8-11-2022

Bioinformatics Analyses of New Genes of Focus in the Research of Autism with Dogs, Rats, and a Variety of Other Model Organisms

Aaron Pennington
adp2992@rit.edu

Bioinformatics Analyses of New Genes of Focus in the Research of Autism with Dogs, Rats, and a Variety of Other Model Organisms

Aaron Pennington

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Bioinformatics

Thomas Gosnell School of Life Sciences

College of Science

Rochester Institute of Technology

Rochester, NY

August 11, 2022

Rochester Institute of Technology
Thomas H. Gosnell School of Life Sciences
Bioinformatics Program

To: Head, Thomas H. Gosnell School of Life Sciences

The undersigned state that Aaron Pennington, a candidate for the Master of Science degree in Bioinformatics, has submitted his thesis and has satisfactorily defended it.

This completes the requirements for the Master of Science degree in Bioinformatics at Rochester Institute of Technology.

Thesis committee members:

Name	**Date**
Gary R. Skuse, Ph.D. Thesis Advisor	25 July 2022
André O. Hudson, Ph.D.	25 July 2022
Julie A. Thomas, Ph.D.	25 July 2022

Abstract:

Autism is a mental disorder in which multiple genes are involved in the development of its various symptoms. However, as a result of the challenges inherent to identifying the responsible genes, many studies are ongoing and inconclusive. [1] To date, studies have shown that two genes, BAG3 and PALLD, were upregulated in autistic individuals. [2–6] BAG3 encodes BAG cochaperone 3 [32]; while PALLD encodes Palladin, Cytoskeletal Associated Protein [33]. This project analyzed the similarities and differences between the human BAG3 and PALLD genes and those in various model organisms. While it is likely that epigenetic modifications affect the expression and activity of these genes, they were not a focus of this work due to the paucity of prior research. [7–11]
It is hypothesized that there is a model organism best suited for studying BAG3 and/or PALLD for a better understanding of its role in human autism.

Comparisons were performed using a variety of bioinformatics tools in order to identify mRNA variants in model organisms. The most similar variants of BAG3 and PALLD were then assessed for the significance of these variations on the predicted structure and expression of the encoded protein. Finally, the same variants' regulatory regions were predicted and compared in order to identify similarities and differences found upstream of the BAG3 and PALLD genes.

The findings of this study suggest that the BAG3 and PALLD protein structures from the model organisms dog and rat are sufficiently similar compared to those in humans so they can be used to better understand the genes in humans diagnosed with autism. Furthermore, there are similarities in the regulatory regions predicted for select model organisms. However, these regulatory regions are in areas where there is very little known and therefore where future research into these genes' effect on autism should be focused. The results from this work provide guidance as well as evidence justifying the need for future research and experimental manipulation of BAG3 and PALLD in these model organisms.

Abbreviation list

BAG3	BAG cochaperone 3
PALLD	Palladin, Cytoskeletal Associated Protein
ZNF804A	Zinc Finger Protein 804A
SHANK3	SH3 and multiple ankyrin repeat domains 3
RCrusI	Right Crus I
FoxP1	Forkhead box P1
Itpr3	Inositol 1,4,5-Trisphosphate Receptor Type 3
BTBR T+Itpr3tf/J	Mice with a deletion in the Itpr3 gene
+/+	2 wild type copies of the gene
+/-	Heterozygous with 1 wild type copy of the gene and 1 mutant copy of the gene that are not expressed
-/-	2 mutant copies of the gene that are not expressed
NCBI	National Center for Biotechnology Information
BLAST	Basic Local Alignment Search Tool
MEGAX	Molecular Evolutionary Genetics Analysis with cross-platform functionality
OAT	Orthologous Average Nucleotide Identity Tool
MUSCLE	Multiple Sequence Comparison by Log-Expectation

Introduction:

Autism is a mental disorder in which multiple genes are involved in the development of its various symptoms.[1] However, as a result of the challenges inherent to identifying the responsible genes, studies are ongoing. Some of these studies involve the use of model organisms as substitutes and preliminary tests before human trials can be proposed.

Autism is part of Autism Spectrum Disorder (ASD) which is a disability which involves abnormal development of the brain during the first 3 years of life. The reason that autism is considered a spectrum is because of the wide variety of effects it can have on individuals with ASD. [1]While not all genes involved have been identified there are a few known factors that increase the chance of ASD developing. First, having ASD in the family increases the odds of an individual developing it. Furthermore, mutant variants and/or differently expressed ZNF804A gene and NMDA receptor associated genes in the glutamate signaling pathway also increase the chances of developing ASD. Finally, being exposed to Thalidomide during pregnancy increases the chance of the offspring developing ASD due to its effects on the development of the brain stem. [12]The first signs of ASD show in the first year of life. These symptoms could include

avoiding eye contact, not showing facial emotions, not responding to others physically or verbally, and using few gestures. This may further develop in the second year to include not sharing objects of interest, not using or responding to pointing, and not noticing when others are hurt or sad. Lastly in the third year of life the individual may have difficulty understanding others' emotions, sharing their own thoughts, and not projecting personalities onto objects when playing. There are also some symptoms that can develop at any time. These can include obsession with maintaining the order of objects, repetition of words or phrases, obsessive focus on particular interests, becoming angered when minor changes occur, repeating routines, and maintaining motion such as spinning rocking or flapping hands. Some other effects that the symptoms can cause are being hyperactive, epilepsy, anxiety, lack of fear, abnormal eating and sleeping habits, abnormal emotional reactions, gastrointestinal issues, various additional psychotic disorders, delayed language, learning, and motor skills. [1,12]As a result of all its variations, ASD for some individuals can be somewhat hindering to crippling on a daily basis. [1]For some individuals ASD can make adjusting into society difficult due to the lack of experience or skills with social interactions relative to their age. [13]While there is no cure for ASD, some medications that decrease psychotic symptoms can reduce these symptoms in some individuals. In adult life some individuals may have improved symptoms. [1]

Previous studies have used a variety of model animals to study autism. Mice and rats were used for experiments involving genetic modifications and/or exposure to chemicals during development in order to create autism like symptoms. [14–16]Zebrafish have also been used due to their well characterized genetics and their extensive use as a model organism to understand human diseases .[17] More uniquely, zebra finches play a key role in autism research due to the ability to study the effect of autism on how they learn their songs. This work provides insights into the auditory effects of autism. [18]Lastly, research is being conducted into the viability of using dogs as models for autism because dogs are more social animals than the other models currently in use. To go along with this, previous studies have found that dogs can express ASD like behaviors. [19]

Blind cave fish, *Astyanax mexicanus,* are a species of fish found in Texas and Mexico which have evolved in lightless caves resulting in the loss of their eyes, the development of extra sensitive lips to taste their surroundings, and other unique behavioral patterns. Studies have found that these behaviors correlate with behavioral traits found in human autism. Furthermore, medication used to reduce autism like symptoms in humans were significantly effective at reducing these behavioral traits in blind cave fish. This finding suggests that some of the same genes that cause these behaviors in blind cave fish may also be responsible for autism in humans. [2]

Rattus norvegicus (rats) are a common model organism used to better understand humans. For human autism the main focus is on their behavioral patterns throughout life when ASD suspected genes are manipulated. In one study, the offspring of heterozygous rats with no SHANK3 expression and wild type were studied in order to observe the development of the offspring. The purpose of this study was to identify what behavioral phenomenon would occur for the rats that were SHANK3 -/- and those that were +/-. This was done by observing the rat's reactions to ultrasonic verbal (USV) patterns that were used in pro-social behavior. They found that in the

juvenile ASD rats their social communications were hindered. In particular, communications from the male SHANK3 -/- were bidirectional in that there is a reduced response to USV. Furthermore, male SHANK3 -/- spend less time playing with other juveniles, social sniffing, or engaging with other rats. The researchers believe that these changes were only detectable in males due to female rats naturally being less social than male rats.[16]

Mus musculus (mouse) is another common model organism for studying autism. Two of the genes that are looked at in mice regarding autism are Right Crus I (RCrusI) and Itpr3. RCrusI had already been associated with functions of inferior parietal lobule in ASD children. Therefore, studying the mutation of this gene in mice provides an area of research for investigating cerebellar abnormalities. This research has resulted in the findings of RCrusI affecting in particular the Purkinje neuron TscI. Furthermore, the inhibition of RCrusI, which was chemogenetically mediated, was able to sufficiently create ASD related behaviors in the mice including repetitive, restricted, and associated social behaviors.[14] Itpr3 in the past had been investigated for researching diabetes-induced nephropathy, insulin-resistance, and phenyloketonuria. However, it was also found to be an effective model for autism in mice due to it producing consistent autistic behaviors. Therefore, a strain of mice with a deletion in the Itpr3 gene called BTBR T+Itpr3tf/J(BTBR) has been used to study the underlying mechanisms for autistic behavior on a molecular level. Some resulting effects include increased production of kynurenic acid, the abnormal expression of multiple genes in the hippocampus, aberrant regulation of actin cytoskeleton, and the down regulation of myelin-related proteins along with stable tubule only polypeptide protein (STOP). [15]

*Danio rerio (*zebrafish) have been another commonly used model organism for studying biology. It has also been used to study SHANK3, although in zebrafish this gene is represented by SHANK3a and SHANK3b. However, a study looking at SHANK3 in zebrafish used CRISPR/Cas9 to stop its expression, in comparison to the study in rats which used biochemicals to remove SHANK3 expression. Specifically, they cause a C terminal frameshift mutation in both versions of SHANK3. In this study, the researchers looked at the side effects on the digestive track of the zebrafish. They found that the mutant zebrafish had a much slower rate of digestive track movement that did not function properly due to a decrease in serotonin-positive enteroendocrine cells. When crossbreeding with wild types the heterozygous individuals expressed a faster digestive track while still not being the same rate as wild type. Furthermore, when these heterozygous zebrafish had offspring that were exposed to mRNA of human SHANK3, they produced digestive tracts that functioned more similar to wild type in terms of the amount of serotonin-positive enteroendocrine cells present but were still slower than normal. [17]

Taeniopygia guttata (zebra finches) have a unique perspective in terms of autism study from observing how they learn songs. In a study they observed how the knockdown of FoxP1 expression affects the learning and memorization of mating songs in juvenile zebra finch. The knockdown was performed by developing a short hairpin RNA against FoxP1. As a result, they found that the juvenile zebra finch were unable to remember the songs in the long-term. However, they were still able to correctly imitate the song immediately after hearing it. Upon further investigation, the researchers determined this was due to the knockdown of Fox P1

decreasing the plasticity of HVCX neurons which are essential to forming memories in zebra finches.[18]

Since *Canis lupus familiaris* (dogs) have not been accepted fully as a model for autism, current research is demonstrating the many benefits that would result from using dogs in autism studies. The first main factor is that dogs are the most social animals out of all the ones discussed previously. This also comes along with needing to identify the mechanisms, systems, and treatments of dogs with ASD like behaviors compared to humans. [19,20] The sociability can be measured with visual cue-response association task and measuring social competence. One paper focused on the impaired social motivation found in autistic humans and looked for similar traits in dogs. Using diagnostic criteria of ASD they developed a system for finding dogs with ASD like behaviors. Next, they tested how easily these dogs were distracted by social and nonsocial stimuli during a task. They found that similar to human ASD with similar experimental setups, wild type dogs were distracted more by social stimuli than dogs with ASD like behavior. [19]Other papers have also looked at the symptoms found in dogs with ASD like behaviors and the underlying etiology. In one paper dogs were found to have more phenotypic similarity to human ASD than rodents. Furthermore, there is evidence suggesting that the neurocognition may also be more similar in dogs compared to rodents. [20]

In a study observing the genetics of blind cave fish relative to autistic humans, scientists analyzed the expression of all genes present in both species compared to the most closely related relatives or wild type forms respectively.[2] The results suggest that BAG3 and PALLD could be genes of interest in the study of autism in various model organisms. Across the three studies cross-referenced from human gene expression to the studies' observations of blind cave fish gene expression, all found that there was a significant increased expression of BAG3 and PALLD.[2–5] Furthermore, both of these genes have been well studied in terms of their function and structure within humans and other model organisms. [21,22]

This project will analyze the genetic differences between the BAG3 and PALLD genes found in humans, blind cave fish, mice, rats, zebrafish, zebra finch, and dogs, with red-bellied piranhas used throughout as a control. This will be done using BLAST [23,24], global alignments, and multiple sequence alignments to identify which variants are most closely related and to find what regions of the gene have the most variation and similarity. Furthermore, genes will be measured for the significance of these variations on the predicted structure of the protein. Lastly, the mechanisms of regulation will be compared to identify similarities and differences. This will provide the evidence needed for future research to start experimenting with BAG3 and PALLD in model organisms for the purpose of identifying their role in autism.

Materials and Methods:

SEQUENCE-BASED SEARCHES USING BLAST

For sequence-based genetic comparisons BLASTN [23] and BLASTP [24] were used to compare model organisms and identify what sequences were most similar. The percentage of similarity and areas with the most variation were identified. BLAST [23,24] stands for "Basic Local

Alignment Search Tool" and is a web program that takes an input sequence and looks through the NCBI nucleotide database with BLASTN [23] or the NCBI protein databases with BLASTP [24]. This is in order to identify the most similar sequences available in the respective databases compared to the input sequence. This project adjusted the settings so that BLAST [23,24] would exclusively compare to the organisms being studied under the mega blast optimization. Other parameters were set to default. This provided a reference point for how similar and how different the model organisms' versions of these genes were. BLAST [23,24] on the default settings were used for the primary transcription of both genes from the organisms being studied. Initially 100 results were given with the default settings. However, this did not include results for all target organisms so this was expanded and redone with the settings adjusted to allow for up to 5000 results. Furthermore, the results were limited to only show matches from the organisms being studied.

The following taxid were used for each of the given organisms; *Homo sapiens* (taxid:9606), *Astyanax mexicanus* (taxid:7994), *Rattus norvegicus* (taxid:10116),*Mus musculus* (taxid:10090), *Danio rerio* (taxid:7955), *Taeniopygia guttata* (taxid:59729),*Canis lupus familiaris* (taxid:9615), and *Pygocentrus nattereri* (taxid:42514) The exact sequences used were from the following NCBI Reference Sequence codes - BAG3: human version NM_004281.4, mouse version NM_013863.5, zebrafish version NM_001003533.2, rat version NM_001011936.1, dog version XM_544046.6, blind cave fish version XM_007252251.3, red-bellied piranha version XM_017696524.2, and zebra finch version XM_030276542.3; PALLD: human version NM_001166108.2, mouse version NM_001293772.2, blind cave fish version XM_007245327.3, red-bellied piranha version XM_017687824.2, zebrafish version XM_021468620.1, zebra finch version XM_032748140.2, dog version XM_038435147.1, and rat version XM_039094916.1

Additionally, well documented genes that have been associated with autism were compared for similarities between mutant and wild type forms if they were available. One gene that was looked into is ZNF804A because it has already been associated with an increased risk in ASD and may be able to act as a reference point for how similar the autism variant of a gene needs to be in humans. [12]

SEQUENCE ALIGNMENTS

With this baseline set, the BAG3 and PALLD genes from dog, rat, mouse, zebrafish, and zebra finch were compared individually to humans, blind cave fish, and red-bellied piranhas. This was done to determine if the genes have enough similarity to be studied in order to provide insights into the genetic basis of human autism. This also included looking to see if there are similar areas of conservation and similar areas carrying mutations. Furthermore, there was a global alignment for all versions of the BAG3 and PALLD genes found in these selected organisms. The global alignment was used to reveal what parts of the genes are conserved the most. This was done using the program Molecular Evolutionary Genetics Analysis (MEGAX) version 10.0.5 [25] with the default settings. MEGAX [25] is a program designed to be a cross-platform program which contains multiple statistical tools for analyzing molecular data. This project primarily used MEGAX [25] for its alignment tool. [26]

MEGAX [25] has two alignment options using the ClustalW [27] or MUSCLE [28] algorithms. ClustalW is designed as a general alignment algorithm that can be applied to multiple types of data sets. [27] MUSCLE is an algorithm optimized for accuracy and in particular to be used with protein sequences, it however can still be used with DNA sequences. [28] These tools were used due to the program being able to generate a global alignment that can be viewed on a nucleotide level. Furthermore, MEGAX [25] can also accept amino acid inputs which were used later. [25] CJ Bioscience's Orthologous Average Nucleotide Identity Tool (OAT) version 0.93.1 [29] allowed for the creation of a dendrogram showing the evolutionary relationship between these genes. The dendrogram was useful for tracking the changes of these genes on an evolutionary scale. It was created by using an optimized version of a simple Average nucleotide identity algorithm called OrthoANI.[30] OrthoANI [30] is an improvement due to it decreasing the number of reciprocal values that were not symmetrical in comparison to ANI. This was done by first breaking the sequence into 1020 bp fragments. Next, the algorithm removes any fragments that are shorter and uses only orthologous DNA fragments to align the full sequences. [30]Lastly the algorithm will also generate a matrix showing the percentage similarity between each pair of inputted genes. Furthermore, OAT [29] provides the GC content of each gene providing another variable for comparison. [31]

There were initial plans to look at the autistic versions of BAG3 and PALLD. However, due to these genes not being confirmed to contribute to autism no such sequences were available on NCBI.[32,33] While not successful, the following methods were used in an attempt to find these sequences. The papers looking at the expression differences between autistic and wild type humans were looked at again for supplementary material and external links that gave access to the raw data used. This gave access to PROBE and Probe ID for both genes. [3–5] Additionally, some information was found on PsychENCODE Knowledge Portal [34]. The next attempt involved performing BLASTN [23] on SRA data sets with the transcripts of BAG3 and PALLD as references. A more general search was attempted by looking for bio projects in NCBI that involved autism in humans. An alternate approach was then taken by looking at ZNF804A, a gene confirmed to have an impact on autism through mutations to the gene's sequence. The website Gene Cards [35] had links to confirmed variations of this gene on the website GWAS Catalog [36–38]. This site was then used to look up variations of BAG3 [39] and PALLD [40].

To further the comparisons between humans and other model organisms for BAG3 and PALLD, the similarities and differences of their amino acid sequences were also analyzed. The key focus was to identify the changes on a base pair level that had a significant impact on the resulting amino acid sequence. This included determining the severity of the amino acid changes. This is because some amino acids are interchangeable in a structure without disrupting the proteins folding significantly. [41] This comparison was also done using MEGAX [25] and analyzed using Point Accepted Mutation (PAM). PAM is a method of measuring how probable an amino acid mutation is to occur and be conserved by natural selection. Higher scores are less likely to disrupt the function of the protein when an amino acid is replaced. [24]

STRUCTURE COMPARISONS

Based on the results from BLAST [23,24], OAT [29], and MEGAX [25], further analyses focused on humans and the two most similar versions of BAG3 and PALLD. In both cases the most similar versions were the dog's followed by the rat's. The same steps used to make the transcription global alignment with MEGAX [25] were performed again but only on humans, dogs, and, rats. This was done to give better clarity on the differences between just these three versions of each gene and for later reference. Furthermore, the full nucleotide sequence with 2,000 additional nucleotides up and downstream from these three organisms were put in a global alignment with MEGAX. [25] This was done to have the ability to compare at the nucleotide, transcription, and translated levels of the genes.

To give further context to the differences among these genes, the purpose of the proteins in the model organisms were compared. This is because some organisms may have multiple uses for the same protein that are different to the use(s) of the same protein in other organisms of [21,22]This was done by researching the mechanisms of these genes in each model organism. In the primary area(s) of conservation, the differences were recorded. These differences were then compared to the same area in the homology comparisons to see the results of these changes. For homology, two different methods of comparing predicted structures were used since only the human version has been confirmed. The first method used the program UGENE [42] which individually predicted what areas of the amino acid sequence would have a secondary structure. These areas were then aligned with the new global alignment made with MEGAX [25] that only had the human, dog, and rat versions of the amino acid sequences to identify what differences were responsible for changing the predicted structure. The second method involved a program called UCSF Chimera [43]which aligned and modeled predictions of the secondary structure for the proteins using the confirmed human model as a template.

A third method using the program SWISS-MODEL Modelling from University of Basel's Biozentrum, The Center for Molecular Life Sciences,[44] was attempted using the default settings in order to measure the structural differences between the three versions of BAG3 and PALLD proteins using PAM. However, this website only produced blank results and therefore was not considered for the rest the project.

It was found that using UGENE version 33.0[42] and its "predict secondary structure" function with the GORIV setting gave a prediction that allowed for comparison between the protein sequences for both genes. The following steps were used.

> With the protein sequences open and selected in UGENE [42] (name will be bolded) go to "actions" → "analyze"→ "predict secondary structure". Ensure that both the whole sequence is selected along with GORIV then click "predict". Save and create the .gd file. The next step is to right-click the new annotations and go to "Export/Import" →"Export Annotations". Then change the file type to CSV, select "Save sequences under annotations", "Save sequence names", and click "OK". Reselect the original sequence,

right-click the original sequence and go to “Export/Import” → ” import annotations from CSV file”. Set up the following settings in the menu, have the file read as a CSV file, set the results file as anything desired, ensure column selector is a “,”(click “preview” to refresh), set to First lines to skip to “Do not skip”, make sure that “Remove quotes” is selected. Next the “results preview” will need to be set up by clicking “[ignored]” to change the variable types as the following titles of each column as the first row in the CSV file, the annotations “[start position]” as start, “[end position (include should be selected)]” as end, “[length]” as length, “[group]” as sequence name, ”[name]” as sec_struct_type, and all other categories should be “[ignored]”. Once “run” is clicked use the “export image” camera in the top right of the screen to export the image of the results using “assumed annotation” and “sequence details”.

Chimera version 1.15[43] was used to take the confirmed human model of each gene to predict the structure of the dog and rat versions. The same human sequence was also performed upon in order to get a control and determine the accuracy of the program. The instructions were used as follows.

With the template structure open, go to tools, sequence, and sequence. In the new menu click edit and “add sequence” followed by entering the sequence as plain text with the default settings for all three sequences. Once all the sequences are in the program, go to structure and click modeler (homology). In the new menu the following steps will need to be repeated for all three sequences. First choose each target sequence in each run, then choose the template as the confirmed model, finally Run Moeller via web service then click OK. With these steps completed for all sequences, re-add all new variations via the sequence tab. Next identify the best matches to remove the unwanted versions. This is done by clicking the edit tab followed by the delete sequences/caps while the unwanted sequences are highlighted and click OK. Finally, save all of the results under the file tab. To save just the optimal results for later use first “save the session”. Next go to “save as a PDB” with the following settings, select all files of interest including the original, add the “$name $number” to the end of the file name you give, set relative to the template model, and save multiple models in multiple files, and click save. Also, repeat these same steps except under the Mol2 format instead of PDB. This session can now be closed and reopened as a new session with only opening the saved PDB files.

ANALYSES OF REGULATORY REGIONS

When looking for the regulatory regions of BAG3 and PALLD multiple methods and tools were used in order to get the best idea due to there being no confirmed regulatory regions for either gene. When looking at the NCBI Nucleotide page for each gene and clicking on “Highlight Sequence” the “Feature” dropdown menu was used to find the Introns, Exons, regulatory class, and the CDS [45–52] of the nucleotide sequence for these genes if available. More information was gathered by clicking on “Show in Genome Data Viewer” to view the gene relative to other

features on the chromosome.[53–58] In order to download the sequences, the following steps were used.

> Hover the mouse over the gene name, once the details menu opens automatically copy the location number range. Next right-click on any sequence and click "Select a Range". This is followed by hovering over the newly selected green area that overlaps with the numbers listed at the top of the menu to bring up a menu automatically where "Modify Range" is selected. The location range is then input into the menu to select the full gene. Going back to the "Modify Range" menu again allows for the range to be modified for other sequences. By once again hovering over the top of the selected area clicking "Download FASTA (Selection)" will download the selected sequence as a FASTA file for later analysis.

For this project two different upstream sequences were looked at. Both went upstream of the start site of the respective gene and overlapped with less than the first 100 nucleotides of the start of the gene to ensure that the full upstream sequence was captured. The first type was the beginning 1000 nucleotides upstream as a standard range of observation. It was also used for looking into the TATA box and any immediate enhancers/promoters. The second type consisted of the 10,000 nucleotides upstream of the start of the gene. However, only results within the first 2,000 nucleotides upstream of the gene were considered, unless the first result was found beyond the 2,000 nucleotide range.

It should be noted that initially, instead of 10,000 nucleotides being observed at once, they were observed in 1000 nucleotides sections at a time. The first 4000 nucleotides were downloaded this way for human BAG3 and PALLD. For BAG3 these four sequence regions were tested on "Berkeley Drosophila Genome Project Neural Network Promoter Prediction" [59] and "DTU Health Tech Promoter - 2.0 Transcription start sites in vertebrate DNA" [60] for promoter sites along with Softberry FPROM [61] for finding the TATA box. Softberry did not test the following sections once the first TATA box was found for that gene. However, due to the results for each website being inconsistent with each other and only 0 to 2 results being found per 1000 nucleotides; the upstream 10,000 nucleotide sequences were then used instead. These each gave the same results as the 1000 nucleotide sections with a few additions if a promoter is predicted to be across two of these 1000 nucleotide sections. The remaining upstream sequences for BAG3 and PALLD from rats and dogs were downloaded from NCBI.

The 1000 nucleotide upstream sequences were first aligned using the same methods as the transcribed sequences with MEGAX.[25] This was done in order to both find areas of conservation between the upstream sequences and to manually add the TATA boxes and enhancers/promoters predicted. Sequences were mainly added by pressing Ctrl + N to create a blank sequence. Then with the enhancers or TATA box sequence copied from the results of other programs, Ctrl + F was used to find the sequence on the alignment. Next it was clicked on where the enhancer would go on the new sequence to match up with its source and then Ctrl + V to paste it into the new line. Some spaces were needed to be added in order to account for gaps in the alignment. This was done by adding spaces like a text document.

For predicting promoters, two web-based programs "Berkeley Drosophila Genome Project Neural Network Promoter Prediction" [59] and "DTU Health Tech Promoter - 2.0 Transcription start sites in vertebrate DNA" [60] were used with their default settings. Both 1000 nucleotide and 10,000 nucleotide upstream sequences were analyzed using these web-based programs. These results were then downloaded for analysis. Analysis was performed by looking at promoters found within the first 1000 nucleotides and adding them to their respective alignment in order to identify their similarity both in sequence and in location. Promoters found within 2,000 nucleotides were also looked at but were not added to the sequence alignment.

For predicting the TATA box, two different web-based programs provided by the website Softberry [61,62] were used. The first was FPROM [61] which takes one input sequence and tries to predict the location and sequence of any TATA box. The other web-based program PromH(G) [62] takes two input sequences and tries to use both in order to predict the TATA boxes for both sequences. Since TATA boxes are expected to be close to the start of the gene, initially only the 1000 nucleotide sequences were used. However due to the poor results from PromH(G) [62], 10,000 nucleotide sequences were also looked at for any TATA boxes, with further upstream results being less credible. It should also be noted that the website has a limited number of runs that can be performed daily, making these results take multiple days to calculate.

Results:

INITIAL RESEARCH

The first question that needed to be addressed was identifying the closest surface relative to *Astyanax mexicanus*. It was found to be *Astyanax aeneus*.[63] However, only whole genome sequences are available with there being little to no research on individual genes in *A. aeneus*.[64] It was determined that it was better to use the organism that was most closely related and had sequences with confirmed orthologs of BAG3 and PALLD to be used instead. This organism was *Pygocentrus nattereri* the red-bellied piranha and it is in the same order as *Astyanax Characiformes*.

Some research has also been done to identify the functions of BAG3 and PALLD. BAG3 in humans is involved in the chain reaction of heat shock regulation. It also plays roles in cytoskeleton dynamics, protein quality control, and a structural role in muscle cells. [22] Particular focus was given to proteins that affect BAG3 and its regulation along with the proteins it directly affects. The same was done for PALLD and its role is as an actin cross-linker for helping cells handle mechanical tension. [21]

BLAST

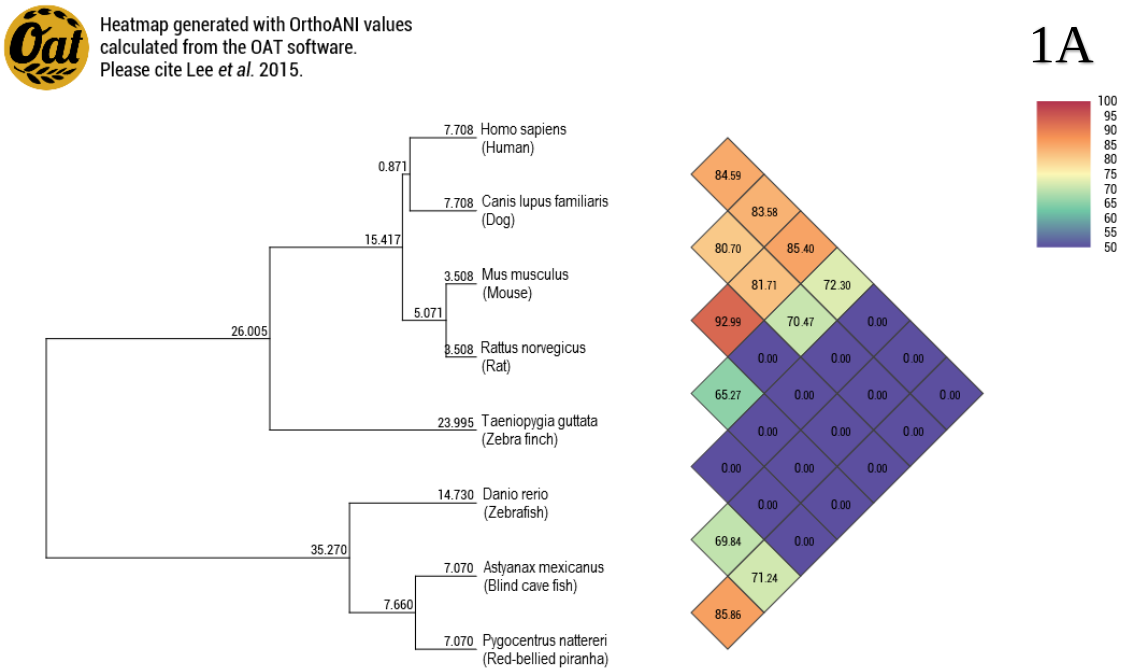
BLASTN [23] and BLASTP [24] were used to answer the question of what species have the most sequence similarity between their BAG3 and PALLD to each other. When looking at the overall

results from BLASTN [23] and BLASTP [24], an initial observation was that while the maximum number of results allowed was set to 5000 no search identified more than 1100 results. This could in part be due to there being a limited number of sequences to analyze between the eight different organisms being searched along with there being a cut off for insufficient results as part of the default for BLASTN [23] and BLASTP [24]. When looking at the overall results, there are some immediate patterns that apply for both genes. The first being that the fish (blind cave fish, zebrafish, red-bellied piranha) are only found together while the other organisms are themselves found together. The next being that of the other group, the zebra finch had the least similarity to the other members. This indicates that there is a separation between the mammals, the bird, and fish. This makes sense considering the evolutionary relationship between these three groups. Furthermore, the organism most similar to humans in both cases appears to be the dog.

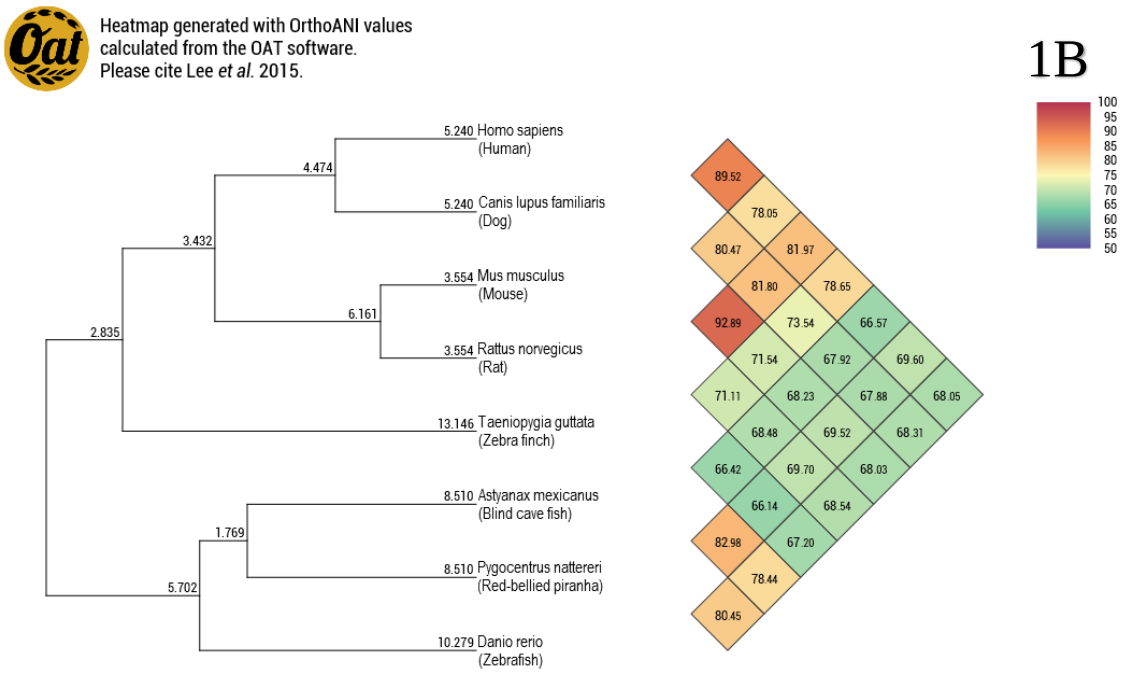
AUTISTIC HUMAN GENES

Another question that needed to be answered before experimentation could start was, are there an available sequences of BAG3 and PALLD from any human diagnosed with autism? Despite the various attempts none of the methods used to get an autistic version of the human BAG3 or PALLD genes were successful. One of the first methods used was to look at the PID found from the same three studies that were cross-referenced for human gene expression in autistic individuals.[3–5] However, the results would either lead to these PID not having the desired information or they were not publicly accessible. The next attempt was to look at the Trace [65] for these genes and try to search for them using the data from NCBI's Sequence Read Archive. This also proved to be unsuccessful. They both gave identical results of only giving 100% fragments of the full sequence, all being only 69 Base pairs long. It was concluded that this tool would not work for finding the autistic version of BAG3 and PALLD. When reviewing the BLASTN [23] and BLASTP [24] results for both genes from humans there were none that related to autism. Searching for Bio projects did not find the desired information either. [66] The last attempts that were able to give partial results were Gene Cards [35] and GWAS Catalog [36–38]. From Gene Cards [35] there were 3 links to the following point mutations on ZNF804A stored at GWAS Catalog [36–38] that were related to autism. The following chromosome position had a point mutation, 184947213 A to T, or C [36]; 184736693 C to A, T, or G [37]; 184668853 T to A, or C [38]. While there was data available for many different recorded mutations for BAG3 [39] and PALLD [40], none were related to autism. This makes sense considering that non-silent mutations are more likely to disrupt the function of a gene instead of increasing its expression level. This is important considering that these genes were originally identified through their up regulation observed in autistic individuals. It should also be noted that the disorders that are most often associated with the mutant version of these genes are as follows - BAG3: electrocardiogram morphology, left ventricular ejection fraction, hypertrophic cardiomyopathy, and dilated cardiomyopathy [39]; PALLD: coronary artery disease, pulse pressure, and systolic blood pressure [40]. It was concluded from this that while there is evidence towards a change in regulation for BAG3 and PALLD, there is no evidence for mutations in the genes contributing to autism.

ALIGNMENTS



1A



1B

Figure 1 - Gene dendrogram and heat map based on pairwise alignments and OrthoANI values [30]: This dendrogram shows the evolutionary history of the BAG3 (1A) and PALLD (1B) along with the pairwise alignment showing sequence similarity between each species' version of the genes represented on a heat map. Both of these were generated using the OAT software [29]. By following the species name to the heat map the immediate upwards and downwards columns represent a modified similarity value called OrthoANI [30]. It represents the probability that the two sequences are from the same species with 95% being the cutoff. The species labels on the dendrogram were replaced for readability.

An example in the PALLD heat map is the most similar sequences being between mice and rats at an OrthoANI value of 92.89%. Another example is the least common non 0% connection to humans which is blind cave fish at an OrthoANI value of 66.57%. The numbers on the branches of the dendrogram are the branch lengths representing the evolutionary difference between the two groups and their most common ancestor. It is clear that the mammals, birds, and fish are all distinct groups. It is interesting that for PALLD the blind cave fish was less similar to humans than to the control red-bellied piranha. Furthermore, the two most similar organisms are the mouse and the rat for both genes. Lastly it is noted that the two most similar organisms to humans were dogs followed by rats.

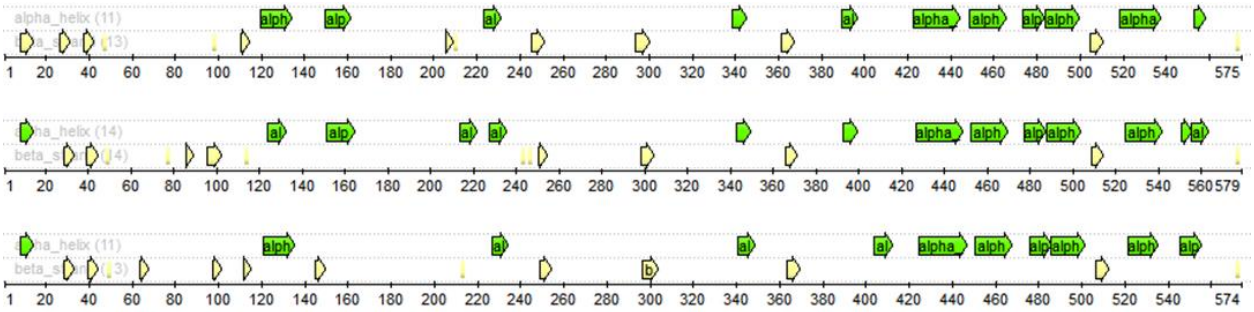
One interesting observation regarding Figure 1A is that despite the BAG3 mouse and rat sequences having an OrthoANI value of 92.99% and zebra finches and rats having an OrthoANI value of 65.27% the program gave the mouse and zebra finch an OrthoANI value of 0%. A possible explanation for this is that in the global alignment, it was shown that the mouse sequence was longer than both the rat and the zebra finch who both were relatively closer in length. This may have made it more difficult to create a proper pairwise alignment with the zebra finch and the mouse than the mouse and the rat which while despite this length difference have a lot of similar nucleotides that are lacking in the zebra finch.

With the sequence alignments done, a global alignment was performed in order to identify what areas of the genes are conserved and most variable. In BAG3 while there are some areas of general conservation on a nucleotide level there was no one section that was conserved across all organisms. However, when looking at the protein global alignment there were two mostly conserved areas. They were 32 amino acids at position 27 and 76 amino acids long at 471 on the global alignment of the BAG3 amino acid sequences. For comparison, the sequence length of the gene ranged from 459 to 579 amino acids and the global alignment was 629 amino acids long. For PALLD there were other smaller areas of conservation across the nucleotide sequences that formed into one primary conserved area of 462 amino acids long starting at 1017 on the PALLD global protein alignment which was 1538 amino acids long. However, the size of PALLD had a much larger variable length between organisms ranging from being 680 to 1453 amino acids with only minor variation in this conserved section. After creating a global protein alignment using MEGAX [25] with the muscle algorithm on the protein sequences of BAG3 and PALLD it was found that the protein sequences were more conserved than the mRNA sequence. Furthermore,

BAG3 was still not very well conserved with the exception of two areas. Meanwhile PALLD had a large area of conservation for a majority of the protein sequence.

SECONDARY STRUCTURE PREDICTION

2A



2B

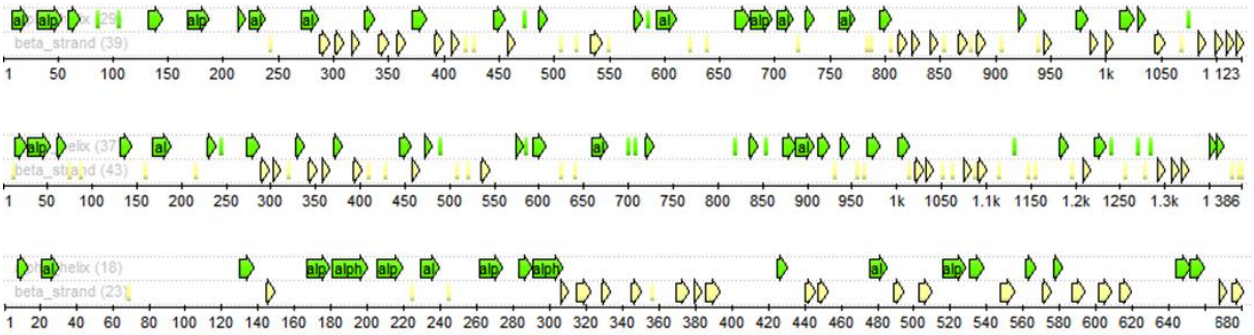
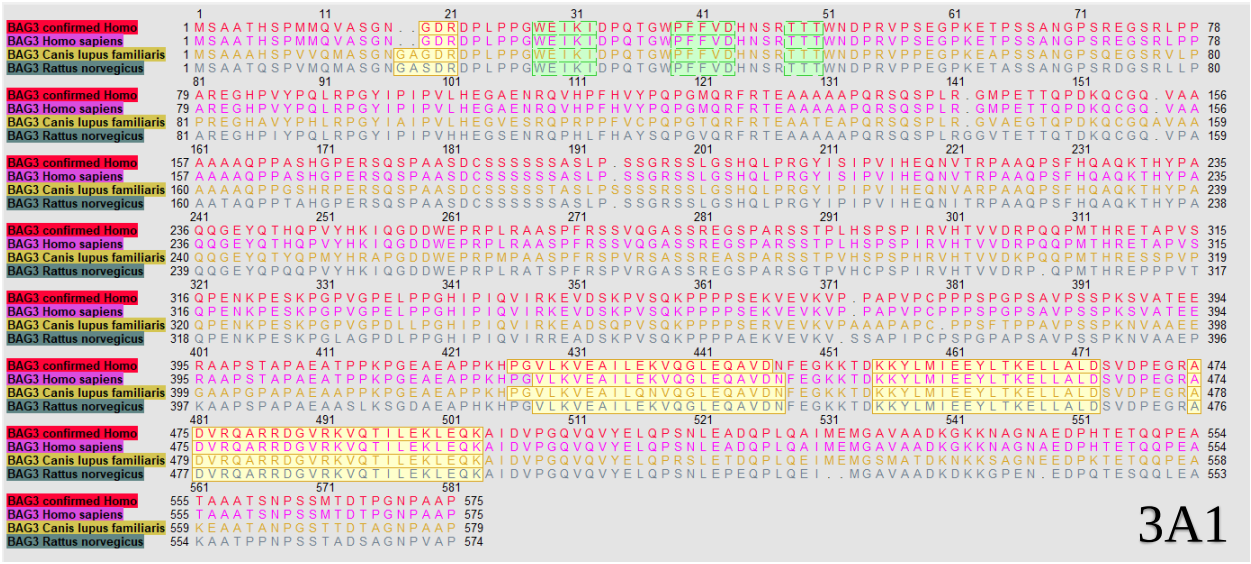


Figure 2 – UGENE [42] secondary structure prediction for BAG3 and PALLD:
This figure shows the predicted secondary structure for BAG3 (2A) and PALLD (2B) made using UGENE [42]. For both 2A and 2B each line represents the human, and dog, and rat sequence in that order. The green arrows represent areas predicted to be alpha helices and the yellow arrows represent the predicted beta sheets. It should be noted that while the BAG3 sequences were all relatively the same length, the sequences for PALLD were more variable and therefore are not as well aligned with each other.


BAG3 confirmed Homo
BAG3 Homo sapiens
BAG3 Canis lupus familiaris
BAG3 Rattus norvegicus
3A1


PALLD Confirmed Homo
PALLD Homo sapiens
PALLD Canis lupus familiaris
PALLD Rattus norvegicus
3B1

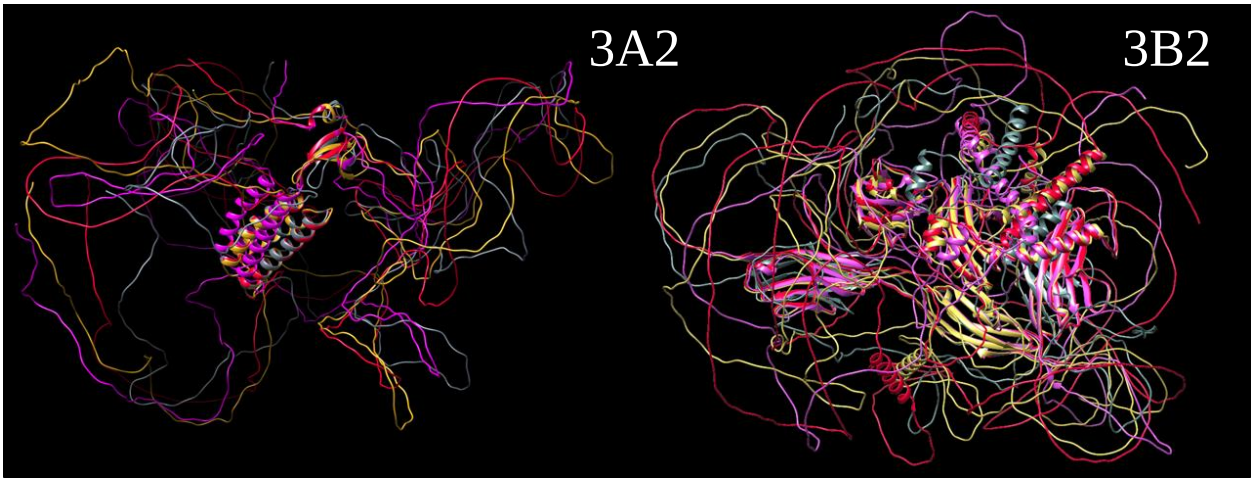

3A2
3B2

Figure 3 - Protein homology models of BAG3 and PALLD made using UCSF Chimera [43]: Chimera [43] aligned and modeled predictions of the secondary structure for the BAG3 (3A1 and 3A2) and PALLD (3B1 and 3B2) proteins using the confirmed human model as a template. 3A1 and 3B1 show the aligned sequences with the sections highlighted in yellow representing the alpha helixes while the green highlights represent the beta sheets. In all images the color scheme used and the order on the alignment was as follows; confirmed human model in red, human amino acid sequence in purple, dog amino acid sequence in gold, and the rat amino acid sequence in bluish gray.

With the model organisms for this study being narrowed down to dogs and rats, the next part was to identify if the predicted 3D structures of their genes were similar enough to humans that they could be used in future studies. When looking at the UGENE [42] results as shown in part in Figure 2 it is observed that several small secondary structures were predicted across the entirety of both genes. Furthermore, despite having multiple sections with the same secondary structure back to back there are a few individually predicted secondary structures of significant length. While this method for the most part was able to correctly predict what type of structure would be present in the same location as the confirmed human model (purple in Figure 3), the large amount of small secondary structures makes the results more difficult to read. Meanwhile the results from Chimera [43] (Figure 3) matched up much better to the confirmed human model with fewer but larger and more accurate secondary structure predictions in all sequences. From looking at the predicted 3D model in Figure 3A2 and 3B2 it is clear that BAG3 is very heavily conserved with minimal changes in the location of the secondary structures and their sizes. This is not true for PALLD (Figure 3B2) where the predicted structures remained consistent, but the variation in the length of the sequences makes it harder to confirm the full sequence as being conserved. Furthermore, the predicted 3D structure (Figure 3B1) has some concerning differences between the sequences. The first being that some secondary structures on the confirmed model were not predicted in the other sequences. While the structures present are in the same general location, not all of them share the same orientation. This could be in part because of the missing secondary structures, the difference in sequence lengths, or some other details in the process of Chimera [43] making these models.

REGULATORY REGIONS

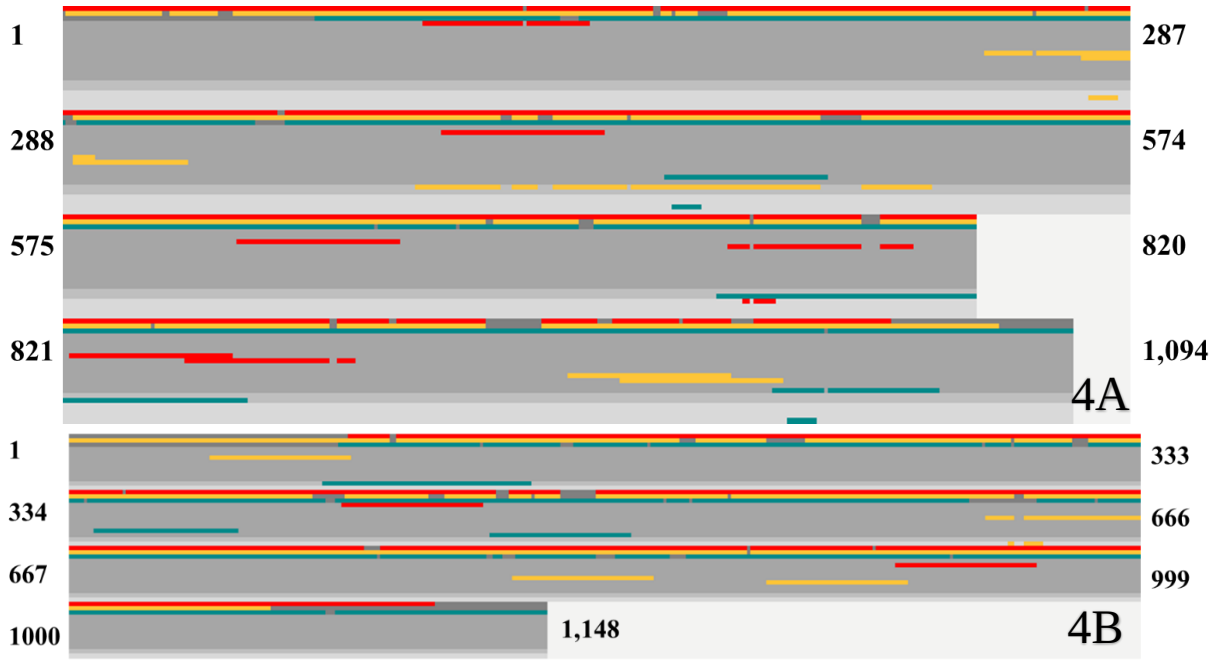


Figure 4 - Graphic comparison between upstream regulatory regions of BAG3 and PALLD: This figure looks at the predicted regulatory regions and promoters from Berkeley [59], DTU [60], and the TATA box from FPROM [61] in comparison to their location upstream of approximately the first 1000 nucleotides of the BAG3 (4A) and PALLD (4B) from humans, dogs and rats. Red sequences are from humans, yellow from rats, and teal from dogs. The numbers on the left and right represent the sequence alignment position with the start of the line at the left and the end of the segment on the right. The first 3 lines in the darkest gray area represent the full upstream sequences. The second darkest gray with the most individual lines represents the promoter results from Berkeley [59]. The second to lightest gray area shows the results from DTU [60], and the lightest section represents the TATA box results from FPROM [61].

The last question that needed to be addressed was, what predicted regulatory regions were consistently found upstream of the genes across different regulation prediction programs? When aligning the first 1000 nucleotides upstream of the two genes for each of the three organisms, the following results from MEGAX [25] were found. For BAG3 while there were multiple conserved nucleotides there was no consistent clusters or locations of conservation. Meanwhile, PALLD was heavily conserved throughout the majority of the upstream sequence. This is consistent with the alignments of the genes themselves where PALLD was more conserved than BAG3 on a nucleotide level. When comparing these results to Figure 4 there are two complications. The first being that while there are some overlapping promoters in BAG3 (Figure 4A), this area is shown to not be conserved from the alignment performed in MEGAX [25]. This means that despite the promoters being in the same location, they are significantly different. The opposite problem is true with PALLD as seen in Figure 4B where the alignment is showing a conserved upstream sequence with minimal promoter overlap with the exception of one rat promoter being in the

same place as rat TATA box. In both genes there is not strong evidence for the promoters upstream of the genes being conserved across these organisms.

Berkeley:

This web-based program seems to give multiple predicted promoters that were compared to other promoters on the same gene. Furthermore, the promoter sequences seem to always be provided in 46 nucleotides sections. However, there were multiple cases where one sequence was found in two different promoters. This indicates that the overlapping area is more likely to be a predictor than the other findings. As a result, the results from this program are considered adequate for consideration. From these results it was found that in BAG3 there were 6 human promoters, 6 dog promoters, and 4 rat promoters within the first 2,000 upstream nucleotides. (Figure 4A) What is interesting is that the two promoters closest to the gene in humans overlapped in the area "5'-GCCCGCGCCCGCC-3'". This is significant because it is similar to the two closest promoters in dogs which overlap with "5'-GGCCCCGCCCCGGCCCGCCCCGCCCCGCCC-3'".

All four of these promoters are also relatively close to each other on the sequence alignment. While this similarity is not found in rats there is a similar promoter that is found in a rat promoter as well as two different human promoters. These promoters start with 4 or 5 "A" followed by a similar pattern of nucleotides for the rest of the predicted promoter. It should be noted that all of these promoters are relatively further away from the start of the gene but were close to each other. Furthermore, the dog promoter, the second furthest away from BAG3, also matched this pattern but was further upstream by about 500 nucleotides. Looking at the results for PALLD there were 2 human promoters, 5 dog promoters, and 2 rat promoters within the first 2,000 upstream nucleotides. (Figure 4B) The only discernible pattern between the different promoters is that there were multiple cases of each nucleotide repeating 3 to 5 times in a row, although none of them were similar. Additionally, while the promoters are all close to each other on the alignment the only overlap occurred between the closest promoters for humans and dogs by four nucleotides with "TTTC". One interesting promoter was found in rats that is not found in any other sequence for both genes where the further away of the two rat PALLD consisted primarily of repeating "TA" which made up 30 of the 46 nucleotides in the predicted promoter. Later experimentation did not find any TATA boxes in this area.

DTU:

When looking at the DTU [60] results there were two things immediately obvious compared to promoter predictions. The first being that there were fewer promoters found overall. Furthermore, the promoters that were found were different from Berkeley [59]. This could indicate that the programs use different methods of applying neural networks to make their predictions, alternatively this could imply that there are no clear promoters based on sequence prediction alone. One complication with this web-based program is that the outputs give a single position that is divisible by 100 with the sequence being presented in 60 nucleotide rows. It is also not clear if this is the start or end position. Although, from aligning 60 nucleotides before, after, and

together from that position it was determined that it was considered a starting point. Therefore, when aligning these results the position number followed by the rest of the 60 nucleotides in the row and in the next row were aligned. It was concluded that based on the low number of results and the lack of similarity between the results on top of the lack of clarity of the promoters' start and end positions that these results should not be considered.

Within the first 2,000 nucleotides for BAG3 there were two promoters from humans, one promoter from dogs, and one promoter from rats. (Figure 4A) When aligning these regions, it was determined that the two human promoters were very similar in their sequence going from the position number predicted forward. Meanwhile the dog and rat promoters, while more similar to each other than to human, were still unique in comparison with each other. Furthermore, all these predicted promoters were hundreds of nucleotides away from each other. The result for PALLD gave one promoter for dogs and rats within 2,000 nucleotides upstream. (Figure 4B) The closest human promoter predicted was 2300 nucleotides upstream of the start of PALLD. While the dog and rats were about 300 nucleotides apart, the human promoter was over one thousand nucleotides away. Once again there was no discernible pattern between these three promoters.

FPROM:

From looking at the overall view of the results from FPROM [61] it appears that the program is able to effectively find the more common/typical types of TATA boxes found in a sequence. However, since the program had some difficulty with the PALLD sequences it could imply that there may either be limits to what the program can predict or that the PALLD gene is regulated in a more atypical manner. The results for BAG3 gave one TATA box for humans and dogs while two TATA boxes were found for the rat sequence. (Figure 4A) When aligning these TATA boxes to the aligned 1000 upstream nucleotides, none were close to each other and all were different from each other. The two closest were the dog TATA box "TTTTTATA" and the further upstream rat TATA box "TATTAAAA". When PALLD was put into FPROM [61], no TATA boxes for humans or rats were found within the first 1000 nucleotides. It was able to find one TATA box for dogs within that range. (Figure 4B) When looking at the 10,000 nucleotide upstream sequence, two human TATA boxes were found at 3,601 nucleotides upstream and 5,475 nucleotides upstream. However, considering their distance this is likely not the actual TATA box. For rats no TATA box was found in the 10,000 nucleotide sequence. All tests that gave zero promoter results were retested with the same outcome. When the one TATA box in range from dogs was added to the 1000 nucleotide alignment, there was a gap in the rat sequence where most of the TATA box would be and the human sequence was "CAGAGTAAGAA" in comparison to the dog TATA box "TA - - - TAAAAA". This indicates that this aligned section of the human sequence does not contain a TATA box.

1A1

BAG3-1000 Nucleotides Upstream						
Organism	Homo sapiens		Canis lupus familiaris		Rattus norvegicus	
Ortholog	Nucleotides upstream	Weight	Nucleotides upstream	Weight	Nucleotides upstream	Weight
H.s.			656	165.47	633	178.47
C.l.f	No results				633	178.47
R.n.	No results		No results			

1A2

BAG3-10,000 Nucleotides Upstream						
Organism	Homo sapiens		Canis lupus familiaris		Rattus norvegicus	
Ortholog	Nucleotides upstream	Weight	Nucleotides upstream	Weight	Nucleotides upstream	Weight
H.s.			1898	209.2	633	178.47
					75	176.93
					7688	205.4
					1407	193.73
C.l.f	7019	184.87			633	178.47
					7688	205.4
					1407	193.73
R.n.	No results		656	209.2		

1B1

PALLD-1000 Nucleotides Upstream						
Organism	Homo sapiens		Canis lupus familiaris		Rattus norvegicus	
Ortholog	Nucleotides upstream	Weight	Nucleotides upstream	Weight	Nucleotides upstream	Weight
H.s.			331	202.87	No results	
C.l.f	No results				No results	
R.n.	No results		331	202.87		

1B2

PALLD-10,000 Nucleotides Upstream						
Organism	Homo sapiens		Canis lupus familiaris		Rattus norvegicus	
Ortholog	Nucleotides upstream	Weight	Nucleotides upstream	Weight	Nucleotides upstream	Weight
H.s.			2714	202.73	6017	171.27
			9185	213.87	8672	171.33
C.l.f	3602	181.8			6017	171.27
	6388	194.53			8672	171.33
	8739	196.2				
R.n.	3602	181.8	331	202.87		

Table 1 - PromH(G) [62] results organized:
This figure shows the TATA box search results from the Softberry program PromH(G) [62]. This program took an input sequence (Organism) followed by a comparison sequence (Ortholog) in order to identify possible TATA boxes in both sequences. Each column represents the input sequence while the row represents the comparison sequence. The tables 1A1 and 1A2 show the

results for BAG3 while the tables 1B1 and 1B2 show the results for PALLD. The tables 1A1 and 1B1 looked at only the first 1000 upstream nucleotides while tables 1A2 and 1B2 looked at the first 10,000 nucleotides. Lastly each result is represented by the first number being how far the TATA box is from the gene, with zero being at the start of the gene and 1000/10,000 being further away. The next number represents the "Weight" of the results or how significant it is. The three organisms compared were humans (*Homo sapiens* (H.s.)), dogs (*Canis lupus familiaris* (C.l.f)), and rats (*Rattus norvegicus* (R.n.)).

PromH(G):

One observation that was immediate is that unlike the other programs used, the results using the 10,000 nucleotide sequence did not always contain the results from the 1000 nucleotide sequence. Furthermore, when rerunning the results that gave zero TATA box predictions, they were consistent between the first and second analysis of the same input. The next observation was that this program had difficulty finding TATA boxes that were relatively close to the gene. This is especially true for humans on both genes in that no TATA box was found within 1000 nucleotides for either gene. (Table 1) In conclusion, the lack of consistency and low number of nearby TATA boxes predicted by this program indicates that it likely should not be used for the rest of the current project.

The only BAG3 result for human was when it was compared to the dog sequence and found one TATA box 7,019 nucleotides upstream of the gene. When looking at the dog results, there is no consistency between 1000 nucleotides and 10,000 nucleotides for BAG3. Compared to both human and rat the 10,000 nucleotide sequences indicated a TATA box 1,898 away, while with the 1000 nucleotide sequence being compared to rats gave no results and compared to humans gave a TATA box 656 nucleotides away. Rat results were both the most abundant and most consistent giving mostly the same results between the human and dog comparison, with their only being one missing TATA box found in the 10,000 nucleotide dog comparison. However, the one TATA box found in all four comparisons focusing on the rat indicated that there was a TATA box 633 nucleotides away from the BAG3 gene. For PALLD the 10,000 nucleotides sequences gave better results compared to the 10,000 nucleotide sequence where only one TATA box was found for dog and none for humans and rats. In the 10,000 nucleotide sequence results there were three unique human TATA boxes, three unique dog TATA boxes, and two unique results in rats. When looking at TATA boxes that appeared multiple times, human TATA boxes compared to dog and rat both agreed that there was a TATA box 3,602 nucleotides upstream. For dog PALLD, it gave the second most consistent results between the 1000 nucleotides and the 10,000 nucleotides when compared to rats. They both identified a TATA box 331 nucleotides upstream. Lastly, analysis of the rat PALLD found two TATA box results with the closest TATA box being was 6,017 nucleotides upstream and the further away one being 8,672 nucleotides upstream. (Table 1)

Discussion:

At the start of this research project, versions of BAG3 and PALLD from an autistic individual were looked for in order to identify what is different from the wild type version. However, due to the lack of autistic variants of BAG3 and PALLD being found and other research it was determined that there might only need to be a single nucleotide change for some genes to cause autism and in other cases a change in the regulation of their expression. This was shown by the accessibility of the autistic variants of the control gene ZNF804A, which has been confirmed to contribute to autism [36–38]. It was found that if any of 3 different locations on ZNF804A have a missense mutation, the mutation will increase the risk of autism. [36–38] Furthermore, four different studies that looked at the expression of all genes in autistic individuals were reviewed. [3–6] All four have found that BAG3 was upwardly expressed while three of the studies found the same for PALLD. [3–5] Another paper found no significant change for PALLD. [6] Considering these findings and the fact that parts of autistic traits are caused by variations in regulation instead of mutant variants of genes, along with an upward expression not being likely explained by a mutation on the genes themselves, it was concluded that there is likely not a difference between the wild type and autistic version of BAG3 and PALLD.

Both BAG3 and PALLD have been found to be up-regulated and therefore it was concluded that it is a safe assumption that, without having access to an autistic genome or methods of verifying the sequence of these genes in autistic individuals, they are identical to wild type human genes. However, there is a difference in the regulation of these genes and more focus was given to the regulation differences between the organisms with less on the sequence differences. As a result, from this it was determined that the project would shift away from primarily looking at the differences between the sequences of BAG3 and PALLD across different model organisms and instead look more at the expression differences between them. It is also likely that epigenetic modifications of BAG3 and PALLD may play a role in regulating their expression in autistic individuals. Without supporting data or the tools to analyze sequences for epigenetic modifications, this project could not explore this notion further. However, it represents a focus for future studies when suitable data for analyses is available. [7–11]

It was found that UCSF Chimera [43] was able to more accurately predict the secondary structure than UGENE [42] which had multiple 1 to 3 nucleotide long secondary structures that were not found in the confirmed model. Furthermore, the BAG3 models all aligned well to each other. Meanwhile, the PALLD proteins were not as cleanly matched but still had some similarities to each other. Although in this case there were predictions for one area of the sequence where there was an alpha helix that is not present in the confirmed model found on all 3 amino acid sequences. There is also an alpha helix on the confirmed model that was only predicted in dogs despite that area of the sequence being the same in all sequences studied. Therefore, the organisms chosen for both BAG3 and PALLD were still worthy of continued investigation.

Based on the sequence alignments and gene tree created, it was determined which organisms should be included and which organisms should be excluded going forward with this project.

When performing the pairwise alignments and creating the gene dendrogram/heatmap, the BAG3 gene in the zebrafish, piranha, and blind cave fish all gave OrthoANI values of 0% by OAT [29] compared to the other model organisms. (Figure 1) Furthermore, these three organisms also had the least similarity to humans in their PALLD genes. Therefore, zebrafish, red-bellied piranha, and blind cave fish were not included going forward. Additionally, mice and rats were the two most similar models with having OrthoANI values of 92.99% for BAG3 and 92.89% for PALLD relative to each other. (Figure 1) Thus, having both would be redundant. The rat version of these genes is more similar to humans with BAG3 at 85.395% and PALLD at 81.97%. (Figure 1) With mice it was BAG3 at 83.58% and PALLD at 78.045%, meaning that they should no longer be included. (Figure 1) When looking at which organism had the most similar version to humans, in both cases it was the dog with OrthoANI values of 84.59% for BAG3 and 89.52% for PALLD. (Figure 1) Therefore, the dog version of these genes was included. This left the zebra finch, whose similarity to humans in both cases was in the middle with 72.3% for BAG3 and 78.65% for PALLD. (Figure 1) It was determined that it did not have enough similarity to look further into so more focus was given to the remaining organisms. These results match and agree with the BLASTN [23] and BLASTP [24] results in terms of the ordering of most similar to least similar genes compared to humans. It was concluded that in order to give better focus and obtain the best results going forward only the dog and rat sequences would be compared to humans with regard to their protein structure and regulatory regions. It was determined that if this was not done there would be too much data to properly analyze. However, the other organisms were no longer included primarily due to them having significant genetic difference from humans as seen in the results from BLASTN [23], BLASTP [24] and OAT [29]. The one exception was the mouse which was the third most similar to humans. It was determined that since the mouse and rat sequences were so similar only one of them needed to be analyzed. In this case both BAG3 and PALLD from rats were more similar to humans than were those genes in mice. (Figure 1)

When observing the global alignments using MEGAX [25], it was clear that PALLD was much more conserved across species in comparison to BAG3. As expected, the protein alignments had more defined areas of conservation in comparison to the nucleotide sequences. This could imply that PALLD is important for the survival of the organism. Changes to this gene may have a significant impact on the ability of the individual to survive in the wild. This is, however, complicated by the results from the predicted 3D protein structure.

When comparing the conserved regions found on the global alignments of all model organisms made in MEGAX [25] to the predicted 3D structure for humans, dogs, and rats created using Chimera [43] (Figure 3), there were some clear connections and overlaps. BAG3 had two areas on the protein sequence that were conserved. The first of these two sections was 32 amino acids long. When looking at this position on the human, dog, and rat structures it contained all three of the known and conserved beta sheets. The other conserved area was 76 amino acids long and consisted of all three of the larger alpha helices with the exception of the last two amino acids on the third. While the last small alpha helix in humans, dogs, and rats was not conserved in all model organisms, it was still conserved in humans, dogs, and rats. This indicates that despite BAG3 initially seeming to have low conservation on the protein level, its 3D structure is very

well conserved especially in the two most similar model organisms. Due to the size difference between the different PALLD proteins it is impossible for the full human structure to be conserved. However, the conserved area of 462 nucleotides was shown to cover a conserved structure between residues 867-1329 in the global alignment which was the last amino acid on the predicted 3D structure.

This shows where these genes were conserved in humans, dogs and rats that were also conserved in other organisms supporting the argument that those areas are essential to the function of the two genes. The areas outside of this conserved area may have either naturally changed due to random mutation or they have changed because the gene had a slightly altered function in that organism. The exact impact and significance of these changes could be a focus of future studies on the significance of evolutionary differences between BAG3 and PALLD. Another area that could be investigated is getting a proven protein model for BAG3 and PALLD from dogs and rats in order to confirm these results to the proven human versions of these proteins. Lastly, epigenetic modifications are prime for future research but the challenges of identifying them and discerning their effect on gene expression is beyond the scope of the current project.

From the comparisons between Chimera [43] and UGENE [42] outputs there are a few possible factors that explain their significant differences in structure prediction results. The first major one is that while UGENE [42] was working exclusively with the sequence, Chimera [43] had a confirmed model to act as a reference. Without that reference model the UGENE [42] method needed to make predictions based on the properties of the amino acids both in isolation and in comparison to the nearby amino acids. This would, however, mean that if there was no confirmed human model, UGENE [42] would be an effective tool to get an initial prediction for an unknown protein. However, Chimera [43] is better suited for comparing similar sequences that have a confirmed model as a reference point. As a result, while both tools are valid, the availability of the confirmed human models for both BAG3 and PALLD made Chimera [43] the superior option for this project.

When initially performing analyses on the upstream sequences of both genes in 1000 nucleotide intervals, two flaws in that approach were identified. The first was that regulatory regions found between the 1000 nucleotide intervals were lost. The other problem was that the programs took less time to run 10,000 nucleotides than to download and individually run 10 different 1000 nucleotide sequences. It was concluded that for the rest of the work only the first 1000 nucleotides would be checked for an alignment and identification of the TATA box, while the 10,000 nucleotide method would be used for finding promoters within 2,000 nucleotides or the closest find but prioritizing the results that are closer to the gene and identifying any similar matches across organisms. If a result was found significantly past 2,000 nucleotides it was considered less likely to be accurate or meaningful with regard to regulating gene expression. This is reasonable given that only about 10 results were given for each human gene and the programs were able to analyze the large sequences in less than a minute.

After using the four different programs to look at upstream sequences there was a problem that was unlike other tools that looked at the same data, there is no consistency across programs. The one exception was that there were some TATA boxes identified from the FPROM [61] results that matched with the predicted promoters from the Berkeley [59] analysis of the same organism. This lack of consistency implies that either these tools are not very effective or that it is difficult to predict features in the upstream sequence of these genes. It was concluded that this is the area that requires the most future research to fill in missing data.

When comparing how well BAG3 and PALLD were conserved against each other there was a better case for BAG3. While PALLD had a stronger case on a nucleotide level, BAG3 provided a better case on the protein structural level. Since the resulting protein has a stronger effect on its effect on autism than the nucleotide sequence, BAG3 serves as a better case for further study than PALLD. Furthermore, since this project found the regulation of both genes to be inconclusive they cannot necessarily be compared against each other and instead should be where research is focused in the future. When comparing which one model organism was the best for studying autism on a genetic level, dogs consistently were more similar to humans in all of the analyses performed. However, rats still had enough similarities to be considered a potential model organism for these genes in autism. This is reassuring considering that rats are one of the most common model organisms for human studies. This means that rats have the most infrastructure, namely DNA sequence and protein structure, already available to perform this research. [16]

Since regulation is so important, another factor that should be looked into is epigenetics and how the environment has a role in effecting the expression of both of these genes. Currently the effects of epigenetics on autism is a new and exploratory field that could be explored with these model organisms and these two genes. Specifically, there is very little research on epigenetics in regard to BAG3 and PALLD in any of the organisms studied. [7–11]

Conclusions:

Autism is a mental disorder in which multiple genes are involved in the development of its various symptoms. However, as a result of the challenges inherent to identifying the responsible genes, many studies are ongoing and all are inconclusive. [1] To date, studies have shown that two genes, BAG3 and PALLD, were upregulated in autistic individuals. [2–6]. However, no research has looked further into the role of these two genes in generating the manifestations of autism or the observed variability among individuals on the autism spectrum. This project analyzed the similarities and differences between the human BAG3 and PALLD genes and their counterparts in various model organisms. While it is likely that epigenetic modifications affect the expression and activity of these genes, they were not a focus of this work due to the paucity of prior research. [7–11] The autism model organisms used were blind cave fish, dogs, mice, rats, zebrafish, and zebra finch with red-bellied piranhas acting as the outgroup. It is hypothesized that there is a model organism best suited for studying BAG3 and/or PALLD that will lead to a better understanding of its role in human autism.

Comparisons were performed using a variety of bioinformatics tools including NCBI BLASTN [23], BLASTP [24], global alignments in MEGAX [25], and multiple pairwise sequence alignments in OAT [29] to identify mRNA variants in model organisms. The most similar variants of BAG3 and PALLD were then assessed for the significance of these variations on the predicted structure and expression of the encoded protein using UCSF Chimera [43]. Finally, the same variants' regulatory regions were predicted and compared in order to identify similarities and differences found upstream of the BAG3 and PALLD genes using MEGAX [25] for alignments, Berkeley Neural Network Promoter Prediction [59] along with DTU Health Tech Promoter - 2.0 for predicting promoter sites [60], and Softberry's web applications [61,62] for finding TATA boxes.

From NCBI BLASTN [23] and BLASTP [24], it was observed that there were three general groups consisting of the mammals (humans, dogs, rats, and mice), a group consisting of the fish (blind cave fish, zebrafish, and red-bellied piranhas), and lastly the zebra finch which had low similarity to both groups. These findings were also reflected in the global and pairwise alignments. It is determined that the organisms most similar to humans were dogs followed by rats. This allowed for a narrowed focus on humans, dogs and rats when looking at the predicted protein structure and predicted regulatory regions. It was found that while the different versions of BAG3 were more similar to each other than the versions of PALLD, the three versions of both genes were sufficiently similar to each other. Lastly, when predicting the regulatory regions it was found that there were similar types of promoters near each other on the aligned human and dog upstream sequences along with humans and rats but none between rat and dog for BAG3. However, minimal similarities existed with PALLD. Furthermore, no similarity was found in the type or location of the predicted TATA boxes in any variant of the gene.

The findings of this study suggest that the BAG3 and PALLD protein structures from the model organisms used, dog and rat, are sufficiently similar compared to those in humans so they can be used to better understand the genes in humans diagnosed with autism.

Furthermore, there are similarities in the regulatory regions predicted for select model organisms. However, these regulatory regions are in areas where there is very little known and therefore where future research into these genes' effect on autism should be focused. The results from this work provide guidance as well as evidence justifying the need for future research and experimental manipulation of the BAG3 and PALLD genes in the model organisms dogs and rats, to identify their role in autism.

Competing interests:

The writer claims no conflict of interest.

Acknowledgments:

I would like to acknowledge Dr. Gary Skuse for his assistance as my advisor on this project and thank him for helping to guide me. I also would like to thank my committee members Dr. Julie Thomas, Dr. Andre Hudson, and Dr. Paul Shipman for their input and advice on how to better approach the project. Finally, I would like to thank my family for their support throughout the entire project.

Literature Cited:

(1) Autism, NCBDDD, CDC. *Cent. Dis. Control Prev.* **2021**.

(2) Yoshizawa, M.; Settle, A.; Hermosura, M. C.; Tuttle, L. J.; Cetraro, N.; Passow, C. N.; McGaugh, S. E. The Evolution of a Series of Behavioral Traits Is Associated with Autism-Risk Genes in Cavefish. *BMC Evol. Biol.* **2018**, *18* (1), 89. https://doi.org/10.1186/s12862-018-1199-9.

(3) Voineagu, I.; Wang, X.; Johnston, P.; Lowe, J. K.; Tian, Y.; Horvath, S.; Mill, J.; Cantor, R. M.; Blencowe, B. J.; Geschwind, D. H. Transcriptomic Analysis of Autistic Brain Reveals Convergent Molecular Pathology. *Nature* **2011**, *474* (7351), 380–384. https://doi.org/10.1038/nature10110.

(4) Parikshak, N. N.; Swarup, V.; Belgard, T. G.; Irimia, M.; Ramaswami, G.; Gandal, M. J.; Hartl, C.; Leppa, V.; Ubieta, L. de la T.; Huang, J.; Lowe, J. K.; Blencowe, B. J.; Horvath, S.; Geschwind, D. H. Genome-Wide Changes in LncRNA, Splicing, and Regional Gene Expression Patterns in Autism. *Nature* **2016**, *540* (7633), 423–427. https://doi.org/10.1038/nature20612.

(5) Ansel, A.; Rosenzweig, J. P.; Zisman, P. D.; Melamed, M.; Gesundheit, B. Variation in Gene Expression in Autism Spectrum Disorders: An Extensive Review of Transcriptomic Studies. *Front. Neurosci.* **2017**, *10*. https://doi.org/10.3389/fnins.2016.00601.

(6) Rahman, M. R.; Petralia, M. C.; Ciurleo, R.; Bramanti, A.; Fagone, P.; Shahjaman, M.; Wu, L.; Sun, Y.; Turanli, B.; Arga, K. Y.; Islam, M. R.; Islam, T.; Nicoletti, F. Comprehensive Analysis of RNA-Seq Gene Expression Profiling of Brain Transcriptomes Reveals Novel Genes, Regulators, and Pathways in Autism Spectrum Disorder. *Brain Sci.* **2020**, *10* (10), 747. https://doi.org/10.3390/brainsci10100747.

(7) Bakulski, K. Environmental Epigenetics In Autism Spectrum Disorder. *Eur. Neuropsychopharmacol.* **2019**, *29*, S747–S748. https://doi.org/10.1016/j.euroneuro.2017.06.087.

(8) Eshraghi, A. A.; Liu, G.; Kay, S.-I. S.; Eshraghi, R. S.; Mittal, J.; Moshiree, B.; Mittal, R. Epigenetics and Autism Spectrum Disorder: Is There a Correlation? *Front. Cell. Neurosci.* **2018**, *12*, 78. https://doi.org/10.3389/fncel.2018.00078.

(9) Loke, Y. J.; Hannan, A. J.; Craig, J. M. The Role of Epigenetic Change in Autism Spectrum Disorders. *Front. Neurol.* **2015**, *6*. https://doi.org/10.3389/fneur.2015.00107.

(10) Tseng, C.-E. J.; McDougle, C. J.; Hooker, J. M.; Zürcher, N. R. Epigenetics of Autism Spectrum Disorder: Histone Deacetylases. *Biol. Psychiatry* **2022**, *91* (11), 922–933. https://doi.org/10.1016/j.biopsych.2021.11.021.

(11) Wiśniowiecka-Kowalnik, B.; Nowakowska, B. A. Genetics and Epigenetics of Autism Spectrum Disorder—Current Evidence in the Field. *J. Appl. Genet.* **2019**, *60* (1), 37–47. https://doi.org/10.1007/s13353-018-00480-w.

(12) Volkmar, F. R. *Encyclopedia of Autism Spectrum Disorders*, 2nd ed.; Springer Nature, 2021.

(13) *Autism in Adulthood*; Lowinger, S., Pearlman-Avnion, S., Eds.; Autism and Child Psychopathology Series; Springer International Publishing: Cham, 2019. https://doi.org/10.1007/978-3-030-28833-4.

(14) Stoodley, C. J.; D'Mello, A. M.; Ellegood, J.; Jakkamsetti, V.; Liu, P.; Nebel, M. B.; Gibson, J. M.; Kelly, E.; Meng, F.; Cano, C. A.; Pascual, J. M.; Mostofsky, S. H.; Lerch, J. P.; Tsai, P. T. Altered Cerebellar Connectivity in Autism and Cerebellar-Mediated Rescue of Autism-Related Behaviors in Mice. *Nat. Neurosci.* **2017**, *20* (12), 1744–1751. https://doi.org/10.1038/s41593-017-0004-1.

(15) Meyza, K. Z.; Blanchard, D. C. The BTBR Mouse Model of Idiopathic Autism – Current View on Mechanisms. *Neurosci. Biobehav. Rev.* **2017**, *76*, 99–110. https://doi.org/10.1016/j.neubiorev.2016.12.037.

(16) Berg, E. L.; Copping, N. A.; Rivera, J. K.; Pride, M. C.; Careaga, M.; Bauman, M. D.; Berman, R. F.; Lein, P. J.; Harony-Nicolas, H.; Buxbaum, J. D.; Ellegood, J.; Lerch, J. P.; Wöhr, M.; Silverman, J. L. Developmental Social Communication Deficits in the *Shank3* Rat Model of Phelan-Mcdermid Syndrome and Autism Spectrum Disorder: Social Communication in the *Shank3* Rat. *Autism Res.* **2018**, *11* (4), 587–601. https://doi.org/10.1002/aur.1925.

(17) James, D. M.; Kozol, R. A.; Kajiwara, Y.; Wahl, A. L.; Storrs, E. C.; Buxbaum, J. D.; Klein, M.; Moshiree, B.; Dallman, J. E. Intestinal Dysmotility in a Zebrafish (Danio Rerio) Shank3a;Shank3b Mutant Model of Autism. *Mol. Autism* **2019**, *10* (1), 3. https://doi.org/10.1186/s13229-018-0250-4.

(18) Garcia-Oscos, F.; Koch, T.; Pancholi, H.; Trusel, M.; Daliparthi, V.; Ayhan, F.; Co, M.; Alam, D. H.; Holdway, J. E.; Konopka, G.; Roberts, T. F. *Autism-Linked Gene FoxP1 Selectively Regulates the Cultural Transmission of Learned Vocalizations*; preprint; Neuroscience, 2020. https://doi.org/10.1101/2020.03.14.992016.

(19) Galambos, Á.; Petró, E.; Nagy, B.; Turcsán, B.; Topál, J. The Effects of Social and Non-Social Distracting Stimuli on Dogs with Different Levels of Social Competence – Empirical Evidence for a Canine Model of Autism. *Appl. Anim. Behav. Sci.* **2021**, *244*, 105451. https://doi.org/10.1016/j.applanim.2021.105451.

(20) Topál, J.; Román, V.; Turcsán, B. The Dog (Canis Familiaris) as a Translational Model of Autism: It Is High Time We Move from Promise to Reality. *WIREs Cogn. Sci.* **2019**, *10* (4). https://doi.org/10.1002/wcs.1495.

(21) Sun, H.-M.; Chen, X.-L.; Chen, X.-J.; Liu, J.; Ma, L.; Wu, H.-Y.; Huang, Q.-H.; Xi, X.-D.; Yin, T.; Zhu, J.; Chen, Z.; Chen, S.-J. PALLD Regulates Phagocytosis by Enabling Timely Actin Polymerization and Depolymerization. *J. Immunol.* **2017**, *199* (5), 1817–1826. https://doi.org/10.4049/jimmunol.1602018.

(22) Marzullo, L.; Turco, M. C.; De Marco, M. The Multiple Activities of BAG3 Protein: Mechanisms. *Biochim. Biophys. Acta BBA - Gen. Subj.* **2020**, *1864* (8), 129628. https://doi.org/10.1016/j.bbagen.2020.129628.

(23) *Nucleotide Blast: Search Nucleotide Databases Using a Nucleotide Query*. National Center for Biotechnology Information, U.S. National Library of Medicine. https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome.

(24) *Protein Blast: Search Protein Databases Using a Protein Query*. National Center for Biotechnology Information, U.S. National Library of Medicine. https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome.

(25) *MEGA Molecular Evolutionary Genetics Analysis*. https://megasoftware.net/.

(26) Kumar, S.; Stecher, G.; Li, M.; Knyaz, C.; Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol. Biol. Evol.* **2018**, *35* (6), 1547–1549. https://doi.org/10.1093/molbev/msy096.

(27) *About ClustalW*. MEGAX-Help. https://megasoftware.net/web_help_11/index.htm#t=Part_II_Assembling_Data_For_Analysis%2FBuilding_Sequence_Alignments%2FClustalW%2FAbout_ClustalW.htm&rhsearch=ClustalW&rhhlterm=ClustalW&rhsyns=%20.

(28) *Multiple Sequence Alignment*. Embl-Ebi. https://www.ebi.ac.uk/Tools/msa/.

(29) *OAT*. EZBioCloud. https://www.ezbiocloud.net/tools/orthoani.

(30) Lee, I.; Ouk Kim, Y.; Park, S.-C.; Chun, J. OrthoANI: An Improved Algorithm and Software for Calculating Average Nucleotide Identity. *Int. J. Syst. Evol. Microbiol.* **2015**, *66* (2), 1100–1103. https://doi.org/10.1099/ijsem.0.000760.

(31) Yoon, S.-H.; Ha, S.-M.; Kwon, S.; Lim, J.; Kim, Y.; Seo, H.; Chun, J. Introducing EzBioCloud: A Taxonomically United Database of 16S RRNA Gene Sequences and Whole-Genome Assemblies. *Int. J. Syst. Evol. Microbiol.* **2017**, *67* (5), 1613–1617. https://doi.org/10.1099/ijsem.0.001755.

(32) Bag3 Orthologs - NCBI. *Natl. Cent. Biotechnol. Inf. US Natl. Libr. Med.*

(33) PALLD Orthologs - NCBI. *Natl. Cent. Biotechnol. Inf. US Natl. Libr. Med.*

(34) (Author Name Not Available). UCLA-ASD. **2016**. https://doi.org/10.7303/SYN4587609.

(35) *ZNF804A Gene - Zinc Finger Protein 804A*. Gene Cards The Human Gene Database. https://www.genecards.org/cgi-bin/carddisp.pl?gene=ZNF804A#snp.

(36) *Variant: rs4380187*. GWAS Catalog. https://www.ebi.ac.uk/gwas/variants/rs4380187.

(37) *Variant: rs11693094*. GWAS Catalog. https://www.ebi.ac.uk/gwas/variants/rs11693094.

(38) *Variant: rs7597593*. GWAS Catalog. https://www.ebi.ac.uk/gwas/variants/rs7597593.

(39) *Gene: BAG3*. GWAS Catalog. https://www.ebi.ac.uk/gwas/genes/BAG3.

(40) *Gene: PALLD*. GWAS Catalog. https://www.ebi.ac.uk/gwas/genes/PALLD.

(41) Pandurangan, A. P.; Blundell, T. L. Prediction of Impacts of Mutations on Protein Structure and Interactions: SDM, a Statistical Approach, and MCSM, Using Machine Learning. *Protein Sci.* **2020**, *29* (1), 247–257. https://doi.org/10.1002/pro.3774.

(42) *Unipro UGENE*. Unipro UGENE. http://ugene.net/.

(43) *UCSF Chimera an Extensible Molecular Modeling System*. RVBI. https://www.cgl.ucsf.edu/chimera/.

(44) *SWISS-MODEL*. University of Basel Biozentrum The Center for Molecular Life Sciences. https://swissmodel.expasy.org/interactive.

(45) *Homo sapiens BAG cochaperone 3 (BAG3), mRNA CDS*. NCBI NIH National Library of Medicine Nucleotide. https://www.ncbi.nlm.nih.gov/nuccore/NM_004281.4?feature=CDS#feature_NM_004281.4_CDS_0.

(46) *Homo sapiens palladin, cytoskeletal associated protein (PALLD), transcript variant 1, mRNA CDS*. NCBI NIH National Library of Medicine Nucleotide. https://www.ncbi.nlm.nih.gov/nuccore/NM_001166108.2?feature=CDS#feature_NM_001166108.2_CDS_0.

(47) *Rattus norvegicus BAG cochaperone 3 (Bag3), mRNA CDS (no regulatory)*. NCBI NIH National Library of Medicine Nucleotide. https://www.ncbi.nlm.nih.gov/nuccore/NM_001011936.1?feature=CDS#feature_NM_001011936.1_CDS_0.

(48) *PREDICTED: Rattus norvegicus palladin, cytoskeletal associated protein (Palld), transcript variant X3, mRNA CDS (no regulatory)*. NCBI NIH National Library of Medicine Nucleotide. https://www.ncbi.nlm.nih.gov/nuccore/XM_039094916.1?feature=CDS#feature_XM_039094916.1_CDS_0.

(49) *PREDICTED: Canis lupus familiaris BAG cochaperone 3 (BAG3), transcript variant X8, mRNA CDS (no regulatory)*. NCBI NIH National Library of Medicine Nucleotide. https://www.ncbi.nlm.nih.gov/nuccore/XM_544046.6?feature=CDS#feature_XM_544046.6_CDS_0.

(50) *PREDICTED: Canis lupus familiaris palladin, cytoskeletal associated protein (PALLD), transcript variant X2, mRNA CDS (no regulatory)*. NCBI NIH National Library of Medicine Nucleotide. https://www.ncbi.nlm.nih.gov/nuccore/XM_038435147.1?feature=CDS#feature_XM_038435147.1_CDS_0.

(51) *Homo sapiens BAG cochaperone 3 (BAG3), mRNA regulatory*. NCBI NIH National Library of Medicine Nucleotide. https://www.ncbi.nlm.nih.gov/nuccore/NM_004281.4?feature=CDS#feature_NM_004281.4_regulatory_0.

(52) *Homo sapiens palladin, cytoskeletal associated protein (PALLD), transcript variant 1, mRNA regulatory*. NCBI NIH National Library of Medicine Nucleotide. https://www.ncbi.nlm.nih.gov/nuccore/NM_001166108.2?feature=CDS#feature_NM_001166108.2_regulatory_0.

(53) *Homo sapiens chromosome 4, GRCh38.p14 Primary Assembly NCBI Reference Sequence: NC_000004.12 Nucleotide Graphic Report*. NCBI NIH National Library of Medicine Nucleotide. https://www.ncbi.nlm.nih.gov/nuccore/568815594?report=graph&tracks=[key:gene_model_track,name:T2420844,display_name:Genes\, MANE Project (release v1.0),id:T2420844,dbname:SADB,category:Genes,subcategory:NCBI Genes,annots:NA000342948.1,Options:ShowAll,CDSProductFeats:false,NtRuler:true,AaRuler:true,HighlightMode:2,ShowLabel:true,shown:true,order:0][key:sequence_track,name:T15993,display_name:Sequence,id:T15993,dbname:GenBank,category:Sequence,subcategory:Assembly,annots:NA,ShowLabel:false,ColorGaps:false,shown:true,order:1][key:SNP_track,name:T2375272,display_name:5' UTR Region\, dbSNP b155 v2,id:T2375272,dbname:VDB,annots:NA000305581.1\236,Layout:Adaptive,shown:true,order:2][key:SNP_track,name:T2375270,display_name:3' UTR Region\, dbSNP b155 v2,id:T2375270,dbname:VDB,annots:NA000305581.1\237,Layout:Adaptive,shown:true,order:3][key:aggregate_feature_track,name:T2430859,display_name:Biological regions\, aggregate\, NCBI NCBI Homo sapiens Annotation Release 110,id:T2430859,subkey:biological_region,dbname:SADB,category:Genes,subcategory:NCBI Other Features,annots:NA000351982.1,Layout:Adaptive,shown:true,order:4][key:six_frames_translation,name:T18197,display_name:Six-frame translations,id:T18197,dbname:GenBank,category:Sequence,subcategory:Features,annots:Six-frame translation,ShowOption:All,OrfThreshold:20,HighlightCodons:true,AltStart:false,SubsetAltStart:false,shown:true,order:5]&assm_context=GCF_000001405.40&app_context=Gene&v=168432343:168993148&c=800080&select=null&slim=0.

(54) *Homo sapiens chromosome 10, GRCh38.p14 Primary Assembly NCBI Reference Sequence: NC_000010.11 Nucleotide Graphic Report*. NCBI NIH National Library of Medicine Nucleotide. https://www.ncbi.nlm.nih.gov/nuccore/568815588?report=graph&tracks=[key:sequence_track,name:T15993,display_name:Sequence,id:T15993,dbname:GenBank,category:Sequence,subcategory:Assembly,annots:NA,ShowLabel:false,ColorGaps:false,shown:true,order:0][key:gene_model_track,name:T2420844,display_name:Genes\, MANE Project (release v1.0),id:T2420844,dbname:SADB,category:Genes,subcategory:NCBI

Genes,annots:NA000342948.1,Options:ShowAll,CDSProductFeats:false,NtRuler:true,AaRuler:true,HighlightMode:2,ShowLabel:true,shown:true,order:1][key:SNP_track,name:T2375272,display_name:5' UTR Region\, dbSNP b155 v2,id:T2375272,dbname:VDB,annots:NA000305581.1\236,Layout:Adaptive,shown:true,order:2][key:SNP_track,name:T2375270,display_name:3' UTR Region\, dbSNP b155 v2,id:T2375270,dbname:VDB,annots:NA000305581.1\237,Layout:Adaptive,shown:true,order:3][key:aggregate_feature_track,name:T2430859,display_name:Biological regions\, aggregate\, NCBI NCBI Homo sapiens Annotation Release 110,id:T2430859,subkey:biological_region,dbname:SADB,category:Genes,subcategory:NCBI Other Features,annots:NA000351982.1,Layout:Adaptive,shown:true,order:4][key:six_frames_translation,name:T18197,display_name:Six-frame translations,id:T18197,dbname:GenBank,category:Sequence,subcategory:Features,annots:Six-frame translation,ShowOption:All,OrfThreshold:20,HighlightCodons:true,AltStart:false,SubsetAltStart:false,shown:true,order:5]&assm_context=GCF_000001405.40&app_context=Gene&v=119647413:119681784&c=99ccff&select=null&slim=0.

(55) *Rattus norvegicus BAG cochaperone 3 (Bag3), mRNA NCBI Reference Sequence: NM_001011936.1 Nucleotide Graphic Report*. NCBI NIH National Library of Medicine Nucleotide. https://www.ncbi.nlm.nih.gov/genome/gdv/browser/nucleotide/?id=NM_001011936.1.

(56) *PREDICTED: Rattus norvegicus palladin, cytoskeletal associated protein (Palld), transcript variant X3, mRNA NCBI Reference Sequence: XM_039094916.1 Nucleotide Graphic Report*. NCBI NIH National Library of Medicine Nucleotide. https://www.ncbi.nlm.nih.gov/genome/gdv/browser/nucleotide/?id=XM_039094916.1.

(57) *PREDICTED: Canis lupus familiaris BAG cochaperone 3 (BAG3), transcript variant X8, mRNA NCBI Reference Sequence: XM_544046.6 Nucleotide Graphic Report*. NCBI NIH National Library of Medicine Nucleotide. https://www.ncbi.nlm.nih.gov/genome/gdv/browser/nucleotide/?id=XM_544046.6.

(58) *Canis lupus familiaris isolate SID07034 breed Labrador retriever chromosome 25, ROS_Cfam_1.0, whole genome shotgun sequence NCBI Reference Sequence: NC_051829.1 Nucleotide Graphic Report*. NCBI NIH National Library of Medicine Nucleotide. https://www.ncbi.nlm.nih.gov/genome/gdv/browser/nucleotide/?id=NC_051829.1.

(59) *Berkeley Drosophila Genome Project Searches Neural Network Promoter Prediction*. BDGP. https://www.fruitfly.org/seq_tools/promoter.html.

(60) *Promoter - 2.0 Transcription start sites in vertebrate DNA*. DTU Health Tech. https://services.healthtech.dtu.dk/service.php?Promoter-2.0.

(61) *FPROM*. Softberry. www.softberry.com/berry.phtml?topic=fprom&group=programs&subgroup=promoter.

(62) *PromH(G) Recognition of human and animal Pol II promoters (Transcription Start Site and TATA-box)*. Softberry. http://www.softberry.com/berry.phtml?topic=promhg&group=help&subgroup=promoter.

(63) Jeffery, W. R. Astyanax Surface and Cave Fish Morphs. *EvoDevo* **2020**, *11*, 14. https://doi.org/10.1186/s13227-020-00159-6.

(64) *SRX958788: Astyanax aeneus Aeneus_Surface*. NCBI NIH National Library of Medicine Nucleotide. https://www.ncbi.nlm.nih.gov/sra/SRX958788[accn.

(65) *GSM2358068: SOX5 OverExp 1; Homo sapiens; RNA-Seq (SRR4444537)*. NCBI's Sequence Read Archive. https://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR4444537.

(66) *BioProject*. NCBI NIH National Library of Medicine Nucleotide. https://www.ncbi.nlm.nih.gov/bioproject.