BIOINFORMATICS
INSTITUTE

Personalised genomics and gene therapy based on 23andMe data

February 15, 2024

Project number 5

by Ilia Popov and Shakir Suleimanov

Introduction

The human genome contains the information that determines our traits and health risks [1]. However, the human genome is variable and influenced by genetic variations, such as SNPs, which can affect gene function and disease susceptibility. To explore and modify these variations, we can use genotyping and genome editing technologies. Genotyping is the process of determining the alleles of specific genetic markers, such as SNPs, using methods such as genotyping chips. Genome editing is the process of altering DNA sequences in living cells using tools such as CRISPR-Cas systems [2]. In this study, we used data from 23andMe, a direct-to-consumer genetic testing service, to analyze and modify the genome of an individual. We identified the mtDNA and Y-DNA haplogroups, the eye color genotype, and five clinically significant SNPs associated with various diseases. We also proposed potential genome editing strategies to correct the deleterious SNPs and reduce the disease risk. Our study shows the utility of 23andMe data for personal genomics analysis and the potential of genome editing for disease prevention and treatment.

Materials and methods

The dataset with raw data from 23andMe was obtained from Google Drive. The program PLINK v1.90b6.22 was used to convert the raw data into the standard .vcf format. All SNPs corresponding to deletions and insertions were removed from the resulting file for further SNP annotation [3].

Haplogroup identification

For the classification of mitochondrial DNA (mtDNA) haplogroups, the online server mthap v0.19b was used with default settings, utilizing haplogroup data from PhyloTree Build 17 [4]. The Cambridge Reference Sequence (CRS) was used as a reference.

The Y-chromosome haplogroup was determined by using the online servers of Y-SNP Subclade Predictor [5] and YSEQ Clade Finder [6] with default parameters.

SNP annotation for searching for clinically significant variants

SNPs were annotated using SnpEff v5.1d with the Genome Reference Consortium Human Build 37 [7], release 75 database. SnpSift v5.1d was employed to compare the obtained data with the ClinVar database [8,9]. Search and confirmation of the obtained SNP data were conducted in the SNPedia database [10].

Results

For the analysis of human genome variations, we started with 610,526 SNPs in the raw 23andMe data file. After removing insertions and deletions using the PLINK v1.90b6.22 program, we obtained 595,401 unique SNPs.

Haplogroup identification

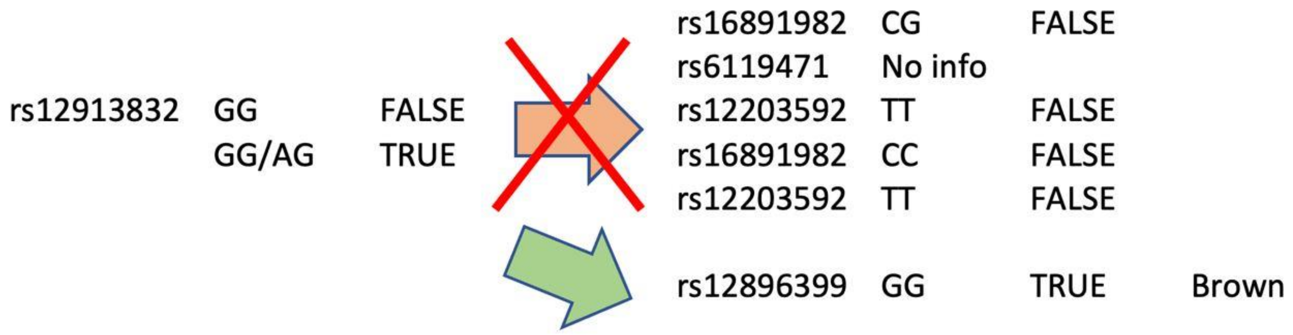
To determine haplogroup based on variations in mitochondrial DNA, we used 3270 markers covering 3268 positions and analyzing up to 19.7% of mtDNA. As a result, the analyzed mtDNA was determined to belong to haplogroup H(T152C). Using the mthap v0.19b, the following markers were identified: four (750G, 1438G, 4769G and 8860G) in the non-coding (control) region and two (152C and 263G) specifically in the hypervariable region II.

Using Y-SNP Subclade Predictor and YSEQ Clade Finder, we classified the Y-DNA haplogroup with 2084 SNPs located on the Y chromosome. Based on 166 detected mutations with positive results, this man was found to belong to haplogroup R1a1a (markers M17, M198), namely R1a1a (R-M417/R-Page-7).

Eye colour

In this individual, the genetic variant rs12896399 on chromosome 14 indicates the GG genotype. This genotype is associated with brown eye colour, which is the result of genetic factors affecting iris pigmentation [11].

Figure 1. Eye colour analysis steps.



Clinically significant variants

A search for clinically significant SNPs using SnpEff and SnpSift revealed a total of 14 SNPs associated with risk factors for various diseases. Of these, we selected 5 for further analysis and suggestions for genome modification.

Table 1. Clinically significant SNPs.

rsID	Location		Genotype	Gene	Associated Condition	Fixed Genotype
	Chromosome	Coordinates				
rs4402960	3	185511687	GT	IGF2BP2	1.2x increased risk for type-2 diabetes	GG [12]
rs7756992	6	20679709	AG	CDKAL1	1.3x increased risk for type-2 diabetes	AA [13]
rs2004640	7	128578301	GT	IRF5	1.4x increased risk for Systemic lupus erythematosus	GG [14]
rs7794745	7	146489606	AT	CNTNAP2	Slightly increased risk for autism	AA [15]
rs13266634	8	118184783	CT	SLC30A8	Increased risk for type-2 diabetes	TT [16]

Discussion

In this study, we analyzed the genome variations of an individual using data from 23andMe, a direct-to-consumer genetic testing service. We identified the mtDNA and Y-DNA haplogroups, the eye colour genotype, and five clinically significant SNPs associated with various diseases. Our findings provide insights into the genetic ancestry, traits, and health risks of the individual, as well as potential applications of genome editing.

The haplogroups of the individual's mitochondrial DNA and Y-chromosome can provide clues about their genetic ancestry and ethnicity [17]. The mtDNA haplogroup H is the most common in Europe, and it originated in Southwest Asia, near present-day Syria, about 20,000 to 25,000 years ago [18]. The individual belongs to the subclade H(T152C), which is relatively rare and has been found mainly in Europe and the Near East [19]. The Y-DNA haplogroup R1a is widespread in Eurasia, and it originated in Southern Siberia about 22,000 to 25,000 years ago. The individual belongs to the subclade R1a1a (R-M417/R-Page-7), which diversified in the vicinity of Iran and Eastern Turkey about 5,800 years ago. This subclade is associated with the Indo-European languages and cultures, and it has been found in various regions such as Central and Eastern Europe, Central and South Asia, and Siberia [20,21]. Based on these haplogroups, the individual may have a mixed ancestry and ethnicity, reflecting the complex history and migrations of human populations in Eurasia.

One of the most striking results of our analysis is that the individual carries three SNPs that are associated with an increased risk of type-2 diabetes (T2D), a chronic metabolic disorder that affects millions of people worldwide [22]. These SNPs are rs4402960 in the IGF2BP2 gene, rs7756992 in the CDKAL1 gene, and rs13266634 in the SLC30A8 gene. These genes are involved in the regulation of insulin secretion, glucose metabolism, and beta-cell function [23]. The presence of these SNPs suggests that the individual has a genetic predisposition to develop T2D, especially if combined with environmental and lifestyle factors such as obesity, sedentary behavior, and poor diet. Therefore, it is important for the individual to monitor their blood glucose levels, adopt a healthy lifestyle, and consult a physician if they experience any symptoms of T2D.

Another implication of our analysis is that the individual could benefit from genome editing, a novel technology that allows precise and targeted modification of DNA sequences. Genome editing could potentially correct the deleterious SNPs that we identified and reduce the risk of developing the associated diseases. For example, genome editing could change the GT genotype of rs4402960 to GG, which is associated with a lower risk of T2D. This could be achieved by using a CRISPR-Cas9 system that targets the specific site of the SNP and introduces a double-strand break, followed by a homology-directed repair mechanism that inserts the desired G allele [2]. Similarly, genome editing could alter the other SNPs in the CDKAL1, IRF5, CNTNAP2, and SLC30A8 genes to the protective alleles, which could reduce the risk of T2D, systemic lupus erythematosus, autism, and T2D, respectively.

However, genome editing is not without challenges and limitations. First, the safety and efficacy of genome editing in humans are still uncertain, and there may be unintended consequences such as off-target effects, immune reactions, and ethical issues. Second, the causal relationship between the SNPs and the diseases is not fully established, and there may be other genetic and environmental factors that influence the disease susceptibility and progression. Third, the individual's consent and preference are essential for any genome editing intervention, and they may have different views on the benefits and risks of modifying their genome.

In conclusion, our study demonstrates the utility of 23andMe data for personal genomics analysis and the potential of genome editing for disease prevention and treatment. However, further research and ethical deliberation are needed before genome editing can be applied in clinical practice.

References

1. Toxicology, N.R.C. (US) C. on D. Human Genetics and the Human Genome Project. In *Scientific Frontiers in Developmental Toxicology and Risk Assessment*; National Academies Press (US), 2000.
2. Bak, R.O.; Gomez-Ospina, N.; Porteus, M.H. Gene Editing on Center Stage. *Trends in Genetics* **2018**, *34*, 600–611, doi:10.1016/j.tig.2018.05.004.
3. Shaun Purcell, C.C. PLINK v1.90b6.22 Available online: www.cog-genomics.org/plink/1.9/ (accessed on 15 February 2024).
4. Van Oven, M.; Kayser, M. Updated Comprehensive Phylogenetic Tree of Global Human Mitochondrial DNA Variation. *Hum. Mutat.* **2009**, *30*, E386–E394, doi:10.1002/humu.20921.
5. Extract Y-DNA Data from an Autosomal Test Available online: https://ytree.morleydna.com/extractFromAutosomal (accessed on 15 February 2024).
6. YSEQ Clade Finder Available online: https://cladefinder.yseq.net/ (accessed on 15 February 2024).

7. Cingolani, P.; Platts, A.; Wang, L.L.; Coon, M.; Nguyen, T.; Wang, L.; Land, S.J.; Lu, X.; Ruden, D.M. A Program for Annotating and Predicting the Effects of Single Nucleotide Polymorphisms, SnpEff. *Fly* **2012**, *6*, 80–92, doi:10.4161/fly.19695.
8. Cingolani, P.; Patel, V.M.; Coon, M.; Nguyen, T.; Land, S.J.; Ruden, D.M.; Lu, X. Using Drosophila Melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. *Front Genet* **2012**, *3*, 35, doi:10.3389/fgene.2012.00035.
9. Landrum, M.J.; Lee, J.M.; Riley, G.R.; Jang, W.; Rubinstein, W.S.; Church, D.M.; Maglott, D.R. ClinVar: Public Archive of Relationships among Sequence Variation and Human Phenotype. *Nucl. Acids Res.* **2014**, *42*, D980–D985, doi:10.1093/nar/gkt1113.
10. SNPedia Available online: https://www.snpedia.com/ (accessed on 15 February 2024).
11. Hart, K.L.; Kimura, S.L.; Mushailov, V.; Budimlija, Z.M.; Prinz, M.; Wurmbach, E. Improved Eye- and Skin-Color Prediction Based on 8 SNPs. *Croat Med J* **2013**, *54*, 248–256, doi:10.3325/cmj.2013.54.248.
12. Rs4402960 - SNPedia Available online: https://www.snpedia.com/index.php/Rs4402960 (accessed on 15 February 2024).
13. Rs7756992 - SNPedia Available online: https://www.snpedia.com/index.php/Rs7756992 (accessed on 15 February 2024).
14. Rs2004640 - SNPedia Available online: https://www.snpedia.com/index.php/Rs2004640 (accessed on 15 February 2024).
15. Rs7794745 - SNPedia Available online: https://www.snpedia.com/index.php/Rs7794745 (accessed on 15 February 2024).
16. Rs13266634 - SNPedia Available online: https://www.snpedia.com/index.php/Rs13266634 (accessed on 15 February 2024).
17. Keats, B.J.B.; Sherman, S.L. Chapter 13 - Population Genetics. In *Emery and Rimoin's Principles and Practice of Medical Genetics (Sixth Edition)*; Rimoin, D., Pyeritz, R., Korf, B., Eds.; Academic Press: Oxford, 2013; pp. 1–12 ISBN 978-0-12-383834-6.
18. Achilli, A.; Rengo, C.; Magri, C.; Battaglia, V.; Olivieri, A.; Scozzari, R.; Cruciani, F.; Zeviani, M.; Briem, E.; Carelli, V.; et al. The Molecular Dissection of mtDNA Haplogroup H Confirms That the Franco-Cantabrian Glacial Refuge Was a Major Source for the European Gene Pool. *The American Journal of Human Genetics* **2004**, *75*, 910–918, doi:10.1086/425590.
19. Roostalu, U.; Kutuev, I.; Loogväli, E.-L.; Metspalu, E.; Tambets, K.; Reidla, M.; Khusnutdinova, E.; Usanga, E.; Kivisild, T.; Villems, R. Origin and Expansion of Haplogroup H, the Dominant Human Mitochondrial DNA Lineage in West Eurasia: The Near Eastern and Caucasian Perspective. *Molecular Biology and Evolution* **2007**, *24*, 436–448, doi:10.1093/molbev/msl173.
20. Underhill, P.A.; Poznik, G.D.; Rootsi, S.; Järve, M.; Lin, A.A.; Wang, J.; Passarelli, B.; Kanbar, J.; Myres, N.M.; King, R.J.; et al. The Phylogenetic and Geographic Structure of Y-Chromosome Haplogroup R1a. *Eur J Hum Genet* **2015**, *23*, 124–131, doi:10.1038/ejhg.2014.50.
21. Garrison, E.; Marth, G. Haplotype-Based Variant Detection from Short-Read Sequencing 2012.
22. Khan, M.A.B.; Hashim, M.J.; King, J.K.; Govender, R.D.; Mustafa, H.; Al Kaabi, J. Epidemiology of Type 2 Diabetes – Global Burden of Disease and Forecasted Trends. *J Epidemiol Glob Health* **2020**, *10*, 107–111, doi:10.2991/jegh.k.191028.001.
23. Fu, Z.; Gilbert, E.R.; Liu, D. Regulation of Insulin Synthesis and Secretion and Pancreatic Beta-Cell Dysfunction in Diabetes. *Curr Diabetes Rev* **2013**, *9*, 25–53.