Title

COITR — novel algorithm for predicting child’s externally visible traits.

Authors

Kira Kozlova [1, 2]
Igor Nizamutdinov [1]
Nikolai Slepov [1, 2]
Ekaterina Surkova [1]
Yaroslav Popov [1]
Kirill Tsukanov [1]
Valery Ilinsky [1, 3, 4, 5]
Alexander Rakitko [1, 2]

1. Genotek Ltd., Nastavnichesky lane 17 build.1, 105120, Moscow, Russia
2. Lomonosov Moscow State University, Faculty of Mechanics and Mathematics, Leninskiye Gory, Main building, 119991, Moscow, Russia
3. Pirogov Russian National Research Medical University, Ostrovityanova St. 1, 117997, Moscow, Russia
4. Institute of Biomedical Chemistry, Pogodinskaya Stt 10 build. 8, 119121, Moscow, Russia
5. Vavilov Institute of General Genetics, Gubkina St. 3, 119333, Moscow, Russia

Corresponding author
Alexandr Rakitko
Nastavnichesky lane 17 build.1, 105120, Moscow, Russia
rakitko@gmail.com

Abstract

We describe COITR (https://github.com/Genotek/COITR) — Calculator Of Infant's TRaits — a novel algorithm for predicting phenotypic features of a child from genetic data of parents. We used data from open sources and a Bayesian approach for the prediction of externally visible traits. Based on genetic data of parents the developed algorithm can determine the probability of having children with specific phenotypic characteristics: eye, hair and skin color, presence of freckles. It might be helpful for parents undergoing in vitro fertilization (IVF) with embryo selection or genome editing as well as for sperm or egg bank customers.

Text

Nowadays advances of the IVF with preimplantation genetic screening allow a selection of embryos of a specific gender or without disease-causing mutations [1]. The same technology might, in theory, be used to select embryos for various non-medical purposes. Recent statements from bioethics tolerate such practice to be introduced shortly [2]. Progress in genome editing technologies will likely allow the affordable creation of babies by design. Prediction of child's traits might soon become valuable for decision-making by future parents. It also might be useful for sperm or egg bank customers.

The association between genetic markers and personal appearance is well studied [3]. Heritability of eye and hair color is 98% [4] and 61-92% [5], respectively. Some phenotypic features are inherited together [6], which indicates that they have common genetic nature. There are several algorithms that predict the external characteristics of a person based on genetic data: HIrisPlex-S [7], Snipper [http://mathgene.usc.es/snipper/], algorithm of Venter and colleagues [8].

Here we describe the development and testing of the algorithm for prediction of phenotypic features of a child from genetic data of the parents. Our research included two stages: 1) development and validation of the algorithm to predict external features of a person from genetic data; 2) development of the algorithm to predict external characteristics of a child from genetic data of the parents.

Genetic studies based on self-reported data have recently gained popularity [17]. Over recent years the number of resources providing free access to data for scientific studies has significantly increased. Whereas externally visible traits are the issue of our study we use OpenSNP database [9] containing the genomic and self-reported phenotypic data of consumers of the direct-to-consumer genetic companies. The data selection criteria and data cleaning procedures are described in Methods.

We did not conduct a separate research to discover new genetic markers associated with phenotypic traits, although OpenSNP provides free genomic and phenotypic data making the GWAS possible. Therefore we showed that an applying of the known SNP sets can produce the result with accuracy characteristics comparable to those of previous studies. We used genetic polymorphisms described in the literature: eye color [11], hair color [12], skin color [13], freckles [14] (Supplementary Table 1). SNP rs28777 was excluded from further analyses due to high linkage disequilibrium (LD) with rs16891982 (R^2 = 0.621). After the export and primary data processing the following groups were distinguished: eye color (blue, green, hazel, brown), hair color (blond, brown, red and black), skin color (white, intermediate and black), freckles presence (yes or no). The total number of samples was 1384.

Based on the genetic data the Naive Bayes approach was used to estimate the probability of different phenotypes. We chose this model because there are cases of successful application of the Naive Bayes approach for the problem of eye and hair color prediction [15, 16] and a limited size of training dataset. However, some studies demonstrated that usage of more complex models may enhance predictive capacity of the algorithm.

The accuracy of the trained model is comparable to HirisPlex-S, published before (Supplementary Table 2). HirisPlex-S allow simultaneous prediction of eye, hair and skin color from DNA data. This algorithm is widely used in forensic genetics [7].

For instance, AUC values calculated with the help of 10-fold cross-validation were 0.95 for blue and brown eye color. The same values were 0.94 and 0.95 for blue and brown eye color, respectively, in HirisPlex-S algorithm. In our model we obtained the following values of sensitivity and specificity - 0.81 and 0.91, respectively, for blue eye color (comparable with 0.92 and 0.88 in HirisPlex-S algorithm). For hair color AUC ranged from 0.75 for brown hair to 0.9 for black hair. We obtained a wide spread of AUC values for skin color (from 0.64 for intermediate color to 0.98 for black color). This AUC range might be caused by inaccuracies in self-determined skin color and as a result a great amount of unreliable skin color data in OpenSNP. In conclusion, our model trained on the open access data reproduced results of previous studies with rather high accuracy power. Therefore this approach might be applied in the more complex task - prediction of child's phenotypic traits based on the parents' genetic data. With a knowledge of both parents' genetic data, we are able to model all possible child's genotypes and calculate their probabilities based on the Mendel's laws of heredity . For every possible child’s genotype we apply the algorithm described above to calculate the probability of having a certain phenotypic trait. Together with the known frequencies of every child’s genotype these probabilities form an assessment of the probability distributions for parents to have a child with certain phenotypic trait. The developed algorithm named COITR (Calculator Of Infant's TRaits) and allows predicting the child’s appearance for one or both parents.

To demonstrate the algorithm operation we estimated population probabilities of having a child with certain phenotypic traits for parents with fixed genotypes. Figure 1 shows the box plots for hair color posterior probabilities. The results for child’s hair color distribution for parents with fixed phenotypes are presented in Figure 2. From this chart, in particular, it follows that if one parent has blond hairs and other parent has brown hairs, the average probabilities to have a child with blond, brown, red and black hairs are 37%, 41%, 16%, 6%, respectively. Wherein 25th and 75th quartiles of probability to have a child with blond hairs are 24% and 48%, respectively. It means that taking into account of parental genotypes may correct the assess the child’s phenotype probability based on the parents’ phenotypic traits. For instance, in a quarter of cases of blond-brown pairs of parents, the predicted probability for a child to have blond hair is below 24%. In addition, the accuracy power is higher essentially if both parents’ genotypes are considered comparing with if only one parental genotype is considered. For example, for probability to have blond hairs population standard deviation are 21.1% and 16.7% for one and for both parents, respectively.

Our tool is freely available under GNU license at https://github.com/Genotek/COITR. We implemented it into genetic tests provided by Genotek Ltd. (Russia), direct-to-consumer genetic company

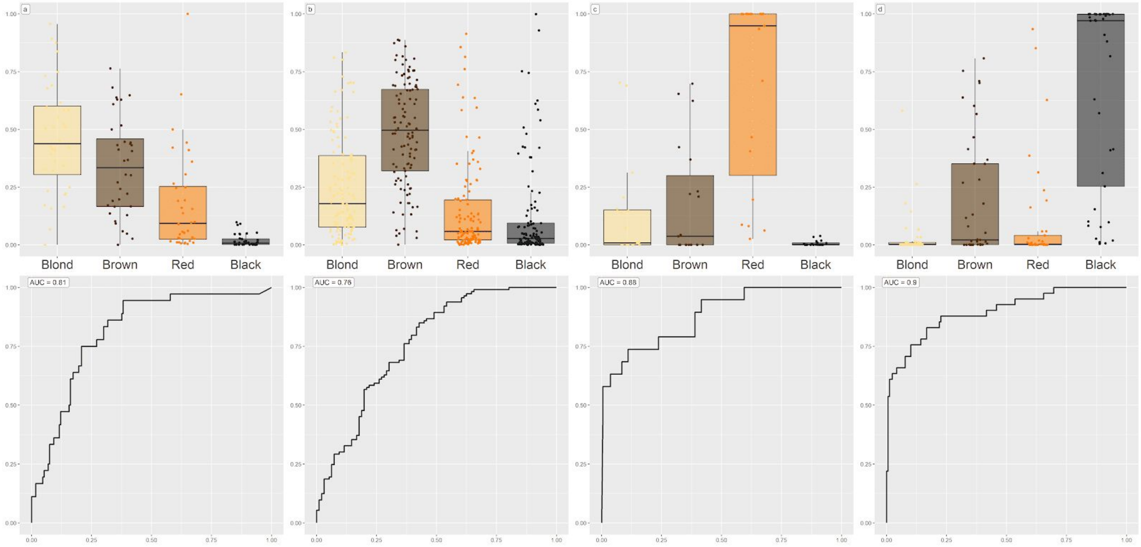

Figure 1. Box-plots for hair color prediction. a,b,c,d represent the probabilities of four hair colors for people with blond, brown, red and black hair, respectively.

a, box-plots for all samples with blond hair color. There were totally 37 samples with blond hair color in our dataset. We calculated four posterior probabilities to have blond, brown, red and black hair color for every sample (in the sum are 1). The leftmost box related to blond hair color was created using 37 probabilities for samples with blond hairs to have blond hair color. The mostright box related to black hair color was created using 37 probabilities for the same samples to have black hair color. The box related to blond hair color is notable higher comparing to others. This indicates the correct work of our algorithm, since the predicted probabilities are greater for real hair color (in case of **a**, it’s blond color). A similar trend is obtained for other colors (**b**, **c**, **d**). ROC-curve and AUC-value calculated for hair color groups “real” vs. “other” are presented under corresponding box plots. Other sections (**b**, **c**, **d**) were created in the same way.

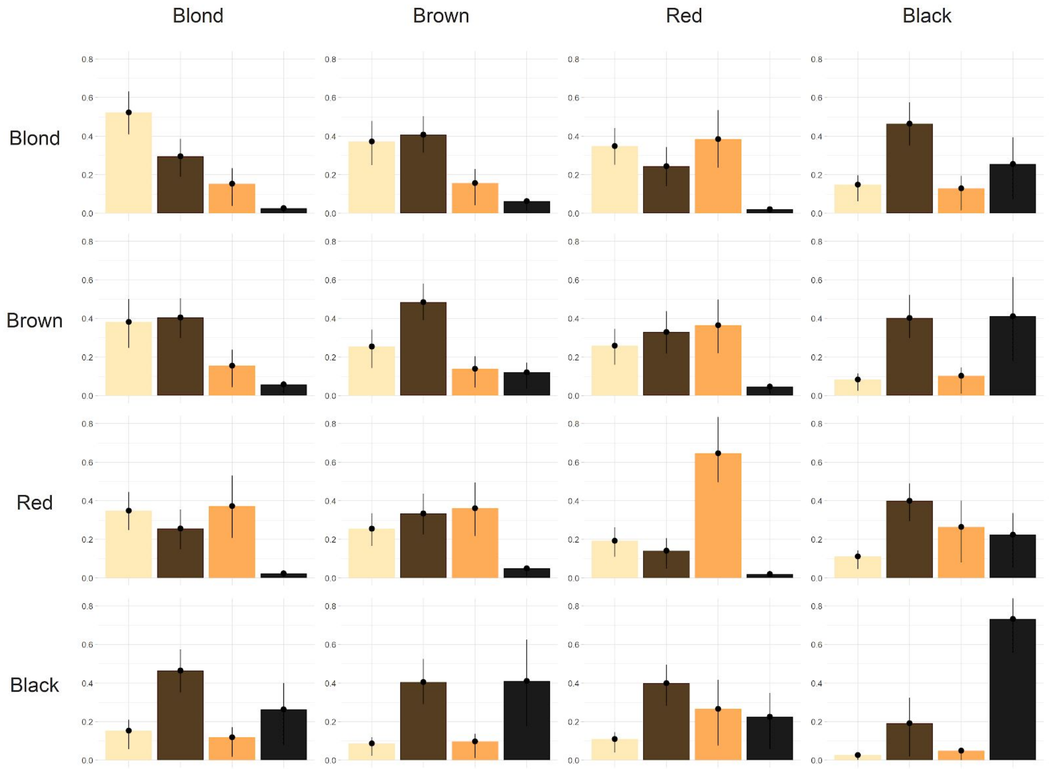

Figure 2. Probabilities of child’s birth with certain hair color, depending on the hair color of the parents (e.g., black bars correspond to black hair probabilities and so on). Bars indicate the average probability whereas whiskers indicate 25% and 75% quartiles (correspond to different genetic profiles of the parents within the same phenotypic groups).

References

1. Handyside AH. 'Designer babies' almost thirty years on. Reproduction. 2018;156(1):F75-F79. doi: 10.1530/REP-18-0157.
2. Nuffield Council on Bioethics (2018) Genome Editing and Human Reproduction: social and ethical issues (London: Nuffield Council on Bioethics)
3. Maroñas O, Söchtig J, Ruiz Y, Phillips C, Carracedo Á, 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.
4. Zhu G, Evans DM, Duffy DL, Montgomery GW, Medland SE, Gillespie NA, Ewen KR, Jewell M, Liew YW, Hayward NK, Sturm RA, Trent JM, Martin NG.A genome scan for eye color in 502 twin families: most variation is due to a QTL on chromosome 15q. Twin Res. 2004;7(2):197-210.
5. Lin BD, Mbarek H, Willemsen G, Dolan CV, Fedko IO, Abdellaoui A, de Geus EJ, Boomsma DI, Hottenga JJ. Heritability and Genome-Wide Association Studies for Hair Color in a Dutch Twin Family Based Sample. Genes (Basel). 2015;6(3):559-76. doi: 10.3390/genes6030559.
6. Lock-Andersen J, Wulf HC, Knudstorp ND. Interdependence of eye and hair colour, skin type and skin pigmentation in a Caucasian population. Acta Derm Venereol. 1998;78(3):214-9.
7. Chaitanya L, Breslin K, Zuñiga S, Wirken L, Pośpiech E, Kukla-Bartoszek M, Sijen T, Knijff P, Liu F, Branicki W, Kayser M, Walsh S. The HIrisPlex-S system for eye, hair and skin colour prediction from DNA: Introduction and forensic developmental validation.Forensic Sci Int Genet. 2018;35:123-135. doi: 10.1016/j.fsigen.2018.04.004.
8. Lippert C, Sabatini R, Maher MC, Kang EY, Lee S, Arikan 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):10166-10171. doi: 10.1073/pnas.1711125114.
9. Greshake B, Bayer PE, Rausch H, Reda J. openSNP--a crowdsourced web resource for personal genomics.PLoS One. 2014;9(3):e89204. doi: 10.1371/journal.pone.0089204. eCollection 2014.
10. 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. doi: 10.1038/nature15393.
11. Pośpiech E, Draus-Barini J, Kupiec T, Wojas-Pelc A, Branicki W. Prediction of eye color from genetic data using Bayesian approach. J Forensic Sci. 2012;57(4):880-6. doi: 10.1111/j.1556-4029.2012.02077.x.
12. Branicki W, Liu F, van Duijn K, Draus-Barini J, Pośpiech E, Walsh S, Kupiec T, Wojas-Pelc A, Kayser M. Model-based prediction of human hair color using DNA variants. Hum Genet. 201;129(4):443-54. doi: 10.1007/s00439-010-0939-8.

13. Walsh S, Chaitanya L, Breslin K, Muralidharan C, Bronikowska A, Pospiech E, Koller J, Kovatsi L, Wollstein A, Branicki W, Liu F, Kayser M. Global skin colour prediction from DNA.Hum Genet. 2017;136(7):847-863. doi: 10.1007/s00439-017-1808-5.
14. Hernando B, Ibañez MV, Deserio-Cuesta JA, Soria-Navarro R, Vilar-Sastre I, Martinez-Cadenas C. Genetic determinants of freckle occurrence in the Spanish population: Towards ephelides prediction from human DNA samples. Forensic Sci Int Genet. 2018;33:38-47. doi: 10.1016/j.fsigen.2017.11.013.
15. Ruiz Y, Phillips C, Gomez-Tato A, Alvarez-Dios J, Casares de Cal M, Cruz R, Maroñas O, Söchtig J, Fondevila M, Rodriguez-Cid MJ, Carracedo A, Lareu MV. Further development of forensic eye color predictive tests. Forensic Sci Int Genet. 2013;7(1):28-40. doi: 10.1016/j.fsigen.2012.05.009
16. Söchtig J, Phillips C, Maroñas O, Gómez-Tato A, Cruz R, Alvarez-Dios J, de Cal MÁ, Ruiz Y, Reich K, Fondevila M, Carracedo Á, Lareu MV. Exploration of SNP variants affecting hair colour prediction in Europeans.Int J Legal Med. 2015;129(5):963-75. doi: 10.1007/s00414-015-1226-y.
17. Hysi, Pirro G., Ana M. Valdes, Fan Liu, Nicholas A. Furlotte, David M. Evans, Veronique Bataille, Alessia Visconti, et al. 2018. “Genome-Wide Association Meta-Analysis of Individuals of European Ancestry Identifies New Loci Explaining a Substantial Fraction of Hair Color Variation and Heritability.” Nature Genetics 50 (5): 652–56.

Supplementary Materials

Methods

<u>*Data collection*</u>

As training dataset we used the data of OpenSNP [9] that allows customers of direct-to-customer genetic companies (23andme, FTDNA, Ancestry.com etc.) to publish their genetic and phenotypic results. In particular, customers can fulfil a survey and report their phenotypic data such as eye, hair and skin color, and freckling. The OpenSNP data were downloaded using the library of the programming language R "rsnps". We used only samples for which genetic and phenotypic data were available. To sort samples according the hair color we used the survey items “*Hair Color*”, “*Hair color*”, “*hair colour*”, “*Hair colour*” and “*hair color*”. We considered only answers “*blonde*”, “*blond*”, “*red*”, “*brown*” and “*black*” both letter-cases.

For eye color classification the survey items “*Eye color*”, “*Eye Color*” and “*eye colour*” were used and the answers “*braun*”, “*brown*”, “*blue*”, “*green*”, “*hazel*” both letter-cases were considered.

For skin color classification we used the survey items “*white skin*”, “*black skin*”, “*Skintype*”, “*Medium brown skin*” and “*Skin - Fitzpatrick Scale*”. The only answers “*caucasian*”, “*palewhite*”, “*fair*”, “*whiteskin(cantan)*”, “*whiteskin*”, “*white*”, “*pale*”, “*typeii*”, “*typei*”, “*mediterranean*”, “*lighttan*”, “*oliveskin*”, “*tanskin*”, “*tan*”, “*typeiv*” and “*brown*” both letter-cases were considered.

For sorting samples according the freckles’ presence the survey item “*Freckling*” was used and answers “*none*”, “*no*”, “*light*”, “*moderate*”, “*some*”, “*heavy*”, “*extensive*”.

The PCA plots were used to visualise sample groups and ensure in feasibility of their clusterization by analyzed SNPs. As a result of clusterization, some self-reported phenotypes were merged in common groups.

Unfortunately, quantity and quality of OpenSNP skin color data were lower essentially. It might be linked to inaccuracies in self-determined skin color and as a result to misclassification. Therefore for model training to predict skin color we use data from 1000 genome project (1000 Genomes Project Consortium et al. 2015). We selected data of populations, in which people usually have the same skin color - white (FIN finnish), intermediate (JPT japanese, MXL mexican and CLM colombian) and black

(GWD gambian and other African populations). Prediction of child's phenotypes was carried out, supposing that GBR british population has a white skin color, ACB African Caribbean in Barbados, ESN Esan in Nigeria - black skin color, and ITU Indian Telegu in the UK - intermediate skin color. Eventually the statistic model was built for the following categories - Eye color (blue / green / hazel / brown), Hair color (blond / red / brown / black), Skin color (white / intermediate / black), Freckles presence (yes / no). After all exclusions the following training sets were formed - Eye color (518 samples, OpenSNP), Hair color (217 samples, OpenSNP), Skin color (318 samples, 1000 Genomes), Freckles presence (170 samples, OpenSNP)

Prediction of phenotypic traits

The Naive Bayes approach was used to estimate the probability of traits. The choice of the approach seems reasonable due to a relatively small number of conditionally independent (the only exception are R and r haplotypes of MC1R gene which are considered as SNPs after recoding) SNPs using in the model. The Naive Bayes approach has already used in previous similar studies [15, 16] and demonstrated its applicability.

Here we describe the model we applied. Let Y be the indicator of the phenotypic group of a person, e.g. it could be the color of eyes. For simplicity, we decode all groups by positive integers. So let Y takes values in {1,...,m}. We denote person's genotypes for k tested SNPs by X_1, X_2, ..., X_k. Our goal is to estimate the posterior probability of each phenotypic group given certain genotypes:

$$P(Y = i | X_1 , X_2, ..., X_k) = \frac{P(X_1 , X_2, ..., X_k|Y = i)P(Y = i)}{\sum_{j=1}^{m} P(X_1 , X_2, ..., X_k|Y = j)P(Y = j)} = \frac{P(X_1|Y = i)...P(X_k|Y = i)P(Y = i)}{\sum_{j=1}^{m} P(X_1|Y = j)...P(X_k|Y = j)P(Y = j)} ,$$

where i is equal to 1,...,m. Conditional probabilities $P(X_1|Y = i)$, ..., $P(X_k|Y = i)$ were estimated from the training sample. For the prior probabilities $P(Y = i)$ we assume uniform distribution as it leads to higher accuracy.

Application of the Naive Bayes approach makes it possible to find the probability distribution of the particular phenotypic trait, however the point prediction value can not be find out this way. We made exact predictions of the hair, skin and eyes colors, and freckles presence based on the value of the posterior probability. If the posterior probability is higher than a particular threshold (0.56 for eyes, 0.75 for hair, 0.8 for skin color, 0.9 for freckles presence), then prediction was one value corresponding to the maximum posterior probability, if less, then two colors corresponding to two maximum probabilities. These thresholds were chosen to maximize the method's sensitivity as well as to minimize the number of cases with more than one predicted color.

Haplotype R is determined if one of the substitutions is present in the homozygous state or a substitution in any two of the four SNPs. A haplotype r is determined if one of the substitutions is present in the homozygous state or a substitution in any two of the four SNPs. In case of detection of genotype RR, the probability of red hair was determined in 100%. Other combinations of these haplotypes were used as genetic markers in the developed algorithm.

Prediction of child's traits

Given the genotypes of two persons we were able to generate all possible unique combinations of genotypes for their synthetic children. Assuming the absence of the linkage disequilibrium between SNPs we are able to calculate the exact frequencies of each combination. It allows us to compute the mean probability of each phenotypic group for "averaged" child given parent's genotypes.

Supplementary information

Table 1. List of genetic markers used to develop an algorithm for predicting the external characteristics of a person.

Phenotype	SNP	Gene/Locus
Eye color	rs12913832	HERC2
	rs1800407	OCA2
	rs12896399	SLC24A4
	rs16891982	SLC45A2
	rs1393350	TYR
	rs12203592	IRF4
Hair color	rs12913832	HERC2
	rs12203592	IRF4
	rs1042602	TYR
	rs4959270	EXOC2
	rs683	TYRP1
	rs1800407	OCA2

	rs2402130	SLC24A4
	rs12821256	KITLG
	rs16891982	SLC45A2
	rs2378249	PIGU
Skin color	rs10777129	KITLG
	rs13289	SLC45A2
	rs1408799	TYRP1
	rs1426654	SLC24A5
	rs1448484	OCA2
	rs16891982	SLC45A2
	rs2402130	SLC24A4
	rs3829241	TPCN2
	rs6058017	ASIP
Freckles	rs4911442	NCOA6
	rs2153271	BNC2
	rs12896399	SLC24A4
	rs16891982	SLC45A2
	rs1393350	TYR
	rs12203592	IRF4
	rs12821256	KITLG
Haplotype R	rs1805006-rs11547464-rs1805007-rs1805008	MC1R

Haplotype r	rs1110400-rs1805005-rs2228479-rs885479	MC1R

Table 2. Comparison of the characteristics of our algorithm and the HirisPlex-S eye, hair and skin color prediction algorithm.

	Our algorithm			HirisPlex-S		
	sensitivity	specificity	AUC	sensitivity	specificity	AUC
Eye colour						
Blue	0.8	0.91	0.95	0.92	0.88	0.94
Green	0.18	0.9	0.72	0*	0.99*	0.74*
Hazel	0.33	0.86	0.72			
Brown	0.8	0.91	0.95	0.91	0.88	0.95
Hair colour						
Blond	0.5	0.83	0.81	0.66	0.78	0.81
Brown	0.59	0.72	0.75	0.66	0.67	0.74
Red	0.68	0.9	0.87	0.62	0.99	0.93
Black	0.64	0.95	0.9	0.35	0.98	0.86
Skin colour						
Very Pale	0.33**	0.98	0.71	0.09	0.99	0.83
Pale				0.67	0.69	0.76
Intermediate	0.98	0.21	0.64	0.58	0.81	0.78
Dark	0.997**	0.7	0.98	0.53	0.99	0.98
Dark-to-Black				0.92	0.99	0.99

* The algorithm HirisPlex-S predicts the color of the eyes of three categories: blue, intermediate, brown.

** Our algorithm divides the skin color into three categories: white (corresponds to very pale or pale), intermediate (corresponds to intermediate) and black (corresponds to dark or dark-to-black).

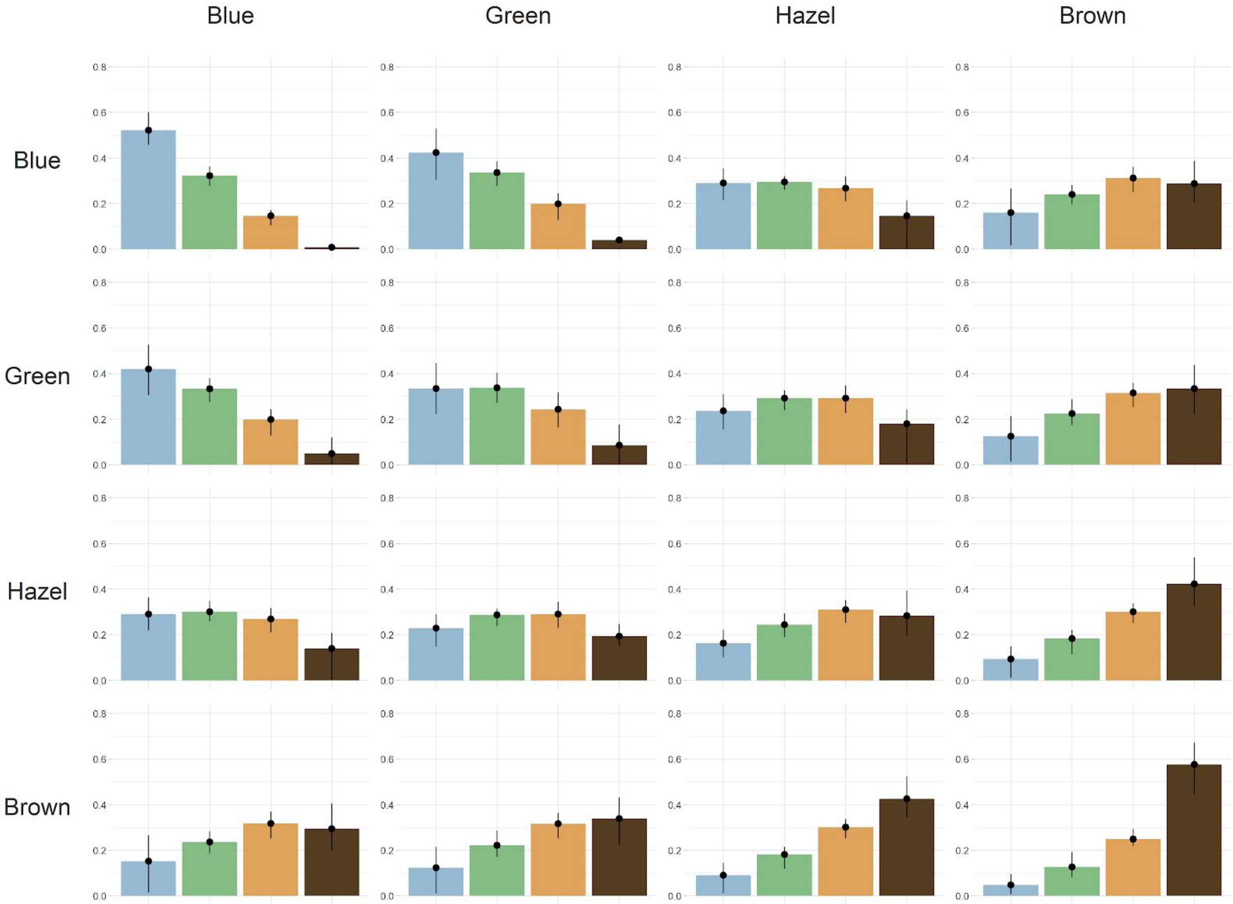

Figure 1. Probabilities of a child's birth with certain eye color, depending on the eye color of the parents. Bars indicate the average probability whereas whiskers indicate 25% and 75% quartiles (correspond to different genetic profiles of the parents within the same phenotypic groups).

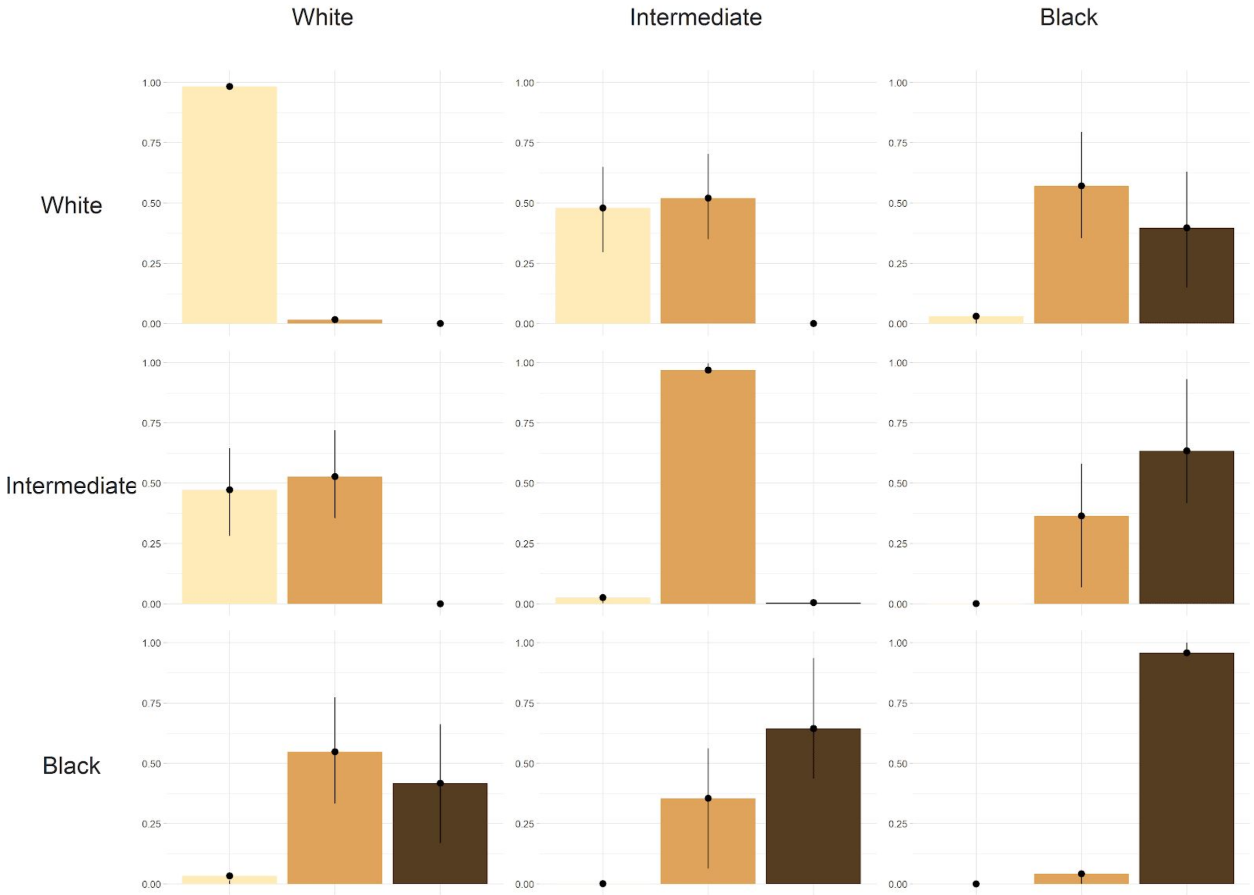

Figure 2. Probabilities of a child's birth with certain skin color, depending on the color of skin of the parents. Bars indicate the average probability whereas whiskers indicate 25% and 75% quartiles (correspond to different genetic profiles of the parents within the same phenotypic groups).

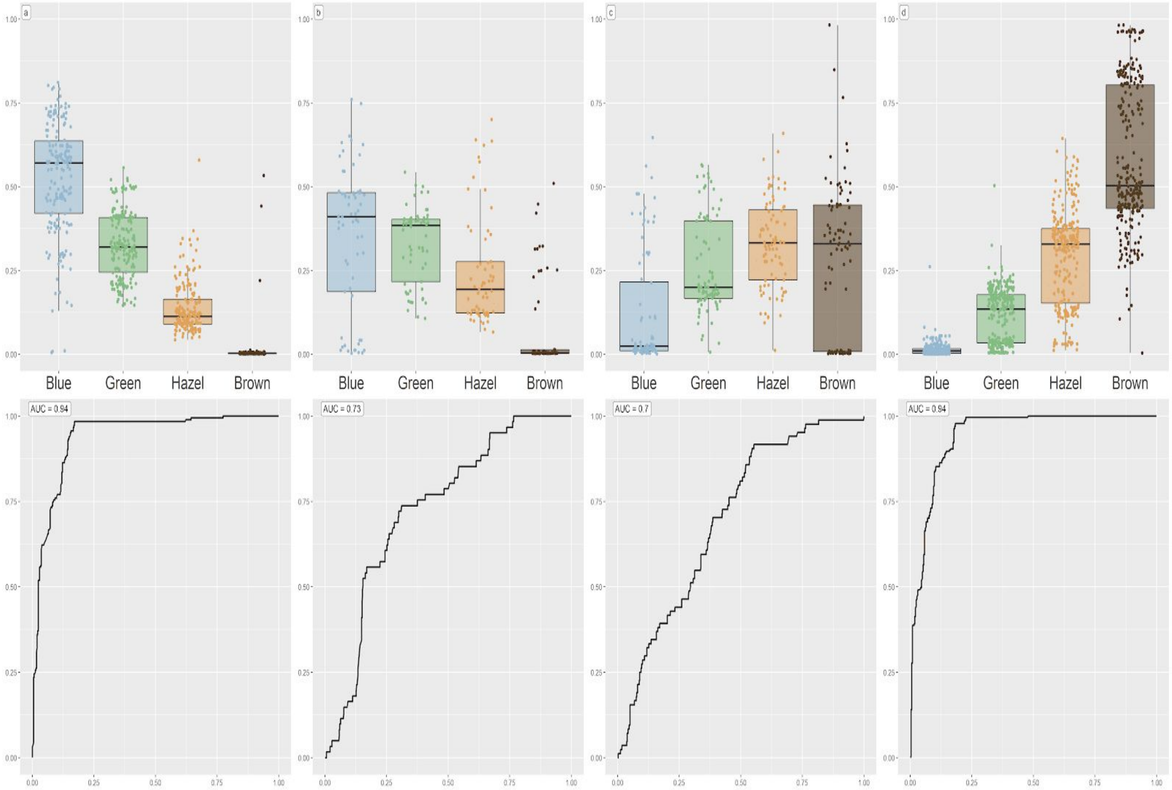

Figure 3. Box-plots for eye color prediction. a,b,c,d represent the probabilities of 4 colors of eyes for people with blue, green, hazel and brown eyes respectively.

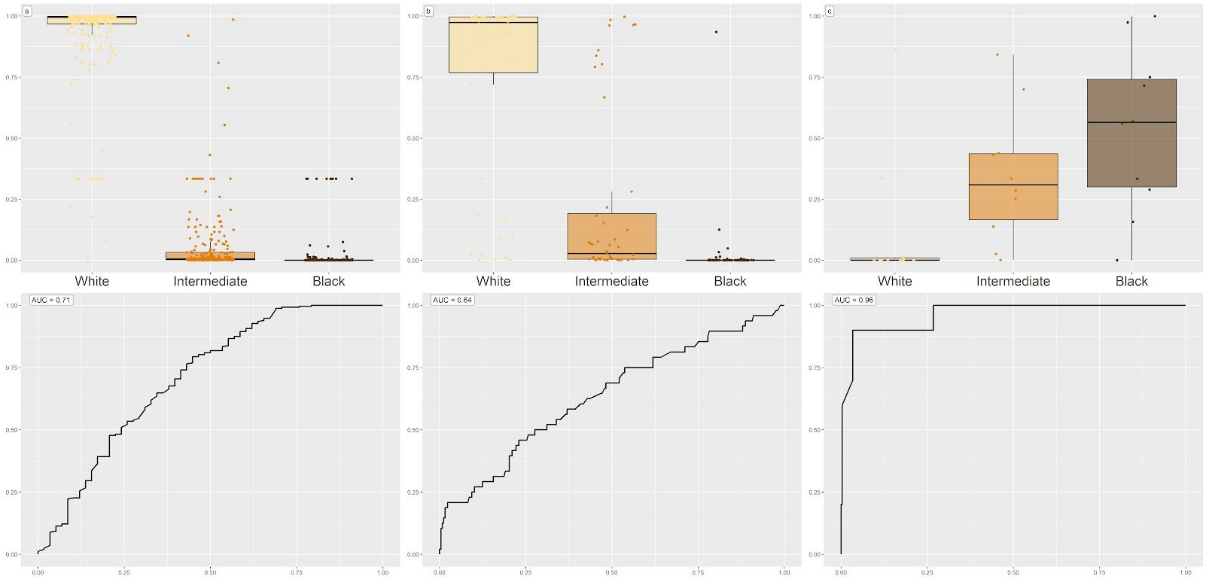

Figure 5. Box-plots for skin color prediction. a,b,c represent the probabilities of 3 colors of skin for people with white, intermediate and black skin respectively.

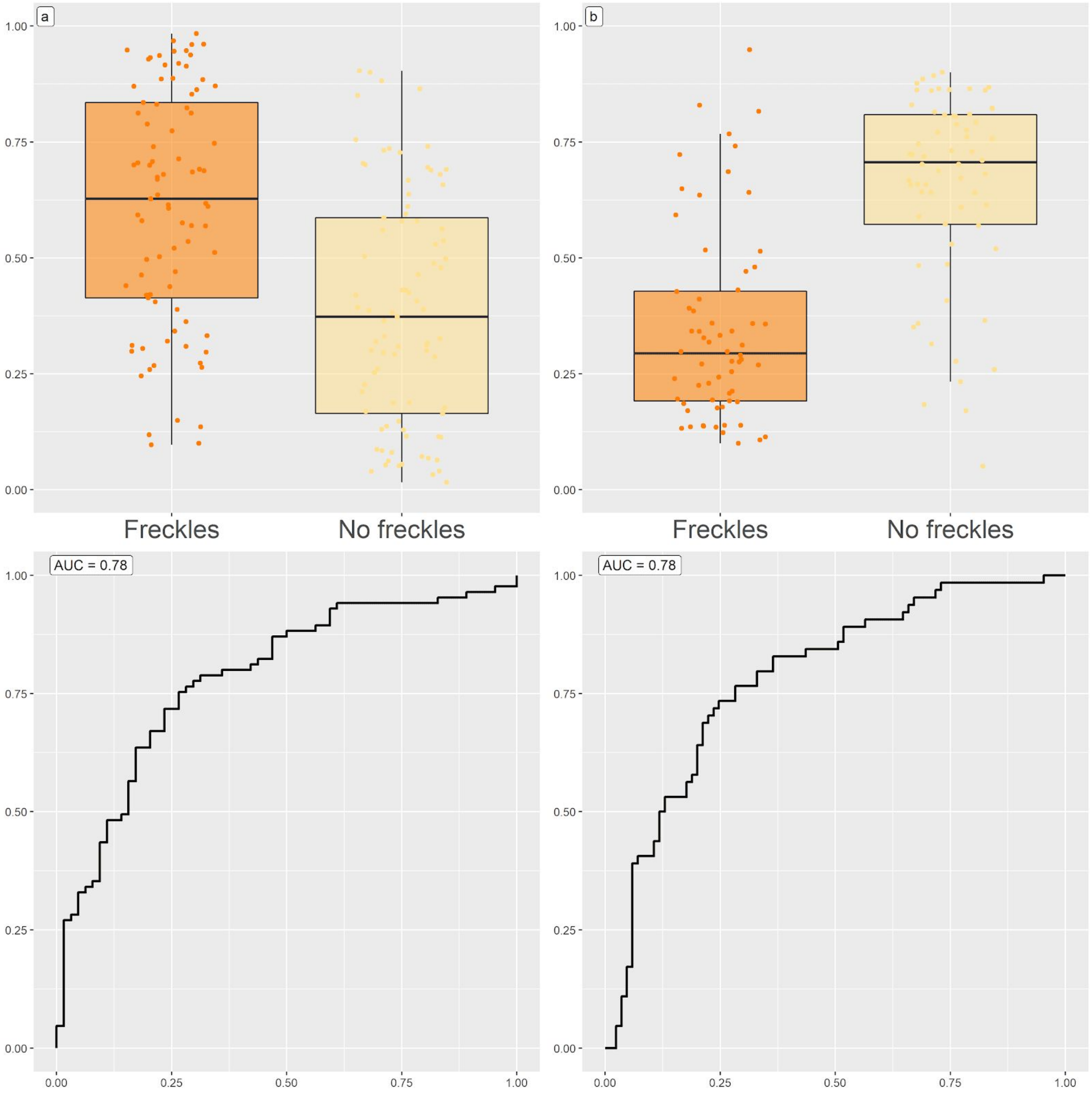

Figure 6. Box-plots for prediction of freckles presence. a,b represent the probabilities of freckles for people with presence or absence of freckles respectively.