# Machine Learning-based Genome Wide Association Studies of Rheumatoid Arthritis

Allie Burton
Advisor: Prof. Sandra Batista

## Introduction

Scoliosis is a disease marked by curvature of the spine. The most common type, adolescent idiopathic scoliosis (AIS), generally occurs right before puberty and has no known causes. According to the Scoliosis Research Society, approximately 30% of all AIS patients have some family history of scoliosis, so many researchers now are looking for a genetic component[1]. To look for these genetic components, researchers perform what is known as a genome-wide association study, or GWAS, which the National Human Genome Research Institute defines as an "approach that involves rapidly scanning markers across the complete sets of DNA, or genomes, of many people to find genetic variations associated with a particular disease"[2]. The results of these studies, however, have been generally inconclusive collectively. For example, Takahashi et. al.'s[3] GWAS studying approximately 1,400 Japanese females shared data on 87 of the top 100 single nucleotide polymorphisms (SNPs) found in Sharma et. al.'s[4] racially diverse GWAS of 419 families, however, only one of those SNPs showed significant association in Takahashi et. al's GWAS. The goal of my project is to study the usage and efficacy of machine learning for GWAS in scoliosis.

# Background

## Related Work

Although there have not been any studies done to date using machine learning for GWAS of idiopathic scoliosis, there have been many studies using machine learning for other phenotypes including IgM and rheumatoid arthritis, as mentioned above, in addition to myocardial infarction, coronary artery calcification, and anti-cyclic citrullinated peptide[5]. These studies will provide the basis for my methodology, specifically D'Angelo et. al.'s[6] and Tang et. al.'s[7] GWASs of rheumatoid arthritis and Stassen et. al.'s GWASs of IgM. Since there are no prior machine learning-based GWASs of AIS, I will replicate their respective methodologies as best as possible, adapting where necessary for the specifics of scoliosis and the data sets I am using.

## Important Terminology

To conduct a genome-wide association study, researchers get DNA samples from two groups of people: those with the trait in question and those without it. Using these samples, each person's entire genome is scanned in a machine looking for single-nucleotide polymorphisms. A single-nucleotide polymorphism, or SNP (prounounced "snip"), is a variation at a single position in an individual's DNA sequence that occurs in one percent or less of the population. If certain SNPs occur more frequently in persons with the disease than without it, then those SNPs are associated with the trait. Although these SNPs can point to places in the human genome that might be related to the source of the trait, the SNPs themselves may not be the cause of the trait itself, so researchers often look at base pairs in the region to see if they are also related[8].

# Methods

## Phase 1

The purpose of Phase 1 was to familiarize myself with using PLINK, an open-source command-line "whole genome association analysis toolset"[9]. For this phase, I used the GAW16 data set from the North American Rheumatoid Arthritis Consortium. This data set contains 868 cases of rheumatoid arthritis and 1,164 controls for a total sample size of 2,062. This data set came in two parts: a CSV file and a MAP file. The CSV file contains 2,062 records each with the following fields:

- subject ID
- Rheumatoid arthritis (RA) affection status
- gender
- HLA-DRB1 allele 1
- HLA-DRB1 allele 2
- number of shared-epitope alleles
- existence of shared-epitope alleles
- anti-CCP
- rheumatoid factor IgM
- 545,080 SNP-genotype fields

Of these fields the most relevant for my purposes are the RA affection status, and the SNP-genotype fields. The .map file is a special type of file used by PLINK that contains information on each of the 545,080 SNPs with these fields:

- SNP name
- chromosome
- SNP position in basepairs

In order to ensure compatibility with PLINK, the CSV file needed to be reformatted to .ped format, the standard file format for PLINK. This pre-processing consisted of removing the fields corresponding to HLA-DRB1 alleles, number of shared-epitope alleles, existence of shared-epitope alleles, anti-CCP, and rheumatoid factor IgM; rearranging columns 1-3; changing the denotation for male and female from "M" and "F" to "1" and "2"; and reformatting the denotation for genotypes.

## References


[1] Scoliosis Research Society, "Adolescent Idiopathic Scoliosis | Scoliosis Research Society."

[2] National Human Genome Research Institute, "Genome-Wide Association Studies Fact Sheet," 2015.

[3] Y. Takahashi, I. Kou, A. Takahashi, T. a. Johnson, K. Kono, N. Kawakami, K. Uno, M. Ito, S. Minami, H. Yanagida, H. Taneichi, T. Tsuji, T. Suzuki, H. Sudo, T. Kotani, K. Watanabe, K. Chiba, N. Hosono, N. Kamatani, T. Tsunoda, Y. Toyama, M. Kubo, M. Matsumoto, and S. Ikegawa, "A genome-wide association study identifies common variants near LBX1 associated with adolescent idiopathic scoliosis," *Nature Genetics*, vol. 43, no. 12, pp. 1237–1240, 2011.

[4] S. Sharma, X. Gao, D. Londono, S. E. Devroy, K. N. Mauldin, J. T. Frankel, J. M. Brandon, D. Zhang, Q. Z. Li, M. B. Dobbs, C. A. Gurnett, S. F. A.

Grant, H. Hakonarson, J. P. Dormans, J. A. Herring, D. Gordon, and C. A. Wise, "Genome-wide association studies of adolescent idiopathic scoliosis suggest candidate susceptibility genes," *Human Molecular Genetics*, vol. 20, no. 7, pp. 1456–1466, 2011.

[5] S. Szymczak, E. Holzinger, A. Dasgupta, J. D. Malley, A. M. Molloy, J. L. Mills, L. C. Brody, D. Stambolian, and J. E. Bailey-Wilson, "r2VIM: A new variable selection method for random forests in genome-wide association studies.," *BioData mining*, vol. 9, p. 7, 2016.

[6] G. M. D'Angelo, D. Rao, and C. C. Gu, "Combining least absolute shrinkage and selection operator (LASSO) and principal-components analysis for detection of gene-gene interactions in genome-wide association studies.," *BMC proceedings*, vol. 3 Suppl 7, no. Suppl 7, p. S62, 2009.

[7] R. Tang, J. P. Sinnwell, J. Li, D. N. Rider, M. de Andrade, and J. M. Biernacka, "Identification of genes and haplotypes that predict rheumatoid arthritis using random forests," *BMC Proceedings*, vol. 3, no. Suppl 7, p. S68, 2009.

[8] Nature Education, "single nucleotide polymorphism / SNP."

[9] C. Chang, C. Chow, S. Vattikuti, L. Tellier, J. Lee, and S. Purcell, "PLINK."