H+, or how to build a perfect human.

Authors: Anna Rybina, Daria Nikanorova

Abstract

In this study we analyzed 23andMe genome-wide data of a european man in order to pinpoint specific single nucleotide variants (SNVs), which may result in the development of complex disorders. After SNPs annotation and effects prediction we compared our data with ClinVar and GWAS catalog databases. Subsequent Y-chromosome and mtDNA haplotype calling might suggest that ancestors could be from Central Asia and/or Europe. We found several variants that were described previously to be associated with Type II Diabetes (rs7756992), heart disease (rs10757274 and rs2383206), osteoporosis (rs3736228) and several autoimmune diseases (rs12150220). In order to promote the health of the carrier we suggest fixing these mutations with the CRISPR/Cas approach. Moreover, we highlighted 5 variants that may significantly improve the well-being of a carrier, including HIV resistance (i3003626), ability to digest milk (rs4988235), and increased empathy (rs53576).

Introduction

Genotyping is a powerful approach, which allows pinpoint specific single nucleotide variants (SNPs) or other mutations in the human genome. Overall, SNP is a single based mutation in a particular locus, usually consisting of two alleles that is present in a sufficiently large fraction of the population (> 1%). Millions of human SNPs have been discovered over the last decades, including over 6 million from The International HapMap Project alone. In high-density oligonucleotide SNP arrays, fluorescently labelled target DNA hybridizes with allele-specific probes that are arrayed on a chip. The most significant clinical application of SNP arrays is determination of disease susceptibility. There are two main categories of SNPs: linked (or indicative) and causative. The linked SNPs are usually situated outside the gene and do not alter the protein function directly. Being phenotypically silent they still can be markers of a disease. By contrast, causative SNPs are located in the gene, changing amino acid sequence or amount of protein produced. Actually, the link between a particular SNP and a complex trait may be statistically confirmed during GWAS analysis, while the precise mechanism of causation may remain unknown.

Genome-wide association study (GWAS) aims to analyze a genome-wide set of variants, with particular focus on SNPs, across different individuals to reveal association of allele with a trait and/or a disease. Relationship between genetic variant and trait could be studied using gene-editing techniques such as CRISPR/Cas9 which allows the deletion of specific genomic regions with high accuracy (Wang, La Russa, and Qi 2016). CRISPR screening approach is based on genome-wide systematic knock down of genes (Koike-Yusa et al. 2014). Moreover, gene-editing tools could be used to examine the non-coding genome. For instance, in CRISPR-interference (CRISPRi), guide RNAs and impaired molecules of the Cas9 enzyme are used to block regulatory elements from binding their target genes (Qi et al. 2013). CRISPR-activation (CRISPRa) operates with transcriptional activator fused to the Cas9 protein to increase transcription (Bikard et al. 2013). Such techniques might be used to identify and validate functions of disease-associated regulatory elements.

Nowadays the tendency in commercialization of genetic testing (e.g. 23andMe) and declining costs of NGS could make genomics studies more available for ordinary people. In the transhumanist future, technologies such GWAS coupled with DNA-editing techniques could allow us to cure genetic diseases, make medicine much more personalized, enhance the human body, and even create “superhuman” with upgraded qualities.

In this study, we aimed to investigate 23andMe genome-wide data on genetic variants, focusing on SNPs, to reveal variants associated with clinical conditions and traits, and suggest some adjustments and improvements for the carrier of genetic material.

Methods

Raw 23andMe data was converted to vcf format using plink v1.90b6.21 (Purcell et al. 2007) with parameters: --output-chr MT --snps-only just-acgt, so that all variants corresponding to deletions and insertions were discarded. SNPs were annotated and effects were predicted with SnpEff v5.0e (Cingolani et al. 2012) using GRCh37.75 database. Resulting data was compared with ClinVar (Landrum et al. 2020) and GWAS catalog (Buniello et al. 2019)metadata using SnpSift. Information of revealed clinical and trait associations was extracted via bash, and verified using dbSNP (Buniello et al. 2019) and SNPedia (https://www.snpedia.com/index.php/SNPedia). Mitochondrial DNA (mtDNA) haplogroup was classified on the web server (https://dna.jameslick.com/mthap/) and via command-line tool haplogrep v2.2.9 (https://github.com/seppinho/haplogrep-cmd) with default parameters. The Cambridge Reference Sequence (CRS) was used as a reference. Y-chromosome haplogroup was identified on the web server https://ytree.morleydna.com/extractFromAutosomal and using command-line tool yhaplo v1.1.0 (https://github.com/23andMe/yhaplo) with default parameters.

Results

Raw data from 23andMe contained 610526 different variants. After discarding deletions and insertions, we obtained 595401 SNPs. Comparing our data with ClinVar, we revealed 862 variants of clinical significance, all of which occurred to be heterozygous. Out of them, 713 variants were non-synonymous. Based on GWAS catalog metadata, we found associated traits and/or diseases for 84 heterozygous genetic loci.

Based on 3270 markers at 3268 positions covering 19.7% of mitochondrial DNA (mtDNA), we identified that mtDNA belongs to H(T152C) haplogroup. The following markers were defined: four (750G, 1438G, 4769G, and 8860G) in the control region (CR) and two (152C and 263G) specifically in the hypervariable region II (HVR2).

We classified the Y-chromosome haplogroup as R1a1a1 (R-M417/R-Page-7) using 2084 SNPs located on the respective chromosome.

Analyzing SNPs, associated with medical conditions and some traits, we suggested several improvements and fixations for a carrier of genetic material (**Table 1**). We discuss possible effects of these changes below.

Table 1. Summarizing data on improvements and adjustments suggested for several SNP variants

rsID	genotype	fixed genotype	acquired feature
Adjustments			
rs3736228	CT	CC	Increased risk for osteoporosis
rs10757274	AG	AA	increased risk for heart disease
rs2383206	AG	AA	increased risk for heart disease
rs7756992	AG	AA	increased risk of Type II Diabetes
rs12150220	AT	TT	increased risk for several autoimmune diseases
Improvements			
rs53576	AG	GG	Optimistic and empathetic; handle stress well
rs4680	AG	GG	An advantage in the processing of aversive stimuli (warrior)
		AA	An advantage in memory and attention tasks (worrier)
i3003626	(I;I)	(D;D)	HIV resistance
rs4988235	GG	AA	Ability to digest milk
rs12913832	AG	GG	Blue eyes

Discussion

Improvements

rs53576 is responsible for G to A substitution in the intron of the oxytocin receptor (OXTR) gene. Oxytocin functions as both a hormone and neurotransmitter, having a broad range of influences on social and emotional state throughout the body and the brain. According to different studies, the polymorphism rs53576 is considered to be linked with social behavior and personality. Particularly, individuals with G allele are more empathetic, generally feel

less lonely and eventually become more sensitive parents (Kim et al. 2010; Rodrigues et al. 2009).

rs4680 is a variant in the catechol-O-methyltransferase (COMT) gene. This polymorphism leads to Val to Met substitution in the 158 position of a protein. COMT is an enzyme, which destroys dopamine in the prefrontal cortex. The most common wild-type allele G is coding for a valine amino acid, while the A allele results in amino acid substitution to a methionine. This substitution alters the structure of the COMT enzyme, leading to the increased level of dopamine in the prefrontal cortex. The outcome depends on the number of replaced alleles. Individuals with heterozygous genotype A/G have intermediate dopamine levels, while the carriers of two alleles acquire advantage in memory and attention tasks. Main drawbacks of such genotype are lower pain threshold and enhanced vulnerability to stress. This phenotype also refers to as “worrier” (Stein et al. 2006). By contrast, wild-type genotype G/G has its own advantages: better stress resiliency, higher pain threshold and advantage in the processing of aversive stimuli, which may be summarized as being a “warrior”. Another striking feature is being more exploratory in terms of trying new strategies, being overall more adventurous (Frank et al. 2009).

Not being a true SNP, rs333 (i3003626 in 23andMe results) is still extremely useful. Most commonly known as CCR5-Δ32, It is a deletion of 32 nucleotides in the CCR5 chemokine receptor gene. This deletion introduces a premature stop codon into the CCR5 receptor, altering its function, which leads to inability of HIV-1 virus to enter the cell. People with this region deleted in both alleles are highly resistant to HIV infection (Stein et al. 2006; Huang et al. 1996).

Another useful adaptation to acquire may be the ability to digest milk in adulthood. Lactose intolerance may be put down to decline in activity of the lactase-phlorizin hydrolase (LPH) in intestinal cells. LPH, coding by LCT gene, hydrolyzes lactose into glucose and galactose, but this function may be lost after childhood. It was found that a DNA variant rs4988235 is significantly associated with lactase non-persistence, despite being situated 14 kb upstream from the LCT locus (Enattah et al. 2002).

The last improvement we can advise is eye-color changing. One particular SNP rs12913832, localized in the first intron of the oculocutaneous albinism (OCA2) gene, is considered to be linked to the eye-color. As regards brown eyes, it is associated with the rs12913832(A;G) or (A;A) genotypes. Less popular blue eyes is a feature of both G alleles carriers (Eiberg et al. 2008).

Adjustments.

In our data, we detected two heterozygous variants A/G, rs10757274 and rs2383206, located near each other on the chromosome. Nucleotide substitutions rs10757274 and rs2383206, both leading to intron variants, were reported to be associated with the predisposition to coronary artery disease (CAD) (Xu et al. 2020) and ischemic stroke (Smith et al. 2009). For instance, rs10757274 increased the probability of coronary heart disease in a half of Caucasian population in ~15 to 20% cases (McPherson et al. 2007). Both SNPs

rs10757274 and rs2383206 are placed at the locus 9p21 in the gene CDKN2B-AS1 (or ANRIL). The CDKN2B-AS1 gene is located between CDKN2A and CDKN2B, known as tumor suppressor genes, which encode the inhibitors of cell-cycle kinases INK4/ARF. CDKN2B-AS1 gene product is CDKN2B antisense RNA 1, functional long non-coding RNA that could bind polycomb repressive complex-1 (PRC1) and -2 (PRC2), resulting in epigenetic silencing of other genes in this cluster (Burd et al. 2010). Probably, rs10757274 and rs2383206 affect ncRNA ANRI-mediated interference that regulates expression of INK4/ARF leading to increased risk of coronary artery disease. As carriers with homozygous variants AA have normal risk of heart medical conditions (McPherson et al. 2007), we would suggest G>A substitution as fixation.

Another adjustment that we propose concerns rs7756992 SNP in CDKAL1 gene leading to intron variant formation. At this locus, we identified an A/G variant which was shown to be significantly associated with lower insulin levels in type 2 diabetic patients (Tabara et al. 2009), (Omori et al. 2008)). CDKAL1 is required for catalyzing 2-methylthio modification of adenosine at position 37 of tRNALys(UUU). Reduced level of such modification could result in decreased insulin secretion (Zhou et al. 2014). SNP rs7756992 might impair splicing and formation of non-coding splicing variant CDKAL1-v1, which increases miRNA-mediated suppression of CDKAL1 as reduced level of CDKAL1-v1 could not outcompete miRNA binding to CDKAL1 (Zhou et al. 2014). Thus, we propose to substitute G nucleotide with the A which would restore the level of CDKAL1-v1 and regulation in normal conditions. (Tabara et al. 2009)

We found a missense variant A/T (rs12150220) in the gene NALP1 encoding NACHT leucine-rich-repeat protein 1, a regulator of the innate immune system. This SNP might be associated with increased risk for several autoimmune diseases, including Addison's disease and rheumatoid arthritis (Magitta et al. 2009). T > A substitution results in replacement of Leucine with Histidine at position 155 on amino acid sequence which could probably affect the function of native protein. We suggest constructing a homozygous variant T/T that might recover the functioning of NALP1 protein and reduce risk of autoimmune diseases.

Heterozygous variant rs3736228 (C/T) in the gene LRP5 is associated with increased risk for osteoporosis (van Meurs et al. 2008). LRP5 gene encodes low-density lipoprotein receptor–related protein 5 (LRP5) that is involved in bone mass accrual and susceptibility to osteoporosis. The substitution C > T leads to Ala1330Val mutation which is positioned within one of low-density lipoprotein (LDL) receptor domains of LRP5. It was demonstrated recently that the Val1330 variant decreased Wnt signaling activity relative to the Ala1330 variant (Kiel et al. 2007). Thus, we advise to replace T with C to restore homozygosity of the rs3736228 variant and decrease the chances of osteoporosis development.

Y-chromosome and mtDNA haplotypes.

Based on the results of mtDNA and Y-chromosome haplogroup calling, we might assume that ancestors of a studied individual were from Central Asia and/or Europe. R1a1a1 haplogroup was revealed among Near/Middle East and Caucasus populations (Underhill et al. 2015), including eastern Slavic populations (Mirabal et al. 2009). Descendants of this haplogroup exhibit geographic localization in Central and South Asia and Europe (Underhill et al. 2015). H2 haplogroup is widely spread in eastern Europe and the Caucasus

populations (Pereira et al. 2005). Distribution of H2a reaches Central Asia, resembling somewhat the phylogeography of Y-chromosomal Hg R1a (Loogväli et al. 2004).

Supplementary materials

Detailed pipeline with code is provided here.

References

Bikard, David, Wenyan Jiang, Poulami Samai, Ann Hochschild, Feng Zhang, and Luciano A. Marraffini. 2013. “Programmable Repression and Activation of Bacterial Gene Expression Using an Engineered CRISPR-Cas System.” *Nucleic Acids Research* 41 (15): 7429–37.

Buniello, Annalisa, Jacqueline A. L. MacArthur, Maria Cerezo, Laura W. Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, et al. 2019. “The NHGRI-EBI GWAS Catalog of Published Genome-Wide Association Studies, Targeted Arrays and Summary Statistics 2019.” *Nucleic Acids Research* 47 (D1): D1005–12.

Burd, Christin E., William R. Jeck, Yan Liu, Hanna K. Sanoff, Zefeng Wang, and Norman E. Sharpless. 2010. “Expression of Linear and Novel Circular Forms of an INK4/ARF-Associated Non-Coding RNA Correlates with Atherosclerosis Risk.” *PLoS Genetics* 6 (12): e1001233.

Cingolani, Pablo, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J. Land, Xiangyi Lu, and Douglas M. Ruden. 2012. “A Program for Annotating and Predicting the Effects of Single Nucleotide Polymorphisms, SnpEff: SNPs in the Genome of Drosophila Melanogaster Strain w1118; Iso-2; Iso-3.” *Fly* 6 (2): 80–92.

Eiberg, Hans, Jesper Troelsen, Mette Nielsen, Annemette Mikkelsen, Jonas Mengel-From, Klaus W. Kjaer, and Lars Hansen. 2008. “Blue Eye Color in Humans May Be Caused by a Perfectly Associated Founder Mutation in a Regulatory Element Located within the HERC2 Gene Inhibiting OCA2 Expression.” *Human Genetics* 123 (2): 177–87.

Enattah, Nabil Sabri, Timo Sahi, Erkki Savilahti, Joseph D. Terwilliger, Leena Peltonen, and Irma Järvelä. 2002. “Identification of a Variant Associated with Adult-Type Hypolactasia.” *Nature Genetics* 30 (2): 233–37.

Frank, Michael J., Bradley B. Doll, Jen Oas-Terpstra, and Francisco Moreno. 2009. “Prefrontal and Striatal Dopaminergic Genes Predict Individual Differences in Exploration and Exploitation.” *Nature Neuroscience* 12 (8): 1062–68.

Huang, Y., W. A. Paxton, S. M. Wolinsky, A. U. Neumann, L. Zhang, T. He, S. Kang, et al. 1996. “The Role of a Mutant CCR5 Allele in HIV-1 Transmission and Disease Progression.” *Nature Medicine* 2 (11): 1240–43.

Kiel, Douglas P., Serge L. Ferrari, L. Adrienne Cupples, David Karasik, Danielle Manen, Alma Imamovic, Alan G. Herbert, and Josée Dupuis. 2007. “Genetic Variation at the Low-Density Lipoprotein Receptor-Related Protein 5 (LRP5) Locus Modulates Wnt Signaling and the Relationship of Physical Activity with Bone Mineral Density in Men.” *Bone* 40 (3): 587–96.

Kim, Heejung S., David K. Sherman, Joni Y. Sasaki, Jun Xu, Thai Q. Chu, Chorong Ryu, Eunkook M. Suh, Kelsey Graham, and Shelley E. Taylor. 2010. “Culture, Distress, and Oxytocin Receptor Polymorphism (OXTR) Interact to Influence Emotional Support Seeking.” *Proceedings of the National Academy of Sciences of the United States of America* 107 (36): 15717–21.

Koike-Yusa, Hiroko, Yilong Li, E-Pien Tan, Martin Del Castillo Velasco-Herrera, and Kosuke Yusa. 2014. “Genome-Wide Recessive Genetic Screening in Mammalian Cells with a Lentiviral CRISPR-Guide RNA Library.” *Nature Biotechnology* 32 (3): 267–73.

Landrum, Melissa J., Shanmuga Chitipiralla, Garth R. Brown, Chao Chen, Baoshan Gu,

Jennifer Hart, Douglas Hoffman, et al. 2020. “ClinVar: Improvements to Accessing Data.” *Nucleic Acids Research* 48 (D1): D835–44.
Loogväli, Eva-Liis, Urmas Roostalu, Boris A. Malyarchuk, Miroslava V. Derenko, Toomas Kivisild, Ene Metspalu, Kristiina Tambets, et al. 2004. “Disuniting Uniformity: A Pied Cladistic Canvas of mtDNA Haplogroup H in Eurasia.” *Molecular Biology and Evolution* 21 (11): 2012–21.
Magitta, N. F., A. S. Bøe Wolff, S. Johansson, B. Skinningsrud, B. A. Lie, K-M Myhr, D. E. Undlien, et al. 2009. “A Coding Polymorphism in NALP1 Confers Risk for Autoimmune Addison’s Disease and Type 1 Diabetes.” *Genes and Immunity* 10 (2): 120–24.
McPherson, Ruth, Alexander Pertsemlidis, Nihan Kavaslar, Alexandre Stewart, Robert Roberts, David R. Cox, David A. Hinds, et al. 2007. “A Common Allele on Chromosome 9 Associated with Coronary Heart Disease.” *Science* 316 (5830): 1488–91.
Meurs, Joyce B. J. van, Thomas A. Trikalinos, Stuart H. Ralston, Susana Balcells, Maria Luisa Brandi, Kim Brixen, Douglas P. Kiel, et al. 2008. “Large-Scale Analysis of Association between LRP5 and LRP6 Variants and Osteoporosis.” *JAMA: The Journal of the American Medical Association* 299 (11): 1277–90.
Mirabal, Sheyla, Maria Regueiro, Alicia M. Cadenas, L. Luca Cavalli-Sforza, Peter A. Underhill, Dmitry A. Verbenko, Svetlana A. Limborska, and Rene J. Herrera. 2009. “Y-Chromosome Distribution within the Geo-Linguistic Landscape of Northwestern Russia.” *European Journal of Human Genetics: EJHG* 17 (10): 1260–73.
Omori, Shintaro, Yasushi Tanaka, Atsushi Takahashi, Hiroshi Hirose, Atsunori Kashiwagi, Kohei Kaku, Ryuzo Kawamori, Yusuke Nakamura, and Shiro Maeda. 2008. “Association of CDKAL1, IGF2BP2, CDKN2A/B, HHEX, SLC30A8, and KCNJ11 with Susceptibility to Type 2 Diabetes in a Japanese Population.” *Diabetes* 57 (3): 791–95.
Pereira, Luísa, Martin Richards, Ana Goios, Antonio Alonso, Cristina Albarrán, Oscar Garcia, Doron M. Behar, et al. 2005. “High-Resolution mtDNA Evidence for the Late-Glacial Resettlement of Europe from an Iberian Refugium.” *Genome Research* 15 (1): 19–24.
Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David Bender, Julian Maller, et al. 2007. “PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses.” *American Journal of Human Genetics* 81 (3): 559–75.
Qi, Lei S., Matthew H. Larson, Luke A. Gilbert, Jennifer A. Doudna, Jonathan S. Weissman, Adam P. Arkin, and Wendell A. Lim. 2013. “Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene Expression.” *Cell* 152 (5): 1173–83.
Rodrigues, Sarina M., Laura R. Saslow, Natalia Garcia, Oliver P. John, and Dacher Keltner. 2009. “Oxytocin Receptor Genetic Variation Relates to Empathy and Stress Reactivity in Humans.” *Proceedings of the National Academy of Sciences of the United States of America* 106 (50): 21437–41.
Smith, J. Gustav, Olle Melander, Håkan Lövkvist, Bo Hedblad, Gunnar Engström, Peter Nilsson, Joyce Carlson, Göran Berglund, Bo Norrving, and Arne Lindgren. 2009. “Common Genetic Variants on Chromosome 9p21 Confers Risk of Ischemic Stroke: A Large-Scale Genetic Association Study.” *Circulation. Cardiovascular Genetics* 2 (2): 159–64.
Stein, Dan J., Timothy K. Newman, Jonathan Savitz, and Rajkumar Ramesar. 2006. “Warriors versus Worriers: The Role of COMT Gene Variants.” *CNS Spectrums* 11 (10): 745–48.
Tabara, Yasuharu, Haruhiko Osawa, Ryuichi Kawamoto, Hiroshi Onuma, Ikki Shimizu, Tetsuro Miki, Katsuhiko Kohara, and Hideichi Makino. 2009. “Replication Study of Candidate Genes Associated with Type 2 Diabetes Based on Genome-Wide Screening.” *Diabetes* 58 (2): 493–98.
Underhill, Peter A., G. David Poznik, Siiri Rootsi, Mari Järve, Alice A. Lin, Jianbin Wang, Ben Passarelli, et al. 2015. “The Phylogenetic and Geographic Structure of Y-Chromosome Haplogroup R1a.” *European Journal of Human Genetics: EJHG* 23 (1): 124–31.
Wang, Haifeng, Marie La Russa, and Lei S. Qi. 2016. “CRISPR/Cas9 in Genome Editing and Beyond.” *Annual Review of Biochemistry* 85 (June): 227–64.

Xu, Lang-Biao, Yi-Qing Zhang, Nan-Nan Zhang, Biao Li, Jia-Yi Weng, Xiao-Yang Li, Wen-Chao Lu, et al. 2020. “Rs10757274 Gene Polymorphisms in Coronary Artery Disease: A Systematic Review and a Meta-Analysis.” *Medicine* 99 (3): e18841.

Zhou, Bo, Fan-Yan Wei, Narumi Kanai, Atsushi Fujimura, Taku Kaitsuka, and Kazuhito Tomizawa. 2014. “Identification of a Splicing Variant That Regulates Type 2 Diabetes Risk Factor CDKAL1 Level by a Coding-Independent Mechanism in Human.” *Human Molecular Genetics* 23 (17): 4639–50.