Published online *Nucleic Acids Research, 0000, Vol. 00, No. 00* **1–9**
doi:10.1093/nar/gkn000

Analyzing the reason behind similar behavioral traits of domestic animals and human with William Syndrome

Nabila Shahnaz Khan and Rocco Clayton DiGiorgio V *

Department of Computer Science, University of Central Florida, Orlando, FL 32816, USA

ABSTRACT

Williams syndrome occurs when people are missing a chunk of DNA on chromosome 7. It results in certain types of physical traits such as hypersocial personalities, heart defects, intellectual disability, and more. Among the genes often deleted in Williams Syndrome are WBSCR17, GTF2I, and GTF2IRD1. In a study it was seen that these same genes were important in dog evolution. This hints towards the possibility that this region of genome might be responsible for the highly friendly behavior of both dogs and also people with William Syndrome. In this project, we tried to determine if there's any evidence supporting this claim. To do so, we test for any similarity between the sequence structure of the affected genes listed above, comparing humans to wolves and dogs, using well-known comparison methods like Longest Common Subsequence (LCS) and Sequence Alignment. We also converted the genes to mRNAs, predicted their secondary structures and compared the structures to see if there is similarity in the mRNA folding of these genes. Later, based on the data, we determined that there is at least marginal evidence for a relationship between human Williams Syndrome and dogs' behavioral evolution from wolves.

INTRODUCTION

Williams syndrome (WS) is a genetic condition that is present from birth and is caused by a contiguous deletion of a subset of 26-28 genes on chromosome 7. This condition is characterized by mild to moderate intellectual disability or learning problems, unique personality characteristics, distinctive facial features, and heart and blood vessel (cardiovascular) problems. Some other typical symptoms are difficulty with visual-spatial tasks such as drawing and assembling puzzles. Mostly people with WS tend to have outgoing, engaging personalities and they seem to be very friendly to other people (1).

According to studies, dogs are found to be more sociable compared to wolves. Researcher vonHoldt and her team in Princeton University have been studying the underlying genetic basis for social behavior in dogs and wolves for several years. The first hint of a link between dogs and WS came in 2010 (2), when biologist vonHoldt et al. found WBSCR17 and other genes near it were important in dog evolution and domestication. These genes are located near or in the same reigon where deletions of WS are found (3). This information hints towards the possibility that there might be some link between the genes affected by human WS and the genes important to the evolution of dogs' behavior.

No significant computational research has been done yet based on this potential link. So in our project we attempted to figure out computationally if this specific region involved in human WS has any similarity with the corresponding reigon in dogs, both at the level of sequence and of secondary structure. We used a global Longest Common Subsequence (LCS) algorithm to figure out length of the common portion between two sequences. We also planned to examine the actual common subsequences found to see if there is any repetitive common pattern. Then we used Global Sequence Alignment, as it is considered to be one of most efficient ways to establish correspondence between a pair of sequences (4).

We also split the genes into smaller overlapping reads and tried to convert them to their possible messenger RNA sequences (mRNA). mRNA are used to convey genetic information from an organism's DNA to the ribosomes, informing the structure of proteins by dictating the sequence of nucleic acids (5). As proteins are connected to structure and function of tissues and organs, and translation to protein is related to mRNA, we determined that comparing these constructed human and dog mRNAs would yield meaningful results.

First we predicted the secondary structure of the mRNAs. Then we compared the structures to see if they have any significant similarities. After analyzing the results of both approaches, we found that there's at least a small amount of evidence supporting a connection between human William's Syndrome and the behavioral evolution of dogs from wolves.

The rest of the paper is organized as follows. Section III discusses the findings in this field and our motivations behind this work. Section IV presents the data collection and preprocessing procedure. Computational method used in this project has been discussed elaborately in Section V and the results have been analyzed in Section VI. Section VII summarizes the whole discussion and future plans. Finally, Section VIII concludes the paper by highlighting the possible implications of this work.

*To whom correspondence should be addressed. Email: nabilakhan@Knights.ucf.edu; cldigiorgio@Knights.ucf.edu

LITERATURE REVIEW

Williams Syndrome, also known as Williams-Beuren Syndrome, is a rare genetic disorder characterized by extreme friendliness, a varying degree of mental deficiency, distinctive facial features, and other symptoms. Most affected individuals have a friendly, outgoing, talkative manner of speech (6). Near about 1 in 10,000 people worldwide are affected by WS (7). Individuals with WS are typically known to be hyper-social, they are sometimes intensely friendly and demand the attention of other people (8).

Causes of WS are sometimes unknown but recently, from an ongoing research, it has been found that it might result from a contiguous deletion of genes located on the long arm (q) of chromosome 7, or more specifically, the 7q11.23 region. According to research, 28 genes such as CLIP2, ELN, GTF2I, GTF2IRD1, and LIMK1 play crucial role in this case (6). The symptoms of Williams Syndrome can appear individually from each other. Individuals with the condition may have all the typical symptoms, or only a few. For example, one study found that deletion of ELN and LIMK1 was linked to the personality trait of WS (9).

Researcher VonHoldt previously had found that certain genes on chromosome 6 in dogs (similar to regions of chromosome 7 in Humans) have an important role in dogs' evolution from wolves. Researchers found that threes genes (WBSCR17, GTF2I , GTF2IRD1) were correlated with the behavioral traits of dogs as compared to wolves (10). In humans, two of these genes, GTF2I and GTF2IRD1, are located in the region typically containing the deletions found in those with WS, 7q11.23 (9, 11). The third gene, WBSCR17, which is named after the syndrome, is located near the typical deletion region in humans and so is thought to also play a role in WS (9, 11). So we designed this project to figure out if this implied claim, that there is similarity in the this particular genomic region between dogs and humans with WS, is reasonable.

DATA COLLECTION AND PROCESSING

We originally intended to collect more data than we did, however finding what we did was reasonably difficult. We searched, but were unable to find any more human chromosome 7 sequences than from the reference genome, and no researchers or institutions were willing to give us access to any chromosome sequences of individuals with WS. Additionally, such data was not publicly available. Locating the accession number for vonHold's data was also fairly difficult, requiring email correspondence with the researcher herself.

Data Collection

Two forms of data were obtained for this study: a human chromosome 7 sequence and a collection of dog and wolf chromosome 6 sequences. For the human chromosome, the human reference chromosome was obtained from NCBI. This was obtained as a raw sequence text file. The dog and wolf chromosomes were obtained from vonHoldt's study. More specifically, they were obtained from NCBI SRA using the accession number found in vonHoldt's paper (10). These chromosomes were downloaded in the form of binary sra files.

The molecular location of three genes WBSCR17, GTF2I, and GTF2IRD1 were found through NCBI for humans and dogs (11) separately. And the genes that, in WS, are associated with the personality-based symptoms were also researched (9).

Three python scripts were written:

1. for breaking up a sequence into reads of a given length with a given overlap
2. for isolating genes with given molecular start and endpoints from multiple input files, each with one input sequence, and saving them in individual files
3. for removing genes with given molecular start and endpoints from one input file, and saving the whole new sequence in a separate file

One bash script was written for converting binary sra files to fastq files and then into raw sequence text files.

Data Reformatting: vonHoldt's Dog and Wolf Data

The first step for the wolf and dog data was to convert from a binary format to a fastq format. This was done from a Linux commandline using the fastq-dump tool from NCBI's SRA Toolkit. The outputted fastq files were then stripped of all information apart from the actual nucleotide sequence using the commandline tool awk. Since this process worked one file at a time, a custom bash script (19) was written to automatically process all 24 sra files.

Data Preprocessing

The data for our study needed to be preprocessed in a few ways.

First, the human chromosome 7 was copied and then a sequence from nucleotide 74,027,772 to nucleotide 74,122,525 was removed to form an artificial example of a chromosome 7 from a person with William's Syndrome. This location contains two genes - ELN, LIMK1 - known to be associated with the personality traits associated with William's (9).

Both chromosome 7s were then broken up into reads of length 1100 base pairs, with each read sharing an overlap of 50 base pairs with the previous read in the sequence, for a total of 1000 unique base pairs per read, excluding the first and last.

Three genes, WBSCR17, GTF2I, and GTF2IRD1 were isolated from the human reference chromosome 7, the dog chromosome 6, and the wolf chromosome 6 and stored as separete files. This was not done for the artificial William's Syndrome chromosome.

Finally the previously isolated WBSCR17 genes from wolf, dog, and human were then broken into reads, again of length 1100 base pairs and 50 overlap. The reads were then converted to complementary RNA and stored in separate, individual files.

METHODOLOGY

In order to figure out if there is any significant similarity between dogs and people with WS, we also needed to check for the possibility of similarity between human and wolves.

Because, if human sequence have almost equal similarity to both dogs and wolves, then that might not have any significant role behind the similar behavioral traits. For that, we had to follow two steps. In 1st step, we had to do cross-species gene comparison and in the second step, we tested similarity between humans with WS and dogs. For these two approaches, we used different computational algorithms such as: Global Linear-Space LCS, Myers-Miller Sequence Alignment Algorithm, Zuker-Sankoff Minimum Energy Folding Algorithm and also used a tool called Efficient Alignment of RNA (ERA) for comparing the secondary structures of mRNA. The whole computational methodology followed in this project (shown in Figure 1) has been described step-by-step below:

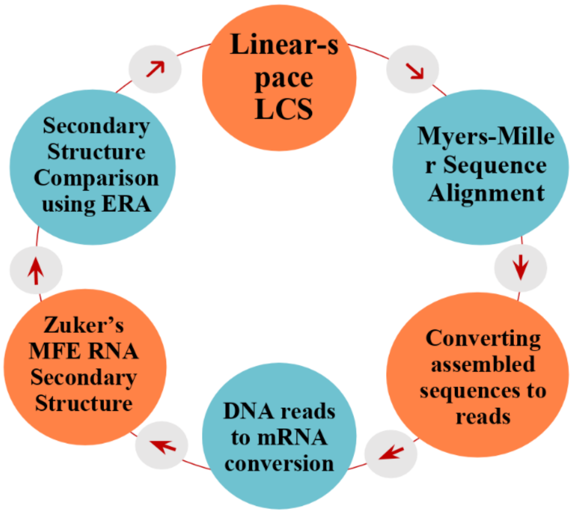
Linear-space LCS
Myers-Miller Sequence Alignment
Converting assembled sequences to reads
DNA reads to mRNA conversion
Zuker's MFE RNA Secondary Structure
Secondary Structure Comparison using ERA

Figure 1. Overall Computational Methodology Used

Cross-species gene comparison

According to study (10), WBSCR17, GTF2I and GTF2IRD1, these three genes are mostly responsible for the evolution of dogs' behavior. These three genes also exist in human chromosome 7. By nature, dogs are more social beings like human. To see if these three genes play any role in the behavioral similarity of dogs and humans with WS, first we compared these three genes of 16 different dogs to the corresponding human genes, found in the human reference genome. Then we compared these same three genes from the human reference to the corresponding three genes of 8 different wolf sequences.

For comparing the three genes, first we used the memory efficient version of the LCS algorithm. Then we used the Myers-Miller sequence alignment algorithm. Due to the huge data size, we used only gene WBSCR17 for further processing. According to these results, the dog sequence with serial no SRR5500858 had highest similarity to human reference gene WBSCR17. So, we split the human WBSCR17 gene and the dog WBSCR17 gene from SRR5500858 into small reads and then converted these DNA reads to mRNA sequences. Later, we predicted the secondary structure of the mRNA sequences using the Zuker-Sankoff algortihm. Lastly, we compared the secondary structure of each read of WBSCR17 gene for that dog to each read of human WBSCR17 gene.

Applying LCS: The Longest Common Subsequence algorithm tries to find the longest common subsequence between two given sequences. Here we used this algorithm to see if humans and dogs have longer common subsequence than humans and wolves. Also, the plan was to align the longest common subsequences found between human and dog sequences to see if there is any conserved structure. We also used LCS to find the longest common subsequence length between a full chromosome 7 of a human with WS and a dog chromosome 6. Then, we also applied LCS between the human reference chromosome 7 and a dog chromosome 6. After that, we compared these scores to see if the chromosome 7 of the human with WS has higher LCS score than the reference human chromosome 7. If this is the case, it would mean that the chromosome 7 of the human with WS has more in common with dog's chromosome 6 than the reference human chromosome 7.

Normally LCS requires two (n×n) arrays, one for calculating the score and another for backtracking. If we are only calculating score, then this can be done using one (2×n) array. But for finding out common patterns between the longest common subsequences of the genes, we also needed the subsequences as output. As the length of the input sequences are really large (near about 500k bp), using two (n×n) array wasn't an option. Even computers with 8Gb/16GB/32GB weren't sufficient enough to allocate this huge amount of memory. So we had to modify the code so that it uses one (2×n) array and two separate hard drive files (one holds the whole 2D matrix for calculating the score and another stores the matrix in reverse order so that it can be used later for backtracking). The code can be found in the 'Supplementary Files' (19) section.

Applying Sequence Alignment: For finding the similarity between two sequences using computational methods, sequence alignment is a very useful technique. Basically two approaches are followed, in one approach, alignment scores are increased with each aligned base pair. In the other approach, penalties are added for each mismatch or gap (insertion or deletion). For global sequence alignment, Gotoh's algorithm (12), introduced in 1982, is very common but requires O(MN) space. This can't be used for very long sequences. Using a modified Gotoh's algorithm, alignment cost between two sequences can be calculated using linear space, but without returning the number of insertions, deletions, matches and mismatches. These values are required for calculating the identity value of two sequences, and so are required for our study. Using Hishberg's concept (13), in 1988, Myers et al. implemented Gotoh's algorithm using linear space and this solution can return the aligned sequence along with alignment cost. This algorithm uses the concept of Merge Sort along with Dynamic Programming. In our project, we implemented this alignment algorithm for calculating the alignment cost between two sequences. Like the LCS problem, we used this alignment algorithm to compare genes WBSCR17, GTF2I, GTF2IRD1 of dogs and humans and then we compared these same three genes of humans and wolves. The implemented code can be found in the 'Supplementary

Files’ (19) section. For measuring the similarity of two sequences, we calculated the identity score of two sequences using the following formula:

$$\text{Identity} = \frac{\text{Match}}{\text{Match}+\text{Mutation}+\text{Insertion}+\text{Deletion}} \qquad \textbf{(1)}$$

Converting Genes to Small Reads: For converting the DNAs to mRNAs, we had to split the human WBSCR17 gene and dog serial number SRR5500858’s WBSCR17 gene (the dog gene with maximum LCS and alignment score) into smaller reads. The human WBSCR17 gene was 588725bp long and the dog SRR5500858’s WBSCR17 gene was 437764bp long. We split the human gene into 554 overlapping reads and the dog gene into 411 overlapping reads. Here size of each gene is 1100bp where any middle read has 50 overlapping base-pairs in the beginning and 50 overlapping base-pairs in the end. The code (readSplitter.py) used for splitting the sequences can be found in the ‘Supplementary Files’ (19) section.

Transcription - converting DNA reads to mRNA: Transcription uses DNA as a template to make a complimentary RNA molecule (known as messenger RNA) as shown in Figure 2. Here, we converted the small reads of DNA to mRNA. The code of conversion (readSplitter.py) is placed in the ‘Supplementary Files’ (19) section.

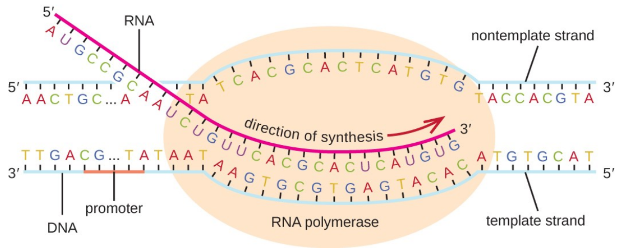

Figure 2. Transcription Process (14)

Secondary Structure Prediction of mRNA: The study of RNA secondary structure is critical for understanding the function and regulation of mRNA, as it plays an important role in the bio-synthesis of proteins (15). There are various secondary structure prediction algorithms, but among them, the Zuker-Sankoff Minimum Free Energy Folding algorithm (16) is the most common one. The recursive equations of this algorithm are shown in Figure 3. Here, W(i) holds the minimum energy of a structure on s[1. . . i], V(i,j) holds the minimum energy of a structure on s[i. . . j] with s[i] and s[j] forming a base-pair, and WM(i,j) holds the minimum energy of a structure on s[i. . . j] that is part of multi-loop. From the recursive equation it can be clearly seen that the run-time for this algorithm is O(n^4), which is huge for large sequences. Because of this runtime restriction, we had to keep the size of reads comparatively smaller. Smaller read size was also preferable as in reality, mRNA size is not that large on average. The ViennaRNA package(17) was used to visualize the secondary structure of the mRNAs. The **Images of Secondary Structures** of the mRNAs for both human and dog reads can be found in the output section of the ‘Supplementary Files’ (19).

$$\boxed{\begin{aligned} W(i) = \min\{&W(i-1),\\ &\min_{0\le k<i}\{W(k)+V(k+1,i)\}\}. \end{aligned}}$$

$$\boxed{\begin{aligned} V(i,j) = \min\{&eH(i,j),\\ &eS(i,j,i+1,j-1)+V(i+1,j-1),\\ &\min_{i<i'<j'<j \text{ and } i'-i+j-j'>2}\{eL(i,j,i',j')+V(i',j')\},\\ &\min_{i+1<k<j}\{WM(i+1,k-1)+WM(k,j-1)+a\}, \end{aligned}}$$

$$\boxed{\begin{aligned} WM(i,j) = \min\{&V(i,j)+b,\\ &WM(i,j-1)+c,\\ &WM(i+1,j)+c,\\ &\min_{i<k\le j}\{WM(i,k-1)+WM(k,j)\}\}, \end{aligned}}$$

Figure 3. Zuker-Sankoff Minimum Free Energy RNA Folding (14)

Secondary Structure Comparison The secondary structures of probable mRNAs of dog SRR5500858’s WBSCR17 were compared to the probable mRNAs of the human WBSCR17 gene using Efficient alignment of RNA (ERA) tool (18). This tool was implemented using the sparse dynamic programming technique and requires a runtime of O(n^3). Here, each dog read compared to each human read. In total, 412 × 554 calculations were done which made the process lengthy. The tool was run on these 412 × 554 combinations using a bash script.

Similarity testing of Human with William Syndrome and Dogs

In the second step, the dog chromosome 6 was compared to the chromosome 7 of a human with WS, using LCS and sequence alignment. As the chromosome size is really large (near about 184,000,000 bp), it was divided into overlapping reads and then compared. Then the dog chromosome was similarly compared to the human reference chromosome 7. The aim of these comparisons was to figure out if the dog chromosome 6 has more similarity to chromosome 7 of a human with William Syndrome than to a normal human chromosome 7.

RESULTS

The results of all the implemented algorithms have been analyzed below in this section:

Result of Cross-species Comparison

In the cross-species compassion method, there are six sets of comparisons. One table for each set has been shown [Table 1-6] where the results of running the LCS and Alignment algorithms have been given. Each table shows one gene of interest, for example WBSCR17, and the results of comparing each dog or wolf instance of it to the human version. Here, three tables hold data of 16 dog each (one table each for genes WBSCR17, GTF2I and GTF2IRD1) and other three tables hold data of 8 wolves each (again for genes WBSCR17,

Table 1. Dog WBSCR17 - LCS and Alignment Scores against Human WBSCR17

Sequence Serial	LCS Score	Matches (Alignment)	Replacements (Alignment)	Insertions (Alignment)	Deletions (Alignment)	Alignment Score	Percent Identity
SRR5500853	315917	240946	178251	13162	162259	446932	40.5211
SRR5500854	316974	241499	176725	14135	163232	447164	40.5478
SRR5500855	315325	239824	178716	13819	162916	447060	40.2879
SRR5500856	317326	241735	177182	13442	162539	446250	40.6347
SRR5500857	316634	241243	177581	13535	162632	445336	40.5457
SRR5500858	318036	242408	176501	13450	162547	447640	40.7473
SRR5500859	315775	240661	177766	13932	163029	448536	40.4209
SRR5500860	317990	241983	176998	13378	162475	445556	40.6808
SRR5500861	315955	240330	178101	13928	163025	451108	40.3655
SRR5500862	316149	240891	177974	13494	162591	447346	40.4893
SRR5500863	315966	240780	177768	13811	162908	447358	40.4491
SRR5500864	315845	240863	177871	13625	162722	446928	40.4757
SRR5500865	316986	241365	177670	13324	162421	446502	40.5806
SRR5500866	314928	239927	177727	14705	163802	448406	40.2453
SRR5500867	315325	239824	178716	13819	162916	447060	40.2879
SRR5500868	317326	241735	177182	13442	162539	446250	40.6347

GTF2I and GTF2IRD1). Each table contains six columns where Sequence Serial holds the serial number of the dog or wolf chromosome and LCS score column holds the length of the longest common subsequence found. The rest of the six columns hold the output parameters of the Alignment algorithm which are number of matches, replacements, insertions, deletions, alignment score and percentage of sequence identity. A comparison was shown between the genes of Dogs and Wolves based on Average LCS score and Average Alignment Score in Table 7. Then the secondary structure of each read of the human WBSCR17 (total 554 reads) was compared to each read of dog SRR5500858’s WBSCR17 (total 412 reads). In total 412×554 pairs of structures were compared. As the data set is huge, we have simply discussed summary values of the output (The full output data can be found in ‘Supplementary Files’ (19) section.). The analysis of the output data for each phase is given below:

Comparing WBSCR1 Gene: First we had to compare human WBSCR17 to dog WBSCR17 and then human WBSCR17 to wolf WBSCR17. The results are shown in Table 1 and 2 respectively. The LCS score value for both dogs and wolves were in the range of 314000 to 319000. Comparatively, the LCS scores for dogs as compared to humans were slightly more than wolves as compared to humans in general. From Table 7 we can see that the Average LCS score value for dogs is 316404 and for wolves it is 316331, so for dogs the mean value is slightly higher. The highest LCS score found among dogs is 318036 (for dog SRR5500858) and highest LCS score for wolves is 317837 (for wolf SRR5500849). The value of Percent Sequence Identity of Alignment is linear to the value of LCS score as shown in Figure 4. Similarly for wolf sequences, the relation between Percent Sequence Identity and LCS score is linear. The average alignment score for Dog sequences is 447215 and average Percent Identity is 40.495, whereas for wolves, the average alignment score and average Percent Identity are 447151 and 40.4938625 respectively. Though the difference is not large, both average LCS score and alignment score are still higher for dogs compared to wolves. This might be an indication of the fact that after mutations, dog WBSCR17 gene is more similar to human WBSCR17 gene than wolf WBSCR17 gene.

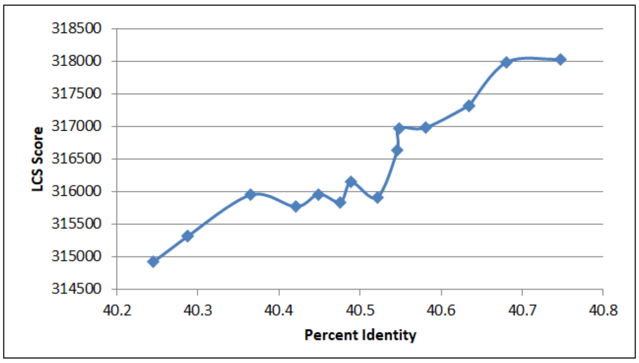

Figure 4. Linear relation between the Percent Alignment Identity and LCS Score of WBSCR17 gene of Dog and Human

Comparing GTF2I Gene: The comparison result for gene GTF2I has been shown in Table 3 and 4. From these two tables it can be seen that the relation between Alignment score and Percent Identity is less linear compared to WBSCR17 gene (for both dogs and wolves). The relation between Alignment Score and Percent Identity for Dogs has been shown in Figure 5. We can see some break points (near x=43 and x=43.8) for which LCS score was high, even though the Alignment Percent Identity was low. For example, Dog SRR5500858 has a higher percent Identity (43.06) than dog SRR5500860 (43.01) but a lower LCS score. From Table 7 we can see that, for this gene, though the LCS score is high for dogs, the alignment is higher for Wolves. So, we couldn’t reach any significant conclusion for this gene.

Comparing GTF2IRD1 Gene: In case of Gene GTF2IRD1 (shown in Table 5 and 6), the value of Average LCS Score and Average Alignment Score is significantly higher for dogs as compared to wolves. This indicates the possibility of more similarity between the GTF2IRD1 gene of human and dogs.

Table 2. Wolf WBSCR17 - LCS and Alignment Scores against Human WBSCR17

Sequence Serial	LCS Score	Matches (Alignment)	Replacements (Alignment)	Insertions (Alignment)	Deletions (Alignment)	Alignment Score	Percent Identity
SRR5500845	316944	241361	177710	13288	162385	446886	40.5823
SRR5500846	316545	241312	177599	13448	162545	446470	40.5632
SRR5500847	316996	241622	177467	13270	162367	447160	40.6274
SRR5500848	314335	238987	179036	14336	163433	446698	40.1125
SRR5500849	317837	242086	176920	13353	162450	446986	40.6998
SRR5500850	314646	239537	178743	14079	163176	448610	40.2222
SRR5500851	316799	241742	176713	13904	163001	446860	40.6043
SRR5500852	316548	241237	177507	13615	162712	447538	40.5392

Table 3. Dog GTF2I - LCS and Alignment Scores against Human GTF2I

Sequence Serial	LCS Score	Matches (Alignment)	Replacements (Alignment)	Insertions (Alignment)	Deletions (Alignment)	Alignment Score	Percent Identity
SRR5500853	69095	53072	42816	18201	7140	90943.5	43.7783
SRR5500854	68132	52411	43049	18629	7568	91974.5	43.081
SRR5500855	69216	53083	42884	18122	7061	91076.5	43.8159
SRR5500856	67899	52215	43446	18428	7367	91474.5	42.9909
SRR5500857	67849	52261	43306	18522	7461	91888.5	42.9955
SRR5500858	68021	52341	43221	18527	7466	91690.5	43.0595
SRR5500859	69122	53149	42688	18252	7191	91575.5	43.8234
SRR5500860	68265	52252	43384	18453	7392	92016.5	43.0125
SRR5500861	67779	51803	43977	18309	7248	92321.5	42.6935
SRR5500862	69169	53239	42400	18450	7389	91379.5	43.826
SRR5500863	68789	52979	42841	18269	7208	91262.5	43.6771
SRR5500864	69288	53265	42732	18092	7031	92745.5	43.977
SRR5500865	68207	52363	42969	18757	7696	92543.5	42.9963
SRR5500866	69317	53265	42635	18189	7128	90426.5	43.9419
SRR5500867	69216	53083	42884	18122	7061	91076.5	43.8159
SRR5500868	67899	52215	43446	18428	7367	91474.5	42.9909

Comparing the mRNA secondary Structures After comparing all possible combinations of reads (mRNAs) of human WBSCR17 to dog SRR5500858's WBSCR17 (selected because it had the highest Sequence Alignment Score), it was seen that Human read 156 and Dog read 208 had the max alignment score 24455 and Optimal Pair matching 1386 with 106 pairs matched. The average alignment score for all the mReads pairs is 8900, which is high. This points towards the possibility that on average there might be some similarity in the folding patterns of the mRNAs. This might be related to their similar behavioral traits. We will be able to draw a more solid conclusion after following the same procedure with other dogs' WBSCR17 sequence and also some wolf WBSCR17 sequences. If the possible dog WBSCR17 mRNAs have more similar folding to the human WBSCR17 mRNAs than the wolves' do, then that might hold a significant meaning behind their similar sociable nature.

Testing Similarity of Dogs and Humans (with and without William Syndrome) Chromosome

The Human Chromosome 7 (184386054 bp long) was split into 151759 reads, Human Chromosome 7 with WS (161241941 bp long) into 151668 reads and Dog SRR5500858 Chromosome 6 (594359308 bp long) into 561608 reads. LCS and Alignemnt algortihm was applied first on 15×151759 combinations of human and dog chromosome reads. Then it was applied on 15×151668 combinations of human with WS and dog chromosome reads. For the combinations of human and dog reads, the read-pairs with highest Alignment score were selected one after another, the cross-pairing alignments

Table 4. Wolf GTF2I - LCS and Alignment Scores against Human GTF2I

Sequence Serial	LCS Score	Matches (Alignment)	Replacements (Alignment)	Insertions (Alignment)	Deletions (Alignment)	Alignment Score	Percent Identity
SRR5500845	67732	52013	43910	18166	7105	90502.5	42.9171
SRR5500846	69068	53098	43003	17988	6927	92278.5	43.8768
SRR5500847	68354	52436	43224	18429	7368	92004.5	43.1725
SRR5500848	67622	51937	43969	18183	7122	91057.5	42.8484
SRR5500849	69001	53079	42606	18404	7343	91929.5	43.7109
SRR5500850	68587	52400	43403	18286	7225	92711.5	43.1937
SRR5500851	68432	52490	43150	18449	7388	91117.5	43.2098
SRR5500852	69136	53037	42847	18205	7144	92055.5	43.748

Table 5. Dog GTF2IRD1 - LCS and Alignment Scores against Human GTF2IRD1

Sequence Serial	LCS Score	Matches (Alignment)	Replacements (Alignment)	Insertions (Alignment)	Deletions (Alignment)	Alignment Score	Percent Identity
SRR5500853	79402	60150	46458	3263	42557	113333	39.4613
SRR5500854	79080	59926	46639	3306	42600	113646	39.3032
SRR5500855	78973	59980	46354	3537	42831	113205	39.2791
SRR5500856	79271	59777	46561	3533	42827	114593	39.1472
SRR5500857	78880	59667	46587	3617	42911	113900	39.0537
SRR5500858	78735	59525	46699	3647	42941	113625	38.9531
SRR5500859	79500	60237	46130	3504	42798	113407	39.4559
SRR5500860	79943	60804	45347	3720	43014	114152	39.7711
SRR5500861	81079	61765	44515	3591	42885	113946	40.4338
SRR5500862	79569	60496	45900	3475	42769	113772	39.6331
SRR5500863	79751	60478	46058	3335	42629	114182	39.6577
SRR5500864	79607	60383	46311	3177	42471	113776	39.6365
SRR5500865	79447	60163	46356	3352	42646	113355	39.4468
SRR5500866	79561	60417	45868	3586	42880	113249	39.5526
SRR5500867	78973	59980	46354	3537	42831	113205	39.2791
SRR5500868	79271	59777	46561	3533	42827	114593	39.1472

Table 6. Wolf GTF2IRD1 - LCS and Alignment Scores against Human GTF2IRD1

Sequence Serial	LCS Score	Matches (Alignment)	Replacements (Alignment)	Insertions (Alignment)	Deletions (Alignment)	Alignment Score	Percent Identity
SRR5500845	79242	59892	46617	3362	42656	113296	39.2665
SRR5500846	78960	59820	46666	3385	42679	112973	39.2134
SRR5500847	79278	60349	46144	3378	42672	113867	39.562
SRR5500848	78932	60077	45967	3827	43121	113455	39.2681
SRR5500849	79222	59821	46664	3386	42680	112871	39.2138
SRR5500850	79541	60471	45673	3727	43021	114490	39.5514
SRR5500851	79239	60251	46115	3505	42799	112700	39.4649
SRR5500852	79173	60346	46141	3384	42678	112858	39.5584

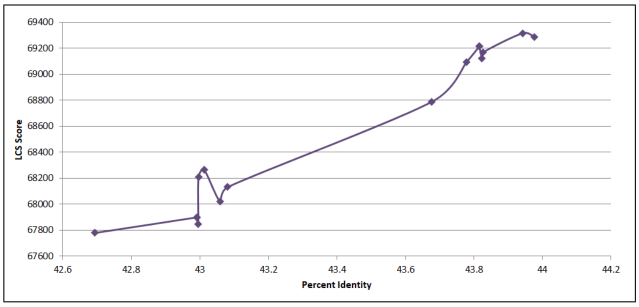

Figure 5. Linear relation between the Percent Alignment Identity and LCS Score ofGTF2I gene of Dog and Human

were discarded. The calculated average alignment score was found to be 1104. After following the similar procedure, the average LCS score was found to be 734. For the comparison of the dog and the human with WS, the average Alignment Score was calculated as 1104 and average LCS score was found to be 728. The average LCS and alignment score between the dog and the human with WS is slightly larger than between the dog and the human without WS. This might mean that they have more similarity, but due to the huge amount of data, we couldn’t finish running the procedure on the whole input data before submission, so the output might not be completely reliable.

Table 7. Comparsion of Average LCS Score and Alignment Score

	Dog		Wolf	
Gene	Average LCS Score	Average Alignment Score	Average LCS Score	Average Alignment Score
Gene WBSCR17	316404	447215	316331	447151
Gene GTF2I	68579	91617	68492	91707
Gene GTF2IRD1	79440	113746	79198	113314

Table 8 contains data from vonHoldt’s study pertaining to the friendliness of the subjects whose DNA sequences are used in this study. ABS stands for attentional bias to social stimuli, HYP for hypersociability, and SIS for social interest in strangers (10). PC1, PC2, and PC3 are the result of dimensionally reducing six other data values to three uncorrelated ”principle components” (10). PC1 was found to represent an autonomous personality phenotype, PC2 was found to represent a tendency towards approaching strangers, and PC3 was tentatively associated with reliance on humans for help solving tasks (10).

Table 9 contains the friendliness scores, calculated for each animal using the following formula:

Friendliness Score = zScore(ABS) + zScore(HYP)+ zScore(SIS) - zScore(PC1) + zScore(PC2) + zScore(PC3)

Table 8. Data from vonHoldt's Study(2)

Read ID	Animal ID	Species	ABS	HYP	SIS	PC1	PC2	PC3
SRR5500868	2769	dog	0.864	362.4	122.4	-1.321	0.453	-0.764
SRR5500867	2771	dog	0.823	311.4	84.96	-1.041	-0.211	-1.03
SRR5500866	2772	dog	0.326	415.2	198	-0.944	2.244	-0.936
SRR5500865	2773	dog	0.424	223.44	10.2	0.214	-1.634	-1.214
SRR5500864	2774	dog	0	180.84	7.92	1.289	-1.681	-1.083
SRR5500863	2775	dog	0.548	376.8	150	-1.412	0.645	-0.802
SRR5500862	2776	dog	1.246	390.36	161.4	-1.973	0.638	-0.001
SRR5500861	2777	dog	1.031	239.4	80.76	-0.52	-0.47	0.089
SRR5500860	2778	dog	0.995	344.4	147.6	-1.32	0.367	-0.106
SRR5500859	2779	dog	1.188	308.16	126.48	-1.92	-0.739	1.395
SRR5500858	2780	dog	1.067	324.36	135.12	-1.518	-0.156	0.563
SRR5500857	2781	dog	0.704	230.4	63.12	-0.423	-1.059	-0.032
SRR5500856	2782	dog	0.863	267.96	79.2	-0.983	-0.987	0.278
SRR5500855	2783	dog	0.585	289.08	141.24	-0.416	0.239	0.397
SRR5500854	2784	dog	0.538	420.36	181.08	-1.33	1.498	-0.984
SRR5500853	2785	dog	1.393	306.24	127.08	-2.545	-1.124	2.356
SRR5500852	2786	wolf	0.167	332.28	168.36	-0.174	1.155	-0.108
SRR5500851	2787	wolf	0	222.6	83.76	1.25	-0.689	-0.187
SRR5500850	2788	wolf	0	87.12	44.76	3.086	0.02	0.532
SRR5500849	2789	wolf	0	111.48	36.36	2.687	-0.496	0.131
SRR5500848	2790	wolf	0	262.08	179.52	2.404	2.168	0.415
SRR5500847	2791	wolf	0	166.32	164.52	2.69	1.662	1.815
SRR5500846	2792	wolf	0	168.6	48.84	2.224	-0.4	-0.474
SRR5500845	2793	wolf	0	140.4	25.2	1.995	-1.442	-0.249

From the table we can see that Dog SRR5500862, SRR5500866, SRR5500859, SRR5500858 and SRR5500854 have highest Friendliness Score. Among these, For gene WBSCR17, dog SRR5500858 had the highest LCS and Alignment Score. For gene GTF2I, dog SRR5500866 had the highest LCS score and high alignment score. And for gene GTF2IRD1, dog SRR5500862, SRR5500866, SRR5500859 and SRR5500854 had comparatively high LCS and Alignment Score. So there might be a possibility that dogs with High LCS and Alignment Score for these three genes are more likely to have a high Friendliness Score. To be certain about that, we need to perform this same procedure on a larger dataset.

DISCUSSION

Scientific Significance

Based on the results found in this study, it appears that there is at least marginal evidence for a relationship between human WS and dogs' behavioral evolution from wolves. For example, the difference in human-wolf and human-dog alignment scores are small, but they are still present. This would mean that the differences between dogs and wolves could serve as a loose model for human WS. If this is the case, it may help to advance research on the condition, lessening its rarity as a limiting factor.

Statistical Significance

It is possible that these results happened by chance. Our sample size was fairly small. While we had data from 16 dogs and 8 wolves, we only had one human example. Since we had only one human chromosome 7, we also only one human copy of each of the three genes of interest. Running these tests again with more human data would lend a lot of support to these findings. Also, the reference genome for Human with William Syndrome wasn't publicly accessible. With sufficient amount of practical data of Human with William syndrome, this computational method might have been able to bring new evidence to light about the recent Hypothesis about the connection between similarity of gene mutations of Domestic Animals and Human with William Syndrome.

Table 9. Friendliness Scores of Each Subject Calculated from vonHoldt's Study(2)

Read ID	Animal ID	Species	Compiled Friendliness Score
SRR5500868	2769	dog	2.252789941
SRR5500867	2771	dog	-0.05032069457
SRR5500866	2772	dog	4.140647103
SRR5500865	2773	dog	-5.272436732
SRR5500864	2774	dog	-7.167794772
SRR5500863	2775	dog	2.393949058
SRR5500862	2776	dog	5.397749335
SRR5500861	2777	dog	0.2601084366
SRR5500860	2778	dog	3.419416897
SRR5500859	2779	dog	4.110253124
SRR5500858	2780	dog	3.537873974
SRR5500857	2781	dog	-1.529441303
SRR5500856	2782	dog	0.204121492
SRR5500855	2783	dog	1.786717456
SRR5500854	2784	dog	3.865644234
SRR5500853	2785	dog	5.612438582
SRR5500852	2786	wolf	1.937274152
SRR5500851	2787	wolf	-3.546582012
SRR5500850	2788	wolf	-5.290439217
SRR5500849	2789	wolf	-5.841544018
SRR5500848	2790	wolf	1.014987594
SRR5500847	2791	wolf	0.6814383966
SRR5500846	2792	wolf	-5.34020485
SRR5500845	2793	wolf	-6.576646176

CONCLUSION

The connection between mutated genes implicated in both the evolution of dogs' behavior and in human William Syndrome, made regarding their similarly hyper-social behavior, might be able to answer some intriguing questions related to the specific functions of some genes. It might also hint towards potential genetic causes of this particular behavioral trait. To be sure, computational analysis of the genomic regions highlighted in this study might find very significant results. In the future, it might lead to substantial research.

REFERENCES

1. Williams syndrome - Genetics Home Reference - NIH. Link: https://ghr.nlm.nih.gov/condition/williams-syndromegenes
2. Author, Bridgett M. vonHoldt and Others (2010) Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. *Nature*, **464**, 898, publisher Nature Publishing Group.
3. Author Nala Rogers Rare Human Syndrome May Explain Why Dogs are So Friendly (2017). Link: https://www.insidescience.org/news/rare-human-syndrome-may-explain-why-dogs-are-so-friendly.
4. Author, Stephen F. Altschul and Author Mihai Pop (2017) Sequence Alignment–Handbook of Discrete and Combinatorial Mathematics. publisher CRC Press/Taylor & Francis.
5. Author Lodish, H., Berk, A., Zipursky, S. L., Matsudaira, P., Baltimore, D., & Darnell, J.(2000) The three roles of RNA in protein synthesis. book title Molecular Cell Biology. 4th edition publisher WH Freeman.
6. Williams Syndrome - NORD. Link: https://rarediseases.org/rare-diseases/williams-syndrome/
7. What is Williams Syndrome? — Williams Syndrome Association. Link: https://williams-syndrome.org/what-is-williams-syndrome
8. Galaburda, A. M., Holinger, D. P., Bellugi, U., Sherman, G. F. (2002). Williams syndrome: neuronal size and neuronal-packing density in primary visual cortex. *Archives of Neurology*, **59(9)**, 1461-1467.
9. Author, Colleen A Morris, MD, FACMG, FAAP (2017). Williams Syndrome - NCBI Bookshelf. Link: https://www.ncbi.nlm.nih.gov/books/NBK1249/
10. Author, Bridgett M. vonHoldt and Others (2017) Structural variants in genes associated with human Williams-Beuren syndrome underlie stereotypical hypersociability in domestic dogs. *Science Advances*, **3**, e1700398, publisher American Association for the Advancement of Science.
11. NCBI Gene. Link: https://www.ncbi.nlm.nih.gov/gene
12. Author, Gotoh and Author, Osamu (1982) An improved algorithm for matching biological sequences. *Journal of molecular biology*, **162**, 705–708, publisher Elsevier.
13. Author, Hirschberg and Author, Daniel S(1975) An improved algorithm for matching biological sequences. *Communications of the ACM*, **18**, 341–343, publisher ACM.
14. RNA Transcription. Mechanisms of Microbial Genetics. Link: https://courses.lumenlearning.com/microbiology/chapter/rna-transcription/
15. Author, Paulo Gaspar, Author, Gabriela Moura, Author Manuel A. S. Santos and Author José Luís Oliveira1 (2013) mRNA secondary structure optimization using a correlated stem–loop prediction. *Nucleic acids research*, **41**, e73–e73, publisher Oxford University Press.
16. Author Michael Zuker and Author David Sankoff (1984) RNA secondary structures and their prediction. *Bulletin of mathematical biology*, **46**, 591–621, publisher Springer.
17. ViennaRNA package. Link: https://www.tbi.univie.ac.at/RNA/
18. Author, Paulo Gaspar, Author, Gabriela Moura, Author Manuel A. S. Santos and Author José Luís Oliveira1 (2013) Efficient alignment of RNA secondary structures using sparse dynamic programmin. *BMC bioinformatics*, **14**, 269, publisher BioMed Central.
19. Supplementary Files. Link: https://drive.google.com/drive/folders/1PjRerc43RCbBVm6eS2tCNqvYPqsK9FJX?usp=sharing