Accurate prediction of multiple pigmentation phenotypes

Babak Alipanahi*, Pierre Fontanillas, Michael Multhaup, Suyash Shringarpure, Catherine Wilson, Michaela Johnson, the 23andMe Research Team, Steven Pitts, Adam Auton, and Robert Gentleman

23andMe, Inc., Mountain View, CA – research.23andme.com *balipanahi@23andme.com

Introduction

Eye, hair, and skin pigmentation are amongst the most heritable of human traits. These traits are highly polygenic, with dozens of loci having been identified as associated [1], many of which show evidence of epistatic interactions [2]. Perhaps not surprisingly, the genetic architectures of these three pigmentation traits are highly overlapping. However, this feature has not been used so far for predicting pigmentation. Here, we applied a custom-designed deep neural network on self-reported pigmentation phenotypes from a large cohort of European ancestry 23andMe research participants for joint-prediction of eye, hair, and skin colors. Our model can be trained on tens of thousands of samples and shows excellent prediction performance on out-of-sample test data.

Methods

Data Collection

Data were collected via online surveys of 23andMe participants that have consented to participate in research. The self-reported pigmentation phenotypes used in this study were collected within a broader survey on skin cancer and associated risk factors. The possible options for the three phenotypes are listed in **Table 1**.

Table 1: Pigmentation phenotype levels

Phenotype	Levels
Eye color	Blue, Blue-Gray, Gray, Gray-Green, Green, Light-Brown, Medium-Brown, Dark Brown/Black
Hair color	Red, Light Blond, Dark Blond, Light Brown, Dark Brown, Black
Skin color	Extremely Fair, Light, Medium, Olive, Brown, Black

To separate the training and testing data, we used the responses collected in the four months of the survey as the training dataset and the subsequent six months as the validation dataset (see **Fig. 1**).

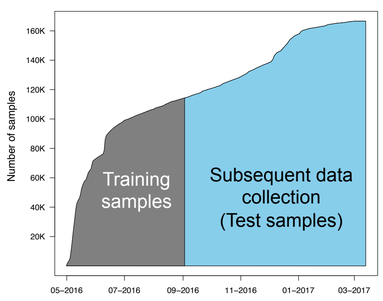


Figure 1: **Data acquisition as a function of date.** To train our pigmentation model, we selected data collected prior to September 2016, with subsequently collected data reserved for model validation purposes. All plots in this poster are made from the validation data

Predictors

To exhaustively identify SNPs that could discriminate between self-reported pigmentation levels, for each pigmentation trait, we performed a three-step process for picking the predictors:

1. Pairwise and one-vs-others genome-wide association studies (GWAS) on all phenotypes levels (e.g., Blue vs Brown and Blue vs Non-blue)
2. Conditional analyses on each of the significant loci from the previous step and picking all significant secondary signals
3. Interaction analysis on all SNP:SNP and SNP:covariate (age, sex and genetic PCs) interactions

This process helped us discover several novel SNPs in highly-significant loci with complex haplotype structures, such as the *MC1R* and *OCA2-HERC2* loci.

The Model

As these phenotypes are correlated, instead of building multiple independent predictive models, we built a joint model to predict all pigmentation traits simultaneously via multitask learning [3]. Pigmentation phenotypes are significantly correlated and their GWASs share tens of genomic loci. Hence, it seems natural to jointly predict all of them.

Our neural network architecture is shown in **Fig 2A**. We adapted a multi-output neural network architecture, used linear node activations for the output layer, and minimized the summation of the three mean squared errors corresponding to each of the phenotypes. For targets, we mapped the first level of each phenotype to 0, the next level to 1 and so on. We used L1 and L2 weight regularization and dropout to regularize the model. Model hyperparameters were determined by searching over 200 random configurations. The model was coded in Keras [4] and optimized and tested on a Tesla K80. The final model was an ensemble of 10 models learned on bootstrapped (random sampling with replacement) samples for training and unsampled data for validation.

To analyze the importance of the input features, we used the gradient of the outputs with respect to input idea proposed in [5]. The top 29 features are shown in **Fig. 2B**. It is noteworthy that several interaction terms have significant contributions to all outputs.

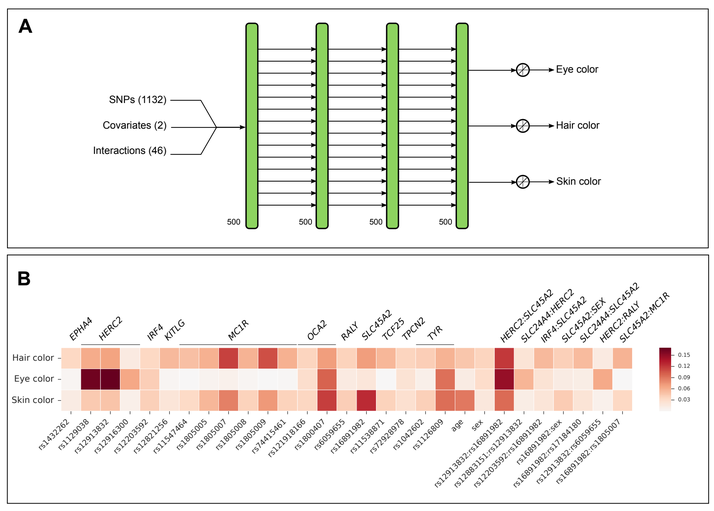


Figure 2: **A. The neural network architecture.** The number of nodes in each hidden layer is shown at the bottom left of that layer. All hidden layer nodes have SeLU activations and the output nodes are linear. The numbers in parenthesis next to inputs denotes their dimensionality. **B.** The estimated importance of the top 29 input features on the three outputs. Darker shades indicate higher contribution.

Results

We computed the model predictions trained on 117k individuals for the 31k held-out test individuals. The outputs of Pigmentor are scalar indices for each of the three pigmentation phenotypes. However, we have the reported level from the respondents and can plot the level distributions (**Fig. 3**) and compute the pairwise and aggregate AUCs (**Table 2**).

We note that in spite of assuming arbitrary uniform spacing between phenotypes levels, the model puts the modes of the level distributions in an intuitively-meaningful manner.

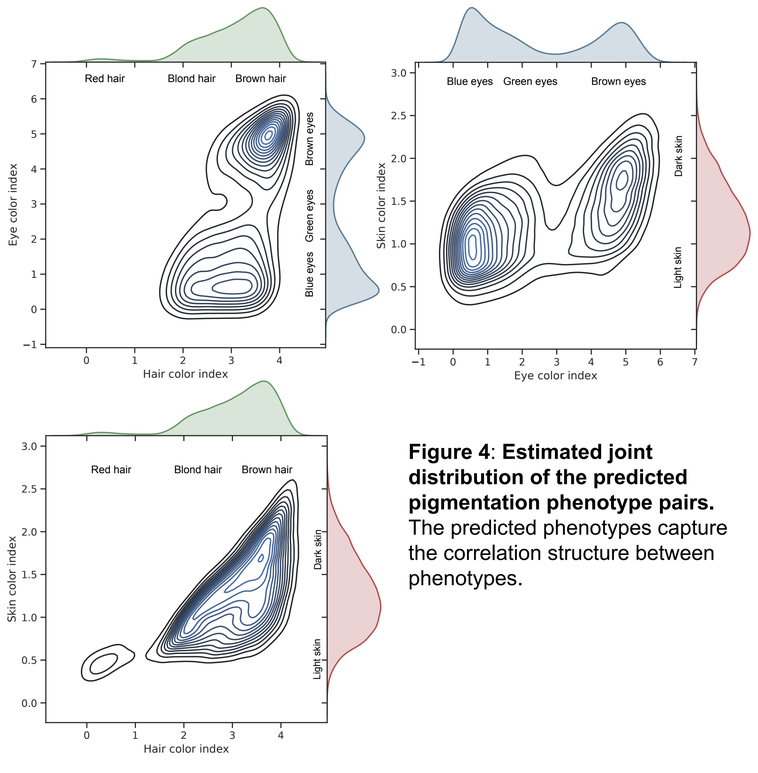
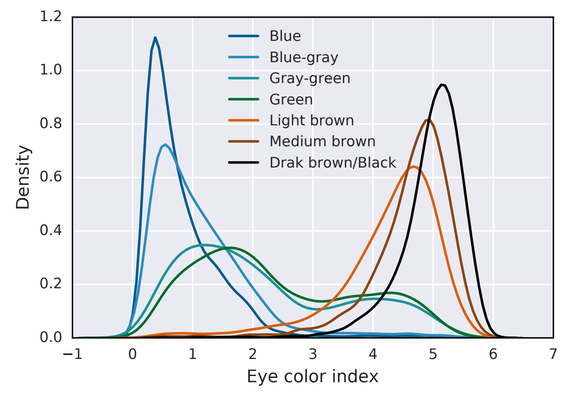
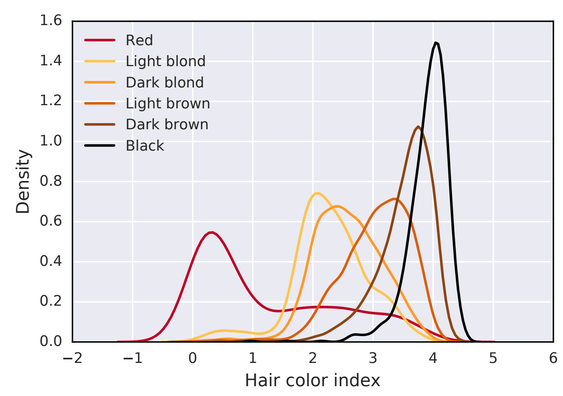


Figure 4: **Estimated joint distribution of the predicted pigmentation phenotype pairs.** The predicted phenotypes capture the correlation structure between phenotypes.





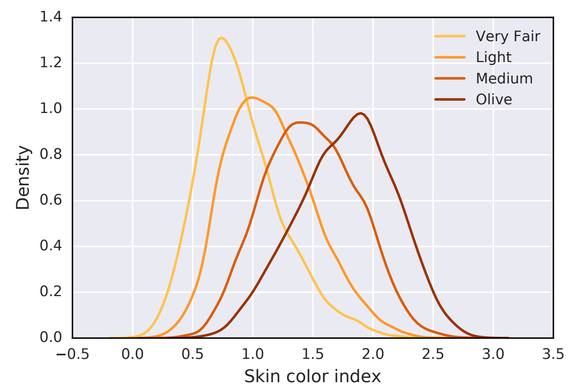


Figure 3: **Estimated probability distribution functions for all levels of the three pigmentation phenotypes.** From top to bottom eye color, hair color and skin color, respectively.

To assess the performance of Pigmentor, for each phenotype, we computed pairwise AUCs between all pairs of levels and then computed the aggregate AUC as follows:

$$\text{Aggregate AUC} = \frac{\sum_{i,j} \pi_i \pi_j \text{AUC}_{i,j}}{\sum_{i,j} \pi_i \pi_j}$$

Where π_i is the frequency of level *i* and $\text{AUC}_{i,j}$ is the pairwise AUC between levels *i* and *j*.

Table 2: Aggregate AUCs of pigmentation phenotypes

Phenotype	Aggregate AUC
Eye color	86.3%
Hair color	81.7%
Skin color	76.4%

Discussion

Pigmentor is a useful tool to predict the pigmentation phenotypes while preserving their inherent correlations (see **Fig. 4**), to understand the architecture of these traits, and to analyze their involvement in skin cancer risk or other pigmentation-associated disorders.

References

[1] Sulem, P. *et al.* (2007) Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nat Genet.*12:1443-52
[2] Duffy, D.L. e*t al.* (2010) Multiple pigmentation gene polymorphisms account for a substantial proportion of risk of cutaneous malignant melanoma. *J Invest Dermatol.* 130(2):520-8.
[3] Caruana, R. (1997) Multitask learning. *Machine Learning.* 28: 41-75.
[4] Chollet, F. (2015) Keras. https://github.com/fchollet/keras.
[5] Simonyan et al. (2013) Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. arXiv:1312.6034v2.

Acknowledgments

We would like to thank the research participants of 23andMe who provided consent to participate in the research, and for enabling this study. We also thank employees of 23andMe who contributed to the development of the infrastructure that made this research possible.