THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

DEPARTMENT OF BIOBEHAVIORAL HEALTH

Characterizing Intergenic Tandem Repeats and Their Implications on Adolescent Idiopathic Scoliosis

CARLY BROGAN
SPRING 2022

A thesis
submitted in partial fulfillment
of the requirements
for a baccalaureate degree
in Biobehavioral Health
with honors in Biobehavioral Health

Reviewed and approved* by the following:

David Vandenbergh
Professor of Biobehavioral Health
Thesis Supervisor

Helen Kamens
Associate Professor of Biobehavioral Health
Honors Adviser

* Electronic approvals are on file.

ABSTRACT

Adolescent Idiopathic Scoliosis (AIS) impacts 4 out of every 100 adolescents (Konieczny et al., 2013). The exact etiology of AIS is currently unknown with ongoing research exploring its genetic basis and environmental influences (Kikanloo et al., 2019). The influence of genetics on the development and severity of AIS can be evaluated in analyzing genetic variation in individuals, particularly single nucleotide polymorphisms and tandem repeats. The variability of a tandem repeat locus within the intergenic region of the *DRD4* and *DEAF1* genes was evaluated via bioinformatic analysis of sixty-four long read sequences and two short-read sequences from the UCSC Genome Browser. The repeat was found to be polymorphic with 34 copy number variants. Additionally, the relation between an allele for SNP rs11604855 and alleles at the tandem repeat were analyzed to assess the linkage of the two variants. The two loci were not found to be in linkage disequilibrium. In other words, the number of copies of the tandem repeat was not associated with the presence of the minor A allele. Despite a significant association between the allele rs11604855 and the copy number of tandem repeats being found of -0.452 ($p < 0.001$), due to the limited sample size, the causal role of the tandem repeat in the development of AIS cannot be inferred.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank Dr. Vandenbergh for his guidance and support in my research endeavors in the Molecular Genetics Lab and his guidance in the writing process of this thesis. I would also like to thank the research assistants in the Molecular Genetics Lab for their collaboration and support in the experimental process. Lastly, I would like to thank Andrew Burich for crafting code which allowed for the trimming of long-read sequences for the necessary analysis.

Chapter 1

Introduction

Adolescent Idiopathic Scoliosis

AIS is characterized by a three-dimensional deviation of the spinal axis (Trobisch, 2010). An individual with a ten-degree curvature or greater on a plain anteroposterior X-ray image meets the diagnostic criteria (Trobisch, 2010). The prevalence of idiopathic scoliosis is about 1-2% among adolescents, 8% in adults and is greater than 50% among persons over the age of 60 (Schwab et al., 2005). Idiopathic scoliosis does not typically cause pain in adolescents and children, though the asymmetry of the spine can cause them to have negative self-image (Freidel et al., 2002). Negative self-image can lead to tendencies of elevated alcohol consumption, depression, and suicidal thought (Freidel et al., 2002). Symptoms in adults can vary depending on the location of the curvature (Trobisch, 2010). For instance, scoliosis located in the lumbar spine can lead to debilitating back pain, whereas adults with thoracic scoliosis typically do not experience pain (Trobisch, 2010); however, thoracic scoliosis can impact pulmonary function when the curvature is severe. In most cases, a brace can be worn to correct for the curvature (Trobisch, 2010). In more severe cases of Idiopathic Scoliosis, defined by a curvature that exceeds 45 degrees, surgery may be necessary (Danielsson and Nachemson, 2001). The goals of treatment are to diminish symptoms experienced by the individuals and to prevent progression of the curvature or additional medical issues that may arise (Danielsson and Nachemson, 2001).

The etiology of scoliosis remains unclear and is predicted to be multifactorial (Kikanloo et al., 2019). Some of these factors such as genetic predisposition, toxin exposure, hormonal imbalances, and diet, have been found to overlap in their contribution to not only the onset of scoliosis, but also the severity of it (Kikanloo et al., 2019). There is presumably a genetic link due to a 70% concordance rate among monozygotic twins and a trend in the development of scoliosis in individuals with near relatives who have the condition (Trobisch, 2010). The heritability of scoliosis has been found to be 38%, thus approximately 62% of the susceptibility of disease is attributed to environmental factors (Grauers et al., 2012). The genetic aspect of the development of AIS can be studied using specific genetic variants that serve as markers for disease association.

Genetic Basis of Adolescent Idiopathic Scoliosis

Multiple genes have been studied to examine their association with AIS. A study found that a *MATN1* SNP at 1p35 was associated with a disorder of chondrocyte distribution and was associated with scoliosis (Chan et al., 2009). A separate *MATN1B* SNP on chromosome 11 was also found to be associated with AIS (Peng et al., 2012). Additionally, genetic variation of the *CHD7* gene has been found to be significantly associated with the development of AIS (Wu et al., 2021). It was discovered the decreased expression of the *CHD7* gene may be the involved in the development of scoliosis due to its involvement in determining individual's bone mass (Wu et al., 2021). Additional SNPs have been found to have possible implications in the development of spinal abnormalities. For example, rs11190870 in the 3' flanking region of the *LBX1* gene on chromosome 10 has been studied to better understand how such polymorphisms contribute to

bone development or formation and have implication in the development of abnormal spinal curvatures (Liang et al, 2014). Over time possible genetic loci on chromosome 6p, 10q, 18q, 19p13.3, 17p11, 19p13, 8q12, 9q31.2-q34.2, 17q25.3-qtel, 12p, and Xq have been explored (Liang et al, 2014). It is evident the development of Idiopathic Scoliosis has a multifactorial genetic basis that may involve both genetic and environment influences.

Genome Wide Association Study

The Human Genome Project in 2003 as well as the HapMap project in 2005 provided whole genome sequence data that could be utilized for extensive genetic research. Genome wide association studies (GWAS) are utilized to test genetic variants for their association with diseases (Witte, 2010). GWAS can identify certain variations in the genome by utilizing sequences in databases such as the UCSC Genome. Once the variants are identified and/or linked to a disease or trait, databases such as the Genome Browser mark them as variable and their possible association. Single nucleotide polymorphisms (SNPs) serve as a type of marker in the genome and can be utilized to test for association with a particular disease or disease characteristics. The investigation of SNPs and their associations can provide a method of assessing heritability, predicting onset of disease, or creating possible treatment methods (Witte, 2010).

Despite being important variants in the genome, SNPs are not the only form of genetic variation. It is important to consider the possibility identified SNPs are linked to other variants within the genome. GWAS may indicate a certain SNP is associated with a disease, though a

second variant's link to the SNP may be the actual causal factor for the phenotype (Gabriel et al., 2002).

Tandem Repeats

Tandem repeats are short segments of repetitive DNA that exist adjacent to each other and are dispersed throughout the genome (Hannan, 2018). Variable Number of Tandem Repeats (VNTRs) are loci with repeats that vary in copy number from individual to individual. If there is a varying copy number of the tandem repeat across individuals, the region is considered polymorphic and potentially can have a functional role in controlling gene expression, and thus phenotype (Hannan, 2018). The repeat itself can play a role in disease development, or possible pathways can exist in which the repeat interacts or is linked with other variants in the genome. The linkage of multiple variants is known as linkage disequilibrium and arises when two, or in some cases more than two, variants in the genome experience this non-random linkage such that the presence of a certain allele at both loci occurs together (Slatkin, 2008). Correlation analyses can be done to evaluate if the presence of a certain allelic variant is associated with the presence of another specific variant/s at a separate place in the genome (Slatkin, 2008). This analysis can help in the understanding of how the genetic variant an individual has is associated with a separate variant and ultimately contribute to their disease. This type of analysis was conducted in this paper to understand the linkage of the rs11604850 SNP and a tandem repeat in the intergenic region of the *DEAF1* and *DRD4* genes.

DEAF1 and *DRD4* Genes Possible Implications

The genes Dopamine Receptor D4 gene (*DRD4)* and Deformed Epidermal Autoregulatory Factor 1 (*DEAF1)* surround the tandem repeat and SNP which were studied in this thesis. These genes have been analyzed which provides context of the SNP rs11604850's regulatory potential on surrounding areas of the genome. *Retinoic Acid Induced 1* (*RAI1*) was found to be regulated by one of the genes flanking the SNP rs11604855 (Chen et al., 2020). Specifically, *DEAF1* binds to *RAI1* and regulates its expression (Chen et al., 2020). This finding is relevant in that RAI1 has been associated with skeletal abnormalities, including scoliosis (Chen et al., 2020). Additionally, a study examined *DRD4* and its impact of environments on behavior and health (Grady et al. 2013). Specifically, Grady et al. analyzed physical activity levels in human populations of varying age groups (Grady et al. 2013). Physical activity plays a role in bone modeling so there may be a connection to scoliosis (Sharkey and Lang, 2007). Genes that are associated with behaviors related to physical activity may impact skeletal integrity by influencing the frequency and magnitude of loading experienced by bones (Sharkey and Lang, 2007). The loading can impact the skeletal phenotype experienced by the modeling and remodeling of the bone in adapting to their mechanical environment (Sharkey and Lang, 2007). Additionally, it was found that in an older population carrying the *DRD4* 7R allele, as opposed to the 4R, correlated with increased levels of physical activity (Grady et al., 2013). Understanding the *DRD4* and *DEAF1* genes, which flank the SNP and tandem repeat of interest, allows for a broader context of how the SNP plays a possible role in regulating gene expression and thus contributing to AIS.

Aim of Research

With lack of evidence of a causal association of SNP rs11604855 and AIS, it was hypothesized that a nearby tandem repeat locus might be variable, and that alleles at this site are in linkage disequilibrium with rs11604855, thus making the tandem repeat a candidate for a causal role in AIS. The goal of this thesis was to characterize tandem repeats in this intergenic region and explore their association with the indicated SNP. In understanding the variability of the tandem repeats in terms of their length and copies numbers, the implications of an individual having a certain allele of the SNP and the tandem repeat can be better understood.

Chapter 2

Methods

Locating Repeat of Interest

The UCSC Genome Browser is an electronic database which contains whole genome assemblies. The Human GRCh38/hg38 sequence within a Genome Browser hosted by the University of California, Santa Cruz (UCSC) was utilized for analysis of the genome. The browser displays the entirety of the genome in providing information regarding genes and a variety of functional elements such as tandem repeats, promoters, and SNPs (Kent et al., 2002). Genomic variation can be seen via tracks, which will provide a visual aid of the location of known variable elements and their position on a chromosome or proximity to certain genes.

The Molecular Genetics Laboratory at Penn State University has a specific focus on evaluating the molecular basis of addiction, and in particular studying tandem repeats and their implications on gene expression. The lab's focus on the dopamine transporter gene, *SLC6A3,* prompted the investigation for tandem repeats within other genes with known association with addiction such as dopamine receptor genes *DRD3, DRD4, DRD5,* and others. The Simple Tandem Repeat by TRF track was used in the Genome Browser to identify variants surrounding the dopamine related genes and notable repeats were considered for analysis.

Investigating Variability within the Tandem Repeat of Interest

The USCS Genome Browser contains two main sequences, the reference and alternate sequences which are publicly available for analysis and are marked with genetic variants and other descriptive information. The reference sequences are limited to parts of the genome with large amount of variation, so that the developers felt an alternative sequence was needed to highlight the variability. The targeted area for study is one such site. The availability of both sequences allowed for a preliminary analysis of the tandem repeat's variability. The intent was to compare the characteristics of the repeats in both the reference and alternate sequence to determine possible variability in number of copies of the repeat. The repeat's sequence, length, and copy number were used for this elementary check for variability.

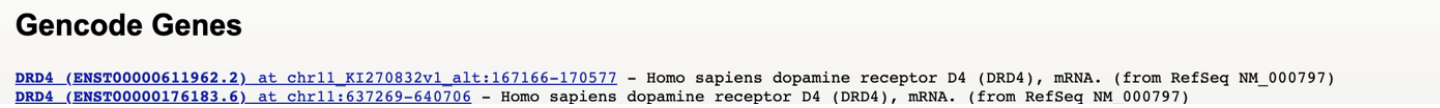

Figure 1: The output from the UCSC Genome Browser when the *DRD4* gene is searched for. The results indicate there is a reference and an alternate sequence available for use to examine the *DRD4* gene. These two sequences indicated were utilized to examine the repeat's possible variability.

Investigating Variability of Repeat of Interest with Bioinformatics

The bioinformatic analysis was intended to analyze variability of the tandem repeat in a larger number of individuals whose sequences were generated through long-read sequencing technology. Sixty-four long read sequences of unrelated individuals were obtained from the Human Genome Structural Variant Consortium for this analysis (Course et al., 2021). The long-read sequences were trimmed using code constructed by Andrew Burich, from the department of Data Science, and was run on the software BioPython. Long-read sequences were uploaded into

the code, along with barcodes containing the start and stop points of the desired shortened sequence. The barcodes used specifically for the trimming of the tandem repeat in the intergenic region of the *DRD4* and *DEAF1* genes are as follows:

3’ flank: CCTGCTGGGTTGGATGAAAAGCCATTGTAGA

5’ flank: CAGCATCTGGAGCGCGTCCTGCAGCAGCCC

Assessing Linkage of Variable Elements

Variations in the genome can be associated with each other in that the presence of one variant may be associated with the presence of a certain variant at a different location on the chromosome (Slatkin, 2008). Therefore, the number of units of the repeat, or copies, of the repeat was evaluated to analyze the association of the copy number with a certain allele at a nearby SNP. The tandem repeats within the long-read sequences were analyzed for their copy number using the Tandem Repeat Finder software. Tandem Repeat Finder provides data regarding the copy number of tandem repeats in each sequence. The number of units of the repeat in each sequence was used for analysis of the sequences. In addition to characterizing tandem repeats within the long-read trimmed sequences, each sequence was evaluated for the identity of the allele at rs11604855 (A or G). Once these data were collected, a correlation analysis was performed to assess the potential link between the SNP allele and the number of repeat units within the sixty-four sequences as well as the reference and alternate sequences. IBM SPSS Statistics Software was utilized for the descriptive analyses of the tandem repeat copy number and allelic frequencies of the rs11604855 SNP. Additionally, SPSS was used for the correlation analysis of the two variants.

PCR Analysis of Tandem Repeat

Polymerase Chain Reaction (PCR) was used to amplify the DNA sequence of interest. The amplified DNA was then visualized using gel electrophoresis. Based on the length the DNA traveled on the gel, it could be determined if the process was successful in amplifying the intended region.

The DNA used for the PCR experiments were samples collected from previous research assistants in the Molecular Genetic lab at Pennsylvania State University. The samples were collected under the IRB STUDY00008832. Epithelial cells were obtained from the group of participants via buccal swabs and their DNA was extracted with the use of a lysis buffer (including the detergent, SDS, and Protease K and RNAase A), followed by extraction with a solvent mix of phenol and chloroform. The DNA was then ethanol precipitated, redissolved in water and stored for analysis.

The PCR reaction requires primers to signal the region of DNA needed for amplification. Primers for the experiments were constructed using the software PrimerBlast which identifies pieces of DNA within the region of interest that are stable and would be optimal to be used for amplification (https://www.ncbi.nlm.nih.gov/tools/primer-blast/). The primers designed to amplify the tandem repeat were:

Left Primer:

- hDRD4642682F
- 5’ TGCTGGGTTGGATGAAAAGC 3’

Right Primer:

- hDRD4643498R
- 5’ TGATGTTTGCCTCTGAAGCG 3’

Various PCR reactions were performed using these primers and reaction conditions were altered in order to optimize the results. The modifications included trying various polymerases, enzymes, and annealing temperature to improve the amplification process.

The PCR reaction consisting of buffer, dNTPs, primer, polymerase, DNA and water in the following quantities:

Table 1: PCR reaction conditions for experiments attempting to amplify 28-BP repeat.

Reagents	1 reaction	5 reactions
Water	5.25 µL	26.25 µL
5X Buffer	2.5 µL	12.5 µL
2 mM dNTP	2.5 µL	12.5 µL
2 mM F Primer	6.25 µL	31.25 µL
2 mM R Primer	6.25 µL	31.25 µL
Q5 Enzyme	0.25 µL	1.25 µL
DNA	2.0 µL	N/A

The reaction mix was created so that five samples of DNA could be amplified if needed. 23 µL of the reaction mix was added to four reaction tubes and 2 µL of DNA was added to each of the tube. In the fourth PCR tube, a control of 2 µL of water was added for later reference on the polyacrylamide gel. The samples were then placed in the GeneAmp PCR System 9700 and left to run for 35 cycles of denaturation, annealing, and extension. The DNA was denatured at 95.0°C for three minutes followed 92°C for thirty seconds. The DNA was then annealed at 63°C

for thirty seconds. Extension took place at 72°C for thirty seconds. These steps were repeated for thirty-five cycles, followed by one cycle of 72°C for ten minutes.

Once the PCR was completed, the samples were visualized via polyacrylamide gel electrophoresis. 10 μL of the five samples were combined with a blue dye to allow for the sample's migration patten in the gel to be see under UV light. In addition to the four samples, a DNA ladder was entered into a well of the gel to serve as a reference of how far the fragments traveled while running. The gel was run for 45 minutes under 200 V current and 50 mA resistance. Upon completion, the gel was stained with ethidium bromide and viewed under a UV light. The DNA fragment's lengths were assessed based on their migration's alignment with the DNA ladder. The larger the DNA fragment, the slower it travels through the gel therefore, the ladder's measurement starts from the bottom with each mark indicating 100 base pairs.

Chapter 3

Results

Locating Repeat of Interest

The UCSC Genome Browser was used to identify genetic variants within dopamine receptor genes. The search was narrowed to the dopamine receptor D4 gene (*DRD4)* which is a heavily studied gene with known association with substance abuse and psychiatric disorders such as attention deficit disorder (Ptacek et al., 2011). The gene contains three tandem repeats with a specific 48-base pair Variable Number Tandem Repeat (VNTR) in exon 3 (Figure 2). The 7R allele of this repeat is known to influence the second messenger system that responds to *DRD4*, through which the polymorphism influences dopamine signaling (Asghari et al., 1995). Alleles at this site have differential effects in the ability to inhibit cyclic AMP formation which provides evidence of the functional consequences of the alleles in altering cell physiology (Asghari et al., 1995). This change is cAMP levels can have implications on phosphorylation states and the number of proteins which can influence gene expression of other genes (Meyer and Miller, 1974).

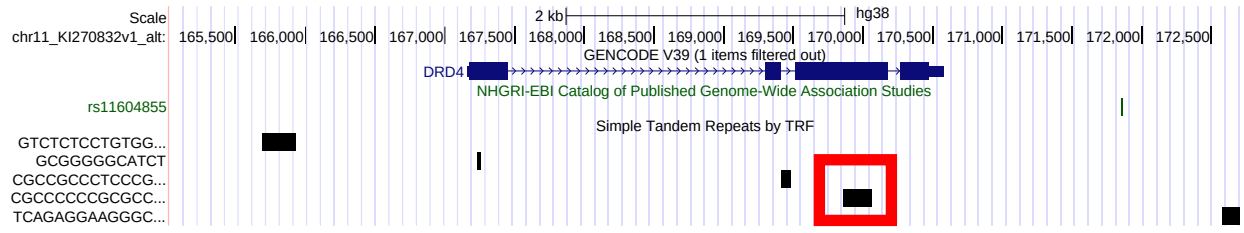

Figure 2: Graphic from the UCSC Genome Browser showing the DRD4 gene. Its three tandem repeats are displayed under the Simple Tandem Repeats by TRF track as black boxes beneath the graphic of the gene. The well-known 48-bp VNTR in exon 3 of DRD4 is boxed in red.

During the analysis of the genetic variability in the *DRD4* gene, a tandem repeat approximately 2000 base pairs from the gene was identified (Figure 3). The repeat is in the intergenic region between *DRD4* and *DEAF1*, with a unit length of 28 base pairs long. The repeat is located on chromosome 11 (p15.5), which is near the telomere of the chromosome. One unit of the repeat is AGGAAGGGCCAGGCGGGCTGAGGGTCAC and 4.9 copies existed in the reference sequence of the Genome Browser, thus the genomic size is 136 base pairs total.

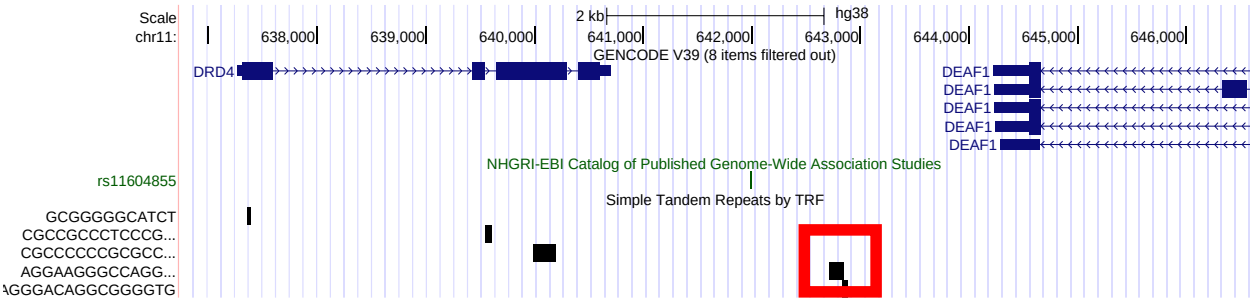

Figure 3: Genome Browser graphic with indication of the intergenic 28-bp repeat between the DRD4 and DEAF1 genes. The repeats approximately 2 kb distance from the DRD4 gene can be seen using the measure at the top of the figure. A small additional repeat is seen overlapping the repeat of interest, though it was not included in the analysis.

First Analysis of Repeat of Interest

The UCSC Genome Browser provides two sequences, both the reference and alternate, for evaluation and analysis. The repeat in both sequences was found and the results were compared based on the descriptive information provided (Figure 4). The repeat in both the reference and alternate sequence was 28 base pairs long, though a difference in the number of copies was found between the sequences. In contrast to the reference sequence, with 4.9 copies, the alternate sequence contained 75 copies of the tandem repeat, making it 2099 bases long.

Figure 5 provides a graphical comparison of the length of the two VNTRs. Note that a second short tandem repeat is located immediately to the right of the 28 base VNTR. The short repeat has a unit length of 16 bases and has 3.2 copies in the reference sequence of the Genome Browser and 6.5 copes in the alternate sequence, indicating this short repeat is also variable. The smaller repeat was also analyzed in the Tandem Repeat Finder software and nine alleles were recognized. The copy numbers included 2.2, 2.5, 4.2, 5.5, 6.2, 6.5, 6.7, 7.5, and 25.5 for this short repeat. Despite its evident variability, further analysis of this repeat was not performed due to a focus on the more variable 28 base pair tandem repeat.

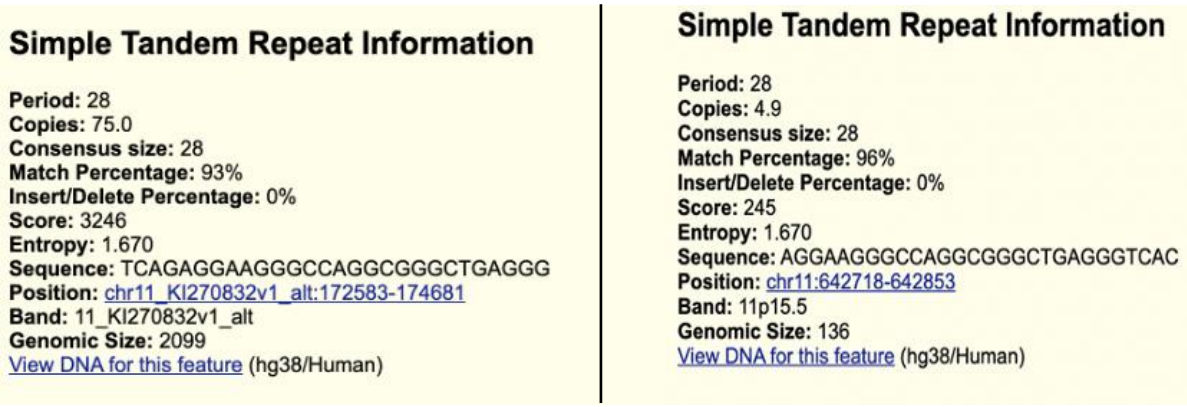

Figure 4: Tandem repeat information for the reference sequence (right) and the alternate sequence (left). Information is provided about the 28-bp repeat in each of the sequences including the difference in copy number and overall genomic size.

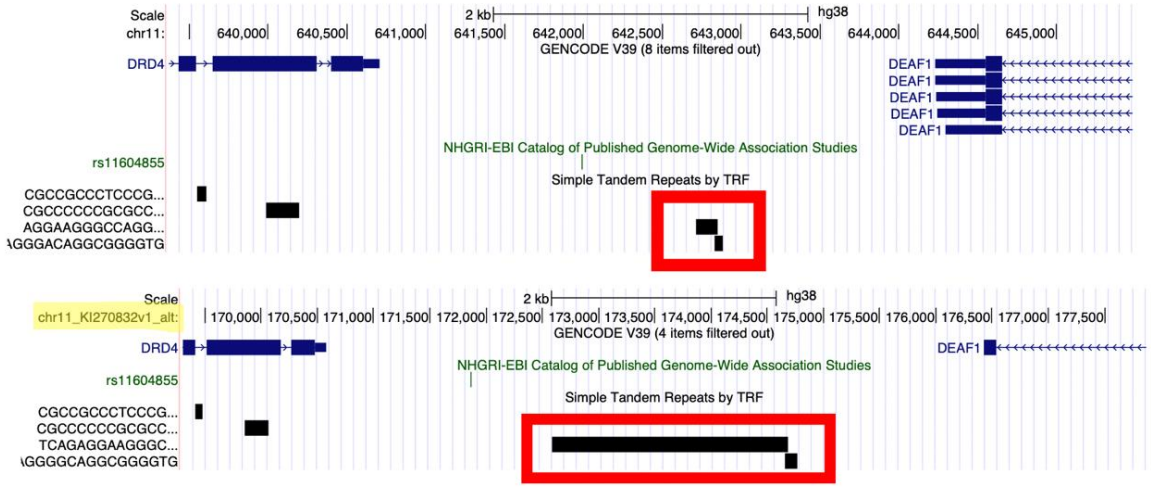

Figure 5: Depictions of the 28-bp tandem repeat of interest in both the reference and alternate sequences. The reference sequence, shown on the top, is visibly shorter appearing approximately 100 bp long. The bottom sequence is the alternate, which is notably longer, appearing to be approximately 2000 bp in length.

Locating rs11604855 SNP

The location of the tandem repeat in the intergenic region of the *DRD4* and *DEAF1* genes prompted exploration of additional variable elements in the surrounding genome. The UCSC Genome Browser contains a track that visualizes the single nucleotide polymorphisms (SNPs) that have been identified by Genome Wide Association Study (GWASs). The track revealed rs11604855, which is an A/G SNP that is less than 1 kB from the tandem repeat of interest (Figure 6 and 7). The SNP was associated with AIS in one GWAS (Liu et al., 2017). Due to the finding of this SNP and its proximity to the variable tandem repeat, association tests were run to assess for relationship between the presence of a certain allele of the SNP and the number of tandem repeats present in individuals (described below).

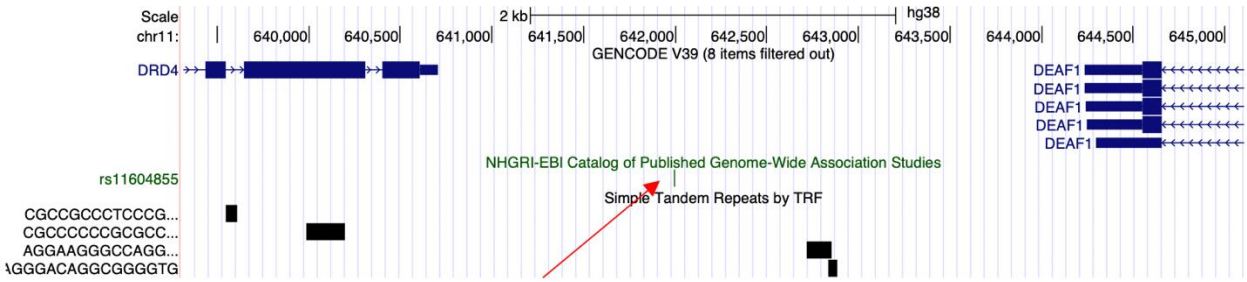

Figure 6: Visual of rs11604855 within the reference sequence of the Genome Browser. The arrow is pointing to the specific nucleotide within this sequence, which is a Guanine base.

NHGRI-EBI Catalog of Published Genome-Wide Association Studies (rs11604855)

dbSNP: rs11604855
Position: chr11:641991-641991
Band: 11p15.5
Genomic Size: 1
View DNA for this feature (hg38/Human)
Reported region: 11p15.5
Publication: Liu J *et al.* The coexistence of copy number variations (CNVs) and single nucleotide polymorphisms (SNPs) at a locus can result in distorted calculations of the significance in associating SNPs to disease. *Hum Genet.* 2018-07-17
Disease or trait: Adolescent idiopathic scoliosis
Initial sample size: 196 cases, 303 controls
Replication sample size: NA
Reported gene(s): *Not reported*
Strongest SNP-Risk allele: rs11604855-?
dbSNP build 144 observed alleles for rs11604855: A/G
Risk Allele Frequency: *Not reported*
p-Value: 9E-10
Odds Ratio or beta: *Not reported*
95% confidence interval: *Not reported*
Platform: Affymetrix [*Not reported*]
Copy Number Variant (CNV)?: No

View table schema

Go to GWAS Catalog track controls

Data last updated at UCSC: 2022-03-02

Figure 7: Visual of SNP information provided by the Genome Browser. It is indicated the SNP is located on chromosome 11 position 15.5. The SNP is an A/G SNP with possible association with Adolescent Idiopathic Scoliosis.

The Genome Browser also provided information regarding the allelic frequency of the SNP. ALFA is NCBI's Allele Frequency Aggregator which provides allele frequency for a population of more than a million individuals. The alleles are reported based on data gathered from studies that have sequenced individual's genomes and are now stored in the NCBI database

of genotypes and phenotypes (dbGaP https://www.ncbi.nlm.nih.gov/gap/). The allele frequencies based on place of origin are reported in the table figure to display the proportion of the population with the A versus G allele. The major allele is the G allele in the total population of 23,860 individuals sampled by NCBI and has a frequency of 96.1%. The minor, or alternate allele, the A allele, has a total frequency of 3.9% across the sample of individuals with various backgrounds (Figure 8).

ALFA Allele Frequency

The ALFA project provide aggregate allele frequency from dbGaP. More information is available on the project page including descriptions, data access, and terms of use.

Release Version: 20201027095038

Search:

Population	Group	Sample Size	Ref Allele	Alt Allele
Total	Global	23860	G=0.96127	A=0.03873
European	Sub	16076	G=0.97406	A=0.02594
African	Sub	3402	G=0.9612	A=0.0388
African Others	Sub	114	G=0.974	A=0.026
African American	Sub	3288	G=0.9608	A=0.0392
Asian	Sub	150	G=0.800	A=0.200
East Asian	Sub	122	G=0.795	A=0.205
Other Asian	Sub	28	G=0.82	A=0.18
Latin American 1	Sub	230	G=0.943	A=0.057
Latin American 2	Sub	2628	G=0.8980	A=0.1020
South Asian	Sub	104	G=0.933	A=0.067

Figure 8: Frequency of the A/G alleles within rs11604955 in populations of various origins. The top row indicates the total allelic frequency taken for the entirety of the population and the other areas of origin and their frequencies of alleles are stated below. Asian and South East Asian populations have the highest percentages of the minority allele

Bioinformatic Findings from Long Read Sequence Data

Tandem Repeat Analysis

The sixty-four long-read sequences were inputted into the python code for trimming and the barcode sequences used to signify the start and end of the intended shortened sequences were found in all the files. The trimmed sequence encapsulated the intergenic tandem repeat with about 30 base pairs of flanking sequence on each end. The shortest trimmed sequence was 144 bases long and the longest was 4,428 bases long. There were two consensus sequences of 28-base pairs provided for the mapping and alignment of the repeats within the trimmed sequences. The consensus sequence of AGGAAGGGCCAGGCGGGCTGAGGGTCAC was the most common and is also seen in the reference sequence in the genome browser. The other consensus sequence was TCAGAGGAAGGGCCAGGCGGGCTGAGGG which is also 28 bases long, though the repeat unit differs slightly from the other consensus sequence in starting with four different bases. After the extra bases (TCAG), found at the beginning of the sequence, the second consensus sequence aligns with the majority consensus sequence. It is of note the identification of two consensus sequences is not a likely occurrence and it is unknown why the program recognizes these units distinctly to use for comparison.

The Tandem Repeat Finder software allowed for input of all sixty-four sequences for analysis of the number of units of the 28-base pair repeat. Table 2 shows that the greatest number of tandem repeat units identified in the sequence was 156, indicating the longest repeat length was 4,368 base pairs total. The smallest copy number was just 3 units, with a total repeat length of 84 base pairs total. The average copy number in the sixty-four sequences and the alternate and

reference sequences from the genome browser was found to be 74.030 with a standard deviation of 33.3621 copies.

Table 2: Information regarding the copy numbers in each of the sixty-four trimmed sequences and the reference and alternate sequences from the Genome Browser. The copy number of the 28-base pair repeat was provided by the software Tandem Repeat Finder. The analysis of the copy numbers was done in SPSS and the frequencies of the copy number variants in each sequence can be found in Appendix A.

Descriptive Statistics of Copy Number of 28-BP Repeat from Trimmed Long Read Sequence Data and Reference and Alternate Sequence from Short-Read Data from Genome Browser

	N	Minimum	Maximum	Mean	Std. Deviation
Repeat Copy Number	66	3.0	156.0	74.030	33.3621
Valid N (listwise)	66				

Rs11604855 Analysis

Each of the sixty-four sequences as well as the reference and alternate sequences from the Genome Browser were analyzed for the presence of either the A or G allele of rs11604855. As shown in Table 3, the G allele was found to be the major allele in that 57 of the 66 total sequences contained the G allele for the SNP. The minor A allele was found in the other nine sequences, with the overall frequency of the A allele being 13.6% of the sample.

Table 3: rs11604855 allele frequency in sixty-four long read sequences and the reference and alternate sequences from the UCSC Genome Browser. SPSS Statistics software was used to generate this table.

SNP Allele (A/G)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	A	9	13.6	13.6	13.6
	G	57	86.4	86.4	100.0
	Total	66	100.0	100.0	

Tandem Repeat and SNP Linkage Analysis

The sequences once analyzed for the A or G allele of the SNP and were then grouped based on the allele they contained, allowing for a comparison of the allele at the SNP locus and the copy number of tandem repeats 1kB away. The 57 sequences that contained the G allele had copy numbers with a large range as seen in Figure 9. The nine sequences that contained the A allele for rs11604855 had a range of copy numbers for the tandem repeat of interest, though with less variation. Six of the nine sequences with the A allele contained 15.9 copies of the repeat with the other 3 sequences containing 21.9, 91.9 and 118 copies of the repeat (Figure 10).

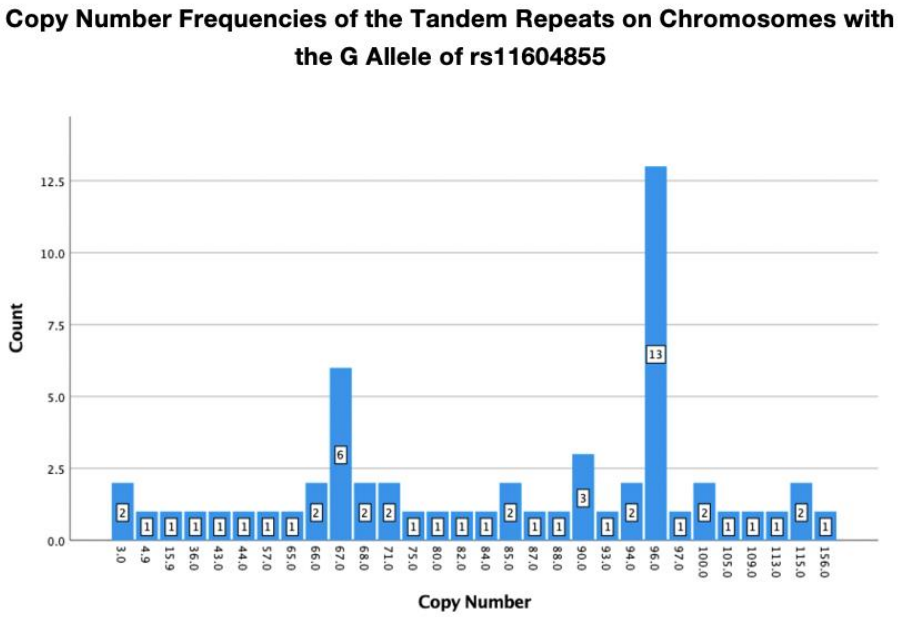

Figure 9: Analysis of sequences containing G allele of rs11604855. The graph shows the copy numbers of the tandem repeat of interest within the sequences with the A allele.

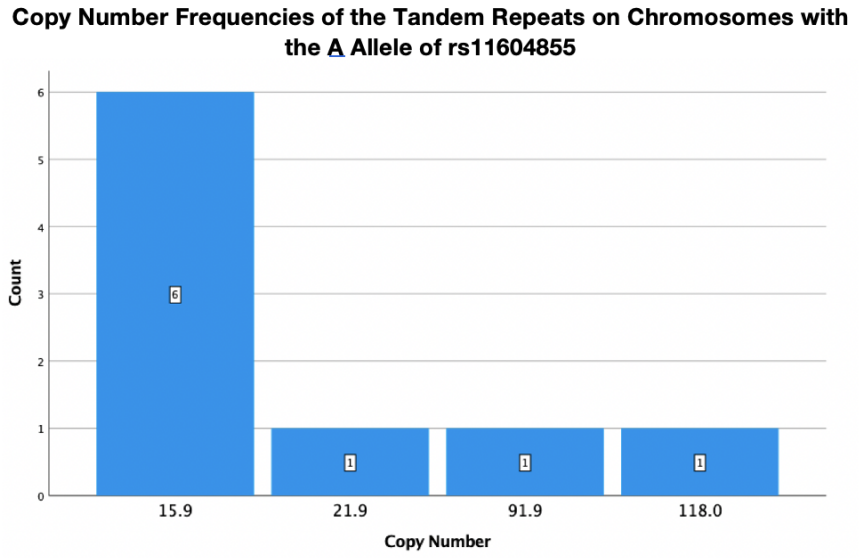

Figure 10: Analysis of sequences containing A allele of rs11604855. The graph shows the copy numbers of the tandem repeat of interest within the sequences with the A allele.

Correlation Test for SNP Allele and Tandem Repeat

Utilizing the SPSS Statistics Software, the A/G SNP and the tandem repeat underwent a correlation analysis to understand their possible chromosomal association. The G allele was coded as 1 and the A allele was coded as 2 to allow for the correlation analysis. The findings result in an r value of -0.452 which was found to be a significant association (p-value < 0.001). The negative r value provides evidence of an inverse correlation of the copy number of repeats and the presence of the A versus G allele on the chromosome. Due to the G allele being coded as 1, the lower value, the G allele was found to be associated with the presence of a higher number of tandem repeat copies.

Correlation of Repeat Copy Number and SNP Allele

		Repeat Copy Number	SNP Allele (A/G)
Repeat Copy Number	Pearson Correlation	1	-.452**
	Sig. (2-tailed)		<.001
	N	66	66
SNP Allele (A/G)	Pearson Correlation	-.452**	1
	Sig. (2-tailed)	<.001	
	N	66	66

**. Correlation is significant at the 0.01 level (2-tailed).

Figure 11: Correlation test of the A/G SNP and the copy number of tandem repeats (r = -0.452, p < 0.01).

DNA Amplification Gel Images

More than ten trials in attempting to amplify the tandem repeat were unsuccessful. The first attempt at improving the reaction was including a GC enhancer, which is a proprietary solution from New England BioLabs that aids in amplifying regions high in GC content. Due to the repeat and its flanking region having a 71% GC content the addition of the enhancer was used to aid in the amplification, but no product was formed.

The tandem repeat was seemingly successfully amplified utilizing enzyme called Q5 High-Fidelity DNA Polymerase from New England BioLabs in the PCR reaction mix. The PCR reaction utilized DNA samples #5, #7 and #8. The expected product length utilizing the primers made by PrimerBlast was 485 base pairs. In Figure 11 there is some indication the three individuals are heterozygotes as seen in the presence of two bands between the 400 and 500 bp marker bands. Despite this possible implication, the bands seen on the gel image are likely background bands. It is highly unlikely all three individuals have identical copies of the tandem

repeat, especially with the knowledge of the high amount of variability of the copy numbers possible for this repeat shown in the bioinformatic analysis.

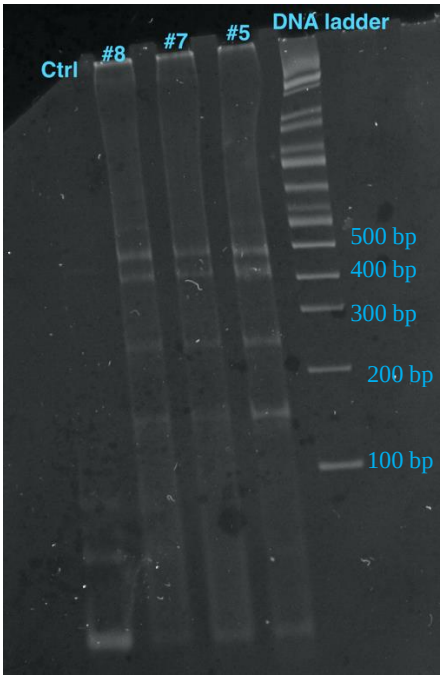

Figure 12: Gel image of the amplification of the intergenic tandem repeat of interest. There are distinct bands at the 450-500 bp mark on the gel, though due to the consistency of bands in all three samples, it is highly unlikely this is an accurate amplification of the region of interest.

Chapter 4

Discussion

Tandem repeats are important genetic variants to study due to their ability to alter gene expression and more specifically for their link to the development of certain diseases (Duitama et al., 2014). Various tandem repeats have been identified for their link with neuropsychiatric disorders such as substance abuse disorder and attention deficit disorder (Duitama et al., 2014). A tandem repeat within exon 3 of the *DRD4* gene has been identified for its probable contribution to the neurological underpinnings of substance use disorder by means of regulating dopamine signaling (Ptacek et al., 2011). It is possible the intergenic tandem repeat about 2 kB from the *DRD4* gene may play a role in the expression of the *DRD4* gene as well or may influence the expression of the other adjacent gene, *DEAF1*. The analysis in this research focused specifically on the potential of the repeat to be in linkage disequilibrium with rs11604855 SNP due to its proximity. The linkage would indicate a potential genetic association with AIS and provide impetus for tests of possible causal role in the development of AIS.

The use of the alternate and reference sequences on the UCSC Genome Browser was an initial method in testing for variability of the tandem repeat of interest. It was helpful in revealing the number of copies of the repeat was different in each of the two sequences, thus providing evidence of the possibility of polymorphism at this locus. Despite this preliminary analysis, the use of just two sequences is not sufficient in identifying large numbers of units of a long tandem repeat utilizing short read sequence technology with which the Genome Browser sequences was generated. Short-read sequencing has shown to be inferior due to limitations in the ability to identify the structural variants and repetitive regions of DNA sequence (Mantere et

al., 2019). The use of long-read based sequence data in this experiment allowed for a more precise and extensive analysis of copy numbers of the tandem repeats.

The trimmed long-read sequences provided a means of assessing variability on a larger scale (Logsdon et al., 2020). Data generated by the Tandem Repeat Finder software provided clear evidence of variability in the tandem repeat across individuals with 34 different copy numbers of the tandem repeats being present in the sixty-four sequences. Genetic variability within human DNA is a significant tool in understanding individual's susceptibility to disease or traits (Duitama et al., 2014). The variability of this tandem repeat is notable and has potential implications on gene expression be it direct or indirect in which it can possibly be in linkage disequilibrium with another structural variant within the genome (Duitama et al., 2014).

The PCR technique is intended to amplify and then visualize the DNA sequence of interest to understand its variability in length. The PCR experiments intending to amplify the intergenic repeat were unsuccessful. A gel utilizing the Q5 enzyme was successful in producing bands around the predicted length of the sequence at 485 bases, but the identical patterns of bands in the three individuals suggests the bands visible were background amplification products. If the PCR had been successful, it would have provided a more rapid and less expensive tool to identify variability of the repeat in a large number of samples.

The SNP analysis utilizing the Genome Browser revealed the A allele of rs11604855 was the minor allele in it only being present in about 4% of the sample population of studies presented in NCBI's databases. The analysis performed on the 66 sequences for the bioinformatic analysis revealed about a 13.6% frequency of the A allele, which is notably larger than the NCBI's allelic frequency report. This inflated allele frequency may be due to the small

sample size of 70 sequences compared to the NCBI's 20,000 sample size. The allele frequency is not entirely unreasonable in that the allele frequency varies quiet drastically depending on individuals' backgrounds. For example, as seen in Figure 8, in Asian and East Asian populations, the A allele is seen in about 20% of individuals. Overall, the A allele was identified as the minor allele in the bioinformatic analysis which aligns with the NCBI database of genotypes and phenotypes, despite the frequencies not being precisely the same.

The ultimate goal of this research was to characterize the tandem repeat's indirect association to the development of AIS through its association with the SNP rs11604855. The SNP analysis was performed to define a possible chromosomal association amongst the tandem repeat copy number with the alleles at rs11604855. The correlation analysis revealed a significant association (-0.452) between the SNP in the sequence with the copy number of the repeat. The negative association of -0.452 indicates a larger number of copies of the tandem repeat is correlated with the G allele of rs11604855. The analysis of the tandem repeat allele and the SNP allele revealed an association of the presence of 15.9 copies of the tandem repeat and the A allele. Due to 66.7% of the chromosomes containing the A allele having 15.9 copies of the repeat, it is possible this association is driving the significant correlation seen. Additionally, a few factors need to be considered in understanding this correlation. First, the sample size used in this experiment is 66 sequences. With this rather small sample, the allele frequency of the minor A allele of the SNP is inflated compared to a more representative sample of the general population. Additionally, there is a wide range of the copy number of repeats (3-156), which prevents a simple relation of one particular copy number being correlated with the A versus G allele for rs11604855. Due to the limited sample size and the high degree of heterozygotes, it

cannot be distinctly stated that the number of units of the tandem repeat that an individual has is associated with an allele at the SNP.

It is important to recognize the variation in copy number in the A versus the G allele. The sequences containing the G allele had a wide array of different copy numbers whereas the A allele only had four different copy numbers. Six out of the nine sequences with the A allele contained 15.9 copies of the tandem repeat and this particular allele of the repeat may be of importance and the association should be looked at further with a larger sample size in which the A allele is present. An alternative approach that is out of the scope of this research would be to genotype the tandem repeat in a large group of individuals with AIS and a control group using long-read sequencing, a methodology in the early stages of use (Stevanovski et al., 2022).

Overall, the hypothesis of the intergenic tandem repeat between the *DRD4* and *DEAF1* genes being variable was proven. Bioinformatic analysis of 66 sequences revealed the repeat is polymorphic and there is a large amount of variation at this locus. In contrast, the predicted linkage disequilibrium of a repeat allele with rs11604855 could not be confirmed. The correlation might be skewed by the small sample size. Due to the lack of linkage disequilibrium, it cannot be inferred the tandem repeat is associated with AIS and thus is not yet a candidate for a causal role in AIS.

The research performed is relevant on a larger scale in characterizing the intergenic repeat between the *DRD4* and *DEAF1* genes. No data can be found on the variability of this repeat in the literature, suggesting the bioinformatic data may be the first to show that this region is polymorphic. Even with the small sample size, it is evident this region is variable amongst individuals and its alleles should be better studied. Utilizing a larger sample size may allow for a

clearer pattern of copy number variants for this repeat and thus may allow for a better understanding of its association with the SNP and AIS. This study was successful in bettering the understanding of a site that might contribute to the complex nature of the genetic basis of AIS. The eventual understanding of the biological underpinnings of AIS has implications that reach farther than just understanding the inheritance pattern or mechanism of the development of a spinal curvature.

Research has shown that AIS has implications for adolescent's mental health in addition to the more obvious physical health implications (Wang et al., 2019). Anxiety and depression have been found to be associated with an AIS diagnosis which can be attributed to the negative self-image or discomfort caused by the spinal curvature (Rainoldi et al., 2015). Not only does AIS influence adolescent's quality of life but it influences their comfortability in social settings which impacts their relationships (Wang et al., 2019). Clinicians are recommended to consider their patient's mental health when in the treatment and diagnosis process due to the ramifications the disease can have (Studner et al., 2013). AIS is evidently a disease that does not just impact the physical health of individuals, and its implications on mental health only emphasize the importance of understanding its biological underpinnings so that diagnostic and treatment methods are appropriate, effective, and efficient.

Chapter 5 Appendix A

Table 4: Frequencies of copy number variants in the sixty-four trimmed sequence files. The table indicates the number of times a certain copy number was seen in the sequence as well as what percentage of the total copy number variants that specific variant account for.

Frequencies of Copy Number Variants

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	3.0	2	3.0	3.0	3.0
	4.9	1	1.5	1.5	4.5
	15.9	7	10.6	10.6	15.2
	21.9	1	1.5	1.5	16.7
	36.0	1	1.5	1.5	18.2
	43.0	1	1.5	1.5	19.7
	44.0	1	1.5	1.5	21.2
	57.0	1	1.5	1.5	22.7
	65.0	1	1.5	1.5	24.2
	66.0	2	3.0	3.0	27.3
	67.0	6	9.1	9.1	36.4
	68.0	2	3.0	3.0	39.4
	71.0	2	3.0	3.0	42.4
	75.0	1	1.5	1.5	43.9
	80.0	1	1.5	1.5	45.5
	82.0	1	1.5	1.5	47.0
	84.0	1	1.5	1.5	48.5
	85.0	2	3.0	3.0	51.5
	87.0	1	1.5	1.5	53.0
	88.0	1	1.5	1.5	54.5
	90.0	3	4.5	4.5	59.1
	91.9	1	1.5	1.5	60.6
	93.0	1	1.5	1.5	62.1
	94.0	2	3.0	3.0	65.2
	96.0	13	19.7	19.7	84.8
	97.0	1	1.5	1.5	86.4
	100.0	2	3.0	3.0	89.4
	105.0	1	1.5	1.5	90.9
	109.0	1	1.5	1.5	92.4
	113.0	1	1.5	1.5	93.9
	115.0	2	3.0	3.0	97.0
	118.0	1	1.5	1.5	98.5
	156.0	1	1.5	1.5	100.0
	Total	66	100.0	100.0	

BIBLIOGRAPHY

Asghari, V., Sanyal, S., Buchwaldt, S., Paterson, A., Jovanovic, V., & Van Tol, H. H. (1995). Modulation of intracellular cyclic AMP levels by different human dopamine D4 receptor variants. *Journal of neurochemistry, 65*(3), 1157–1165. https://doi.org/10.1046/j.1471-4159.1995.65031157.x

Course, M. M., Sulovari, A., Gudsnuk, K., Eichler, E. E., & Valdmanis, P. N. (2021). Characterizing nucleotide variation and expansion dynamics in human-specific variable number tandem repeats. *Genome research, 31*(8), 1313–1324. https://doi.org/10.1101/gr.275560.121

Chen, L., Jensik, P. J., Yuan, X., Liu, M., Saffen, D., & Elsea, S. H. (2020). Evidence for a Potential Common Gene Network of Smith-Magenis and Potocki-Lupski Syndromes, DAND (DEAF1-Associated Neurodevelopmental Disorder) and ZEB1-Associated Neurodevelopment Disorder.

Danielsson, A. J., & Nachemson , A. L. (2001). Radiologic findings and curve progression 22 years after treatment for adolescent idiopathic scoliosis: comparison of brace and surgical treatment with matching control group of straight individuals. *Spine, 26*(5), 516–525. https://doi.org/10.1097/00007632-200103010-00015

Duitama, J., Zablotskaya, A., Gemayel, R., Jansen, A., Belet, S., Vermeesch, J. R., Verstrepen, K. J., & Froyen, G. (2014). Large-scale analysis of tandem repeat variability in the human genome. *Nucleic acids research, 42*(9), 5728–5741. https://doi.org/10.1093/nar/gku212

Eichler, E. E. (2019). Human-specific tandem repeat expansion and differential gene expression during primate evolution. Proceedings of the National Academy of Sciences of the United States of America, 116(46), 23243–23253. https://doi.org/10.1073/pnas.1912175116

Freidel, K., Petermann, F., Reichel, D., Steiner, A., Warschburger, P., & Weiss, H. R. (2002). Quality of life in women with idiopathic scoliosis. *Spine*, *27*(4), E87–E91. https://doi.org/10.1097/00007632-200202150-00013

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S. N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E. S., Daly, M. J., & Altshuler, D. (2002). The structure of haplotype blocks in the human genome. *Science (New York, N.Y.)*, *296*(5576), 2225–2229. https://doi.org/10.1126/science.1069424

Grady, D. L., Thanos, P. K., Corrada, M. M., Barnett, J. C., Jr, Ciobanu, V., Shustarovich, D., Napoli, A., Moyzis, A. G., Grandy, D., Rubinstein, M., Wang, G. J., Kawas, C. H., Chen, C., Dong, Q., Wang, E., Volkow, N. D., & Moyzis, R. K. (2013). DRD4 genotype predicts longevity in mouse and human. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, *33*(1), 286–291. https://doi.org/10.1523/JNEUROSCI.3515-12.2013

Grauers, A., Rahman, I., & Gerdhem, P. (2012). Heritability of scoliosis. European spine journal: official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society, 21(6), 1069–1074. https://doi.org/10.1007/s00586-011-2074-1

Green, M. R., & Sambrook, J. (2018). The Basic Polymerase Chain Reaction (PCR). *Cold Spring Harbor protocols*, *2018*(5), 10.1101/pdb.prot095117. https://doi-org.ezaccess.libraries.psu.edu/10.1101/pdb.prot095117

Hannan, A. Tandem repeats mediating genetic plasticity in health and disease. *Nat Rev Genet* **19,** 286–298 (2018). https://doi.org/10.1038/nrg.2017.115

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. (2002). The human genome browser at UCSC. *Genome research*, *12*(6), 996–1006. https://doi.org/10.1101/gr.229102

Kikanloo, S. R., Tarpada, S. P., & Cho, W. (2019). Etiology of Adolescent Idiopathic Scoliosis: A Literature Review. *Asian spine journal*, *13*(3), 519–526. https://doi.org/10.31616/asj.2018.0096

Konieczny, M. R., Senyurt, H., & Krauspe, R. (2013). Epidemiology of adolescent idiopathic scoliosis. *Journal of children's orthopaedics*, *7*(1), 3–9. https://doi.org/10.1007/s11832-012-0457-4

Kreig, A., Calvert, J., Sanoica, J., Cullum, E., Tipanna, R., & Myong, S. (2015). G-quadruplex formation in double strand DNA probed by NMM and CV fluorescence. *Nucleic acids research*, *43*(16), 7961–7970. https://doi.org/10.1093/nar/gkv749

Liang, J., Xing, D., Li, Z., Chua, S., & Li, S. (2014). Association Between rs11190870 Polymorphism Near LBX1 and Susceptibility to Adolescent Idiopathic Scoliosis in East Asian Population: A Genetic Meta-Analysis. *Spine*, *39*(11), 862–869. https://doi-org.ezaccess.libraries.psu.edu/10.1097/BRS.0000000000000303

Linthorst, J., Meert, W., Hestand, M. S., Korlach, J., Vermeesch, J. R., Reinders, M., & Holstege, H. (2020). Extreme enrichment of VNTR-associated polymorphicity in human subtelomeres: genes with most VNTRs are predominantly expressed in the brain. Translational psychiatry, 10(1), 369. https://doi.org/10.1038/s41398-020-01060-5

Liu, J., Zhou, Y., Liu, S., Song, X., Yang, X. Z., Fan, Y., Chen, W., Akdemir, Z. C., Yan, Z., Zuo, Y., Du, R., Liu, Z., Yuan, B., Zhao, S., Liu, G., Chen, Y., Zhao, Y., Lin, M., Zhu, Q., Niu, Y., … Wu, N. (2018). The coexistence of copy number variations (CNVs) and single nucleotide polymorphisms (SNPs) at a locus can result in distorted calculations of the significance in associating SNPs to disease. *Human genetics*, *137*(6-7), 553–567. https://doi.org/10.1007/s00439-018-1910-3

Liu, Sen MD∗,†,‡; Wu, Nan MD∗,†,‡; Zuo, Yuzhi MD∗; Zhou, Yangzhong MD∗; Liu, Jiaqi MD∗; Liu, Zhenlei MD∗; Chen, Weisheng MS∗; Liu, Gang MS∗; Chen, Yixin MS∗; Chen, Jia MS∗; Lin, Mao MS∗; Zhao, Yanxue MS∗; Ming, Yue MD†; Yuan, Tangmi MD∗; Li, Xiao MD∗; Xia, Zenan MD∗; Yang, Xu MS∗; Ma, Yufen MS∗; Zhang, Jianguo MD∗; Shen, Jianxiong MD∗; Li, Shugang MD∗; Wang, Yipeng MD∗; Zhao, Hong MD∗; Yu, Keyi MD∗; Zhao, Yu MD∗; Weng, Xisheng MD∗,†,‡; Qiu, Guixing MD∗,†,‡; Wu, Zhihong MD†,‡ Genetic Polymorphism of LBX1 Is Associated With Adolescent Idiopathic Scoliosis in Northern Chinese Han Population, SPINE: August 1, 2017 - Volume 42 - Issue 15 - p 1125-1129 doi: 10.1097/BRS.0000000000002111

Logsdon, G. A., Vollger, M. R., & Eichler, E. E. (2020). Long-read human genome sequencing and its applications. *Nature reviews. Genetics*, *21*(10), 597–614. https://doi.org/10.1038/s41576-020-0236-x

Mantere, T., Kersten, S., & Hoischen, A. (2019). Long-Read Sequencing Emerging in Medical Genetics. *Frontiers in genetics*, *10*, 426. https://doi.org/10.3389/fgene.2019.00426

Meyer, R. B., Jr, & Miller, J. P. (1974). Analogs of cyclic AMP and cyclic GMP: general methods of synthesis and the relationship of structure to enzymic activity. *Life sciences*, *14*(6), 1019–1040. https://doi-org.ezaccess.libraries.psu.edu/10.1016/0024-3205(74)90228-8

Peng Y, Liang G, Pei Y, Ye W, Liang A, Su P: Genomic polymorphisms of G-Protein Estrogen Receptor1 are associated with severity of adolescent idiopathic scoliosis. Int Orthop 36:671–677, 2012 73.

Ptácek, R., Kuzelová, H., & Stefano, G. B. (2011). Dopamine D4 receptor gene DRD4 and its association with psychiatric disorders. *Medical science monitor: international medical journal of experimental and clinical research*, *17*(9), RA215–RA220. https://doi.org/10.12659/msm.881925

Rainoldi, L., Zaina, F., Villafañe, J. H., Donzelli, S., & Negrini, S. (2015). Quality of life in normal and idiopathic scoliosis adolescents before diagnosis: reference values and discriminative validity of the SRS-22. A cross-sectional study of 1,205 pupils. *The spine journal : official journal of the North American Spine Society*, *15*(4), 662–667. https://doi.org/10.1016/j.spinee.2014.12.004

Schwab, F., Dubey, A., Gamez, L., El Fegoun, A. B., Hwang, K., Pagala, M., & Farcy, J. P. (2005). Adult scoliosis: prevalence, SF-36, and nutritional parameters in an elderly volunteer population. *Spine, 30*(9), 1082–1085. https://doi.org/10.1097/01.brs.0000160842.43482.cd

Sharkey, N. A., & Lang, D. H. (2007). Genes in context: probing the genetics of fracture resistance. *Exercise and sport sciences reviews, 35*(3), 86–96. https://doi.org/10.1097/jes.0b013e31809ff2ca

Slatkin M. (2008). Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nature reviews. Genetics, 9*(6), 477–485. https://doi.org/10.1038/nrg2361

Stevanovski I, Chintalaphani SR, Gamaarachchi H, Ferguson JM, Pineda SS, Scriba CK, Tchan M, Fung V, Ng K, Cortese A, Houlden H, Dobson-Stone C, Fitzpatrick L, Halliday G, Ravenscroft G, Davis MR, Laing NG, Fellner A, Kennerson M, Kumar KR, Deveson IW. Comprehensive genetic diagnosis of tandem repeat expansion disorders with programmable targeted nanopore sequencing. Sci Adv. 2022 Mar 4;8(9):eabm5386. doi: 10.1126/sciadv.abm5386. Epub 2022 Mar 4. PMID: 35245110.

Stundner, O., Kirksey, M., Chiu, Y. L., Mazumdar, M., Poultsides, L., Gerner, P., & Memtsoudis, S. G. (2013). Demographics and perioperative outcome in patients with depression and anxiety undergoing total joint arthroplasty: a population-based study. *Psychosomatics, 54*(2), 149–157. https://doi.org/10.1016/j.psym.2012.08.009

Sulovari, A., Li, R., Audano, P. A., Porubsky, D., Vollger, M. R., Logsdon, G. A., Human Genome Structural Variation Consortium, Warren, W. C., Pollen, A. A.,

Chaisson, M., & Trobisch, P., Suess, O., & Schwab, F. (2010). Idiopathic scoliosis. *Deutsches Arzteblatt international*, *107*(49), 875–884. https://doi.org/10.3238/arztebl.2010.0875

Trobisch, P., Suess, O., & Schwab, F. (2010). Idiopathic scoliosis. *Deutsches Arzteblatt international*, *107*(49), 875–884. https://doi.org/10.3238/arztebl.2010.0875

Wang, H., Li, T., Yuan, W., Zhang, Z., Wei, J., Qiu, G., & Shen, J. (2019). Mental health of patients with adolescent idiopathic scoliosis and their parents in China: a cross-sectional survey. *BMC psychiatry*, *19*(1), 147. https://doi.org/10.1186/s12888-019-2128-1

Witte J. S. (2010). Genome-wide association studies and beyond. *Annual review of public health*, *31*, 9 following 20.–. https://doi.org/10.1146/annurev.publhealth.012809.103723

Wu, Z. , Dai, Z. , Yuwen, W. , Liu, Z. , Qiu, Y. , Cheng, J. , Zhu, Z. & Xu, L. (2021). Genetic Variants of CHD7 Are Associated with Adolescent Idiopathic Scoliosis. *SPINE*, *46* (11), E618-E624. doi: 10.1097/BRS.0000000000003857.

Yaman, O., & Dalbayrak, S. (2014). Idiopathic scoliosis. *Turkish neurosurgery*, *24*(5), 646–657. https://doi.org/10.5137/1019-5149.JTN.8838-13.0

Academic Vita

Carly Brogan

EDUCATION

Bachelor of Science in Biobehavioral Health
The Pennsylvania State University, University Park, PA — *Dean's List (7 Semesters)*
Schreyer Honors Scholar
Graduation: Spring 2022

RESEARCH

Molecular Genetics Laboratory — **Spring 2019 - Present**
Student Researcher
- Perform PCR reactions to amplify tandem repeats located in the intergenic region of the dopamine receptor D4 and the deformed epidermal autoregulatory genes to identify links to Idiopathic Scoliosis
- Coded a method to view and analyze high genomic coverage areas in exons and introns of genetic assembly sequences

WORK EXPERIENCE

Einstein Medical Center — **Summer 2019**
Patient Advocate Intern
- Worked with the hospital's Patient Advocate in the Emergency Department
- Interacted with patients in the department and spoke with them about their experience in the hospital and worked to accommodate any of their needs or requests
- Took patient vitals upon entering the ER
- Sent letters and made phone calls to previous patients to collect information on how to improve the department
- Set up and attended meetings with the Patient Experience team of the hospital

Penn State Biobehavioral Health Department — **Spring 2020**
Undergraduate Teaching Intern for BBH311
- Aided in the student's review of literature regarding scientific domains and how they apply to Biobehavioral Health
- Held weekly office hours

Movie Tavern
Server
- Welcome guests and take orders
- Clean and organize the theaters between movie times

COMPUTER SKILLS

Bioinformatics: UCSC Genome Browser, Bioconda, PrimerBLAST, SnapGene
General applications: Microsoft Office Suite

ACTIVITIES

- Student Red Cross Club | 2019- 2021
- Phi Eta Sigma Honors Society member | 2019-2021
- Schreyer Mentoring with Honors program | 2020-2021
- Schreyer Honors College Orientation Mentor | 2020-2021
- ICNA Relief SHAMS Clinic Volunteer | 2020-present

AWARDS

- Provost Award Recipient | 2018
- Student Engagement Network Innovation Grant Recipient | 2021