**Analysis of Protein from the Thousand Genomes Project**

**Kaitlin Lisle, Kymberleigh Pagel, Alex Wu**

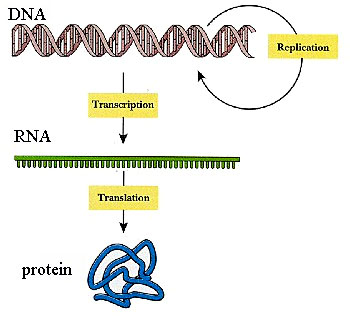
**Abstract**

Each person is born with their own specific genetic variations, which is called deoxyribonucleic acid, or more commonly known as DNA. In studying these variations, scientists can determine how each variation results in a specific human trait or characteristic. However, the process of deciphering these variations has been notoriously difficult since the methods needed have historically been very time consuming and expensive. Recent technological advancements have allowed for vast improvements in this field, resulting in the availability of the populations’ genetic information at a fraction of the original cost.

The 1000 Genome Project is the first project of its kind to take a significantly large number of participants’ DNA information and publish it publicly to the world in order to provide a detailed source for sequencing genomes. It is a collaboration of numerous international research teams, with the end goal of finding the most variations in at least 1% of the population studied. They enlisted people from areas across the world in order to incorporate specific regions and cultural differences, and to ensure that they are not using a homogeneous population. In doing so, they have provided the opportune data set for our project to elaborate on.

**I. Introduction and Statement of the Research Problem**

The human genome is composed of DNA. From here, the DNA can be turned into RNA, or ribonucleic acid, through a process known as transcription, and finally RNA can be translated into proteins. Proteins are extremely important in maintaining the life, health, and longevity of the human body. They are essential pieces of machinery that depend upon the stability of genetic code to be produced. DNA also helps configure the specific characteristics within a person. This could mean the differences between how diseases impact a person to smaller mechanics like freckles and allergies. The importance of proteins and DNA are not to be underestimated, as they are essentially what makes each person unique and different than the person next to them.

    In order to understand the basic biology of our project, we must first explain the specific ways in which scientists are able to configure protein placement from a sequenced strand of DNA. It is a three step process between DNA (nucleotides), RNA (nucleotides), and the final protein strand (amino acids).   
Image 1. The Central Dogma  
  
    The image above illustrates the process DNA must go through in order to determine the resulting protein. The gene strand first goes through transcription to go from the double stranded DNA to the single stranded RNA. Both DNA and RNA are composed of nucleotides, except RNA substitutes the nucleotide called uracil for thymine. Then the RNA is translated into a stand of amino acids, which folds into a protein. RNA is produced by copying the DNA nucleotide chain and inserting a uracil in place of the thymines. Once the DNA is transcribed into an RNA strand, the code for creating proteins has been created. The synthesis of proteins is called translation. Every combination of three nucleotides in the RNA chain is called a codon, and each codon can be mapped to a specific amino acid. By chugging through the string of codons the body produces a string of amino acids. Once the string of amino acids is complete it folds into its own unique shape and then is let loose into the cell to perform its functions.

A slight difference in the DNA chain, even just one small part, can have a monumental effect. If a certain area of the genome which happens to be a region that codes for a protein is damaged, it could result in a mutant protein or even stop the protein from being produced entirely. For example, phenylketonuria is an example of a disease caused by protein malfunction. People who suffer from this disease are missing from their genome a very important enzyme, a type of protein, which is vital for the breakdown of meat. Unless they avoid these types of food for their entire life then they will suffer as their bodies slowly accumulate massive amounts of unneeded meat byproducts in their cells which can lead to mental retardation.

Information about proteins can at times be inferred from the human genome, but we don’t always have access to enough data to draw these conclusions. A solution to the problem would be to look at the data provided by the Thousand Genomes Project. The goal of the Thousand Genomes Projects is to produce a list of all the genetic variants which are present in at least one percent of the human population.  At first glance, one percent does not seem like a very large amount, but at one percent these variants would still affect millions of people. In order to save time, the genomes which are sequenced by the project are done only at a more basic level. This is accomplished by breaking each person’s genome down into many pieces and sequencing each separately. According to the project, in order to completely sequence a person’s genome they would need to repeat their process 28 times. However, they are only doing it 4 times, which is enough to find the 1% of genetic variation. The unfortunate side effect of this process is that less than one hundred percent of each individual’s genome is sequenced. Although not every genome of every person is available for all individuals in the study, the amount of data that is provided by the Thousand Genomes Project is still a valuable resource for studying human genetic variation. The information published by the project is freely available for download online, which can be accessed through their website.

With the DNA sequence provided by the Thousand Genomes Project, we are able to decode it into strands of protein that we need. Thanks to this advancement in the project, the data sets provided by the Thousand Genome Project seemed to fit perfectly into our plan. We thought this project would be interesting, as it would allow us to determine the difference between proteins within people from around the world. More specifically, we thought that if we took two different populations like those of European decent and those of Japanese descent, we could visualize a difference alongside the already existing cultural differences.

The goal of this project was to evaluate the data provided by the Thousand Genomes Project in relation to protein function to see if there were significant differences in protein structure across populations. The Thousand Genome project provided the perfect resources for our research because it provided a large data set that yielded the most accurate results. When more volunteers are included as part of the data set, then more concrete and broadly applicable results can be formulated. This type of work is important because it can lead to the discovery, understanding, and treatment of diseases, as well as other health problems resulting from protein insufficiency or incompatibility with intended functions. Additionally, it helps us to better understand the human body and can shed some light on human population genetics.

**II. Background and Related work**

The Thousand Genome project within itself is an ongoing field of research. Scientists are still trying to sequence many more participants in the hopes that they can establish a more concrete database of genomic data. The goal of the project is to pinpoint and understand all the different types of genetic variation. The hope is that with this data, it can be easier to bridge the gap between genotype, the genetic code, and phenotype, which is the physical expression of the genotype. The hope is also to see how the gene is expressed in the body. This attempt is complicated by the fact that not every gene has a distinct correlation to a phenotype. It could be that many genes are combined in order to produce a genotype or that some genes do not produce a genotype that we have recognized yet. By evaluating the diverse variations within particular portions of the genome, or the lack thereof, conclusions can be made about the importance and function that region.

Genome sequencing has been expensive and time consuming, but there have been recent developments which have greatly increased the speed. An alternative to complete genome sequencing is to preferentially sequence select portions of the genome or to instead run sequencing in multiple smaller segments. These two factors have lead to a great boom in the amount of genetic data which has been available to use in research for the past decade. As the process of sequencing gets less financially and temporally exhaustive we can expect to see far greater amounts of genomes sequenced which will result in more data provided.

There have been many diseases’ genes that have discovered because of the information published by this project. By comparing the diseases of known donors and their genetic mutations to other unaffected individuals, it becomes possible to track down the genes that are responsible for the disease. The data provided by the Thousand Genomes Project has led to the discovery of the base mutation rate in humans. One of the methods used in the sequencing of genomes has been to take donors in triads. Each of the triad groups is composed of two parents and one child. The base mutation rate for humans was determined by evaluating the genetic relationships and variance within each of these triad groups.  It was determined that per generation, one in every one hundred million base pairs will end up having a mutation.

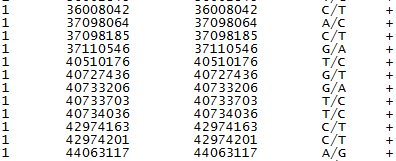
As their research expanded, they were able to provide basic data for other research projects in the same vicinity as ours. These groups can strive, as we have, to grasp a more thorough understanding of the differences between populations in accordance with their lineage and global setting. There is a lot potential in this data but it must be analyzed in order for it to actually mean anything. Our project took the Thousand Genomes resource and tried to make sense of it on a protein level, but there are several different avenues that this kind of research can take. We hope that our particular project can serve as a base for other projects that can go more in depth than ours.

**III. Research Methodology**

Once our data sets were chosen and our input was created from the preexisting data, we found a program that would evaluate the genetic mutations to see if they would have any real effect on the production of proteins. The program produced an output which cataloged all the different variations and where they were relative to genes. Although many of the mutations were shown to be completely harmless in terms of protein production, not all of them were. By looking at the mutations within each individual within the data set, we were able to determine the percentage of mutations which would have a negative impact. The output information was mapped back to the original data to determine which individuals were affected by which mutations. We then calculated the percentage of mutations that were negative and used it in our analysis.

The first step of our project was to determine which data set from the Thousand Genomes Project would be the best to use for our purposes. There was exon data, or triad data. Triad data is composed of individuals in sets of three: two parents and a child, while exon data consisted of data that was part of the coding region in DNA. Beyond that, exon data also contained information from all the participants, and was more general than the triad data.. We chose to use the exon data rather than the triad data because we thought it would yield more interesting and statistically significant results. We used the Japanese and European data sets for our project because we were limited on time and there seemed to be a clear contrast between groups. In order to sequence one person it would take a long time. We used the sequenced genome of ninety European participants, and one hundred and five Japanese participants. The more samples that we have from each region, the more accurate our analysis will be since we are trying to determine our results in reference to lineage in particular. We parsed through the data provided by the project for Europeans and Japanese and removed superfluous data in order to create the input files for our program.

**1. Input File**

**Chromosome    Start Location       End Location           Change        Strand** 

Above is a sample input of the stripped data strand we ran in our virtual machine, or VM. The necessary input data for the program includes the chromosome number, start position, end position, mutation, and strand. The human genome is split up and stored in 23 different chromosomes all of which are different and contain information vital to the mechanisms of the human body. Giving the chromosome number to the program allows it to decide which portion of the DNA it is going to be looking at. The start and end positions of the mutations reveal to the program the exact location of the mutation along the nucleotide chain. Because the mutations we looked at were only the replacement of one nucleotide with another, also called single-nucleotide polymorphisms or SNPs, the start and end locations were the same. The mutations (e.g. C/T) represent the change from one nucleotide to another. In the figure above, the first change is a transition from a Cytosine to a Thymine. Since DNA is composed of two identical strands, the strand refers to which of the two the mutation is affecting. Since these are all SNPs, it was sufficient to input for all cases that it was the same strand.

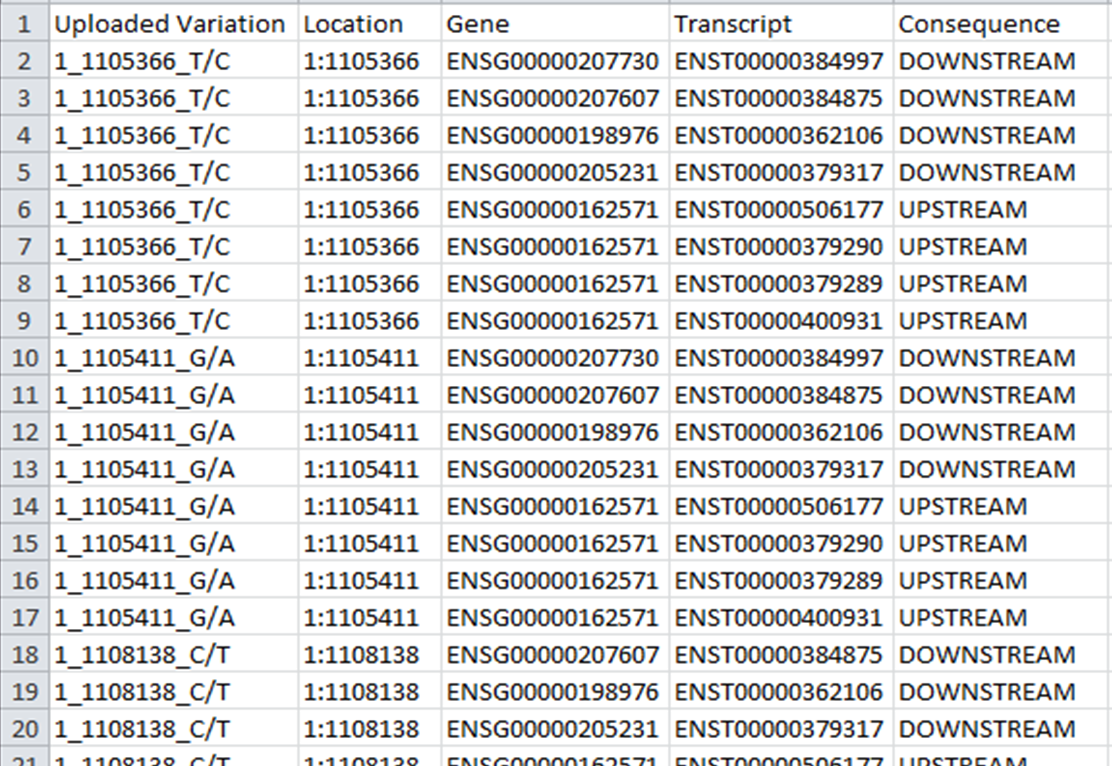
The program we used is a variant effect predictor script file written in perl. The program takes an input of a list of variants and their exact positions, as shown in the figure above. After being given this input the program then references the Ensembl Variation database to match the exact variation to how it might affect the genetic code. The script returns an output in the form of a table that shows the possible effects of the variation on the genome. Most importantly the data provided in the output can be linked to specific gene regions.

Unfortunately, even running a few lines of input in this program took a few minutes. Our input consisted of several hundreds and thousands of lines. Running this much data (each person’s genome would take a significant amount of time to run and we used hundreds of people), would take much more time than this projects term allotted, so we figured out a way to bypass a lot of that time. We decided that in order to successfully gather results, we would need to use parallel processing, which would significantly reduce the amount of time we spent waiting for results as it would allow us to run our data concurrently.

We created a VM or virtual machine in order to solve the problem of having a large amount of data, and we used a perl program in order to run the input we had created to obtain our desired output. We created an account on the FutureGrid program, which allowed us access to their supercomputers. This access gave us the capacity to run our program through different virtual machines, but on the face of one. The VMs allowed us to be able to run our programs on the clusters at school from virtually anywhere with internet access, such as the comfort of your own home, or at the hotel of a nice beach. Specifically, we utilized the machine in the system that is called India. India, shown with our group in the image below, is a cluster which has 50 nodes, each of which can support up to 8 VMs. With the aid of the Future Grid systems, we ran the VM on a cloud computing system, which allowed for the specific data sequencing sets to be processed simultaneously. Once completed, the output file would need to be uploaded to a remote machine, and then finally the protein differences could be analyzed for its significance.

 Image 2. Our team in front of Big Red

**2. Output File**



This chart above represents an example of the output file that was given once the VM was able to successfully run the input through the specified program. The first column of the output shows the chromosome number, location, and mutation. The second column shows the location of the mutation within the nucleotide chain. The third column tells you the ID of the gene that was affected by the mutation. The fourth column is the ID of the specific transcript affected by the mutation. The fifth and most important column shows you the consequence of the mutation. Upstream and downstream results mean that the mutation was not in the coding region, so it most likely did not affect the production of the protein. The consequences we were interested in are: Essential splice site, Stop gained, Stop lost, Complex in/del, Non-synonymous coding, Frameshift coding, Splice site, Partial codon, Within mature miRNA, and NMD transcript.

A mutation in an essential splice site would most likely change how the portion of DNA which creates the protein is separated from the genome. This could mean that some of the protein making code is deleted or that some material has been added unnecessarily. A stop gained or a stop lost can significantly affect the protein by increasing or decreasing its length. Adding on extraneous length can kill the protein in the same way that cutting out the functional parts can. A complex in/del refers to an insertion or deletion of a string of genetic data. A nonsynonymous mutation changes one of the protein building blocks into something different. A frameshift mutation augments the mRNA code which changes every building block that makes up the protein. A mutation of this type can completely change the physical structure of the protein.

We utilized several resources throughout the course of this project. Credit should be given to the Thousand Genomes Project for providing the exon mutation and for providing population data which were analyzed in this study. Futuregrid is credited for the use of their servers, because without their resources our project would not have been possible. The program used in this project was the Variant Effect Predictor which was obtained from ensembl.org**.**

**IV. Analysis**

Our process of analysis took us back to the work that was already done by the Thousand Genome Project, since we needed to pinpoint the mutations of each individual. A portion of the original data provided from the website contained a list for every individual of the populations which indicated the presence of those mutations within their genomes. Every person had two chances to have a copy of each mutation, some people had two copies of the mutation, others had only one copy, and many had no copies of the mutation at all. By mapping our output data to this data that was already established, we were able to determine the relative amounts of harmless genetic mutations and deleterious mutations for each person.

It is possible for an individual to have the mutation while at the same time not having the mutation. Since humans are diploid organisms, we all have two copies of our DNA. One copy we inherited from our mother and the other copy came from our father. An example from the data which would indicate the presence or absence per mutation per individual would be 1/0. The one on the left side of the slash indicates that the first copy of DNA has this mutation. The zero on the right side of the slash indicates that in this copy of DNA does not have this specific mutation. If a person is 0/0 then it means that neither copy of their DNA has the mutation. Alternative if the person is 1/1 it would mean that both copies of their DNA have the mutation.

We took the list of mutation consequences from the output of the program and applied them towards the data provided by the Thousand Genomes Project.  By taking the count of negative mutations overall and dividing by the total number of mutations we could determine the percentage of mutations which were detrimental.

**V. Results**

|  |  |  |
| --- | --- | --- |
|  | **European** | **Japanese** |
| Average | 10.3% | 11.4% |
| Standard Deviation | 1.02% | 1.31% |
| Minimum | 6.91% | 7.98% |
| Maximum | 12.4% | 14.8% |
|  |  |  |

Table 1.

The table above, Table 1, shows the distribution of deleterious protein mutations amongst all genomic mutations annotated in the input file per individual. These are very interesting numbers, because although they are in the same neighborhood they are very different. The averages show that only about ten percent of mutations have a significant negative effect on protein synthesis. There is a 1.1% difference in the averages of the two populations. The standard deviation for both populations is small, showing that in general there was not a lot variation between individuals on the proportion of negative mutations. Both the minimum and maximum of the European sample data are lower than the Japanese. The maximum for Japanese is three standard deviations above the average but for Europeans the maximum is only two standard deviations higher. Likewise, the gap between minimum and average is larger in the Japanese set, four standard deviations, than it is for the European set, three standard deviations. This, along with the thirty percent higher standard deviation, shows that there is more variation in the Japanese data set than in the European set.

In order to determine whether or not the differences specifically between populations were significant, we decided as a group to use the statistical student’s T-test. The students T-test is a statistical measure used to compare the means of two samples. The output of a student’s T-test is a p-value. The p-value represents the probability that we could obtain these results at random, meaning that there is NOT a significant difference between the two sets. If the p-value is low then it means that there is a very low probability of the two sets being similar. Generally the cut off for a p-value is .05 or .01, at which point you can say that the two sets are statistically different.

In this case, we compared those of the European participants and the Japanese participants. After running the student’s t-test we got a result of p=3.08848x10^-09. A p value this low indicates that the Japanese have a significantly higher likelihood of having protein mutations within their genetic makeup. It is far lower than the recommended .05 or .01, so we can be very certain that these results are significant.

We had not anticipated results that would be this clear cut. Our assumption was that there would be a very slight difference, if at all. It might be due in part to the fact that we selected two groups that had very diverse backgrounds throughout the course of human history. Although it seems like in many cases humans are very similar across the world, we can see that this is not the case. This brings up very many interesting questions, and a lot of additional research that could be done to determine the real world effects of these genomic differences.

**VI. Conclusions**

**1. Limitations of our method.**

When presenting these results it important that we should discuss the limitations of our particular method and how that might have had an effect on our results.  It is difficult to do an analysis which is completely accurate. Unless the sample size happens to be one hundred percent of the population we cannot claim that our percentages will apply to all individuals within the population, however, our sample size can be large enough that we can say that it is statistically significant, and can say with a certain confidence that it affects the population.

Beyond that there are ways in which our method can be improved to produce more accurate results. The first is that the sample size might not be large enough. If the sample size were one thousand or ten thousand individuals then it is very likely that the mutation distribution could have been vastly different which would yield a different result. A sample size of only about one hundred is not generally considered a sufficient amount of individuals to justify very broad conclusions, especially when the population in question is so large.

Additionally it needs to be taken into account that these numbers only reflect the mutations in a small subset of the coding region of DNA. The genome is billions of base combinations long, our analysis only involved several thousands of bases. There are several more genes which were not sequenced and could have a different distribution as well. It also needs to be emphasized that we selected mutations which were perceived as the most likely to change the protein. In general genetic mutations produce negative results for the organism. Some, albeit a very small percentage, of the mutations we retained probably did not have a negative effect on the individual. Unfortunately without more in depth coverage and analysis this could not be determined.

**2. Understanding the results.**

There are several possible reasons to explain why the rate of functional mutations is higher in the Japanese population, all dealing with evolution and population genetics. The three reasons that we address are merely possibilities and our speculation. In order to be more conclusive, more specific studies of Japanese population genetics would have to be conducted in order to determine which is the most viable. It is possible that there are other forces we have not considered that could affect the mutation rate in Japan.

The first possibility is that there is less migration within the Japanese population. Japan is a relatively secluded island and historically has not been subject to a lot of contact or interbreeding from different populations of the world. Breeding within a population over the course of time can lead to the accumulation of deleterious mutations. The accumulation takes place because there is less genetic variation to mix up within the gene pool. This is somewhat of an oversimplification, especially considering the population size of Japan, but we believe the principle can be applied because this is a very long period of evolutionary development. While the population size is not small enough to produce drastic results in terms of genetic mutations, there could be some subtle effects present in the genome as a result of this county’s genetic stability.

The second possibility that we considered is that without strong selection, there would be less reason for those detrimental mutations to be removed from the gene pool. In human societies the principle of “survival of the fittest” does not always apply. This could at least be partly due to the incredible medical advances in the past few hundred years. In the past, diseases which used to kill of thousands of people now only kill hundreds, or even less, and many chronic illnesses can now be treated. If the collective population of an area is not exposed to stresses then there is less reason to adapt. If people can still survive and reproduce even though they have the mutation, then it will persist in the population throughout their future generations. Since Japan is relatively secluded it could mean that historically they were exposed to fewer of those diseases which could have killed off a greater amount of the more physiologically vulnerable individuals.

The third possibility is that these mutations have not yet been fixed within the Japanese population. Fixation is a concept within population genetics which refers to the tendency of mutations to be removed from the gene pool over time. There are several properties of a mutation which can lead to its fixation, like population size and heritability. The process of fixation can take a very long time which can help to explain why these mutations present in the gene pool.

**3. Future Implications of Research**

Analysis of only two populations opens to the door to comparing several different populations at the same time. The data of genomic sequencing is available for Europeans, East Asians, West Africans, and Americans through the Thousand Genomes Project. Taking one step beyond our plan would be to look at the diseases or protein disorders which commonly affect each sub-population. An evaluation could be done on a population level to look at the differences between disease prevalence and see if any can be correlated with a particular gene mutation. Another avenue would be to analyze the mutations on a person by person basis. It would be great to look at the diseases and mutations of each individual donor to the Thousand Genomes Project. Hopefully with enough data there are a few diseases that can be linked to one gene if not to a specific gene mutation.

**Acknowledgements**

We thank Andrew Younge and Fuxiao Xin, our mentors, for helping us with finding and using tools and for coaching us on how to be a better team while utilizing our individual strengths. We would like to thank the FutureGrid systems for allowing us to utilize their machinery and the Thousand Genomes Project for publicising the data. We would also like to thank Geoffrey Fox and Maureen Biggers for giving us the opportunity to work on this project.

**References**   
The Thousand Genomes Project Consortium. "A Map of Human Genome Variation from Population-scale Sequencing : Nature : Nature Publishing Group." *Nature Publishing Group : Science Journals, Jobs, and Information*. Nature, 28 Oct. 2010. Web. 30 Apr. 2011. <http://www.nature.com/nature/journal/v467/n7319/full/nature09534.html>.

Durbin, Richard, David Altshuler, Lisa D. Brooks, Adam Felsenfeld, Jean McEwen, and Paul Flicek. "About | 1000 Genomes." *Home | 1000 Genomes*. NHGRI, 15 Nov. 2010. Web. Spring 2011. <http://www.1000genomes.org/about>.  
  
"Home Page | FutureGrid Portal." *Home Page | FutureGrid Portal*. NSF, Lilly Endowment, Indiana University Pervasive Technology Institute, 15 Sept. 2010. Web. Spring 2011. <http://futuregrid.org>.