Forensic Science International: Genetics 7 (2013) 444–452

Contents lists available at SciVerse ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

SNP model development for the prediction of eye colour in New Zealand

Julia S. Allwood [a,b,*], SallyAnn Harbison [a]

[a] Institute of Environmental Science and Research (ESR Ltd.), Mt Albert Science Centre, Private Bag 92-021, Auckland Mail Centre, Auckland 1142, New Zealand
[b] School of Chemical Sciences, The University of Auckland, Private Bag 92019, Auckland 1142, New Zealand

ARTICLE INFO

Article history:
Received 24 May 2012
Received in revised form 28 February 2013
Accepted 12 March 2013

Keywords:
Eye colour
SNPs
Classification tree
Prediction
Forensic

ABSTRACT

The ability to predict externally visible characteristics (EVCs) from DNA has appeal for use in forensic science, particularly where a forensic database match is not made and an eye witness account is unavailable. This technology has yet to be implemented in casework in New Zealand. The broad cultural diversity and likely population stratification within New Zealand dictates that any EVC predictions made using anonymous DNA must perform accurately in the absence of knowledge of the donor's ancestral background. Here we construct classification tree models with SNPs of known association with eye colour phenotypes in three categories, blue vs. non-blue, brown vs. non-brown and intermediate vs. non-intermediate. A set of nineteen SNPs from ten different known or suspected pigmentation genes were selected from the literature. A training dataset of 101 unrelated individuals from the New Zealand population and representing different ancestral backgrounds were used. We constructed four alternate models capable of predicting eye colour from the DNA genotypes of SNPs located within the HERC2, OCA2, TYR and SLC24A4 genes using probability calculation and classification trees. The final model selected for eye colour prediction exhibited high levels of accuracy for both blue (89%) and brown eye colour (94%). Models were further assessed with a test set of 25 'blind' samples where phenotype was unknown, with blue and brown eye colour predicted correctly where model thresholds were met. Classification trees offer an aesthetically simple and comprehendible model to predict blue and brown eye colour.

1. Introduction

Predicting phenotypes from SNP data is an advancing field, both in medical and consumer genomics, and more recently forensic biology [1,2]. The emergence of genome-wide association studies (GWAS) and the HapMap Project [3] resulted in vast amounts of SNP data being available, with the majority of SNPs located in the intron regions of the genome. Many different phenotypes have since been investigated for SNP association (e.g. height [4,5]; eye colour [6,7] and many medical conditions [8,9]), often with the intent of utilising the SNPs contributing the greatest effect for eventual prediction of phenotype. Developing the ability to use SNPs to predict EVCs of humans has obvious appeal to forensic biology. The capacity to predict aspects of an individual's appearance using EVC SNPs from DNA of an unknown individual could provide information comparable to that of an eye witness account. This technology would be particularly useful in instances where classic DNA profiling failed to generate a database match from crime scene DNA.

Eye colour is a highly heritable, complex, genetic trait that exhibits a variety of alternate phenotype states [10,11]. Eye colouration is the product of melanin stored within melanocyte cells of the iris, with variations in colour primarily being a product of melanin type [12]. Melanogenesis is a complex process responsible for the majority of mammalian pigmentation, including hair, skin, and eye colour in humans. The genes and biochemical processes involved in the synthesis of melanin are numerous and have been identified through different means. The role of the OCA2 and MC1R genes in pigmentation was revealed through murine phenotype studies, and in the case of OCA2, its causative effect in oculocutaneous albinism (OCA2 [13,14]; MC1R [15,16]). Early linkage analysis also demonstrated a strong association between iris colouration and two loci on chromosome

* Corresponding author at: Institute of Environmental Science and Research (ESR Ltd.), Mt Albert Science Centre, Private Bag 92-021, Auckland Mail Centre, Auckland 1142, New Zealand. Tel.: +64 9 815 3910; fax: +64 9 849 6046.
E-mail addresses: julia_a_713@hotmail.com, Julia.allwood@esr.cri.nz (J.S. Allwood), Sallyann.harbison@esr.cri.nz (S. Harbison).

1872-4973/$ – see front matter © 2013 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.fsigen.2013.03.005

15 [17]. The majority of the remaining pigmentation genes, including those associated with eye colour, were discovered in GWAS. GWAS identified strong associations between eye colour and the SNPs of the OCA2 and the upstream HERC2 genes [18,19]. Further fine mapping GWAS investigations of these regions revealed that a single HERC2 SNP (rs12913832) was responsible for the majority of effect towards eye colour being blue or brown [19–22]. Other significantly associated pigmentation genes and SNPs of lesser effect have since been identified, a selection of which include: TYR, SLC45A2, SLC24A5, SLC24A4, MYO5A, TYRP1 and IRF4 [18,23–29].

The original ancestral state of human eye colour is thought to be brown, with blue and other lighter colour variants having evolved over time. Individuals of European ancestry show the most variation in pigmentation traits, including eye colour [10,30,31]. SNPs associated with eye colour (both exonic and intronic) are likely to have variable frequencies amongst different individuals of alternate ancestral backgrounds in response to forces acting upon the gene pool neutrally (e.g. genetic drift) or otherwise (e.g. selection) [32]. Therefore in order for SNP association with eye colour to be applied to individuals of a range of ancestral backgrounds, any association should ideally be with the trait in question, here eye colour, rather than influenced by underlying genetic substructure. This is of particular importance if eye colour associated SNPs were to be applied for forensic use in New Zealand. New Zealand is a culturally diverse country, with individuals of European descent forming the majority of the population (~68%), the remainder of the population is made up of a collection of diverse minority groups that include (but is not limited to) Māori (New Zealand's indigenous people) (~15%), Pacific Islanders (Polynesians) (~6.5%), and multiple Asian ethnicities (including Indian) (~9%) [33]. Such cultural diversity is likely to suggest population stratification across individuals within the New Zealand population, therefore a model for forensic use on anonymous DNA in New Zealand requires the capacity to accurately predict eye colour, regardless of ancestral origins and subsequent genetic substructure.

This is the first study of its kind in New Zealand therefore a selection of 19 SNPs found to be associated with eye colour in the literature were used to create predictive models. These SNPs were utilised in an attempt to build interactive classification tree models across three eye colour categories: blue, brown and intermediate. We have investigated four alternate models capable of predicting eye colour in New Zealand. The final model recommended for use predicts blue and brown eye colour with a high degree of accuracy on an exemplar subset of a New Zealand population.

2. Materials and methods

2.1. Sample collection and DNA extraction

Study participants were recruited and enrolled in the study in accordance with the University of Auckland Human Participants ethics committee stipulations. The observed eye colour frequencies of the training set of unrelated individuals resulted in the following phenotype groups: brown ~31% (*N*: 31/101); intermediate (green and hazel) ~30% (*N*: 30/101) and blue ~40% (*N*: 40/101). Phenotype classification was performed by visual assessment of all study participants by a single observer. DNA samples from all participants were taken via duplicate buccal swabs and extracted using a phenol–chloroform method followed by ethanol precipitation (adapted from [34]) followed by centrifugation for DNA pellet formation, ethanol removal and pellet rehydration in ddH_2O. Blind testing samples used to test model performance in a replicate forensic setting were classified in the same manner as above by the observer and later recoded by an external party to ensure sample anonymity to the original observer and subsequent analyst. The blind testing test set comprised 25 individuals that were not used within the training set. This test set resulted in the following phenotype groups: brown 32% (*N*: 8/25); intermediate (green and hazel) 28% (*N*: 7/25) and blue 40% (*N*: 10/25). The blind testing test set was applied to the three best performing models. Genotypes for the SNPs required for each model were discerned using the same methods described below. Following genotyping and eye colour probability calculation, phenotype details of each sample were revealed and probabilities were assessed against phenotype observations to assess prediction accuracy.

2.2. Genotype screening

A total of 19 SNPs from ten different genes were included: HERC2 gene – rs12913832, rs1129038, rs916977; OCA2 gene – rs7495174, rs4778138, rs1800401, rs1800407; MC1R gene – rs2228479, rs1805007, rs1805008; SLC45A2 gene – rs26722, rs16891982; SLC24A5 gene – rs1426654; SLC24A4 gene (not located within SLC24A4, flanking position) – rs12896399; MYO5A gene – rs1724630; TYR gene – rs1393350, rs1042602; TYRP1 gene – rs2733832 and IRF4 gene – rs12203592 (Table 1). This SNP selection includes all of the SNPs employed in the validated IrisPlex [32] system, plus many highly ranked SNPs found in the notable eye colour association study of Liu et al. [6]. Primers were designed to amplify and sequence the SNP region from the 101 participants in the training set to obtain genotypes (Table 1). This same procedure was followed for blind testing samples for those SNPs relevant to the models produced using the training set. Briefly, SNP regions were amplified in a standard PCR reaction with 2.5 μl buffer, 2.5 μl dNTPs (2 mM), forward and reverse primers at a final concentration of 0.2 μM, 0.8ul $MgCl_2$, ddH_2O, and 1 unit of Platinum®*Taq* polymerase (Invitrogen) using a Mastercycler® EP Gradient S (Eppendorf) thermocycler. Thermocycling conditions were as follows: denaturing at 94 °C for 2 min, 30 s for annealing, extension for 30 s to 1 min at 72 °C (depending on amplicon size) for 28 cycles, followed by a final extension at 72 °C for 7 min. Amplified fragments were visualised on 2% E-Gels (E-Gel® EX system (Invitrogen)) and purified using enzyme digestion with Exonuclease I and Antarctic shrimp phosphatase (New England Biolabs). Purified amplicons were sequenced using BigDye™ Terminator v. 3.1 (Applied Biosystems), purified using Agencourt®CleanSEQ® (Agencourt Bioscience) following manufacturer's instructions and analysed via capillary electrophoresis using Applied Biosystems 3130-Avant Automated Sequencer (Applied Biosystems). SNP genotypes were determined from sequence data in Geneious Pro 4.7.6 [35].

2.3. Statistical analysis

Classification trees and partition modelling were used to create alternate models for eye colour prediction. Partition modelling is an analytical method used to reduce the dimensionality of detailed data resulting in the association of predictor variables to response variables [36].

SNP genotypes were used as predictors. A series of binary questions were asked of each observation, grouping the responses based on their similarity. Each observation was channelled to a final predictive class based on the responses to these questions. The final SNPs chosen for use in the classification trees were determined on the strength of their capacity to segregate the data into homogenous nodes (groups), referred to as a predictor's 'goodness-of-split' [36,37]. The rest were discarded by the model. The final groups formed in the tree, known as 'terminal nodes' [38] classify the observational data with an accompanying probability of likely class inclusion [36].