My Internship at Texas Biomedical Research Institute

I conducted an internship at the Texas Biomedical Research Institute in the spring of 2013. In this report I will discuss the history of the Texas Biomedical Research Institute and the type of research that is conducted there. I will also discuss one of the major studies I was able to contribute to and the work I completed while interning.

What is the Texas Biomedical Research Institute?

Originally named the Foundation of Applied Research, Texas Biomedical Research Institute was founded in 1941 by Thomas Slick. Today, Texas Biomedical Research Institute is one of the world’s leading independent biomedical research institutions. The 200-acre campus is located in northwest San Antonio, and consists of more than 400 employees, including over 85 scientists. Texas Biomed comprises several departments working both separately and together to fight various diseases from AIDS and parasites to cancer and diabetes. The departments include Southwest National Primate Research Center, the Department of Virology and Immunology, and the Department of Genetics. The Southwest National Primate Research Center of Texas Biomed is home to the world’s largest baboon colony and the Department of Virology and Immunology contains a biosafety level 4 laboratory. This biosafety level 4 laboratory allows researchers to work with deadly pathogens in order to further our understanding of the pathogens as well as treatment options for their respective diseases. In fact, Texas Biomed was responsible for testing the safety of the Merck Hepatitis B vaccine that is currently being used in humans.
The Genetics Department of Texas Biomed uses statistical analysis in order to research the genetic components responsible for disease susceptibility in animal and human populations. Due to the complexity of the statistical analysis, Texas Biomed is also the home of the AT&T Genomics Computing Center that contains 3,000 computer processors and is the largest genomic research computer cluster in the world. A group of scientists led by Dr. John Blangero design the software packages used on campus. One of these packages is SOLAR (Sequential Oligogenic Linkage Analysis Routines).

In addition to its many scientists with degrees in genetics, the genetics department employs many anthropologists from diverse backgrounds. The scientists work together on a number of ongoing projects, most taking years to complete. Many of the scientists at Texas Biomed are currently working on developing new phases of current projects as well as new studies.

My internship supervisor, Dr. Anthony Comuzzie, is currently working on projects that study the genetics of coronary artery disease in Alaskan Natives, the genetics of atherosclerosis in Mexican Americans, and a study on diet and genotype in primate atherosclerosis. These are only a few of the many studies to which Dr. Comuzzie has been a contributor.

**San Antonio Family Heart Study**

The San Antonio Family Heart Study (SAFHS) focuses on the genetics of atherosclerosis in Mexican-Americans. Despite the fact that heart disease is the number one cause of death among Mexican-Americans, SAFHS is the only population-based genetics study in Mexican-Americans. Atherosclerosis (heart disease) is associated with many risk factors such as obesity, glucose intolerance, and
hypertension. SAFHS tests the hypothesis that the genes that influence variation in these risk factors are likely to predispose individuals to atherosclerosis (Mitchell et. al. 1996). So far SAFHS has identified genes associated with insulin, fat mass and many others (Guowen et al 2004, Comuzzie et al. 2007). In the future, the scientists of Texas Biomed hope to be able to identify people with at-risk genes in order to tailor treatment to their specific genotype.

SAFHS began in 1992 and by 1995 it consisted of 1,236 individuals from more than forty families in San Antonio. Researchers chose these families from a single census tract that included a low-income neighborhood of San Antonio. Researchers approached households in this tract at random, and individuals were invited to participate in the study if they were older than forty and had at least six offspring or siblings older than sixteen living in San Antonio. All first, second, and third degree relatives of enrolled individuals and their spouses were invited to participate.

This type of study is called a pedigree study because it includes participants and their relatives. Researchers can create a pedigree diagram that looks similar to a family tree (Figure 1.). Any instances of disease or risk factors can be noted on this diagram, this especially useful for teasing out the genetic variants associated with a particular trait or disease because individuals share DNA.
Figure 1. Example of SAFHS Pedigree Diagram

The first phase of SAFHS was between 1992 and 1996 (Mitchell et al. 1996). In the first phase of the study, researchers collected immense amounts of data. Participants received a medical examination, glucose tolerance test and an interview. Each exam included anthropometric measurements such as skinfold thicknesses, waist circumference, ratio of waist to hip circumference, body mass index, and fat mass. Individuals also gave blood samples in order to measure lipids, glucose, and insulin. Those with diabetes were diagnosed according to the World Health Organization’s criteria for plasma glucose levels (Mitchell et al. 1996). Interviews were given to determine social, behavioral, and lifestyle factors that may be associated with heart disease. These interviews included questionnaires regarding family medical history, level of physical activity and dietary intake (Mitchell et al. 1996).

Analysis of phase one data focused on comparing subjects with and without a family history of heart disease risk factors such as diabetes, hypertension as well as the
presence of heart disease itself. First, the scientists used descriptive statistics to
determine the prevalence of these risk factors in the families of SAFHS. Next,
participants were assigned a family history score that represented the magnitude of
heart disease risk based on the presence of risk factors in their family. Family history
scores were calculated based on the number of family members with associated risk
factors and their relatedness to the participant.

Phase one of SAFHS study found significant associations between family history
of diabetes and a range of cardiovascular risk factors (Mitchell et al. 1999). The
relatives of diabetic participants who did not have diabetes or heart disease often had
elevated BMI and cholesterol (Mitchell et al. 1999). These findings are significant
because they indicate that one gene may control multiple cardiovascular risk factors.

The second phase of SAFHS focused more on genome-wide linkage analysis
which uses polymorphisms across the genome to identify Quantitative Trait Loci (QTLs)
associated with phenotypic traits. These polymorphisms, or microsatellites, are single
base pair repeats in DNA that can be found in individuals across populations.
Microsatellites serve as mile markers on the chromosome by which we can map the
location of possible genes linked with a trait such as obesity. QTLs are given a
Logarithm of Odds score (LOD) which is the chance that a particular trait is present
randomly. A LOD score greater than two suggests that the trait is the result of genetic
inheritance, or that there is a 1 and 100 chance that that trait is present randomly. Once
certain traits are linked to locations on a chromosome, researchers can begin to study
genes within that region to determine if they contribute to the phenotype of interest.
The second phase of SAFHS discovered QTLs for fat mass on chromosome two and BMI on chromosome 8 (Comuzzie et al. 2003). This helps to narrow down the location of possible genes that influence BMI and fat mass that will later help to determine what gene variations may contribute to unhealthy BMI or fat mass.

The current and third phase of SAFHS also uses genome wide linkage analysis but focuses on certain inflammatory factors associated with heart disease in addition to obesity related attributes. These inflammatory factors are also called cytokines, their presence in blood can indicate the presence of infection or something else that has triggered an inflammatory response.

**Statistical Analysis of the San Antonio Family Heart Study Data**

My work on the San Antonio Family Heart Study began with data management. Before I could conduct any statistical analyses on the dataset from SAFHS, I had to clean-up or adjust the data so that it resembled a normal distribution curve. In this case, the data consisted of various anthropometric measurements associated with obesity (BMI, fat mass, waist circumference), glucose, insulin, and levels of inflammatory cytokines taken from blood tests. I used PEDSYS, a UNIX based software package to remove the outliers so that the data fell within a normal distribution.

Removing outliers is a very important step in this process because outliers are usually the result of an error that occurred when researchers put data into a computer file. For example, weight must fall within a certain range. It is unlikely that you will find an individual who weighs more than 2,000 pounds. Including this type of data could have negative effects on your statistical analysis, especially if the tests assume that
your dataset falls in a normal distribution. In this case, it is more beneficial to remove outliers from your dataset.

Next, I had to adjust the data so that it was compatible with SOLAR. SOLAR is also a UNIX based software package but it was designed for complex statistical analysis. In order for SOLAR to display data properly, rows and columns of a dataset can only be a certain number of characters long. If I made any changes to the variables of a data set, I had to make sure that these changes were compatible with SOLAR. This process is usually more time consuming than the statistical analysis.

Once I cleaned up the data and made sure that the dataset was compatible with SOLAR, I began loading files necessary for statistical analysis. SOLAR uses two types of files when it runs statistical tests. The first file is the pedigree file that contains information about families and their relatedness. I had to use the copy command to copy the pedigree file from Dr. Voruganti’s master file into my master file. You need only load the pedigree file into SOLAR once. As long as the pedigree file is in your master file, SOLAR will be able to access it.

The second file is the phenotype file that contains information on the traits you are studying. My phenotype file contained anthropometric measurements and various measurements taken in the laboratory such as blood glucose levels, and fat mass measured by computer tomography scan (CT scan). Phenotype files are usually converted from imported excel datasets. This information is arranged according to each individual’s ID number. Unlike the pedigree file, the phenotype file must be loaded each time you want to run a statistical test.
I began my analysis of the SAFHS data by compiling basic descriptive statistics. I used the PEDSYS tally command on my phenotype file to obtain the means of obesity related measurements such as BMI, fat mass, fat free mass, and waist circumference grouped by sex. PEDSYS automatically performed calculations and I compiled the results in data tables.

Next, I loaded my phenotype file into SOLAR, which allowed me to calculate heritabilities for my traits of interest. Heritability ($h^2$), or narrow-sense heritability, is the proportion of total phenotypic variance that is due to additive genetic effects. Additive genetic variation is the variation that is the result of multiple genes contributing a small effect on the phenotype. Narrow-sense heritability requires a pedigree study because it uses measurements from both parents and offspring to measure the proportion of phenotypic variance that is the result of genetic factors. SOLAR is able to differentiate these values based on the pedigree file, and I simply specified the trait I wanted heritability for and typed in the command that adjusts for the variance due to sex and age (Figure 2). Heritability always ranges from zero to one.

```
Solar> load phen phenotype.out
Solar> trait BMI
Solar> covar age^sex#1,2
Solar> polygenic
```

**Figure 2. Example of Solar Command for Heritability Calculation**
In addition to heritability, SOLAR calculates a p-value for each trait which allows you to determine whether the $h^2$ is statistically significant. These p-values must equal less than 0.05 in order to be considered significant. This means that there is less than a 5% chance that these results occurred randomly. A heritability closer to one and a p-value less than 0.05 indicates that a large proportion of the phenotypic variation of a trait is due to genetic factors rather than environmental factors.

After obtaining heritabilities for traits, such as BMI and cytokine levels, I ran Quantitative Linkage Analysis on those same traits. Quantitative Linkage Analysis looks for associations between regions of a chromosome and a trait of interest. This results in a Quantitative Trait Loci. In order to run quantitative linkage analysis I had to load my phenotype file, make sure I had previously calculated heritabilities and then give SOLAR the command to search for associations on the chromosomes. It takes SOLAR a very long time to search for these associations as it scans the chromosome at intervals of five centimorgans. A centimorgan is a somewhat arbitrary measurement for distance on a chromosome; it allows us to map areas of the chromosome relative to one another rather than in an actual distance. When SOLAR discovers a LOD score greater than 1.0, it begins to scan that region of the chromosome at each centimorgan. Once SOLAR finishes the linkage scan, it gives the QTL with the greatest LOD score. QTLs are denoted by the chromosome and centimorgan associated with the trait of interest.

Lastly, after obtaining QTLs for all traits, it was my responsibility to translate these QTLs from arbitrary measurements into base pairs increments. I began by locating QTLs with LOD scores greater than 2.0. Then I noted the location on the chromosome equal to the LOD score, plus and minus 1.0. This ensured the region of
I was using contained the genes of interest. Next, I looked up the nearest satellite marker to the beginning and end chromosome loci. Each marker has a specific name that consists of letters and numbers. Once I located the nearest satellite markers, I used the University of California Santa Cruz (UCSC) genome browser to note the location and length of the markers in base pairs. At the end, I had a start and stop satellite marker and the range of base pairs that ran between them. The genes that may influence the trait of interest, are located within this range of base pairs. Once I had the area of interest narrowed down to base pairs, I used the National Center for Biotechnology Information (NCBI) website to locate the genes within this range of base pairs. The NCBI website allows you to type in a range of base pairs and it gives you all of the known genes contained in that region. Once these genes are known, the scientists at Texas Biomed can keep them in mind for follow up studies that test whether those specific genes are associated with traits of interest.

I compiled all of my results into tables suitable for publication in a scientific journal; however, the results of this study will not be published for quite some time. In the future, the tables will be incorporated into journal articles that discuss what information we discovered from my analysis of obesity and inflammatory factors in families from the SAFHS.

My internship at Texas Biomed allowed me to see how scientific research is conducted across multiple fields such as genetics, epidemiology, and anthropology. I learned that identifying the genetic causes of disease is a complex and extensive process that often takes years to complete. This internship allowed me to utilize my
knowledge of genetics and better understand the usage of statistics in population studies focusing on ethnic minorities and their cultures.

My internship at Texas Biomed has encouraged me to pursue graduate studies in epidemiology and medical anthropology. I plan to attend graduate school in the near future. An internship at Texas Biomedical Research Institute provides an opportunity to participate in the research process and learn how cultural aspects affect epidemiology.
Works Cited


