Inference of Biogeographical Ancestry and Pigmentation Phenotype using Single Nucleotide Polymorphisms

Charmain Vanessa Castel

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

Centre for Forensic Science

University of Technology, Sydney, NSW, Australia

2014

Certificate of Original Authorship

I certify that work in this thesis has not previously been submitted for a degree nor has it been submitted as part of the requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged, In addition, I certify that all information sources and literature used are indicated in the thesis.

Charmain Vanessa Castel
July 2014

Acknowledgements

First and foremost, I would like to sincerely thank my supervisor Associate Professor Anita Piper for her ongoing support and invaluable guidance throughout my candidature. I have thoroughly enjoyed working with you over the years and have learnt so much, not only about the 'science' but also myself and my capabilities. Thank you for helping me to grow into an independent researcher and for your patience and continuous encouragement.

Thank you to Prof. Claude Roux (Centre for Forensic Science) and the Faculty of Science at the University of Technology, Sydney (UTS) for their funding and support of this project. I would also like to thank the UTS Graduate School who have awarded me with an Australian Postgraduate Award and have provided funding support to attend national and international conferences.

Thank you to Dr Runa Daniel (Victoria Police Forensic Services Centre) for the advice she provided in the early stages of this project, particularly for showing me the ins and outs of the ABI 310. Special thanks also extend to Connie Akratos and Zofia Winiarski from Technical Services (UTS) for kindly providing me with some of the materials and equipment I needed to complete the assay validation phase of this project. I would also like to thank all of the volunteers who provided DNA samples and without whom this project would not have been possible.

Thank you to all my friends and colleagues at UTS, past and present. In particular, Dr Michael Johnson, Dr Marilyn Katrib, Dr Rob Walker, Dr Michael Lees, Dr Nicholas Archer and Prue Gatt for convincing me to do a PhD. Whilst this has at times been a tough journey, I am so glad I took the bait and could not have done it without your encouragement, guidance and wisdom. Thank you all so much for the fun times and memories that I will cherish for a lifetime. Most importantly, I would like to thank Dr Stephanie Dowdell for being my office buddy, for all the teas and invaluable words of wisdom, for your kindness and generosity and for all the guidance, moral support and encouragement that you have given me over the years. I am so lucky to have made a lifetime friend like you!

Thank you to my dearest and oldest friend Mohini Saini for 15 years of friendship, laughter and fun. Despite the distance that now separates us, you have not ceased to provide me with your ongoing support and encouragement all the way from the other side of the World and I truly appreciate all that you have done.

I would especially like to thank (fellow soon to be Dr) Michael Strauss for sharing this incredible journey with me and for being by my side from day one. Thank you for being an amazing and understanding person, for your patience and for always supporting, and encouraging me through it all, particularly the difficult times. I would not have made it without you and I look forward to starting the next chapter of our lives together.

Finally, I would like to thank my family, especially my parents Yves Castel, Lou and Tony Doyle and my sister Karen for their wisdom and ongoing support and throughout my life. Each of you have inspired me in your own special way, but you have all taught me to never give up and that no challenge is too great. Your unwavering belief in me and my abilities astounds me and I would not have accomplished this, one of the greatest intellectual challenges of my life, without you all. I am truly grateful to have been blessed with a family like you.

Table of Contents

Table of Figures

Table of Tables

Publications

Castel C.V. and Piper A.A (2011) Development of a SNP multiplex assay for the inference of biogeographical ancestry and pigmentation phenotype. *Forensic Sci Int Genet Suppl Series,* **3***,* e411-412

Presentations

Castel, C.V. and Piper, A.A. Development of a SNP multiplex assay for the inference of biogeographical ancestry and pigmentation phenotype. Poster presented at the 24th World Congress of the International Society of Forensic Genetics, Vienna, Austria, August 2011. Abstract P014.

Castel, C.V. and Piper, A.A. Mitochondrial single nucleotide polymorphisms: Informative markers of biogeographical ancestry. Poster presented at the 20th International Symposium on the Forensic Sciences: Forensic Science on trial, Sydney, Australia, September 2010. Abstract 238.

List of Abbreviations

AIM	Ancestry Informative Marker
α-MSH	α-Melanocyte Stimulating Factor
ASCCEG	Australian Standard Classification of Cultural and Ethnic Groups
ASIP	Agouti Signalling Protein
ATPase	ATP Synthase
AUC	Area Under the receiver characteristic operating Curve
BGA	Biogeographical Ancestry
bp	Base Pair
cAMP	Cyclic Adenosine Monophosphate
CEPH	Centre d'Etude du Polymorphisme Humain
CI	Confidence Interval
CIP	Calf Intestinal Phosphatase
CO	Cytochrome C oxidase
Cytb	Cytochrome b
DARC	Duffy Antigen/Receptor for Chemokine
dbSNP	SNP database
DCT	Dopachrome Tautomerase
δ	Absolute Allele Frequency Difference
ddNTP	Dideoxyribonucleotide Triphosphate
dH_20	Deionised water
DHPLC	Denaturing High Performance Liquid Chromatography
DNA	Deoxyribonucleic Acid
dNTP	Deoxyribonucleotide Triphosphate
DTT	Dithiothreitol
EDTA	Ethylinediaminetetraacetic Acid
EVC	External Visible Characteristic
ExoI	Exonuclease I
FIC	Fisher Information Content
[F]ddNTP	Fluorescently labelled Dideoxyribonucleotide Triphosphate
F_{ST}	F-Statistic

FUT	Fucosyltransferase
GDA	Genetic Distance Analysis
gDNA	Genomic Deoxyribonucleic Acid
H_e	Expected Heterozygosity
HGDP	Human Genome Diversity Project
HLTF	Helicase-Like Transcription Factor
H_o	Observed Heterozygosity
HWE	Hardy Weinberg Equilibrium
I_n	Rosenberg's Informativeness for Assignment measure
IL	Interleukin
LD	Linkage Disequilibrium
LE	Linkage Equilibrium
LEF1	Lymphoid Enhancer Binding Factor 1
MA	Minor Allele
MALDI-TOF	Matrix Assisted Laser Desorption Ionisation-Time Of Flight
MATP	Membrane Associate Transporter Protein
MC1R	Melanocortin-1 Receptor
MCMC	Markov Chain Monte Carlo
$MgCl_2$	Magnesium Chloride
MITF	Microphthalmia Transcription Factor
MLE	Maximum Likelihood Estimate
mM	Millimolar
mtDNA	Mitochondrial Deoxyribonucleic Acid
NCBI	National Center for Biotechnology Information
ND	NADH dehydrogenase
nDNA	Nuclear Deoxyribonucleic Acid
ng	Nanogram
NJ	Neighbour-Joining
NRY	Non-Recombining region of the Y-Chromosome
nt	Nucleotide
OCA2	Oculocutaneous Albinism Type 2
OR	Odds Ratio

PAR	Pseudoautosomal Region
PCoA	Principal Co-ordinates Analysis
PCR	Polymerase Chain Reaction
pg	Picogram
RFLP	Restriction Fragment Length Polymorphism
RFU	Relative Fluorescence Unit
rRNA	Ribosomal Ribonucleic Acid
SAP	Shrimp Alkaline Phosphatase
SBE	Single Base Extension
SIC	Shannon Information Content
SLC24A5	Solute Carrier Family 24 member 5
SNP	Single Nucleotide Polymorphism
STR	Short Tandem Repeat
TBE	Tris Borate EDTA
TE	Tris EDTA
T_m	Theoretical Melting Temperature
TMRCA	Time of the Most Recent Common Ancestor
TYR	Tyrosine
TYRP1	Tyrosine Related Protein-1
µg	Microgram
µL	Microlitre
USA	United States of America
UTS	University of Technology, Sydney

Abstract

Conventional DNA profiling of Short Tandem Repeats (STR) provides little evidentiary value in the absence of reference profiles or in the case of a non-match. Recently, the forensic DNA intelligence field has flourished to provide investigators with valuable information from DNA samples that can narrow the collection of potential matches by identifying previously unknown reference individuals. Intelligence data of special interest includes the biogeographical ancestry (BGA) and external visible characteristics (EVC) such as the eye, hair and skin colour of unknown DNA samples donors.

Innovative technological advances like next-generation sequencing and microarrays have been crucial to the establishment of population diversity repositories comprising millions of DNA markers, the most abundant of which include single nucleotide polymorphisms (SNPs). Large-scale SNP studies of global populations have enabled reconstructions of mitochondrial (mtDNA) and non-recombining Y-chromosome (NRY) phylogenies, providing highly comprehensive population specific patterns of maternal and paternal genetic variation. Similarly, numerous patterns of autosomal genetic variation have been identified between different populations. These studies have culminated in panels of markers capable of resolving ancestry at the continental level. The identification of autosomal SNPs associated with human pigmentation variation has also resulted in the discovery of specific SNPs capable of predicting EVCs. Several DNA intelligence and phenotyping assays for the inference of BGA and for the prediction of eye, hair and skin colour have subsequently been developed. However, most of these intelligence tools have primarily focused on the analysis of one class of SNPs, hence limiting the amount of ancestry intelligence that could be obtained. The scarcity and often environmentally compromised nature of forensic biological evidence means that performing numerous individual intelligence tests is not optimal and a consolidated DNA intelligence diagnostic test is very much needed.

This study aimed to develop a SNP genotyping system that combined autosomal, NRY and mtDNA markers for comprehensive predictions of BGA and EVCs. Candidate SNPs were selected through literature and database searches to identify loci exhibiting skewed allele frequency differences between Sub-Saharan African, North African, Middle Eastern, European, South and East Asian populations. A hierarchical

arrangement comprising five separate multiplexes was implemented, in which SNP typing was performed by single-base extension assays. The haploid mtDNA and NRY SNPs were grouped into Multiplex 1 to 4, with SNPs defining maternal and paternal lineages (haplogroups) affiliated with the same geographic region grouped in the same reaction. The markers defining basal haplogroups were included in Multiplex 1, which is then used to identify the subsequent multiplex(es) required to achieve further haplogroup resolution and to minimise the number of tests required. The autosomal SNPs are typed separately in Multiplex 5.

A performance evaluation of the 5-multiplex SNP assay was undertaken on 146 individuals originating from the six major population groups of interest. Population genetic analyses of the mtDNA and NRY haplotypes and autosomal genotypes revealed that a greater degree of population differentiation was achieved with the selected NRY and autosomal SNPs than with the mtDNA SNPs. Moreover, the results indicated that the assay primarily allowed for the differentiation of continental ancestry, with populations in close proximity within continents, such as Europe, the Middle East and South Asia, often difficult to distinguish. However, the observed correlation between the declared and inferred geographic regions of maternal and paternal origin was high; 73-100% for maternal and 79-100% for paternal regional BGA. The bi-parental BGA predictions ranged from 85 to 95%, provided Middle Easterners and Europeans were grouped into a single Western Eurasian population. In 99% of cases, two of the three SNP classes correctly predicted the same ancestry from one of the five broad geographical regions (Sub-Saharan Africa, North Africa, Western Eurasia, South Asia and East Asia). High prediction accuracies were also observed for the inference of EVCs including hair (86-88%) and eye colour (81-95%). The DNA intelligence assay also demonstrated advanced performance with low starting amounts of genomic DNA, with full profiles observed for up to 100pg of template and for the analysis of routine casework biological samples. Consequently, this study presented the successful development of a novel, consolidated DNA intelligence tool that has displayed high performance for the inference of regional (continental) BGA and EVC in preliminary tests. Further validations of the assay are required; however the developed 5-multiplex SNP assay remains a valuable DNA intelligence diagnostic tool for the forensic science community.