Tweaks and fixes: releasing a new fast and stable human version (not LTS)

Aigul Nugmanova[1, 2 *] and Nikita Vaulin[1, 3, * †]

[1]Bioinformatics Institute, Saint-Petersburg, Russia
[2]ITMO university, Saint-Petersburg, Russia
[3]Skolkovo Institute of Science and Technology, Moscow, Russia

[*]Contributed equally.
[†]vaulin@ro.ru

Abstract

In this study we analysed the 23andMe VCF results of a private person. We identified the1 Y chromosome and mitochondrial DNA haplotypes. Also, we performed a scanning for the SNPs associated with dangerous diseases or, *vice versa*, favorable traits. As a result, we provide a short list of 5 candidate SNPs to introduce or remove in both above-mentioned categories.

Key words: SNP, genotype, haplotype, allele, disease, traits

Introduction

The study of the genome can help a person in various areas of life, from the detection and treatment of genetic diseases to forensics. One of the most accessible options for obtaining information about the genome of a particular person is genotyping chips. They allow you to quickly determine the presence of SNP mutations in genes [1].

Also, a very important recent achievement was the CRISPR-Cas9 gene editing technology [2, 3]. CRISPR-Cas systems are part of the prokaryotic adaptive immunity which consists of repetitive sequences, interrupted by unique «spacer» sequences. In most of the studied cases, spacers represent the viral DNA sequences. Such a system serves as a genetic memory that helps the cell to detect and destroy viruses. Similarly, CRISPR-Cas9 works to edit the genome of more complex organisms. CRISPR spacers are transcribed into short RNA sequences capable of directing the system to match DNA sequences. When a target DNA is found, one of the enzymes, Cas9, binds to the DNA and cuts it, turning off the target gene or modifying it.

In this paper, we analyze the results of the 23andme DNA genetic test to find specific variants and show how genome editing technology can be applied to improve the quality of life.

Materials and methods

Raw data

The 23andMe results of Illumina OmniExpress Plus Beadchip genotyping of private person were used. The GRCh37 human genome assembly was used as a reference. Raw 23andMe data was converted into standard VCF format with excluding non-SNP variants using *plink* [4, 5].

Variants analysis

To establish the haplogroups basing on mitochondrial DNA or Y chromosome the James Lick Mthap and *yhaplo* v. 1.1.2. Predictor correspondingly were used [6, 7].

Annotation and filtering of SNPs was performed with *Ensemble Vatiant Effect Predictor* (VEP) [8] against gnomaAD exome database with default parameters. The resulting variants were analysed by searching in ClinVar database and SNPedia [9, 10].

Results

Haplogroups

Using the mitochondrial DNA sequence there were found 3270 haplogroup markers. The main defining markers found (152C 263G 750G 1438G 4769G 8860G) result in European H2a2a1 mtDNA haplogroup best match as a branch of H haplogroup (see James Lick Mthap results in). With the Y chromosome DNA from the 13569 unique filtered SNPs the haplogroup R1a1a1 R-M417 was defined.

Proposed SNP changes

With the VEP almost 30% of the SNPs were found to be missence variants. By the way, only 13% of all SNPs were intergenic. The

Compiled on: February 17, 2023.
Draft manuscript prepared by the author.

Ch	Position	ID	Current	Proposed	Gene	Description
7	128938247	rs2004640	GT	GG	IRF5	Systemic lupus erythematosus, rheumatoid arthritis increased risk
8	117172544	rs13266634	CT	TT	ZNT8	2.5 times increased risk for type-2 diabetes
3	133775510	rs1049296	CT	CC	TF	Slightly higher risk for Alzheimer's disease
2	203867991	rs231775	AG	AA	CTLA4	1.5 times increased risk of autoimmune thyroiditis
6	32636595	rs9272346	AA	GG	HLA-DQA1	18.5 times increased risk for type-1 diabetes

Table 1. Suggested SNPs to fix

Ch	Position	ID	Current	Proposed	Gene	Description
7	17244953	rs4410790	CT	TT	AHR	Lower caffeine dependence
9	98542066	rs2184026	CT	CC	GABBR2	Lower nicotine dependence
11	27658369	rs6265	CT	GG	BDNF	Slightly lower risk for ADHD or depression
6	28662914	rs13194504	GG	AG	ZBED9	Slightly lower risk for schizophrenia
16	31096368	rs9923231	CC	CT	VKORC1	Reduced warfarin dose if treated for VTE

Table 2. Suggested SNPs to optimize

suggested changes for the fixes and optimisations are provided in the tables 1 and 2 correspondingly.

Discussion

Bug fixes

The **rs2004640** SNP is a singe one SNP that was found by VEP to be a risk factor and have a high impact at the same type. It is located in the IRF5 gene (Interferon regulatory factor 5) in chromosomal region 7q32.1. The rs2004640 is one of several SNPs associated with systemic lupus erythematosus (SLE) and also may have a pathogenic effect.

The **rs13266634** SNP is associated with type-2 diabetes in several studies. Also it was shown that the C allele is associated with younger age of onset of type-1 diabetes.

The **rs1049296** SNP encodes C1/C2 subtypes of the transferrin TF gene. The C allele encodes the C1 subtype, and the rarer T allele encodes the C2 subtype. It is involved in the transport of iron, transferrin and its subtypes have been linked at times to various conditions, perhaps most notably Alzheimer's disease.

The **rs231775** SNP is a polymorphism of the CTLA4 gene. Polymorphisms are associated with several autoimmune diseases, especially autoimmune thyroiditis. Also it was shown to be associated with susceptibility to Graves' disease.

The **rs9272346** SNP has been reported to be associated with type-1 diabetes. The (A;A) genotype can lead to 18,5 times increased risk.

Optimizations

The **rs6265**, responsible for the brain-derived neurotrophic factor BDNF gene, slightly increased the risk for ADHD and depression, and for Alzheimer patients increase mental decline.

People with A allele in **rs13194504** have a lower risk of schizophrenia compared to those one with the common (G;G) genotype. We hope it will help the instructor to better endure work with the students.

In [11] was shown that related to the treatment of venous thromboembolism with the blood thinner warfarin for this SNP is that carriers of the **rs9923231**(T) allele require significantly reduced doses of warfarin. This turns out to be very important, since warfarin is a life-saving but insanely dangerous drug, an increased dose of which can lead to death from blood loss.

The **rs2470890** SNP has a significant correlation with the CYP1A2 gene, who influences the speed of caffeine metabolizers.

H2a2a1 and R1a1a1 haplogroups

We identified that sample under the study belongs to the H2a2a1 branch oh H2 sub-haplogroup of haplogroup H. Haplogroup H is the the most frequent haplogroup in the western Eurasia [12]. H2 sub-haplogroup is typical for the eastern Europe population [12]. H2a2a1 was found to be significantly associated with high altitude pulmonary edema [13]. At the same time it was shown, that H2a2a1 may have a possible protective effect against Leber's hereditary optic neuropathy [14].

R1a haplogroup also occurs to be more often in the eastern Europe [15, 16]. By the way, it may appear in Europe with the Scythians expeditions [16]. R1a1a1 is the most common large subgroup within the R1a, which stands out with the M417 sub-group marker (appeared 5800 years ago) [16]. Some result may indicate, that the R1a1a1 haplogroup is associated with higher risks of severe COVID-19 [17].

Eye color

Our results indicate that the private person under the study possibly have brown eyes [18]

References

1. Tsuchihashi Z, Dracopoli N. Progress in high throughput SNP genotyping methods. The pharmacogenomics journal 2002;2(2):103–110.
2. Hsu PD, Lander ES, Zhang F. Development and applications of CRISPR-Cas9 for genome engineering. Cell 2014;157(6):1262–1278.
3. Ormond KE, Mortlock DP, Scholes DT, Bombard Y, Brody LC, Faucett WA, et al. Human germline genome editing. The American Journal of Human Genetics 2017;101(2):167–176.
4. Shaun Purcell CC, PLINK 1.9;. [Online; accessed 17-February-2023]. www.cog-genomics.org/plink/1.9/.
5. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 2015;4(1):s13742–015.
6. Van Oven M, Kayser M. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Human mutation 2009;30(2):E386–E394.
7. Poznik GD. Identifying Y-chromosome haplogroups in arbitrarily large samples of sequenced or genotyped men. BioRxiv 2016;p. 088716.
8. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, et al. The ensembl variant effect predictor. Genome biology 2016;17(1):1–14.
9. Cariaso M, Lennon G. SNPedia: a wiki supporting personal

genome annotation, interpretation and analysis. Nucleic acids research 2012;40(D1):D1308–D1312.

10. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic acids research 2018;46(D1):D1062–D1067.
11. Rieder MJ, Reiner AP, Gage BF, Nickerson DA, Eby CS, McLeod HL, et al. Effect of VKORC1 haplotypes on transcriptional regulation and warfarin dose. New England Journal of Medicine 2005;352(22):2285–2293.
12. Roostalu U, Kutuev I, Loogväli E, Metspalu E, Tambets K, Reidla M, et al. 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(2):436–448.
13. Sharma S, Singh Y, Sandhir R, Singh S, Ganju L, Kumar B, et al. Mitochondrial DNA mutations contribute to high altitude pulmonary edema via increased oxidative stress and metabolic reprogramming during hypobaric hypoxia. Biochimica et Biophysica Acta (BBA)-Bioenergetics 2021;1862(8):148431.
14. Qiao C, Wei T, Hu B, Peng C, Qiu X, Wei L, et al. Two families with Leber's hereditary optic neuropathy carrying G11778A and T14502C mutations with haplogroup H2a2a1 in mitochondrial DNA. Molecular Medicine Reports 2015;12(2):3067–3072.
15. Underhill PA, Poznik GD, Rootsi S, Järve M, Lin AA, Wang J, et al. The phylogenetic and geographic structure of Y-chromosome haplogroup R1a. European Journal of Human Genetics 2015;23(1):124–131.
16. Cynarski WJ, Maciejewska A. The proto-Slavic warrior in Europe: the scythians, sarmatians and lekhs. Ido Movement for Culture Journal of Martial Arts Anthropology 2016;3(16):1–14.
17. Dalton D. Why People of Bangladeshi Origin Are Most at Risk of Dying from COVID-19 in the UK, ie, Why They Carry the Highest Level of Those Six Deleterious Neanderthal-Inherited Genes. Why They Carry the Highest Level of Those Six Deleterious Neanderthal-Inherited Genes (August 10, 2020) 2020;.
18. Hart KL, Kimura SL, Mushailov V, Budimlija ZM, Prinz M, Wurmbach E. Improved eye-and skin-color prediction based on 8 SNPs. Croatian medical journal 2013;54(3):248–256.

Supplemental materials

Working Notes: Notion

GitHub repository of the project: Human_project_BI_2022