Supplemental data

Eye color and the prediction of complex phenotypes from genotypes

Fan Liu, Kate van Duijn, Johannes R. Vingerling, Albert Hofman,

André G. Uitterlinden, A. Cecile J.W. Janssens, and Manfred Kayser

Supplemental experimental procedures

Population characteristics

The Rotterdam Study is a population-based prospective study of subjects aged 55 years or older [S1,S2]. Collection of eye (iris) colour data and purification of DNA have been described in detail previously [S3]. In brief, each eye was examined by slit lamp examination by an ophthalmological medical researcher, iris color was graded by standard images showing various degrees of iris pigmentation and categorized into blue, brown and non-blue / non-brown called here intermediate. The Medical Ethics Committee of the Erasmus Medical Center approved the study protocol, and all participants provided written informed consent. Individuals identified as outliers using an identity-by-state analysis as described previously [S4] have been excluded because they most likely represent individuals of non-European ancestry.

SNP selection and genotyping

We selected 37 SNPs that were statistically significantly associated with human iris colour in previous studies [S3,S5-S13] (Supplementary Tables 1 and 2). Multiplex genotype assay design was performed with the software MassARRAY Assay Design version 3.1.2.2 (Sequenom Inc., San Diego, CA). We designed two 17-plex iPlex multiplexes, sequences of forward, reverse and extension primers are provided in Supplementary Table 1. The Sequenom genotyping was performed on 5 ng of dried genomic DNA in 384-well plates (Applied Biosystems Inc. Foster City, CA) in a reaction volume of 5 µl containing 1x PCR Buffer, 1.625 mM MgCl, 2.5 µM dNTPs, 100 nM each PCR primer, 0.5 U PCR enzyme (Sequenom). The reaction was incubated in a GeneAmp PCR System 9700 (Applied Biosystems) at 94 ºC for 4 minutes followed by 45 cycles of 94 ºC for 20 seconds, 56ºC for 30 seconds, 72ºC for 1 minute, and finalized by 3 minutes at 72ºC. To remove the excess dNTPs, 2 µl SAP mix containing 1x SAP Buffer and 0.5 U shrimp alkaline phosphatase (Sequenom) was added to the reaction. This was incubated in a GeneAmp PCR System 9700 (Applied Biosystems) at 37ºC for 40 minutes followed by 5 minutes at 85ºC for deactivation of the enzyme. Then 2 µl of Extension mix is added containing a concentration of adjusted extend primers varying between 3.5-7 µM for each primer, 1x iPLEX buffer (Sequenom), iPLEX termination mix (Sequenom) and iPLEX enzyme (Sequenom). The extension reaction was incubated in a GeneAmp PCR System 9700 (Applied Biosystems) at 94ºC for 30 seconds followed by 40 cycles of 94ºC for 5 seconds, 5 cycles of 52ºC for 5 seconds, and 80ºC for 5 seconds, and finalized at 72ºC for 3 minutes. After the extension reaction desaltation was carried out by adding 6 mg Clean Resin (Sequenom) and 16 µl water followed by rotating the plate for 15 minutes. The extension product was spotted onto a G384 + 10

SpectroCHIP (Sequenom) with the MassARRAY Nanodispenser model rs1000 (Sequenom). The chip was then transferred into the MassARRAY Compact System (Sequenom) where the data was collected, using TyperAnalyzer version 4.0.3.18 (Sequenom), SpectroACQUIRE version 3.3.1.3 (Sequenom), GenoFLEX version 1.1.79.0 (Sequenom) and MassArrayCALLER version 3.4.0.41 (Sequenom). For quality control reasons, the data was checked manually after data collection. In addition, rs6058017, was typed with the commercially available Taqman assay C__22275334_10 as recommended by the manufacturers (Applied Biosystems) and data for two other SNPs (rs12203592 and rs1408799) were used from microarray genotyping performed in the whole Rotterdam Study cohort using the Infinium II HumanHap550K Genotyping BeadChip® version 3 (Illumina Inc. San Diego, CA) as described in detail previously [S4].

Association and linkage disequilibrium testing

Single SNP association was verified using a linear model where blue, intermediate, and brown were coded as 1, 2, and 3 quantitatively, and SNP genotypes were coded as 0, 1, or 2 minor alleles. Notably, rs12913832 in the *HERC2* gene showed the largest effect (beta = 1.13, $P < 1.0\times10^{-300}$), in agreement with previous findings [S7,S8]. Adjusting for the effect of rs12913832 led to multiple SNPs in the *HERC2/OCA2* region becoming less or non-significant (Supplementary Table 3), as expected due to the existing linkage disequilibrium (LD). However, rs1800407, a non-synonymous SNP in *OCA2* (Arg419Gln), displayed considerably stronger significance after adjustment ($P = 1.7\times10^{-28}$ adjusted versus 7.7×10^{-13} unadjusted), indicating an independent effect. Interestingly,

this SNP was reported to act as a penetrance modifier of *HERC2* rs12913832 [S7]. We performed a tagging SNP analysis excluding markers in strong LD (pair-wise r2>0.8) using software package Haploview 4.1 [S14]. Thirteen SNPs in strong LD were excluded from the *OCA2-HERC2* region (Supplementary Figure 1A and B). Thus, a total of 24 SNPs were included in prediction analyses (Supplementary Table 3).

Prediction modelling

The Rotterdam Study cohort was randomly split into a model-building set consisting of 3804 individuals and a model verification set consisting of the remaining 2364 individuals. Five models were constructed in the model-building set described in detail below.

Ordinal regression

Ordinal regression is often used when the response is categorical with ordered outcomes. The model provides predicted probabilities, inside the probability space, for each level of the response without assuming constant variance. Consider eye colour, *y*, to be three ordinal levels "blue," "intermediate", and "brown", which are determined by the genotype, *x*, of *k* SNPs. Let π1, π2, and π3 denote the probability of "blue," "intermediate", and "brown", respectively. The ordinal regression can be written as

$$\text{logit}(\Pr(y \le \text{blue} \mid x_1 \ldots x_k)) = \ln\left(\frac{\pi_1}{1-\pi_1}\right) = \alpha_1 + \sum \beta_k x_k$$

$$\text{logit}(\Pr(y \le \text{inter} \mid x_1 \ldots x_k)) = \ln\left(\frac{\pi_1 + \pi_2}{1-(\pi_1 + \pi_2)}\right) = \alpha_2 + \sum \beta_k x_k \ ,$$

where α and β can be derived in the model-building set.

Eye colour of each individual in the model-verification set can be probabilistically predicted based on his or her genotypes and the derived α and β,

$$\pi_1 = \frac{\exp(\alpha_1 + \sum \beta_k x_k)}{1 + \exp(\alpha_1 + \sum \beta_k x_k)},$$

$$\pi_2 = \frac{\exp(\alpha_2 + \sum \beta_k x_k)}{1 + \exp(\alpha_2 + \sum \beta_k x_k)} - \pi_1, \text{ and}$$

$$\pi_3 = 1 - \pi_1 - \pi_2.$$

Multinomial logistic regression

Multinomial logistic regression is often used for categorical outcomes, where the model does not assume ordinary data, which can be written as:

$$\text{logit}(\Pr(y = \text{blue} \mid x_1 ... x_k)) = \ln\left(\frac{\pi_1}{\pi_3}\right) = \alpha_1 + \sum \beta(\pi_1)_k x_k$$

$$\text{logit}(\Pr(y = \text{inter} \mid x_1 ... x_k)) = \ln\left(\frac{\pi_2}{\pi_3}\right) = \alpha_2 + \sum \beta(\pi_2)_k x_k ,$$

and the probabilities for each individual being a certain colour category can be estimated as:

$$\pi_1 = \frac{\exp(\alpha_1 + \sum \beta(\pi_1)_k x_k)}{1 + \exp(\alpha_1 + \sum \beta(\pi_1)_k x_k) + \exp(\alpha_2 + \sum \beta(\pi_2)_k x_k)}$$

$$\pi_2 = \frac{\exp(\alpha_2 + \sum \beta(\pi_2)_k x_k)}{1 + \exp(\alpha_1 + \sum \beta(\pi_1)_k x_k) + \exp(\alpha_2 + \sum \beta(\pi_2)_k x_k)}, \text{ and}$$

$$\pi_3 = 1 - \pi_1 - \pi_2.$$

For ordinal and multinomial logistic regressions, the colour category with the $\max(\pi_1, \pi_2, \pi_3)$ was considered as the predicted colour.

Fuzzy c-means clustering (FCM)

There have been increasing interests in the methods based on machine-learning techniques, such as fuzzy logic and artificial neural networks. These methods can conveniently map an input space to an output space that is related through nonlinear functions which sometimes can be statistically complex. In this study we also constructed two prediction models based on fuzzy C-means clustering (FCM) and pattern-reorganization neural networks. FCM clustering is the most frequently used algorithm in generating a fuzzy inference system (FIS). It is based on iterative minimization of an objective function wherein each data point belongs to a cluster to some degree that is specified by a membership grade [S15]. A Sugeno-type FIS structure was generated based on FCM clustering in the model-building set. The input space was defined as a *k*-marker by *N*-individual matrix of the number of minor alleles plus one. The target variable was defined as a 3 by *N* matrix, where each row vector represents yes-no of the corresponding colour type. The generated FIS was subsequently used to predict eye colours in the model-verification set, returning 3 membership grades of values between 0 and 1 for each individual indicating his or her colour type. The colour type with the maximal membership grade was considered as the predicted colour.

Neural networks

Neural networks have been used to characterize gene-gene interactions [S16], find SNP-phenotype associations [S17,S18], and predict genetic phenotypes [S19]. A feed-forward network for pattern recognition was initialized in the model-building set, by specifying

tan-sigmoid transfer functions in both the hidden and output layers. The hidden layer contained 10 arbitrary neurons and the output later contained three output neurons, each represents yes-no for one colour type. The pattern recognition network was then trained using scaled conjugate gradient algorithm where the inputs and targets followed the same format described in the FCM section. During training, the model-building data set was randomly divided into three subsets, 60% were used for training, 20% were used to control for over-fitting by comparing the mean squared errors. The last 20% were used as an independent test of network generalization. The derived pattern recognition network was subsequently used to predict colours in the model-verification set, returning 3 numeric vectors with values between 0 and 1. The colour type with the maximal value was considered as the predicted colour.

Classification Tree

Classification tree, one of the main data mining techniques, is used to predict membership of categorical objects from one or more predictors [S20]. Compared to multiple regression that simultaneously analyzes multiple predictors, the classification tree hierarchically and recursively conducts single regression analyses, where the next regression on a different predictor is conducted in the samples not classified in a previous regression. The assumptions regarding the level of measurement of predictors are less stringent compared to multiple regression. In the current study, the classification tree was trained in the model-building set and was used to predict eye colours in the model-verification set, returning an outcome with 3 categories representing each colour.

Model evaluation

We evaluated the performance of the five prediction models in the model-verification set. A 2 by 2 confusion table was derived for each colour type. The predicted colour types were classified as true positives (TP), true negatives (TN), false positives (FP), or false negatives (FN). Four measurements of the prediction performance were derived:

1) Sensitivity = TP/(TP+FN)×100 is the percentage of correctly predicted colour type among the observed colour type.

2) Specificity = TN/(TN+FP)×100 is the percentage of correctly predicted non-colour type among the observed non-colour type.

3) Positive predictive value (PPV) = TP/(TP+FP)×100 is the percentage of correctly predicted colour type among the predicted positives.

4) Negative predictive value (NPV) = TN/(TN+FN)×100 is the percentage of correctly predicted non-colour type among the predicted negatives.

Additionally, we measured the area under the receiver operating characteristic (ROC) curves, or AUC [S21]. AUC is the integral of ROC curves which ranges from 0.5 representing total lack of prediction to 1.0 representing perfect prediction. AUC measures the predicted outcomes that are numeric or probabilistic values between 0 and 1. Because the classification tree gives categorical predictions or training frequencies that are non-accurate conditional probability estimates, the performance of classification tree was not evaluated using AUC. Because AUC is robust against the prevalence of each colour type, we consider it as an overall measurement of model performance.

To access the contribution of each SNP to the predictive accuracy, we performed a step-wise analysis by iteratively excluding one SNP from the models. For each

iteration, the lowest contributor in the model-building set was excluded; a model was then rebuilt; and subsequently used to re-predict colours in the model verification set. The contribution of each SNP was measured by the AUC loss of the models with and without that SNP.

Model building and verification procedures were programmed in MATLAB version 7.6.0 (The MathWorks, Inc., Natick, MA).

Supplemental References

S1. Hofman, A., Grobbee, D.E., de Jong, P.T., and van den Ouweland, F.A. (1991). Determinants of disease and disability in the elderly: the Rotterdam Elderly Study. Eur J Epidemiol *7*, 403-422.

S2. Hofman, A., Breteler, M.M., van Duijn, C.M., Krestin, G.P., Pols, H.A., Stricker, B.H., Tiemeier, H., Uitterlinden, A.G., Vingerling, J.R., and Witteman, J.C. (2007). The Rotterdam Study: objectives and design update. Eur J Epidemiol *22*, 819-829.

S3. Kayser, M., Liu, F., Janssens, A.C., Rivadeneira, F., Lao, O., van Duijn, K., Vermeulen, M., Arp, P., Jhamai, M.M., van Ijcken, W.F., den Dunnen, J.T., Heath, S., Zelenika, D., Despriet, D.D., Klaver, C.C., Vingerling, J.R., de Jong, P.T., Hofman, A., Aulchenko, Y.S., Uitterlinden, A.G., Oostra, B.A., and van Duijn, C.M. (2008). Three genome-wide association studies and a linkage analysis identify HERC2 as a human iris color gene. Am J Hum Genet *82*, 411-423.

S4. Richards, J.B., Rivadeneira, F., Inouye, M., Pastinen, T.M., Soranzo, N., Wilson, S.G., Andrew, T., Falchi, M., Gwilliam, R., Ahmadi, K.R., Valdes, A.M., Arp, P., Whittaker, P., Verlaan, D.J., Jhamai, M., Kumanduri, V., Moorhouse, M., van Meurs, J.B., Hofman, A., Pols, H.A., Hart, D., Zhai, G., Kato, B.S., Mullin, B.H., Zhang, F., Deloukas, P., Uitterlinden, A.G., and Spector, T.D. (2008). Bone mineral density, osteoporosis, and osteoporotic fractures: a genome-wide association study. Lancet *371*, 1505-1512.

S5. Sulem, P., Gudbjartsson, D.F., Stacey, S.N., Helgason, A., Rafnar, T., Jakobsdottir, M., Steinberg, S., Gudjonsson, S.A., Palsson, A., Thorleifsson, G., Palsson, S., Sigurgeirsson, B., Thorisdottir, K., Ragnarsson, R., Benediktsdottir, K.R., Aben, K.K., Vermeulen, S.H., Goldstein, A.M., Tucker, M.A., Kiemeney, L.A., Olafsson, J.H., Gulcher, J., Kong, A., Thorsteinsdottir, U., and Stefansson, K. (2008). Two newly identified genetic determinants of pigmentation in Europeans. Nat Genet *40*, 835-837.

S6. Han, J., Kraft, P., Nan, H., Guo, Q., Chen, C., Qureshi, A., Hankinson, S.E., Hu, F.B., Duffy, D.L., Zhao, Z.Z., Martin, N.G., Montgomery, G.W., Hayward, N.K., Thomas, G., Hoover, R.N., Chanock, S., and Hunter, D.J. (2008). A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation. PLoS Genet *4*, e1000074.

S7. Sturm, R.A., Duffy, D.L., Zhao, Z.Z., Leite, F.P., Stark, M.S., Hayward, N.K., Martin, N.G., and Montgomery, G.W. (2008). A single SNP in an evolutionary conserved region within intron 86 of the HERC2 gene determines human blue-brown eye color. Am J Hum Genet *82*, 424-431.

S8. Eiberg, H., Troelsen, J., Nielsen, M., Mikkelsen, A., Mengel-From, J., Kjaer, K.W., and Hansen, L. (2008). 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 *123*, 177-187.

S9. Sulem, P., Gudbjartsson, D.F., Stacey, S.N., Helgason, A., Rafnar, T., Magnusson, K.P., Manolescu, A., Karason, A., Palsson, A., Thorleifsson, G., Jakobsdottir, M., Steinberg, S., Palsson, S., Jonasson, F., Sigurgeirsson, B.,

Thorisdottir, K., Ragnarsson, R., Benediktsdottir, K.R., Aben, K.K., Kiemeney, L.A., Olafsson, J.H., Gulcher, J., Kong, A., Thorsteinsdottir, U., and Stefansson, K. (2007). Genetic determinants of hair, eye and skin pigmentation in Europeans. Nat Genet *39*, 1443-1452.

S10. Kanetsky, P.A., Swoyer, J., Panossian, S., Holmes, R., Guerry, D., and Rebbeck, T.R. (2002). A polymorphism in the agouti signaling protein gene is associated with human pigmentation. Am J Hum Genet *70*, 770-775.

S11. Duffy, D.L., Montgomery, G.W., Chen, W., Zhao, Z.Z., Le, L., James, M.R., Hayward, N.K., Martin, N.G., and Sturm, R.A. (2007). A three-single-nucleotide polymorphism haplotype in intron 1 of OCA2 explains most human eye-color variation. Am J Hum Genet *80*, 241-252.

S12. Graf, J., Hodgson, R., and van Daal, A. (2005). Single nucleotide polymorphisms in the MATP gene are associated with normal human pigmentation variation. Hum Mutat *25*, 278-284.

S13. Frudakis, T., Thomas, M., Gaskin, Z., Venkateswarlu, K., Chandra, K.S., Ginjupalli, S., Gunturi, S., Natrajan, S., Ponnuswamy, V.K., and Ponnuswamy, K.N. (2003). Sequences associated with human iris pigmentation. Genetics *165*, 2071-2083.

S14. Barrett, J.C., Fry, B., Maller, J., and Daly, M.J. (2005). Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics *21*, 263-265.

S15. Bezdek, J.C. (1981). Pattern Recognition with Fuzzy Objective Function Algorithms (New York: Kluwer Academic Publishers).

S16. Ritchie, M.D., White, B.C., Parker, J.S., Hahn, L.W., and Moore, J.H. (2003). Optimization of neural network architecture using genetic programming improves detection and modeling of gene-gene interactions in studies of human diseases. BMC Bioinformatics *4*, 28.

S17. Moore, J.H. (2003). The ubiquitous nature of epistasis in determining susceptibility to common human diseases. Hum Hered *56*, 73-82.

S18. North, B.V., Curtis, D., Cassell, P.G., Hitman, G.A., and Sham, P.C. (2003). Assessing optimal neural network architecture for identifying disease-associated multi-marker genotypes using a permutation test, and application to calpain 10 polymorphisms associated with diabetes. Ann Hum Genet *67*, 348-356.

S19. Penco, S., Buscema, M., Patrosso, M.C., Marocchi, A., and Grossi, E. (2008). New application of intelligent agents in sporadic amyotrophic lateral sclerosis identifies unexpected specific genetic background. BMC Bioinformatics *9*, 254.

S20. Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984). Classification and regression trees (Monterey, Calif., U.S.A.: Wadsworth, Inc.).

S21. Janssens, A.C., Pardo, M.C., Steyerberg, E.W., and van Duijn, C.M. (2004). Revisiting the clinical validity of multiplex genetic testing in complex diseases. Am J Hum Genet *74*, 585-588; author reply 588-589.

Table S1. PCR and extension primer sequences from Sequenom SNP genotyping

iPLEX	SNP	2nd-PCRP	1st-PCRP	UEP_SEQ
1	rs3794604	ACGTTGGATGATGCCCTCCTGGCTTTGTG	ACGTTGGATGCACTTTTCTAGGGCTTTCAC	GCTTTGTGGCCTCTCAC
1	rs3935591	ACGTTGGATGACTGAGGTCCAGGTTCCTTG	ACGTTGGATGTGGCTTTCGTGGAGGAACAG	TCCTTGCTGGCTGAGCTA
1	rs4778232	ACGTTGGATGAACAGTTTCTTGCCCATGCC	ACGTTGGATGAAGAACCAAGGGATCTAGGG	CTGCCCTCTTCTTCAACAG
1	rs8041209	ACGTTGGATGAGAACTTGGTGGAGGATAGC	ACGTTGGATGTCTTAGAGACAAAATTCCC	TGGAGGATAGCCTACAGAT
1	rs1667394	ACGTTGGATGCCATTAAGACGCAGCAATTC	ACGTTGGATGGTCTTTTTCTCCTTTCAGTTC	AGCAATTCAAAACGTGCATA
1	rs16950987	ACGTTGGATGAATTACCCAGCATGCATGAC	ACGTTGGATGCTTGTTACTTTATCTTCCTC	tAGCATGCATGACTCATGAA
1	rs2346050	ACGTTGGATGGAGCCCAGCTGATTTTTCTC	ACGTTGGATGGGAATTCTTCCACTTAATG	tTGATGACTTAGGGTTGGTG
1	rs1800407	ACGTTGGATGACTCTGGCTTGTACTCTCTC	ACGTTGGATGATGATGATCATGGCCCACAC	cCCAGGCATACCGGCTCTCCC
1	rs1129038	ACGTTGGATGCTTCTCATCAGACACACCAG	ACGTTGGATGTCGTGAGATGAGAGCCTGAG	CTACAGTCTACACAGCAGCGAG
1	rs728405	ACGTTGGATGACCCCCATGGAAGAATGAGC	ACGTTGGATGACATAGGATGCGTGAGTGTG	gGGAAGAATGAGCCAAAAAAAA
1	rs2240202	ACGTTGGATGTGGCCTCTTACAGGACTTAG	ACGTTGGATGAGTCCTTTAAGCCCGGCTAC	aCTCTTACAGGACTTAGTAACCGC
1	rs12592730	ACGTTGGATGAGACAGAAAAGCTGCCAAG	ACGTTGGATGATTCTGCTGTTATTGGCTGG	tACTGGATCCAATCAAAATTTACA
1	rs7179994	ACGTTGGATGGGCTCTAACCATAGCATCTC	ACGTTGGATGCCAACAACCACACAGATGAG	gaaggGTTCAGCTGGAGCAAGGTC
1	rs7495174	ACGTTGGATGTAGGTCGGCTCCGTCGCAC	ACGTTGGATGGGCTTAGGAAGCAAGGCAAG	aTCCGTCGCACCCGTCTGTGCACACT
1	rs1448485	ACGTTGGATGAGCTTCAGCAAGAGCCTAAC	ACGTTGGATGCCCCACCATATTATTACCAG	CCATGGTTGTTATTAATACTCATCAA
1	rs7183877	ACGTTGGATGCTGTCTCATGGGTAGTAATC	ACGTTGGATGACACTTGAAGCAGTATACA	GGTAGTAATCAAAGAAACGACAAGTA
1	rs683	ACGTTGGATGCCTTCTTTCTAATACAAGC	ACGTTGGATGTTCTGAAAGGGTCTTCCCAG	CTTCTTTCTAATACAAGCATATGTTAG
2	rs8028689	ACGTTGGATGTTGTGCTGCTACTCATCTCC	ACGTTGGATGAGTGCTAGCAATGCTAGGTC	CTCAGTGTTCCACTTCC
2	rs12593929	ACGTTGGATGAGGACACCTGCCAGGACTAC	ACGTTGGATGGAAGCACCTGAGAGTGTCTG	GGGCCCACCTGCCACACG
2	rs16891982	ACGTTGGATGTCTACGAAAGAGGAGTCGAG	ACGTTGGATGAAAGTGAGGAAAACACGGAG	GGTTGGATGTTGGGGCTT
2	rs4778138	ACGTTGGATGCCTCCCATCACTGATTTAGC	ACGTTGGATGGAAAGTCTCAAGGGAAATCAG	CTGATTTAGCTGTGTTCTG
2	rs12896399	ACGTTGGATGGATGAGGAAGGTTAATCTGC	ACGTTGGATGTCTGGCGATCCAATTCTTTG	tgTCTGCTGTGACAAAGAGA
2	rs4778241	ACGTTGGATGAGGAGTGCAATTGTTGGCTG	ACGTTGGATGTGTACAGCCACTCTGGAAAG	aggGGCTGGTAGTTGCAATT
2	rs916977	ACGTTGGATGTTCTGTTCTTCTTGACCCCG	ACGTTGGATGGGTGTGGGATTTGTTTTGGC	ttCAGCCTTGGCCAGCCTTCT
2	rs12913832	ACGTTGGATGCGAGGCCAGTTTCATTTGAG	ACGTTGGATGAAAACAAAGAGAAGCCTCGG	CCAGTTTCATTTGAGCATTAA
2	rs8024968	ACGTTGGATGCAGGGAGAGTACAGATTCAC	ACGTTGGATGTTGGTGCCTTAGATGGACTG	GAGAGTACAGATTCACAGACTT
2	rs16950979	ACGTTGGATGGCTCTGCTGCTCTTCTTCCA	ACGTTGGATGAGGAAGCAGACGATAAGGAG	gtttaCTCTTCTTCCAGCTCTTC
2	rs2240203	ACGTTGGATGTCTATATTAGCCTCATCAG	ACGTTGGATGGAAGATCTTGCTTCCAAAGG	TGTCTTAATGTTTACATTCCTTA
2	rs2594935	ACGTTGGATGGCCACACAACTTGGATCTTC	ACGTTGGATGCCACAGGAAAACCTGCAATG	TGGATCTTCTTGTAGCAAGTAAC
2	rs1597196	ACGTTGGATGAACTCTCCGTGCCTTCCTCC	ACGTTGGATGGCATGAGTTCACGTGTATGA	ccCAGGCTCTGGAACCTGCAATTT
2	rs1393350	ACGTTGGATGGGAAGGTGAATGATAACACG	ACGTTGGATGTACTCTTCCTCAGTCCCTTC	ggtgGTAAAAGACCACACAGATTT
2	rs26722	ACGTTGGATGGATGGAATGTACGAGTATGG	ACGTTGGATGTTTTTGCTCCCTGCATTGCC	gggagTGTACGAGTATGGTTCTATC
2	rs7170852	ACGTTGGATGATTTGTAGCAGCTGTGCGTC	ACGTTGGATGACCAGGCCTTCTCTTTCATC	TTTGTAGCAGCTGTGCGTCTGTTTCC
2	rs1635168	ACGTTGGATGAATCTCAGAGATCTTACCCG	ACGTTGGATGACTTTGCCTGAGCACACAAG	cctccCAGAGATCTTACCCGTACCTGA

Table S2. 37 SNPs with significant iris color association as ascertained from previous studies with details from the previous studies and the present one

					Previous studies		Rotterdam Study			
SNP-ID	**Chr**	**Position**	**Gene**	**Allele**	**Reference**[1)]	**P-value**	**N**	**CR**	**MA**	**MAF**
rs16891982	5	33987450	SLC45A2(MATP)	CG	[S6]	5.0E-03	6420	0.99	C	0.03
rs26722	5	33999627	SLC45A2(MATP)	CT	[S6]	2.0E-03	6428	0.99	T	0.01
rs12203592	6	341321	IRF4	CT	[S6]	6.1E-13	5971[2)]	1.00	T	0.08
rs1408799	9	12662097	TYRP1	CT	[S5]	1.5E-09	5964[2)]	1.00	T	0.17
rs683	9	12699305	TYRP1	AC	[S13]	<0.01	6367	0.98	C	0.32
rs1393350	11	88650694	TYR	AG	[S9]	3.3E-12	6410	0.99	A	0.23
rs12896399	14	91843416	SLC24A4	GT	[S5,S6,S9]	4.1E-38	6409	0.99	G	0.50
rs2594935	15	25858633	OCA2	AG	[S3]	1.5E-10	6417	0.99	A	0.25
rs728405	15	25873448	OCA2	AC	[S3]	3.8E-09	6308	0.98	C	0.18
rs1800407	15	25903913	OCA2	CT	[S7]	5.0E-10	6219	0.97	T	0.04
rs3794604	15	25945660	OCA2	CT	[S3]	8.5E-12	6418	0.99	T	0.11
rs4778232	15	25955360	OCA2	CT	[S3]	2.5E-13	6411	0.99	T	0.22
rs1448485	15	25956336	OCA2	GT	[S3]	3.4E-08	6392	0.99	T	0.13
rs8024968	15	25957284	OCA2	CT	[S3]	1.5E-11	6430	0.99	T	0.10
rs1597196	15	25968517	OCA2	GT	[S3]	9.1E-18	6387	0.99	T	0.18
rs7179994	15	25997365	OCA2	AG	[S3]	5.4E-13	6417	0.99	G	0.14
rs4778138	15	26009415	OCA2	AG	[S3,S11]	5.4E-221	6421	0.99	G	0.12
rs4778241	15	26012308	OCA2	AC	[S3,S8,S11]	2.8E-267	6426	0.99	A	0.15
rs7495174	15	26017833	OCA2	AG	[S3,S5,S9, S11]	1.4E-239	6407	0.99	G	0.06
rs1129038	15	26030454	HERC2	CT	[S7,S8]	6.1E-46	6412	0.99	C	0.18
rs12593929	15	26032853	HERC2	AG	[S8]	---[3)]	6427	0.99	G	0.06
rs12913832	15	26039213	HERC2	AG	[S6-S8]	6.1E-46	6420	0.99	A	0.18
rs7183877	15	26039328	HERC2	AC	[S3]	6.2E-11	6407	0.99	A	0.05
rs3935591	15	26047607	HERC2	CT	[S8]	1.5E-25	6413	0.99	T	0.11
rs7170852	15	26101581	HERC2	AT	[S8]	1.1E-17	6421	0.99	T	0.13
rs8041209	15	26117253	HERC2	GT	[S3]	6.6E-22	6415	0.99	T	0.05
rs8028689	15	26162483	HERC2	CT	[S3]	1.2E-21	6426	0.99	C	0.05
rs2240203	15	26167797	HERC2	CT	[S8]	8.9E-17	6424	0.99	C	0.05
rs2240202	15	26184490	HERC2	AG	[S3]	2.2E-22	6412	0.99	A	0.05
rs916977	15	26186959	HERC2	CT	[S3,S7,S8]	<1E-300	6420	0.99	T	0.12
rs16950979	15	26194101	HERC2	AG	[S3]	7.0E-11	6394	0.99	G	0.05
rs2346050	15	26196279	HERC2	CT	[S3]	6.3E-19	6413	0.99	C	0.05
rs16950987	15	26199823	HERC2	AG	[S3]	8.3E-11	6414	0.99	A	0.05
rs1667394	15	26203777	HERC2	CT	[S3,S5,S7, S9]	8.5E-31	6405	0.99	C	0.13
rs12592730	15	26203954	HERC2	AG	[S3]	2.6E-22	6409	0.99	A	0.05
rs1635168	15	26208861	HERC2	AC	[S3]	1.5E-11	6397	0.99	A	0.06
rs6058017	20	32320659	ASIP	AG	[S10,S13]	2.2E-03	6186	0.97	G	0.12

P-value for eye color association obtained from the largest previous study in case included in several studies; CR: call rate in the current study; MA: minor allele; MAF: minor allele frequency; [1)] see Supplemental Reference list, [2)] data from Infinium II HumanHap550K Genotyping arrays, [3)] in haplotype association with eye color

Table S3. Single-SNP association with human iris color variation from the Rotterdam Study with and without adjustment for the largest effect contributed by *HERC2* rs12913832, Tagging SNP selection and priority rank in prediction analysis

SNP	Gene	Chr	Position	minor	beta1	P1	beta2	P2	Tag	Rank
rs16891982	SLC45A2	5	33987450	C	0.45	**1.1E-30**	0.08	**3.7E-03**	1	4
rs26722	SLC45A2	5	33999627	T	0.32	**4.6E-06**	0.13	**4.1E-03**	1	
rs12203592	IRF4	6	341321	T	-0.07	**7.5E-03**	-0.07	**2.9E-05**	1	6
rs1408799	TYRP1	9	12662097	T	0.05	**3.3E-03**	0.05	**5.3E-05**	1	12
rs683	TYRP1	9	12699305	C	0.07	**5.6E-06**	0.03	**3.3E-03**	1	15
rs1393350	TYR	11	88650694	A	-0.05	**8.8E-03**	-0.05	**3.8E-06**	1	5
rs12896399	SLC24A4	14	91843416	G	0.09	**1.2E-08**	0.08	**6.5E-14**	1	3
rs2594935	OCA2	15	25858633	A	0.21	**1.1E-34**	-0.06	**2.1E-06**	1	
rs728405	OCA2	15	25873448	C	0.27	**1.2E-42**	-0.07	**4.1E-08**	1	
rs1800407	OCA2	15	25903913	T	0.27	**7.7E-13**	-0.29	**1.7E-28**	1	2
rs3794604	OCA2	15	25945660	T	0.40	**2.5E-60**	0.02	1.4E-01	0	
rs4778232	OCA2	15	25955360	T	0.30	**2.9E-62**	-0.01	6.8E-01	1	11
rs1448485	OCA2	15	25956336	T	0.39	**2.5E-68**	0.01	6.6E-01	1	
rs8024968	OCA2	15	25957284	T	0.45	**6.3E-74**	0.03	1.3E-01	1	13
rs1597196	OCA2	15	25968517	T	0.33	**5.4E-63**	0.01	5.1E-01	1	
rs7179994	OCA2	15	25997365	G	0.31	**2.0E-45**	-0.01	3.7E-01	1	
rs4778138	OCA2	15	26009415	G	0.73	**4.7E-239**	0.07	**4.8E-05**	1	
rs4778241	OCA2	15	26012308	A	0.75	**<1.0E-300**	-0.04	**3.6E-02**	1	
rs7495174	OCA2	15	26017833	G	1.05	**4.7E-274**	0.13	**1.5E-07**	1	8
rs1129038	HERC2	15	26030454	C	1.12	**<1.0E-300**	-0.03	8.6E-01	0	
rs12593929	HERC2	15	26032853	G	1.07	**9.1E-265**	0.11	**1.8E-05**	0	
rs12913832	HERC2	15	26039213	A	1.13	**<1.0E-300**	1.13	**<1.0E-300**	1	1
rs7183877	HERC2	15	26039328	A	0.89	**5.7E-166**	-0.15	**9.0E-09**	1	10
rs3935591	HERC2	15	26047607	T	1.03	**<1.0E-300**	-0.04	7.6E-02	1	
rs7170852	HERC2	15	26101581	T	0.92	**<1.0E-300**	-0.01	7.2E-01	0	
rs8041209	HERC2	15	26117253	T	1.03	**2.6E-226**	0.09	**5.7E-04**	0	
rs8028689	HERC2	15	26162483	C	1.06	**4.7E-236**	0.09	**7.0E-04**	0	
rs2240203	HERC2	15	26167797	C	1.04	**2.0E-230**	0.09	**9.1E-04**	0	
rs2240202	HERC2	15	26184490	A	1.03	**3.7E-221**	0.09	**8.6E-04**	0	
rs916977	HERC2	15	26186959	T	1.05	**<1.0E-300**	-0.02	5.5E-01	0	
rs16950979	HERC2	15	26194101	G	1.05	**2.8E-227**	0.09	**3.9E-04**	0	
rs2346050	HERC2	15	26196279	C	1.04	**2.0E-229**	0.08	**1.1E-03**	0	
rs16950987	HERC2	15	26199823	A	1.05	**2.1E-238**	0.09	**6.8E-04**	0	
rs1667394	HERC2	15	26203777	C	1.06	**<1.0E-300**	0.02	4.7E-01	1	9
rs12592730	HERC2	15	26203954	A	1.05	**5.3E-223**	0.09	**5.1E-04**	1	7
rs1635168	HERC2	15	26208861	A	1.03	**3.0E-258**	0.09	**2.0E-04**	0	
rs6058017	ASIP	20	32320659	G	-0.01	7.9E-01	-0.02	2.7E-01	1	14

beta1, P1: betas and P-values derived from single SNP association tests unadjusted for rs12913832; beta2, P2: betas and P-values derived from single SNP association tests adjusted for rs12913832; P values smaller than 0.05 are indicated in bold; Tag: tagging SNPs were selected based on pair-wise r2 <0.8; Rank: 15 SNPs are ranked according to their contribution to eye color prediction when all 24 tagging SNPs were included in a multinomial logistic regression model, the smallest number represents highest prediction value, the 9 SNPs without number code did not contribute to the prediction accuracy, see main text and Figure 1.

Table S4. Performances of four alternative models for DNA-based prediction of human iris color using 24 associated single nucleotide polymorphisms in Dutch Europeans of the Rotterdam Study*

Model	Measure	Blue	Intermediate	Brown
Neural Network	Sensitivity	92.9	0	91.7
	Specificity	79.4	100.0	87.0
	PPV	90.6	---[1]	66.3
	NPV	83.9	90.0	97.4
	AUC	0.89	0.65	0.91
Fuzzy C-Means Clustering	Sensitivity	93.0	0	85.2
	Specificity	75.8	100.0	86.8
	PPV	89.2	---[1]	64.3
	NPV	83.6	90.0	95.5
	AUC	0.91	0.67	0.93
Ordinal Regression	Sensitivity	93.5	0	88.5
	Specificity	77.0	100.0	87.7
	PPV	89.7	---[1]	66.6
	NPV	84.7	90.0	96.5
	AUC	0.91	0.73	0.93
Classification Tree	Sensitivity	91.5	13.0	74.9
	Specificity	75.3	95.0	90.3
	PPV	88.8	22.4	68.3
	NPV	80.6	90.7	92.8
	AUC	---[2]	---[2]	---[2]

*For results of the multinomial logistic regression model, see Table 1. AUC: Area Under the receiver operating characteristic (ROC) Curves, PPV: Positive Predictive Value, NPV: Negative Predictive Value, [1]zero denominator, [2]categorical outcomes were not measured by AUC.

Figure S1. Haplotype blocks in the *HERC2-OCA2* region in the Rotterdam Study cohort. SNPs were aligned according to chromosomal positions (NCBI B36 assembly). The triangles surrounded by solid black lines are suggested haplotype blocks. The redness of the colour represents pair-wise D'/LOD and the values are pair-wise r^2. The left bottom legend lists all inferred haplotypes with frequencies greater than 1% in the cohort and the multiallelic D' between blocks. A: all 28 ascertained SNPs associated with eye color in the chromosomal region. B: Sixteen tagging SNPs were selected after excluding the SNPs in strong LD using a threshold of pair-wise $r^2 = 0.8$.