Exploring the Genetic Basis of Congenital Heart Defects

Sanjay Siddhanti **Jordan Hannel** **Vineeth Gangaram**

szsiddh@stanford.edu jfhannel@stanford.edu vineethg@stanford.edu

1 Introduction

The Human Genome Project[1], completed in 2003, opened the door for Big Data approaches to studying Mendelian disorders[2]. In the decade since, hundreds of millions of dollars have gone towards funding Genome Wide Association Studies[3] (GWAS), which have yielded interesting results but have almost completely failed at actually explaining the genetic basis of any diseases[4]. Because of monetary constraints, most GWAS only consider approximately 1-2 million commonly occurring mutations seen on SNP arrays[5] in order to avoid the costs of sequencing all 3 billion base pairs in the genome. Because researchers in the field are slowly abandoning GWAS, much of this SNP array data is now ignored as people see more promise in cutting edge techniques such as Next Generation Sequencing[6].

Congenital Heart Defects (CHD) are an example of a disease for which GWAS was only able to explain a small percentage of cases. CHD continues to be the most common birth defect in the United States, affecting almost 1% of live births[7] and increasing in prevalence as time passes[8]. In this project we set out to salvage results from forgotten SNP arrays for CHD patients by building a classifier that predicts whether a patient has a congenital heart defect based on his or her genome sequence at 2.5 million SNPs[1]. Because the diagnosis of CHD upon birth is fairly good, achieving 100% classifier accuracy is not critical here; instead, our main goal is to identify a minimal feature set of loci (locations in the genome) that warrant further biological investigation as potentially causal agents of CHD. This is a challenging problem that many researchers have approached in various forms, and any progress would be considered notable in its own right. In order to avoid potentially missing some interesting signal that past researchers have overlooked, we consciously try to minimize the amount of false negatives emitted by our classifier.

2 Methods

2.1 Data Preprocessing

The human genome is full of noisy components that can potentially confound our analysis. In order to maintain confidence that any results we find actually reflect a genetic link to congenital heart defects, we take a couple steps to control for confounding variables.

2.1.1 Segmentation of Patients by Ethnic Group

Ethnic differences between patients are very easily detected in the genome. If the genetic makeup of the positive and negative class individuals differs in any way, a classifier on this data would easily be able to attain high accuracy simply by giving high weight to the variants that differ between the racial groups; this classifier would not learn anything useful about CHD. To control for this, we group all CHD infants by ethnicity, as reported by their parents at the time of the study. We choose to focus on the largest ethnic group, White infants with CHD, consisting of 73 individuals.

2.1.2 Introduction of Negative Class via the Thousand Genomes Project

We now seek negative class samples with a couple key requirements. First, these patients should be healthy individuals who do not have any form of CHD. Second, these patients must be ethnically similar to our positive class samples, so that the classifier does not simply learn to detect ethnic differences. Third, we must be able to access genotype data for these individuals at the same 2.5 million SNPs that we have for the positive class patients.

[1] SNP, pronounced "snip," stands for Single Nucleotide Polymorphism. A 2.5 million SNP array surveys patient genotypes at 2.5 million locations in the genome at which a mutation is common. We say that a patient "has" a SNP if he or she has the common mutation at that particular location.

We notice that all of these requirements appear to be satisfied by the 543 individuals in the 1000 Genomes Project[9] who are of European descent. To confirm that our positive and negative class are ethnically similar, we performed a Principal Components Analysis as shown in Figure 1 below.

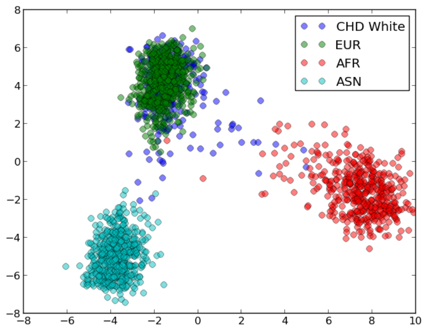

Figure 1. Principal Components Analysis of positive class samples and negative class samples (who we assume to all be White individuals) in addition to Asian and African individuals from the 1000 Genomes Project. Since the positive and negative class samples cluster together by race, but separately from the other two races, this is an indication that racial factors will not confound our classifier.

2.2 Feature Selection

As our primary goal is to identify a minimal feature set that still achieves reasonable classification accuracy on the data, feature selection is the key problem in this project.

2.2.1 Reduction of Feature Set Using Domain Knowledge

Prior GWAS studies have already attempted to identify SNPs that may be relevant to CHD. The intent of this project is to improve on those studies, not to recreate them. Prior efforts in the Bejerano Lab have identified a list of 1179 SNPs over 10 biological pathways that we believe may be relevant to CHD. Therefore we immediately reduce our feature set to these 1179 SNPs and set out to identify key SNPs within these pathways. It is not uncommon in genetics for a handful of SNPs to cause a large change in phenotype.

2.2.2 Filter Feature Selection

We implement filter feature selection using two separate metrics, one of which integrates domain knowledge of genetics.

2.2.2.1 Filtering by Mutual Information (MI)

In this version of filter feature selection, we calculate the Mutual Information of every SNP with respect to the output variable (disease or no disease) and select the top N SNPs as our feature set, where N is a threshold that we determine.

2.2.2.2 Filtering by Transmission Disequilibrium Test (TDT) Score

The Transmission Disequilibrium Test[10] detects alleles that are in linkage disequilibrium in a population of patients with a particular disease. Let's define A and a to be the major (common) and minor (mutation) alleles, respectively, at SNP i. If SNP i is not associated with CHD, then allele transmission at SNP i should be uniformly random according to standard genetic laws. That is, we should see that parents with genotype Aa at SNP i should be equally likely to transmit either allele (A or a) to their CHD offspring. However, if allele a at SNP i is strongly linked to CHD, we would expect to see not uniformly random transmission at SNP i, but rather a preponderance of transmission of allele a at this SNP.

At SNP i we look at the genotypes of each CHD child and their parents in order to determine the specific allele that each parent transmitted to the child. We then increment one number in the following table for each parent:

	Non-transmitted Allele	
Transmitted Allele	A	a
A	b	c
a	d	e

Figure 2[11]: Illustration of Transmission Disequilibrium Test. No information is given by entries b and e in the table since homozygous parents will always transmit one copy of an allele and not transmit one copy of the same allele.

The test then uses a chi-squared test with one degree of freedom to test the hypothesis that allele transmission is random at this locus. $\chi^2 = \frac{(c-d)^2}{c+d}$. We compute this chi-squared statistic at all SNPs in our feature set and filter by the associated p-value.

2.2.2.3 Filtering by a Combination of Mutual Information and Transmission Disequilibrium

The most straightforward method of combining two filter feature selection metrics is to "chain" the methods by first filtering by one statistic, and then by the other. We also study the correlation of MI and TDT scores over all SNPs.

Our most effective method examines the overlap of the top N features by MI and the top N features by TDT, for any threshold N. We choose the N at which the size of the overlap, relative to N, is most significant when testing against the null hypothesis that TDT and MI are unrelated. We then examine the SNPs in this overlap for biological significance and for efficacy as a reduced feature set.

2.3 Algorithm Selection

After analyzing the performance of several classifiers, including linear regression, logistic regression, Naïve Bayes, and Support Vector Machines (SVM) with various kernels and regularization parameters, we converged on a regularized SVM (penalty coefficient 0.5) with a Gaussian kernel. We found empirically that this algorithm did not overfit, and produced the best classifier given our preference for false positives over false negatives. Because we have 73 cases and 543 controls, we weight each sample inversely proportional to class frequency to account for this.

3 Results

We now present some of the more notable results that illustrate how feature selection impacts the performance of our classifier.

3.1 Filter Feature Selection

Figure 3 below shows how classifier accuracy changes as a function of feature set size. Interestingly, classification accuracy is fairly robust to reduced feature set sizes. This may be partially because of linkage disequilibrium, a known genetic phenomenon that posits that inheritance at nearby loci is closely linked. Therefore, knowing information about only one SNP in a region is often sufficient to extrapolate and obtain information about the remaining SNPs in that same region of the genome.

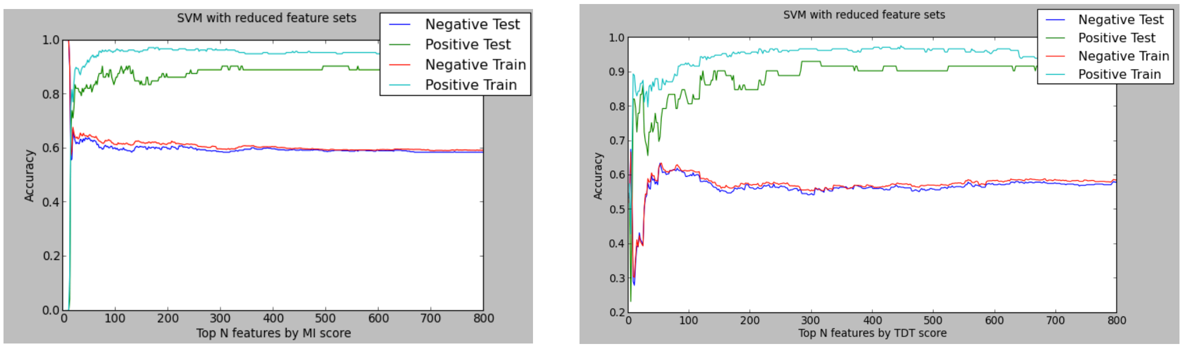

Figure 3. Classifier accuracy vs. feature set size for top features as determined by Mutual Information (left) and Transmission Disequilibrium Test (right).

Further investigation shows that using the top 15 SNPs by mutual information score as a feature set results in reasonable accuracy (Figure 4). At any feature set size below 15, accuracy rapidly decreases, indicating that each SNP being removed from the feature set is critical to classification accuracy.

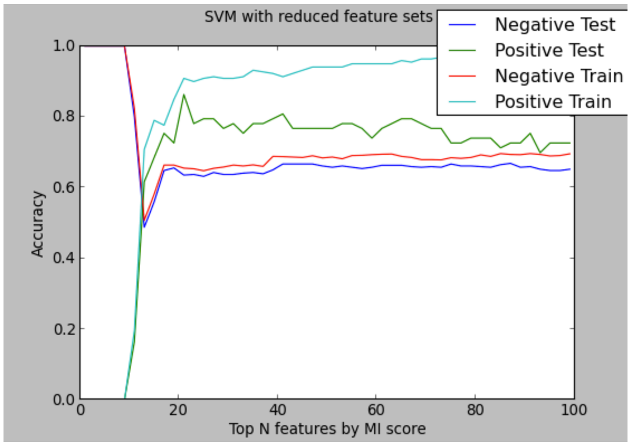

Figure 4. Zoom in on the Mutual Information chart from Figure 3. Classification accuracy on a feature set of size 15 is reasonable, and accuracy decreases sharply for each feature that is subsequently removed from the feature set.

The level of accuracy that the classifier achieves using only 15 SNPs is quite high, leading us to ask questions about the underlying distribution of the mutual information and TDT scores. These distributions are shown in Figure 5, where it is clear that 9 SNPs have much higher mutual information scores than the rest of the feature set.

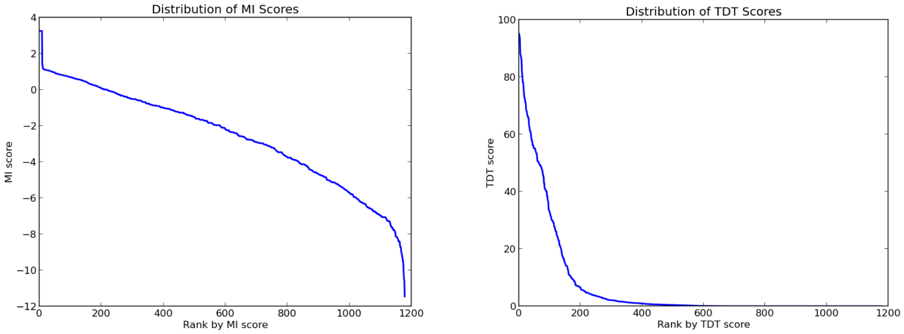

Figure 5. Distribution of Mutual Information (left) and TDT (right) scores.

Predicting using only the top 9 SNPs by mutual information score results in zero classification accuracy on the positive class, as represented in Figure 4. However, we still seek a slightly smaller feature set, but without the fall in accuracy that is seen in Figure 4. To solve this problem we explore methods of combining mutual information and TDT in order to select an even smaller feature set. Using the overlap method discussed in Section 2.2.2.3, we identify that the top 78 features by mutual information and the top 78 features by TDT share 9 SNPs in common, and these 9 SNPs alone have reasonable predictive power as a feature set. In fact, a classifier using this feature set of 9 SNPs has effectively the same accuracy on the positive class as a classifier using the top 100 features by mutual information does. These results are shown in Figure 6.

Feature Set	Pos Class Accuracy	Neg Class Accuracy
Top 15 MI	68.5%	56.0%
Top 15 TDT	56.2%	59.5%
Top 9 MI	0%	100%
Top 9 TDT	27.4%	82.3%
Top 9 TDT + MI overlap	75.3%	59.7%
Top 100 MI	76.7%	66.1%

Figure 6. SVM accuracy over various small feature sets.

Having identified a set of 9 features that may be related to CHD, our final step is to analyze the distribution of these 9 SNPs in the original 10 biological pathways that we started with. Biological pathways can share SNPs, especially if the pathways correspond to similar function. This distribution is shown in Figure 7, which clearly points to the second and tenth pathways as targets for future investigation.

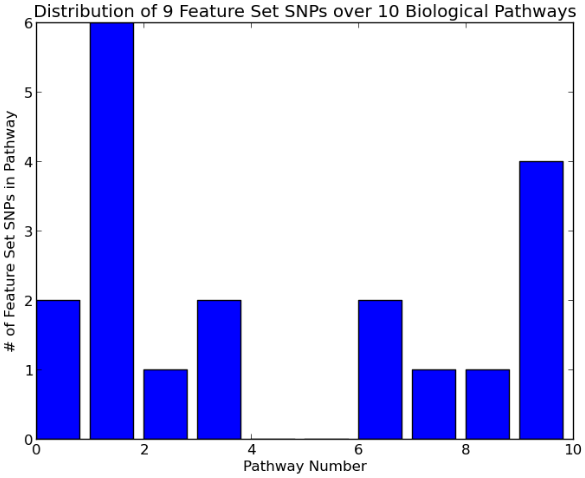

Figure 10. Distribution of the 9 feature set SNPs over the original 10 biological pathways from the Mammalian Phenotype Ontology[12].

4 Discussion

By integrating domain knowledge of genetics and genomics with traditional machine learning techniques we were able to improve on previous results and identify 9 SNPs and 2 biological pathways that warrant further investigation as potentially causal agents of CHD. These links cannot be validated until further biological experimentation is carried out. If any of these SNPs or pathways does show a new link to CHD, we will have successfully salvaged new information from data that many researchers have long forgotten.

Acknowledgements

We acknowledge Professor Gill Bejerano for access to data and lab computing resources, and both Professor Bejerano and Sandeep Chinchali for guidance on this project.

References

1. Lander et al."Initial Sequencing and Analysis of the Human Genome." *Nature* (2001): Web.
2. Kennedy, Martin A. "Mendelian Genetic Disorders." *Mendelian Genetic Disorders*(n.d.): Web. 11 Dec. 2014.
3. Visscher, Peter M. et al. "Five Years of GWAS Discovery." *American Journal of Human Genetics* 90.1 (2012): 7–24. *PMC*. Web. 11 Dec. 2014.
4. Jeffreys, Sir Alex. "Interviews." *European Human Genetics Conference 2010:* Web. 11 Dec. 2014.
5. LaFramboise, Thomas. "Single Nucleotide Polymorphism Arrays: a Decade of Biological, Computational and Technological Advances." *Nucleic Acids Research* 37.13 (2009): 4181–4193. *PMC*. Web. 11 Dec. 2014.
6. "Next-generation Sequencing." *Nature.com*. Nature Publishing Group, n.d. Web. 11 Dec. 2014.
7. "Fact Sheets." *The Children's Heart Foundation*. N.p., n.d. Web. 11 Dec. 2014.
8. Botto, Lorenzo D., Adolfo Correa, and David Erickson. "Racial and Temporal Variations in the Prevalence of Heart Defects." *Pediatrics* (2001): Web. 11 Dec. 2014.
9. "An Integrated Map of Genetic Variation from 1,092 Human Genomes."*Nature.com*. Nature Publishing Group, 31 Oct. 2012. Web. 11 Dec. 2014.
10. Spielman, R S, R E McGinnis, and W J Ewens. "Transmission Test for Linkage Disequilibrium: The Insulin Gene Region and Insulin-Dependent Diabetes Mellitus (IDDM)." *American Journal of Human Genetics* 52.3 (1993): 506–516. Print.
11. "Transmission Disequilibrium Test." *Wikipedia*. Wikimedia Foundation, 29 Nov. 2014. Web. 12 Dec. 2014.
12. Blake JA, Bult CJ, Eppig JT, Kadin JA, Richardson JE; The Mouse Genome Database Group. 2014. The Mouse Genome Database: integration of and access to knowledge about the laboratory mouse. Nucleic Acids Res. 2014 42(D1):D810-D817.