens*)	AF005068 (Holt, 1997)
Orangutan (*Pongo pygmaeus*)	AF019077 (Hacia, 1997)
Gorilla (*Gorilla gorilla*)	AF019076 (Hacia, 1997)
Chimpanzee (*Pan troglodytes*)	AF019075 (Hacia, 1997)
Macaque (*Macaca mulatta*)	AF019078 (Hacia, 1998)
Howler Monkey (*Alouatta seniculus*)	AF019079 (Haccia, 1998)

Table 1: Accession Codes of Organism Protein Sequences Used

Sequences were selected based on the organism (primates known to be evolutionarily close to humans), and on the availability of exon 11 protein sequence data.

B. Sequence Alignment Using MAFFT

Before phylogenetic analysis of the protein sequences can be completed we first needed to align them using MAFFT (Multiple Alignment using Fast Fourier Transform). For this, we used an online service that can be found at https://www.ebi.ac.uk/Tools/msa/mafft/.

The alignment was performed using the BLOSUM62 matrix with a 1.53 gap open penalty, 0.123 gap extension, maxitrate of 2, no performed FFTS, and a tree rebuilding number of 2.

C. PAUP

The first analysis we ran using PAUP created the distance matrix from the alignment data. This is simply done by executing the alignment file and then typing the “showdist” command.

After we have obtained a distance matrix we can move to completing Neighbor-Joining Analysis. In this analysis we were looking for the minimum evolution neighbor-joining tree with bootstrapping (1000). This allows us to show confidence scores on branches of the tree obtained.

We then once again used the distance matrix found from the alignment data using PAUP and created a UPGMA tree.

D. Bayesian Analysis using Mrbayes

We completed an unpartitioned analysis using the MrBayes program. MrBayes uses a two-phased Markov Chain Monte Carlo (MCMC) approach to find trees. In phase one the program works to find the parameters that bring the analysis near the maximum likelihood, then in phase two the program works to explore the settings near those parameters (Pederson). This approach allows for the approximation to become highly accurate.

When running the analysis we continued the runs until they showed a difference of less than 0.01. This let us know that we had reached stationarity and it was fine to move forward with the analysis.

We then used the sumt command to generate a tree with the posterior probability values. The trees show the estimated branch lengths and posterior probabilities of each tree.

The settings used were as shown below:

```
MrBayes > showmodel

   Model settings:

      Data not partitioned --
         Datatype  = Protein
         Aamodel   = Poisson
                     Substitution rates are fixed to be equal
         Covarion  = No
         # States  = 20
                     State frequencies are fixed to be equal
         Rates     = Gamma
                     The distribution is approximated using 4 categories.
                     Likelihood summarized over all rate categories in each gene
ration.
                     Shape parameter is exponentially
                     distributed with parameter (0.05).

   Active parameters:

      Parameters
      --------------------
      Statefreq           1
      Shape               2
      Ratemultiplier      3
      Topology            4
      Brlens              5
      --------------------

      1 --  Parameter  = Pi
            Type       = Stationary state frequencies
            Prior      = Fixed (equal frequencies)

      2 --  Parameter  = Alpha
            Type       = Shape of scaled gamma distribution of site rates
            Prior      = Exponential(0.05)

      3 --  Parameter  = Ratemultiplier
            Type       = Partition-specific rate multiplier
            Prior      = Fixed(1.0)

      4 --  Parameter  = Tau
            Type       = Topology
            Prior      = All topologies equally probable a priori
            Subparam.  = V

      5 --  Parameter  = V
            Type       = Branch lengths
            Prior      = Unconstrained:GammaDir(1.0,0.1000,1.0,1.0)
```

Figure 3: Settings used for MrBayes Analysis

These settings were based off of the in-class lab activity but modified for being an Amino acid sequence. The model was kept simple to avoid human error as we are inexperienced with the program and did not want to introduce human error.

IV. Results

A. Distance Matrix Generated Using PAUP

```
paup> showdist

Uncorrected ("p") distance matrix
  1801 characters are included
  All characters have equal weight

                    1         2         3         4         5         6
1 AAB61673 1        -
2 AAC39584 1  0.02103         -
3 AAC39583 1  0.02279   0.01227         -
4 AAC39585 1  0.03856   0.02892   0.02892         -
5 AAC39586 1  0.07375   0.06497   0.06497   0.07024         -
6 AAC39587 1  0.12841   0.12137   0.12049   0.12665   0.12324         -
```

Figure 4: Distance Matrix for Primate Protein Sequences

B. Neighbor Joining Unrooted Consensus Tree

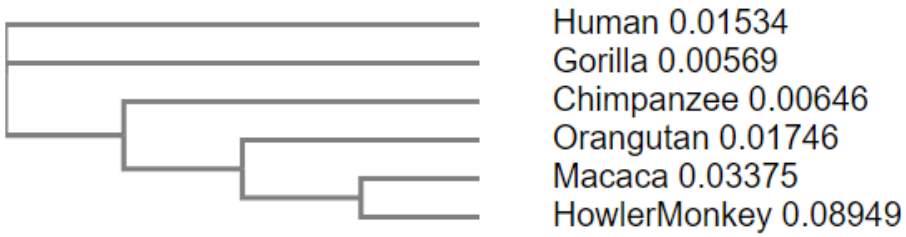

Figure 5: Unrooted Neighbor Joining Consensus Tree Using Uncorrected p-Distance Values

C. PAUP

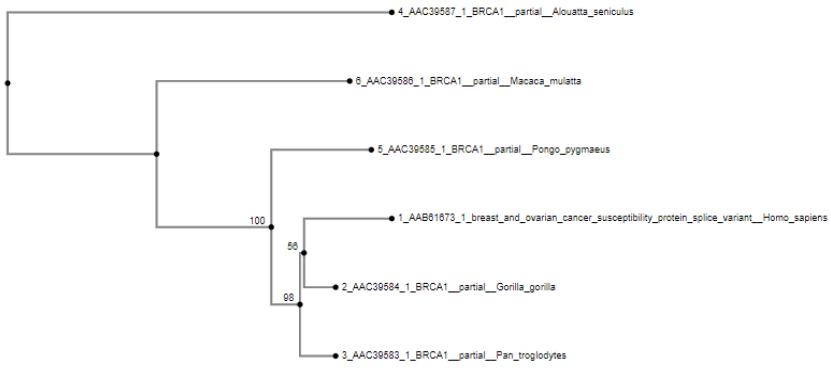

Figure 6: Neighbor Joining Tree Based on Uncorrected p-Distances

Distances calculated from the alignment.

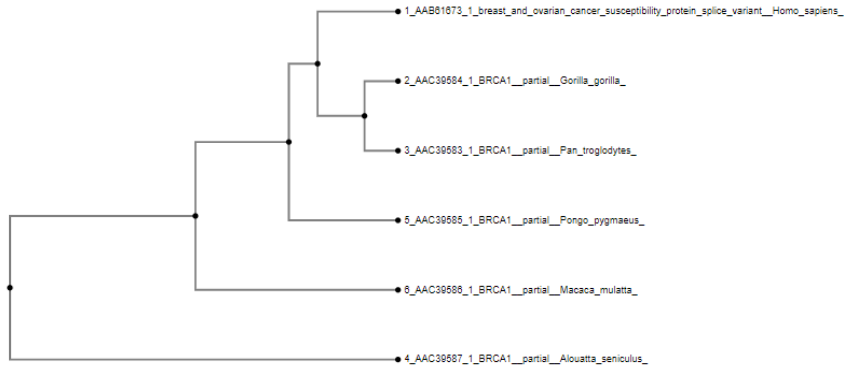

Figure 7: Alignment tree made using UPGMA Algorithm

D. MrBayes

```
Clade credibility values:

/------------------------------------------------------------------ AAB61673_1 (1)
|                                                                    HUMAN
|------------------------------------------------------------------ AAC39584_1 (2)
+                                                                    GORILLA
|               /---------------------------------------------------- AAC39583_1 (3)
|               |                                                     CHIMPANZEE
\-------98------+               /---------------------------------- AAC39585_1 (4)
                |               |                                     ORANGUTAN
                \------100------+               /------------------ AAC39586_1 (5)
                                \------100------+                     MACACA
                                                \------------------ AAC39587_1 (6)
                                                                    HOWLER MONKEY

Phylogram (based on average branch lengths):

/-------- AAB61673_1 (1) HUMAN
|
|--- AAC39584_1 (2) GORILLA
+
|/--- AAC39583_1 (3) CHIMPANZEE
||
\+  /--------- AAC39585_1 (4) ORANGUTAN
 |  |
 \--+         /------------------ AAC39586_1 (5) MACACA
    \---------+
              \------------------------------------------------- AAC39587_1 (6)
                                                                  HOWLER MONKEY
|---------| 0.020 expected changes per site

Calculating tree probabilities...

Credible sets of trees (3 trees sampled):
   99 % credible set contains 2 trees
```

Figure 8: MrBayes Trees: Clade Credibility Values and Distance Phylogram.

V. Discussion of Results

A. PAUP Distance Analysis, Neighbor-Joining, and UPGMA Analysis

For these analyses we used the uncorrected p-distances. In future work we would ideally correct these using a substitution matrix. This would increase the credibility of our tree.

Using Mafft and PAUP, we were able to align and analyze the distance matrix of our data sets. Our PAUP distance analysis after "showdist" showed us a distance matrix that would be used to configure our NJ and UPGMA trees. When performing the NJ analysis we used a Jukes-Cantor distance model as well as bootstrapping to form a tree from the distance matrix. The resulting neighbor joining tree was an unrooted consensus tree. This tree shows that maccas (*Macaca mulatta*) and howler monkeys (*Alouatta seniculus*) are closely related, yet are the most distantly related to humans (*Homo sapiens*). Despite being unrooted, the initial tree does not shift dramatically when it becomes rooted in the UPGMA and NJ Minimum Evolution trees. An interesting thing to note is that on this tree the human and gorilla are grouped together even though there is low support for the grouping with a score of 56. There are many possibilities for why this may have happened. The first is that we were expecting results to the species tree (chimpanzee closest to human) and based on this gene that simply may not be the case. Another reason is that there may be deep coalescence that causes the grouping but would also explain the low support.

When looking at the UPGMA tree, one is able to establish the differences between it and the NJ Minimum Evolution tree. While the two trees are remarkably similar, their arrangement of branches at a specific node is different. The neighbor joining tree has humans (*Homo sapiens*) and gorillas (*Gorilla gorilla*) being more closely related and sharing a common ancestor with

chimpanzees (*Pan troglodytes*). In the UPGMA tree, however, gorillas (*Gorilla gorilla*) and chimpanzees (*Pan troglodytes*) are considered more closely related to each other and they share a common ancestor with humans (*Homo sapiens*). Based on these results, gorillas and chimpanzees are shown to be the most similar to humans when comparing BRCA1: Exon 11 which makes them excellent candidates to continue to study how the gene and its variants function. However, it is important to note that the UPGMA is known to have large error due to the fact that this algorithm assumes that the rate of change is the same for all clades when we know that this is not the case.

B. Bayesian Analysis

When examining the known species trees for primates, we expected that chimpanzees should show the closest relationship to humans with gorillas following close behind and extending out from there. This is indicated by the expected changes per site, however, as the tree is unrooted this is relative information and must be interpreted as such. Relative to humans, our Bayesian analysis found that the Macaca and Howler Money had the most expected changes per site which was expected as they diverged earlier than other primates in the species tree. We were surprised to find that Chimpanzees when compared to humans showed more expected changes per site than the gorillas. This data argues that the human and gorilla variants of the BRCA1: Exon 11 show the closest relationship on the level of this specific gene. This conclusion was confirmed by high clade credibility values (98, 100, 100).

VI. Conclusion

We hypothesized that if chimpanzees are the most closely related to humans on the species tree, then they would show the least changes from humans at the level of

BRCA1:Exon 11. We have proven our hypothesis to be incorrect. Through a thorough examination of our Neighbor Joining Tree (uncorrected p-distances) and the trees generated through our BrBayes analysis we showed that *Homo sapiens* were most closely related to gorillas instead of chimps at the level of BRCA1: Exon 11.

There are many possible reasons for these results. Firstly, the support for the grouping of humans and gorillas was very low (56) indicating little support. This lack of support indicates that the result may not be completely accurate. As outlined in the Discussion of Results, this low support grouping may be due to deep coalescence or it may be accurate and be an example of how the species tree can differentiate from the gene specific tree. When looking at our Bayesian analysis tree is difficult to interpret with complete certainty because we did not root our tree at the Howler Monkey as would have been suggested from the species tree. Overall it is important to remember that the examination of trees is a subjective process and is highly susceptible to human error in interpretation.

Future work that we believe could strengthen this study would be to examine the alignments and analyze the human only changes compared to sites conserved in the other primates. For this type of analysis PAML could be used to observe possible patterns of synonymous and nonsynonymous substitutions. Another way we could strengthen the study would be to use the corrected p-distance values. In our current study we used the uncorrected values and this causes lower confidence in the results.

Through the duration of this project, the class and its content came full circle in a way. We were able to laugh over when we struggled with logging onto HPC class or spent way too long developing an alignment by hand using the Needleman-Wunch algorithm and the Blosum 62 matrix as we successfully mastered MAFFT, PAUP, and MrBayes on our own. Our confidence and familiarity with phylogenetic analysis has skyrocketed along with our experiences with literature review and understanding. Using the NCBI

GEO database, PubMed, and several other sources, we were able to develop an understanding of BRCA1 molecular markers and specifically compare exon 11 conservation across species. The results were interesting and applicable as cancer is a prominent topic and impacts thousands of people every day. We anticipate that we will be able to apply the knowledge gained from this project as well as this class in our future endeavors.

Bibliography

Hall MJ, Reid JE, Burbidge LA, et al. BRCA1 and BRCA2 mutations in women of different ethnicities undergoing testing for hereditary breast-ovarian cancer. *Cancer* 2009; 115(10):2222–2233.

Holt, J. T. (1997, June 25). *Homo sapiens breast and ovarian cancer susceptibility protein splice variant (BRCA1) mRNA, complete cds*. NCBI GenBank. Retrieved April 7, 2021, from https://www.ncbi.nlm.nih.gov/nuccore/AF005068

Kuchenbaecker KB, Hopper JL, Barnes DR, et al. Risks of breast, ovarian, and contralateral breast cancer for BRCA1 and BRCA2 mutation carriers. *JAMA* 2017; 317(23):2402–2416.

National Cancer Institute at the National Institutes of Health. (2020, November 19). *BRCA Gene Mutations: Cancer Risk and Genetic Testing*. National Cancer Institute. Retrieved April 15, 2021, from https://www.cancer.gov/about-cancer/causes-prevention/genetics/brca-fact-sheet

Nelson HD, Fu R, Goddard K, Mitchell JP, Okinaka-Hu L, Pappas M, Zakher B. Risk Assessment, Genetic Counseling, and Genetic Testing for BRCA-Related Cancer: Systematic Review to Update the U.S. Preventive Services Task Force Recommendation. Evidence Synthesis No. 101. AHRQ Publication No. 12-05164-EF-1. Rockville, MD: Agency for Healthcare Research and Quality; 2013.

Pavlicek, A. (2004, November 15). *Evolution of the tumor suppressor BRCA1 locus in primates: implications for cancer predisposition*. Oxford Academic. Retrieved April 7, 20201, from https://academic.oup.com/hmg/article/13/22/2737/610248

Pederson, A. *Bayesian Phylogenetic Analysis*. DTU Bioinformatics. Retrieved April 25th, 2021

from http://www.cbs.dtu.dk/dtucourse/cookbooks/gorm/27615/bayes1.php.

Tommaro, C., Raponi, M., Wilson, D., & Baralle, D. (2012). *BRCA1 exon 11 alternative splicing, multiple functions and the association with cancer* (4th ed., Vol. 40). Biochemical Society Transactions. https://portlandpress.com/biochemsoctrans/article/40/4/768/85490/BRCA1-exon-11-alternative-splicing-multiple

Acknowledgements

Cameron Vannoy:

As my part of the project I helped with the research process and development of the research question. I completed the alignment and the analysis using PAUP. I also ran the Bayesian Analysis and prepared the groundwork for Kaitlyn to finish the trees from that analysis. I completed the results section for the Bayesian analysis. I checked the citations to ensure that each time we used data it was cited properly.

Kaitlyn Barnes:

I spent time utilizing online scholarly resources in order to learn more about the BRCA1 gene as well as exon 11 specifically in order to develop a background/introduction section. I used Mr. Bayes alongside Cameron to develop the trees for further analysis. I also unsuccessfully attempted RAxML-NG as well as rerooting the trees with FigTree. I made corrections to the paper after peer reviews were completed. I developed the concluding thoughts on both the project and the class as a whole. In this section, I also went over the other possible questions and future work that could be done. I developed a new, more specific question for our paper to focus on after our discussion on Monday.

Rachel Rompala:

As my contribution to the project I gathered journal articles related to the BRCA1 gene in order to write a developed background. Using the data we collected from MAFFT and PAUP, I unsuccessfully attempted to run PHYLIP and ran one of the initial MrBayes programs that had errors resulting in no tree being formed. I wrote the discussion on the results of PAUP and helped edit our paper based on the suggestions given to us by our peer reviewers and professor.

Thank you to Dr. Lavrov and our peer reviewers for making this project successful.