

Table of Contents

[1. Background 2](#_gjdgxs)

[2. Overview 2](#_30j0zll)

[2.1. 23andMe 2](#_1fob9te)

[2.2. Parser for 23andMe 5](#_3znysh7)

[3. Project Goals 6](#_2et92p0)

[3.1. Our Aim/Goal 6](#_tyjcwt)

[3.2. How to achieve it 6](#_3dy6vkm)

[3.3. Purpose 8](#_gohba0jlj8o5)

[3.4. Databases Scope 8](#_1t3h5sf)

[4. Development for other Potential Data Analysis/Processing Options 9](#_17dp8vu)

[4.1. DNA.Land 9](#_3rdcrjn)

[4.2. Promethease 9](#_26in1rg)

[4.3. Genomelink 10](#_lnxbz9)

[5. Implementation 10](#_44sinio)

[5.1. What we did 10](#_rxj4s960ype2)

[5.2. What we struggled with 13](#_1ksv4uv)

[5.3. Improvement Cases 15](#_bcz0qqnl8q3j)

[6. Advancement in Technology 16](#_2jxsxqh)

[7. Future Trending 17](#_z337ya)

[8. Final Recommendations 17](#_3j2qqm3)

[9. References 18](#_1y810tw)

# Background

Single nucleotide polymorphism (SNP) is also known as a simple DNA variation that causes changes that can be transition or transversion type that occur in around 1 in 1,000 base pairs (Shastry, 2009). SNPs are responsible for genome evolution, common familial traits, and interindividual differences in drug response and diseases (Shastry, 2009). In molecular biology, SNPs can change the encoded amino acids known as nonsynonymous, can be silent known as synonymous, or can affect the noncoding regions which this can influence promoter activity where it will affect gene expression, mRNA conformation, and subcellular localization of mRNA or proteins which can result in disease (Shastry, 2009). The importance of determining the identity of these variations in genes including the analysis of their aftermath can help in better understanding of how SNPs impact gene function and their health (Shastry, 2009). There are a bunch of services who can sequence these SNPs and give you a report based on all SNPs that were found. There are now over 10 million SNPs that are being reported and contain limited genetic information due to the limited number of alleles (Lee, 2007). However, for people that do not have a science background can struggle on trying to understand their reports. One company that will be focused on is 23andMe and how UMGC can help in better understanding what these reports mean.

# Overview

# 23andMe

23andMe is a genetic sequencing company that will offer kits that people use to send out their DNA. Once 23andMe receives the kit, they will go ahead and extract the DNA and sequence it. Once sequenced, 23andMe will send out a file that consists of single-nucleotide polymorphism (SNP) variant information. At the height of their popularity, 23&me was valued as high as $6 billion with a whopping 14 million customers (Allyn, 2024). Each of these customers would have undergone this DNA analysis. Other companies have also attempted to provide similar services like ancestry.com who claims to have racked up some 3.5 million customers (Ancestry, n.d.). In every case, customers who undergo this DNA analysis receive an output from the company, generally with information regarding ethnic background and distant relatives. Some, like 23&me, also provide a health based output, but because of certain HIPAA restrictions, there is limited information they are able to provide. In all cases, these companies will provide their customers with their raw genetic data, should they be interested.

SNP analyses have been an emerging technology over the last few decades because they allow for the derivation of important genetic information without the requirement of doing full DNA sequencing. The most recent type of analysis done by 23&me uses an Illumina GSA V5 chip which is capable of identifying roughly 650,000 SNPs (Odake, 2023). While this number seems significant, consider also that over 335 million total SNPs have been identified in humans (Odake, 2023) and that the total genome size is a little more than 3 billion bases (Brown, 2002). This means our SNP analyses cover about 0.02% of the total human genome. The SNPs analyzed, however, have been selected based on variability and information gathering potential. Some regions of the genome may be variable person to person and the allele that is present may tell us something about health or genealogical background. When this is done 650,000 times, 23&me is able to give a summary about where someone's genetic lineage is from and maybe even who they are related to. So how does the SNP analysis actually work?

The GSA (global screening array) V5 chip from Illumina is a specific type of microarray chip. It contains hundreds of thousands of beads sitting in etched microarrays. Each microarray contains many copies of an oligonucleotide which is made to bind to a chosen locus in the human genome. As the customers’ DNA fragments pass over these arrays, complementary sequences will bind to the oligonucleotides. The oligonucleotides are made to end one base prior to the locus of interest. A single base extension is performed using a fluorescence linked nucleotide. After extension, each microarray is laser excited and the fluorescence linked nucleotides emit a specific wavelength that can be detected by a scanner (Illumina, 2012). This nucleotide will be complementary to the customer genotype so 23&me knows the actual genotype is the complementary base. Since each GSA chip contains hundreds of thousands of beaded microarrays, this process is iterated over that many times to give a total of over 600,000 SNPs analyzed.

Obviously for the ordinary individual, the synopsis that 23&me provides is more than enough to satisfy. Depending on what product is chosen, the customer should receive a detailed and user friendly report that contains the information they desire. Most people don’t even realize that the raw data from the SNP analysis is available to everyone. Even if this fact were understood, the format from which raw data is received is confusing at best. The raw data is found in a tab-delimited text file with 4 columns; rsid, chromosome, position, and genotype. Rsid refers to a unique identifier given to every SNP (or at least those that have been studied). Chromosome and position are references to where in the genome this locus can be found and genotype is the customers alleles (A,T,G,C) at that locus. All people have two sets of chromosomes so these genotypes are given as 2 alleles (they can be different).

Information regarding SNPs and the research behind them can be found in a variety of places with the most relevant being NCBI (National Center for Biotechnology Information) and ClinVar (a subsect of NCBI). Searching by rsid on these databases will net a user with information like location, most common alleles, clinical significance, and links to research that cite that SNP. For example, searching NCBI for ‘rs121908177’ would inform you that the A allele at this locus is associated with a disease known as ‘Bardet-Biedl syndrome’. It would also tell you that this allele only occurs around .02% of the time globally. No other allele is listed as having any sort of condition associated with it (NCBI). Since the data file that is provided by 23&me contains both rsid and genotype, you may be able to gain some sort of insight using this process.

So while this process in and of itself is not too overwhelming, also consider that the file is over 600,000 lines long. It would take a science background for someone to truly understand what a single line is referencing, but to actually gather any information from the file would require the assistance of a program that can analyze the information.

# Parser for 23andMe

University of Maryland Global Campus (UMGC) parser will be used to design, read, and publish a program with the ability to insert a variant file from 23andMe and create it into a legible report that is user friendly. This parser is meant to be free and user friendly where someone who is not familiar with coding can easily use this. This program will allow a raw data file such a SNP file to be run locally and will export a PDF report.

# Project Goals

# Our Aim/Goal

The aim of this project is to develop a program that is capable of accepting a SNP analysis raw data file from 23&me, interpreting the content, and returning a report that summarizes the findings. This program should utilize the variant information that is located within the file to provide a report that contains clinical health outcomes/predispositions associated with said variant. Because raw data files are quite large and the information gathering potential in one is significant, this report should only list the top 20 SNPs and the associated clinical significance (ie. disease associated). This program should be accessible to all (github) and act as a mechanism to obtain additional information after someone pays for a program like 23&me or ancestry.com.

# How to achieve it

In order to achieve the successful construction of such a program, the process was divided into several individual sub-sections:

**Step 1: Import Libraries** - Import necessary libraries for HTTP requests, data structures, logging, regular expressions, and GUI functionality, enabling access to essential functions used throughout the program.

**Step 2: Configure Logging** - Set up logging to capture warning-level messages, which helps manage and display program feedback effectively.

**Step 3: Define Custom Exception** - Create a custom exception to flag when the limit of pathogenic diseases has been reached, allowing for specialized error handling in the program.

**Step 4: Define Helper Functions** - Define functions that perform specific tasks, including allele extraction, SNP data retrieval, allele observation extraction, and disease information printing.

* **Step 4.1: Extract Allele Function** - Define a function to parse and return the alternate allele from a given preferred name, helping standardize allele identification.
* **Step 4.2: Retrieve SNP Information Function** - Define a function to fetch ClinVar data for a given SNP ID from an external API, enabling access to genetic variant information.
* **Step 4.3: Extract Observed Alleles Function** - Define a function to gather observed alleles from the retrieved ClinVar data, identifying which alleles appear in each entry.
* **Step 4.4: Print Disease Information Function** - Define a function to print and format disease information based on observed alleles, including pathogenic disease details, to present findings in a readable format.

**Step 5: Process SNP IDs Function** - Processes each SNP ID by retrieving data, extracting and printing pathogenic diseases, and maintaining a cumulative list of unique diseases found across all SNPs.

**Step 6: Extract Pathogenic Diseases Function** - Define a function to identify and return diseases that are marked as pathogenic in each entry, allowing easy extraction of clinically significant diseases.

**Step 7: File Selection and Processing Function** - Create a function to open a file dialog, read a selected SNP data file, and process each SNP entry, ultimately displaying the pathogenic diseases found.

**Step 8: Create User Interface** - Build and display a GUI with labels, instructions, and an upload button, allowing users to select a raw SNP data file for processing.

**Step 9: Run Application** - Start the application by initializing the GUI, which allows users to interact with the program and analyze their SNP data file.

# Purpose

The purpose of this program is to provide the public, and especially customers of 23&me or similar services, a mechanism by which they can utilize genetic information obtained from such services to gather personalized health data. The program is unique because it functions locally (on one’s own computer), which allows it to avoid restrictions put forth by HIPAA, as compared to other companies that require files to be shared (which triggers HIPAA restrictions).

# Databases Scope

The databases that are most commonly used for single nucleotide polymorphism (SNPs) consist of dbSNP and ClinVar. dbSNP is the database of single nucleotide polymorphism which is a repository of variations that contains data of common and rare single nucleotide variations (Sayers et al., 2021). dbSNP has over 1 million reference SNP cluster ID since 2021 (Sayers et al., 2021). Moreover, ClinVar is another database for SNPs. ClinVar is a library of submitted reports that contain information of human variations and phenotypes with supporting evidence (Sayers et al., 2021). The National Institutes of Health maintains ClinVar, which includes the genetic variants, the importance of a disease, and how significant the disease can be (Landrum et al., 2018). These diseases are ranked from benign to pathogenic. Furthermore, another tool that can be commonly used is web application programming interfaces (APIs). Some examples will consist of MyGene.info, MyVariant.info and MyChem,info for annotating genes, variants, and even chemical components (Lelong et al, 2022). MyVariant.info is a high performance annotation service that aggregates and will contain metadata for variants (Lelong et al., 2022). MyVariant.info allows to improve bioinformatics pipelines by providing their own data storage and data fragmentation for downloading, updating and tracking data (Lelong et al., 2022). UMGC parser will focus on using ClinVar for any SNP files that are used and will use API such as MyVariant.info in order to prevent downloading a database and keeping that database updated.

# Development for other Potential Data Analysis/Processing Options

# DNA.Land

DNA.Land was first founded in October 2015 where this is a free upload site that is meant to be an open-access source to help scientists find new discoveries and have a better understanding for cancer (DNA, 2017). Their system works differently compared to Promethease and Genomelink. It is similar as a file will need to be uploaded and you’ll have access to their tools (DNA, 2017). The difference is if someone participates, consent is required, and essentially your data will be shared except any personal identifiers (DNA, 2017). After a file is uploaded, research questions are followed, and then you’ll have access to a personal page. This personal page will contain ancestry data and trait prediction report (DNA, 2017). The trait prediction report will have a genotype summary of effect sizes and SNP locations (DNA, 2017). DNA.Land provides matching information and they impute results unlike other data analysis like Promethase (DNA, 2017).

# Promethease

Promethease is a system that is based on retrieving documents which create a personalized DNA report based from a variant (SNP) file (Promethease, n.d.). Next this will use that file and search through scientific findings which are cited in SNPedia (Promethease, n.d.). SNPedia is similar to wiki, however this wiki inquiries human genetics in which will display information about any effects of variations in DNA which will contain peer-reviewed scientific publications that are cited (SNPedia, 2017). Any person who has taken DNA testing with any public service such as 23andMe can use Promethease to find out more of their variations (Promethese, n.d.). Their reports cost around $12 in which most reports will be produced in a timely manner depending on how large the data file is (Promethese, n.d.). An additional cost of $4 is charged if a combined report is required (Promethese, n.d.).

# Genomelink

Genomelink requires an upload of the SNP file from any genetic testing service which can produce a free report up to 100 traits (Genomelink, n.d.). Genomelink will provide a dashboard which will contain all your genetic information and have the ability to learn more about your DNA based on current science (Genomelink, n.d.). Genomelink also provides a summary of each study that contains a scientific reliability score, in depth reports with information such as nutrition and fitness (Genomelink, n.d.). Genomelink also has the capability to look at other DNA reports such as ancestry and wellness and health (Genomelink, n.d.).

# Implementation

# What we did

This section will act as an outline of what exists in the code in its current state, how it works, and the rationale behind it. It will be broken into the project sub-sections as described in section 3.2. For all of the below programming, it was determined that python would be the most suitable language for this purpose.

**Creation of a parser -** The creation of a parser first started with an understanding of what the input would be and what we needed to extract. The raw data files contain 4 tab-separated columns that include rsid, chromosome, position, and genotype. All that is important to obtain from this is rsid and genotype since all we care about is if the rsID is associated with the disease and if the individual has the disease variant(genotype).

For the actual code, the parser begins with an open dialog, which allows the user to input their file. The program then skips over any lines that start with ‘#’ (these are comment lines and are only found at the beginning of the files). For each of the other lines in the file, the program will strip and split the line, which means each column will be held as its own variable. Because rsid is column 1 and genotype is column 4, variables ‘snp\_id’ and ‘genotype’ are defined as columns[0] and columns[3] respectively. This information is then forwarded to a database which will be discussed in the ‘Obtain SNP information’ sub-section.

The final, especially relevant piece of the parser is that there is a line of code that makes it so that if the rsid doesn’t start with ‘rs’ the code will ignore it and move to the next line. The reason this is important is that it was discovered that actual raw data from 23&me files contains SNP IDs beginning with rs but it also contains SNP ID’s that start with ‘i’ which are 23&me’s proprietary SNP loci (this basically just means other’s have yet to study these loci). Because of their proprietary nature, the SNPs that start with ‘i’ cannot be found in a database and therefore no information can be pulled on them. It was found that it was better to skip over them entirely, as it saves time and computing power.

**Obtain SNP information** - This subsection is by far the largest. First and foremost there needs to be a way by which information is automatically obtained with only an rsid and genotype variable. The API “MyVariant” (<https://myvariant.info/v1/variant/>) is a database that holds all of the genetic data we are sourcing for this project. The API connection allows for data to be searched online without a local database downloaded.

The parsed rsID is used to obtain information regarding disease, clinical significance, and disease associated allele. It uses the link ‘[https://myvariant.info/v1/variant/{snp\_id](https://myvariant.info/v1/variant/%7Bsnp_id)}’ to obtain all this through the API. As data is obtained, it goes through a few filters. These will be described in the next section.

**SNP organization** - After rsID obtains the relevant information from the API, it needs to undergo filtering, so that only relevant information is retained and shared. The first filter is that there is the ‘pathogenic’ tag in the clinical significance field. It was decided that only pathogenic diseases would be included as we felt like we may have more than needed. It was also known that this field could be altered to include additional tags like ‘likely pathogenic’ so for the current version only pathogenic was included.

The second important filter the data needed to go through was the allele filter. The user should receive a disease for which they have the allele associated with it. For example, if disease X is associated with a A>G SNP, this means the wild type is A and the disease allele is a G. If our user has a genotype AA, we wouldn’t want to include disease X in their list because they don’t have the disease causing allele. This particular filter is what caused the greatest issues.

After undergoing the filtering, there should be a list of pathogenic diseases. This list should then be taken on to the creation of a report.

**Report Design** - Because the main functions of the program were never fully completed, it was challenging for some of the supporting functions to be fully completed either. Report design is one of those supporting functions. FPDF is a python library that is capable of designing pdf output files from within a program. In the case of this project, FPDF was used to create a simple pdf that contained the list of pathogenic diseases. In its fully complete form, this report should also include disease information as well as a link to research.

**UI design** - UI is another area that could not be fully completed because of the status of the rest of the project. Tkinter is a python library that is used to create user-friendly GUI’s. Because the requirement for this section was something simple, it was decided that Tkinter would support the needs of the project well. Initially, there was a goal to have 3 parts to the UI - welcome/file input screen, loading screen, and completion screen. The only part that was fully completed was the welcome/file input screen. This part utilized the function ‘create\_ui’ which made a simple UI that describes what the program is and how to use it. The window includes a button that utilizes the command ‘open\_file’ which opens a file dialogue to start the parsing process.

# What we struggled with

**API connection to databases**

An application programming interface (API) is a way software can communicate with each other. In our program, an API had to be used to connect to the database(s) that contain information about the SNPs and diseases so that a local version of the database did not have to be downloaded. We could not use a downloaded database because of the size of the databases we needed to pull data from. One of the main goals of this project was to make it easy for the majority of individuals to use our program. Having to download a local version of the database(s) would make it virtually impossible for anyone with an average amount of computer storage (1T or less) to use the program.

Not all APIs are created equal and while they make the software connection easier, it can be hard to find the data you are looking for in the API. Our original plan was to use the NCBI API called Entrez Programming Utilities (E-utilities) (<https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esummary.fcgi>). This would have been ideal to connect to the NCBI database directly which was our goal since NCBI is the standard for SNP data. We had problems with this API mostly stemming from not having a clear attribute name for the data we were looking for. The API made it difficult to find and print specific data that was connected to a rsID. Our final code up to this point uses the API from MyVariant.info (<https://myvariant.info/v1/variant/>). This website is created and maintained by researchers at the Department of Integrative, Structural, and Computational Biology at Scripps Research. This API is easier to use than E-utilities making it easier to pull specific data from rsIDs. The information used by MyVariant.info is collected from more databases than just NCBI so it is also more frequently updated and has more data than the NCBI API.

**Reading of the Alleles**

In the final version of our program for the completion of this course, the most outstanding bug in our code has to do with the alleles that are searched when looking for diseases. Once a rsID is found, the possible diseases are listed by the allele the data has. Different allele SNPs can cause different genetic diseases. To be accurate, the diseases that are returned to the user need to be only caused by the allele they have. The allele that the program is searching for is not the allele that is listed in the user’s data. Therefore, the diseases that the program is returning are not accurate for the user’s data. This problem has something to do with the way the program is looking for the allele in the SNP data. The program is looking for the allele in an ID name that has several pieces of data strung together. Even though the ID sequence is the same for all ID numbers, the program is struggling to find the correct allele. This is a major bug that needs to be fixed before the program is used.

# Improvement Cases

Once the bugs are fixed in the program, there are several other features that could be added. The possibilities are very extensive, however, some features we would like to see added include the:

1. Disease severity ranking: In the output file, list the diseases the person is at risk of in order to severity. We would recommend using the mortality rate in order to rank the diseases. For example, being susceptible to lung cancer that has high mortality would be higher on the disease list than non-lethal eczema.
2. Listing Cancer risk in a separate table: Cancer is a unique class of diseases that can be assumed to be a special group of interest for users. Cancer risk could be separated into a different table than all other diseases found so that it is easier for users to find the information they want. Many people who use this program may only want to know cancer risks and won’t care as much for other disease risks.
3. Disease Probability: Not all diseases are guaranteed to show symptoms if you have the gene. For genes that only increase the risk of developing the disease, any information on the rate of disease occurrence would be helpful to print with the disease information result for the user to see.
4. Disease mitigation: List recommendations on how to mitigate disease symptoms for each disease.
5. Modern GUI - Reformat the GUI for a more modern feel.
6. Sorting by disease frequency in the population: Sorting the disease output by how common these diseases are would help users quickly see common vs rare diseases they have genes. This would be another easy way to organize the results in a way that is more meaningful for the user.

In addition to those recommendations listed above, the current format of the program receives data from MyVariant for each rsID independently. Because of this, the program runs slow. It feels impractical in its current form to utilize a full raw data file (it simply takes too long). It is important that a future version addresses this issue. One possibility (among many) is to bundle the rsIDs together and gather information on many at the same time (because the longest part of the process is obtaining the data from the API).

# Advancement in Technology

An advancement in technology that can potentially replace pipelines built to analyze and give reports of diseases is artificial intelligence (AI). Artificial intelligence has promising results in simplifying and speeding up genome interpretation by using predictive methods as there is growing knowledge of genetic diseases (De La Vega et al., 2021). To begin with, detecting a rare genetic disease by determining a disease-causing variant among four million benign variants is an issue (De la Vega et al., 2021). To interpret variants, a trained genomic analyst, genetic counselors, and laboratory directors are needed to justify that a variant causing disease is real and can be time-consuming since review for 100 variants can take up to 50 to 100 hours per patient (De La Vega et al., 2021). Using artificial intelligence for genome interpretation methods is being developed to remove the clinical decision support system (De La Vega et al., 2021). However, speed and accuracy of interpretation of data are crucial so artificial intelligence that can work on diagnosis is still being investigated (De La Vega et al., 2021). As of now, artificial intelligence issues arise consist of what methods are most suitable, the comparison between current variant prioritization approaches without losing accuracy, their diagnostic performance among different clinical scenarios and variant types, the potential to offer new sets of decision support, and how they integrate with automated patient phenotyping and decision making (De La Vega et al., 2021). Potentially artificial intelligence can replace any parser or code online that can offer more information and be accurate in the future. As of now, this is being investigated and not currently offered. However, once this has the capability to interpret results and give accurate diagnoses, parsers can become obsolete.

# Future Trending

Epidemiology-based models of disease risk are informed by family history (Ho et al., 2019). To improve the risk model, diseases and phenotypes associated with SNP improved the accuracy (Ho et al., 2019). Currently, developing a risk genetic model in order to successfully gain accurate predictive power for noticing at-risk individuals in a timely robust manner is a main focus (Ho et al., 2019). In addition, adding SNPs into a risk prediction model constructed by polygenic risk scoring or machine learning can help in creating a robust manner (Ho et al., 2019). Polygenic risk scores use a fixed model approach to add the contribution of a set of risk alleles to a complex disease which can consist of unweighted or weighted scores (Ho et al., 2019). In addition, machine learning approaches take account of sophisticated statistical and computational algorithms such as support vector machines or random forests in order to make predictions by mathematically finding a link between the complex association between SNPS and complex disease phenotypes (Ho et al., 2019). The use of random forest and neural networks enables large-scale processing with essentially large genomic data which can excel at determining SNPs for any traits, meanwhile improving prediction accuracy (Oualikene-Gonin et al., 2024).

# Final Recommendations

It would be great to see this program move forward. First, bugs need to be fixed in order to get a solid working core program with an accurate output. Once that is completed, any number of features that have been recommended could be added. Starting with organizing the already given output would be a good place to start, and then move on to adding more data to the output. This will make sure that even while adding features, the resulting output is accurate and usable at each stage of the process even while new features are being added.

# References

Allyn, B. (2024, September 25). 23andMe is in trouble. what happens to all the DNA data?. NPR. https://www.npr.org/2024/09/25/nx-s1-5123633/23andme-is-in-trouble-what-happens-to-all-the-dna-data

Brown TA. Genomes. 2nd edition. Oxford: Wiley-Liss; 2002. Chapter 1, The Human Genome. Available from: https://www.ncbi.nlm.nih.gov/books/NBK21134/

De La Vega, F. M., Chowdhury, S., Moore, B., Frise, E., McCarthy, J., Hernandez, E. J., Wong, T., James, K., Guidugli, L., Agrawal, P. B., Genetti, C. A., Brownstein, C. A., Beggs, A. H., Löscher, B. S., Franke, A., Boone, B., Levy, S. E., Õunap, K., Pajusalu, S., Huentelman, M., … Kingsmore, S. F. (2021). Artificial intelligence enables comprehensive genome interpretation and nomination of candidate diagnoses for rare genetic diseases. Genome medicine, 13(1), 153. https://doi.org/10.1186/s13073-021-00965-0

DNA. (2017). DNA.Land. Imputation Analysis Utilizing Promethease. <https://dna-explained.com/2017/10/11/imputation-analysis-utilizing-promethease/>

DNA. (2017). DNA.Land. DNA. <https://dna-explained.com/2017/12/07/dna-land/>

Genomelink. (n.d.). How it works. https://genomelink.io/product/how-it-works

Ho, D. S. W., Schierding, W., Wake, M., Saffery, R., & O'Sullivan, J. (2019). Machine Learning SNP Based Prediction for Precision Medicine. Frontiers in genetics, 10, 267.<https://doi.org/10.3389/fgene.2019.00267>

Illumina. (2012, March 2). Illumina Infinium Assay: An Overview. YouTube. https://www.youtube.com/watch?v=lVG04dAAyvY

Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., Karapetyan, K., Katz, K., Liu, C., Maddipatla, Z., Malheiro, A., McDaniel, K., Ovetsky, M., Riley, G., Zhou, G., Holmes, J. B., … Maglott, D. R. (2018). ClinVar: improving access to variant interpretations and supporting evidence. Nucleic acids research, 46(D1), D1062–D1067. https://doi.org/10.1093/nar/gkx1153

Lee J. E. (2007). High-throughput genotyping. Forum of nutrition, 60, 97–101. https://doi.org/10.1159/000107078

Lelong, S., Zhou, X., Afrasiabi, C., Qian, Z., Cano, M. A., Tsueng, G., Xin, J., Mullen, J., Yao, Y., Avila, R., Taylor, G., Su, A. I., & Wu, C. (2022). BioThings SDK: a toolkit for building high-performance data APIs in biomedical research. Bioinformatics (Oxford, England), 38(7), 2077–2079. https://doi.org/10.1093/bioinformatics/btac017

Odake, S. (2023, March 22). How many snps does 23andme test?. SelfDecode Resources. https://resources.selfdecode.com/blog/how-many-snps-does-23andme-test/

Oualikene-Gonin, W., Jaulent, M. C., Thierry, J. P., Oliveira-Martins, S., Belgodère, L., Maison, P., Ankri, J., & Scientific Advisory Board of ANSM (2024). Artificial intelligence integration in the drug lifecycle and in regulatory science: policy implications, challenges and opportunities. Frontiers in pharmacology, 15, 1437167.<https://doi.org/10.3389/fphar.2024.1437167>

Our story: Ancestry corporate. Ancestry®. (n.d.). https://www.ancestry.com/corporate/about-ancestry/our-story#:~:text=Ancestry%20reaches%20over%203.5%20million,subscribers%20and%2027%20billion%20records.?msockid=1c6b3bd826196bc308c62ff327806a61

Promethease. (n.d.). Promethease. https://www.promethease.com/ondemandagreed

Sayers, E. W., Beck, J., Bolton, E. E., Bourexis, D., Brister, J. R., Canese, K., Comeau, D. C., Funk, K., Kim, S., Klimke, W., Marchler-Bauer, A., Landrum, M., Lathrop, S., Lu, Z., Madden, T. L., O'Leary, N., Phan, L., Rangwala, S. H., Schneider, V. A., Skripchenko, Y., … Sherry, S. T. (2021). Database resources of the National Center for Biotechnology Information. Nucleic acids research, 49(D1), D10–D17. https://doi.org/10.1093/nar/gkaa892

Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., Connor, R., Funk, K., Kelly, C., Kim, S., Madej, T., Marchler-Bauer, A., Lanczycki, C., Lathrop, S., Lu, Z., Thibaud-Nissen, F., Murphy, T., Phan, L., Skripchenko, Y., Tse, T., … Sherry, S. T. (2022). Database resources of the national center for biotechnology information. Nucleic acids research, 50(D1), D20–D26. https://doi.org/10.1093/nar/gkab1112

Shastry B. S. (2009). SNPs: impact on gene function and phenotype. Methods in molecular biology (Clifton, N.J.), 578, 3–22. <https://doi.org/10.1007/978-1-60327-411-1_1>

SNPedia. (2017). SNPedia. https://www.snpedia.com/

U.S. National Library of Medicine. (n.d.). RS121908177 RefSNP report - dbSNP - NCBI. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/snp/rs121908177