How to build a perfect professor (instruction for beginners)

Nastya Dudkovskaya[1,2*†] and Bogdan Sotnikov[2,3*†]

[1*]International Laboratory of Statistical and Computational Genomics, HSE, 34 Tallinskaya ul, Moscow, 100190, Russia.
[2*] Bioinformatics institute, Kantemirovskaya street 2A, Saint Petersburg, 197342, Russia.
[3*] Medical faculty, Kyrgyz-Russian Slavic university, Kievskaya street 44, Bishkek, 720000, Kyrgyzstan.

*Corresponding author(s). E-mail(s): n-dudkovskaya@mail.com; bogdan.sotnikov.1999@mail.com;
†These authors contributed equally to this work.

Abstract

Our work is performed in consistency with the main postulates of transhumanism and the Earth Planet Transhumanism Party(EPTP). The main goals of EPTP are protection from unhealthy habits, enlargement of life quality and longevity. Using genome sequencing technologies we found five pathogenic-like SNPs and five loci with potential for improvement in the genome of patient M.R. We have changed clinically dangerous SNPs: rs4961, rs3780422, rs7034200, rs10974944, rs755383 for the neutral alleles. It will decrease the risk of arterial hypertension, diabetes mellitus, myeloproliferative neoplasms, testicular cancer and nicotine addiction. We added SNPs rs6046, rs5355, rs7412, rs10989591. They are responsible for improving memory skills in older age and decreasing the probability of Alzheimer's disease and arterial hypertension. We also realised another part of the EPTP - collecting genealogical data about all citizens. We detected, our patient's haplogroup is R1a1a, distributed from Scandinavia and Central Europe to southern Siberia and South Asia.

Keywords: Genome editing, ethnicity identifying, 8-plex eye-colour identifying protocol

1 Introduction

Nowadays, there are a lot of services that could provide information about anybody's genome. It can be done by a whole genome sequencing obtained by an Next-Generation Sequencing (NGS) instrument or just like a collection of Single Nucleotide Polymorphisms (SNPs) obtained using a genotyping chips. Mostly genotyping is using to predict likelihood of having some phenotypic trait, and, more important, likelihood of disease.

Most popular genotyping chips provided by 'Illumina'. There are several options for human's genotyping chips, such as Illumina HumanOmniExpress-24, that used by 23&Me and could test about 700k known SNPs. It finds single nucleotide variants, SNPs which must be present in at least 1% of the population. The main SNPs search methods use microarrays.

According to a genotype information, person could change their lifestyle to prevent development of diseases or feel better. It is a good way to make a life more pleasant, but, to be honest, the best way is to modify genome to get away with such struggles. And there exists a method that could provide it - CRISPR-Cas9 protocol. This development let people change genome - cut, remove and insert the necessary nucleotides.

CRISPR, Clustered Regularly Interspaced Palindromic Repeats, is the immune memory of bacteria that store information about viruses that have been infected by; Cas9 is a nuclease, an enzyme that can cut DNA. This enzyme is directed to a specific segment in the DNA of a bacteriophage, determined by CRISPR, sits on it and cuts like scissors, which disrupts the reproduction of the virus. This bacterial approach scientists use in the same sense to change DNA in other organisms. [1]

In our work we want to find SNPs from human's DNA and extract from this data answers on the questions: "Where do we come from?", "Who are we?" and "Where are we going?" as information about haplogroup, appearance and suggest some modifications that may make life of this human more pleasant.

2 Methods

2.1 Raw data

We have used the 23andMe DNA sequencing results of patient M.R. by courtesy of him. Data are available by the below-written link The data was a list of SNPs and their features in the plain text format, so we have converted them to vcf format, using plink (version 1.90b6.21) [2, 3]. We have used plink with the next options (for better perception options are listed in a column):

```
plink --23file <23andMe_file_name>
--recode vcf
```

```
--out snps_clean
--output-chr MT
--snps-only just-acgt
```

For exact compatibility we used GRCh37 version release of human genome assembly for annotation and SNPs' selection because our 23andMe sequencing results were made after August 9, 2012 according with 23andMe rules.

2.2 Haplogroups identification

For haplogroup identification, we use [4] tool. As our data provided by 23&Me, we chose this option during using, FTDNA format (transferred Geno 2.0 results) and Experimental Tree as a Tree version.

2.3 Identification of appearance

For identification of the patient's sex, we used below-written bash script:

```
cat <vcf_file_with_SNPs> |
awk '{print $1}' |
sort -u |
grep-v "#"
```

If we found Y chromosome we may conclude, the patient is a man, else – she is a woman.

For the identification of eye and skin colour, we have used an 8-plex system protocol, which was described in Hart et al. publication in detail [5]. We have looked at SNPs rs12913832, rs12203592, rs16891982, rs6119471 and rs12896399 for eye colour identification. SNPs rs12913832, rs16891982, rs1426654, rs 1545397 and rs885479 were used for skin colour identification.

2.4 Clinically significant SNPs

For clinically significant SNPs selection and further description, we used tools snpEff (version 5.1d) [6] and SnpSift (version 5.1d) [7]. We have downloaded the GRCh37.75 release of the human genome, and have annotated the converted vcf file, using snpEff (for rational space using we started the process with -Xmx4G option, other parameters were default). We have filtered the output of snpEff with the aid of SnpSift with default options. Sources of information were the ClinVar database and the EBI GWAS catalogue.

For filtering only clinically significant data we used the grep utility. We include only rows with ”CLNDN” and ”CLNVC=single_nucleotide_variant” and exclude ”Benign”, ”Uncertain_significance”, ”Likely_benign”, ”CLNDN=not_provided”, ”CLNREVSTAT=no_assertion_criteria_provided” and ”Congenital”. We have filtered only SNPs with clinical diagnosis (outside of polymorphisms with ”not provided” diagnosis). We have excluded

Table 1 SNPs responsible for the eye colour

SNP	Genotype	Interpretation
rs12913832	AG	Not blue
rs12203592	CT	Ambigous
rs16891982	CG	Ambigous
rs6119471	Absent	Absent
rs12896399	GG	Brown

benign, likely benign SNPs and SNPs with uncertain significance because the main goal of our investigation is improving the patient's genome (excluding pathogenic and adding "well" mutations). In the cause, our patient is an adult we have excluded all mutations, which determined congenital disease. We didn't use in further analysis mutation without assertion criteria and non-SNP mutations.

For GWAS-filtered data, we have used only one ("GWASCAT_TRAIT") filter for select clinically significant mutations.

The next stage of the investigation was SNPs' hand sorting and information searching. For searching data about selected SNPs' we have used NCBI Entrez through package reutils (version 0.2.3) for R (version 4.2.2) and SNPedia with recursion links from it.

2.5 Other

All the tools (plink, snpEff and SnpSift) were downloaded from bioconda [8]. Every tool (except SnpSift) was installed in its own conda virtual environment. More detailed information about technical processes may be found in laboratory journals.

3 Results

3.1 Where do we come from

As a result of using `morleydna.com` tool, the most likely haplogroup for Mikhail Raiko is R1a1a; R1a-L168 (R1a-M17, R1a-M198). It is distributed in a large region in Eurasia, extending from Scandinavia and Central Europe to southern Siberia and South Asia.

3.2 Who are we

We have identified that person, whose genome data we analysed, is a man because we have found a Y chromosome in his sequencing data.

Also, we analysed eye and skin colour-responsible polymorphism. Results showed in tables 1 and 2.

Table 2 SNPs responsible for the skin colour

SNP	Genotype	Interpretation
rs12913832	AG	Ambigous
rs1426654	AA	Light/non-dark
rs16891982	CG	Ambigous
rs1545397	Absent	Absent
rs885479	GG	Ambigous
rs6119471	Absent	Absent

Table 3 Disease-associated SNPs-candidates for potential replacing

SNP	Genotype	Gene	Effect
rs4961	GT	ADD1	Increased risk for high blood pressure
rs3780422	CT	GABBR2	Increased risk for nicotine addiction
rs7034200	AC	GLIS3	Increased risk for type 2 diabetes melitius
rs10974944	CG	JAK2	Increased risk for myeloproliferative neoplasms
rs755383	CT	DMRT1	Increased risk for testicular cancer

Table 4 SNPs-candidates for potential replacing that provide advantages

SNP	VG	BG	Gene	Effect
rs6046	GA	AA	F7	Decreased blood pressure levels
rs5355	TG	TT	SELE	Decreased blood pressure levels
rs7412	CC	TT	APOE	Lower risk for Alzheimer's disease
rs10989591	GG	TT	NR3A	Better episodic memory in older age

We can conclude, that our patient has brown eyes, but we can't predict his skin colour, because we didn't arrive at any criteria from the appropriate article.

3.3 Where are we going

We selected five SNPs for further replacement. They are briefly described in table 3. More detailed information about affected products and their clinical significance is situated in the next section of the paper.

Also, we selected five SNPs that could be replaced and give some advantages. They are briefly described in table 4. More detailed information about affected products and their clinical significance is situated in the next section of the paper. VG is for viewed genotype, BG is for 'Better' genotype.

4 Discussion

In this section, we will describe probable mechanisms of action for disease-related and advantage-related genes.

4.1 Correction of disease-related genes

4.1.1 Hypertension

rs4961 is a SNP in adducin 1 (ADD1) gene. ADD1 is responsible for many processes in the human organism. Some of them are cell signal transduction, regulation of actin cytoskeleton and ion transport across the cell membrane. The last function is been of interest to us in the context of hypertension. It has been shown that Gly460Trp polymorphism is associated with primary hypertension and faster proximal tubular resorption through the activation of Na,K-ATPase [9]. GT genotype is responsible for at least 1.8-fold increasing arterial hypertension risk. It has been shown in many investigations [10, 11].

We have changed this SNP nucleotide sequence, specifically GT to GG, which hasn't increased the risk of arterial hypertension.

4.1.2 Nicotine addiction

rs3780422 is located in GABBR2 - gene of gamma-aminobutyric acid type B receptor Subunit 2 [12]. It performs a receptor for gamma-aminobutyric acid - a neurotransmitter in CNS. Ligand-receptor cooperation activates GIRK-type potassium channels [13]. The GABA neurons are part of the mesolimbic dopamine system, critically important in mediating the reinforcing properties of nicotine, alcohol and drugs of abuse [14]. Mutations in this locus may lead to increasing in nicotine addiction probability [15, 16]

The correction of this locus may decrease the probability of having nicotine addiction in our patient. It is important for his health because smoking leads to chronic obstructive pulmonary disease and lung cancer.

4.1.3 Diabetes mellitius

rs7034200 is a SNP in a GLIS3 gene [17]. GLIS3 encodes the transcription factor GLIS family zinc finger 3, which activates or represses transcription and participates in beta cell ontogeny [18]. GLIS3 also plays a role in β-cell survival and likely in insulin secretion[19]. It is nominally associated with impaired proinsulin-to-insulin conversion and insulin secretion [20]. CA and CC genotypes enlarge risk of diabetes melitius[21].

We have changed this SNP nucleotide sequence, specifically CA to AA, which doesn't increase the risk of diabetes mellitius.

4.1.4 Myeloproliferative neoplasms

One of the polymorphisms in JAK2 gene is rs10974944. It is a non-receptor tyrosine kinase that activates cytokine-mediated signals by the JAK–STAT signal pathway [22]. JAK/STAT pathway plays a central role in the signaling of cytokines by regulating cell proliferation, survival, differentiation and immune response. Consequently, JAK/STAT pathway also has an important role in oncogenesis [23]. Different mutations in components of JAK/STAT pathway and in particular JAK2 lead to neoplasms development. rs10974944 was associated with the development of V617F-positive myeloproliferative neoplasms [24]. G allele of rs10974944 significantly increased the chance of getting essential thrombocythemia, primary myelofibrosis, and polycythemia vera (most common variants myeloproliferative neoplasms) [25].

We have corrected the genotype of the patient and converted this locus from GC to CC state.

4.1.5 Testicular cancer

rs755383 is situated into DMRT1 gene sequence. This gene's function is a differentiation of Sertoli cells and germ cells of the testis [26]. Some polymorphisms of the above-written SNP may cause testicular germ-cell tumor development [27]. The risk allele for this disease is C [28], so we have changed CT genotype to TT.

4.2 Adding advantages

4.2.1 Decreased blood pressure levels

From researches it only could be seen an epistatic interaction that associated with Decreased blood pressure levels between rs6046 and rs5355 with a quite significant statistical proof [29].

4.2.2 Lower risk for Alzheimer's disease

Rs7412 is a SNP in genome of the apolipoprotein E (ApoE), which is a protein involved in the metabolism of fats in the body of mammals. A subtype is implicated in Alzheimer's disease and cardiovascular disease. APOE interacts significantly with the low-density lipoprotein receptor (LDLR), which is essential for the normal processing (catabolism) of triglyceride-rich lipoproteins. The most preferable isoform of this protein is ϵ_2, which consist rs7412(T;T) and rs429358(T;T) genotypes, and the last is already contained in Mikhail's genome. This mutation provides ϵ_2 isoform of this protein, which associated with less risk for of Alzheimer's disease.

4.2.3 Better episodic memory in older age

Variations of the dopamine D2 receptor gene (rs6277, C957T) and the N-methyl-D-aspartate 3A (NR3A) gene, coding for the N-methyl-D-aspartate 3A subunit of the glutamate N-methyl-D-aspartate receptor (rs10989591, Val362Met), modulate a reliable gene-gene interaction, which was observed in older adults only: older individuals carrying genotypes associated with greater

D2 and N-methyl-D-aspartate receptor efficacy showed better episodic performance [30]. In our case person has rs10989591 (C;C) variant, whereas we need change nucleotides in rs6277 to provide better episodic memory in older age.

Declarations

- All authors declare that they have no conflicts of interest.
- All authors also declare, they are tired a lot.

References

[1] Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A., Charpentier, E.: A programmable dual-rna;guided dna endonuclease in adaptive bacterial immunity. Science

[2] Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., Sham, P.C.: PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet **81**(3), 559–575 (2007)

[3] Purcell, S. http://pngu.mgh.harvard.edu/purcell/plink/

[4] Morley, C. https://ytree.morleydna.com/extractFromAutosomal

[5] Hart, K.L., Kimura, S.L., Mushailov, V., Budimlija, Z.M., Prinz, M., Wurmbach, E.: Improved eye- and skin-color prediction based on 8 SNPs. Croat Med J **54**(3), 248–256 (2013)

[6] Cingolani, P., Platts, A., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., Ruden, D.M.: 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 (2012)

[7] Cingolani, P., Patel, V.M., Coon, M., Nguyen, T., Land, S.J., Ruden, D.M., Lu, X.: Using drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, snpsift. Frontiers in Genetics **3** (2012)

[8] ning, B., Dale, R., din, A., Chapman, B.A., Rowe, J., Tomkins-Tinch, C.H., Valieris, R., ster, J.: Bioconda: sustainable and comprehensive software distribution for the life sciences. Nat Methods **15**(7), 475–476 (2018)

[9] Glorioso, N., Filigheddu, F., Cusi, D., Troffa, C., Conti, M., Natalizio, M., Argiolas, G., Barlassina, C., Bianchi, G.: alpha-Adducin 460Trp allele is associated with erythrocyte Na transport rate in North Sardinian primary hypertensives. Hypertension **39**(2 Pt 2), 357–362 (2002)

[10] Zhang, Y., Chang, P., Liu, Z.: Single Nucleotide Polymorphisms Are Associated With Essential Hypertension Among Han and Mongolian Population in Inner Mongolia Area. Front Genet **13**, 931803 (2022)

[11] Watanabe, Y., Metoki, H., Ohkubo, T., Katsuya, T., Tabara, Y., Kikuya, M., Hirose, T., Sugimoto, K., Asayama, K., Inoue, R., Hara, A., Obara, T., Nakura, J., Kohara, K., Totsune, K., Ogihara, T., Rakugi, H., Miki, T., Imai, Y.: Accumulation of common polymorphisms is associated with development of hypertension: a 12-year follow-up from the Ohasama study. Hypertens Res **33**(2), 129–134 (2010)

[12] Jugessur, A., Wilcox, A.J., Murray, J.C., Gjessing, H.K., Nguyen, T.T., Nilsen, R.M., Lie, R.T.: Assessing the impact of nicotine dependence genes on the risk of facial clefts: An example of the use of national registry and biobank data. Nor Epidemiol **21**(2), 241–250 (2012)

[13] Jones, K.A., Borowsky, B., Tamm, J.A., Craig, D.A., Durkin, M.M., Dai, M., Yao, W.J., Johnson, M., Gunwaldsen, C., Huang, L.Y., Tang, C., Shen, Q., Salon, J.A., Morse, K., Laz, T., Smith, K.E., Nagarathnam, D., Noble, S.A., Branchek, T.A., Gerald, C.: GABA(B) receptors function as a heteromeric assembly of the subunits GABA(B)R1 and GABA(B)R2. Nature **396**(6712), 674–679 (1998)

[14] Cui, W.Y., Seneviratne, C., Gu, J., Li, M.D.: Genetics of GABAergic signaling in nicotine and alcohol dependence. Hum Genet **131**(6), 843–855 (2012)

[15] Beuten, J., Ma, J.Z., Payne, T.J., Dupont, R.T., Crews, K.M., Somes, G., Williams, N.J., Elston, R.C., Li, M.D.: Single- and multilocus allelic variants within the GABA(B) receptor subunit 2 (GABAB2) gene are significantly associated with nicotine dependence. Am J Hum Genet **76**(5), 859–864 (2005)

[16] Wei, J., Chu, C., Wang, Y., Yang, Y., Wang, Q., Li, T., Zhang, L., Ma, X.: Association study of 45 candidate genes in nicotine dependence in Han Chinese. Addict Behav **37**(5), 622–626 (2012)

[17] Dou, H.Y., Wang, Y.Y., Yang, N., Heng, M.L., Zhou, X., Bu, H.E., Xu, F., Zhao, T.N., Huang, H., Wang, H.W.: Association between genetic variants and characteristic symptoms of type 2 diabetes: A matched case-control study. Chin J Integr Med **23**(6), 415–424 (2017)

[18] e, V., Chelala, C., Duchatelet, S., Feng, D., Blanc, H., Cossec, J.C., Charon, C., Nicolino, M., Boileau, P., Cavener, D.R., res, P., Taha, D., Julier, C.: Mutations in GLIS3 are responsible for a rare syndrome with neonatal diabetes mellitus and congenital hypothyroidism. Nat Genet **38**(6), 682–687 (2006)

[19] Wen, X., Yang, Y.: Emerging roles of GLIS3 in neonatal diabetes, type 1 and type 2 diabetes. J Mol Endocrinol **58**(2), 73–85 (2017)

[20] Wagner, R., Dudziak, K., fer, S.A., Machicao, F., Stefan, N., Staiger, H., ring, H.U., Fritsche, A.: Glucose-raising genetic variants in MADD and ADCY5 impair conversion of proinsulin to insulin. PLoS One **6**(8), 23639 (2011)

[21] Miranda-Lora, A.L., az, M., Cruz, M., nchez-Urbina, R., guez, N.L., nez, B., nder, M.: Genetic polymorphisms associated with pediatric-onset type 2 diabetes: A family-based transmission disequilibrium test and case-control study. Pediatr Diabetes **20**(3), 239–245 (2019)

[22] Sopjani, M., Morina, R., Uka, V., Xuan, N.T., rmaku-Sopjani, M.: JAK2-mediated Intracellular Signaling. Curr Mol Med **21**(5), 417–425 (2021)

[23] Vainchenker, W., Constantinescu, S.N.: JAK/STAT signaling in hematological malignancies. Oncogene **32**(21), 2601–2613 (2013)

[24] Tanaka, M., Yujiri, T., Ito, S., Okayama, N., Takahashi, T., Shinohara, K., Azuno, Y., Nawata, R., Hinoda, Y., Tanizawa, Y.: JAK2 46/1 haplotype is associated with JAK2 V617F-positive myeloproliferative neoplasms in Japanese patients. Int J Hematol **97**(3), 409–413 (2013)

[25] Ngoc, N.T., Hau, B.B., Vuong, N.B., Xuan, N.T.: JAK2 rs10974944 is associated with both V617F-positive and negative myeloproliferative neoplasms in a Vietnamese population: A potential genetic marker. Mol Genet Genomic Med **10**(10), 2044 (2022)

[26] Zarkower, D., Murphy, M.W.: DMRT1: An Ancient Sexual Regulator Required for Human Gonadogenesis. Sex Dev **16**(2-3), 112–125 (2022)

[27] Lessel, D., Gamulin, M., Kulis, T., Toliat, M.R., Grgic, M., Friedrich, K., Zunec, R., Balija, M., rnberg, P., Kastelan, Z., gel, J., Kubisch, C.: Replication of genetic susceptibility loci for testicular germ cell cancer in the Croatian population. Carcinogenesis **33**(8), 1548–1552 (2012)

[28] Poynter, J.N., Hooten, A.J., Frazier, A.L., Ross, J.A.: Associations between variants in KITLG, SPRY4, BAK1, and DMRT1 and pediatric germ cell tumors. Genes Chromosomes Cancer **51**(3), 266–271 (2012)

[29] Functional epistatic interaction between rs6046 g / a in f7 and rs5355 c / t in sele modifies systolic blood pressure levels. PLoS One. 2012;7(7):e40777

[30] Papenberg, G., Li, S.-C., Nagel, I.E., Nietfeld, W., Schjeide, B.-M., Schröder, J., Bertram, L., Heekeren, H.R., Lindenberger, U., Bäckman, L.: Dopamine and glutamate receptor genes interactively influence episodic

memory in old age. Neurobiology of Aging **35**(5), 1213–312138 (2014). https://doi.org/10.1016/j.neurobiolaging.2013.11.014