The proper implementation of SNP analysis results: building a perfect human

Kupaeva Daria, Ivanova Eugenia

Abstract

Rapid development of sequencing technologies and genetic engineering in recent years has expanded the horizons of practical application of genetic data. This paper considers one interesting usage of genetic data - human genome editing with the purpose of improvement in terms of health, skills or something else. Forgetting for a moment about ethics we analysed raw data from the 23andMe project in order to find the regions for correction with CRISPR-Cas9 technology. Using SnpEff/SnpSift tools along with Ensembl Variant Effect Predictor we performed functional annotation and effect prediction and extracted informative variants. Finally we revealed pathogenic SNPs and SNPs associated with risk of a certain disease onset and suggested several changes for the person upgrade.

Introduction

Since ancient times people have tried to know more about yourself and nowadays we have the capability to examine our genome in many different ways. It is possible to sequence the human genome completely or partially, for example, only exome or single genes. Apparently whole-genome sequencing provides a lot of information but is it always indispensable for routine examination of the diseases, risks and personal traits? Complex tests based on single nucleotide polymorphism (SNP) detection were developed to meet those needs[1]. SNP represents a change of one nucleotide that can cause a phenotypic manifestation from harmless or even useful traits of a person, like resistance to certain pathogens, to predisposition to clinical condition. SNP databases are renewed every day and provide more and more comprehensive characterisation of a person's genome regardless of the less price of those tests.

One of the popular technologies for getting information about SNPs is microarray. SNP microarrays belong to hybridization-based methods in which many thousands of probes corresponding to different SNPs with several alleles are located on the solid surface of a microchip or on the polystyrene beads[2]. DNA of a sample hybridizes with probes and fluorescent signals are detected from bonded probes after washing off non-specifically bonded sequences. It is a particular case of genotyping which measures more general genetic variation. In this investigation we used raw data from the 23andMe project where Illumina Infinium OmniExpress-24 kit is applied. It is a BeadChip array described as Genome-Wide Genotyping Array and High-Throughput Genotyping Array that provides an overview of the entire genome, enabling genome-wide discoveries and associations[3]. Infinium technology works on primer extension which is a two-step process. Firstly the probes hybridize on the sample sequences upstream from the SNP and then DNA polymerase extends the hybridized primer by adding SNP hapten-labeled nucleotides. This label is recognized by antibodies that are coupled to a detectable signal[4].

Obtaining the list of SNPs and using special tools one can easily analyze it and recognize the unpleasant variants. For now ethical rules forbid direct manipulations with the human genome. However the systems for genome editing already exist and widely used on other organisms. The CRISPR-Cas9 system originated in bacteria *Streptococcus pyogenes* as an

adaptive immune system against DNA viruses and plasmids[5]. Cas9 is an endonuclease bound to the guide RNA and associated with the Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR). The CRISPR acts as a storage of the sequences of the previously met hostile agents and Cas9 cleaves the matching DNA of the invader. People take over this gorgeous technology and modernize it for their needs. Current methods allow not only the cleavage of the target DNA but also direct change of the sequence by one-letter transformation[6] or insertion of a small sequence[7]. Probably in the future all these technologies will find an application in human genome editing in genetic disease treatment or human improvement.

Materials and methods

In the start of our study we have had raw SNP data in format 23andMe. We converted it to vcf format with the help of plink[8] tool (--out snps_clean, --output-chr MT --snps-only just-acgt options).
For annotation we used the SnpEff[9] tool that searches for the annotation of the SNPs in the human genome base, GRCh37 version. We selected clinical significant variants with the use of SnpSift[10] tools that perform search in ClinVar base.
Also we annotated our data with help of Ensembl web-server Variant Effect Predictor with default settings. We extracted variations, which were annotated as “risk_factor” or “pathogenic”, using standard bash tools (awk, grep, cut). Additional information about predicted clinical significant variations we got in the base dbSNP.
We have determined the patient's haplotype using haplogrep[11]. Also this result was confirmed with mthub[12].
For determining Y-chromosome haplogroup we used web-tool Y-SNP Subclade Predictor[13].

All the commands that we used in this project, presented in the Supplementary materials.

Results

From the results of the analysis of potentially clinically significant variables, we selected 52 SNPs, which clinical significance was annotated by Variant Effect Predictor as “risk_factor” or “pathogenic”. Of the 52 variations obtained, we discarded those that are not marked as “risk_factor” or “pathogenic” in the dbSNP database. All these observations presented in table 1. Full list of the extracted variations is shown in table 1 of the Supplementary materials.
Among the variations associated with clinical pathologies, we did not find rare polymorphisms (with a frequency of <20% according to GnomAD). We also did not find any homozygous variation.

Table 1 - Potentially pathogenic SNPs and SNPs referred to as risk factors.

SNP ID	Position	Nucleotide	Consequence	Significance	Diagnosis	Frequency (by GnomAD)
rs1024611	17:32579788-32	G	CCL2:	Pathogenic,	Spina_bifida,_s	0.28232

	579788		Promoter region	risk factor	usceptibility_to|Mycobacterium_tuberculosis,_susceptibility_to|Coronary_artery_disease,_modifier_of|Coronary_artery_disease,_development_of,_in_hiv	
rs12150220	17:5485367-5485367	T,	NLRP1 : Missense Variant	risk factor	Vitiligo-associated_multiple_autoimmune_disease_susceptibility_1	0.327519
rs13266634	8:118184783-118184783	T	SLC30A8 : Missense Variant	risk factor	Diabetes_mellitus_type_2,_susceptibility_to	0.265589
rs1801197	7:93055753-93055753	G	CALCR : Missense Variant	risk factor	Bone_mineral_density_quantitative_trait_locus_15	0.357403
rs1801275	16:27374400-27374400	G	IL4R : Missense Variant	risk factor	Atopy,_susceptibility_to	0.343606
rs2004640	7:128578301-128578301	T	IRF5 : Splice Donor Variant	Pathogenic, risk factor	Rheumatoid_arthritis|Systemic_lupus_erythematosus_10	0.430568 (ALFA)
rs2073658	1:161010762-161010762	T	USF1 : Intron Variant	risk factor	Hyperlipidemia,_familial_combined,_susceptibility_to	0.225855
rs4402960	3:185511687-185511687	T	IGF2BP2 : Intron Variant	risk factor	Diabetes_mellitus_type_2,_susceptibility_to	0.376215
rs5174	1:53712727-53712727	T	LRP8 : Missense Variant	risk factor	Myocardial_infarction_1	0.288519
rs7794745	7:146489606-146489606	T	CNTNAP2 : Intron Variant	risk factor	Autism_15	0.49374
rs909253	6:31540313-31540313	G	LTA : Intron Variant LOC100287329 : Intron Variant	risk factor	Myocardial_infarction|Psoriatic_arthritis,_susceptibility_to	0.395429
i6058764	16:27356203-27356203	G	IL4R: Missense Variant	Pathogenic,_protective	Acquired_immunodeficiency_syndrome,_slow_progression_to|Atopy,_resistance_to	No info in dbSNP

We also analyzed the mitochondrial haplogroup and the Y-haplogroup.
According to haplogrep data, the subject's mitochondrial group is H2a2a1. According to the familytreedna.com database, this haplotype is now most common in Germany, USA, England and Sweden[14].

Table 2 - Current distribution of haplogroup H2a2a1.

Maternal Origin*	Branch Participants H2a2a1	Downstream Participants H2a2a1 and Downstream (Excluding other Letters)	All Downstream Participants H2a2a1 and Downstream (Including other Letters)	Distribution
Germany	80	94	94	13.18%
United States	66	92	92	12.90%
England	64	92	92	12.90%
Sweden	32	89	89	12.48%
Norway	40	50	50	7.01%
Ireland	39	47	47	6.59%
United Kingdom	24	36	36	5.05%
France	27	29	29	4.07%
Finland	5	27	27	3.79%
Poland	21	21	21	2.95%

According to the web service ytree.morleydna.com, the subject's Y-DNA haplogroup is R1a1a. At the same time, a large number of replacements remained unanalyzed. The signs by which the subject was assigned to these groups were the presence of mutations M17, M198 / PF6238, M514, M515.
Full information on the SNPs that were used to determine the haplotype can be found in Figure 1.

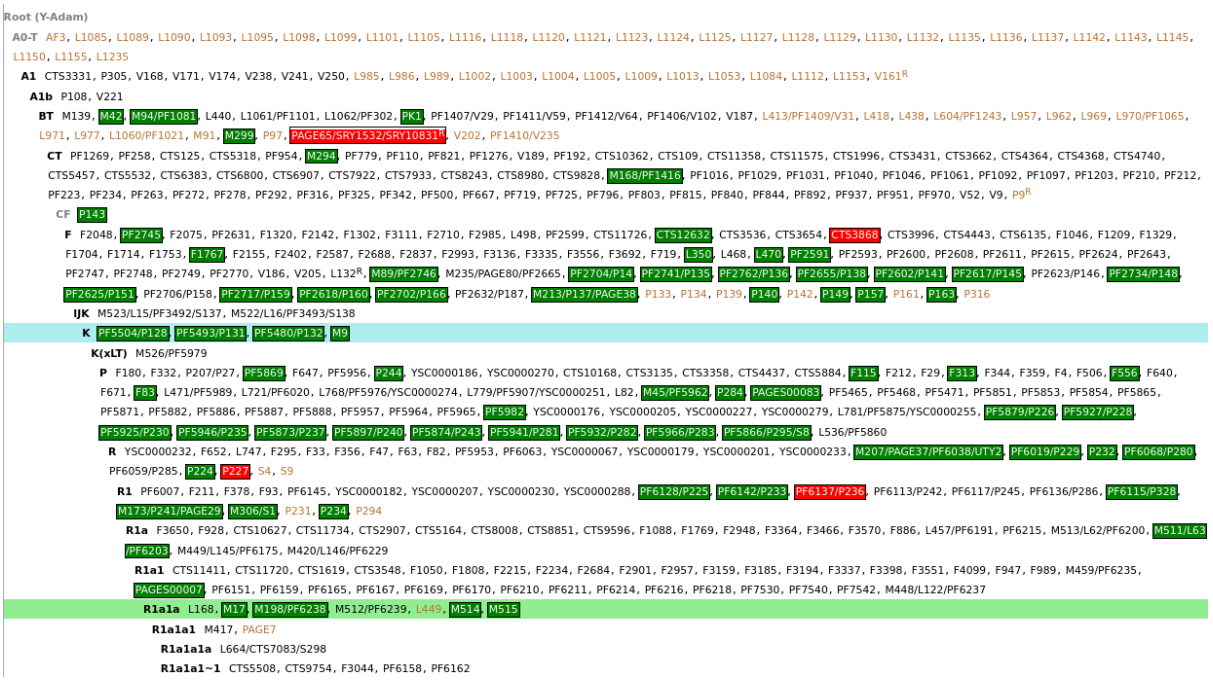

Figure 1. ISOGG tree for Y-DNA haplotype

Discussion

To analyze the most serious clinically significant polymorphisms, we need to take into account the frequency of substitution in the population, the effect of polymorphism on the risk of disease, the prevalence of the disease in the population (Table 3). All the substitutions we found are frequent and widespread in the population. All the substitutions we found, except for rs13266634, affect patients in both homozygous and heterozygous states.

Table 3 - Epidemiological information about potential diseases of the analyzed person.

Disease	**Epidemiology**
Autism	62 per 10 000[15]
Coronary heart disease	37,5% for men, 18,3% for women in USA[16]
Addison's disease	0.9 to 1.4 per 10 000[17]
Type 2 Diabetes	606 per 10 000[18]
Systemic lupus erythematosus	2-7 per 10,000[19]

Accordingly, the most serious clinically significant polymorphisms present in the subject are those associated with diseases common in the population is the rs1024611. The nucleotide changing is placed in the promoter region of the monocyte chemoattractant protein-1 coding gene and causes change in its production. It is associated with a 1.86 increased chance of coronary heart disease[20]. 37.5% of men and 18.3% of women in the USA suffer from ischemic heart disease and it is also one of the top 3 causes of death in the world[16]. This polymorphism affects patients with one allele of variation. This mutation is a good example of a mutation that should be removed from the subject's genome.

Another polymorphism that caught our attention is the rs13266634. This SNP represents a missense variant in the zinc transporter protein SLC30A8 coding gene and might affect protein function. It is associated with an increased risk of type 2 diabetes: 46% of offspring of type 2 diabetic patients homozygous for this gene have a predisposition to diabetes[21]. In a heterozygous state, the risk of acquiring type 2 diabetes with this variation is estimated at 2.5[22]. In combination with the high incidence of this disease in the population (606 per 10,000), this polymorphism is also undesirable and dangerous. In addition to decreasing the risks of diabetes we recommend the correction of the rs4402960 variant which referred to a significant 1.2x risk of type-2 diabetes in Japanese patients in heterozygous and homozygous states. It is positioned in the intron of Insulin-like growth factor 2 mRNA-binding protein 2 and might affect the process of splicing and therefore alter the IGF2 translation.

Although other clinically significant mutations are associated with diseases with a low incidence and probably do not contribute much to their occurrence, we would advise to pay attention to them - especially in the aspect of family planning.

SNP rs7794745 is also an intron located variant in the contactin-associated protein-like 2 which means that again it can affect the splicing. It is associated with increased risk for autism based on a study of 148 affected children from families with two more autistic children[23]. Although it is a more rare condition then diabetes, the number of affected people per 10000 is still quite sensible, so we choose this variant for the correction.

The rs2004640 is a splice donor variant in Interferon regulatory factor 5 protein coding gene which indicates the apparent mechanism of its action. Since the rs2004640 variant is one of several SNPs associated with systemic lupus erythematosus (SLE) we decided to discard this risk factor too. SLE is a severe autoimmune disease and this association has been observed in multiple populations (Caucasian, Asian, and African-Americans)[24] .

However none of these variants described in detail and their direct effect on the human organism remains unknown. With regard to all other variations, we have no reason to believe that they will have a significant effect on the subject, since they have a high level of prevalence in the population, the frequency of the disease associated with them in the population is low, and, moreover, they do not significantly affect the risks of the disease.

If we talk about new variations that, in our opinion, could be useful for the research subject, then we propose, first of all, to turn to polymorphisms that affect the patient's susceptibility to HIV. The most significant contribution to resistance is provided by RS333, especially in the homozygous form[25] . In addition, there are suggestions that it provides resistance to smallpox[26] . Besides positive predictions, there are also reports of its correlation with abdominal aortic aneurysm[27]. However, in Russia, where the patient lives, the HIV epidemic is currently continuing (1,476,023 cases, including those who died in 2020, per 144,500,000 of the population of Russia)[28], the risks of contracting HIV significantly exceed the risks of abdominal artery aneurysm.

We also believe that the rs28931573 mutation in the APOA1 gene, which is involved in the metabolism of low density lipoproteins, is useful. This variation is associated with a reduced risk of developing arthrosclerosis[29].

Next we decided to provide two changes in variant rs4680 for a person's choice. It is a missense mutation in the COMT gene involved in dopamine metabolism[30]. Currently the person has AG genotype, and it is possible to create AA or GG variation. AA person would have lower COMT enzymatic activity, therefore higher dopamine levels; lower pain threshold, enhanced vulnerability to stress, yet also more efficient at processing information under most conditions. GG person, conversely, would have higher COMT enzymatic activity, therefore lower dopamine levels; higher pain threshold, better stress resiliency, albeit with a modest reduction in executive cognition performance under most conditions.

Also we have found a SNP that reduces the risk of type-2 diabetes - rs3816873 in MTTP gene. MTTP encodes the large subunit of the heterodimeric microsomal triglyceride transfer protein[31]. Our person has genotype CT in this position and to protect him a bit more we can edit one allele and get CC.

Playing with appearance we can try to change the hair color. The rs3829241 variant with AA genotype corresponds to blonde hair. Changing the one allele nucleotide in this position (the person has AG genotype) on adenine could lead to alteration in hair color[32].

Conclusion

Hereby we briefly examined a human genome using microarray approach and revealed a huge area for the investigation and confirmation because many of the described SNPs have questionable connection to the corresponded diseases. Despite this the science progresses by large steps and in the nearest future we will collect sufficient information to predict diseases more precisely and cure them by genome editing.

1. Kwok PY, Chen X. Detection of single nucleotide polymorphisms. Curr Issues Mol Biol. 2003 Apr;5(2):43-60. PMID: 12793528.
2. Harbron S; Rapley R (2004). Molecular analysis and genome discovery. London: John Wiley & Sons Ltd. ISBN 978-0-471-49919-0
3. https://www.illumina.com/products/by-type/microarray-kits/infinium-omni-express.html
4. Gunderson KL, Steemers FJ, Ren H, Ng P, Zhou L, Tsan C, Chang W, Bullis D, Musmacker J, King C, Lebruska LL, Barker D, Oliphant A, Kuhn KM, Shen R. Whole-genome genotyping. Methods Enzymol. 2006;410:359-76. doi: 10.1016/S0076-6879(06)10017-8. PMID: 16938560.
5. Heler R, Samai P, Modell JW, Weiner C, Goldberg GW, Bikard D, Marraffini LA. Cas9 specifies functional viral targets during CRISPR-Cas adaptation. Nature. 2015 Mar 12;519(7542):199-202. doi: 10.1038/nature14245. Epub 2015 Feb 18. PMID: 25707807; PMCID: PMC4385744.
6. Gaudelli, N., Komor, A., Rees, H. et al. Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. Nature 551, 464–471 (2017). https://doi.org/10.1038/nature24644
7. Roy, K., Smith, J., Vonesch, S. et al. Multiplexed precision genome editing with trackable genomic barcodes in yeast. Nat Biotechnol 36, 512–520 (2018). https://doi.org/10.1038/nbt.4137

8. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559-575. doi:10.1086/519795
9. Cingolani P, Platts A, Wang le L, et al. 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 (Austin). 2012;6(2):80-92. doi:10.4161/fly.19695
10. Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, Ruden DM, Lu X. Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. Front Genet. 2012 Mar 15;3:35. doi: 10.3389/fgene.2012.00035. PMID: 22435069; PMCID: PMC3304048.
11. Weissensteiner H, Pacher D, Kloss-Brandstätter A, et al. HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. Nucleic Acids Res. 2016;44(W1):W58-W63. doi:10.1093/nar/gkw233
12. https://dna.jameslick.com/mthap/
13. https://ytree.morleydna.com/extractFromAutosomal
14. https://www.familytreedna.com/public/mt-dna-haplotree/H;name=H2a2
15. Elsabbagh M, Divan G, Koh YJ, et al. Global prevalence of autism and other pervasive developmental disorders. Autism Res. 2012;5(3):160-179. doi:10.1002/aur.239
16. Sanchis-Gomar F, Perez-Quilis C, Leischik R, Lucia A. Epidemiology of coronary heart disease and acute coronary syndrome. Ann Transl Med. 2016;4(13):256. doi:10.21037/atm.2016.06.33
17. Brandão Neto RA, de Carvalho JF. Diagnosis and classification of Addison's disease (autoimmune adrenalitis). Autoimmun Rev. 2014;13(4-5):408-411. doi:10.1016/j.autrev.2014.01.025
18. Khan MAB, Hashim MJ, King JK, Govender RD, Mustafa H, Al Kaabi J. Epidemiology of Type 2 Diabetes - Global Burden of Disease and Forecasted Trends. J Epidemiol Glob Health. 2020;10(1):107-111. doi:10.2991/jegh.k.191028.001
19. Danchenko N, Satia JA, Anthony MS. Epidemiology of systemic lupus erythematosus: a comparison of worldwide disease burden. Lupus. 2006;15(5):308-318. doi:10.1191/0961203306lu2305xx
20. Kim MP, Wahl LM, Yanek LR, Becker DM, Becker LC. A monocyte chemoattractant protein-1 gene polymorphism is associated with occult ischemia in a high-risk asymptomatic population. Atherosclerosis. 2007;193(2):366-372. doi:10.1016/j.atherosclerosis.2006.06.029
21. Boesgaard TW, Zilinskaite J, Vänttinen M, et al. The common SLC30A8 Arg325Trp variant is associated with reduced first-phase insulin release in 846 non-diabetic offspring of type 2 diabetes patients--the EUGENE2 study. Diabetologia. 2008;51(5):816-820. doi:10.1007/s00125-008-0955-6
22. Omori S, Tanaka Y, Takahashi A, et al. Association of CDKAL1, IGF2BP2, CDKN2A/B, HHEX, SLC30A8, and KCNJ11 with susceptibility to type 2 diabetes in a Japanese population. Diabetes. 2008;57(3):791-795. doi:10.2337/db07-0979
23. Arking DE, Cutler DJ, Brune CW, Teslovich TM, West K, Ikeda M, Rea A, Guy M, Lin S, Cook EH, Chakravarti A. A common genetic variant in the neurexin superfamily member CNTNAP2 increases familial risk of autism. Am J Hum Genet. 2008 Jan;82(1):160-4. doi: 10.1016/j.ajhg.2007.09.015. PMID: 18179894; PMCID: PMC2253968.

24. https://www.snpedia.com/index.php/Rs2004640
25. Huang Y, Paxton WA, Wolinsky SM, et al. The role of a mutant CCR5 allele in HIV-1 transmission and disease progression. Nat Med. 1996;2(11):1240-1243. doi:10.1038/nm1196-1240
26. Galvani AP, Slatkin M. Evaluating plague and smallpox as historical selective pressures for the CCR5-Delta 32 HIV-resistance allele. Proc Natl Acad Sci U S A. 2003;100(25):15276-15279. doi:10.1073/pnas.2435085100
27. Ghilardi G, Biondi ML, Battaglioli L, Zambon A, Guagnellini E, Scorza R. Genetic risk factor characterizes abdominal aortic aneurysm from arterial occlusive disease in human beings: CCR5 Delta 32 deletion. J Vasc Surg. 2004;40(5):995-1000. doi:10.1016/j.jvs.2004.08.014
28. http://www.hivrussia.info
29. Wang L, Tian F, Arias A, Yang M, Sharifi BG, Shah PK. Comparative Effects of Diet-Induced Lipid Lowering Versus Lipid Lowering Along With Apo A-I Milano Gene Therapy on Regression of Atherosclerosis. J Cardiovasc Pharmacol Ther. 2016;21(3):320-328. doi:10.1177/1074248415610216
30. https://www.snpedia.com/index.php/Rs4680
31. Rubin D, Helwig U, Pfeuffer M, Schreiber S, Boeing H, Fisher E, Pfeiffer A, Freitag-Wolf S, Foelsch UR, Doering F, Schrezenmeir J. A common functional exon polymorphism in the microsomal triglyceride transfer protein gene is associated with type 2 diabetes, impaired glucose metabolism and insulin levels. J Hum Genet. 2006;51(6):567-574. doi: 10.1007/s10038-006-0400-y. Epub 2006 May 24. PMID: 16721486.
32. https://www.snpedia.com/index.php/Rs3829241