Practice

BIOINFORMATICS INSTITUTE

# Variants analysis of a patient 23andMe genotyping data: what is hidden behind mysterious SNPs?

Ekaterina Grigorieva[1, 2] and Anna Chechenina[1, 3]

[1]Bioinformatics Institute
[2]Moscow Institute of Physics and Technology
[3]Pompeu Fabra University

## Abstract

There is a significant advance being reached in the field of genotyping during this century. The growing amount of genotyping array techniques and approaches to identify variations in the whole genome sequencing data gave the opportunity to explore sets of SNPs for many individuals with high precision for a reasonable price. Single nucleotide polymorphism analysis allows us to identify many traits, diseases, and phenotypic characteristics that might be inherited. These features might be essential for clinical predictions, children planning and for many other health-related purposes. In this study, we performed an analysis of SNP variants on a certain person's data. We tried to infer the most important phenotypical features and describe the most essential SNP from the dataset.

**Keywords:** genotyping, SNP, 23andMe, variants, DNA

## Introduction

As we know, genetic variants are one of the main features defining a phenotype of an organism. These single nucleotide polymorphisms (SNPs) could play a crucial role in the development of diseases and be responsible for many traits. There are several ways of identifying sets of SNPs for the sample. The first approach is utilizing genotyping arrays (chips), that perform scanning of a sample for a given set of SNPs. Since 2000s, genotyping arrays have been used for many applications, including clinical diagnostics of chromosomal abnormalities, genome-wide association studies (GWAS), fine mapping of known loci, and linkage studies Verlouw *et al.* (2021). The number of SNPs tested in a single array varies from hundreds to millions of variants. The most advanced arrays are produced by Affymetrix/Thermofisher and Illumina (for instance, HumanOmniExpress-24 chip) companies Ha *et al.* (2014). The second approach consists of extracting information about variations from the whole genome sequencing (WGS) data. However, this method requires more computational power, time and verification, since the data is very diverse and extensive. The interpretation of abundant mutation information from WGS, especially non-coding and structure variants, requires the analysis of large-scale WGS data integrated with RNA-Seq, epigenomics, immuno-genomic and clinic-pathological information Nakagawa and Fujita (2018).

Nowadays, the CRISPR-Cas9 system is a long-standing trend with many projects going on. This technology allows to precisely edit the DNA molecule and it has been already demonstrated that the creation of synthetic RNA molecules and supplying them alongside Cas9 into the genome of eukaryotes, since distinct specific regions of the genome can be manipulated and targeted Abdelnour *et al.* (2021); Uddin *et al.* (2020). Thus the therapeutic potential of CRISPR-Cas9 and similar systems is very high, but facing many ethical problems. The SNV exploration could give us a hint of which variant might be essential for some diseases. And this is where CRISPR-Cas9 system can come in handy, because we can target the right parts of the genome, knowing which variants are more acceptable.

In this research, we will try to infer the most significant phenotypic traits, describe the most essential SNPs, and discuss potential applications of this information. As the experimental data, we will use the 23andMe genotyping data for a certain person (see Materials for the details).

## Materials and methods

### Data

In this research, we used the DNA-sequencing dataset of Mike Rayko by the 23andMe platform. This data contains information about 595401 different SNPs. The total genotyping rate is 0.989246. For each SNP there are details on chromosome, position, and a unique identifier. For the raw data, please, contact the authors.

### Data handling

To convert the 23andMe data format to the standard vcf format we used the Plink tool Chang *et al.* (2015); Shaun Purcell (2023) version v1.90b7 64-bit. Using it we removed all SNPs corresponding to deletions and insertions to make the file compatible with annotation tools.

### Haplogroup analysis

In order to assess the ethnicity of the given sample we analyzed maternal (mtDNA) and paternal (Y chromosome) haplotypes. For the mitochondrial data analyses we used the mthub platform van Oven and Kayser (2009). To identify Y-chromosome

haplogroups we used the website https://ytree.morleydna.com/extractFromAutosomal.

### Annotation

In this study, we performed an annotation for sex and eye color. To determine the sex of the patient we looked for genes on the Y chromosome in the data. Since these genes were found, we concluded that the sample was from a male person. To analyze variants related to eyes color we used the approach described in Hart *et al.* (2013).

For further annotation we used VEP (Variant Effect Predictor) tool with a corresponding version of the genome. For the description we used Snpedia database.

The code, raw data and results are available at https://github.com/grigorievaekaterina/BI_Project_5. You can also consult Supplementary section of this report for the full table of SNP variants.

## Results

Ethnicity analysis based on mitochondrial DNA showed that our sample carried the markers of 8 main haplogroups (see Table 1). Overall, there were 3270 markers found at 3268 positions covering 19.7% of mtDNA. Analysis of the Y chromosome SNPs showed 5 groups as the most probable for our data (see Table 3).

**Table 1** Main mtDNA haplogroups for the sample

| mtDNA | Area |
|---|---|
| H(T152C) | An umbrella group, Europe, Near East and in the Caucasus region Behar *et al.* (2012) |
| H1(T152C) | Spain, Iberia, Marocco, Sardinia Achilli *et al.* (2004) |
| H | Europe Ghezzi *et al.* (2005) analyze |
| H46 | Ireland and Germany Europedia (2022) |
| H69 | European/Caucasus Behar *et al.* (2012) |
| H16(T152C) | Primarily in Germanic countries and Poland Europedia (2022) |
| H3(T152C) | Spain, Gallia, Sardinia Achilli *et al.* (2004) |
| H52 | Britain and Russia (Karelia) Europedia (2022) |
| H9 | Italy Doan *et al.* (2017) |

You can find results related to the eyes and skin color in the Table 2. This combination of variants represents white skin color and a general European phenotype. However, the data is ambiguous in eye color, therefore we can only assume that person has not blue color of eyes.

**Table 2** SNPs associated with skin and eye color

| rs_id | gene | genotype | associated with (based on SNPedia) |
|---|---|---|---|
| rs12203592 | IRF4 | C/T | Primarily in Europeans; high sensitivity of skin to sun exposure |
| rs12913832 | HERC2 | A/G | brown eye color |
| rs12896399 | SLC24A4 | T/T | Lighter hair color & blue eyes more likely |
| rs16891982 | SLC45A2 | C/G | if European 7x more likely to have black hair |
| rs1426654 | SLC24A5 | A/A | probably light-skinned European ancestry |
| rs885479 | MC1R | G/G | darker skin color |

Overall results by the VEP tool can be seen on the Figure 2 and detailed table with SNP variants that were classified as risk-factors can be found in the supplements (Table 4). We identifyed the SNP related to the susceptibility to diseases like Coronary artery disease, Three Vessel Coronary Disease, Diabetes mellitus type 2, Dystonia and others. For the full information see 4.

## Discussion

Overall we can say that this person is male and probably European according to the ethnicity data obtained by the mitochondrial and Y chromosome data. The H haplotype has a wide area of coverage that was expanded thanks to multiple migration, but the main region is still central Europe (see Figure **??**). We can assume that the eye color is brown and the overall phenotype belongs to the European group.

We found out that there are several diseases with a high risk for this person, this diseases included many blood-related diseases such as Coronary artery disease and Three Vessel Coronary Disease (rs4977574, rs10757274, rs1024611). Apart from it we found several variants related to the insulin biology (rs1169288, rs13266634). The research Rosta *et al.* (2017) explains the protective effect of the minor SNPs alleles rs7578326(G) rs13266634(T) for Gestational diabetes mellitus. The former may affect insulin signaling, while the latter may affect zinc homeostasis, which, according to the authors of the article, may alter insulin production and beta-cell function.

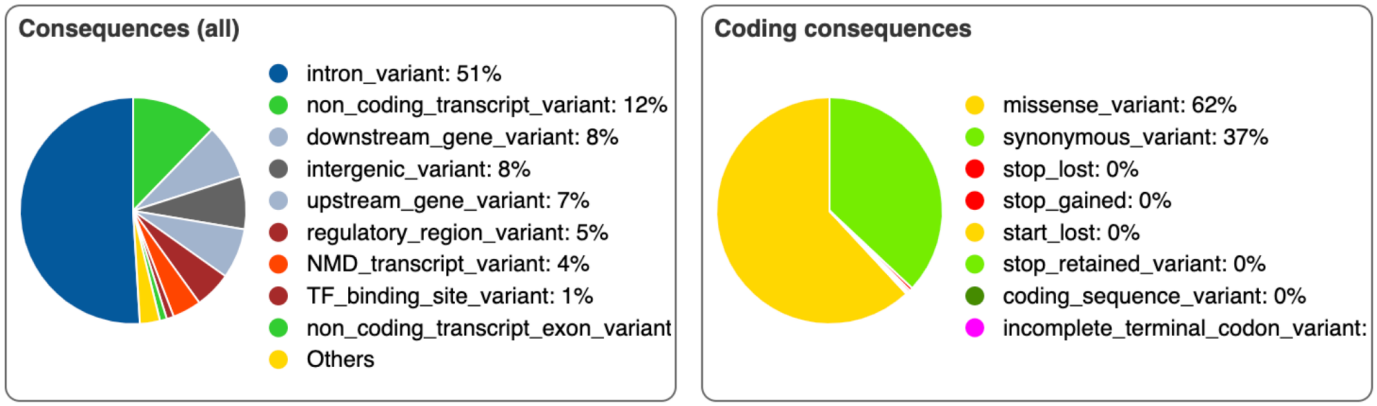


**Figure 1** Overall statistics on the genotyping data by VEP tool

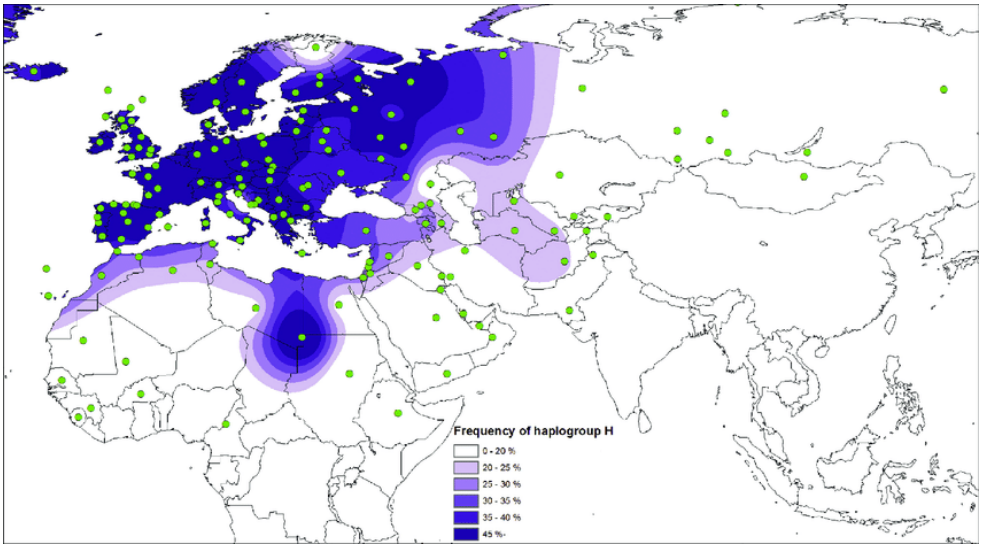


**Figure 2** Frequencies of mitochondrial haplogroup H in modern populations

**Table 3** Main Y chromosome haplogroups for the sample

| Y chromosome | Area Underhill *et al.* (2014) |
|---|---|
| R1a1a {[}R1a-L168 (R1a-M17, R1a-M198){]} | East and North Europe, Central Asia, Hindustan |
| M3 {[}M-P117 (M-P118){]} | New Britain, Melanesia |
| K {[}K-PF5504 (K-PF5493, K-PF5480){]} | Island South East Asia, Melanesia |
| R1b1a2a1a2a1b3∼2 {[}R1b-L421 (R1b-L433, R1b-L88){]} | Western Europe |
| N1a {[}N-M96 (N-CTS7095, N-P189){]} | Northern Eurasia |

## Acknowledgments

We thank Mike Rayko for supervising and supporting us, providing guidelines and recommendations.

## Literature cited

Abdelnour SA, Xie L, Hassanin AA, Zuo E, Lu Y. 2021. The potential of CRISPR/cas9 gene editing as a treatment strategy for inherited diseases. Frontiers in Cell and Developmental Biology. 9.

Achilli A, Rengo C, Magri C, Battaglia V, Olivieri A, Scozzari R, Cruciani F, Zeviani M, Briem E, Carelli V *et al*. 2004. 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. 75:910–918.

Behar DM, van Oven M, Rosset S, Metspalu M, Loogväli EL, Silva NM, Kivisild T, Torroni A, Villems R. 2012. A "copernican" reassessment of the human mitochondrial DNA tree from its root. The American Journal of Human Genetics. 90:675–684.

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. 4.

Doan K, Zachos FE, Wilkens B, Vigne JD, Piotrowska N, Stanković A, Jędrzejewska B, Stefaniak K, Niedziałkowska M. 2017. Phylogeography of the tyrrhenian red deer (cervus elaphus corsicanus) resolved using ancient DNA of radiocarbon-dated subfossils. Scientific Reports. 7.

Europedia. 2022. Europedia. https://www.eupedia.com/europe/Haplogroup$_{Hm}tDNA.shtml$.

Ghezzi D, Marelli C, Achilli A, Goldwurm S, Pezzoli G, Barone P, Pellecchia MT, Stanzione P, Brusa L, Bentivoglio AR *et al*. 2005. Mitochondrial DNA haplogroup k is associated with a lower risk of parkinson's disease in italians. European Journal of Human Genetics. 13:748–752.

Ha NT, Freytag S, Bickeboeller H. 2014. Coverage and efficiency in current SNP chips. European Journal of Human Genetics. 22:1124–1130.

Hart KL, Kimura SL, Mushailov V, Budimlija ZM, Prinz M, Wurmbach E. 2013. Improved eye- and skin-color prediction based on 8 SNPs. Croatian Medical Journal. 54:248–256.

Nakagawa H, Fujita M. 2018. Whole genome sequencing analysis for cancer genomics and precision medicine. Cancer Science. 109:513–522.

Rosta K, Al-Aissa Z, Hadarits O, Harreiter J, Nádasdi Á, Kelemen F, Bancher-Todesca D, Komlósi Z, Németh L, Rigó J *et al*. 2017. Association study with 77 SNPs confirms the robust role for the rs10830963/g of MTNR1b variant and identifies two novel associations in gestational diabetes mellitus development. PLOS ONE. 12:e0169781.

Shaun Purcell CC. 2023. Plink 1.9. www.cog-genomics.org/plink/1.9/.

Uddin F, Rudin CM, Sen T. 2020. CRISPR gene therapy: Applications, limitations, and implications for the future. Frontiers in Oncology. 10.

Underhill PA, Poznik GD, Rootsi S, Järve M, Lin AA, Wang J, Passarelli B, Kanbar J, Myres NM, King RJ *et al*. 2014. The phylogenetic and geographic structure of y-chromosome haplogroup r1a. European Journal of Human Genetics. 23:124–131.

van Oven M, Kayser M. 2009. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Human Mutation. 30:E386–E394.

Verlouw JAM, Clemens E, de Vries JH, Zolk O, Verkerk AJMH, am Zehnhoff-Dinnesen A, Medina-Gomez C, Lanvers-Kaminsky C, Rivadeneira F, Langer T *et al*. 2021. A comparison of genotyping arrays. European Journal of Human Genetics. 29:1611–1624.

## Supplementary data

**Table 4** SNPs and their meaning

| ID | Location | Allele | Desease name | Significance |
|---|---|---|---|---|
| rs1024611 | 17:32579788-32579788 | G | Coronary artery disease | Risk-Factor |
| rs1049296 | 3:133494354-133494354 | T | Transferrin variant c1/c2 | Association |
| rs10757274 | 9:22096055-22096055 | G | Three Vessel Coronary Disease | Risk-Factor |
| rs1169288 | 12:121416650-121416650 | C | Insulin resistance | Risk-Factor |
| rs12150220 | 17:5485367-5485367 | T | Vitiligo-associated multiple autoimmune disease susceptibility 1 | Risk-Factor |
| rs13266634 | 8:118184783-118184783 | T | Diabetes mellitus type 2, susceptibility to | Risk-Factor |
| rs1801197 | 7:93055753-93055753 | G | Bone mineral density quantitative trait locus 15 | Risk-Factor |
| rs1801274 | 1:161479745-161479745 | G | Lupus nephritis | Risk-Factor |
| rs1801275 | 16:27374400-27374400 | G | Atopy | Risk-Factor |
| rs1801394 | 5:7870973-7870973 | G | Down syndrome | Risk-Factor |
| rs1801968 | 9:132580901-132580901 | G | Dystonia 1 | Risk-Factor |
| rs2073658 | 1:161010762-161010762 | T | Hyperlipidemia | Risk-Factor |
| rs2184026 | 9:101304348-101304348 | T | Tobacco addiction | Risk-Factor |
| rs2241880 | 2:234183368-234183368 | G | Inflammatory bowel disease 10 | Risk-Factor |
| rs231775 | 2:204732714-204732714 | G | Hashimoto thyroiditis | Risk-Factor |
| rs4402960 | 3:185511687-185511687 | T | Diabetes mellitus type 2 | Risk-Factor |
| rs4880 | 6:160113872-160113872 | G | Microvascular complications of diabetes 6 | Risk-Factor |
| rs4961 | 4:2906707-2906707 | T | Hypertension | Risk-Factor |
| rs4977574 | 9:22098574-22098574 | G | Three Vessel Coronary Disease | Risk-Factor |
| rs5174 | 1:53712727-53712727 | T | Myocardial infarction 1 | Risk-Factor |
| rs5186 | 3:148459988-148459988 | C | Renal dysplasia | Benign |
| rs61747071 | 16:53720436-53720436 | T | Retinitis pigmentosa in ciliopathies | Risk-Factor |
| rs6265 | 11:27679916-27679916 | T | Memory impairment | Risk-Factor |
| rs6280 | 3:113890815-113890815 | T | Schizophrenia | Risk-Factor |
| rs6504649 | 17:48437456-48437456 | G | Pseudoxanthoma elasticum | Risk-Factor |
| rs699 | 1:230845794-230845794 | G | Preeclampsia | Risk-Factor |
| rs763110 | 1:172627498-172627498 | T | Lung cancer | Risk-Factor |
| rs7794745 | 7:146489606-146489606 | T | Autism | Risk-Factor |
| rs909253 | 6:31540313-31540313 | G | Myocardial infarction | Risk-Factor |