Magdalena Kukla-Bartoszek

Doctoral dissertation title: **Genetic prediction of human pigmentation characteristics – implementation of the HIrisPlex-S predictive tool and searching for improvements**

Abstract

The advent of the genome-wide DNA analysis technologies has made it possible to conduct research into the genetics of complex traits including human appearance traits. In particular, Genome Wide Association Studies (GWAS) have been very successful in identifying genes and DNA variants associated with specific phenotypes. A better understanding of the genetic architecture of appearance traits has spawned a new research area in forensic genetics known as forensic DNA phenotyping (FDP). The purpose of the FDP is to predict human appearance from DNA and, by describing physical characteristics, narrow down the number of potential suspects in criminal investigations as well as to provide useful information in the study of human remains.

The first aim of this dissertation was implementation, validation, and forensic application of the HIrisPlex-S tool which was developed for the purpose of DNA-based prediction of eye, hair and skin colour. The second aim was to identify weaknesses and search for improvements of this pigmentation predictive method.

The study on the implementation and validation of HIrisPlex-S, conducted within the international collaboration, led to the development of a new SNaPshot-based multiplex assay which complements the previously developed 24-plex used in the HIrisPlex tool. The HIriSplex-S assays provide comprehensive genetic tool for collection of SNP data used in the next step in predictive algorithms. The assay proved high sensitivity allowing full DNA profiles to be obtained from as little as 63 pg of DNA and met all the requirements for its application to forensic casework. The validation studies, in which I was mostly involved, revealed high concordance of the genotypes obtained using the newly developed assay, between the 5 research laboratories.

Although the newly developed DNA test proved to be very robust, the SNaPshot technology has low capacity. Therefore, our further research aimed to transit the HIrisPlex-S data collection method to the Massively Parallel Sequencing (MPS) standard offered by the two popular MPS technologies, Ion Torrent and Illumina. My research helped in the development of the two genetic tests for both technologies combining all 41 HIrisPlex-S markers in a single assay. The international validation I participated in showed

differences in the sensitivity of these tests, showing the superiority of the Ion AmpliSeq technology-based test analysed with Ion Torrent technology. In addition, an integrated bioinformatics analysis pipeline was provided as a part of the comprehensive workflow, allowing semi-automated data analysis, to make predictive analysis more straightforward. Furthermore, an innovative 2-person mixture separation tool was proposed, providing genotype reliability assessment and the most probable mixture scenario. Although further validation testing was recommended, the study has revealed good performance of the developed DNA tests and their potential applicability in real casework.

In the next step, I evaluated a more sensitive AmpliSeq-based HIrisPlex-S method in a real identification study that involved 63 challenging bone samples. The analysis allowed to establish full DNA profiles for 35 (56%) samples, from as little as 49 pg, while in 5 cases (8%) no DNA profiles were generated. The study showed underperformance of 3 SNPs used in the algorithm for skin colour prediction. The results revealed that although degradation of DNA constitutes the most serious problem in DNA phenotyping of skeletal remains, AmpliSeq solution developed for the HIrisPlex-S can be successful in the study of highly degraded DNA samples. Notably, in most cases, predictive analysis could be successfully accomplished due to the ability of HIrisPlex-S to predict phenotypes from partial DNA profiles. Based on the obtained results, we also formulated some recommendations for challenging samples analysis with the use of HPS-MPS-ION workflow.

The identified problems with accuracy of predictions prompted me to conduct a search for improvements in predicting the pigmentation phenotype. At first, I evaluated the impact of age-dependent hair colour darkening observed in childhood on hair colour prediction accuracy. The main conclusion of the analysis conducted within a set of ~470 Polish children was that in most instances (71%) HIrisPlex predicts the pre-adolescent hair colour of individuals who experienced hair colour darkening early in life, although in some cases (28%) correct predictions of the darkened hair colour phenotype were also observed. Overall, the study provided evidence that DNA-based hair colour prediction is affected by age-dependent hair colour changes occurring in childhood. Understanding the underlying causes of this phenomenon is apparently of great importance for more accurate prediction of hair colour.

In order to improve the description of the pigmentation phenotype of an unknown DNA donor, I conducted a study to develop a predictive model for freckles. Ninety-seven SNP variants previously associated with pigmentation traits were subjected to the predictive modelling in 960 Polish individuals. As a result, a simplified 12-variable binomial model,

and extended 14-variable multinomial model were developed. They proved relatively good predictive performance and were able to correctly predict freckling status in 68%-79% of evaluated cases, depending on the model used. The developed predictive models for freckles improved DNA-based prediction of human pigmentation phenotype and showed complexity of the genetics of freckles, indicating the need for further studies.

Finally, I explored the possibility of improving the accuracy of eye colour prediction. In particular, by using various strategies, I identified 27 new candidate DNA markers and subjected them to predictive modelling in >800 Polish samples together with 114 known DNA variants associated with pigmentation. Very high prediction accuracies were revealed for blue and brown eye colours (AUC=0.93; 0.96, respectively), and increased for intermediate eye colour (AUC=0.85). Importantly, the study identified a new potential predictor of eye colour, rs2253104 in *ARFIP2* and confirmed the important role of rs74653330 in *OCA2*. Interestingly, none of the sophisticated machine learning approaches outperformed classic logistic regression, although some revealed increased sensitivity of intermediate eye colour prediction. The study provided additional knowledge on the genetics of human eye colour and suggested that probably the advanced statistical methods require high dimensional data to demonstrate their potential better.

Summarizing, my research allowed for development and validation of the most advanced method designed to predict human pigmentation traits from DNA. At the same time, my research allowed for a better understanding of the genetic architecture of the human eye and hair color and freckles occurrence, and the obtained results provided the basis for further improvement of prediction of human pigmentation phenotype.

Signed by /
Podpisano przez:

Wojciech Branicki

Date / Data: 2021-06-21 08:25

Magdalena Kulla-Jarosz

Magdalena Kukla-Bartoszek

Tytuł dysertacji: **Genetyczna predykcja cech pigmentacyjnych u człowieka – implementacja narzędzia HIrisPlex-S i poszukiwanie usprawnień**

Streszczenie

Pojawienie się metod umożliwiających analizę całego genomu pozwoliło na rozwój badań nad genetyczną determinacją cech wielogenowych, w tym cech wyglądu człowieka. W szczególności badania GWAS (ang. *Genome Wide Association Studies*) okazały się bardzo skuteczne w identyfikacji genów i wariantów DNA związanych z określonymi fenotypami. Lepsze zrozumienie architektury genetycznej cech wyglądu zaowocowało pojawieniem się nowego obszaru genetyki sądowej, zwanego kryminalistycznym fenotypowaniem DNA (ang. *Forensic DNA Phenotyping*, FDP). Celem FDP jest przewidywanie wyglądu człowieka na podstawie DNA i tym samym zawężanie kręgu osób potencjalnie podejrzanych w sprawie kryminalnej, a także dostarczenie przydatnych informacji w badaniu szczątków ludzkich.

Pierwszym celem niniejszej rozprawy doktorskiej było wdrożenie, walidacja oraz praktyczna aplikacja narzędzia HIrisPlex-S, które zostało opracowane do przewidywania koloru oczu, włosów i skóry na podstawie analizy DNA. Drugim celem było zidentyfikowanie słabych jego stron oraz poszukiwanie możliwości udoskonalenia.

Badania nad implementacją i walidacją narzędzia HIrisPlex-S, prowadzone w ramach współpracy międzynarodowej doprowadziły do opracowania nowego testu multiplexowego opartego na technologii SNaPshot, który uzupełnił wcześniej opracowany 24-plex, stosowany w narzędziu HIrisPlex. Testy HIrisPlex-S stanowią kompleksowe narzędzie genetyczne do gromadzenia danych SNP, wykorzystywanych na kolejnym etapie analizy predykcyjnej. Nowo opracowany test wykazał wysoką czułość, pozwalającą na uzyskanie pełnych profili DNA już z 63 pg DNA i spełnił wszystkie wymagania umożliwiając jego zastosowanie w sprawach kryminalnych. Badania walidacyjne, w które byłam głównie zaangażowana, wykazały wysoką zgodność genotypów uzyskanych za pomocą nowo opracowanego testu pomiędzy 5 laboratoriami badawczymi.

Mimo, że nowo opracowany test genetyczny okazał się bardzo wydajny, technologia SNaPshot, w której został on zaprojektowany, charakteryzuje się niską przepustowością. Dlatego też nasze dalsze badania skoncentrowały się na opracowaniu rozwiązań dostosowanych do dwóch najczęściej stosowanych technologii masowego równoległego sekwencjonowania DNA: Ion Torrent i Illumina. Moje badania pomogły w opracowaniu testów

genetycznych dla obu technologii, pokrywających wszystkie 41 markerów HIrisPlex-S w pojedynczym teście. Międzynarodowe badania walidacyjne, w których uczestniczyłam, wykazały różnice w czułości nowo opracowanych paneli genetycznych oraz przewagę na korzyść testu opracowanego w oparciu o metodę AmpliSeq i analizę z wykorzystaniem technologii Ion Torrent. Dodatkowo, opracowany został protokół bioinformatyczny, umożliwiający półautomatyczną analizę danych, w celu jej usprawnienia. Ponadto, zaproponowano innowacyjne narzędzie do separacji mieszanin DNA pochodzących od dwóch osób, zapewniające ocenę wiarygodności genotypu oraz najbardziej prawdopodobny scenariusz dotyczący składu mieszaniny. Chociaż dalsze testy walidacyjne są zalecane, badania wykazały optymalne działanie opracowanych testów genetycznych i ich potencjalne zastosowanie w sprawach kryminalnych.

W moich dalszych badaniach, oceniłam użyteczność charakteryzującego się wyższą czułością testu genetycznego HIrisPlex-S opartego na technologii Ion AmpliSeq, w badaniu identyfikacyjnym 63 próbek kostnych. Analiza pozwoliła na uzyskanie pełnych profili DNA w przypadku 35 (56%) próbek, o stężeniach ≥49 pg, podczas gdy w 5 przypadkach (8%) nie wygenerowano żadnych profili DNA. Badanie wykazało również niską wydajność amplifikacji 3 fragmentów DNA, zawierających 3 polimorfizmy pojedynczego nukleotydu (SNP), wykorzystywane przez algorytm do przewidywania koloru skóry. Wyniki ujawniły, że chociaż degradacja DNA stanowi najtrudniejszy problem w analizie predykcyjnej zdegradowanego materiału kostnego, zaproponowane rozwiązanie AmpliSeq opracowane dla HIrisPlex-S może być z powodzeniem stosowane nawet w silnie zdegradowanych próbkach DNA. Ponadto, w większości przypadków analiza predykcyjna może być z powodzeniem przeprowadzona ze względu na zdolność HIrisPlex-S do przewidywania fenotypów na podstawie częściowych profili DNA. Na podstawie uzyskanych wyników sformułowaliśmy również zalecenia dotyczące analizy trudnego materiału z wykorzystaniem protokołu HIrisPlex-S w technologii wysokoprzepustowego sekwencjonowania.

Zidentyfikowane problemy dotyczące dokładności predykcji skłoniły mnie do poszukiwania możliwości poprawy dokładności predykcji fenotypu pigmentacyjnego u człowieka. Na początku oceniłam wpływ zjawiska ciemnienia włosów, obserwowany w okresie dzieciństwa, na dokładność predykcji koloru włosów. Głównym wnioskiem płynącym z badania przeprowadzonego na grupie ok. 470 dzieci jest fakt, że w większości przypadków (71%) HIrisPlex przewiduje kolor włosów obserwowany we wczesnym dzieciństwie, u osób, które doświadczyły ciemnienia włosów z wiekiem, chociaż w niektórych

przypadkach (28%) zaobserwowano także prawidłowe predykcje ciemniejszego koloru włosów. Przeprowadzone badanie dostarczyło dowodów na to, że zmiany koloru włosów w dzieciństwie mają wpływ na poprawność przewidywania koloru włosów w późniejszym okresie życia w oparciu o DNA, a wyjaśnienie podstaw tego zjawiska ma duże znaczenie dla dokładniejszego przewidywania koloru włosów.

W celu udoskonalenia opisu fenotypu pigmentacyjnego nieznanego dawcy DNA, przeprowadziłam badania mające na celu opracowanie modelu predykcyjnego do przewidywania piegów. Dziewięćdziesiąt siedem polimorfizmów typu SNP, pierwotnie zasocjowanych z cechami pigmentacyjnymi poddałam modelowaniu predykcyjnemu u 960 osób z Polski. W rezultacie opracowałam dwa niezależne modele predykcyjne: dwumianowy, zawierający 12 zmiennych i rozszerzony, wielomianowy, zawierający 14 zmiennych. Modele wykazały relatywnie dobry poziom dokładności predykcji, były w stanie poprawnie przewidzieć fenotyp w 68-79% ocenianych przypadków, w zależności od zastosowanego modelu. Opracowane modele do predykcji piegów wzbogaciły analizę predykcyjną cech pigmentacyjnych i wykazały złożone genetyczne podstawy powstawania piegów, co wskazuje na potrzebę dalszych badań.

Na koniec zbadałam możliwość poprawy dokładności predykcji koloru oczu. Aby to osiągnąć zastosowałam kilka strategii, a w efekcie zidentyfikowałam 27 nowych markerów kandydackich, które poddałam modelowaniu predykcyjnemu w ponad 800 polskich próbkach wraz z 114 innymi, znanymi wariantami genetycznymi, zasocjowanymi z cechami pigmentacyjnymi. Wysoka dokładność predykcji odnotowana została w przypadku niebieskiego i brązowego koloru oczu (AUC=0.93; 0.96, odpowiednio), a także zwiększona w przypadku pośredniego koloru oczu (AUC=0.85). Co ważne, w badaniu zidentyfikowany został nowy potencjalny predyktor koloru oczu, rs2253104 w genie *ARFIP2* i potwierdzona istotna rola rs74653330 w genie *OCA2*. Co ciekawe, żadne z zaawansowanych metod uczenia maszynowego nie poprawiło wyników uzyskanych z zastosowaniem klasycznej regresji logistycznej, chociaż niektóre metody wykazały zwiększoną czułość przewidywania pośredniego koloru oczu. Przeprowadzone badanie dostarczyło dodatkowej wiedzy na temat genetycznych podstaw koloru oczu u człowieka a ponadto ujawniło, że prawdopodobnie zastosowane metody statystyczne wymagają danych wielkoskalowych, aby lepiej wykazać swój potencjał.

Podsumowując, moje badania pozwoliły na opracowanie i walidację najbardziej zaawansowanej metody umożliwiającej predykcję cech pigmentacyjnych człowieka na podstawie DNA. Jednocześnie badania te wzbogaciły wiedzę o genetycznej architekturze

koloru oczu i włosów a także piegów, a uzyskane przeze mnie wyniki dały obszar do dalszych badań nad predykcją cech pigmentacyjnych u człowieka.

Signed by /
Podpisano przez:

Wojciech Branicki

Date / Data: 2021-06-21 09:34

Magdalena Kulula-Bartosek