


RESEARCH

Open Access



# Optimizing the genetic prediction of the eye and hair color for North Eurasian populations

Elena Balanovska<sup>1,2†</sup>, Elena Lukianova<sup>3†</sup>, Janet Kagazezheva<sup>1,3,4</sup>, Andrey Maurer<sup>5</sup>, Natalia Leybova<sup>6</sup>, Anastasiya Agdzhoyan<sup>1,3</sup>, Igor Gorin<sup>3,7</sup>, Valeria Petrushenko<sup>3,7</sup>, Maxat Zhabagin<sup>8</sup>, Vladimir Pylev<sup>1</sup>, Elena Kostryukova<sup>9</sup> and Oleg Balanovsky<sup>1,2,3\*</sup> 

From 11th International Young Scientists School “Systems Biology and Bioinformatics” – SBB-2019 Novosibirsk, Russia. 24-28 June 2019

## Abstract

**Background:** Predicting the eye and hair color from genotype became an established and widely used tool in forensic genetics, as well as in studies of ancient human populations. However, the accuracy of this tool has been verified on the West and Central Europeans only, while populations from border regions between Europe and Asia (like Caucasus and Ural) also carry the light pigmentation phenotypes.

**Results:** We phenotyped 286 samples collected across North Eurasia, genotyped them by the standard HirisPlex-S markers and found that predictive power in Caucasus/Ural/West Siberian populations is reasonable but lower than that in West Europeans. As these populations have genetic ancestries different from that of West Europeans, we hypothesized they may carry a somewhat different allele spectrum. Thus, for all samples we performed the exome sequencing additionally enriched with the 53 genes and intergenic regions known to be associated with the eye/hair color. Our association analysis replicated the importance of the key previously known SNPs but also identified five new markers whose eye color prediction power for the studied populations is compatible with the two major previously well-known SNPs. Four out of these five SNPs lie within the *HERC2* gene and the fifth in the intergenic region. These SNPs are found at high frequencies in most studied populations. The released dataset of exomes from Russian populations can be further used for population genetic and medical genetic studies.

**Conclusions:** This study demonstrated that precision of the established systems for eye/hair color prediction from a genotype is slightly lower for the populations from the border regions between Europe and Asia than for the West Europeans. However, this precision can be improved if some newly revealed predictive SNPs are added into the panel. We discuss that the replication of these pigmentation-associated SNPs on the independent North Eurasian sample is needed in the future studies.

**Keywords:** Population genetics, Exome sequencing, Gene pools, Pigmentation, DNA markers, Eye color, Hair color, Appearance

\* Correspondence: [balanovsky@inbox.ru](mailto:balanovsky@inbox.ru)

†Balanovska Elena and Lukianova Elena contributed equally to this work.

<sup>1</sup>Research Centre for Medical Genetics, Moscow, Russia

<sup>2</sup>Biobank of North Eurasia, Moscow, Russia

Full list of author information is available at the end of the article



© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

## Background

Predicting the eye and hair color from DNA became an important part of forensic genetic investigation. The genome-wide association studies [1–3] identified some key genes and sites within these genes which influence the pigmentation of the eye and hair color, as well as skin color [4, 5] phenotypes. These genes have been widely used for predicting pigmentation from genotype, mainly in the forensic context [6–8]. The most important sites have been included into HIRISplex-S system [9–12]. Genotyping the 24 markers (SNPs and indel) [10] allows the rapid and reliable prediction of the eye and hair color (HIRISplex system); additional 17 markers predict the skin color as well (HIRISplex-S system).

The prediction has been shown to be reliable for populations of European descent and the system itself has been developed on European populations (mainly on Dutch). Its precision for populations from other regions of the world has not been tested extensively. Most non-European populations have brown eyes and dark hair only. However, some populations from border regions between Europe and Asia populations (for example, groups from Altai region in Siberia, some populations from the Caucasus) are also known to carry lighter eye/hair phenotypes and these populations do not exhibit close relation with West Europeans on the genome level [13]. Even populations from Ural region, being more related to West Europeans, than Caucasus and West Siberians, are nevertheless much more genetically distant from Dutch than populations of Ireland, Poland, and Greece, used for verification of the HIRISplexSystem [10, 11]. It is therefore possible that some Asian populations carry alleles of the pigmentation-related genes which are not included in HIRISplex-S but affects the appearance phenotypes in these populations. If this is a case, such additional alleles might be useful for eye/color prediction in these populations and have therefore practical importance for genetic forensic investigations in Russia, or when investigating individuals of Russian ancestry in Europe.

We aimed to estimate the precision of HIRISplex-S on different populations from North Eurasia, to search for new alleles within known pigmentation genes, and to estimate the impact of these alleles on the eye and hair color. To do so, we collected the DNA samples and photos from 300 individuals from indigenous communities from Russia and neighboring countries. The sampling covered European part of Russia, Caucasus region, Kazakhstan, and some populations from various parts of Siberia. We performed exome sequencing rather than genotyping to be able identify alleles which were not reported previously and therefore have not been included into the GWAS panels. As many key SNPs are known to be located in intronic regions, we developed the custom exome panel which includes both, exonic and intronic

regions of the 53 genes and intergenic regions known to be involved in the pigmentation traits.

## Results and discussion

### Assembling the dataset

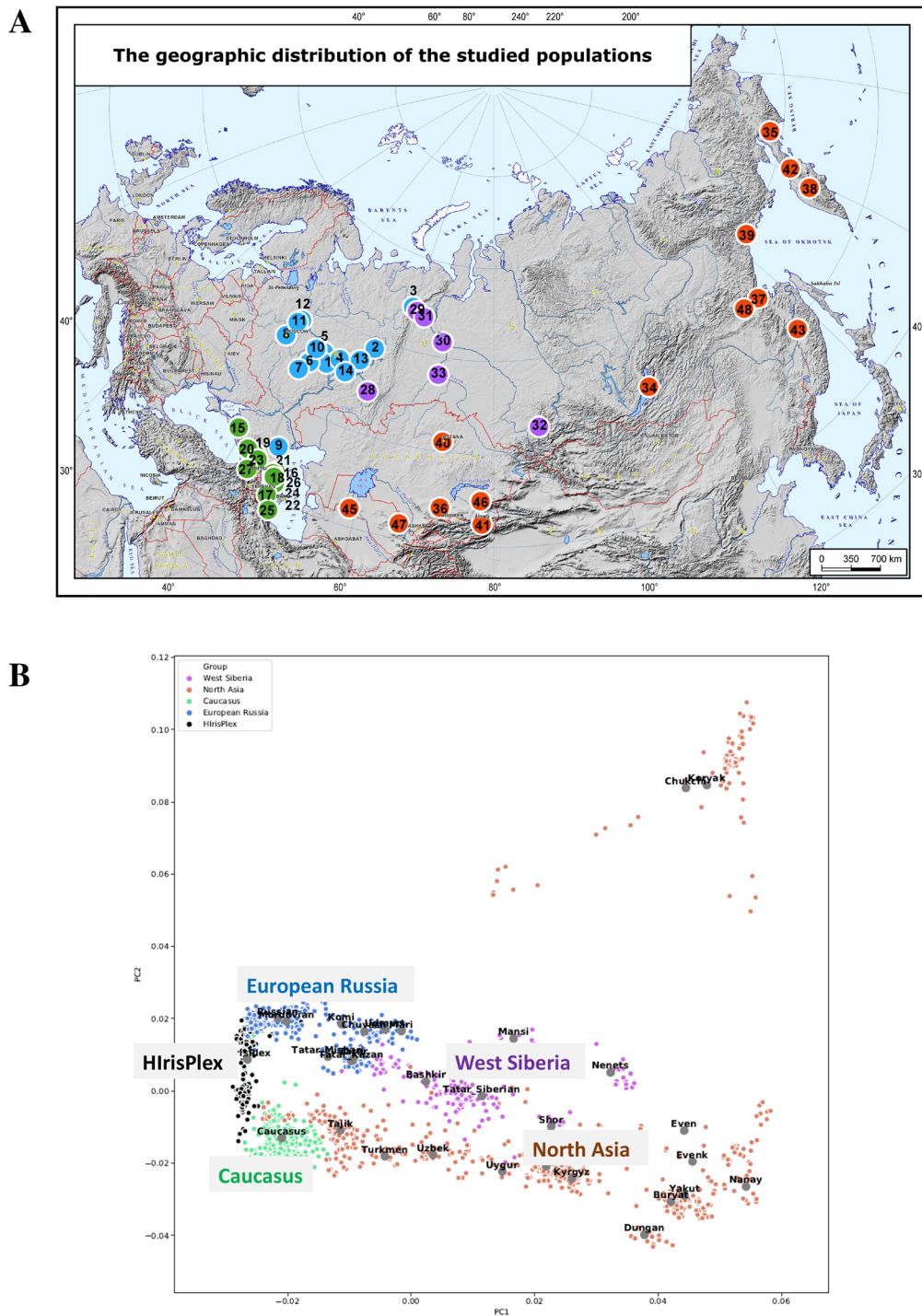
We phenotyped 300 individuals from 48 populations of Russia and neighboring countries by identifying their eye and hair colors. Independent phenotyping by three experts and availability of photos for revisiting made the phenotyping reliable and reproducible. Populations were grouped into four regional datasets: European Russia, West Siberia, Caucasus, and North Asia; Fig. 1a presents the sampling locations and grouping into the regional datasets. In correspondence with the large area sampled, the regional metapopulations have contrasting genetic background. We performed the PC analysis of the populations included into this study to illustrate these findings (Fig. 1b). We note, that the populations on which the HIRISplex-S has been developed and validated (Dutch, Polish, Irish, and Greek) occupy the narrow zone on the “western” extreme of the PC plot, while populations present in our study, particularly North Asian, Caucasus and West Siberia are pronouncedly different from West Europeans and from one another. Thus, all downstream analyses were performed for each regional dataset and for the pooled dataset.

DNA samples from these 300 individuals were sequenced using the specially designed exome capture which included, in addition to the standard Roche exome capture, the intronic and intergenic regions known to carry pigmentation-related polymorphic sites (see Methods for details).

The combined dataset included phenotypic calls and genotypic calls for all individuals. Phenotypic calls included five categories of the hair darkness, three categories of the hair redness, and five categories of the eye darkness. Genotypic calls included genotypes of all polymorphic sites identified within the 53 genes and intergenic regions known to be involved in eye/hair pigmentation. The downstream analyses were performed on the subsets of this combined dataset.

### Validating the precision of HIRISplex on north Eurasian populations

We started with estimating the precision of standard eye/hair prediction system in the newly phenotyped populations. From the combined dataset we extracted the phenotypic and genotypic calls for 24 SNPs included in the HIRISplex-S. Then we predicted the eye and hair color from genotypes using the online HIRISplex-S tool and compared the predicted phenotypes with the real phenotypes (Table 1). Table 2 presents the results for eye color prediction in different metapopulations (excluding the North Asia where the frequency of light eyes



**Fig. 1** The studied populations. Panel **a**: The map of the studied populations. Numbers on the map refers to the following studied populations: 1 - Chuvashes, 2 - Komi Permyaks, 3 - Komi Zyrians, 4 - Mari Meadow, 5 - Mari Mountain, 6 - Mordvins Erzya, 7 - Mordvins Moksha, 8 - Russians, 9 - Russians Nekrasov's Cossacs, 10 - Russians of Nizhny Novgorod region, 11 - Russians of Tver region, 12 - Russians of Yaroslavl'sky region, 13 - Udmurts, 14 - Volga Tatars, 15 - Adyghe, 16 - Avars, 17 - Azeri, 18 - Dargins, 19 - Kabardinians, 20 - Karachays, 21 - Kumyks, 22 - Lezgins, 23 - Ossets, 24 - Rutuls, 25 - Talysh, 26 - Tsakhur, 27 - Turks Meskhetian, 28 - Bashkirs, 29 - Forest Nenets, 30 - Khanty, 31 - Mansi, 32 - Shors, 33 - Siberian Tatars, 34 - Buryats, 35 - Chukchis, 36 - Dungans, 37 - Evenks of Far East, 38 - Evens of Kamchatka, 39 - Evens of Okhotsk coast, 40 - Kazakhs, 41 - Kirghiz, 42 - Koryaks, 43 - Nanais, 44 - Tajiks, 45 - Turkmens, 46 - Uyghurs, 47 - Uzbeks, 48 - Yakuts of Far East. Panel **b**: The principal components plot for this study populations and for the populations used for Hlris-plex-S developing/validation. Hlris-plex populations are in black. Colors refers to the regional datasets present on the Panel A

**Table 1** The AUC and accuracy of the eye color prediction using HirisPlex-S system and North Eurasian set of SNPs for the pooled North Eurasian dataset

	AUC				Accuracy		
	HirisPlex-S on West/Central European populations	HirisPlex-S on North Eurasian populations	North Eurasian SNPs (7 SNPs for eye and 11 SNPs for hair)	North Eurasian SNPs (36 SNPs for eye and 33 SNPs for hair)	HirisPlex-S on North Eurasian populations	North Eurasian SNPs (7 SNPs for eye and 11 SNPs for hair)	North Eurasian SNPs (36 SNPs for eye and 33 SNPs for hair)
Blue eye	0,94	0,93 (93)	0,96 (93)	0,9 (93)	0,86 (93)	0,83 (93)	0,97 (93)
Intermediate eye	0,74	N/A (6)	N/A (6)	N/A (6)	N/A (6)	N/A (6)	N/A (6)
Brown eye	0,95	0,93 (190)	0,86 (190)	0,97 (190)	0,86 (190)	0,79 (190)	0,98 (190)
Red hair	0,93	0,84 (18)	0,91 (18)	0,92 (18)	0,95 (18)	0,97 (18)	0,97 (18)
Blond hair	0,81	0,81 (40)	0,79 (40)	0,8 (40)	0,84 (40)	0,85 (40)	0,94 (40)
Brown hair	0,74	0,65 (70)	0,76 (70)	0,74 (70)	0,66 (70)	0,74 (70)	0,8 (70)
Dark hair	0,86	0,88 (156)	0,92 (156)	0,89 (156)	0,75 (156)	0,86 (156)	0,92 (156)

Note: number of samples in each phenotypic class is indicated in the parentheses

is low). We found (Table 1, Additional file 1) that the AUC value in the pooled North Eurasian dataset is only slightly lower than in the West/Central Europeans (particularly for the brown and red hair). However, when we analyzed the results for each region separately (Table 2), we found that performance of HirisPlex-S panel for predicting eye color is lower for individuals from Caucasus region (AUC values are 0.83 and 0.78, for blue and dark eyes). Especially, the recall for blue eyes in Caucasus is significantly lower in comparison with the other North Eurasian regions - only 47% (Additional file 2). It might indicate that genes of the pigmentation metabolic pathways in the Caucasus populations carry allele spectrum somewhat different from that in Europe. When partitioning the dataset according to the phenotypic class (Table 1 and Table 2) we found that predicting the both, blue and brown eyes in Russian population is much less effective. In particular, the HirisPlex-S systems tends to misclassify blue eyes as brown.

**Eye and hair color prediction in north Eurasian populations: searching for new informative alleles. The general workflow**

Our genetic data on the phenotyped individuals included the full sequencing of the pigmentation-associated genes

and relevant intergenic regions rather than previously known SNPs only. Thus, we were potentially able to reveal the new informative alleles in the known genes. In total, we called 117,012 SNPs in the 53 genes and intergenic regions.

For eye color prediction we performed feature selection algorithms in order to obtain new informative alleles for North Eurasian populations for 4 datasets:

1. Pooled North Eurasian dataset
2. European Russia
3. Caucasus
4. West Siberia

For hair color prediction we used 5 datasets:

1. Pooled North Eurasian dataset
2. European Russia
3. Caucasus
4. West Siberia
5. North Asia

North Asian dataset was analyzed only for hair color prediction due to the fact for this region there is an

**Table 2** The AUC and accuracy of the eye color prediction using HirisPlex-S set of SNPs for the regional North Eurasian datasets

	AUC			Accuracy		
	Caucasus region	West Siberia	European Russia	Caucasus region	West Siberia	European Russia
Blue eye	0,83 (15)	0,9 (17)	0,85 (60)	0,74 (15)	0,86 (17)	0,77 (60)
Intermediate eye	N/A (2)	N/A (1)	N/A (2)	N/A (2)	N/A (1)	N/A (2)
Brown eye	0,78 (38)	0,87 (26)	0,87 (32)	0,69 (38)	0,84 (26)	0,79 (32)
Red hair	N/A (0)	N/A (0)	0,81 (18)	N/A (0)	N/A (0)	0,88 (18)
Blond hair	0,77 (5)	0,58 (6)	0,75 (27)	0,84 (5)	0,79 (6)	0,72 (27)
Brown hair	0,41 (20)	0,55 (10)	0,6 (32)	0,41 (20)	0,65 (10)	0,61 (32)
Dark hair	0,77 (25)	0,81 (28)	0,76 (17)	0,53 (25)	0,77 (28)	0,79 (17)

Note: number of samples in each phenotypic class is indicated in the parentheses

observed variation in hair color while for eye color there is no such variation.

Each dataset has been divided in 60:40 ratio into training and test samples with preserving the percentage of samples for each class. For the pooled dataset we controlled that samples from different regions included in pooled dataset were split in the same proportion (60:40) to avoid region-related bias.

Feature selection procedure has been performed on the training dataset (Figure S2). Feature selection procedure consisted of applying three algorithms:

- 1) *f*\_regression
- 2) mutual\_info\_regression
- 3) Lasso feature selection with different alphas (0.7, 0.5, 0.2, 0.1, 0.05, 0.01, 0.005, 0.001, 0.0005)

When analyzing the distribution of F score (from *f*\_regression) and MI (from mutual\_info\_regression) the thresholds for the most effective features with highest scores were set for each dataset individually. When performing the Lasso feature selection we tested different choices of the alpha parameter. For each value of alpha we calculated *r*<sup>2</sup> scores on training dataset for corresponding subset of SNPs that have non-zero coefficients.

Among these subsets we selected the most important ones according to obtained *r*<sup>2</sup> scores for each dataset individually.

Based on results from three algorithms of feature selection all selected SNPs were combined in the top SNPs lists for each dataset.

In each top SNPs list, we selected SNPs which have the best predictive power. These SNPs formed best SNPs lists which we used to build a classifier. To select the best SNPs, we used the same scale as HIRISplex-S classifier:

1. blue, intermediate and brown for eye color
2. red, blond, brown and dark for hair color

We considered these classes independent from each other and tried to build the classifier with the best power and the smallest SNPs set.

We used separate ranking systems for eye and hair color prediction to estimate the importance and prediction power of each SNP in order to narrow down the SNPs lists.

The performance of the best selected features was validated on the test dataset. To evaluate the quality of the model we calculated R<sup>2</sup> score (coefficient of determination regression score function) ([https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2\\_score.html](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html)), AUC score, precision, recall and accuracy metrics.

## Eye color prediction

### Identifying the top SNPs in the pooled north Eurasian dataset

To identify the top SNPs associated with the eye color in our sample we applied three algorithms: *f*\_regression (F score), mutual\_info\_regression (MI), and Lasso feature selection with different alphas (0.7, 0.5, 0.2, 0.1, 0.05, 0.01, 0.005, 0.001, 0.0005).

We analyzed F (*f*\_regression) and MI (mutual\_info\_regression) scores distributions across the samples and selected the top 30 SNPs with the highest scores.

According to results from Lasso feature selection we decided to include in top SNPs list the most crucial ones - the ones having non zero coefficients for alpha = 0.5 (2 SNPs for 'eye color' dataset and 2 SNPs for 'hair color' dataset) and alpha = 0.2 (8 SNPs for 'eye color' dataset and 8 SNPs for 'hair color' dataset) - these SNPs carry the most prediction power according to *r*<sup>2</sup> score values distribution over different alphas. We also included SNP sets for alphas 0.1, 0.01 and 0.005.

The final top SNPs list consisted of 256 SNPs (Additional file 3).

### Narrowing the list of SNPs and building classifier for eye color based on it

We assigned to each SNP a score from 0 to 3. The score 3 is assigned only for SNPs from the pooled dataset top SNPs list because of the results made for that dataset are much more robust than for regional datasets (sample sizes for the regional datasets are present in the Additional file 4). The score 3 is assigned to SNPs that are in top 5 with highest F score or have coefficients more or equal to 0.1 in absolute value in Lasso models for alpha 0.2 or have non-zero coefficients in Lasso models for alpha 0.5. For the pooled sample the score 2 is assigned to SNPs that are in top 10 with highest F or MI scores or have non-zero coefficients in Lasso model for alpha 0.2. The score 1 is assigned to SNPs that have coefficients greater or equal 0.1 in Lasso model for alpha 0.005. To all other SNPs we assigned the score 0. All 36 SNPs with non-zero scores formed the best SNPs list and were used for the classifier.

The five SNPs had the highest score 3. Two of them were well-known eye color-causing SNPs (rs1129038 and rs12913832) while the remaining three have not been reported previously as powerful eye color predictive alleles.

### Variation of the best SNPs list across geographic regions

The entire analysis performed for the pooled North Eurasian dataset has been repeated for the populations from the three following regions separately: European Russia, Caucasus, and West Siberia. For regional datasets the score 2 was assigned to SNPs that were in top 5 with highest F and MI scores or had coefficients more or

**Table 3** The list of 36 best North Eurasian SNPs for eye color prediction

SNP_ID	Caucasus Score	European Russia Score	West Siberia Score	Pooled Dataset Score	HirisPlex-S	dbSNP RSID	Gene
<b>chr15:28356859_C_T</b>	2	2	2	3	rs1129038	rs1129038	HERC2
<b>chr15:28365618_A_G</b>	2	2	2	3	rs12913832	rs12913832; 4745	HERC2
<b>chr15:28392261_G_A</b>	2	2		3		rs12898729	HERC2
<b>chr15:28410491_C_T</b>		2	2	3		rs12916300	HERC2
<b>chr15:28495956_A_G</b>		2	2	3		rs12912427	HERC2
<b>chr15:28562998_T_C</b>		1		2		rs1614575	HERC2
<b>chr20:39272620_A_G</b>		1		2		rs4812447	Intergene spacer
chr1:119406130_C_T				2		rs1779446	Intergene spacer
chr1:3331899_A_G				2		rs1999528	PRDM16
chr15:28145024_T_C				2		rs2871886	OCA2
chr15:28364059_A_G				2		rs7494942	HERC2
chr15:28380518_T_A				2		rs4778249	HERC2
chr15:28383565_T_C				2		rs7403279	HERC2
chr15:28513364_T_C				2		rs916977;4744	HERC2
chr15:28530182_C_T				2	rs1667394	rs1667394;4743	HERC2
chr15:28566122_A_G				2		rs751089833	HERC2
chr19:7570978_T_C				2		rs685034	C19orf45
chr3:189429301_G_T				2		rs6804480	TP63
chr6:45136347_G_A				2		rs1324530	SUPT3H
chrX:66405249_C_T				2		rs34191540	Intergene spacer
chr10:87576467_C_T				1		rs7923503	GRID1
chr14:92909309_T_C				1		rs12588868	SLC24A4
chr15:28419048_T_G				1		rs35946704	HERC2
chr17:9107969_G_A				1		rs17742781	NTN1
chr19:7578733_A_T				1		rs586243	ZNF358
chr3:189552236_T_C				1		rs7653443	MIR944
chr3:33035542_T_C				1		rs4586761	GLB1
chr3:33111182_T_G				1		rs72856153	GLB1
chr3:54251172_G_A				1		rs11283625	CACNA2D3
chr3:54636061_G_A				1		rs34983676	CACNA2D3
chr4:87847613_T_G		1		1		rs10022539	LOC100506746
chr4:87851083_C_T				1		rs72667724	LOC100506746
chr5:60947483_A_G				1		rs1501841	C5orf64
chr5:73959526_A_G				1		rs2454846	HEXB
chr6:45419110_C_G				1		rs2820339	RUNX2
chr7:42032565_C_T				1		rs2237427	GLI3

Indicated in bold - new SNPs which demonstrated the high prediction power for the eye color

Columns: SNP\_ID – SNP ID in format: chromosome:position in GRCh37\_allele 1\_allele 2. Caucasus score, European Russia Score, West Siberia Score, Pooled Dataset Score – scores as described in section “Eye color prediction” for corresponding datasets

HirisPlex-S – RS ID if used in HirisPlex-S. Otherwise empty

dbSNP RSID – RS ID in dbSNP database

Gene – Nearest gene for this SNP

equal to 0.1 in absolute value in Lasso model for alpha 0.5 or non-zero coefficients in Lasso model for alpha 0.7. The score 1 was assigned to SNPs that were in top 6 with highest F and MI scores or have coefficients non-zero coefficients in Lasso models for alpha 0.7 and 0.5. Additional file 5 presents the resulting best SNPs sets for all three regions. The comparison of the regional lists and the list for the pooled sample is present in the Additional file 6. In general, the set of best SNPs is stable across the regions: the SNPs with the highest scores are included in the most lists, while among the other SNPs there are both, identified within every region and region-specific. Further study on the additional phenotyped samples is necessary to replicate the significance of the region-specific SNPs.

The merged SNPs list was ranked by total score (as sum of all scores for 4 samples: Caucasus, West Siberia, European Russia, and pooled) (Additional file 6). Top 7 SNPs have the highest total score and

occurred in more than one dataset, which is an additional confirmation that these SNPs have a strong predictive power (Table 3). Two of those SNPs (rs1129038 and rs12913832) are already included in HIRisPlex-S panel, while other five SNPs are new candidates for eye color predicting in the North Eurasian populations. We estimated the frequencies of these five SNPs in North Eurasian populations (Additional file 7). Each SNP was detected with polymorphic frequencies in every regional population, so these SNPs are common rather than rare ones.

**The north Eurasian SNPs set performance**

We estimated the performance of the SNPs which demonstrated the highest predictive power in our North Eurasian sample. The minimal set included 7 SNPs, two of which were previously included into the HIRisPlex-S panel. The optimal set included 36 SNPs which received the highest scores on the pooled North Eurasian dataset. We tested the classification performance of both sets of North Eurasian SNPs. Figure 2 presents the ROC curves and AUC scores for the prediction of three eye colors. The accuracy of 7 SNPs set is almost as effective as prediction based on the 41 HIRisPlex-S SNPs, while the set of 36 North Eurasian SNPs slightly outperforms 41 HIRisPlex-S SNPs on our sample (Fig. 2, Table 1).

**Hair color prediction**

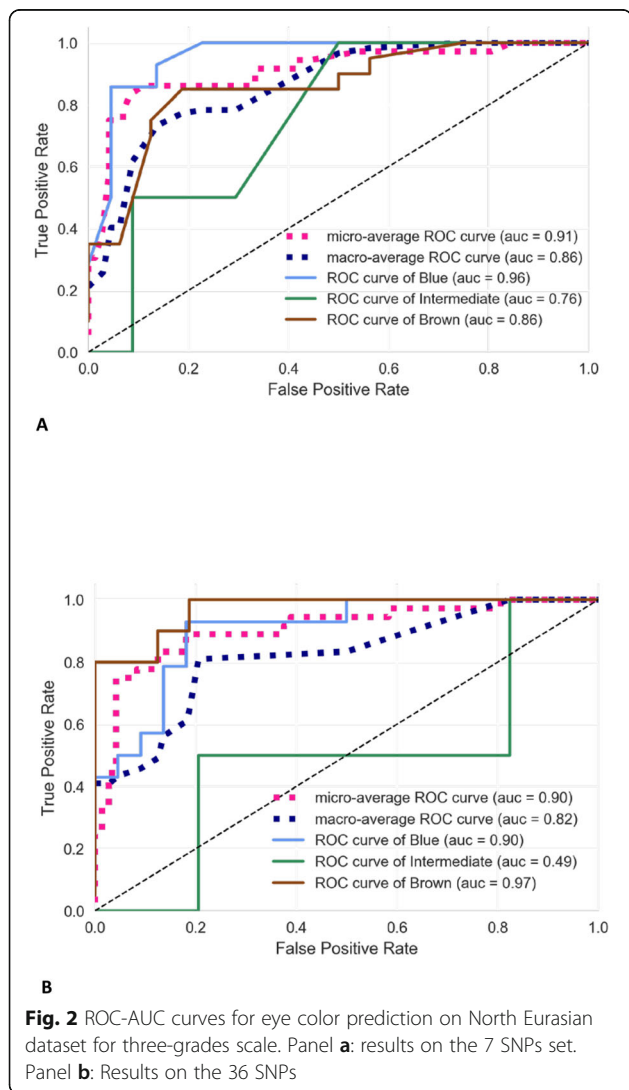
We performed the same feature selection analysis to find and evaluate top SNPs list for hair color prediction for pooled North Eurasian sample, which includes populations from the following regions: Caucasus, European Russia, West Siberia and North Asia.

We selected top 322 SNPs and narrowed the list to 33 best SNPs that have the strongest performance for 4-grade classification: red, blond, brown and dark hair color, the same scale as HIRisPlex-S (Additional file 8).

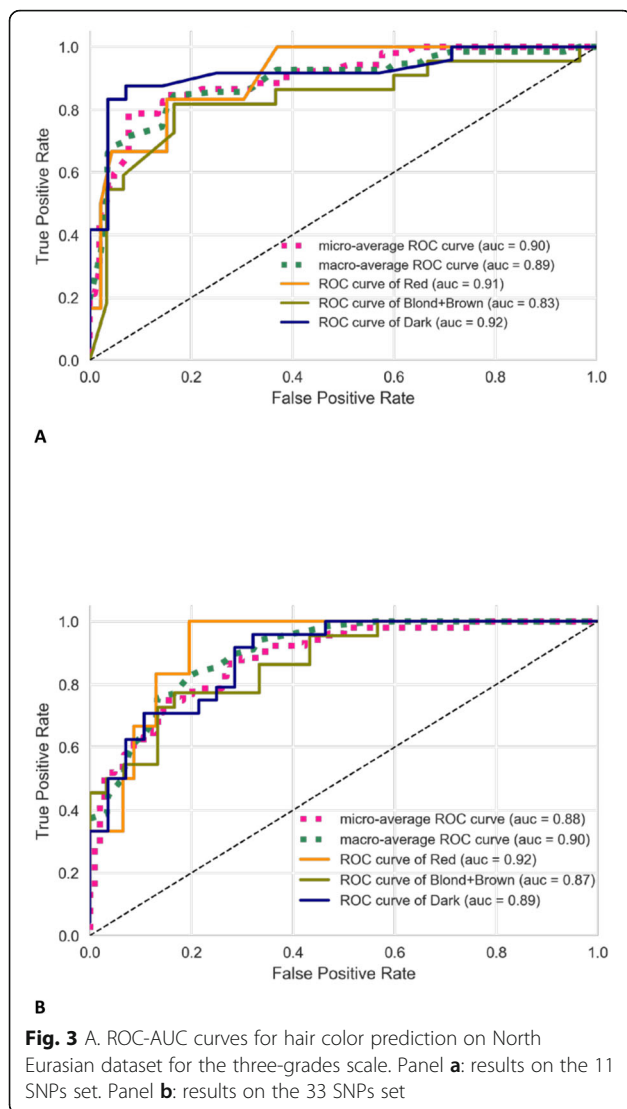
We assigned significance scores to select the minimum set of SNPs in following way:

- 1) The score 3 has been assigned to SNPs that are in top 5 with highest F or MI scores or have coefficients more than 0.05 in absolute value in Lasso models for alpha 0.2 or have non-zero coefficients in Lasso models for alpha 0.5
- 2) The score 2 has been assigned to SNPs in top 10 with highest F or MI scores
- 3) The rest SNPs of 33 best SNPs list have the score 1

We were able to detect the most powerful 11 SNPs that have the highest score (3), three of them are included in HIRisPlex-S panel (rs16891982, rs12913832, and rs1129038).



**Fig. 2** ROC-AUC curves for eye color prediction on North Eurasian dataset for three-grades scale. Panel a: results on the 7 SNPs set. Panel b: Results on the 36 SNPs



We checked the performance of the classifier based on 11 SNPs set and tried to estimate its ability to distinguish between 4 independent classes (the same as for HIrisPlex-S): red, blond, brown and dark hair (Additional file 9).

Additionally, we tried to merge 2 classes of hair color - blond and brown - because algorithm does not have enough power to distinguish them, and checked the performance of selected SNPs for 3 grade scale. As we can see from the results (Fig. 3) the classifier performance improved significantly for both sets of SNPs: the most powerful 11 SNPs and 33 best SNPs.

#### The new potentially informative SNPs

Our analysis identified five new SNPs which demonstrated the high prediction power for the eye color. These SNPs were revealed on the pooled North Eurasian sample and were replicated on the most regional

subsamples. Four of these SNPs are located in *HERC2* gene, and one (rs4812447) is in intergenic region. *HERC2* (HECT And RLD Domain Containing E3 Ubiquitin Protein Ligase 2) gene belongs to the HERC gene family that encodes a group of unusually large proteins, which contain multiple structural domains. Genetic variations in this gene are associated with skin/hair/eye pigmentation variability [1, 14, 15].

#### Limitations of the used approach

We analyzed the performance of the known pigmentation predictive SNPs and looked for the new SNPs in previously unstudied populations from different geographic areas. This regional-based approach allowed identify SNPs which are informative for the particular populations but made the sample sizes from each region quite limited. Therefore, we were not able to subdivide our sample into the training dataset and validation dataset – this would result in reducing sample sizes to numbers not allowing the statistically significant analysis. Therefore, our approach forced us to use the same dataset for SNPs discovery, building the classification model, and also for the validation, which might result in prediction overestimation. Therefore, the performance of our SNPs should be considered as an upper estimate, and the identified SNPs as candidate ones until verification on the independent sample in the future studies. Though stability of the top eye color predictive SNPs across geographic regions partly verifies the effectiveness of the newly identified predictive SNPs.

#### Conclusion

We analyzed the gene-phenotype correlation in the populations from the border regions between Europe and Asia which carry light pigmentation phenotypes but have the contrasting genetic ancestries with the West Europeans. We replicated the effectiveness of the classical HIrisPlex-S panel for these previously unstudied populations, though the accuracy is slightly lower than for the West European groups the classifier has been developed on. Such decrease in accuracy might result from the population-specific SNPs which are present in North Eurasian populations but are rare in West Europeans and thus have not been included in the HIrisPlex-S panel. We analyzed the pigmentation genes and relevant intergenic regions in the phenotyped individuals and performed the association analysis between all identified polymorphic sites and pigmentation phenotypes. Note, that our target sequencing included not only the standard exome, but also intronic regions of the 53 pigmentation-related genes. Thus, the released dataset can be used for further population genetics or medical genetics studies, because it presents the exome variation in many indigenous groups which were previously



studied by SNPs arrays only but not by the sequencing approach. As an additional by-product, the dataset allows to estimate the frequencies of the mutations with uncertain pathogenicity in the North Eurasian populations.

Our analysis of the pigmentation replicated the importance of the key previously known SNPs but also identified five new markers whose eye color prediction power on our North Eurasian dataset is compatible with the two major previously known SNPs. We note, that HIRisPlex-S recall for the blue eye phenotype is lowest in the Caucasus region, and our analysis identified a set of SNPs with prediction power specific for the Caucasus. We note, that all the SNPs revealed are candidate ones, as the same dataset (North Eurasian sample) has been used for both procedures: the feature selection and the classification. To avoid the overtraining effect, the replication of the new SNPs on the independent North Eurasian sample is needed in the future studies.

## Methods

### Populations studied

The dataset consisted of 300 samples from 48 local populations. Additional file 10 presents linguistic affiliation of these populations, while Fig. 1a indicates their geographic locations. The populations were geographically grouped into 4 major regions of North Eurasia (Fig. 1a, Additional file 10, Additional file 11).

To illustrate the genetic relations of these populations, we used the genome-wide datasets on the same ethnic groups available in the GG-base ([www.gg-base.org](http://www.gg-base.org)) and run the principal components analysis using PLINK software (Fig. 1b). We also added into the analysis the populations used for developing and validating the HIRisplex-S system: Dutch, Irish, Polish, and Greek. As there were no data on Dutch populations in the GG-base, we used Northwest French and West Germans as a suitable proxy; this has had a minor impact on the plot, as PC has not identified much differentiation among the HIRisplex populations, as expected for West/Central Europeans.

### Phenotyping

The high-quality photos of members of indigenous North Eurasian communities were obtained during the field trips coordinated by the Biobank of North Eurasia [16]. The eye and hair color phenotypes were called based on these photos by three experts: two were physical anthropologists with deep experience in phenotyping, and the third was the specially trained geneticist. All the experts performed the phenotyping independently, and the cases when the calls were different became a subject of a thorough investigation until the consensus calls were achieved. We identified 99 individuals with light eyes and 187 with dark eyes, 128 individuals with light hair and 156 with dark hair, 76 individuals with red

hair and 209 with not red hair. Additional file 11 presents the individual phenotyping calls including intermediate values.

### Library preparation and sequencing

Genomic DNA from both blood or saliva was extracted using an organic extraction method. List of genes and intergenic regions which can have potential polymorphisms associated with hair color and eye color traits was created based on detailed literature analysis [1, 14, 15, 17–25] including genome-wide association studies (GWAS) catalog (<https://www.ebi.ac.uk/gwas/>). For example, not exons only, but also introns of HERC2 and CACNA2D3 have been included into the sequencing capture. As a result, we developed the custom target sequencing panel which includes the 53 genes or intergenic regions. Fragmentation was performed by the Hydrodynamic Shearing System (Covaris). DNA fragments with ligated adapter molecules were selectively enriched by PCR, and then exons of genes were captured. The exome DNA enrichment was performed with custom SeqCap EZ Exome Plus Library Kit (Roche) with SeqCap Adapter Kit (Roche) and SeqCap HE-Oligo Kit (Roche) and sequencing libraries were generated using KAPA HyperPlus Library Preparation Kit (Roche), according to the manufacturer's recommendations. Products were purified using the AMPure XP system (Beckman Coulter) and quantified using the Agilent high sensitivity DNA assay on the Agilent Bioanalyzer 2100 system. Sequencing was performed on an HiSeq 2500 sequencer (Illumina) with HiSeq SBS v4 250 Kit (Illumina) following the manufacturer's recommendations and yielded 125-bp paired-end reads.

### Bioinformatics analysis

Raw data from high-throughput sequencing in fastq.gz format were aligned to hg19 reference human genome using bwa mem software. The resulting files in bam format were sorted and deduplicated using the SAMtools program package. Mutation calling was performed using freebayes software with filtration (quality (QUAL) > 40 & read depth (DP) > 5) of identified variants with vcfliib program package. Annotation of variants was performed using SnpSift of snpEff program package. Databases dbSNP [26], dbNSFP [27, 28], ClinVar [29], 1000 Genomes Project [30], and ExAC [31] were used as information resources for identified variants. Samples with low genotyping rate have been excluded from further analyses (minimal genotyping rate is 90%), resulting in 286 samples dataset. Then polymorphisms in selected genes were analyzed and characterized for 286 samples.

### Prediction from HIRis-Plex-S SNPs

We called 41 polymorphic sites from HIRis-Plex-S forensic panel in all 286 analyzed samples and calculated hair

and eye color predictions for all samples using online tool of the Department of Genetic Identification of Erasmus MC (<https://hirisplex.erasmusmc.nl>). The predicted phenotypes were then compared with the true phenotypes, and the performance statistics were calculated for the pooled North Eurasian dataset and the regional datasets. Our five-grades scales have been converted into three-grades scales to make phenotypic call fully comparable with the HIRISplex-S calls.

### Identifying the potentially informative SNPs for north Eurasian populations

Our dataset included 48 populations from North Eurasia. For eye color prediction we omitted the populations that are less polymorphic in eye color phenotypes (Additional file 11). This allowed us to achieve the better balance between different phenotypic classes. Populations which don't have at least 4 grades of the our five-grades scale were eliminated from further analyses.

We ran feature selection algorithms in order to find most informative SNPs correlated with eye and hair color according to our 5-grade scales used for quantitative estimate of dark pigment in eyes (the highest value corresponds to the highest concentration of pigment) and hair (where '0' is a red hair, '1' is blond hair, and '4' is the dark hair).

We considered our 5-grade scales (both for eye and hair color) continuous as they reflect the concentration of dark pigment and those classes are not independent.

Three feature selection methods were applied to select the most informative features associated with pigmentation traits for the pooled North Eurasian dataset and for each region separately.

Each dataset has been divided in 60:40 ratio into training and test samples using stratified K folds cross-validator that preserves the percentage of samples for each class.

Feature selection methods were applied to training dataset while the quality metrics for selected features were calculated using test dataset. The minimum set of SNPs with the most predictive power has been identified on test dataset. Also, we built classifier based on these final SNPs.

To evaluate the quality of the model we calculated r2 score ([https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2\\_score.html](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html)), AUC, accuracy, precision and recall metrics using scikit-learn package [32].

### Feature selection algorithms

We used the three following algorithms for feature selection which are suitable for regression tasks (Additional file 12):

- 1) `f_regression` ([https://scikit-learn.org/stable/modules/generated/sklearn.feature\\_selection.f\\_regression](https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.f_regression)).

[https://scikit-learn.org/stable/modules/generated/sklearn.feature\\_selection.f\\_regression](https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.f_regression)) - univariate regression test. Univariate feature selection works by selecting the best features based on univariate statistical tests. It uses linear model for testing the individual effect of each of many regressors. This is a scoring function to be used in a feature selection procedure.

This was done in 2 steps:

1. The correlation between each regressor and the target is computed, that is,  $((X[:, i] - \text{mean}(X[:, i])) * (y - \text{mean}_y)) / (\text{std}(X[:, i]) * \text{std}(y))$ .
2. It is converted to an F score then to a  $p$ -value

We selected only those features that has  $p$ -value < 0.01. Then we sorted them in F-score descending order. Features that have the highest F score values are considered the most promising and potentially informative features.

- 1) `mutual_info_regression` ([https://scikit-learn.org/stable/modules/generated/sklearn.feature\\_selection.mutual\\_info\\_regression.html#sklearn.feature\\_selection.mutual\\_info\\_regression](https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.mutual_info_regression.html#sklearn.feature_selection.mutual_info_regression)). It estimates mutual information (MI) for a continuous target variable. Mutual information between two random variables is a non-negative value, which measures the dependency between the variables. It is equal to zero if and only if two random variables are independent, and higher values mean higher dependency. The function relies on nonparametric methods based on entropy estimation from k-nearest neighbors' distances as described in [33, 34]. Both methods are based on the idea originally proposed in [35].
- 2) L1- based feature selection -- Lasso technique ([https://scikit-learn.org/stable/modules/generated/sklearn.linear\\_model.Lasso.html#sklearn.linear\\_model.Lasso](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html#sklearn.linear_model.Lasso)). Linear models penalized with the L1 norm have sparse solutions: many of their estimated coefficients are zero. The Lasso is a linear model that estimates sparse coefficients. It is useful in some contexts due to its tendency to prefer solutions with fewer parameter values, effectively reducing the number of variables upon which the given solution is dependent.

It consists of a linear model trained with L1 prior as regularizer. The objective function to minimize is:

$$\min_w \frac{1}{2n_{\text{samples}}} \|Xw - y\|_2^2 + \alpha \|w\|_1 \quad (1)$$

The lasso estimate thus solves the minimization of the least-squares penalty with  $\alpha \|w\|_1$  added, where  $\alpha$  (alpha)

is a constant,  $\|w\|_1$  is the L1-norm of the parameter vector,  $X$  is training data and  $y$  is target values.

The parameter alpha controls the sparsity: the higher the alpha parameter, the fewer features selected. For our purposes we tested a range of alphas: 0.7, 0.5, 0.2, 0.1, 0.05, 0.01, 0.005, 0.001, 0.0005. The best features were found using the biggest alphas: 0.7, 0.5 and 0.2.

#### Parameters for selecting the most significant features (SNPs)

##### The pooled dataset

To avoid the situation of finding SNPs associated with particular population of North Eurasia rather than with a phenotypic trait we excluded from analyses those geographic regions in which we didn't find the variation in phenotype. Hence, for eye color prediction the final dataset included Caucasus, European Russia and West Siberia regions (Additional file 11), while the dataset for hair color prediction consisted of populations from all four regions - Caucasus, European Russia, North Asia, and West Siberia.

##### Identifying the top SNPs lists

Top of a few hundred SNPs most significantly associated with phenotypic traits has been chosen using the following thresholds (Additional file 12):

- 1) top 30 SNPs with highest F scores for  $f_{\text{regression}}$
- 2) top 30 SNPs with highest MI scores for  $\text{mutual\_info\_regression}$
- 3) SNPs with non-zero coefficients for Lasso models with alphas 0.5, 0.2, 0.1 0.01 and 0.005

##### Selecting the best SNPs from the top lists

The top lists included hundreds of SNPs, and to narrow down the lists we selected the best SNPs from each top-SNPs list (Additional file 12). We used corresponding thresholds to obtain these lists for both, eye and hair color prediction:

- 1) top 10 SNPs with highest F scores for  $f_{\text{regression}}$
- 2) top 10 SNPs with highest MI scores for  $\text{mutual\_info\_regression}$
- 3) SNPs with non-zero coefficients for Lasso models with alphas 0.5 and 0.2
- 4) SNPs with coefficients more or equal to 0.1 in absolute value for Lasso model with alpha 0.005

##### Regional datasets

To select best SNPs for each region we also performed 3 types of feature selection analyses and looking at the distribution of scores and considering the sample size for each region we set the following thresholds:

- 1) top 6 SNPs with highest F scores for  $f_{\text{regression}}$
- 2) top 6 SNPs with highest MI scores for  $\text{mutual\_info\_regression}$
- 3) SNPs with non-zero coefficients for Lasso feature selection with parameters alpha 0.7 and 0.5.

##### Building the classifier

For building the classifier we used a linear regression algorithm. We used genotypes for SNPs from best SNPs lists converted to values 0, 1 or 2 (2 for genotype '1/1', 1 for genotypes '1/0' or '0/1' and 0 for '0/0'). Model was trained on the training dataset. For quality estimation we calculated  $r^2$  score, AUC, accuracy, precision and recall metrics on the test dataset.

##### Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1186/s12864-020-06923-1>.

**Additional file 1 Table S1.** Performance characteristics of North Eurasian SNPs sets and HlrisPlex-S panel on North Eurasian pooled dataset for eye and hair color prediction.

**Additional file 2 Table S2.** The performance of HlrisPlex-S panel on North Eurasian regional datasets.

**Additional file 3 Table S3.** Top SNPs list for North Eurasian pooled dataset for eye color prediction.

**Additional file 4 Table S4.** Sample sizes of different phenotypic classes in the regional datasets.

**Additional file 5 Table S5.** Best SNPs lists for regional datasets for eye color prediction.

**Additional file 6 Table S6.** Summary table of best SNPs lists for the North Eurasian pooled and regional datasets for eye color prediction.

**Additional file 7 Table S7.** Frequencies of the five new eye color predictive SNPs in North Eurasian populations.

**Additional file 8 Table S8.** Top SNPs list for North Eurasian pooled dataset for hair color prediction.

**Additional file 9 Figure S1.** ROC-AUC curves for hair color prediction on North Eurasian dataset for the four-grades scale. Panel A: results for the 11 SNPs set. Panel B: results for the 33 SNPs set.

**Additional file 10 Table S9.** The studied populations.

**Additional file 11 Table S10.** The analyzed North Eurasian samples.

**Additional file 12 Figure S2.** SNP selection scheme.

##### Abbreviations

AUC: Area under ROC curve; ROC curve: Receiver operating characteristic; CACNA2D3: Calcium channel, voltage-dependent, alpha 2/delta subunit 3 gene; DNA: Deoxyribonucleic acid; GWAS: Genome-wide association studies; F: F-regression score; HERC2: HECT and RLD domain containing E3 ubiquitin protein ligase 2; MI: Mutual info regression score; PC: Principal component analysis; PCR: Polymerase chain reaction; R2 score: Coefficient of determination regression score function; SNP: Single nucleotide polymorphism

##### Acknowledgements

We thank all sample donors whose participation made this study possible. The DNA collections and collections of the anthropological photos were granted by the Biobank of North Eurasia. We cordially thank professor Manfred Kayser for the helpful discussions. The custom exome sequencing kit was developed by Roche and Alamed companies. We also thank Oksana Selezneva for technical assistance with exome sequencing.

**About this supplement**

"This article has been published as part of BMC Genomics Volume 21 Supplement 7, 2020: Selected Topics in "Systems Biology and Bioinformatics" - 2019: genomics. The full contents of the supplement are available online at <https://bmcgenomics.biomedcentral.com/articles/supplements/volume-21-supplement-7>".

**Authors' contributions**

EB designed and coordinated the study. EK generated the genetic dataset, while JK, AM and NL phenotyped the same samples. JK and AA handled the samples and photographs. MZ collected part of the Asian samples. EL, IG and VP analyzed the data. OB and EL drafted the text. All authors read and approved the final manuscript.

**Funding**

This research was supported by the Russian Ministry of Science and Higher Education (state contract # 011-17, September 26, 2017) within the framework of the Union State research program "Developing the innovative gene geographical and genomic technologies for identification and revealing the personal features by studying the gene pools of the regional populations from the Union State" ("DNA-identification") and via the State Task for the Research Centre for Medical Genetics. The publication costs have been funded by authors. The funding body played no role in the design of the study, research, writing and publication of the paper.

**Availability of data and materials**

The exome sequencing dataset used in this study has been deposited in European Nucleotide Archive database (<https://www.ebi.ac.uk/ena>) under PRJEB35224 (ERP118250) identification number of the study.

**Ethics approval and consent to participate**

The study was approved by the Ethics Committee of the Research Centre for Medical Genetics, Moscow, Russia. All procedures performed in studies involving human participants were in accordance with the ethical standards and with the Helsinki declaration (1964). The written informed consent was obtained from all individual participants included in the study.

**Consent for publication**

Not applicable.

**Competing interests**

The authors declare that they have no competing interests.

**Author details**

<sup>1</sup>Research Centre for Medical Genetics, Moscow, Russia. <sup>2</sup>Biobank of North Eurasia, Moscow, Russia. <sup>3</sup>Vavilov Institute of General Genetics, Moscow, Russia. <sup>4</sup>Krasnodar State Medical University, Krasnodar, Russia. <sup>5</sup>Research Institute and Museum of Anthropology, Lomonosov Moscow State University, Moscow, Russia. <sup>6</sup>Institute of Ethnology and Anthropology of Russian Academy of Sciences, Moscow, Russia. <sup>7</sup>Moscow Institute of Physics and Technology, Moscow, Russia. <sup>8</sup>National Center for Biotechnology, Nursultan, Kazakhstan. <sup>9</sup>Federal Research and Clinical Centre of Physical-Chemical Medicine, Moscow, Russia.

Received: 14 July 2020 Accepted: 17 July 2020

Published: 10 September 2020

**References**

- Kayser M, Liu F, Janssens AC, Rivadeneira F, Lao O, van Duijn K, Vermeulen M, Arp P, Jhamai MM, van Ijcken WF, den Dunnen JT, Heath S, Zelenika D, Despriet DD, Klaver CC, Vingerling JR, de Jong PT, Hofman A, Aulchenko YS, Uitterlinden AG, Oostra BA, van Duijn CM. Three genome-wide association studies and a linkage analysis identify HERC2 as a human iris color gene. *Am J Hum Genet.* 2008;82(2):411–23.
- Han J, Kraft P, Nan H, Guo Q, Chen C, Qureshi A, Hankinson SE, Hu FB, Duffy DL, Zhao ZZ, Martin NG, Montgomery GW, Hayward NK, Thomas G, Hoover RN, Chanock S, Hunter DJ. A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation. *PLoS Genet.* 2008;4(5):e1000074.
- Candille SI, Absher DM, Beleza S, Bauchet M, McEvoy B, Garrison NA, Li JZ, Myers RM, Barsh GS, Tang H, Shriver MD. Genome-wide association studies of quantitatively measured skin, hair, and eye pigmentation in four European populations. *PLoS One.* 2012;7(10):e48294.
- Liu F, Visser M, Duffy DL, Hysi PG, Jacobs LC, Lao O, Zhong K, Walsh S, Chaitanya L, Wollstein A, Zhu G, Montgomery GW, Henders AK, Mangino M, Glass D, Bataille V, Sturm RA, Rivadeneira F, Hofman A, van Ijcken WF, Uitterlinden AG, Palstra R-JTS, Spector TD, Martin NG, TEC N, Kayser M. Genetics of skin color variation in Europeans: genome-wide association studies with functional follow-up. *Hum Genet.* 2015;134(8):823–35.
- Stokowski RP, Pant PV, Dadd T, Fereday A, Hinds DA, Jarman C, Filzell W, Ginger RS, Green MR, van der Ouderaa FJ, Cox DR. A genome wide association study of skin pigmentation in a south Asian population. *Am J Hum Genet.* 2007;81(6):1119–32.
- Lippert C, Sabatini R, Maher MC, Kang EY, Lee S, Arkan O, Harley A, Bernal A, Garst P, Lavrenko V, Yocum K, Wong T, Zhu M, Yang WY, Chang C, Lu T, Lee CWH, Hicks B, Ramakrishnan S, Tang H, Xie C, Piper J, Brewerton S, Turpaz Y, Telenti A, Roby RK, Och FJ, Venter JC. Identification of individuals by trait prediction using whole-genome sequencing data. *Proc Natl Acad Sci U S A.* 2017;114(38):10166–71.
- Liu F, van Duijn K, Vingerling JR, Hofman A, Uitterlinden AG, Janssens AC, Kayser M. Eye color and the prediction of complex phenotypes from genotypes. *Curr Biol.* 2009;19(5):R192–3.
- Maronas O, Sochtig J, Ruiz Y, Phillips C, Carracedo A, Lareu MV. The genetics of skin, hair, and eye color variation and its relevance to forensic pigmentation predictive tests. *Forensic Sci Rev.* 2015;27(1):13–40.
- Chaitanya L, Breslin K, Zuñiga S, Wirken L, Pośpiech E, Kukla-Bartoszek M, Sijen T, de Knijff P, Liu F, Branicki W, Kayser M, Walsh S. The HlrisPlex-S system for eye, hair and skin colour prediction from DNA: introduction and forensic developmental validation. *Forensic Sci Int Genet.* 2018;35:123–35.
- Walsh S, Liu F, Wollstein A, Kovatsi L, Ralf A, Kosiniak-Kamysz A, Branicki W, Kayser M. The HlrisPlex system for simultaneous prediction of hair and eye colour from DNA. *Forensic Sci Int Genet.* 2013;7(1):98–115.
- Walsh S, Chaitanya L, Clarisse L, Wirken L, Draus-Barini J, Kovatsi L, Maeda H, Ishikawa T, Sijen T, de Knijff P, Branicki W, Liu F, Kayser M. Developmental validation of the HlrisPlex system: DNA-based eye and hair colour prediction for forensic and anthropological usage. *Forensic Sci Int Genet.* 2014;9:150–61.
- Walsh S, Kayser M. A practical guide to the HlrisPlex system: simultaneous prediction of eye and hair color from DNA. *Methods Mol Biol.* 2016;1420:213–31.
- Pagani L, Lawson D, Jagoda E, Mörseburg A, Eriksson A, Mitt M, Clemente F, Hudjashov G, DeGiorgio M, Saag L, Wall J, Cardona A, Mägi R, Wilson S, Kaewert S, Inchley C, Scheib C, Järve M, Karmin M, Jacobs G, Antao T, Iliescu M, Kushniarevich A, Ayub Q, Tyler-Smith C, Xue Y, Yunusbayev B, Tambets K, Mallick CB, Saag L, Pocheshkhova E, Andriadze G, Muller C, Westaway MC, Lambert DM, Zoraqi G, Turdikulova S, Dalimova D, Sabitov Z, GNN S, Lachance J, Tishkoff S, Momyaliev K, Isakova J, Damba LD, Gubina M, Nymadawa P, Evseeva I, Atramantova L, Utevska O, Ricaut FX, Brucato N, Sudoyo H, Letellier T, Cox MP, Barashkov NA, Skaro V, Mulahasanovic L, Primorac D, Sahakyan H, Mormina M, Eichstaedt CA, Lichman DV, Abdullah S, Chaubey G, JTS W, Mihailov E, Karunas A, Litvinov S, Khusainova R, Ekomasova N, Akhmetova V, Khidiyatova I, Marjanović D, Yepiskoposyan L, Behar DM, Balanovska E, Metspalu A, Derenko M, Malyarchuk B, Voevoda M, Fedorova SA, Osipova LP, Lahr MM, Gerbault P, Leavesley M, Migliano AB, Petraglia M, Balanovsky O, Khusnutdinova EK, Metspalu E, Thomas MG, Manica A, Nielsen R, Villemers R, Willerslev E, Kivisild T, Metspalu M. Genomic analyses inform on migration events during the peopling of Eurasia. *Nature.* 2016;538(7624):238–42.
- Eiberg H, Troelsen J, Nielsen M, Mikkelsen A, Mengel-From J, Kjaer KW, Hansen L. 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. *Hum Genet.* 2008; 123(2):177–87.
- Mengel-From J, Borsting C, Sanchez JJ, Eiberg H, Morling N. Human eye colour and HERC2, OCA2 and MATP. *Forensic Sci Int Genet.* 2010;4(5): 323–8.
- Balanovska EV, Zhabagin MK, Agdzhoyan AT, Chukhryaeva MI, Markina NV, Balaganskaya OA, Skhalyakho RA, Yusupov YM, Utevska OM, Bogunov YV, Asilguzhin RR, Dolinina DO, Kagazezheva ZA, Damba LD, Zaporozhchenko VV, Romanov AG, Dibirova KD, Kuznetsova MA, Lavryashina MB,

- Pocheshkhova EA, Balanovsky OP. Population biobanks: organizational models and prospects of application in gene geography and personalized medicine. *Russ J Genet.* 2016;52(12):1227–43.
17. Jacobs LC, Hamer MA, Gunn DA, Deelen J, Lall JS, van Heemst D, Uh HW, Hofman A, Uitterlinden AG, Griffiths CEM, Beekman M, Slagboom PE, Kayser M, Liu F, Nijsten T. A genome-wide association study identifies the skin color genes IRF4, MC1R, ASIP, and BNC2 influencing facial pigmented spots. *J Invest Dermatol.* 2015;135(7):1735–42.
  18. Kenny ETN, Sikora M, Yee MC, Moreno-Estrada A, Eng C, Huntsman S, Burchard EG, Stoneking M, Bustamante CD, Myles S. Melanesian blond hair is caused by an amino acid change in TYRP1. *Science.* 2012;336(6081):554.
  19. Sochtig J, Phillips C, Maronas O, Gomez-Tato A, Cruz R, Alvarez-Dios J, de Cal MA, Ruiz Y, Reich K, Fondevila M, Carracedo A, Lareu MV. Exploration of SNP variants affecting hair colour prediction in Europeans. *Int J Legal Med.* 2015;129(5):963–75.
  20. Soejima M, Koda Y. Population differences of two coding SNPs in pigmentation-related genes SLC24A5 and SLC45A2. *Int J Legal Med.* 2007;121(1):36–9.
  21. Sulem P, Gudbjartsson DF, Stacey SN, Helgason A, Rafnar T, Magnusson KP, Manolescu A, Karason A, Palsson A, Thorleifsson G, Jakobsdottir M, Steinberg S, Palsson S, Jonasson F, Sigurgeirsson B, Thorisdottir K, Ragnarsson R, Benediktsdottir KR, Aben KK, Kiemenev LA, Olafsson JH, Gulcher J, Kong A, Thorsteinsdottir U, Stefansson K. Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nat Genet.* 2007;39(12):1443–52.
  22. Sulem P, Gudbjartsson DF, Stacey SN, Helgason A, Rafnar T, Jakobsdottir M, Steinberg S, Gudjonsson SA, Palsson A, Thorleifsson G, Palsson S, Sigurgeirsson B, Thorisdottir K, Ragnarsson R, Benediktsdottir KR, Aben KK, Vermeulen SH, Goldstein AM, Tucker MA, Kiemenev LA, Olafsson JH, Gulcher J, Kong A, Thorsteinsdottir U, Stefansson K. Two newly identified genetic determinants of pigmentation in Europeans. *Nat Genet.* 2008;40(7):835–7.
  23. Branicki W, Brudnik U, Kupiec T, Wolanska-Nowak P, Szczerbinska A, Wojas-Pelc A. Association of polymorphic sites in the OCA2 gene with eye colour using the tree scanning method. *Ann Hum Genet.* 2008;72(Pt 2):184–92.
  24. Duffy DL, Montgomery GW, Chen W, Zhao ZZ, Le L, James MR, Hayward NK, Martin NG, Sturm RA. A three-single-nucleotide polymorphism haplotype in intron 1 of OCA2 explains most human eye-color variation. *Am J Hum Genet.* 2007;80(2):241–52.
  25. Dimisianos G, Stefanaki I, Nicolaou V, Sypsa V, Antoniou C, Poulou M, Papadopoulos O, Gogas H, Kanavakis E, Nicolaidou E, Katsambas AD, Stratigos AJ. A study of a single variant allele (rs1426654) of the pigmentation-related gene SLC24A5 in Greek subjects. *Exp Dermatol.* 2009;18(2):175–7.
  26. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 2001;29(1):308–11.
  27. Liu X, Jian X, Boerwinkle E: dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum Mutat.* 2011;32(8):894–9.
  28. Liu X, Wu C, Li C, Boerwinkle E. dbNSFP v3.0: a one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Hum Mutat.* 2016;37(3):235–41.
  29. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Jang W, Karapetyan K, Katz K, Liu C, Maddipati Z, Malheiro A, McDaniel K, Ovetzky M, Riley G, Zhou G, Holmes JB, Kattman BL, Maglott DR. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* 2018;46(D1):D1062–7.
  30. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. A global reference for human genetic variation. *Nature.* 2015;526(7571):68–74.
  31. Karczewski KJ, Weisburd B, Thomas B, Solomonson M, Ruderfer DM, Kavanagh D, Hamamsy T, Lek M, Samocha KE, Cummings BB, Birnbaum D. The exome aggregation consortium, Daly MJ, MacArthur DG: the ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Res.* 2017;45(D1):D840–5.
  32. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: machine learning in Python. *J Mach Learn Res.* 2011;12:2825–30.
  33. Kraskov A, Stögbauer H, Grassberger P. Estimating mutual information. *Phys Rev E Stat Nonlinear Soft Matter Phys.* 2004;69(6 Pt 2):066138.
  34. Ross BC. Mutual information between discrete and continuous data sets. *PLoS One.* 2014;9(2):e87357.
  35. Kozachenko LF, Leonenko NN. Sample estimate of the entropy of a random vector. *PRIT.* 1987;23(2):9–16.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

