*Article*

**A Study on COL5A1 in the CHS (Han Chinese South) Population**

**Rowan Whelan 1 and Christopher A Schmitt 2**

|  |
| --- |
| **Citation:** To be added by editorial staff during production.  Academic Editor: Firstname Lastname  Received: date  Revised: date  Accepted: date  Published: date    **Copyright:** © 2024 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/). |

1 *Department of Computer Science, Boston University*; rwhelan3@bu.edu

2 *Department of Anthropology, Boston University*; caschmitt@bu.edu

**Abstract: Background/Objectives**: The goal of this study was to investigate the SNPs of interest from *Correlations Between the Genetic Variations in the* *COL1A1, COL5A1, COL12A1, and β-fibrinogen Genes and Anterior Cruciate Ligament Injury in Chinese Patients* **[**1} within the Chinese population. These SNPs are linked with histories of injury within patients and study into them could help advance healthcare outcomes. **Methods**: Genetic Analysis with a focus on HWE, LD, Phylogenic Trees, Tajima’s D and iHS scores. **Results**: Based on the available evidence, it is very unlikely the COL5A1 gene within the 1000 Genomes Project CHS project is under selective pressure. Linkage blocks represent high linked areas which are likely linked to maintain genetic traits which help the human body function.

**Keywords:** Linkage Disequilibrium, Selective Pressure, Phylogenetic Tree, Tajima’s D, Extended Haplotype Homozygosity

**1. Introduction**

The gene of interest of this study, COL5A1, encodes the instructions for the body to create collagen V. Collagen V promotes growth and repair in connective tissue and other import soft tissue, and mutations on this gene can lead to many illness and health issues. For example, a study by Dr. Anne De Paepe et al. found that mutations within COL5A1 are causal in both types of Ehlers-Danlos Syndrome. The CHS (Han Chinese south) population was initially selected for study because of an article which found correlations between genetic variation in COL5A1 (and two other COL \_ A1 genes) with ligament injuries in this population. However, in further review the SNP on CO5A1 mentioned in this article was found to be missing from the 1000 Genomes Project’s samples. This deterrent didn’t stop the study however, this study will focus on the overall selective pressures on the COLA51 to see whether these health implications can constitute a selective pressure. The primary hypothesis of this study is that COL5A1 would shows signs of selective pressure because of its essential function. Furthermore, the primary points of interest are the genetic diversity exists within this population and comparative genetic diversity within other closely related populations to try to determine historical changes in the gene.

**2. Materials and Methods**

**Data Retrieval:** The gene of interest, COL5A1, was extracted from chromosome 9 for all populations within the 100 genomes project with the region of COL5A1 defined as chr9:134641803-134844843. **Population Filtering:** specific population subsets were downloaded and filtered on using Unix commands – CHS (Han Chinese South), JPT (Japanese in Tokyo) and CHB (Han Chinese in Beijing). Then these lists were used to filter the COL5A1 VCF fil. **Linkage Disequilibrium (LD) Analysis:** Using R studio, the VCF files were converted to SNP matrixes and Pairwise LD values were calculated. Then a LD heatmap with some SNPs of interest highlighted was created using the LDheatmap package for RStudio. Finally, these graphs were tweaked for readability and some regions of interest were highlighted with the BigLD package. **Phylogenetic Analysis:** The VCF files were then converted to DNAbin objects for further analysis. Pairwise distances were calculated using a K80 substitution model, and a Neighbor Joining (NJ) was constructed. The tree for the population of interest was further annotated for readability. **Tajima’s D:** The Tajima’s D statistic was calculated for the gene region to assess neutrality. **Selective Sweep Analysis:** Haplotypes were filtered and Extended haplotype homozygosity (EHH) was scanned. Using these scans, Integrated Haplotype Scores (iHS) were computed, and candidate selective sweep region were identified. Manhattan Plots were generated to visualize the iHS values and highlight candidates for further analysis. These candidates were then filtered through and cross referenced against the Ensemble genetic database.

**3. Results**

3.1 Hardy Weinberg Equilibrium:

A Hardy Weinberg Equilibrium (HWE) test was done to gauge the amount of selection acting upon the COL5A1 gene region. SNPs that are out of HWE fail to meet the assumptions this measure makes about a population. One of these assumptions is that the population under HWE is not actively under selection, so this metric can be an effective flag for areas which are under selection. After running HWE on the CHS populations VCF there were 53 SNPs whose P values were significant (P < 0.05).A graph of a bar chart

Description automatically generated with medium confidence

Figure 1

3.2 Linkage Disequilibrium:

Pairwise LD values were calculated using the R2 metric. The physical span of the COL5A1 gene is 202.9 kb and this is reflected in the following figures. Figure 1 is a simple breakdown of the calculated LD values in a heatmap format. 2 SNPs of interest from the HWE tests are highlighted. These two SNPs were highlighted because they had widely varied results. The SNP at 9:134651496 had a P score of .016, while the SNP at A black ruler with yellow lines

Description automatically generated9:134654639 had a much more extreme P score of 0. The physical proximity of these two SNPs and different levels of LD are interesting and could provide insight into the inheritance patterns of this region. The second Figure is a similar LD but this time it has been annotated to show the 4 most extreme LD blocks. It should be noted that there were 17 different blocks of LD, but for readability, these were filtered to the 4 biggest blocks.A blue and black tree with many letters

Description automatically generated with medium confidence

Figure 2

Figure 3

3.3 Phylogenetic Analysis:

The phylogenetic analysis of the COL5A1 gene provided insights into the genetic relationships within and between different 1000 genomes populations. The distance metric chosen to represent the genetic distances was the Kimura 2-paramater (K80). A neighbor-joining (NJ) tree was constructed based on this distance parameter (Figure 3). Finally, a multi tree was constructed to compare the diversity within the CHS population to that in other populations (Figure 4).A graph of different colored lines

Description automatically generated with medium confidence

Figure 4

3.4 Tajima’s D:

To assess patterns of genetic variation in the COL5A1 gene, Tajima’s D values were calculated for the CHS population. This analysis resulted in a Tajima’s D score of D = 0.262 which had a P value of 0.793.Figure 5


Figure 5

3.5 Selective Sweep Analysis:

To investigate evidence of selective sweeps in COL5A1 region, Extended Haplotype Homozygosity (EHH) was scanned. Using this Integrated Haplotype Scores (iHS) were calculated and plotted, then regions with extreme his scores were highlighted (Figure 5). 

**4. Discussion**

The goal of this study was to investigate the SNPs of interest highlighted in Zhao et al. [1] and Guiseppe Longo et al. [2]. However, due to the lack of evidence in the 1000 Genomes Project, these SNPs were abandoned to focus on a more comprehensive study of the gene. The primary conclusion that can be made from the results is a lack of evidence for strong selective pressures on the COL5A1 gene in the CHS population. Upon further investigation, the HWE scores which were thought to be due to selection occurring on the gene were due to confounding variables. The highlighted LD blocks suggest that certain haplotypes are conserved regularly, but this is not surprising given the fact that this gene is responsible for one of the building blocks of human locomotion. For example, if the phenotype of COL5A1 was incapable of producing Collagen, then the reproductive success of that subject would be very low during the early evolutionary period for humans. The phylogenic trees for the CHS population show a couple of branches from which densely packed leaves follow, this implies a regular level of genetic diversity and provides no evidence of a bottleneck effect in recent evolutionary history. The Tajima’s D value of 0.262 and non-significant P value suggest that there isn’t any evidence for a departure from neutrality. This means that recent selective sweeps or population bottlenecks are very unlikely. This is also supported by only 4 points having extreme iHS scores. An extreme iHS score would imply that there had been a selective sweep over the gene. The combinations of these factors provide evidence for a relatively stable genetic structure with few recent selective pressures.

**Informed Consent Statement:** This study did not involve the collection of any new data from individuals.

**Appendix A**

The code used for this study can be found at https://github.com/rowanwhelan/A-Study-on-COL5A1-in-the-CHS-Han-Chinese-South-Population

**References**

1. Daohong Zhao, Qi Zhang; Correlations Between the Genetic Variations in the *COL1A1*, *COL5A1*, *COL12A1*, and *β-fibrinogen* Genes and Anterior Cruciate Ligament Injury in Chinese Patients. NIH 2020 [pg 1-4]
2. Umile Giuseppe Longo, Katia Margiotti; Genetics of rotator cuff tears: no association of col5a1 gene in a case-control study. NIH 2018 [pg 1-10]

in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.