11-21-2020

My Genetic Code and What It Says About Me

Josiah Christian Lane

My Genetic Code and What It Says About Me

by

Josiah Christian Lane

An Honors Capstone
submitted in partial fulfillment of the requirements
for the Honors Diploma
to
The Honors College
of
The University of Alabama in Huntsville
11/21/2020
Honors Capstone Director: Professor Judy Cooper
Lecturer, Biological Sciences Department

Josiah Lane		11/21/2020
Student	Date	
Judy E. Cooper		11/21/2020
Director	Date	
Paul Wolf		11/21/2020
Department Chair	Date	
William S Wilkerson	Digitally signed by William S Wilkerson Date: 2020.11.30 12:28:59 -06'00'	
Click or tap here to enter text.		Click or tap to enter a date.
Honors College Dean	Date	

Honors College
Frank Franz Hall
+1 (256) 824-6450 (voice)
+1 (256) 824-7339 (fax)
honors@uah.edu

Honors Thesis Copyright Permission

This form must be signed by the student and submitted with the final manuscript.
In presenting this thesis in partial fulfillment of the requirements for Honors Diploma or Certificate from The University of Alabama in Huntsville, I agree that the Library of this University shall make it freely available for inspection. I further agree that permission for extensive copying for scholarly purposes may be granted by my advisor or, in his/her absence, by the Chair of the Department, Director of the Program, or the Dean of the Honors College. It is also understood that due recognition shall be given to me and to The University of Alabama in Huntsville in any scholarly use which may be made of any material in this thesis.

Josiah Lane

Student Name (printed)

Josiah Lane

Student Signature

11/21/2020

Date

Table of Contents

Abstract

We all have our own unique genetic code and as a microbiology major this is especially interesting to me. My capstone project focused on analyzing my genetic code after having it sequenced by the company Nebula Genetics. I took this information and checked it for certain segments of code that corresponded to physical traits. This was accomplished by writing a program in Bash using online tools and my own work. The majority of the work was researching the segments of code using online databases and coding the program to analyze what it meant. I had expected to learn a lot about myself through this project and hope I can put this knowledge to good use.

Introduction

There are over 3 billion nucleotides making up more than 20,000 genes in the human genome across 23 pairs of chromosomes (1). Nucleotides are the building blocks of the human genome, when stacking together they form strands of DNA, genes, and chromosomes. The Human Genome Project was completed in 2003, but the work on the human genome was far from over (1). As scientist continue to narrow down the number of genes, this project was undertaken to look at just a few of the genes that have already been identified over the years.

This project has several facets to it including research into the genes themselves, research into which computer language was best for analyzing the genomic data, which company would be best suited for gathering the raw genomic data, deciding which secondary tools would be useful, and putting the program together. The project was designed to be primarily research into genes with a program that could compare the nucleotide sequence of a gene with the genomic data. The work was done primarily over the summer of 2020 with the final work done in the fall of 2020.

Chapter 1: Nebula Genetics, Research, and Coding

This project began with setting out a plan. First, a company had to be selected from which to receive the genomic data. Second, research would have to be conducted throughout the project. Using the National Center for Biotechnology Information, NCBI, database, search terms were entered, and the results were examined for useful genes. The final phase was writing the program itself and deciding if any secondary software packages were required. Initially, Perl was proposed to be the programming language of choice, but this was changed to installing Ubuntu, a Linux distribution, in order to write the program as a Bash script.

Nebula Genetics

There were several options for having my genomic data collected and the results delivered. Nebula Genetics was selected because it was a reasonably priced alternative to some of the more expensive options. Nebula Genetics was founded by George Church, a Harvard and MIT professor, in 2018 (2). There were two options, the .4X and the 30X test kits. The .4X kit tests the 40% of your genome where most of your genes that create your physical characteristics reside and the 30X kit tests your entire genome thirty times. Initially, only the .4X kit was ordered and much of the work was done with only this option available. During the later segment of the project the 30X kit was purchased and results were received at the end of the summer. This meant that most of the project was conducted using the data from the .4X and was rechecked at the end with the 30X.

Research

The next phase of the project was conducting the research. This was started at the beginning of the summer and continued throughout the project. The NCBI database become the

primary resource, though early attempts at collecting data did not take advantage of this resource and produced no noticeable results as many results would list the genes but not the nucleotide code. Search results from the NCBI database included the genes name, chromosomal location, nucleotide sequence and provided all of the genes looked at throughout the project.

One gene looked at was interleukin 13, otherwise known as IL13 (3). This gene is related to asthma and can be found on chromosome 5. OCA2 melanosomal transmembrane protein, OCA2, is found on chromosome 15 and is related to brown or blue eye color (4). Melanocortin 1 receptor, MC1R, is related to blond hair and can be found on chromosome 16 (5). Tyrosinase related protein 1, TYRP1, is involved in skin tone, defective ones can cause albinism, and this gene can be found on chromosome 9 (6).

Coding

Creating the program was begun after research had been ongoing for several weeks. Perl was the initial choice because of familiarity with it and its use as a biological coding language. The choice was made to install Ubuntu to create a Bash script instead because of the ability to use tools such as Samtools. This was a software package that would allow for formatting of the files received from Nebula Genetics. The files received were in the formats of BAM, Binary Sequence Alignment/Map, and BAI, which are index files for their respective BAM files. BAM files are compressed SAM, Sequence Alignment/Map, files. These BAM files require the use of Samtools to change them into SAM files, which are a readable format. Samtools also require the BAI files to correctly convert the BAM files into the SAM format.

Chapter 2: Problems and Solutions

The initial problem, which involved how to find an efficient way to conduct the research aspect of the project, was solved by using the NCBI database. Searching the internet for generic terms such as, “genes for hair color” or “genes for asthma” sometimes resulted in finding the genes, but not the nucleotide code that was required to use with the program. The NCBI allowed for more generic search terms such as, “human, hair” or “human, asthma”. Other problems also began to arise as the coding got underway, including the challenge of learning to write in Bash and how to use Samtools. The problems that occurred over the summer included: not being able to install Samtools, not being able to reformat the genomic data, not being able to read in the nucleotide sequence for various reasons, and not being able to program for things such as mutations in the genetic code. A foundational problem with the project was that the presence of a gene did not indicate the gene was active. This meant that having a gene, such as related to asthma, did not mean you had asthma. The program was unable to tell the difference between active and inactive genes which meant the program could not return meaningful data.

Not being able to install Samtools was solved after switching to Bash, instead of Perl, and moving some file directories from the Windows OS to the Linux shell. Using a new programming language often results in many errors in coding and this caused many of the delays in reformatting the genetic data. These mistakes were corrected in due time by learning more tips and tricks of Bash. The program code only received the nucleotide data when it was imputed from the keyboard. This was quite an issue because the length of genes can be in the thousands of nucleotides, and there was no confidence in being able to accurately input one gene let alone testing several. Because of using the Linux shell, copying the data into the program was not possible, and the program was unable to compare data from one file to data from another. The

program was also unable to have the possibility of mutations coded into it because that would require allowing for the genetic code to not line up with the reference genome. These problems with inputting data and accounting for mutations meant the program could work only on a very small scale.

The program did accept lines of code imputed from the keyboard and could accurately check the string of code against the reference genome, but strands of code hundreds of nucleotides long offered no confidence of being imputed correctly. Because the program would not give the desired results, another solution was needed. Using IGV, the Integrative Genomic Viewer, a free software used for visualizing genomic data, the genomic data from Nebula Genetics was able to be reviewed. The previously mentioned genes were all able to be located. The premise of this project was flawed because not being able to account for mutations or whether genes were active or inactive meant there was little to no chance of success as originally envisioned.

Conclusion

This project looked at only four of the more than 20,000 genes but being able to examine my own genetic code was fascinating. Nebula Genetics may have taken longer than was expected in regard to returning the data, but the kit was remarkably easy to use. Research had its lulls but was successful in finding nucleotide sequences to search for. The program ultimately did not work as anticipated, but I found it to be an engaging project that reminded me of the enjoyment I find from coding. Not being able to account for the active/inactive status of a gene or for mutations meant the program could not work successfully.

I find myself amazed at the depth of knowledge we have accumulated on the human genome. Even with our knowledge, there is still so much more to be discovered and cataloged. The fact an undergrad student can spend a summer researching, coding, and using freely available software to look into their own genomic data is a testament to how far science has come. I greatly enjoyed the experience of the research and coding. I hope to take what I have learned from this project and use it to continue conducting research throughout my career in the biological sciences.

References

1) Human Genome Project- https://www.genome.gov/human-genome-project/Completion-FAQ

2) Nebula Genetics- https://nebula.org/george-church/

3) IL13 gene- https://www.ncbi.nlm.nih.gov/gene/3596

4) OC2A gene- https://www.ncbi.nlm.nih.gov/gene/4948

5) MC1R gene- https://www.ncbi.nlm.nih.gov/gene/4157

6) TYRP1 gene- https://www.ncbi.nlm.nih.gov/gene/7306