Mining human genetic variation from the 1000 Genomes Project

Abstract

The 1000 genomes project aims to identify genetic variations that are present in at least one percent of a certain population and the project has had a great impact on our knowledge about genetic variations and how genetics affect human health and disease and what factors that create the genetic diversity. This project is focused on the genomic region between 1:114,356,433-114,414,375 on chromosome 1 where the gene PTPN22 et.al. are located. The gene encodes a protein that is inhibiting T-cell activation, and mutations in this gene can be connected to different types of autoimmune diseases, for example diabetes type 1 and rheumatoid arthritis. Programming is a powerful tool when dealing with large amounts of biological data so by combining knowledge about genetics and programming in python we can analyze the PTPN22 gene and its variations. The dataset is in the Variant Call Format and includes the gene region from 2504 individuals from 26 different populations in Africa, America, East Asia, Europe and South Asia. The data was processed by several algorithms in python to produce figures to visualize and analyze the variant-data and to compare the results between the different populations. This type of analysis can for example be used to gain greater understanding of how genetic variations can give rise to different diseases and this is very important since there still are many links between genetic diseases and genetic variations that have not been discovered.

Introduction

The 1000 Genomes Project is a global project which aims to identify all the different gene variants that are existent in at least one percent of a certain population. This would not only increase the knowledge about how different genetic variations are present in different populations, but also provide new information to for example help the understanding of how genetics affect human health and disease and what factors that create the genetic diversity. I.e. a very important asset to have when performing several types of studies. The project includes reconstructed genomes from (at that point) 2504 individuals from 26 different populations in Africa, America, East Asia, Europe and South Asia and are investigating both single nucleotide polymorphism and structural variants[1].

The data is presented in the Variant Call Format (VCF) and the population's genomes are compared to a human reference genome GH19. In VCF the columns represent the samples, starting at column 10, and the row represents the different variations. The sample columns contain data about whether the sampleID contains any allele of the given variant. The value 0 means that the sampleID has the reference allele and the value 1 means that the sampleID has the first alternative allele, so 1|1 means that the sampleID is homozygous for the first alternative allele, 1|0 or 0|1 means that the sampleID is heterozygous for the first alternative allele and 1|1 means that the sampleID is homozygous för the first alternative allele[2].

In this project the genomic region 1:114,356,433-114,414,375 for chromosome 1 will be analyzed, and especially the gene of interest PTPN22. PTPN22 is short for protein tyrosine phosphatase non-receptor type 22 and it is located on the human chromosome 1 and takes up around 10% of the genomic region. The gene encodes a cytoplasmic protein, often referred to as lymphoid phosphate or LYP, that inhibits T-cell activation[3]. It has been discovered that mutations in the PTPN22 gene can be connected to different types of autoimmune diseases, for example Type 1 diabetes and rheumatoid arthritis[4]. One type of single nucleotide polymorphism that changes the protein sequence is when cystine is substituted to thymidine at nucleotide 1858, leading to that the amino acid at codon 620 is substituted from arginine to tryptophan and that affects the binding of LYP and makes it much more active phosphate. Exactly how this mutation

---

[1] *A global reference for human genetic variation*, Nature VOL 526, 2015.
[2] *Variant Call Format*, IGSR: The International Genome Sample Resource https://www.internationalgenome.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-40/
[3] *PTPN22: Its role in SLE and autoimmunity,* S. A. Chung & L. A. Criswell, 2010 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2875134/
[4] *PTPN22 protein tyrosine phosphatase non-receptor type 22 [Homo sapiens (human)],* National Libary of Medicine, 2023 https://www.ncbi.nlm.nih.gov/gene/26191

leads to the different diseases are not determined yet[5] but it is clear that this type of mutation has been associated with many autoimmune diseases.

By using bioinformatics and various programming softwares we can analyze large entities of biological data and potentially find out more information about for example how genetic variations of PTPN22 can lead to the occurrence of different autoimmune diseases. The purpose of this project is therefore to discover different ways to analyze the genetic variations in different populations in the gene PTPN22 using programming and data from the 1000 genomes project.

Methods

The data for the genomic region was presented in the VCF format under the name "PTPN22.hg19_multianno.vcf" and information about the age, gender, and population for all the samples in the data set was found in the file "1000genomes_sampleinfo.txt". Python version 3 was used to analyze the data by formulating the appropriate algorithms.

Firstly the number of columns were counted to find the *total number of samples* in the VCF file for the genomic region. This was executed in python by splitting the header line on diameter into strings, excluding the first nine strings and then creating a list with the strings from the header line, where the strings are the sample names. The length of the list was determined and that gives the total number of samples in the PTPN22.hg19_multianno.vcf file. To reduce the number of outputs from the "read_vcf" function the total number of samples was determined in the code as the length of the dictionary of sample to number of variations since that dictionary has keys that are the sample names.

The *number of variants* in the VCF file is determined as the number of lines in the file, except the metadata lines and the header line. So by counting the number of lines that does not start with "#" we get the number of variants in the PTPN22.hg19_multianno.vcf file. To reduce the number of outputs from the "read_vcf" function the total number of variants was determined in the code as the length of the dictionary of positions to allele frequencies since that dictionary has keys of the same length as the number of variants.

To find the number of *nonsynonymous variants* in the VCF file the number of variants that is tagged with "nonsynonymous" is determined. The variants have been annotated with Annovar and by that they have the tag in the "info" column and the number of nonsynonymous variants is received by looping over each line in the VCF file and adding 1 to a count variable for every line

[5] *PTPN22: Its role in SLE and autoimmunity,* S. A. Chung & L. A. Criswell, 2010
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2875134/

where Func.refGene=nonsynonymous_SNV is found. To also find how many of the *nonsynonymous variants in the file that alter the amino acid sequence of PTPN22* another count variable is established and increased by one for each line where both Func.refGene=nonsynonymous_SNV and Gene.refGene=PTPN22 is found. The number of *nonsynonymous variants in PTPN22 is normalized by length* by dividing the number of nonsynonymous variants in PTPN22 by the length of the coding region for the gene.

To collect the *alternative allele frequency* in the entire 1000 Genomes population for every variant a function named allele_freq was constructed. The function takes a line (variant) as input and creates a dictionary where the key is the position for the variant and the value is the alternative allele frequency. The dictionary is then used to create a histogram for the alternative allele frequency on the x-axis and the number of variants on the y-axis, where the number of variants are converted into logarithmic scale for better visualization.

The collection and plot of the *number of variants per sample*, the *number of singletons per sample* and the *number of common variants per sample* is all done in a similar way. The function named x_per_sample is constructed to take a line as input and count the number of variants, number of singletons and number of common variants for that line (variant). The input is also three dictionaries that are updated in the function and then returned. The dictionaries are for the first line in the vcf file three dictionaries where the keys are the sample names and all of their values are zero. The value is then updated with one for the sample name when the variant has at least one copy of the alternative allele, singleton or a common variant.

To create a bar plot that shows the average number of variants per sample in each population the function x_per_sapmple_plot is created. The function takes the dictionary from the function x_per_sample as input together with one dictionary with the sample names as key and what population they tillhör as value and one dictionary of countries as keys and their continent as value, both created from information from the 1000genomes_sampleinfo.txt. Then the function loops over the dictionary from x_per_sample and for each sample the total value is updated in a new dict for that corresponding population to that sample, and one is added to another dictionary that later is used to calculate the average. A new dict where the country is the key and the average number of variants per sample is the value. Then a bar plot is generated by plotting the countries on the x-axis and the average number of variants per sample on the y-axis.

Results

The code 1000genomes.py gives the following output when executed with the files PTPN22.hg19_multianno.vcf and 1000genomes_sampleinfo.txt.

Total number of samples: **2504**

Number of variants: **25786**

Number of nonsynonymous variants: **337**

Number of nonsynonymous variants in PTPN22: **44**

Number of nonsynonymous variants in PTPN22 normalized by gene length: **0.018151815181518153**

The code 1000genomes.py also generate the following figures:

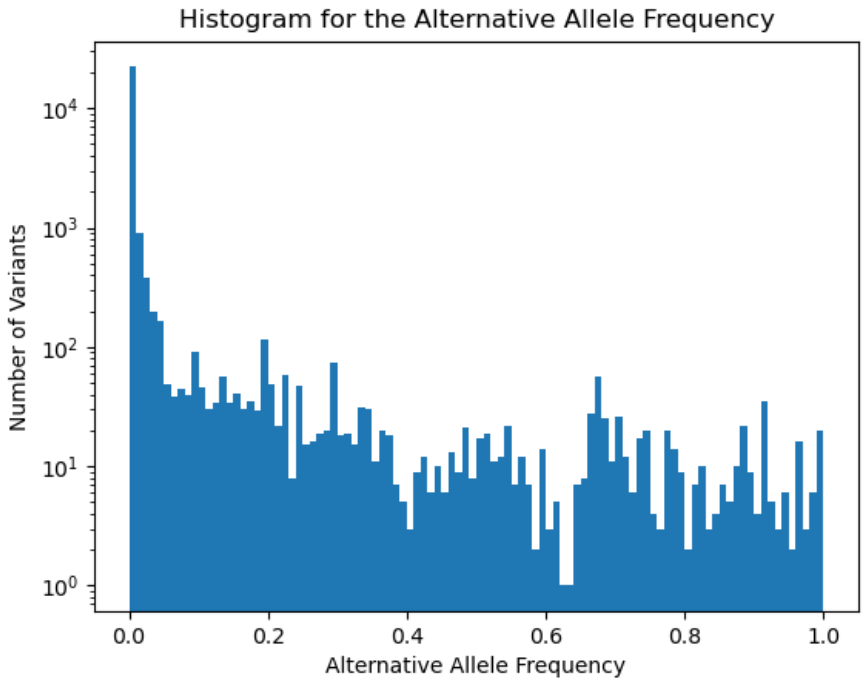

Figure 1: Histogram of the number of variants as a function of alternative allele frequency.

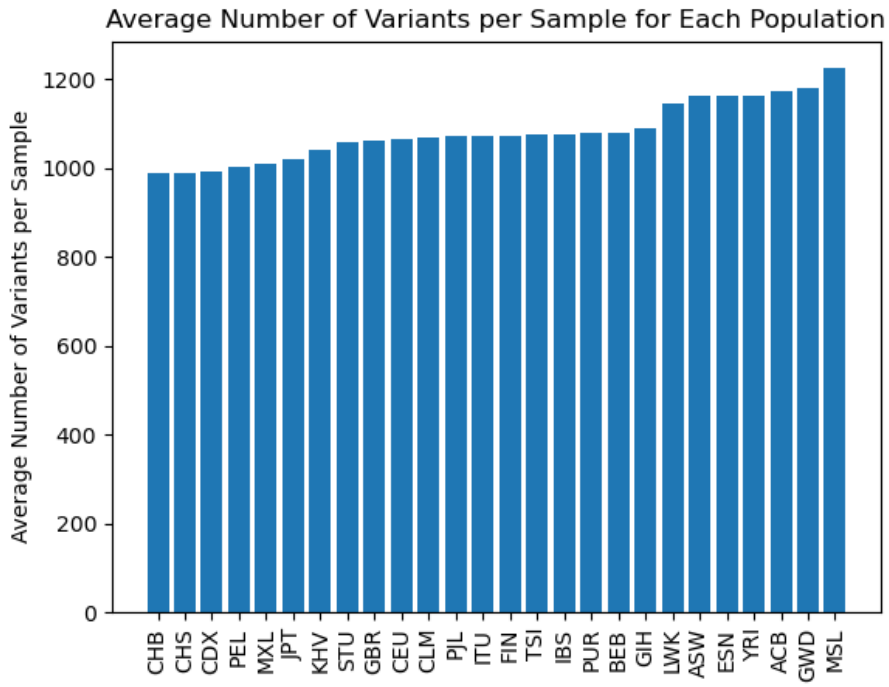

Figure 2: Bar plot of the average number of variants per sample in each of the populations.

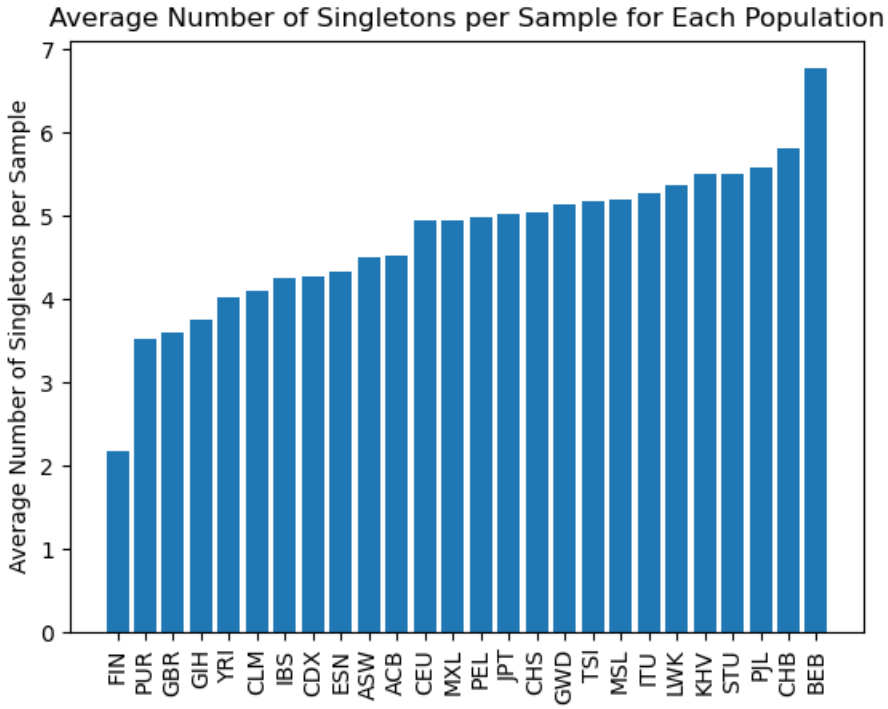

Figure 3: Bar plot of the average number of singletons per sample in each of the populations.

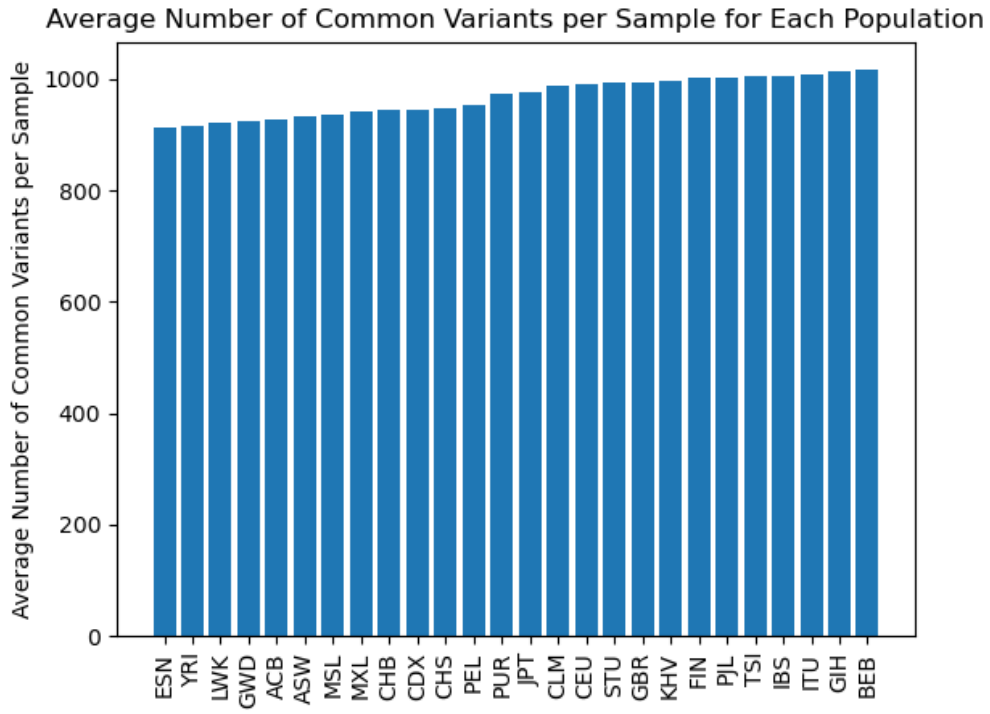

Figure 4: Bar plot of the average number of variants with an allele frequency over 0.05 per sample in each of the populations.

Discussion

When comparing the result with the result from the article "A global reference for human genetic variation"[6] one can see both similarities and differences between them. The Article is analyzing the entire genomes of the individuals while this project as mentioned only analyzes a genomic region which includes the gene PTPN22.

Regarding figure 1, the histogram for the alternative allele frequency, compared to the "Extended Data Figure 3A" in the article shows some differences, and since the dataset in this project is not as big as the one in the article, figure 1 shows a more irregular pattern. That could be caused due to the fact that we are plotting at a higher resolution. The main pattern is quite the same, where we have the significant highest number of variants for the alternative allele frequencies close to zero and then they are decreasing for higher alternative allele frequencies. One difference is that in the figure in the article the number of variants increases after 0.9 and around the alternative allele frequency close to one the number of variants reach a peak, and this peak can not be seen

[6] *A global reference for human genetic variation*, Nature VOL 526, 2015.

in figure 1. This could for example be due to the fact that there are differences between the genomic regions for the allele frequency, which seems likely.

For figure 2, the bar plot of the average number of variants per sample in each of the populations, compared to Figure 1A “The number of variant sites per genome” in the article shows pretty similar results. In general the article shows that Africa and America have the biggest amounts of variants per genome and that the European and East Asian have the lowest amount of variants per genome. For example MSL (Mende in Sierra Leone) showed the highest number of variants per sample respectively genome in both the article and this project, i.e. both for the entire genome and the genomic region. GWD (Gambian in Western Division, Mandinka), ACB (African Caribbean in Barbados) and ESN (Esan in Nigeria) also had high values both for the entire genome and for the genomic region for the number of variations. CHB (Han Chinese in Beijing, China), CHS (Southern Han Chinese), CDX (Chinese Dai in Xishuangbanna, China) showed the lowest amount of variations in the genomic region and that is also shown for the entire genome in the article. Some values that differed a bit for the number of variations between the entire genome and the genomic region is MXL (People with Mexican Ancestry in Los Angeles, CA, USA) and PEL (Peruvians in Lima, Peru) that had some of the lowest average number of variants per sample in the genomic region but for the entire genome they showed a bit higher values (or at least more differences between the individuals since the values where more spread out).

By comparing figure 3 in this project, the bar plot of the average number of singletons per sample in each of the populations, with figure 1.C in the article, the average number of singletons per genome, we can also see both similarities and differences. In the article the LWK (Luhya in Webuye, Kenya), GWD (Gambian in Western Division, Mandinka), BEB (Bengali in Bangladesh) and CHB (Han Chinese in Beijing, China) show some of the highest amount of singletons per genome and FIN (Finnish in Finland), PUR (Puerto Ricans in Puerto Rico), GBR (British in England and Scotland) and GIH (Gujarati Indians in Huston, TX, USA) show some of the lowest amounts of singletons per genome. For the genomic region the MSL (Mende in Sierra Leone) has the highest amount and MS also has one of the higher values in the article as well. The next highest amount of singletons in the genomic region was found in the GWD population which was also the next highest value in the article. The lowest amount of singletons in this project was found in STU (Sri Lankan Tamil in the UK) which differs a lot from in the article since it had one of the higher values there. Another difference between this project and the article is for FIN and PUR that have the lowest values in the article but for the genomic region they show a pretty average amount of singletons compared to the other populations. GBR and GIH that had some of the lowest amount of singletons for the entire genome also had some of the lowest amount of singletons for the genomic region.

Regarding figure 3, the bar plot of the average number of variants with an allele frequency over 0.05 per sample in each of the populations, shows mainly that most of the variants observed in the genomic region are common variants, i.e. have an allele frequency over 0.05 or 5%. This is in line with the article “A global reference for human genetic variation” where it is stated that “the majority of variants observed in a single genome are common: just 40,000 to 200,000 of the variants in a typical genome (1–4%) have a frequency ,0.5%”.

By investigating allele frequency distribution we can also collect important knowledge about human evolution. For example, the variants with the higher allele frequency have generally occurred earlier in the human evolution since they have had more time to become common in the population while many of the variants with low allele frequency tend to have occurred more recently in the human evolution. This is especially for variants that lead to harmful mutations since they do not tend to spread that much, but if a rare variant has an evolutionary advantage that can spread pretty quickly even and show high apple frequency even if they are newer than some other that do not have as high allele frequency. But in general does high allele frequency imply that they occurred earlier in the human evolution than the ones with low allele frequency.

Conclusion

To conclude, the 1000 genomes project is a great recourse for understanding genetic variations and how they differ between populations. This project has had the primary focus on the PTPN22 gene which is a gene where mutations could lead to several autoimmune diseases. By using our knowledge about bioinformatics, genetics and programming we have been able to analyze and draw conclusions about the gene regarding for example allele frequency distribution, number of variants and singletons. This type of knowledge of how to use bioinformatic tools and programming can be used to learn more about genetic variations and its part in different diseases and thereby improve medical science and human health.

References

A global reference for human genetic variation, Nature VOL 526, 2015. https://www.nature.com/articles/nature15393

Variant Call Format, IGSR: The International Genome Sample Resource https://www.internationalgenome.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-40/

PTPN22: Its role in SLE and autoimmunity, S. A. Chung & L. A. Criswell, 2010 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2875134/

PTPN22 protein tyrosine phosphatase non-receptor type 22 [Homo sapiens (human)], National Libary of Medicine, 2023 https://www.ncbi.nlm.nih.gov/gene/26191

PTPN22: Its role in SLE and autoimmunity, S. A. Chung & L. A. Criswell, 2010 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2875134/