Title

COITR — a novel algorithm for predicting externally visible traits of a child

Authors

Kira Kozlova [1, 2]
Igor Nizamutdinov [1]
Nikolai Slepov [1]
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 str 1, 117997, Moscow, Russia
4. Institute of Biomedical Chemistry, Pogodinskaya street 10 build. 8, 119121, Moscow, Russia
5. Vavilov Institute of General Genetics, Gubkina str 3, 119333, Moscow, Russia

Corresponding author
Surkova E.I.
Nastavnichesky lane 17 build.1, 105120, Moscow, Russia
esurkova@genotek.ru

Abstract

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

Text

Advances of IVF with preimplantation genetic screening allows now a selection of embryos of a specific gender or without disease-causing mutations [1]. The same technology may, in theory, be used to select embryos for variety 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.

Forecasting of possible children traits might soon become essential for decision-making by future parents. 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 with colleagues [8].

Here, we describe the development and testing of the algorithm for predicting phenotypic features of a child from genetic data of the parents.

The development of the algorithm included two stages: 1) development and validation of the algorithm to predict external features of a person from genetic data; 2) development and validation of the algorithm to predict external characteristics of a child from genetic data of the parents.

The data was obtained from Opensnp [9] using the library of the programming language R "rsnps" and 1000Genomes Phase 3 [10]. We used genetic polymorphisms described in the literature: eye color [11], hair color [12], skin color [13], freckles [14], hair structure [15] (Supplementary Table 1). Using PCA-plot, 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) and hair texture (straight, wavy and curly).

The Naive Bayes approach was applied to predict the external characteristics. 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 X1, X2, ..., Xk. Our goal is to estimate the posterior probability of each phenotypic group given certain genotypes:

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

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. We made exact predictions of the hair and eyes colors based on the value of the posterior probability. If the posterior probability is higher than a particular threshold (e.g., 0.56 for eyes and 0.75 for hair), then prediction was one value corresponding to the maximum posterior probability, if less, then two colors corresponding to two maximum probabilities. 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.

The accuracy of the trained model corresponds 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].

We applied our model to predict posterior probabilities of the infant’s features for a large number of pairs. It gives us possibility to estimate the average chance for two persons from the certain phenotypic groups have a blue-eye or black-hair child.(Fig. 1, Fig. 2, Fig. 3).

Our tool, named “COITR”, is freely available under GNU license.. We implemented it to predict future children appearance in "Children planning" direct-to-consumer genetic test done by Genotek Ltd. (Russia). Our algorithm can be further improved using more data samples.

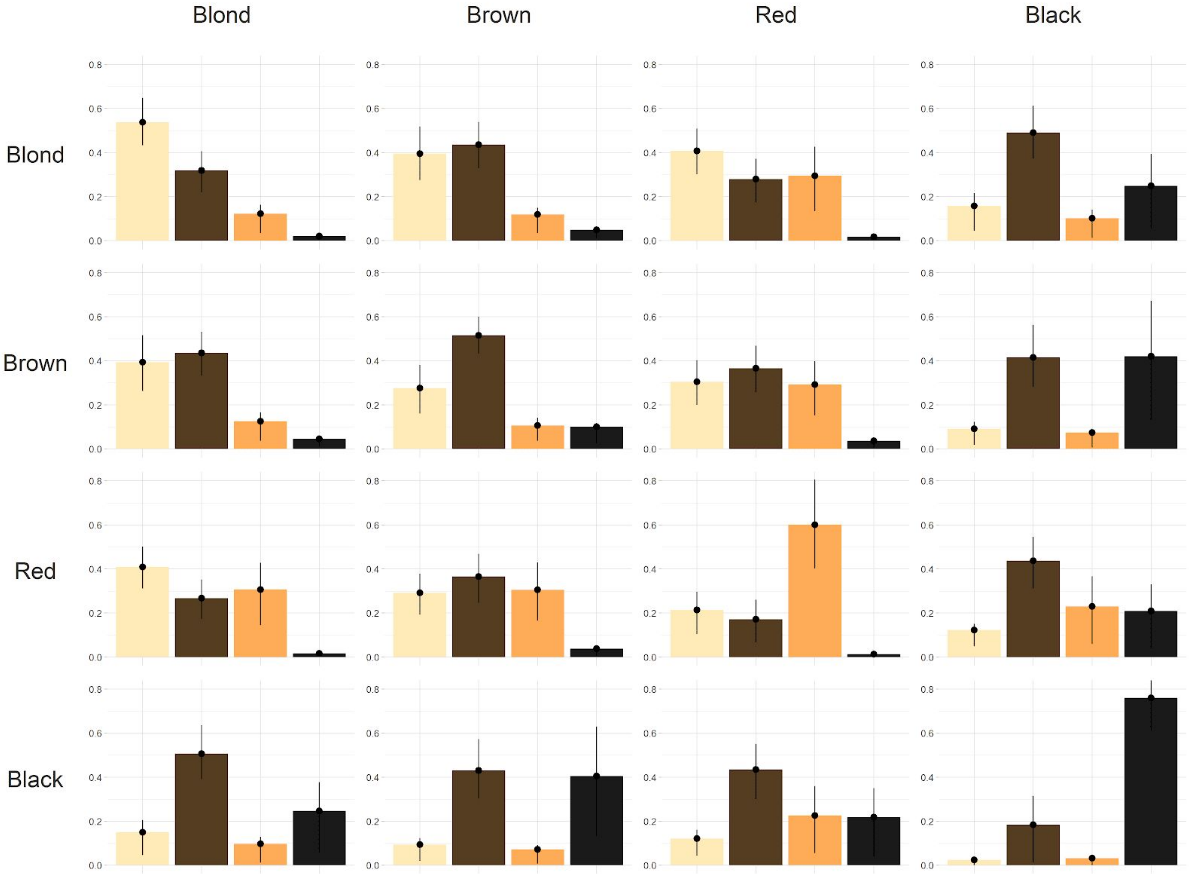

Figure 1. Probabilities of child's birth with particular hair color, depending on the hair color of the parents. Bars indicate the average probability whereas whiskers indicate 25% and 75% quartiles (correspond to different genetic profiles of the parents within the same phenotypic groups).

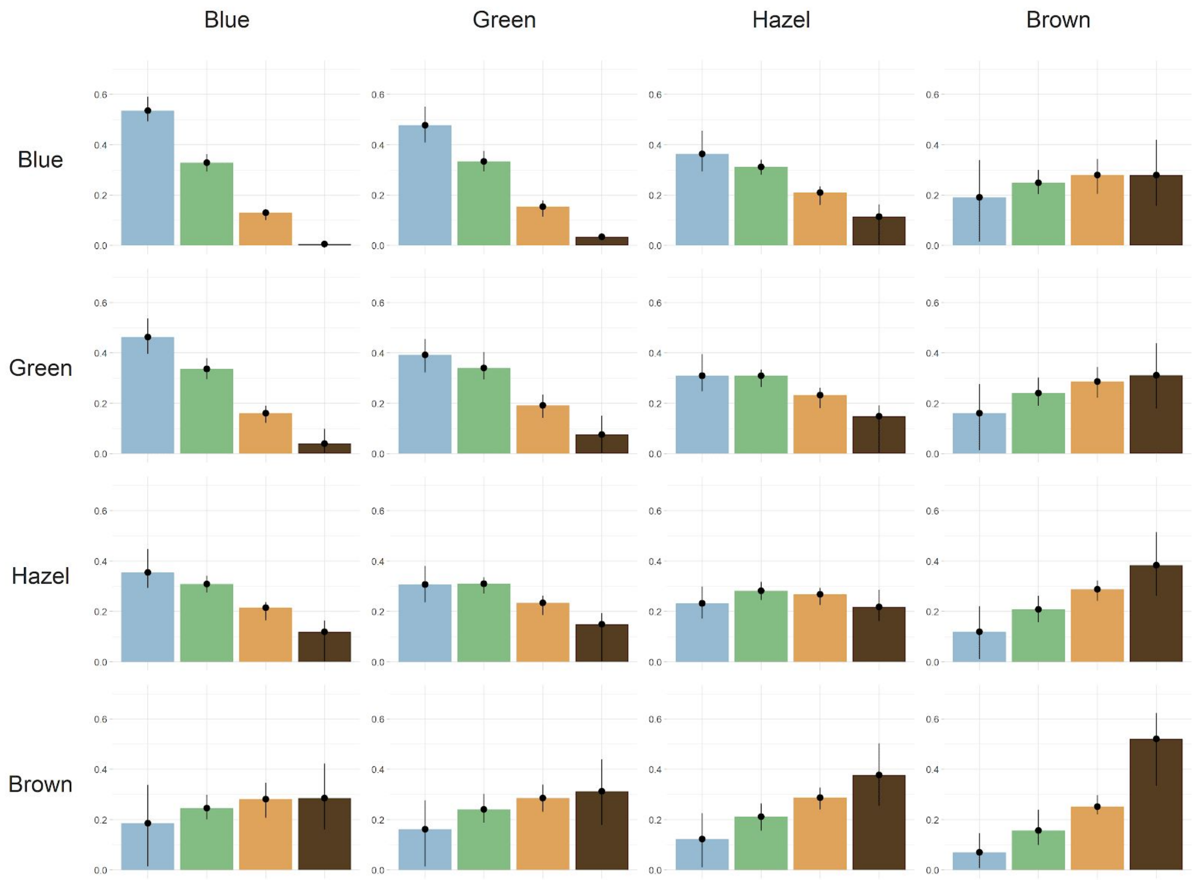

Figure 2. Probabilities of a child's birth with one or another eye color, depending on the eye color of the parents. Bars indicate the average probability whereas whiskers indicate 25% and 75% quartiles (correspond to different genetic profiles of the parents within the same phenotypic groups).

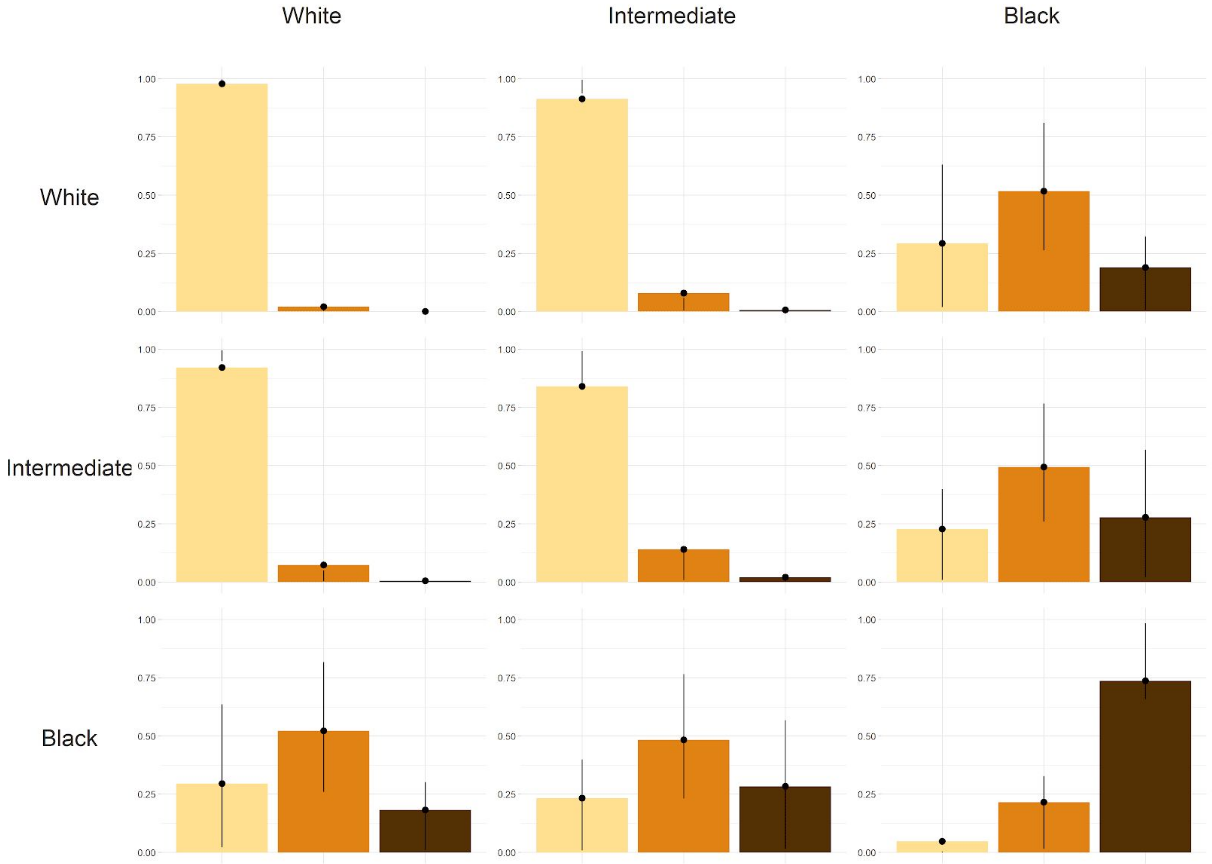

Figure 3. Probabilities of a child's birth with one or another skin color, depending on the color of skin of the parents. Bars indicate the average probability whereas whiskers indicate 25% and 75% quartiles (correspond to different genetic profiles of the parents within the same phenotypic groups).

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. Pośpiech E, Karłowska-Pik J, Marcińska M, Abidi S, Andersen JD, Berge MVD, Carracedo Á, Eduardoff M, Freire-Aradas A, Morling N, Sijen T, Skowron M, Söchtig J, Syndercombe-Court D, Weiler N, Schneider PM, Ballard D, Børsting C, Parson W, Phillips C, Branicki W; EUROFORGEN-NoE Consortium. Evaluation of the predictive capacity of DNA variants associated with straight hair in Europeans.Forensic Sci Int Genet. 2015;19:280-288. doi: 10.1016/j.fsigen.2015.09.004.

Supplementary materials

Methods

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.

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

Phenotype	SNP
Eye color	rs12913832
	rs1800407
	rs12896399
	rs16891982
	rs1393350
	rs12203592
Hair color	rs12913832
	rs12203592
	rs1042602
	rs4959270
	rs683
	rs1800407
	rs2402130
	rs12821256
	rs16891982

	rs2378249
Skin color	rs10777129
	rs13289
	rs1408799
	rs1426654
	rs1448484
	rs16891982
	rs2402130
	rs3829241
	rs6058017
Freckles	rs4911442
	rs2153271
	rs12896399
	rs16891982
	rs1393350
	rs12203592
	rs12821256
Haplotype R	rs1805006-rs11547464-rs1805007-rs1805008
Haplotype r	rs1110400-rs1805005-rs2228479-rs885479

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).

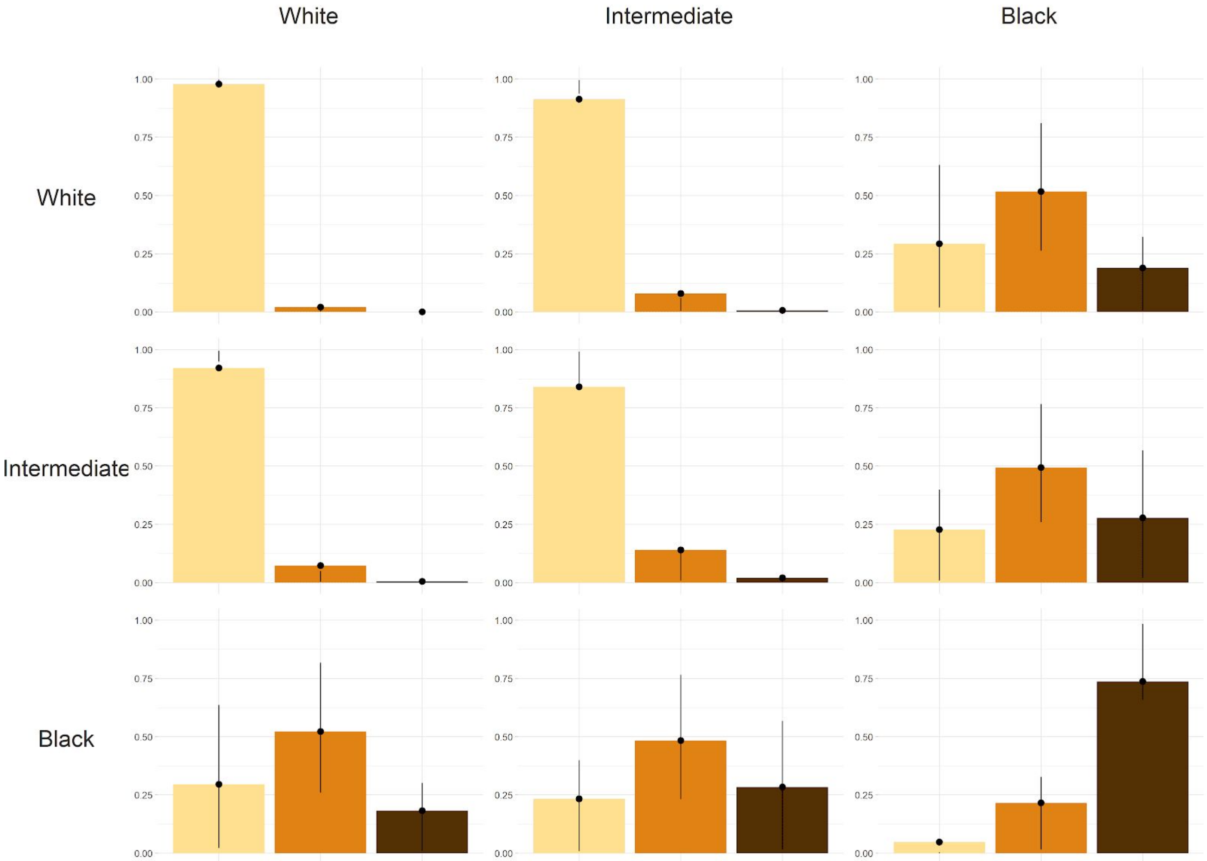

Figure 1. Probabilities of a child's birth with one or another skin color, depending on the color of skin of the parents.

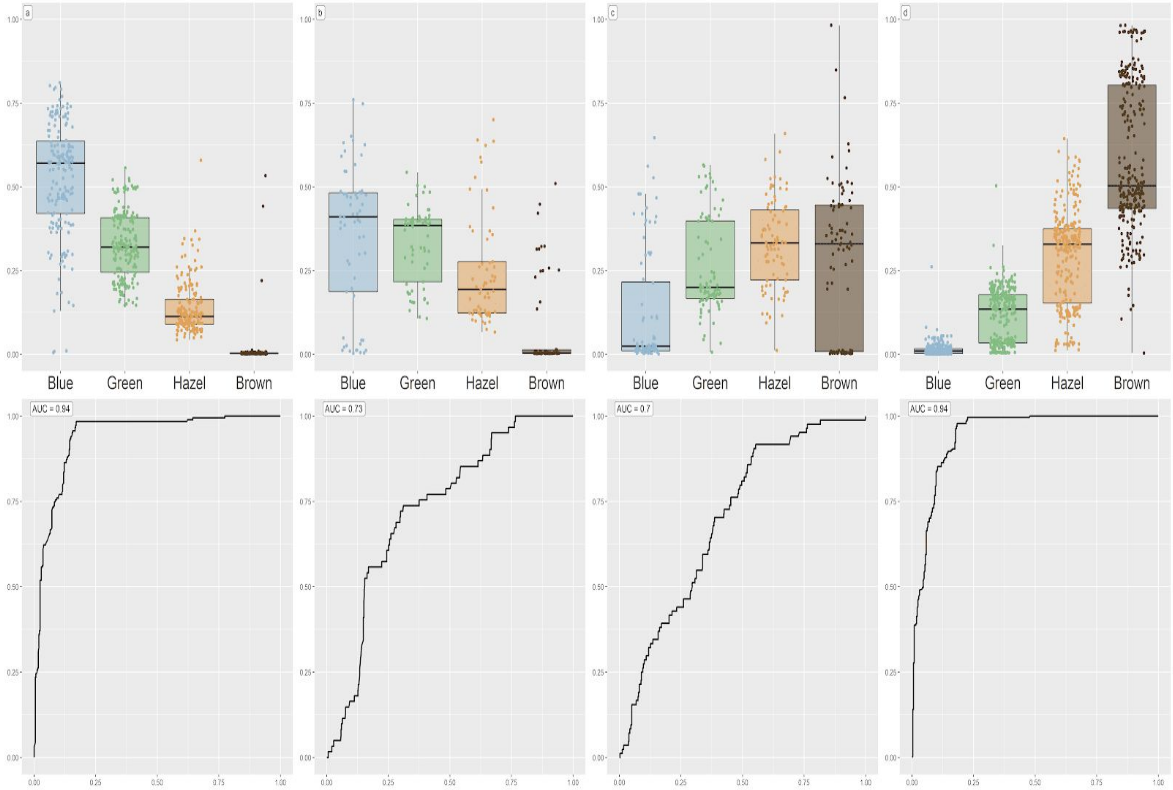

Figure 2. Box-plots for eye color prediction. a,b,c,d represent the probabilities of 4 colors of eyes for people with blue, green, hazel and brown eyes respectively.

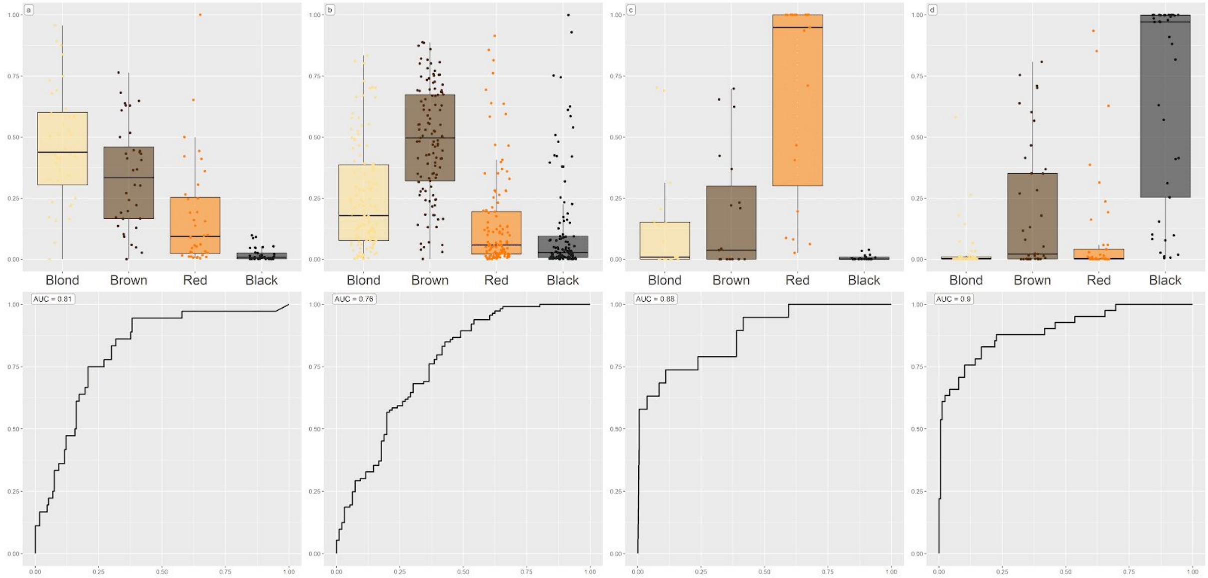

Figure 3. Box-plots for hair color prediction. a,b,c,d represent the probabilities of 4 colors of hair for people with blond, brown, red and black hair respectively.

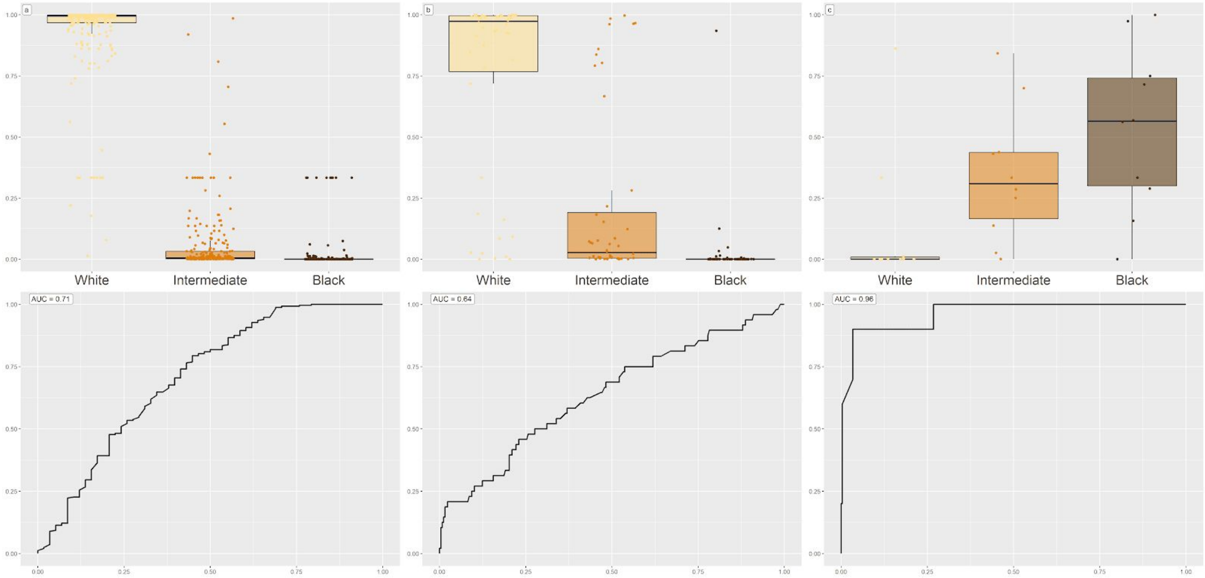

Figure 4. Box-plots for skin color prediction. a,b,c represent the probabilities of 3 colors of skin for people with white, intermediate and black skin respectively.

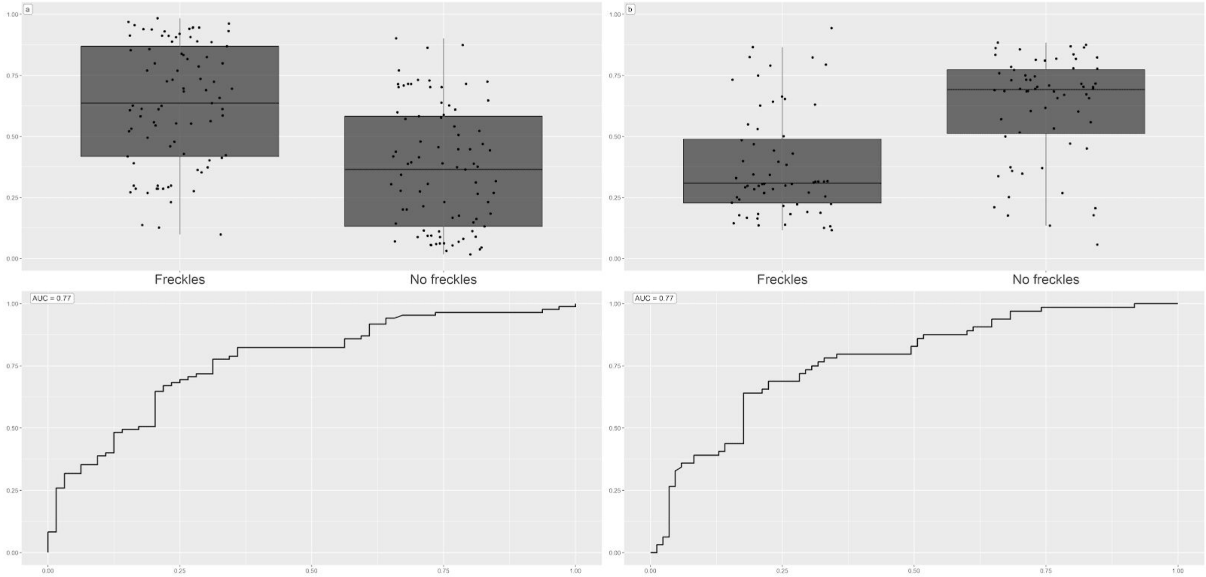

Figure 5. Box-plots for prediction of freckles presence. a,b represent the probabilities of freckles for people with presence or absence of freckles respectively.