

Yale University

## EliScholar – A Digital Platform for Scholarly Publishing at Yale

---

Public Health Theses

School of Public Health

---

1-1-2019

### The Potential Of Next Generation Whole Genome Sequencing Using Dogs As A Model To Understand Human Diseases

Emily Xie  
emilyxie@gmail.com

Follow this and additional works at: <https://elischolar.library.yale.edu/ysphtdl>

---

#### Recommended Citation

Xie, Emily, "The Potential Of Next Generation Whole Genome Sequencing Using Dogs As A Model To Understand Human Diseases" (2019). *Public Health Theses*. 1851.  
<https://elischolar.library.yale.edu/ysphtdl/1851>

This Open Access Thesis is brought to you for free and open access by the School of Public Health at EliScholar – A Digital Platform for Scholarly Publishing at Yale. It has been accepted for inclusion in Public Health Theses by an authorized administrator of EliScholar – A Digital Platform for Scholarly Publishing at Yale. For more information, please contact [elischolar@yale.edu](mailto:elischolar@yale.edu).

**The Potential of Next Generation Whole Genome Sequencing Using Dogs as a Model to  
Understand Human Diseases**

By  
Emily Xie

Year of Completion: 2019

In Candidacy for the Degree of  
Master of Public Health  
2019

A Thesis Presented to  
The Faculty of the Yale School of Public Health  
Department: Chronic Disease Epidemiology  
Committee Chair: Josephine Hoh, PhD  
Committee Member: Andrew Dewan, PhD

## **Abstract**

*Background:* There has been extensive research in the study of the dog genome and comparative genomics to human diseases. Dogs were proposed as a candidate primarily because of the relatively low variation within breeds but high variation between breeds. In this study, we used a conserved gene among dogs, *SMN*, to understand the genetic variability across dog breeds and to compare with human *SMN1*.

*Methods:* Using a sequential method design, new dog samples are added into the database as they become available. This paper is an application of the newly developed algorithms in studying the similarities and differences in genes between dogs and humans.

*Results:* In our analysis, we isolated the sequence for *SMN* across three dogs (two English Bulldogs and one French Mastiff) and compared the sequence to the reference dog *SMN* sequence. We identified a number of SNPs but the differences were located in the intron region suggesting that potential difference in the exon regions may exhibit deleterious effects in the animal. The total genetic variation we observed in our three dog samples is less than 1%. When comparing the reference dog *SMN* to human *SMN1*, we observed conserved sequences predominantly within the exon region of *SMN1*. The conserved sequences located in the intron region between *SMN1* and dog *SMN* may suggest that those regions may serve a regulatory function in gene expression.

*Conclusion:* There were no meaningful genetic variation within the exon region among our dog samples for *SMN*. As additional dog sequences from different breeds are acquired, the comparison of *SMN* will be conducted to better understand the genetic variation of the conserved gene.

## Table of Contents

<b>Introduction .....</b>	<b>3</b>
<b>Methods.....</b>	<b>4</b>
<b>Results.....</b>	<b>8</b>
<b>Discussion.....</b>	<b>10</b>
<b>Table 1: Dogs in the dissertation study (first three samples in the sequential collection).....</b>	<b>12</b>
<b>Table 2: NanoDrop measurements of sample genomic DNA.....</b>	<b>13</b>
<b>Table 3: Qubit Fluorometer Measurements of Sample genomic DNA .....</b>	<b>14</b>
<b>Table 4: Output for majority consensus sequence for each position across the SMN gene for the individual sample .....</b>	<b>15</b>
<b>Table 5: Output of analysis program that compares the overall majority sequence for each sample to the reference SMN sequence .....</b>	<b>16</b>
<b>Table 6: Total genetic variation across samples 1Y, 2O, and MW for SMN .....</b>	<b>17</b>
<b>Table 7: Total genetic variation in SMN across 2,000 bp blocks for samples 1Y, 2O, and MW when compared to the reference SMN.....</b>	<b>18</b>
<b>Table 8: Total genetic variation across samples 1Y, 2O, and MW for SMN when compared with the reference SMN.....</b>	<b>18</b>
<b>Figures 1a-1c: An example of read depth for Samples 1Y, 2O, and MW across the dog SMN gene. ....</b>	<b>19</b>
<b>1a: .....</b>	<b>19</b>
<b>1b: .....</b>	<b>20</b>
<b>1c:.....</b>	<b>21</b>
<b>Figure 2: The reference dog SMN sequence was compared with the human SMN sequence..</b>	<b>22</b>
<b>References.....</b>	<b>23</b>

## Introduction

In leveraging the application of NGS technology, the genome of *Canis lupus familiaris* or dogs was selected as a model for studying human conditions. Dogs have been proposed as a potential candidate for comparative gene mapping and disease identification among humans (Deschenes et al. 1994; Parker et al. 2010). There has been extensive research in the study of the dog genome and comparative genomics to human diseases by Elaine Ostrander and her team in the past two decades. Ostrander proposes that dogs serve as an excellent model for investigating complex human diseases. Firstly, dogs and humans share similar health diseases that may be caused by similar genes between the two species such as cancer and autoimmune diseases (Sutter and Ostrander et al. 2004; Parker et al. 2010). Secondly, the unique breeding patterns for dogs have led to a relatively phenotypic homogeneous population thereby reducing genetic diversity within breeds (Ostrander et al. 2010). Lastly, dogs and humans share similar environmental exposures and so the history of disease onset and progression may be similar between these two species (Shearin et al. 2010).

The dog genome contains approximately 3 billion base pairs and 39 pairs of chromosomes (NHGRI 2018). It is estimated that about 25% of the dog sequence aligns to the human genome (Kirkness et al. 2003) making the dog an attractive model for comparative analysis in understanding the genetic basis of diseases between dogs and humans. In this project, the Illumina NGS technology will be used to sequence the genome of the English Bulldog and French Mastiff with 30-times read coverage. Not only has this technique never been applied in this form of research and it will expand upon existing knowledge about the dog genome, and will be used as a comparison to the existing reference genome. Compared to other NGS technologies (SOLiD system, Roche 454 system, and Illumina HiSeq system), Illumina's HiSeq system generates the greatest sequencing output at the lowest operating cost (Liu et al. 2012).

Through employing the Illumina sequencing system to generate reads for the dog genome, we will investigate a highly conserved gene across dog breeds, comparing it with the human gene equivalent. The study of comparative genomics is a growing field and with advancing technology, there is a need for new algorithms to handle large datasets. Comparing the conserved genes across species to identify regions of similarity and difference may allow researchers to better understand the structure and function of human genes, health conditions, and ultimately improve treatment options. Identifying DNA sequences that have been preserved in different dog breeds and in humans may contribute to the development of innovative treatments for complex human diseases that would enhance health.

The conserved gene of interest within the dog genome in this study is *SMN*. *SMN* or survival motor neuron is a gene associated with the human condition, spinal muscular atrophy or SMA (Nizzardo et al. 2015). Humans have two copies of the gene, *SMN1* and *SMN2*, both of which express the survival motor neuron (SMN) protein. This protein is highly expressed in the spinal cord and is critical for the function of motor neurons. By studying the gene in another species, we may decode the mystery and understand the structure and function of the conserved gene. This knowledge may be valuable in developing newer therapies in addressing the medical needs for patients with SMA.

## Methods

Using a sequential method design, new samples are added into the database as they become available. This paper is an application of the newly developed algorithms in studying the similarities and differences in genes between dogs and humans. The overall aim of the project is to study the genetic variation among conserved genes in dogs and to understand the function it may serve in order to investigate human diseases. In this study, we will be studying a highly conserved gene, *SMN*, to study the genetic variability across dog breeds and to compare the gene

with the human equivalent. Refer to Table 1 for characteristics of the dogs. There are three sub-aims

1. In a sequential manner, collect and prepare blood sample from dogs for Next Generation Whole Genome Sequencing.
2. Utilize Illumina platform to perform WGS technology to sequence the samples.
3. Apply the analysis of *SMN* and identify conserved regions, biological relevant SNPs and indels within promoter and/or exon regions if applicable.

The analysis is conducted on the initial three dog samples acquired.

#### *DNA Extraction*

Blood samples were purified and extracted using the QIAamp Blood Midi Kit (Spin Protocol).

Two milliliters of whole blood were collected per tube and 1x PBS was used to bring the volume of the sample up to 2 ml if necessary before proceeding to purify and extract genomic DNA.

Purified and extracted DNA samples were stored at -20°C. The concentration of the purified

DNA was determined via the Nanodrop spectrophotometer and by PicoGreen Assay before

library preparation. These samples were then sent to Yale Center for Genome Analysis (YCGA)

for Next Generation DNA sequencing.

#### *NanoDrop DNA Quantification*

Table 3 provides the recorded concentration for five representative dog samples via the

NanoDrop. The three dog samples (1Y, 2O, and MW) used for analysis was not incorporated in the table.

The Nanodrop spectrophotometer is a common lab instrument used to measure the concentration of DNA and the A260/280 ratio indicates potential sources of contamination among the samples.

Blood samples from five dogs were extracted and purified. Depending on the amount of blood available for extraction, some samples were divided into two or four vials. As we do not have

control over the amount of leftover blood for collection, we maximized the amount of genomic DNA extracted when possible. Samples 1 and 2 were divided into four vials for DNA extraction and purification. Sample 3,4 and 5 were divided into two vials for DNA extraction and purification.

#### *PicoGreen Assay*

Table 4 provides the recorded concentration for five representative dog samples via the PicoGreen Assay. The three dog samples (1Y, 2O, and MW) used for analysis was not incorporated in the table.

The Pico Green DNA quantification assay is a sensitive method to detect small amounts of double stranded DNA in samples through a DNA binding fluorescent dye. The assay reports an exact concentration of the purified genomic DNA samples.

#### *Whole Genome Sequencing by NGS technology*

2ug of genomic DNA as measured by the PicoGreen assay was required for the PCR-FREE library prep approach. Sequencing was run on the Illumina NovaSeq 6000 and 30X depth. The Illumina platform used bridge amplification for sequencing by synthesis. Paired end reads were generated to maximize coverage and confidence in determining sequence.

#### *Alignment and Extraction of aligned reads*

The Burrow-Wheeler Aligner (BWA) software package was used for mapping sequences to a reference genome (CANFAM3.1) (Heng et al. 2009). The BWA-MEM algorithm was used and alignments were outputted in the SAM format and converted to the BAM output file for analysis. BWA-MEM is an alignment algorithm that performs mapping of paired-end reads to a reference genome (Heng et al. 2013). Tolerance for sequencing errors is 3% error using the BWA-MEM algorithm for 200bp. Reads are 151bp long (Heng et al. 2009). The alignment file is outputted in



the standard SAM format. The Samtools software package via the *Samtools view* command line was used to extract reads for regions that covered the *SMN* and outputted in a text file.

#### *Analysis Programs Developed By Yitao Yan, Hang Li, and Josephine Hoh*

The software was written by computational scientists from Dr. Josephine Hoh's lab and is still currently being revised for improvement. The initial program assesses all of the reads for each position and summarizes the sequence with the majority of reads in agreement with each other and highlights any mismatches. The output file contains information of the general sequence for the gene after consolidating information about all the reads for each position to propose a potential single consensus sequence for the sample. The major limitation in this program is that it is not sensitive to indels, gaps, and does not take into consideration the quality score of the reads. Therefore, the accuracy of the consensus sequence may be reduced.

The follow up program takes the output from Program 1 and conducts a comparison of the sequence file for each sample. This comparison is conducted between the publicly available reference dog sequence (CANFAM3.1) and our sequenced samples. Here, all proposed majority sequences for the gene were compared among all samples and to the reference. The summary output describes any mismatches in the majority sequence between each sample and to the reference sequence.

#### *Annotation of the Gene*

Annotation for *SMN* was extracted from NCBI and used for annotating the gene through the Integrated Genome Browser software. The annotation is based on the publicly available reference genome of the Boxer (CANFAM3.1). This is essentially a "copy and paste" mechanism leveraging the position of the nucleotides which is compared to the reference dog genome.

## Results

### *DNA Quantification as a Measure for Quality:*

The Nanodrop spectrophotometer is a common lab instrument used to measure the concentration of DNA and the A260/280 ratio indicates potential sources of contamination among the samples. Among the five dog samples, the mean DNA concentration ranged from 41.2 ng/ul to 50.5 ng/ul. The A260/280 ratio is used to determine the presence of protein contamination in the sample. Pure DNA samples should have a ratio of 1.8 and a lower value suggests potential contamination which may impact downstream application of the DNA for sequencing. Overall, the ratio for all of the purified DNA samples ranged between 1.66 to 1.83. Two samples (2d and 6b) suggested potential protein contamination and were excluded for WGS sequencing.

The following measures were obtained via the PicoGreen assay. The range of the concentrations for the five samples were between 25.2 ng/ul to 50.7 ng/ul (Table 4). When compared to the NanoDrop measurements, the Pico Green reported concentrations at a lower value suggesting that the NanoDrop overestimates the actual DNA concentration of the extracted samples. It is essential that the quantity and concentration of DNA is sufficient for WGS.

### *Data Quality of the Assembly*

The highly conserved *SMN* gene was used as an example for conducting the comparative analysis across dog samples. Read depth was used as a rough measure of the quality for the assembly. Using NGS, the whole genome of the dog samples was sequenced with 30X coverage and reads of length 151bp. The average read depth across dog *SMN* for samples 1Y, 2O, and MW are  $24.6 \pm 6.76$ ,  $17.35 \pm 5.30$ , and  $37.12 \pm 9.86$  respectively (Figures 1-3)

### *Genetic Variation among the three Dog samples in relation to the Reference Gene.*

The *SMN* region for each of the three samples were extracted and the overall majority sequence for the *SMN* gene was generated. The majority sequence for a position was determined by calculating the number of bases for each position. The position with the greatest number of consenting reads was the proposed majority sequence for the position. The majority sequence for the three samples were compared with the reference *SMN* sequence.

Based on the pairwise comparison of the majority sequence between each of the samples to the reference *SMN* sequence, there were no genetic variations identified within the exon regions. All of the genetic differences were identified in the intron region. When conducting a pairwise comparison of the gene between the three samples and the reference sequence, MW had the most genetic differences than 1Y or 2O. This is expected as 1Y and 2O are English Bulldogs and is hypothesized to have similar genetic patterns within the gene. MW is a French Mastiff and contains the most genetic variation within *SMN* when compared to the reference which is a boxer. The total genetic variation in 1Y, 2O, and MW are 0.12%, 0.10%, and 0.58% respectively.

### *Comparison of Reference Dog SMN to human SMN1*

Currently we are still developing an algorithm to compare the dog *SMN* to human *SMN1* and *SMN2*. Publicly available programs such as UCSC blat tools were used to conduct the comparative analysis between dogs and humans. The reference dog *SMN* gene was extracted and compared to the human *SMN1* gene via USCS blat software. *SMN1* is located on Chr5: 70,925,030-70,953,012 (27,983 bp) and the dog *SMN* is located on Chr2: 54,596,006-54,636,762 (40,757 bp). The overall genetic similarity between dog *SMN* and human *SMN1* is 89.5%, mapping to about 7,413 base pairs in human *SMN1*. There appears to be conserved regions across exons 2 to 7 between human *SMN1* and dog *SMN*. This suggests that these conserved

regions may serve an important function in survival among dogs and humans. Among intron 2, 4, and 5 in human *SMN1* there are conserved sequences between dogs and humans suggesting potential regulatory functions within those regions.

## Discussion

The overarching goal of the project on a large scale is to utilize the dog genome as a surrogate to study what is unknown in gene variability in relation to function and diseases in understanding human conditions. Dogs were proposed as a candidate primarily because of the relatively low variation within breeds but high variation between breeds. The project is hypothesis-free, and data are generated in a sequential manner as additional dog samples are acquired and added into the growing genome database. In this project, we used a conserved gene among dogs, *SMN*, to understand the genetic variability across dog breeds and to compare with human *SMN1*.

In our study, we aimed to compare a highly conserved gene across dog breeds to identify potential genetic variations that may have biological implications for the progression of SMA. In our analysis, we isolated the sequence for the gene across three dogs and compared the sequence to the reference dog sequence. We identified a number of SNPs but the differences were located in the intron region suggesting that potential difference in the exon regions may exhibit deleterious effects in the animal. As additional dog sequences from different breeds are acquired, the comparison of *SMN* will be conducted to better understand the genetic variation of the conserved gene. Literature suggests that the clinical progression of hereditary canine spinal atrophy is phenotypically similar to human SMA, but the condition is not associated with dog *SMN* (Blazej et al. 1998). Although *SMN* in dogs may not be associated with HCSMA, the gene may play a critical role in survival as it is highly conserved across the three dog samples.

The major limitation in the study is that the programs used to generate the majority sequence across the gene has flaws. The program does not consider the quality of the reads generated by the NGS system. The quality of the sequence is reported as a character that corresponds to the level of confidence in calling the sequence for the region. Therefore, the major limitation of this analysis is the lack of quality control measures to assess the accuracy or precision of the majority sequence that is generated by the program. The reported sequence may not be accurate with high certainty. Further research is recommended to develop robust algorithms and software programs to process the sequencing data generated by NGS for comparative analysis of dog and human genome. Lastly, future approaches in the project will include sequencing more dogs across different breeds to build an extensive dog genome database. Other potential genes associated with canine and human cancers will be analyzed to explore the genetic variation across dog breeds.

Table 1: Dogs in the dissertation study (first three samples in the sequential collection)

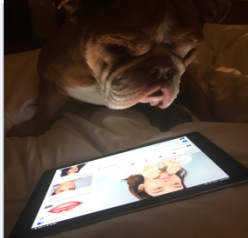


ID	Breed	Age in 2019	Gender	
1Y	English Bulldog	11	Male	
2O	English Bulldog	5.5	Male	
MW (MW030918)	French Mastiff	9	Female	

Table 2: NanoDrop measurements of sample genomic DNA: Five representative dog samples were extracted and purified. Samples 1Y, 2O, and MW were not incorporated in the table. The amount of blood acquired varied and were divided into two or four vials depending on the quantity of blood available. The five samples are numbered (1-5) and divided into vials labeled (a-d) if applicable. Columns 1 and 2 refers to the ID for each tube. Column 3 represents the concentration of DNA measured. Columns 4 and 5 refer to the 260/280 and 260/230 ratios respectively for each sample tube. Columns 6-9 represent the average DNA concentration, standard deviation, minimum concentration, and maximum concentration for each sample (1-5).

Samples concentration ( NanoDrop)					Concentration (ng/ul)			
sample	code on tube	concentration	260/280	260/230	mean	sd	min	max
1a	Anna-Ch-1-DW-091218	52.3	1.8	3.76	46.05	20.39	23.4	71
1b	Anna-Ch-2-DW-091218	37.5	1.74	2.82				
1c	Anna_YW1	71	1.84	0.84				
1d	Anna_YW2	23.4	1.72	3.59				
2a	Scotia-GI-1-DW-091218	72.5	1.83	3.22	50.5	16.66	35.1	72.5
2b	Scotia-GI-2-DW-091218	53.9	1.79	2.81				
2c	Scotia_YW1	40.5	1.81	4.74				
2d	Scotia_YW2	35.1	1.67	3.09				
3a	Chevy-GI-CL-091318	38.7	1.73	2.03	41.35	3.75	38.7	44
3b	Chevy-GI-YW-091318	44	1.81	0.98				
4a	Unch-Hoh-CL-091318	49.3	1.87	2.08	45.15	5.87	41	49.3
4b	Unch-Hoh-YW-091318	41	1.78	2.19				
5a	GS-PO-DSIII_Wu	64.3	1.83	2.85	41.2	18.19	21.1	64.3
5b	GS-PO-DSIII_Wu2	21.1	1.66	2.4				
5c	GS-PO-DSIII_Li	44.7	1.87	2.78				
5d	GS-PO-DSIII_Li2	34.7	1.78	2.41				

**Table 3: Qubit Fluorometer Measurements of Sample genomic DNA:** The DNA concentration of the five representative dog samples were quantified via the Qubit Fluorometer. Samples 1Y, 2O, and MW were not incorporated in the table. Columns 1 and 2 refer to the identification coding for the tube and corresponds to the ID coding used for the NanoDrop measurements. Column 3 refers to the reported DNA concentration. Columns 4-5 refers to the quantity of DNA obtained and the corresponding net DNA concentration yield.

<b>Sample</b>	<b>Code on tube</b>	<b>Qubit (ng/ul)</b>	<b>Volume (ul)</b>	<b>Yield (ng)</b>
1a+1b	Anna-CH-1-DW	30.1	150	4515
2a+2b	Scotia-GL-1-DW	50.7	150	7605
3a	Chevy-GL-CL	25.2	150	3780
4a	RB-DW-110718	45.2	300	13560
5a	GS-PO-DS111-WU	49.2	150	7380



Table 4: Output for majority consensus sequence for each position across the SMN gene for the individual sample. This is a pairwise comparison between the sample *SMN* sequence and the reference dog *SMN* sequence. Column 1 refers to the overall consensus sequence and is determined based on the majority sequence for the position. Column 2 refers to the position of the sequence relative to the reference dog sequence. Column 3 provides the total number of reads identified within each position. Column 4 outputs the different base pairs relative to the reference sequence for that position. Column 5 refers to the number of different base pairs for the position.

## Program 1: Output for Sample 1

- Compares each sample to reference individually, output is text file

Overall consensus sequence	Position rel. to Ref	# reads	Diff BP	# Diff BP
A	54595214	20	-	(0)
T	54595215	20	-	(0)
G	54595216	19	-	(0)
T	54595217	20	C	(1)
T	54595218	20	A	(1)
A	54595219	20	T	(1)
A	54595220	19	T	(1)
T	54595221	19	A	(1)
T	54595222	17	A	(1)
G	54595223	17	A/T	(1/1)
A	54595224	17	T	(1)
A	54595225	17	T	(1)
T	54595226	17	-	(0)

\*\*\*Programs 1 and 2 were developed by Yitao and Josephine\*\*\*  
 \*\*\*Heng is currently revising programs 1 and 2 \*\*\*

Table 5: Output of analysis program that compares the overall majority sequence for each sample to the reference SMN sequence. Any differences in base pairs are denoted by a star. The differences are potential genetic variants or SNPs within the *SMN* gene across the three samples and the dog reference sequence. Column 1 refers to the position of the sequence across the *SMN* gene. Column 2 provides the reference dog sequence for the position. Columns 3-5 provides the overall majority sequence for samples 1Y, 2O, and MW respectively. Column 6 may provide a star to indicate potential SNPs at the position.

## Program 2

- Compares all the samples to each other and to the reference sequence, output is text file.

Position	Reference	1Y	2O	MW	
54597720	T	T	T	A	* → Denotes a difference in base pair to reference
54597721	T	T	T	T	
54597722	T	T	T	C	*
54597723	T	T	T	T	
54597724	T	T	T	T	
54597725	T	T	T	G	*
54597726	T	T	T	C	*
54597727	T	T	T	A	*
54597728	T	T	T	A	*
54597729	T	T	T	T	
54597730	T	T	T	T	
54597731	T	T	T	A	*
54597732	T	T	T	T	
54597733	T	T	T	G	*
54597734	T	T	T	C	*
54597735	T	T	T	T	

\*\*\*Programs 1 and 2 were developed by Yitao and Josephine\*\*\*  
 \*\*\*Heng is currently revising programs 1 and 2 \*\*\*

Table 6: Total genetic variation across samples 1Y, 2O, and MW for SMN. The overall majority sequence for each sample is compared to the reference *SMN*. Reference *SMN* is obtained from the publicly available dog genome (CanFam3.1) through the National Center for Biotechnology Information (NCBI).

## Gene of Interest: *SMN*

<i>SMN</i> Region (54,596,000- 54,635,999)	1Y vs REF	2O vs REF	MW vs REF
Total genetic variation (SNPs)	61	51	290
Percent variation across <i>SMN</i> (%)	0.12	0.10	0.58

Table 7: Total genetic variation in SMN across 2,000 bp blocks for samples 1Y, 2O, and MW when compared to the reference SMN. The x-axis refers to the 2,000 base pair block across the *SMN* gene. The y-axis refers to the total number of genetic variation within the 2,000 bp region/ total base pairs in the *SMN* region. Samples 1Y, 2O, and MW are represented by the blue, red, and gray bars respectively.

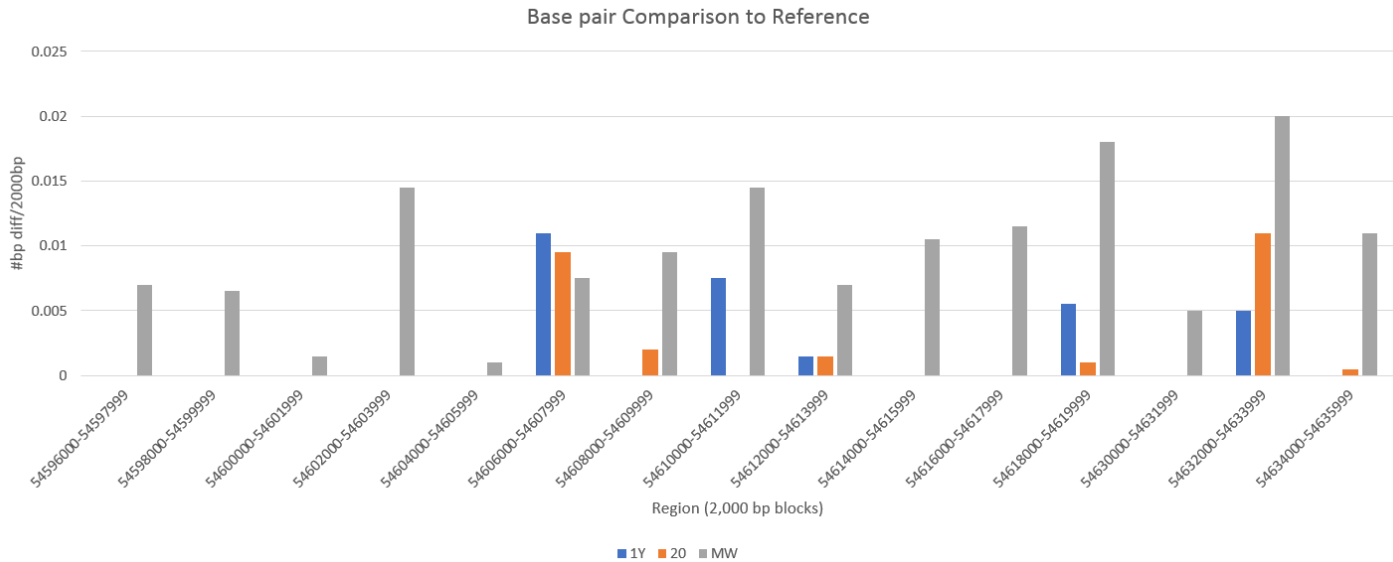
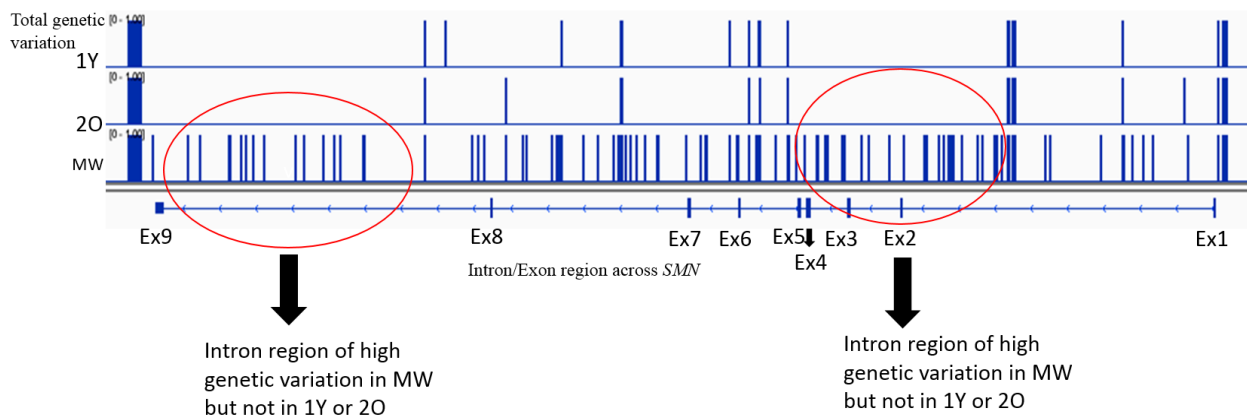


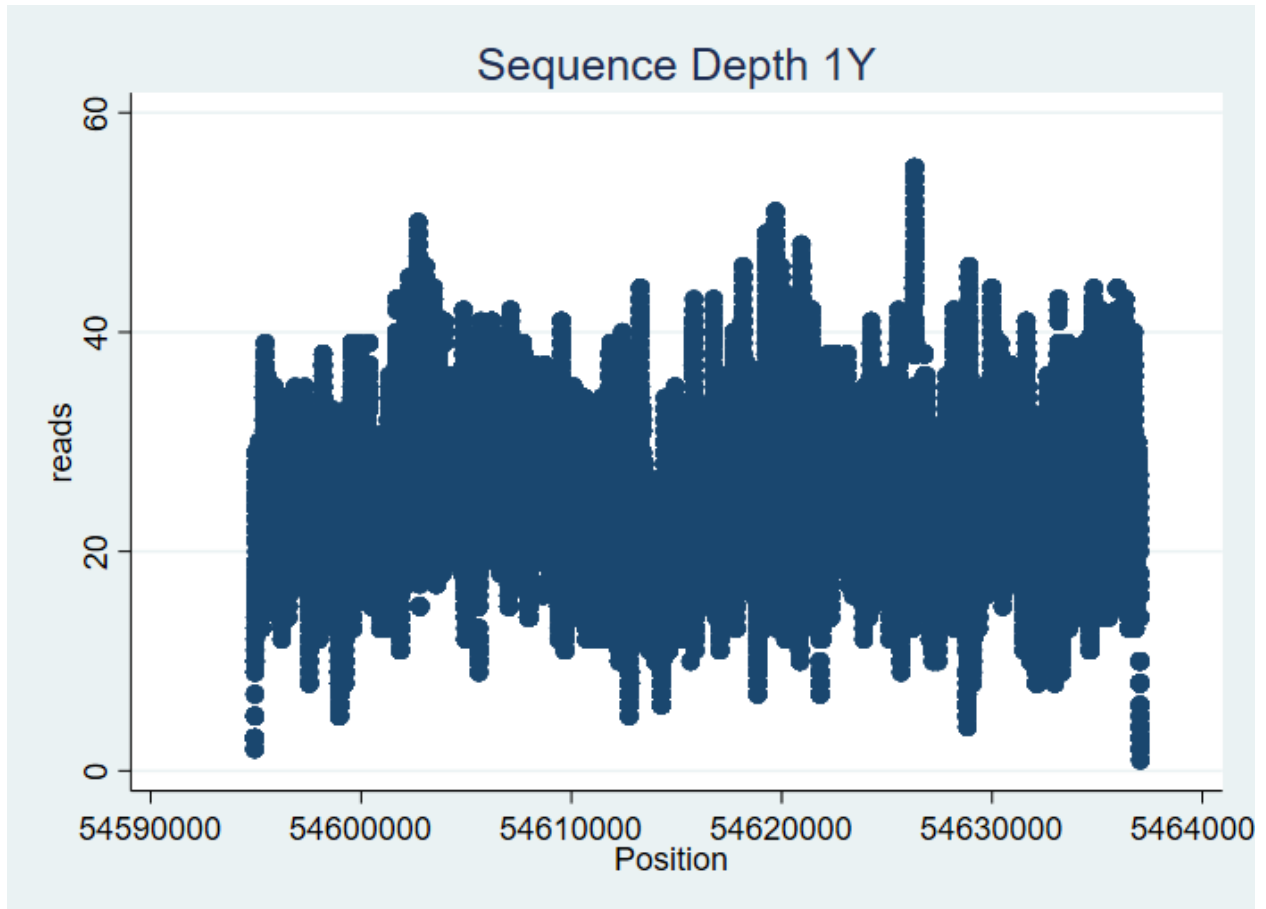
Table 8: Total genetic variation across samples 1Y, 2O, and MW for SMN when compared with the reference SMN. The x-axis refers to the position across the gene while the y-axis represents the total genetic variation for the position.

### Comparison of 1Y, 2O, MW to Reference



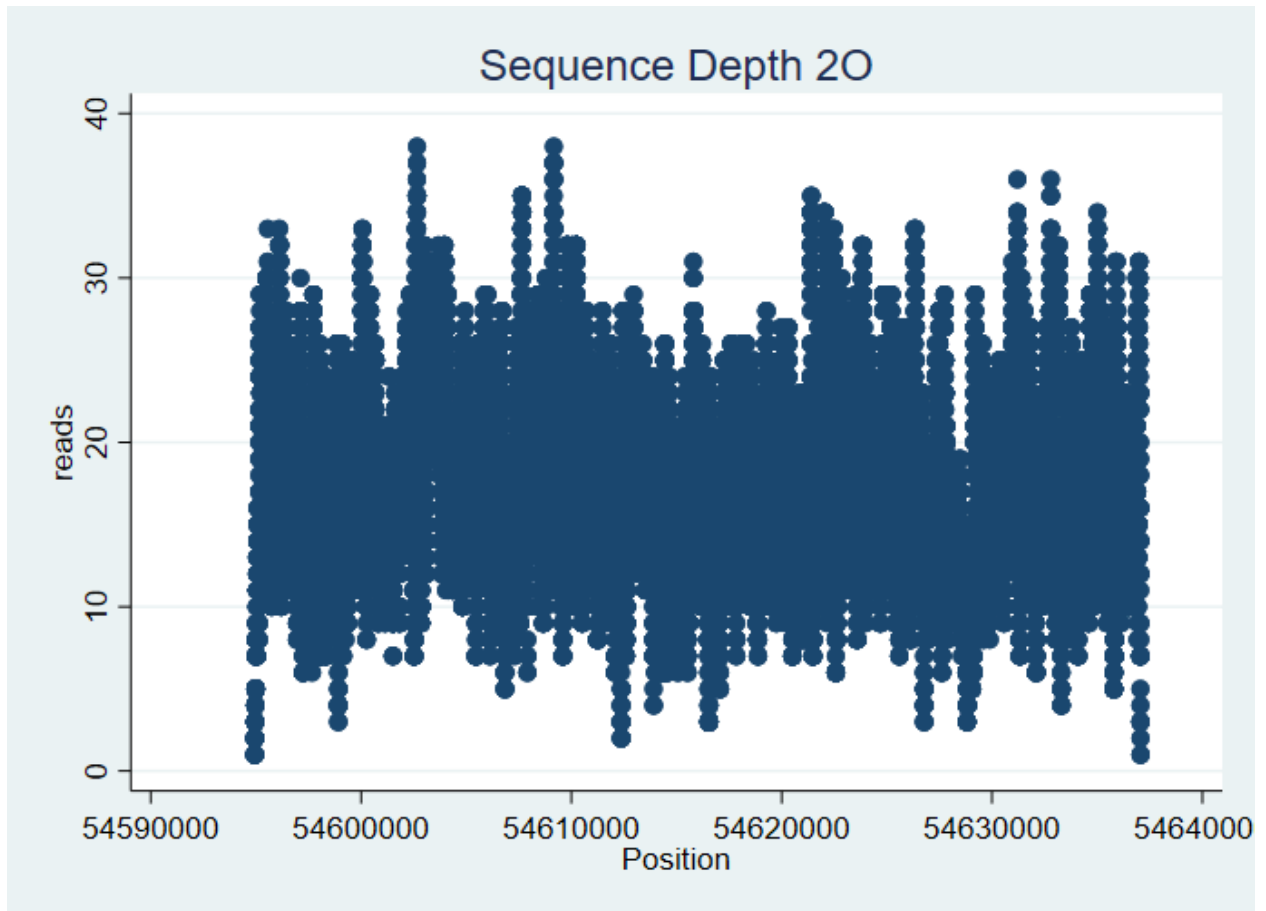
Figures 1a-1c: An example of read depth for Samples 1Y, 2O, and MW across the dog SMN gene. The x-axis refers to the position within the SMN gene and the y-axis refers to the number of reads covering the position. Read depth refers to the number of sequenced reads that aligned to the reference sequence for each region. The average read depth for sample 1Y was roughly 24-25 reads, for sample 2O was 17-18, and for sample MW was 9-10. The table below provides details about the total number of reads generated within the SMN gene for each sample, the average read depth across the gene, the standard deviation of read depth, and the maximum/minimum reads through the gene.

1a:



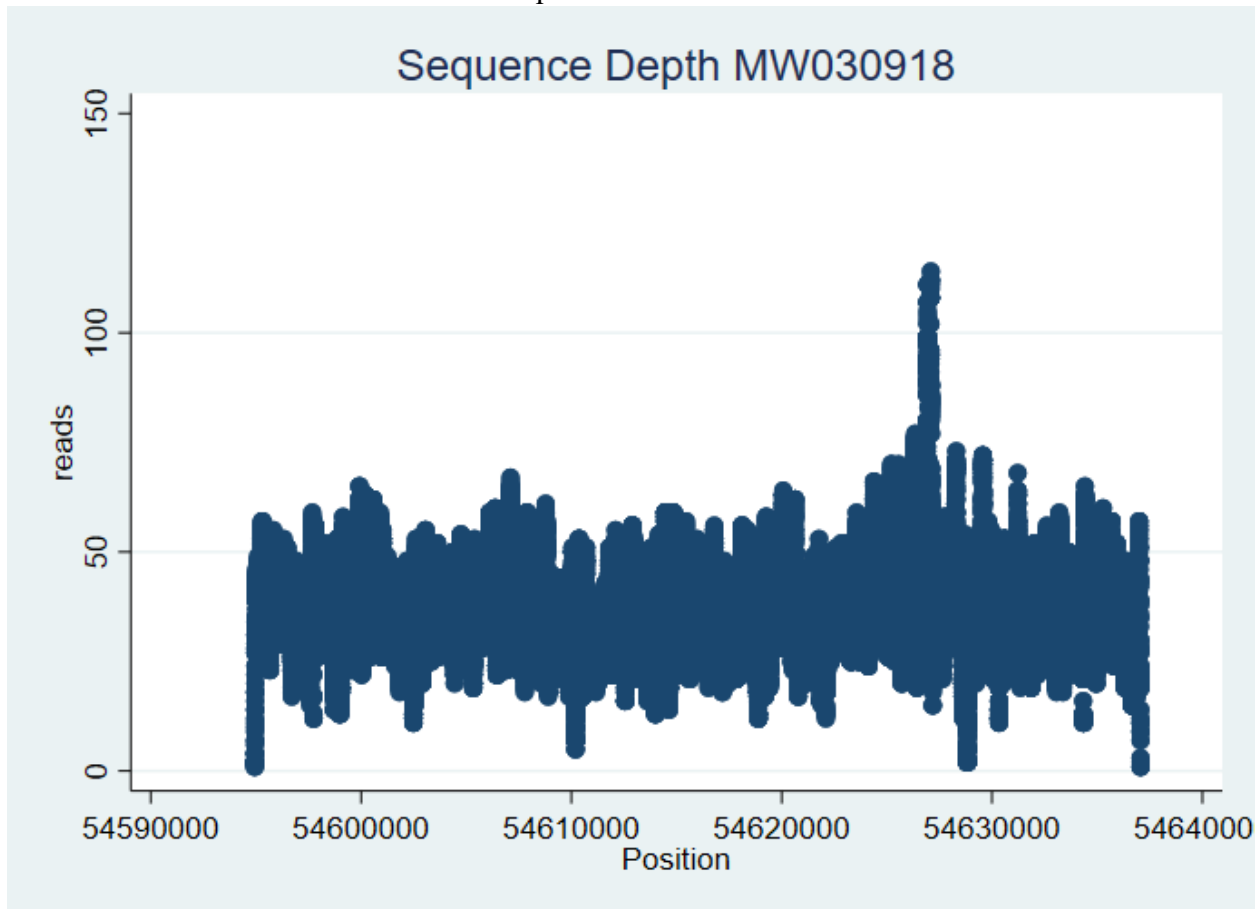
Variable	Obs	Mean	Std. Dev.	Min	Max
reads	168,456	24.60	6.76	1	55

1b:



Variable	Obs	Mean	Std. Dev.	Min	Max
reads	42,121	17.35	5.30	1	38

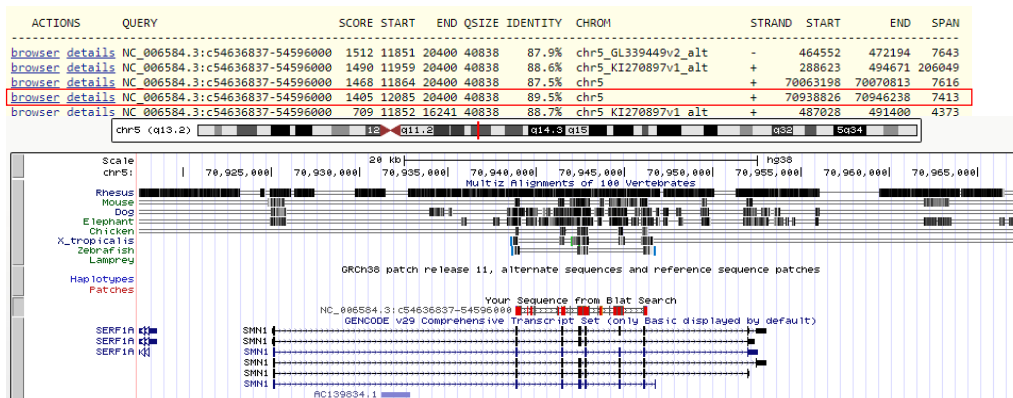
1c: Note that MW030918 and MW are equivalent.



Variable	Obs	Mean	Std. Dev.	Min	Max
reads	42,125	37.12	9.86	1	114

Figure 2: The reference dog SMN sequence was compared with the human SMN sequence. About 7,413 base pairs aligned with the human SMN sequence on Chr5. The percentage of similarity between the two is 89.5%.

## Comparison of dog SMN to human SMN





## References

1. A. Ostrander, E., Lindblad-Toh, K., & S. Lander, E. (2004). *Sequencing the Genome of the Domestic Dog Canis Familiaris*.
2. Deschenes, S. M., Puck, J. M., Dutra, A. S., Somberg, R. L., Felsburg, P. J., & Henthorn, P. S. (1994). Comparative mapping of canine and human proximal Xq and genetic analysis of canine X-linked severe combined immunodeficiency. *Genomics*, 23(1), 62-68. doi:10.1006/geno.1994.1459
3. Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, 107(1), 1-8. doi:10.1016/j.ygeno.2015.11.003
4. Institute, N. H. G. R. (2015). Comparative Genomics. Retrieved from <https://www.genome.gov/11509542/comparative-genomics-fact-sheet/>
5. Institute, N. H. G. R. (2018). The NHGRI Dog Genome Project. Retrieved from [https://research.nhgri.nih.gov/dog\\_genome/](https://research.nhgri.nih.gov/dog_genome/)
6. Kirkness, E. F., Bafna, V., Halpern, A. L., Levy, S., Remington, K., Rusch, D. B., . . . Venter, J. C. (2003). The Dog Genome: Survey Sequencing and Comparative Analysis. *301*(5641), 1898-1903. doi:10.1126/science.1086432 %J Science
7. Li, H. (2013). *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM* (Vol. 1303).
8. Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14), 1754-1760. doi:10.1093/bioinformatics/btp324
9. Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., . . . Law, M. (2012). Comparison of Next-Generation Sequencing Systems %J Journal of Biomedicine and Biotechnology. 2012, 11. doi:10.1155/2012/251364
10. Mellersh, C., Ostrander, E., Cork, L., & Blazej, R. (1998). Hereditary canine spinal muscular atrophy is phenotypically similar but molecularly distinct from human spinal muscular atrophy. *Journal of Heredity*, 89(6), 531-537. doi:10.1093/jhered/89.6.531 %J Journal of Heredity
11. Nizzardo, M., Simone, C., Dametti, S., Salani, S., Ulzi, G., Pagliarini, S., . . . Corti, S. (2015). Spinal muscular atrophy phenotype is ameliorated in human motor neurons by SMN increase via different novel RNA therapeutic approaches. *Sci Rep*, 5, 11746. doi:10.1038/srep11746
12. Parker, H. G., Shearin, A. L., & Ostrander, E. A. (2010). Man's best friend becomes biology's best in show: genome analyses in the domestic dog. *Annu Rev Genet*, 44, 309-336. doi:10.1146/annurev-genet-102808-115200
13. Shearin, A. L., & Ostrander, E. A. (2010). Leading the way: canine models of genomics and disease. *Dis Model Mech*, 3(1-2), 27-34. doi:10.1242/dmm.004358
14. van Dijk, E. L., Auger, H., Jaszczyszyn, Y., & Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends Genet*, 30(9), 418-426. doi:10.1016/j.tig.2014.07.001
15. Wurster, C. D., & Ludolph, A. C. (2018). Nusinersen for spinal muscular atrophy. *Therapeutic advances in neurological disorders*, 11, 1756285618754459-1756285618754459. doi:10.1177/1756285618754459