Title: Analyzing and suggesting improvements to a particular person's genome using SNPs data

Abstract: In this work we analyzed the SNPs data of a particular person obtained via 23andme service for possible improvements, including those mitigating disease risks and enhancing the organism. For each of the two categories, multiple SNPs of interest were analyzed and five modifications were suggested. Also, MT and Y haplotypes were analyzed and racial identity of the person was established.

Introduction:
Some years ago genetic testing cost thousands of dollars – a huge amount of money almost for nothing. However the past few years have seen enormous advances in genotyping technology, including chips that accommodate in excess of 1 million single nucleotide polymorphism (SNP) assays [1]. The use of genotyping chips made genomic analysis inexpensive, and nowadays there are many services that allow everyone to obtain raw genome data at an affordable price. If you have any level of curiosity about your genetic makeup such data could let you learn more about yourself, e.g. your family history or predisposition to illnesses. It is important personal information to have for everyone who wants to be well informed. This information also allows you to connect with genetic relatives that you have never met before that are also using the same (or any other) service. However, the amount of information that these services provide is very limited. In this article we analyzed raw genome data obtained from 23andme in order to get additional insight into the DNA and all the information it contains. We compared DNA sequence to the “standard” human DNA, found differences (polymorphisms) and made some suggestions based on these differences.

Methods:

Input data is a text file obtained directly from 23andme, with proprietary format.

plink v1.90b6.24 [2] command line tool was used to convert the input data to .vcf keeping SNPs only (flags `--recode vcf --out snps_clean --output-chr MT --snps-only just-acgt`).

Variant Effect Predictor (VEP) [3] web interface was used to annotate the obtained SNPs. For further analysis, SNPs annotated with either `risk_factor` or `likely_pathogenic` were considered.

Selected SNPs were then checked against the dbSNP database [4] to find ones with major effects. We also used SNPedia [5] to find promising SNPs not present in the current genome.

For MT haplogroup prediction, HaploGrep 2.1.21 [6] `classify` command line tool and mtHAP web interface (mtHAP version 0.19b (2015-05-11), haplogroup data version PhyloTree Build 17 (2016-02-18) +mods) [7] were used.

For Y haplogroup prediction, YSEQ Clade Finder [8] and MorleyDNA Subclade Predictor [9] web interfaces were used.

Results:

For target mtDNA sequence H haplogroup, particularly H(T152C), was predicted. H haplogroup is a predominantly European group, which probably evolved in Western Asia c. 25,000 years ago from the HV haplogroup. Another prediction was made with haplogrep – the H2a2a1 subgroup.

According to the Y-chromosome DNA sequence, R1a-M198 haplogroup was predicted with both web services. YSEQ Clade Finder could make an even deeper prediction – R1a1a1-M417 is the most likely haplogroup. It originated from R1a (M420) 22,000 - 25,000 years ago and then diversified into 2 subclades about 5800 years ago.

Unfortunately, both mtDNA and Y-DNA haplogroup prediction is not deep enough to establish ethnicity more precisely than that the considered individual is Caucasian.

Annotation of SNPs with VEP provides the following distribution of variant locations (fig. 1).

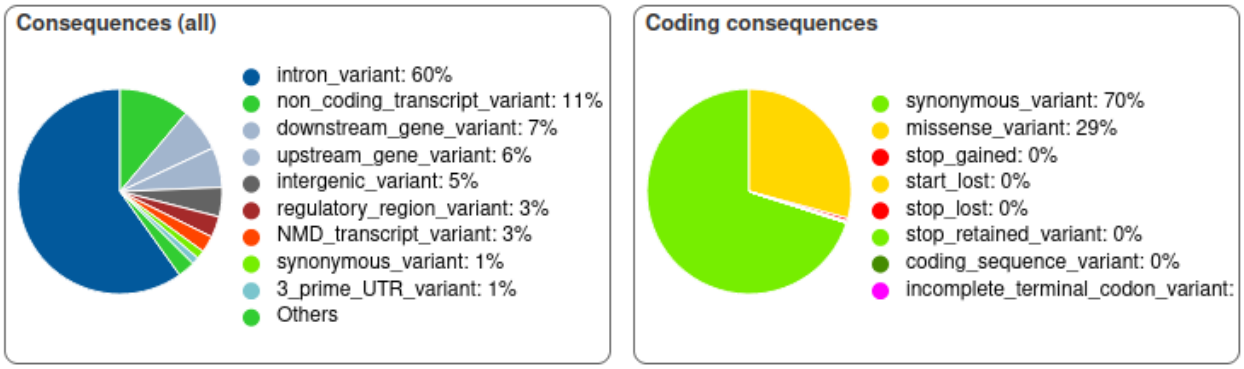

Figure 1. Distribution of variant locations predicted by VEP.

Based on our data investigation we suggest following corrections:

SNP ID(s)	Description
rs1024611	Susceptibility to Spina bifida, a birth defect that may affect the patient's children and require surgery after birth.
rs5174, rs909253	Risk factors of heart attack.
rs763110	Contribution to risk of lung cancer.
rs13266634, rs4402960	Susceptibility to type 2 diabetes.

rs1799986	Likely cause of Keratosis pilaris, a rather common skin condition.

And the following enhancements:

SNP ID(s)	Possible improvement
rs1406844918, rs121912617	Short sleeper – reduce sleep time needed without any known negative medical consequences.
rs1229984	0.56x decreased risk of oral/throat cancers, reduced risk for alcoholism.
rs17070145	Greatly increased memory performance.
rs13333226	~15-20% lower risk for hypertension or cardiovascular events
rs3843763	Normal high-density lipoprotein (HDL) cholesterol plasma levels.

Discussion:
In this article we analyzed SNP data of one human individual. Haplogroup predictions were made based on mtDNA and Y-chromosome DNA polymorphisms. SNPs data was annotated, and possible corrections and improvements were suggested. Although for many SNPs it is not completely clear how they work, for some it is possible to try to describe the mechanism of action.

Starting with major fixes, rs1024611 SNP leads to -2578A>G change in the promoter of the monocyte chemoattractant protein-1 MCP-1 CCL2 gene. Such polymorphism influences the production of its corresponding protein, a chemokine involved in inflammatory responses [10]. rs13266634, also known as Trp325Arg, is a SNP in the zinc transporter protein member 8 SLC30A8 gene that has primarily been associated with type-2 diabetes. The major alleles of the rs13266634 associate with reduced insulin secretion [11].

rs1229984 encodes a form of the alcohol dehydrogenase ADH1B gene that significantly reduces the clearance rate of alcohol from the liver. Polymorphism leads to Arg48His change, which increases ADH1B activity (meaning more rapid oxidation of ethanol to acetaldehyde). Individuals with one or especially two ADH1B*2 alleles are more likely to find drinking unpleasant and have a somewhat reduced risk for alcoholism [12].

Citations:

1. Ragoussis, J. (2009). Genotyping technologies for genetic research. Annual review of genomics and human genetics, 10, 117-133.

2. Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience, 4(1), s13742-015.
3. McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., ... & Cunningham, F. (2016). The ensembl variant effect predictor. Genome biology, 17(1), 1-14.
4. Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic acids research, 29(1), 308-311.
5. Cariaso, M., & Lennon, G. (2012). SNPedia: a wiki supporting personal genome annotation, interpretation and analysis. Nucleic acids research, 40(D1), D1308-D1312.
6. Weissensteiner, H., Pacher, D., Kloss-Brandstätter, A., Forer, L., Specht, G., Bandelt, H. J., ... & Schönherr, S. (2016). HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. Nucleic acids research, 44(W1), W58-W63.
7. mtHAP haplogroup analysis, online resource. https://dna.jameslick.com/mthap/
8. YSEQ Clade Finder, online resource. https://cladefinder.yseq.net/
9. MorleyDNA Subclade Predictor, online resource. https://ytree.morleydna.com/extractFromAutosomal
10. Kim, M. P., Wahl, L. M., Yanek, L. R., Becker, D. M., & Becker, L. C. (2007). A monocyte chemoattractant protein-1 gene polymorphism is associated with occult ischemia in a high-risk asymptomatic population. Atherosclerosis, 193(2), 366-372.
11. Staiger, H., Machicao, F., Stefan, N., Tschritter, O., Thamer, C., Kantartzis, K., ... & Häring, H. U. (2007). Polymorphisms within novel risk loci for type 2 diabetes determine β-cell function. PloS one, 2(9), e832.
12. Muramatsu, T., Zu-Cheng, W., Yi-Ru, F., Kou-Bao, H., Heqin, Y., Yamada, K., ... & Kono, H. (1995). Alcohol and aldehyde dehydrogenase genotypes and drinking behavior of Chinese living in Shanghai. Human genetics, 96(2), 151-154.