Assessment of 13 Forensic Molecular Markers for skin colour in South Africa

Gavin Pharo

PHRGAV001

SUBMITTED TO THE UNIVERSITY OF CAPE TOWN

In partial fulfilment of the requirements for the degree

MPhil (Biomedical Forensic Science)

Division of Forensic Medicine and Toxicology, Department of Pathology

Faculty of Health Sciences

UNIVERSITY OF CAPE TOWN

06/11/2017

Supervisor

Laura Heathfield

Word Count

Abstract: 312

Body: 15925

DECLARATION

I, Gavin Pharo, hereby declare that the work on which this dissertation/thesis is based is my original work (except where acknowledgements indicate otherwise) and that neither the whole work nor any part of it has been, is being, or is to be submitted for another degree in this or any other university.

I empower the university to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever.

Signature: .... Signed by candidate

Date: 06/11/2017

turnitin **Turnitin Originality Report**

phrgav001:PHRGAV001_THESIS_V3_turnitin.doc
by Gavin Pharo

From For Turnitin Submission (99e58091-4ed3-452e-ae8f-f82d0b847550)

Similarity Index	Similarity by Source	
4%	Internet Sources:	2%
	Publications:	2%
	Student Papers:	2%

Processed on 06-Nov-2017 11:20 SAST
ID: 875144115
Word Count: 20952

sources:

1 1% match (student papers from 14-Aug-2016)
Submitted to University of Cape Town on 2016-08-14

---

2 < 1% match (publications)
A. Ibrahim Safaa, E. Ali Amal, K. Ahmady Ali, "Phenotypic and genotypic identification of extended spectrum -lactamases (ESBLs) among clinical Isolates of Escherichia coli", African Journal of Microbiology Research, 2014

---

3 < 1% match (student papers from 31-May-2017)
Submitted to University of Cape Town on 2017-05-31

---

4 < 1% match (student papers from 02-Oct-2007)
Submitted to Cranfield University on 2007-10-02

---

5 < 1% match (student papers from 28-Oct-2017)
Submitted to University of Cape Town on 2017-10-28

---

6 < 1% match (Internet from 27-Oct-2017)
http://scholar.sun.ac.za/bitstream/handle/10019.1/97967/vergotine_molecular_2015.pdf?IsAllowed=y&sequence=2

---

7 < 1% match (Internet from 09-Nov-2015)
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0096886

---

8 < 1% match (publications)
Lucassen, Anton, Karen Ehlers, Paul J. Grobler, and Adeline L. Shezi, "Allele frequency data of 15 autosomal STR loci in four major population groups of South Africa", International Journal of Legal Medicine, 2014.

---

9 < 1% match (Internet from 14-Apr-2016)
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0015740

---

10 < 1% match (publications)
Soundararajan, Usha, Libing Yun, Meisen Shi, and Kenneth K. Kidd, "Minimal SNP overlap among multiple panels of ancestry informative markers argues for more international collaboration", Forensic Science International Genetics, 2016.

---

11 < 1% match (Internet from 28-May-2016)
http://uhra.herts.ac.uk/bitstream/handle/2299/15623/08174518%20Choong%20Melissa%20final%20PhD%20submissequence=1

---

12 < 1% match (Internet from 24-May-2010)

Table of Contents

Abbreviations

3': Three Prime

5': Five Prime

AIM: Ancestry Informative Marker

Bp: Base Pair

DNA: Deoxyribonucleic Acid

CODIS: Combined DNA Index System

EtBr: Ethidium Bromide

EVC: Externally Visible Characteristics

ExoI: Exonuclease I

Fst: Fixation Index

GWAS: Genome Wide Association Study

MBG: Molecular Biology Grade

MI: Melanin Index

PCR: Polymerase Chain Reaction

RFU: Relative Fluorescence Units

RGB: Red Green Blue

RPM: Rotations Per Minute

rSAP: recombinant Shrimp Alkaline Phosphatase

SAPS: South African Police Service

SNP: Single Nucleotide Polymorphism

STR: Short Tandem Repeats

TAE: Tris base, acetic acid and EDTA

TBE: Tris base, boric acid and EDTA

Tm: Melting Temperature

UV: Ultraviolet

List of Figures

Figure F: electropherogram showing the sequenced region surrounding *OCA2* rs1800414 A>G.

Figure G: electropherogram showing the sequenced region surrounding *PROCR* rs2069945 A/C/G.

Figure H: electropherogram showing the sequenced region surrounding *SLC24A5* rs1426654 G>A.

Figure I: electropherogram showing the sequenced region surrounding *SLC45A2* rs16891982 C>G.

Figure J: electropherogram showing the sequenced region surrounding *SLC45A2* rs26722 G>A.

List of Tables

Abstract

Molecular phenotyping is the use of informative genetic variation to estimate appearance. This concept can be applied in a forensic context to predict the appearance of suspects or decayed deceased individuals, which would otherwise remain unidentifiable. This concept has importance in a local context, as approximately 300 individuals remain unidentified, after conventional identification techniques, at Salt River Mortuary, every year. Ancestry Informative Markers (AIMs) are genetic variants with DNA which have been commonly associated with pigmentation phenotypes, and thus has value in predicting skin tone, hair colour and eye colour. This research study aimed to design and optimise an assay to genotype 13 AIMs associated with pigmentation, and then demonstrate the value of this assay by applying it to a case example and qualitatively predicting appearance. Primers were designed and PCR assays optimised to amplify each region, followed by Sanger sequencing on a case example. The case was that of an abandoned neonate, with unknown sex and ancestry. A comparison of the obtained genotypes to previous literature was performed to qualitatively estimate the skin tone, eye colour and hair colour of the decedent, which was not only in agreement with the forensic pathologist's interpretation of sex and ethnicity, but provided richer detail with regards to ancestry, skin tone, eye colour and hair colour. The PCR assays were then further optimised into four multiplex assays with the intention of genotyping these AIMs by two SNaPshot® PCR assays (Applied Biosystems) in a larger control cohort to model the relationship between these AIMs and melanin index more objectively. Unfortunately, the scope of this research project did not allow for the completion of this additional aspect. Overall, these results indicate that these 13 AIMs have potential to predict pigmentation phenotypes of South African individuals. However, genotyping and modelling of the effects of these AIMs should be performed on a large cohort to further strengthen this conclusion.

Acknowledgements

I would like to acknowledge many contributions to the completion of this Master's thesis. Firstly, many thanks must go to Laura Heathfield, in her role as supervisor of this research, as well as her assistance in the writing process. Her knowledge, encouragement and patience were all invaluable during the period of my Master's studies. To the Division of Forensic Medicine, UCT, I am appreciative for the provision of space and resources towards the completion of this research. Lastly, thanks must go to the NRF for providing funding, without which neither the research nor my Master's studies could have occurred.

Chapter 1: Introduction and Literature Review

The applicability of Ancestry Informative Marker studies performed in homogenous populations to admixed populations such as South Africa

1.1 Introduction

Ancestry informative markers (AIMs) are polymorphisms in the genome that show substantial differences in frequency between populations of different geographical origin [1]. Most commonly one variant is of high frequency in a population as compared to others, so the presence of that variant denotes an increased probability of having ancestry from that particular population [2]. Therefore, collections or panels of AIMs can be used to estimate genetic ancestry through genotyping and subsequent prediction of ancestry or ancestries present. Since populations from different locations are often distinguishable due to variations in phenotype it follows that genetic variants that are associated with phenotypic differences between populations are often the strongest markers for determining ancestry [3].

An example of this is the correlation between UV intensity, pigmentation and the frequency of the AIM rs16891982C>G in the human *solute carrier family 45 member 2* gene (*SLC45A2*). The intensity of UV radiation can vary based on location, with UV intensity being highest near the equator and lowest near the poles [4]. There is also variation between the northern and southern hemisphere, with UV intensity being higher in the southern hemisphere [5]. In Hispanics and Europeans a significant association ($p<0.05$) between constitutive skin pigmentation and burning response was found [6]. Individuals with greater skin pigmentation had a reduced burning response, which supports the idea that increased melanin provides a protective effect against skin damage caused by UV radiation. Consequently, there should be a selective pressure towards higher melanin content in indigenous populations that exist in areas of higher UV exposure.

Indeed, there is a gradient of increasing skin reflectance from the equator towards the poles, although this gradient is steeper in the northern hemisphere than the south (likely related to the higher UV levels in the south) [7]. AIMs associated with pigmentation should follow similar gradients, since they would be selected for depending on the UV intensities. Such an AIM is rs16891982, a non-synonymous mutation in *SLC45A2* (a gene involved in the pigmentation pathways) that has been associated with pigmentation levels [8].

The frequency of rs16891982 followed similar trends as seen in UV and melanin levels: the rs16891982 G allele frequency increased further from the equator, between Northern Africa and West Europe [9]. This supports the idea that AIMs that vary between populations of differing phenotypes are often associated with those phenotypes, such as pigmentation. These AIMs can therefore potentially be used to predict melanin levels in populations in which the AIM shows an association with skin, hair or eye pigmentation.

The appropriateness of such AIMs in a population would require studies to be performed in the population of interest. Many AIM selection and association studies have been performed in relatively homogenous populations, such as in Europeans or Caucasian Americans. Therefore the applicability of these AIMs to admixed populations could be questioned, since such nations have a mixture of ancestries from different geographical regions. A good example is Brazil. Brazil has a prolonged and interesting history of admixture of various population groups. Prior to discovery of the country by Europeans, the area had been colonised by a large population of varied Amerindian groups. Subsequent to European discovery, there was an influx of African slaves between the 1500's and 1900's. This was followed by immigration of European settlers in the 19th and 20th century. The history suggests that there would be an admixture of Amerindian, African and European ancestry in the population. This has been confirmed through analysis of STR markers in this population [10].

Within a forensic context, AIMs can be used to aid investigations in cases where crime DNA samples do not provide DNA profile matches within police databases. One way is through the prediction of biogeographical ancestry using AIM panels, which can be used to predict at least continental origin [11]. Another method is the prediction of the externally visible traits of an individual using AIMs associated with traits such as pigmentation. An example is the HIrisPlex, which utilises the genotypes of 24 DNA variants to predict both eye and hair colour [12]. Predictions of externally visible characteristics (EVCs) and ancestry could provide investigative leads by narrowing search criteria for unknown DNA donors (such as crime scene DNA donors) or unrecognizable human remains.

However, a thorough investigation of the literature regarding AIMs, AIM-phenotype associations and admixture determination is required to assess the applicability of AIMs (mostly selected in homogenous populations) to admixed populations. Therefore, within this review the following will be discussed:

i. The frequencies of different AIMs will be compared between homogenous populations and admixed populations to assess the informativeness of the AIMs in the latter

ii. The use of AIM panels to determine admixture proportions (and the use thereof) will be discussed in both a general and South African context
iii. The AIM-phenotype association with regards to pigmentation in homogenous and admixed populations will be compared
iv. Once an understanding of the usability of the AIMs in admixed populations is gained, the future of selection of AIM panels for both global and local usage will be discussed

The AIMs that are discussed within this review were selected because they were studied relatively frequently, preferably in different populations. The lack of a standardised AIM panel resulted in surprisingly few AIMs being shared between studies, as will be discussed further in this review.

1.2 AIMs in Admixed populations and Homogenous populations

A fundamental property of an AIM is that it shows allele frequency differences between different populations [2]. The ability to distinguish different populations would be dependent on the differences in frequencies between the populations. Since most AIM studies have been performed primarily in relatively homogenous populations, the AIMs were selected to be able to distinguish between those homogenous populations. AIMs that are informative in two homogenous populations might be of less use to distinguish admixed populations, in that an admixed population may well have more intermediate frequencies if it is a mixture of the homogenous populations. The applicability of an AIM to distinguish an admixed population would require an examination of AIM allele frequencies in both homogenous and admixed populations.

The applicability of AIMs will be discussed within the following categories: (i) able to distinguish an admixed group from some populations, but not all; (ii) distinguishes the admixed population from a single ancestral group, but groups it with others; (iii) not informative enough to distinguish populations; and (iv) not enough populations examined to accurately determine applicability.

1.2.1 Distinguishes the admixed group from some populations

SLC45A2 rs16891982C>G is an AIM that could potentially distinguish an admixed population from a number of other populations. It has been extensively studied across a wide variety of populations. It's applicability to admixed populations can therefore be determined with greater certainty. The frequency of the G allele of rs16891982 is consistently high in Europeans, varying from 0.725-1 [13-17], but its frequency within other geographical groups is more variable. Within the Asian populations the frequency is consistently low: 0.147 in Indians [15], 0-0.113 in East Asians [13-15].

There is a large departure from this trend in Thailand however, where the frequency of the G allele is 0.996 [18]. This would suggest strong selection for this allele, a strong European admixture, or perhaps a less significant reason: large departures from expectations in AIMs with G/C variants may simply result from confusion over whether one is naming according to the plus or minus strand of the gene [19]. In two indigenous groups the frequency was reported to be zero: Native Americans [20] and New Guinea Islanders [14]. While the frequency was also reported to be zero in some African cohorts [9,15], it was reported to be 0.44 in African Americans [13] and, surprisingly, 0.61-0.709 in Northern Africans [9] which suggests extensive European admixture in these African populations. Other admixed populations include Brazilians, 0.58-0.94 [19,21] and Turks, 0.615 [15].

These frequencies suggest that this AIM could distinguish Europeans from Africans (not North Africa), Native Americans and most Asian nations. Thailand is an exception, as it appears to have equivalently high frequencies to Europe. The wide range of (high) frequencies in an admixed population suggests that, while rs16891982 could be used to distinguish Brazilians from Asians, most Africans and Native Americans, it could potentially not distinguish Brazilians from Europeans, nor from admixed groups such as Turks and African Americans, since these groups also have frequency ranges around 50%.

Similarly, *solute carrier family 24 member 5* (*SLC24A5*) rs1426654G>A is an AIM SNP that can easily distinguish European ancestry from East Asian, African and Native American ancestry. The A allele's shows high frequency in Europe (0.975-1) [20,22,23], and low frequency in Africans (0.033) [23], Native Americans (0-0.09) [20,23] and East Asians (0.016-0.019) [22,24]. In Brazil the A allele frequency varies substantially between different cohorts (0.56-0.9) [19,21,25]. This suggests that the Brazilian population would be able to be distinguished from East Asian, African and Native American Populations in most cases, but not from the European population. Interestingly, other populations also have frequencies near 50%: the Sinhalese Sri Lankans (0.5) and the Central Asian Urgur population (0.536) [22]. The intermediate values may suggest Eurasian admixture. Regardless, this SNP would be less informative in distinguishing these populations from the East Asian, African and Native American and European population. It would therefore be challenging to distinguish some Brazilian population subgroups from the Sri Lankans and Urgurs, because they share an A allele frequency around 50%. This highlights the potential difficulties of distinguishing admixed populations using an AIM.

1.2.2 Distinguishes the admixed population from one ancestral group

Agouti signaling protein (*ASIP*) rs6058017G>A could potentially distinguish an admixed population from a singular other population. The G allele is present at very low frequencies in Europeans (0.12-0.13) [16,17,26,27] and East Asians (0.28) [27]. It is present at relatively high frequencies in admixed populations such as African Americans (0362-0.66) [27,28] and very high frequencies in West Africans (0.8) [27]. In the admixed Brazilian population it varies from 0.16 to 0.35, dependent on cohort location [19,21,25]. This AIM should be useful to distinguish Africans from non-Africans, but not African Americans from Africans or Brazilians from Europeans or Asians. This highlights how the potential ancestry proportion can influence which populations an admixed population can be distinguished from.

HECT and RLD Domain Containing E3 Ubiquitin Protein Ligase 2 (*HERC2*) rs1129038C>T can distinguish European from non-European nations. The C allele has a low frequency (0.18-0.224) in European nations [16,23] but has higher frequencies (0.996 in Africans, 0.983 in Native Americans) in non-European nations [23]. De Cerqueira *et al*. (2014) examined its frequency in two Brazilian populations: in the Baiano subgroup the frequency is 0.816, therefore this AIM would be useful in distinguishing this subgroup from Europeans, but not Africans or Native Americans [19]. In the Gaucho subgroup the frequency is 0.563. While this AIM could still be useful to distinguish this subgroup from the other ancestries on a population level (since there is still a large frequency difference), it would not be as informative.

Contrasting the presence of *ASIP* rs6058017 and *HERC 2* rs1129038 in multiple populations, *oculocutaneous albinism 2* (*OCA2*) rs1800414A>G appears to have been selected for in the Asian populations only. The frequency of the G allele in East Asian populations varied from 0.546-0.602 [24,29], while it is has a frequency of <0.01 in Europeans, Africans and the admixed Brazilians [15,19]. It would be useful to distinguish an admixed population (without significant Asian admixture) from the Asian population, but not from Europeans or Africans.

1.2.3 Not informative enough to distinguish populations

Some AIMs cannot meaningfully distinguish admixed populations from homogenous populations. *SLC45A2* rs26722G>A represents an interesting case in that the A allele is not present at high frequency in any of the examined populations, however it shows a higher frequency in Asians as compared to other continental groups. It has a frequency of 0.339-0.3835 in Asians [13,15], 0.01-0.028 in Europeans [13,16,17,30], 0.058 in Africans [15] and 0.029 in Australian Aboriginals [13]. In

the admixed Brazilians the frequency is 0.089-0.14 [19,21] and 0.25 in African Americans [13]. This AIM would be useful to distinguish Asians from other homogenous ancestral groups, since the frequency difference is reasonably high (30% and greater). However, the frequency in the admixed Brazilians and African Americans is intermediate between the Asians and other groups, so it wouldn't be informative to distinguish the admixed populations from the homogenous populations.

1.2.4 Insufficient number of populations examined

Some AIMs have not been investigated in enough populations to make an informed decision as to their applicability in admixed populations. Examples include *OCA2* rs1800401G>A, rs1800407G>A and *Tyrosinase* (*TYR*) rs1042602C>A, rs1126809G>A. The A alleles of these AIMs have low frequencies in Europeans and admixed Hispanic populations. The *OCA2* rs1800401 A allele has a frequency of 0.06-0.068 in Europeans [17,31] and a frequency of 0.059-0.107 in Brazilians [19,21]. The *OCA2* rs1800401 A allele has a frequency of 0.04-0.082 in Europeans [16,17,30,31] and a frequency of 0.06-0.062 in Brazilians [19,21]. The *TYR* rs1042602 A allele has a frequency of 0.35 in Europeans [16], 0.23-0.403 in Brazilians [19,21], 0.234-0.26 in Mexicans [32,33] and 0.29 in Puerto Ricans [32]. The *TYR* rs1126809 A allele is found at a frequency of 0.28 in Europeans [17] and a frequency of 0.131-0.3 in Brazilians [19,21].

These four AIMs would not be informative in distinguishing admixed Hispanic populations from European populations, but there has not been enough research in other populations to determine their applicability in distinguishing an admixed population from a more diverse range of ancestries.

An examination of all of these AIMs leads to the conclusion that, while some AIMs selected in homogenous nations can be used to distinguish an admixed population from at least one other ancestral group, often the admixed nations will be indistinguishable from a group of other ancestries. These AIMs are therefore not useful as a singular measure to distinguish an admixed population from all other continental ancestries. This is expected given the mixture of ancestries present in such populations.

1.3 Ancestry proportion estimation

1.3.1 Variation in Admixture proportions

As discussed, AIMs that are selected for usage in homogenous populations tend to be less informative in distinguishing admixed populations. However, a panel of AIMs that distinguish

strongly between the ancestral components of an admixed population can be used to determine the proportions of the various ancestral groups' contribution to the genome of admixed populations or individuals [34].

Any variants that show large differences between populations could be used (but aren't equally informative). For example, 15 Short Tandem Repeat (STR) markers from the AmpFISTR Identifiler Kit (used in CODIS) have been used to determine genetic admixture and diversity estimations in the Mexican Mestizo population [35]. They could predict a 5% African, 26% European and 69% Native American contribution. Furthermore, they determined significant differences ($p<0.003$) between this population and American Hispanic, North Eastern Hispanics, Caucasian Americans and African Americans (but not Native Americans or South Western Hispanics).

Even such a simplistic analysis can provide useful information, as a starting estimate of ancestral proportions which can inform the selection of larger panels of AIMs. Studies using larger panels have been performed in a number of admixed American nations: these distinguish far finer differences in proportions than the 15 marker panel.

These analyses revealed important information regarding variation in proportions between even neighbouring admixed countries. A study using 30 informative AIM SNPs found extensive variation in ancestry proportions between Brazil, Chile, Colombia, Mexico and Peru, with Native American Ancestry ranging from 0.09 in Brazil to 0.64 in Peru, African Ancestry ranging from 0 In Peru to 0.11 in Columbia and European ancestry varying from 0.29 in Peru to 0.82 in Brazil [34]. Variation between ancestry proportions in Argentina, Brazil and Colombia has also been found using STR profiling [36], and by meta-analysis of AIM studies performed in Brazil, Argentina, Colombia, Mexico, Peru, Puerto Rico and Venezuela [37]. Therefore, using a single database to describe the ancestry proportions across the admixed South American nations would not be advised, since there are substantial differences in ancestry proportions (and therefore AIMs) between even neighbouring countries [36].

Moreover, studies performed within Brazil, Argentina and the Cape Verde Islands have shown that ancestry proportions can differ significantly within different regions of a single country. In Brazil, African ancestry is highest in the North East (0.27), European in the south (0.77) and Native American in the North (0.32) [37]. Intra-country variation was reported as significant in Brazil, as well as in Chile, Colombia, Mexico and Peru ($p<0.02$) [34]. Avena *et al*. (2012) reported significant variation in European and Native American Ancestry within Argentina, with European ancestry varying between 0.54 and 0.76 ($p<0.001$) [38]. In the Cape Verde Islands, a nation that has had

historical European and West African admixture, West African ancestry has also been found to vary significantly across the archipelago ($p < 0.001$)[39].

Given the significant variation in ancestry proportions between and within admixed countries, it becomes clear that, while admixture studies at the level of countries provide important information regarding ancestry proportions, the information provided may not be sufficient. In admixed countries it would be suggested that all regions within the country must be sampled to provide a truly representative sample, or that each region within a country should be examined [39]. This is necessary so that proper precautions can account for the differing admixture proportions in any genetic studies performed in these admixed nations [40].

1.3.2 Admixed populations in the South African context

Many studies have been performed within the admixed populations of South America, but only a few have been performed within South Africa, a country containing individuals of numerous ancestries, and an admixed population, locally known as the Coloured community**.** The South African population was reported in 2016 to be composed of 80.7% Black African, 8.8% Coloured, 2.5% Asian/Indian and 8.1% White individuals, making the Coloured community the largest of the non-Black African groups in RSA (Statistics South Africa 2016, www. statssa.gov.za).

Studies have utilised different forms of potential AIMs, such as STR markers or SNPs. Like the previously mentioned Mexican study [35], Lucassen *et al.* (2013) utilised the 15 STR markers of the AmpFISTR Identifiler Kit to attempt to distinguish populations groups in South Africa [41]. A comparison of the South African Coloured population and the South African Black population showed an average number of 14.5 loci with significant allele frequency differences. The South African Coloured population and the South African white population showed an average number of 12 loci with significant allele frequency differences [41].

Patterson *et al*. (2010) analysed 900 000 SNPs in 20 unrelated AIM genes in the Coloured population of South Africa [42] . They compared this population to cohorts from ancestral populations that were historically known to have contributed to the Coloured population. Analysis of the comparisons using the STRUCTURE program revealed a four-way admixture in this cohort: 0.231 European ancestry contribution, 0.221 South Asian contribution, 0.180 Indonesian contribution and 0.369 IsiXhosa contribution [42]. Moreover, since the analysis indicated that the IsiXhosa subgroup showed admixture between Khoisan and Bantu (though primarily Bantu), the Coloured population could be considered to be a 5-way admixture.

De Wit *et al*. (2010) show a different picture, in that their results showed a lower Asian contribution [43]. They genotyped 75 000 autosomal SNPs in 959 self-identified Coloured individuals in the Western Cape, and compared their cohort to the HapMap populations. Their model (which took account of the linkage disequilibrium which occurs in admixed populations) estimated 0.21 European, 0.323 Khoisan, 0.359 African and 0.108 Asian ancestral contribution proportions. While the European and African contributions were similar to the results of Patterson *et al*. (2010) [42], the Asian contribution (South Asian and Indonesian in Patterson *et al*. (2010)) was considerably smaller. This could be explained in a number of ways. As mentioned by de Wit *et al*. (2010), the Cape Malay subgroup of the Cape Coloured population is known to have a strong historical influence from Indonesia [43]. Since Patterson *et al*. (2010) only recruited 20 Coloured people, this subgroup could be overly represented. This would indicate substantial population stratification occurring within the coloured community. An alternative is that because de Wit *et al*. (2010) examined the Khoisan population as an ancestral group, the AIMs that were aligned to Asian ancestry in the previous study may have been more accurately aligned as Khoisan in this study.

Despite the differences, it is clear that there is at the least a four-way admixture occurring in the Coloured community.

In cases of relatively recent admixture (350 years as predicted by de Wit *et al*. (2010)) [43], population stratification can occur which can lead to coincidental frequency differences between cases and controls in association studies, due to differences in frequency in subgroups of a population [44]. More studies in the Coloured community should be performed to understand the stratification occurring within it.

1.3.3 Use of admixture proportions to correct for spurious associations

The presence of admixture in an association study population can be both boon and burden. Relatively recent admixture within a population results in larger blocks of the genome that are in linkage disequilibrium (i.e. are inherited together more than would be expected) than would be found in non-admixed populations [2]. This allows for a technique called admixture mapping to be used to more easily determine disease causing variants in a population by only needing to search within regions of the genome in admixed cases that are contributed by the ancestry group which has the higher disease risk. The larger blocks of linkage disequilibrium in admixed populations make this easier because comparatively fewer markers are needed to map the genome, since large areas are linked together [2]. However, recent admixture can result in population stratification (where

different subgroups of a population have differences in frequencies of markers by chance), which can lead to false associations between markers and phenotypes, since markers that are informative AIMs are often strongly linked to traits that vary between populations [3].

The admixture proportions in a population (and subpopulation) need to be accounted for. Sinha *et al.* (2006) compared self-reported race and genetic ancestry analysis in Cleveland [45]. They found that self-reported race was clearly associated with ancestry proportions, and that they could clearly separate the population into groups that corresponded with the self-reported "white" and "black" groups. However, their two population model was simplistic and 94 of the "black" individuals were of mixed ancestry. Using overly simplistic models and broad self-classification may not be useful, since individuals broadly labelled as "Hispanic" can show European or Native American ancestries across the entire proportion spectrum, and the grouping includes numerous South and Central American populations [46]. Individuals can also show bias in self-estimations, with 84% of Hispanic individuals underestimating their levels of Native American ancestry (and Native Americans underestimating their levels of European ancestry) [47].

However, studies performed in admixed populations like Mexico and Puerto Rico have shown that individual admixture estimation using general markers or AIMs can account for and be used to compensate for admixture and population stratification. Beuten *et al.* (2011)[33] examined the associations between 64 AIMs and admixture in two Mexican cohorts. To determine the presence of confounding admixture, they tested the linkage disequilibrium between the 64 markers. A linkage disequilibrium of 5% would be expected, but 10.5% of unlinked marker pairs (on different chromosomes) in one cohort and 27.4% in the other showed significant association with each other, indicating recent admixture. Ancestry proportion estimation using the AIMs revealed significant differences in European and Native American ancestry between cohorts, indicating population stratification. To account for the confounding effects of admixture and stratification, for each AIM-phenotype association test, Beuten *et al*. (2011) used individual ancestry (estimated using the 63 other AIMs) as a covariate in analysis. Using Native American ancestry proportion as a covariate reduced significantly associated AIMs from 7 to 1, with none being significant after Bon-Ferroni corrections. Therefore, not accounting for admixture proportions can lead to an artificial inflation on the significance of AIM–phenotype associations [33].

Knowing the proportions of ancestry in admixed individuals can provide a more accurate reflection of marker–phenotype association significance, both in revealing false positive associations and revealing potential false negatives in association [32]. The use of ancestry matched case-control studies are not sufficient to eliminate stratification [32], so analysis using sufficiently sized panels of

AIM markers should be used to assess stratification and admixture in all studies in admixed populations. This would assist in identifying true associations between genotypes and diseases or phenotypes such as pigmentation.

1.4 AIM–phenotype associations in homogenous and admixed populations

An examination of AIM-phenotype associations requires that certain important caveats should be noted: not all associations detected between variants and phenotypes using GWAS studies exhibit a causal relationship [48]. It is conceivable therefore that the linkage between a causal variant and the detected variant may be broken in admixed populations due to the genome being a combination of various ancestral groups. Additionally, complex phenotypes like pigmentation are multigenic, and interactions between different genes can influence pigmentation [49]. An AIM's effect may be modulated in an admixed population due to the presence of other variants in other pigmentation genes [49]. Therefore AIM-phenotype associations should be examined in the admixed population, to see whether the associations seen in an ancestral population are still present, as the associations in the ancestral population may not necessarily be extrapolated.

The comparison of associations of AIMs and phenotypes such as pigmentation could provide a few categories, dependent on whether an AIM shows association in homogenous population/s, admixed population/s, or both.

1.4.1 Associated in a Homogenous population but not admixed populations

Some AIMs show associations within a homogenous population, but not an admixed one. An example is *AFG3 Like Matrix AAA Peptidase Subunit 1* (*AFG3L1*) rs4785763A>C, which shows significant association with a minor skin pigmentation trait, the presence of freckling (p=3.0X10^-86) in Icelanders [50], but no associations with skin pigmentation in a Brazilian population [19]. While European populations are a contributor to Brazilian ancestry, certain traits (like freckling, red hair) may be relatively common in the Icelander population [50], but would not be expected to be seen within a Hispanic population.

OCA2 1800401 also showed associations in Europeans and Rebbeck *et al.* (2002) [31] found the T allele to be associated with darker eyes, while Nan *et al*. (2009) [17] found it to be marginally associated with darker skin (p=0.04). No associations with skin colour were seen in two Brazilian subpopulations however [19]. This suggests that AIMs associated with even major components of an

admixed ancestry may not show associations in admixed populations, but there may be alternative explanations. Nan *et al.* (2009) [17] looked at self-reported colour categories, while de Cerqueira *et al.* (2014) examined melanin index (MI), so the differences in the type of data examined may have had an effect. Furthermore, Rebbeck *et al*. (2002) [31] examined eye colour, a variable that de Cerqueira *et al*. (2014) [19] did not examine, which may show that associations in some aspects of pigmentation do not necessarily prove associations in all pigmentation.

A similar situation is seen in *OCA2* rs1800407. The rs1800407 T allele was associated with pigmentation traits such as darker eyes (p=7.7E-13), green eyes and red hair in a number of European cohorts [16,30,31], but not all [17]. It was also significantly associated with total hair melanin and skin reflectance in a cohort composed of various ancestries [51]. De Cerqueira *et al.* (2014) [19], found no associations with skin pigmentation though. This further suggests that associations cannot necessarily be extrapolated to the admixed population, without further extensive research.

OCA2 rs1800414 represents another case of an association in a homogenous population only. This SNP was associated with total hair melanin in various ethnicities ($p<0.05$) [51], and the G allele was negatively associated with melanin index in East Asian populations ($p<0.05$) [24,29]. It was not significantly associated with pigmentation in Brazilians [19]. These results aren't unexpected, since the Brazilian population is an admixture of European, African, and Native American ancestry, so an AIM only found at high frequency in Asian populations [15,19,24,29], would not likely show similar associations. De Cerqueira *et al.* (2014) [19] did not assess hair colour, so a comparison to the results of Valenzuela *et al*. (2010) could not be performed.

1.4.2 Associated in one Homogenous population and admixed populations

In contrast to the SNPs discussed previously, *HERC2* rs1129038 provides an example of an AIM associated with an ancestral population and an associated admixed population of which is it a component. The *HERC2* rs1129038 A allele was completely associated with blue eyes in a Danish cohort, while the G allele was strongly and significantly associated with darker eye colour in a Dutch cohort [23,52]. Within two Brazilian subgroups, this AIM showed no significant association, but showed significant associations with melanin index when the cohorts were combined [19], suggesting that the individual sample sizes may not have had the statistical power to detect the association.

Similarly, *TYR* rs1126809 has shown significant associations in a European population, with the A allele being significantly associated with paler self-categorised skin colour [17]. In two Brazilian subgroups this AIM showed no significant association, but showed significant associations with melanin index when the cohorts were combined (p=0.001), with the G allele being associated with higher melanin index [19]. These two SNPs may provide an example of an AIM associated with pigmentation in an ancestral component population, which has a more modest effect in the admixed population, given that a larger sample size (and therefore a higher statistical power) was required to detect the association.

SLC45A2 rs26722 showed interesting patterns of association. It was significantly associated with total hair melanin, skin reflectance and eye colour in a cohort of various ancestries [51], and the A allele was significantly associated with darker eyes, skin and hair in Europeans (including Polish and Dutch) [13,16,17,30]. However, the associations show less clarity: the G allele was associated with darker eyes, yet the A allele was associated with darker hair [53]. In contrast, de Cerqueira *et al.* (2014) [19] found no significant associations with skin pigmentation. Given the variation in ancestry proportions in different regions of Brazil, it is unsurprising that the associations may not be present in all examined cohorts. The associations (verified in European populations) may be dependent of European ancestry levels. A viable alternative is that in the cohort examined by Fracasso *et al*. (2013) [53] the population stratification present in the admixed population has led to false positive associations, since they did not use individual ancestries as a covariate. Further studies would need to be performed to verify the applicability of this AIM for determining phenotype in admixed populations.

1.4.3 Associated in Homogenous populations and admixed populations

An AIM may not show associations in all the populations in a particular ancestral group. For example, *ASIP* rs6058017 was significantly associated with hair, eye and skin pigmentation in a mixed cohort [51], while the G allele was associated with increased melanin index in African Americans [28] and darker hair, eyes and reduced chance of freckling in Europeans [26,50]. However, in other European cohorts no significant associations were found [16,17]. While Nan *et al*. (2009) admit their cohort had limited statistical power, the 6168 person cohort of Liu *et al.* (2009) [16] should have had more statistical power. Similarly to Europeans, in Brazilians the G allele of the AIM was significantly associated with increased skin pigmentation in some cohorts [21,25] but not others [19]. An AIM showing inconsistent associations in an ancestral population and an admixed

population would need to be assessed in the regional population of interest before being used, since one could not necessarily extrapolate the data in one region of a country or continent to all of it.

An AIM with more consistent associations across different populations is *SLC24A5* rs1426654. The relative abundance of research for it has revealed strong associations in the admixed Brazilian population and others: it was associated with self-assessed colour category [21] and the A allele was negatively associated with melanin index, eye pigmentation and hair pigmentation levels in Brazilians [19,23,25] it has also been associated with hair, skin and eye pigmentation in a mixed cohort [51] . The A allele was negatively associated with melanin index in the admixed Hispanics and Cubans [20,54] , as well as in East Asians [24]. This AIM shows consistent, significant association with pigmentation across a number of admixed (and some non-admixed) populations, and so should be considered a good candidate for use in these populations.

SLC45A2 rs16891982 has also had its associations with pigmentation studied in a number of admixed cohorts, showing consistent associations. It was significantly associated with total hair melanin, skin reflectance and eye colour ($p<0.05$) in a mixed cohort [51] and the C allele was significantly associated with darker hair, skin and eyes in various European cohorts [13,16,17,30]. The C allele has also been associated with reduced melanin (non-significantly) in an East Asian cohort [24]. While there appears to be a dearth of research regarding its associations in African populations, it should be noted that the G allele (associated with lighter pigmentation) is completely absent within a homogenous African population, the Ghanaians [14]. In admixed Brazilian and Hispanic cohorts the AIM consistently showed association with skin pigmentation (self-assessed and melanin index), eye colour and hair colour [19-21,53,55] and was significantly associated with melanin index in an admixed Cuban cohort [54]. These consistent associations with various forms of pigmentation across a variety of admixed cohorts suggest that this AIM would be a strong and informative candidate to use in phenotypic prediction in admixed populations. There is a discrepancy across the studies as to whether the G or C allele is associated with increased pigmentation, although discrepancies in G/C SNP variants may simply result from confusion as to whether the plus or minus strand of the allele was being analysed [19].

1.4.4 Associated in Admixed populations only:

TYR rs1042602 provides an interesting example, in that it is significantly associated with self-assessed colour categories in a Brazilian cohort from Rio de Janeiro ($P<0.05$) [21] but not in two other Brazilian populations [19]. Additionally, it was not significantly associated with self-reported

colour in Americans of European descent [17]. The differences in association within the different Brazilian cohorts may be explained by the differences in pigmentation measures: Durso *et al*. (2014) [21] used self-reported categories, which may be both broad and biased, whereas de Cerqueira *et al*. (2014) [19] used an objective measure of pigmentation, melanin index. Further studies would need to be performed in other ancestral population that contribute to Brazilian admixture, such as African and Native American populations.

A comparison of the associations of these AIMs and pigmentation phenotypes in relatively homogenous and admixed populations clearly revealed that the associations in homogenous populations may not necessarily be extrapolated to admixed populations. AIMs that showed consistent associations across a variety of ancestral populations appeared to show more consistent associations in admixed populations, so these AIMs should be considered as candidates for association studies in the admixed populations. An example of this can be seen in *SLC45A2*, which showed associations with pigmentation in Europeans, Asians, Hispanics, Cubans and Brazilians. Alternatively, initial GWAS studies should be performed in the admixed populations themselves, to identify relevant variables, independently of those variants being associated in ancestral populations.

1.5 Development of new AIM panels:

Within this review it has been discussed that AIMs that can distinguish homogenous populations are not necessarily as informative to distinguish admixed populations from other populations. Nevertheless, some of these AIMs could still be useful in AIM panels to distinguish the ancestral proportions in admixed populations. This information is important to account for population stratification in association studies in admixed populations, which may not show the same strength of associations.

One of the primary issues associated with current AIM panels is that the markers used in different panels are often different, so making meaningful comparisons between the results of different panels can be difficult. Soundararajan *et al.* (2016) performed a systematic review and identified 21 different AIM SNP panels, with a combined total of 1397 SNPs [56]. They determined that only 3% of the SNPs were present in four or more panels, while 87.5% were only present in one panel. *SLC24A5* rs1426654 and *SLC45A2* rs16891982 (associated with pigmentation) were the only SNPs to appear in six panels, which both had high Fst values (0.658 and 0.595 respectively). They analysed the top 20 SNPs contributing to distinguishing each of the five-six major clusters, *SLC24A5* rs1426654 was in the

list for Africa, SW Asia, Europe, South Central Asia, East Asia and America and *SLC45A2* rs16891982 was in the list for SW Asia, Europe and South Central Asia. The authors noted that all of the 21 panels could distinguish at least three continental groups, Europe, sub-Saharan Africa, and East Asia.

Their suggestions for future AIM panel development were that the forensic community should collaborate to create panels that are more capable of nuanced distinction of biogeographical regions, beginning with a common panel of the best overlapping AIMs from multiple panels. Creation of these new panels should focus on a number of ideals: identifying AIMs which have high informative value to distinguish different populations and balancing the marker panels such that the different population groups show equal levels of differentiation [57] .

This work should be performed on two levels: a broader spectrum panel of relatively few SNPs which can distinguish larger biogeographical groups, as well as panels that utilise SNPs of a narrower biogeographic value, which can distinguish subsets within the broader groupings [56].

An example of a narrow biogeographical panel would be those designed to be used in admixed populations, to determine admixture proportions therein. AIMs that are selected to distinguish the ancestry proportions would only be effective if they are informative with regards to distinguishing the contributory ancestral populations of the admixed population [58]. The accurate inference of ancestry proportions in admixed populations is important, since recent admixture can lead to false positive AIM-phenotype associations [32]. Sufficiently large or informative AIM panels are important to account for this effect. In an admixed Brazilian cohort, using AIMs that had an Fst (a measure of population) of 0.5 or higher was found to reduce the AIMs required to account for admixture by 75% [59]. For panels created for use in admixed populations, the AIMs must be informative in that particular population, and selected such that the ability to differentiate all present ancestral proportions is balanced. This would ensure that a particular ancestral proportion is not overestimated due to an excess of markers informative for that ancestry [58].

The creation of both the global and local AIM panels would require testing in large number of populations (homogenous and admixed) to ensure adequate coverage. This may require substantial collaboration between research groups and organisations, as well as extensive research into informative AIMs in admixed populations around the world [56].

1.6 Conclusions

AIMs are defined by large frequency differences between populations, but these markers are usually studied in homogenous populations, and selected because they are informative within those

populations. How informative such AIMs would be in admixed populations needed to be examined. A comparison of AIM frequencies and associations in different populations was difficult because there are very few markers that are common between studies. Nevertheless, an examination of a few selected markers between populations revealed that markers that are informative in even the ancestral populations of admixed populations can be less informative with regards to admixed populations. Similarly, AIMs that show associations with phenotypes such as pigmentation in homogenous populations do not necessarily show significant associations in the admixed populations. Strong consistent associations across multiple populations appear to result in a greater likelihood of significant associations in an admixed population. Future association studies should either focus on these AIMs or on AIMs elucidated within the admixed populations themselves. However, association studies in admixed populations can result in inflated false positive AIM-phenotype associations if the ancestry proportions of the cohort are not accounted for. AIMs selected in homogenous populations can be used in this manner, in the context of AIM panels which are created to distinguish between the ancestral components relevant to the admixed population of interest. Creation of new AIM panels should occur on two levels: collaborative creation of globally relevant panels using AIMs which are shared between numerous AIM panels, and creation of locally relevant panels, using AIMs that are informative in that specific context. It is clear that, although data from homogenous populations should not be extrapolated to admixed populations, informative AIMs found in these populations may still be able to be used in the context of admixed populations. More research needs to be performed in these populations however, preferably using AIMs that allow for meaningful comparisons between studies.

1.7 Rationale for the Research

There is a dearth of research regarding forensically relevant genetic variation in pigmentation pathways in populations that exhibit admixture, such as South Africa. The majority of research in this field is performed in Europe and America (North and South), and may not be applicable to the South African population. This study aims to design and optimise a PCR assay to amplify 13 AIM SNPs and, through sequencing of these amplicons, examine associations of the SNPs and pigmentation phenotypes (as reported in literature) using a South African forensic case study. Indications that these 13 AIMs are informative within a South African context would inspire the optimisation of this assay in a multiplex format could provide a panel of SNPs that could be used in the South African context to model melanin index in future (as well as other pigmentation phenotypes such as hair colour and eye colour). The prediction of these pigmentation phenotypes may prove useful in the

identification of unrecognisable decedents or assist in identifying the donors of DNA samples from crime scenes.

1.8 Aims and Objectives

The aim of this study is to examine the associations of 13 AIM SNPs and pigmentation in South Africa.

This will be achieved through the following objectives:

- Developing 13 external primer sets to amplify the 13 AIM SNPs
- Using these primer sets to amplify and sequence to genotype the 13 SNPs in a forensic case study, and perform a qualitative analysis of pigmentation.
- Contributing towards the development of multiplex PCR reactions and SNaPshot® assays to sequence the 13 SNPs in a larger cohort.
- Recruiting 100 South African individuals towards this ongoing study and collecting DNA samples, demographic data and melanin index data from the participants.
- Contributing towards the determination of the association between these SNPS and the melanin index (as determined using a Derma spectrometer), as well as other pigmentation phenotypes.

Chapter 2: Methodology

2.1 Assay design: phase 1

2.1.1 AIM selection

13 informative AIMs were selected for this study, these included 11 of the 18 AIM SNPs utilised by de Cerqueira *et al.* (2014) (all of which had been associated with pigmentation) [60], as well as 2 trialleleic SNPs used by de la Puente *et al.* (2016), which showed differing allele frequencies in multiple population groups [61]. The selected AIMs can be seen below in Table 1:

Table 1: ***The 13 AIM SNPs investigated in this study***

Gene	Rs number *	Base pair change	Type of change
ADAM17 (NG_029873.1)	rs1524668	A>C	Non-coding, upstream
AFG3L1P (NR_003226.1)	rs4785763	A>C	Non-coding
ASIP (NG_011439.1)	rs6058017	A>G	3'UTR
HERC2 (NG_016355.1)	rs1129038	A>G	3'UTR
OCA2 (NG_009846.1)	**rs1800407**	G>A	Missense, Arg>Gln
OCA2 (NG_009846.1)	rs1800401	C>T	Missense, Arg>Trp
OCA2 (NG_009846.1)	**rs1800414**	A>G	Missense, His>Arg
SLC24A5 (NG_011500.1)	**rs1426654**	A>G	Missense, Thr>Ala
SLC45A2 (NG_011691.2)	rs26722	G>A	Missense, Glu>Lys
SLC45A2 (NG_011691.2)	**rs16891982**	C>G	Missense, Phe>Leu
TYR (NG_008748.1)	rs1126809	G>A	Missense, Arg>Gln
PROCR 9 (NG_032899.2)	rs2069945	C>G/A	Non-coding
TLR1 (NG_016228.1)	rs4540055	T>A/G	Non-coding

*SNPs in **bold** are utilised in the Illumina ForenSeq DNA Signature Prep kit.

2.1.2 External primers

The regions containing the 13 AIMs were selectively amplified using the polymerase chain reaction (PCR) [62]. To this end, design of external PCR primer pairs was performed for 9 different genomic regions. Initial primer design was performed using Primer3 v4.0.0 (http://primer3.ut.ee/) [63]. Primers were designed such that their length was 18-24bp, GC content was 40-60% and primer melting temperature (Tm) was 50-65^{o}C. Analysis of homodimers and heterodimers of individual

primers and matched primer pairs were performed using Oligoanalyzer v3.1 (https://eu.idtdna.com/calc/analyzer) (Integrated DNA Technologies Inc., USA). Primers were selected such that the ΔG>-0.4kcal/mol of all homodimers and heterodimers where possible, in all cases the primers were designed so as to minimise these interactions. The Tm of the hairpins was required to be <35^{o}C so as to minimise their inhibition of primer-template binding. Heterodimer formation between all primers was assessed using Autodimer v1 [64]. Primers were selected such that the ΔG>-0.4kcal/mol for the heterodimers of primers that were intended to be used in multiplex PCR.

The specificity of the primer pairs was assessed using Primer-BLAST (https://www.ncbi.nlm.nih.gov/tools/primer-blast/) [65]. Specificity was deemed adequate if an amplicon of the correct size from the sequence of interest would be amplified, with no non-specific products of <1000bp being produced. Non-specific products of >1000bp were considered to be acceptable because the amplification of these larger products could be eliminated during the optimisation process (smaller amplicons would be preferentially amplified). The presence of SNPs in primer sequences was assessed using SNPCheck3 (secure.ngrl.org.uk/SNPCheck/snpcheck.htm; date accessed 10/05/2016). The presence of SNPs close to the 3' end of the primers was not considered to be acceptable, unless that SNP was present at a low frequency (<0.001).

The primers for *SLC24A5* rs1426654 and *SLC45A2* rs16891982 had previously been designed during the course of an in-house pilot study; while the primer sequences for *PROCR* rs2069945 and *TLR1* rs4540055 were taken from de la Puente *et al* (2016). [61]. The sequences for all external primers can be seen in Table A (Appendix 3).

2.1.3 Optimisation of Singleplex PCR

Optimisation of the external primer pairs began with optimising each pair singularly**.** The general PCR reagent setup included 2X KAPA HiFi HotStart Taq ReadyMix (to a final concentration of 1X) (Kapa Biosystems, USA), appropriate forward and reverse primers (each to a final concentration of 0.3µM) and molecular biology grade (MBG) water, to a final volume of 25µl per reaction. The reaction mix was set up as suggested in the HiFi HotStart ReadyMix PCR Kit Technical sheet, with 100ng of DNA being used as the template for the reaction. MBG water was added to the no template control instead.

PCR was performed in the Bio Rad T100 Thermocycler (Bio Rad Laboratories, USA).Conditions included initial denaturation at 95°C for 5 minutes; 30 cycles of: denaturation at 98 °C for 20s, primer annealing at temperatures from 55-65°C for 30s and extension at 72°C for 30s; and a final extension step at 72°C for 10 minutes.

The initial reaction conditions involved optimisation using a temperature gradient between 55-65°C. The same initial temperature gradient was used for all external primer pairs because this would facilitate multiplexing later by allowing comparison across the same temperatures. If a specific, singular band of the expected PCR fragment for the primer pair was not produced at any of the temperatures of the initial temperature gradient, the temperature range was increased from 55-65°C to 60-70°C.

2.2 Case Study

2.2.1 Case history

A case-study was utilised to demonstrate a proof of concept of the value of the 13 AIMs in predicting pigmentation phenotypes within a South African context. In May 2017, a decedent was admitted to Salt River Mortuary, Cape Town, as an alleged abandoned neonate. The neonate was found between a dumping site and a toilet; the identification of the neonate, including its sex and race, was unknown. During the autopsy, a blood sample was obtained, from which DNA was extracted using a standard in-house salting out method.

2.2.2 PCR and sequencing

A working solution of the DNA sample, diluted to 100ng/µl, was used for each singleplex PCR. For each of the 13 AIM SNPs, the reaction setup and optimised conditions were used. PCR reactions which did not show bright, clean product bands on the agarose gel were subsequently amplified using GoTaq® Green Master Mix (Promega, USA).

Post-PCR Cleanup, BigDye® Terminator sequencing and capillary electrophoresis of the amplicons was performed at the Central Analytical Facility of Stellenbosch University. Sequencing was performed using the appropriate forward or reverse external PCR primer (Table 2). According to the BigDye™ Terminator v3.1 Cycle Sequencing Kit User Guide, the 5’ 35bp of a sequencing reaction had a high chance of poor sequencing quality [66]. Therefore the primer (forward or reverse) furthest from the SNP was selected as the primer for sequencing. This would minimise the chance of initial

poor sequence quality obscuring the identity of the SNP of interest. The sequencing data produced was visualised and analysed using Bioedit Sequence Alignment Editor v7.2.5, and the genotype of the appropriate SNP for each sequenced fragment was determined.

Table 2: AIMs sequenced for case study with their corresponding amplicon size and primer used for sequencing

***Gene* (SNP)**	**PCR amplicon size (bp)**	**Forward or Reverse Primer used**
***SLC24A5* (rs1426654 A/G)**	114	Reverse
***OCA2* (rs1800414 G/A)**	258	Forward
***HERC2* (rs1129038 G/A)**	380	Reverse
***PROCR* (rs2069945 G/C/T)**	95	Reverse
***OCA2* (rs1800407 C/T)**	254	Forward
***SLC45A2* (rs26722 G/A)**	374	Forward
***OCA2* (rs1800401 C/T)**	469	Reverse
***TLR1* (rs4540055 A/C/T)**	76	Reverse
***ADAM17* (rs1524668 A/C)**	341	Reverse
***AFG3L1* (rs4785763 G/T)**	658	Reverse
***SLC45A2* (rs16891982 G/C)**	128	Forward
***TYR* (rs1126809 C/T)**	341	Forward
***ASIP* (rs6058017 C/T)**	586	Forward

2.3.3 DNA profiling

The assay designed in this study cannot predict sex, but routine DNA profiling does. Therefore, 5ng of DNA underwent STR profiling using the Investigator 24plex GO! Kit (Qiagen) as per the manufacturer's protocol. Capillary electrophoresis was performed on the 3500 Genetic Analyser (Applied Biosystems, city). Results were analysed using GeneMapper ID-X software.

2.2.4 Data Analysis

These genotypes and the Amelogenin marker (used in the DNA profiling assay to determine sex) were compared to the literature to form a qualitative assessment of variables such as skin colour, hair colour and eye colour for the deceased individual. These results were compared to the autopsy of the deceased individual to establish the veracity of the assessment.

2.3 Assay design: Phase 2

To objectively assess the relationship between the EVC and genotypes, modelling would be required, which is an ongoing larger project within the research group. It was therefore required that the PCR and sequencing assay was developed into something more cost- and time- effective to carry out the assay on a larger cohort. A power calculation revealed that 385 individuals would be required.

2.3.1 PCR multiplexes

The 13 external primers pairs were separated into four different multiplexes, of three to four primers each. The multiplexes were selected based on three criteria: all primers within the multiplex must not have unacceptable interactions with each other; the amplicons of the primers must produce products differing by at least 30 base pairs so that the products can be distinguished on an agarose gel; the primers must produce amplicons that are associated with internal primers which could also be multiplexed together, preferably with the products of one other PCR multiplex as well. Autodimer v1(http://strbase.nist.gov/AutoDimerHomepage/AutoDimerProgramHomepage.htm) was used to assess the multiplex primer interactions, with the same criteria used as previously used to assess the single primer pairs [64]. The multiplex AIM sets can be seen in Table 3.

2.3.2 SNaPshot® Multiplexes

The 13 internal primers were separated into two different multiplexes, with AIMs from PCR multiplex 1 and 2 being used in SNaPshot® multiplex 1, and the AIMs from PCR multiplex 3 and 4 being used in SNaPshot® multiplex 2. The AIMs in the PCR multiplexes had to be considered when constructing the SNaPshot® multiplexes because the PCR products from the PCR multiplexes are used as the template for the SNaPshot® reactions. Other considerations included heterodimer interactions between the primers, assessed using Autodimer; as well as the potential binding of internal primers non-specifically to PCR amplicons other than their assigned one [64]. This was assessed by using Blastn (www.blast.ncbi.nlm.nih.gov) to assess similarities between the internal primers and the PCR amplicons. If binding would occur at the 3' end of the SNaPshot® primer, this was considered unacceptable.

Table 3: AIMs grouped into PCR Multiplex reactions

	***Gene* (SNP)**	**PCR amplicon size (bp)**	**SNaPshot PCR amplicon size (bp)**
SNaPshot® multiplex 1			
PCR multiplex 1	*SLC24A5* (rs1426654 A/G)	114	46
	OCA2 (rs1800414 G/A)	258	31
	HERC2 (rs1129038 G/A)	380	16
PCR multiplex 2	*PROCR* (rs2069945 G/C/T)	95	53
	OCA2 (rs1800407 C/T)	254	27
	SLC45A2 (rs26722 G/A)	374	39
	OCA2 (rs1800401 C/T)	469	35
SNaPshot® multiplex 2			
PCR multiplex 3	*TLR1* (rs4540055 A/C/T)	76	62
	ADAM17 (rs1524668 A/C)	341	48
	AFG3L1 (rs4785763 G/T)	658	27
PCR multiplex 4	*SLC45A2* (rs16891982 G/C)	128	41
	TYR (rs1126809 C/T)	341	34
	ASIP (rs6058017 C/T)	586	16

2.3.3 PCR Optimisation

Once the external primer pairs had been optimised in singleplex, they were optimised in multiplex using the same reaction mix setup, with the total concentration and volume of the forward and reverse primer mixes for each multiplex being equal to the total concentration of the singular primers in singleplex, and all primers in the mix being of equal proportions. The previously mentioned initial PCR conditions were used for the multiplex temperature gradients, with the exception that the gradient was 60-70°C. If an expected band in a multiplex was not produced at all, the temperature gradient was reduced to 55-65°C to reduce the stringency of the annealing conditions, and thereby promote product formation. If an expected product was considerably fainter than the others in a multiplex, the primer ratios in the primer mix were altered to increase the proportion of the primers for that particular product.

The cycling conditions for the multiplexing were later changed from the aforementioned conditions by increasing the number of PCR cycles to 40. New temperature gradients were also performed

using these conditions. This was changed due to low peak heights in the SNaPshot® reactions performed using the products of the former conditions. Increasing the number of PCR cycles was therefore used to increase the number of PCR product templates produced by the multiplex PCRs.

2.3.4 SNaPshot® Primer Design

The AIMs were to be genotyped using the SNaPshot® microsequencing technique (Applied Biosystems, USA). To this end internal primers were designed and assessed using the same tools as the external primers, with the exceptions that the primer sequences were manually selected and not selected using Primer3. The internal primers for *SLC24A5* rs1426654 and *SLC45A2* rs16891982 had previously been designed during the course of the pilot study. Similarly, the primer sequences for *PROCR* rs2069945 and *TLR1* rs4540055 had been designed by de la Puente *et al.* (2016) [61]. These four internal primers were also assessed and adapted. The sequences for all internal primers can be seen in Table B (Appendix 3)

Primers were selected so that a minimum of 14bp would bind directly adjacent to the SNP of interest. This binding sequence was greater if possible, but the length was selected based on the effect of the sequence on homodimer, heterodimers, hairpins and Tm. If the primer was required to be of greater length than the selected binding sequence provided, a poly-A tail was added to the 5' end to achieve the desired length. The addition of a tail sequence also allowed the multiplexes to be created such that products with the same potential SNP identity could have the length of the products adjusted such that they could be properly separated and distinguished during analysis. A poly-A tail was selected for this purpose because the addition of Adenines or Thymines would cause a less drastic change in the primer Tm than the addition of Guanines or Cytosines. Given the large final size difference between the different primers in the multiplex, the use of a poly-C or poly-G tail would have resulted in primer sets with a very broad range in Tms, which would have impeded optimisation.

2.3.5 SNaPshot® Optimisation

Optimisation of the SNaPshot® reaction using the internal primers began with optimising each reaction singularly (i.e. performing the reaction using the appropriate "clean" singleplex PCR product as template and using the corresponding internal primer). "Cleaning" of each PCR product involved the treatment of 5µl of product with 1µl of recombinant Shrimp Alkaline Phosphatase (rSAP) and 0.5ul of Exonuclease I (*ExoI*) enzyme (to a total volume of 10µl) for 37°C for 1hr and 75°C for 15 min. This treatment degrades unused primers and dNTPs present in the PCR product, which could

interfere with any sequencing involving ddNTP extension, therefore including SNaPshot® minisequencing [67].

The SNaPshot® reaction setup was as follows: 1µl "cleaned" PCR product, 1µl of appropriate internal primer (20µM), 1µl of SNaPshot® Reaction Mix and MBG water added to a total volume of 10µl. All set-ups involving the SNaPshot® reaction and products were performed under minimal light conditions. The reaction conditions were: 96°C for 10 seconds, 50°C for 5 seconds and 60 °C for 30 seconds, repeated for 25 cycles. Post reaction clean-up was performed by the addition of 1µl of rSAP to the reaction tube, followed by 37°C for 1hr and 75°C for 15 min. The product was analysed by combining 1µl of cleaned product with 0.2µl of GeneScan™ 120 LIZ® Size Standard (Thermo Fisher Scientific, MA, USA) and 8.8µl of Hi-Di Formamide (Thermo Fisher Scientific, MA, USA), and performing capillary electrophoresis using the 3130 Genetic Analyzer (Applied Biosystems, USA). The electropherograms were analysed using the GeneMapper™ Software v.4.1 (Applied Biosystems, USA).

Optimisation of each of the two SNaPshot® multiplex reactions was initially performed using similar conditions to the singleplex reactions: 0.5µl of each of the two appropriate "cleaned" PCR product, 1µl of appropriate internal primer mix (the mix composed of equal proportions of the appropriate internal primers, 20µM), 1µl of SNaPshot® Reaction Mix and MBG water added to a total volume of 10µl. The reaction conditions, clean-up and analysis were the same as for the singleplex reactions. Analysis revealed that these conditions did not promote high peak heights for the different internal primer products, nor were the products of similar intensity (unequal Relative Fluorescence Units, or RFU).

Various changes to the reactions conditions were attempted to ameliorate these issues including:

- increasing the amount of PCR product in the initial clean-up step by 1.5X
- performing 50-55°C temperature gradients of the primer annealing step of the SNaPshot® multiplex reactions
- increasing the volume of each appropriate cleaned PCR product added to the reaction mix from 0.5µl to 2µl
- increasing the total volume of internal primers added to the reaction mix to 4µl (20µM)
- increasing the volume of SNaPshot® Reaction Mix added to 2µl and
- altering the ratios of the internal primers added to the primer mix for each reaction

Despite these extensive optimisation steps, the SNaPshot® assay was unfortunately not optimised within the timeframe of this project. It was the intention that this assay would be applied to a subset of participants (Section 2.4), however, due to time spent on optimisation, this was not performed.

2.4 Study population and sample taking

2.4.1 Characteristics of the study population

In parallel to the design and optimisation of the assay, a total of 105 volunteers, representing diverse ancestries and ethnicities, were recruited for the prospective, quantitative, cross-sectional aspect of the larger study. These participants were added to the 194 participants previously recruited for this project, making the cohort n=299 to date. The power calculation $N = (P(1-P)Z^2_{95\%})/(d^2)$ performed using the potential population frequencies of the AIMS utilised in the assay revealed that a cohort of 385 individuals would be required to provide the sufficient power to perform analyses of adequate significance. It was aimed that the melanin indices of the cohort would reflect a normal or potentially bimodal distribution along the MI scale. Ethics clearance was obtained for this study from the Faculty of Health Science Human Research Ethics Committee of UCT (HREC number 158/2016) (Appendix 1).

Participants were selected based on these inclusion criteria: they had to be living in South Africa; they had to give informed consent, they had to be over the age of 18; they could not have participated in the pilot study. Exclusion criteria were: participants with skin diseases or pigmentation anomalies were excluded, because their melanin index would not accurately reflect the genetic influence of the AIMs; participants who had used self-tanner, tanning beds or had used skin lightening treatments were excluded for similar reasons.

All individuals were staff or students of the University of Cape Town, primarily on the medical campus. The cohort was recruited from July to August, when their melanin index would be least affected by skin being darkened by sun-exposure. These precautions were taken so that the melanin index reading would most closely reflect constitutive pigmentation.

2.4.2 Demographic Data

Demographic data were collected through the use of a questionnaire (Appendix 2). The questions were explained to the participant and the participant filled in the questionnaire themselves. These variables are shown in Table 4.

Table 4: Variables that were examined in the study

Variable	**Type of Variable**
Age	Numerical
Sex	Binary
South African Official Census Category	Categorical
Eye Colour	Categorical
Hair Colour	Categorical
Ancestral Origin	Categorical
Self-declared Ancestry	Categorical
Father's ethnicity	Categorical
Mother's ethnicity	Categorical
Paternal grandparent's ethnicity	Categorical
Maternal grandparent's ethnicity	Categorical

Melanin Index (MI) readings were obtained from each individual using the DSM II Skin Colormeter (Cortex Technology, Denmark) a spectrometer that measures both erythema and melanin index.

Before taking the readings, the accuracy of the equipment was measured by taking a measurement from a pure white surface provided by the manufacturer. The Red, Green and Blue (RGB) reflectance values were determined to ensure that they were balanced, as should be expected for a white surface. The melanin index and RGB reflectance were measured in duplicate on the inner forearm and inner upper arm, but not over a vein or mole. These regions were less likely to be exposed to the sun, and so more accurately reflect the constitutive pigmentation of the individual. All four readings were averaged and this average value was used in further analyses. A duplicated reading was also taken of the forehead, to provide a means of comparing an area more likely to be exposed to the sun.

2.4.3 Sample Collection

Saline solution was prepared as a 0.9% solution using NaCl and deionised water. 0.9% saline solution is considered to be isotonic with regards to cell contents, so the use of an isotonic solution like 0.9% would not lead to bursting of the cells and premature release of the cell contents. Participants were asked to lightly chew their cheeks, then vigorously swish 10ml of the saline solution around their mouths for 20 seconds as described previously [68-70]. The solution was expectorated into a 50ml tube and kept in ice until extraction. Extraction commenced within an hour of the sample being collected.

2.4.4 DNA Isolation

Extraction was performed as per the isolation protocol for Buccal Cells and Swabs (Rinse Method) provided with the Zymogen G-DNA Miniprep extraction kit (Zymo Research, U.S.A, CA) with two modifications: the initial centrifugation of the 50ml tubes was performed at 2500rpm for 20minutes, since this resulted in more consistent pelleting of cellular material. The initial centrifugation was performed in the Eppendorf 5417C Centrifuge (Eppendorf, USA). Additionally, in the case that there was more than 500µl of the lysis product, 500µl was loaded into the column, it was then centrifuged for 1 minute at 10 000g in the Neofuge 15R centrifuge (Heal Force, China), the rest of the product was loaded into the column and the centrifuge step was repeated. The entirety of the >500µl lysis product was not loaded at once because this could lead to an overfilling of the spin column and consequent leakage during the centrifuge steps. The extracted DNA was used for all downstream assays.

2.4.5 DNA Quantification

The DNA concentration of the extracted samples was assayed by loading 2µl of DNA solution onto the NanoDrop 2000 UV-Vis Spectrophomoter (Thermo Fisher Scientific Inc., USA). This measured the spectrophotometric absorbance of the solution at 260, 280 and 230nm wavelengths to quantify the concentration of nucleic acids, proteins and salts respectively. DNA samples were considered to be of sufficient quantity if they had a DNA concentration of 20ng/µl or greater.

2.4.6 Statistical Analysis

All statistical analysis was performed using the statistical program STATA® 14.0 (StataCorp LP, TX, USA). Normality of the melanin index data as a whole (those recruited in this project and the pilot project) was assessed using the Shapiro-Wilk test, which was subsequently performed on individual groups such as the sexes and population groups. An absence of normality resulted in non-parametric

tests being utilised. The differences between the median MI of the two sexes was assessed using the Wilcoxon Rank-sum test. Differences between the MI of the different populations groups was initially assessed using the Kruskall-Wallis H test, followed by pair-wise assessment of MI differences between individual population groups using the Wilcoxon Rank-sum test.

Chapter 3: Results

3.1 Singleplex PCR Optimization for Sequencing

All PCRs were first optimised in singleplex, using the reagent setup and reaction conditions seen in section 2.7.1.

Fig. 1 shows an example of a singleplex optimisation PCR as visualised on an agarose gel. The PCRs were initially optimised using control DNA samples collected using the saline swish method. The agarose gel image displays the PCR products of the temperature gradient of *TYR* rs1126809, product size 341bp. The level of non-specific amplification decreases as the temperature conditions increase, with lane 5 (61.1°C) showing no non-specific amplification, and increased yield of the *TYR* rs1126809 PCR product compared to subsequent temperatures.

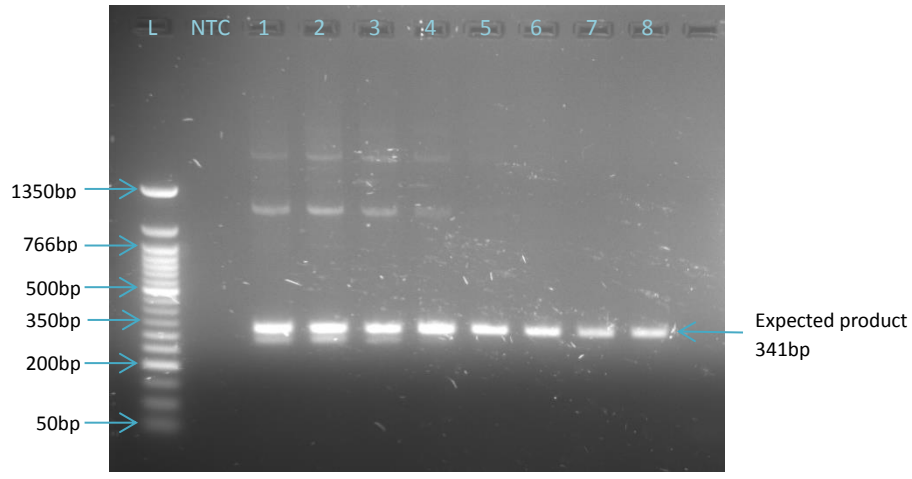

Figure 1: 2% TAE Agarose gel showing the PCR products of the temperature gradient of *TYR* rs1126809. The expected product size is 341bp. NTC is the no template control, while lanes 1-8 contain the products in ascending temperature conditions (55-65°C). Gel electrophoresed at 100V for 1.33h, visualised using EtBr.

All of the 13 AIMs (except *ASIP* rs6058017) were optimised and visualised in the same fashion, revealing the optimum temperatures for each (the temperatures at which no non-specific

amplification and high amplification occurred). *ASIP* rs6058017 showed non-specific amplification under all of the initial conditions however. Further optimisation of *ASIP* rs6058017 initially was attempted by reducing the extension times of the reaction to reduce the amplification of larger, non-specific bands. This failed however, and further optimisation revealed that *ASIP* rs6058017 required a temperature gradient from 60-70°C to reveal its optimum temperature.

Figure 2 and 3 show the PCR products of each of the 13 AIMs amplified at their optimum temperatures. Some faint non-specificity was present in some bands, possibly due to an increased sensitivity of the visualising agent (GelRed) as compared to Ethidium Bromide (EtBr).

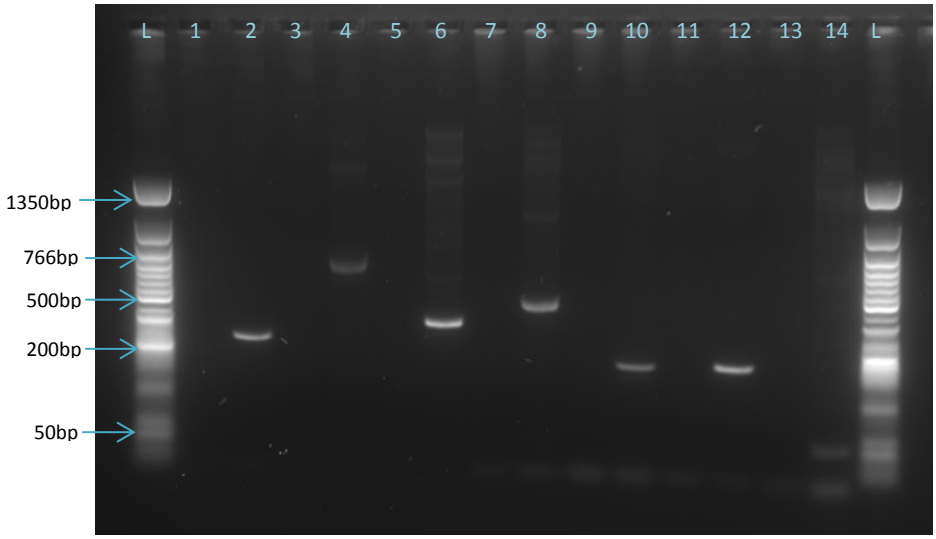

Figure 2: 2% TAE Agarose gel showing 7 PCR products at the optimum temperature of each respective AIM PCR Product. Gel electrophoresed at 120V for 1h, visualised using GelRed. Lanes 1,3,5,7,9,11 and 13 show the NTC for the following lane. Lane 2: *ADAM17* rs1524668, 341bp (61.1°C); Lane 4: *AFG3L1* rs4785763, 658bp (61.1°C); Lane 6: *HERC2* rs1129038, 380bp (64.3°C); Lane 8: *OCA2* rs1800401, 469bp (64.3°C); Lane 10: *OCA2* rs1800407, 254bp (61.1°C); Lane 12: *OCA2* rs1800414, 258bp (63°C); Lane 14: *PROCR* rs2069945, 96bp (64.3°C)

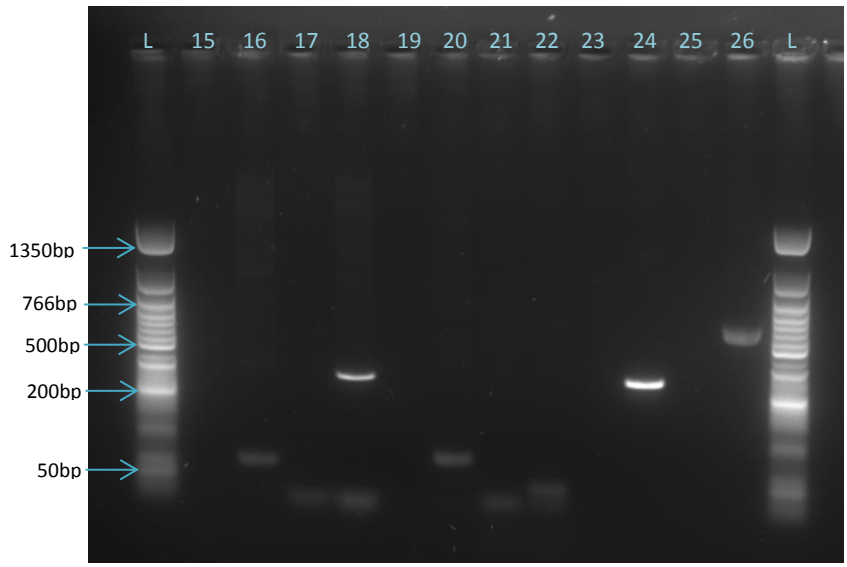

Figure 3: 2% TAE Agarose gel showing 6 PCR products at the optimum temperature of each respective AIM PCR Product. Lanes 15, 17, 19, 21, 23 and 25 show the NTC for the following lane. Lane 16: *SLC24A5* rs1426654, 114bp (56.9°C); Lane 18: *SLC45A2* rs26722, 374bp (61.1°C); Lane 20: *SLC45A2* rs16891982, 128bp (63°C); Lane 22: *TLR1* rs4540055, 76bp (61.1°C); Lane 24: *TYR* rs1126809, 341bp (61.1°C); Lane 26: *ASIP* rs6058017, 586bp (66.1°C).

Once the initial singleplex PCR optimisation had been completed on the control buccal DNA sample, the PCR reactions were attempted using the forensic case DNA sample that had been extracted from blood. Figure 4 shows the amplification of four of the 13 AIMs using the ideal temperatures and conditions that had been determined using the buccal sample. For some of the reactions (e.g. Figure 4, lane 6), non-specific amplification can be seen. This could potentially be caused by difference in DNA sample quality between the buccal and blood DNA. The products in lanes 2 and 4 were considered to be bright enough and specific enough to send for sequencing.

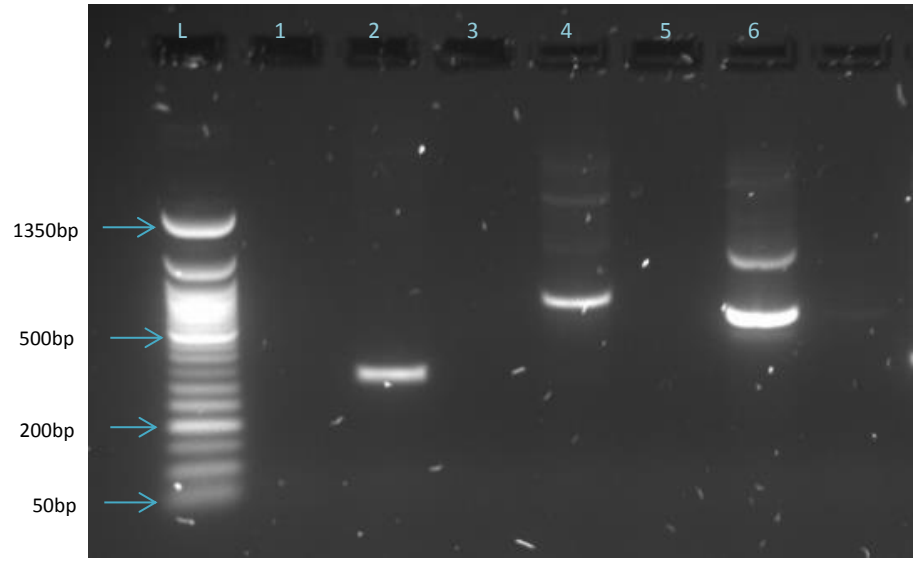

Figure 4: 2% TBE Agarose gel showing 4 PCR products a previously determined optimum temperature of each respective AIM PCR Product. Lanes 1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23 and 25 show the NTC for the following lane. Lane 2: *ADAM17* rs1524668, 114bp (61.1°C); Lane 4: *AFG3L1* rs4785763, 658bp (61.1°C); Lane 6: *ASIP* rs6058017, 586bp (66.1°C); Lane 8: *HERC2* (rs1129038, 380bp (64.3°C).

Those 6 AIMs that did not have satisfactory products to use for sequencing underwent new, four temperature gradients to ascertain the ideal temperatures for them using the new blood sample DNA. Two of these were considered to be bright enough and specific enough to send for sequencing. The four AIMs that still showed unsatisfactory amplification had new, full temperature gradients performed on them using GoTaq® Green Master Mix (Promerga Corporation, USA). All four showed sufficiently good and specific amplification to be sent for sequencing thereafter.

3.2 Sequencing of Case Sample

The clean, bright bands from the aforementioned optimisation process were sequenced. Figure 5 shows the sequencing electropherogram for *HERC2* rs1129038, as an example of a homozygous genotype. The SNP identity is G/G, but since the sequencing was performed using the reverse primer, the true identity is C/C. This sequencing electropherogram showed very little background noise.

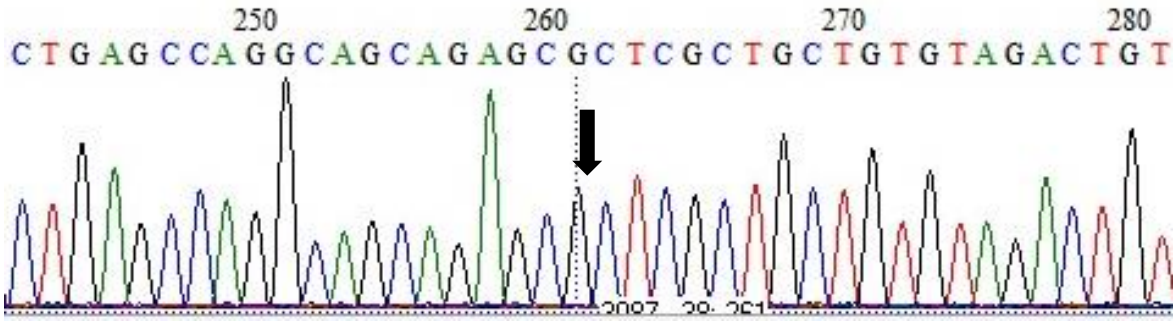

Figure 5: electropherogram showing the sequenced region surrounding *HERC2* rs1129038 G>A. The peak corresponding to rs1129038 is indicated by a black arrow.

Figure 6 shows the sequencing electropherogram for *TYR* rs1126809, as an example of a heterozygous genotype. The SNP identity is R, or G/A, this is one of two heterozygous results. Some of the surrounding sequencing appears to have relatively high levels of background noise, but the true peaks can still be discerned.

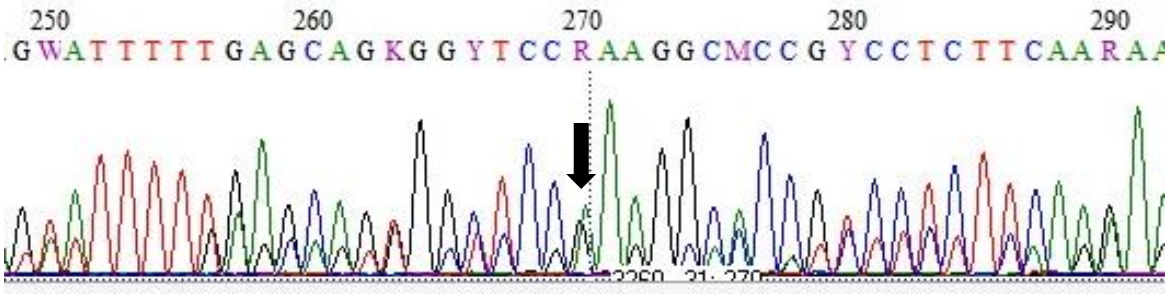

Figure 6: electropherogram showing the sequenced region surrounding *TYR* rs1126809 G>A. The peak corresponding to rs1126809 is indicated by a black arrow.

The sequencing electropherograms for the other sequencing reactions can be found in the appendix (Figures A-J). Only one AIM's sequencing failed completely: *TLR1* rs4540055. The true identity of the SNP could not be discerned due to the messy sequencing results for it. It was not repeated because this AIM is one of three (*ADAM17* rs1524668, *PROCR* rs2069945 and *TLR1* rs4540055) which have not been shown to have associations with pigmentation. Therefore the AIM SNP identity was not needed for the qualitative assessment of the decedent. These three SNPS were included in the assay design because it is hypothesised that these SNPs may have association with pigmentation in South Africa – they do not have predictive value on their own, to our knowledge. To test the hypothesis would require large scale testing in a statistically significant cohort, which did not form part of this study, but will form part of the larger ongoing study. The successful results are shown below in Table 5, along with their previously established associations. The interpretation of these genotypes is discussed in terms of relevant literature in Chapter 4.

Table 5: Genotypes and associations for 10 of the 13 sequenced AIMs

Gene	rs number	Genotype	Phenotype	References
AFG3L1	rs4785763	A/A	A: presence of freckles in European populations	[28,50]
ASIP	*r*s6058017	G/A	G: associated with a higher melanin index, darker skin and darker hair	[26]
*HER*C2	rs1129038	G/G	G: associated with higher MI and darker eye colour	[16,19]
OCA2	rs1800401	C/C	C: associated with lighter eyes and skin	[17,31]
OCA2	rs1800407	G/G	G: associated with lighter eyes	[16,31]
OCA2	rs1800414	A/A	A: associated with a higher MI	[24,29]
SLC24A5	rs1426654	G/G	G: associated with higher MI, darker skin, darker hair and darker eyes	[19,20,24,25,54,55]
SLC45A2	rs26722	C/C	C: associated with lighter eye colour, lighter skin and lighter hair	[13,16,17,30,53]
SLC45A2	rs16891982	C/C	C: associated with darker skin, darker hair, darker eyes, and a higher MI	[54,55]
TYR	rs1126809	G/A	G: associated with higher MI, darker skin	[17,60]

Figure 7 shows part of a DNA profile that was performed on the forensic sample. The entire DNA profile is not shown to preserve anonymity. The marker of relevance to the case study is the Amelogenin marker. The presence of a single, X peak for this marker indicates that the decedent was female. Examining some of the other markers shows additionally that the sample does is not significantly degraded nor contains high levels of PCR inhibitors, due to the absence of the “ski slope effect”. This effect exhibits as the smaller markers having higher peaks than the larger markers, due to the smaller markers replicating more efficiently in degraded DNA than the larger markers.

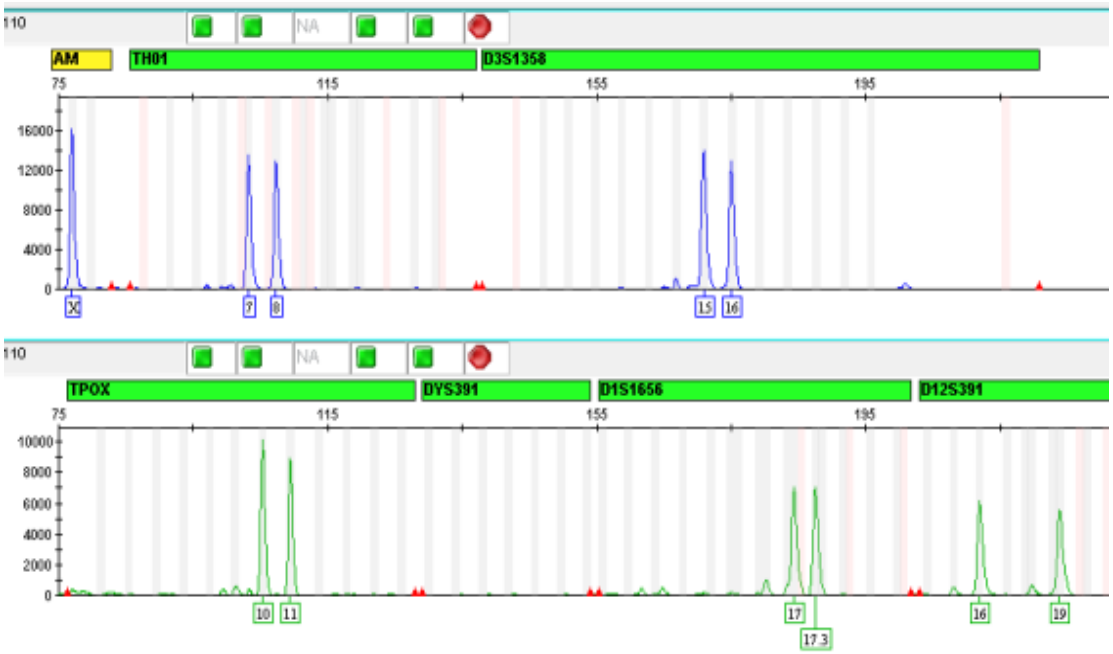

Figure 7: electropherogram showing a portion of the DNA profile of the forensic sample. Amelogenin (AM) is the marker used to determine sex.

3.3 Study Cohort Demographics

A total of 105 individuals were recruited for this ongoing study, which added to the 194 individuals who have already been recruited. The frequencies of variables such as Melanin Index (MI) were determined using the entire cohort.

The average MI values across the combined cohort (Fig. 8) show a primarily positively skewed data distribution, with the majority of the individuals lying in the lower MI ranges.

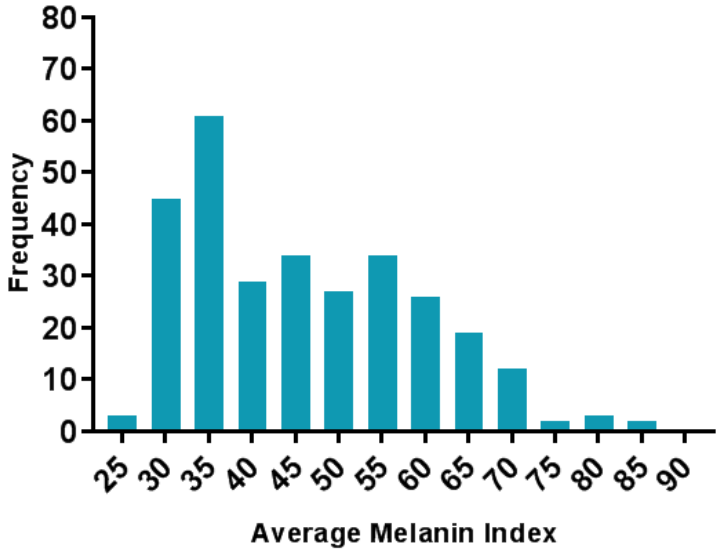

Figure 8: frequencies (number of individuals) of the different average melanin index readings in the combined cohort, the bin size is 5 MI units.

The Shapiro-Wilk test indicated a non-normal data distribution, with a $p<0.001$. Further analyses were therefore performed using non-parametric tests.

Examination of the average melanin index values between the male and female individuals of the combined cohort revealed that males tended to have a higher average melanin index than females. Analysis using the Wilcoxon Rank-sum test revealed a significant difference in MI between the sexes ($p=<0.01$).

Examination of the average MI values of the different population groups in the combined cohort reveal overlapping MI values between all of the groups (Fig. 9). The greatest overlap occurred between the Coloured, Indian/Asian and Other groups.

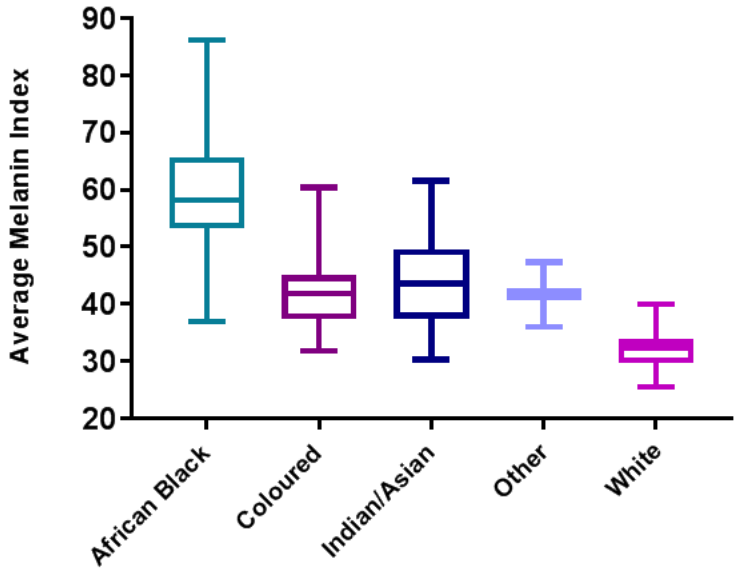

Figure 9: boxplots of the average melanin indexes of the different population groups of the combined cohort.

The Kruskall-Wallis H test revealed that there were significant differences between the population groups in the combined cohort, $p<0.001$. Pairwise Wilcoxon Rank-sum tests between the groups (excluding the 'Other' group, since it only contained 2 individuals), revealed significant differences between individual groups (Table 6). All differences were significantly different even after Bonferroni correction ($p<0.001$). African 67 coloured 29 Indian Asian 28 White 69

Table 6: P-values for the pairwise Wilcoxon Rank-sum tests between population groups

Population Group	African Black	Coloured	Indian/Asian
African Black (n=67)	-	-	-
Coloured (n=29)	P<0.001*	-	-
Indian/Asian (n=28)	P<0.001*	P=0.001*	-
White (n=69)	P<0.001*	P<0.001*	P<0.001*

*** indicates a significant value even after Bonferroni correction.**

3.4 Multiplex PCR optimisation

Optimisation of the PCR multiplex assays was performed in a similar fashion to the singleplexes, with two exceptions: the initial temperature gradients extended from 60-70°C (to increase the stringency of the binding conditions); instead of a single forward and reverse primer being added, a forward primer mix and reverse primer mix were added (1ul of each), with the relevant primers being added to the mixes in equal ratios.

Figure 10 shows a temperature gradient for PCR Multiplex 1, containing *SLC24A5* rs1426654 (114bp), *OCA2* rs1800414 (258bp) and *HERC2* rs1129038 (380bp). It shows a temperature gradient for Multiplex 1 that extends from 55-65°C. While amplification of all 3 bands is present in lanes 1-4, non-specific amplification is also present. Lanes 5-8 show no non-specificity, but SLC24A5 rs1426654 (114bp) is not amplified.

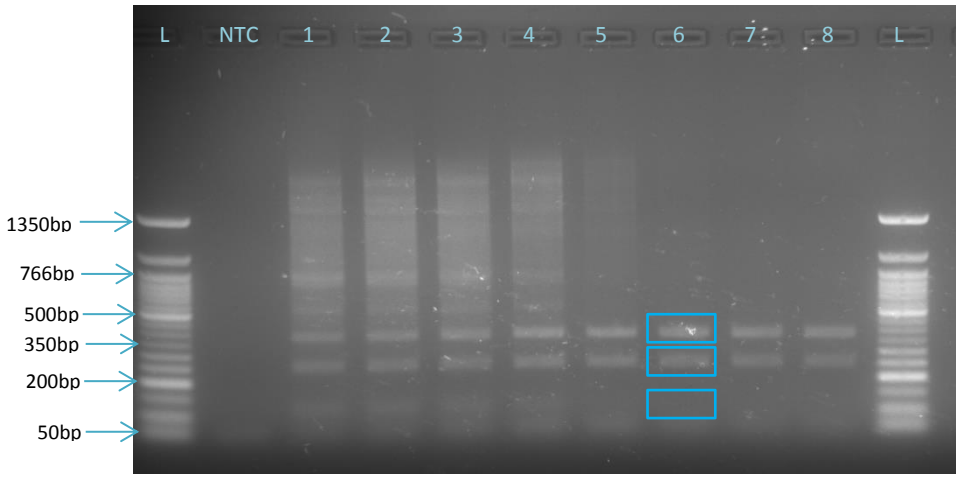

Figure 10: 2% TAE Agarose gel showing the PCR products of the temperature gradient of Multiplex 1. This multiplex contains *SLC24A5* rs1426654 (114bp), *OCA2* rs1800414 (258bp) and *HERC2* rs1129038 (380bp). The temperature gradient extends from 55-65°C. Gel electrophoresed at 120V for 1h, visualised using EtBr.

Further optimisation of Multiplex 1 was attempted by reducing the extension times of the reaction to reduce the amplification of larger, non-specific bands. This failed however, since this alteration reduced *SLC24A5* rs1426654 amplification further as well.

Figure 11 shows the next step in the optimisation of Multiplex 1. The extension time was returned to the default (30 seconds), but the ratios of the primer mixes were altered: the initial ratio of *SLC24A5* rs1426654 F/R: *OCA2* rs1800414 F/R: *HERC2* rs1129038 F/R of 1:1:1 was changed to 2:1:1, while keeping the same volume. Lane 2 in Fig. R10 (59°C) shows the presence of all 3 expected products with no non-specific bands.

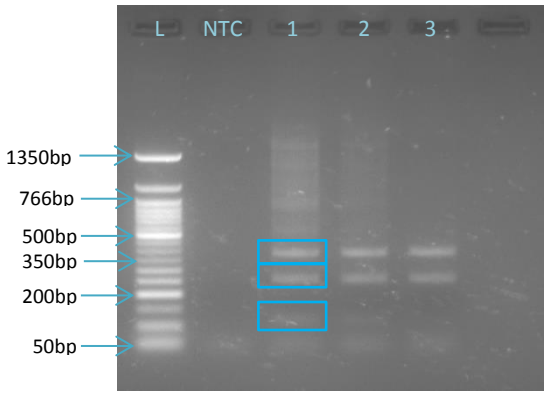

Figure 11: 2% TAE Agarose gel showing the PCR products of a partial temperature gradient of Multiplex 1, after altering primer ratios. This multiplex contains *SLC24A5* rs1426654 (114bp), *OCA2* rs1800414 (258bp) and *HERC2* rs1129038 (380bp). Lane 1: 56.9°C, lane 2: 59°C, lane 3: 61.1°C. Gel electrophoresed at 120V for 1h, visualised using EtBr.

Figure 12 shows a temperature gradient for Multiplex 2, containing *PROCR* rs2069945 (95bp), *OCA2* rs1800407 (254bp), *SLC45A2* rs26722 (374bp) and *OCA2* rs1800401 (469bp). The temperature gradient extended from 60-70°C and amplification of all four bands is present in lanes 1-3, with no non-specific amplification present. Lanes 5-8 show no non-specificity, but *SLC24A5* rs1426654 (114bp) is not amplified. Lane 1 (60°C) was selected as the optimum temperature for amplification.

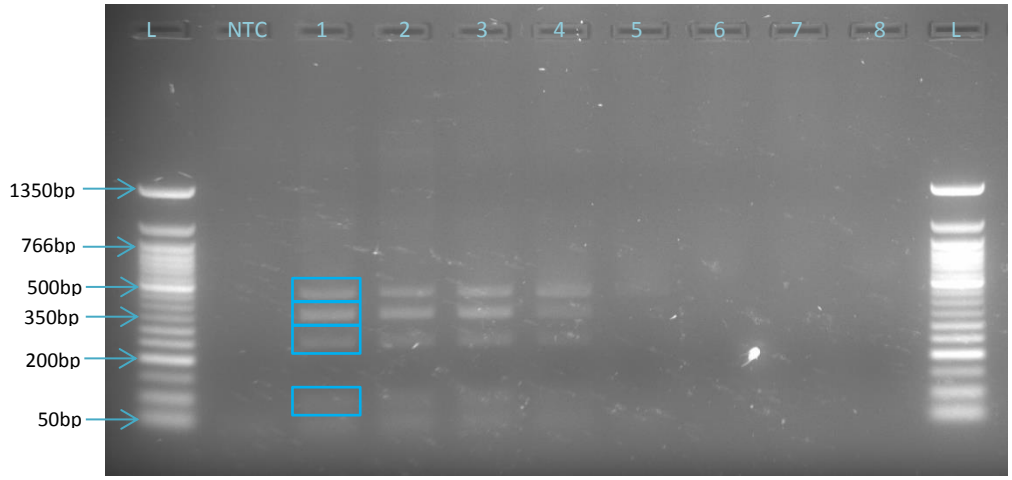

Figure 12: 2% TAE Agarose gel showing the PCR products of the temperature gradient of Multiplex 2. This multiplex contains *PROCR* rs2069945 (95bp), *OCA2* rs1800407 (254bp), *SLC45A2* rs26722 (374bp) and *OCA2* rs1800401 (469bp). The temperature gradient extends from 60-70°C. Gel electrophoresed at 120V for 1h, visualised using EtBr.

Figure 13 shows a temperature gradient for Multiplex 3, containing *TLR1* rs4540055 (76bp), *ADAM17* rs1524668 (341bp) and *AFG3L1* rs4785763 (658bp). The temperature gradient extends from 60-70°C and amplification of all 3 bands is present in lanes 1-2, with no non-specific amplification present. Lanes 5-8 show no non-specificity, but not all bands are amplified. Lane 1 (60°C) was selected as the optimum temperature for amplification.

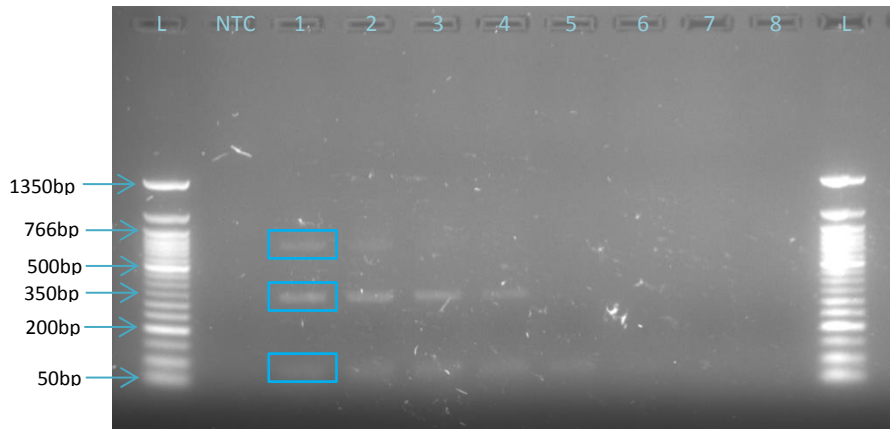

Figure 13: 2% TAE Agarose gel showing the PCR products of the temperature gradient of Multiplex 3. This multiplex contains *TLR1* rs4540055 (76bp), *ADAM17* rs1524668 (341bp) and AFG3L1 rs4785763 (658bp). The temperature gradient extends from 60-70°C . Gel electrophoresed at 120V for 1h, visualised using EtBr.

Figure 14 shows the temperature gradient for Multiplex 3, containing *SLC45A2* rs16891982 (128bp), *TYR* rs1126809 (341bp) and *ASIP* rs6058017 (586bp). The temperature gradient extends from 60-70°C and amplification of all 3 bands is present in lanes 1-3, with no non-specific amplification

present. Lanes 5-8 show no non-specificity, but not all bands are amplified. Lane 1 (60°C) was selected as the optimum temperature for amplification.

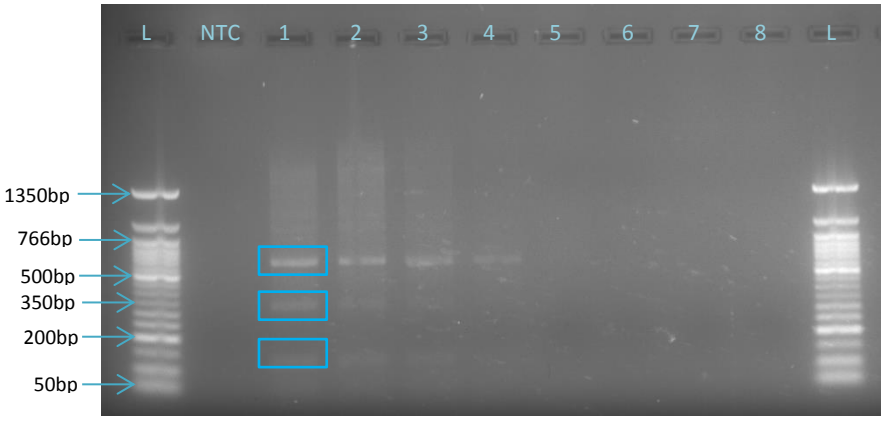

Figure 14: 2% TAE Agarose gel showing the PCR products of the temperature gradient of Multiplex 4. This multiplex contains *SLC45A2* rs16891982 (128bp), *TYR* rs1126809 (341bp) and *ASIP* rs6058017 (586bp). The temperature gradient extends from 55-65°C. Gel electrophoresed at 120V for 1h, visualised using EtBr.

Attempts to use the aforementioned optimised PCR products in SNaPshot® reactions revealed that a higher concentration of products was required. This was attempted by increasing the number of cycles of the PCRs from 30 to 40, so that more product amplicons could potentially be produced. Increasing the number of cycles lead to increase in nonspecific amplification, as can be seen in Figure R15, showing the first attempt at a 40 cycle PCR Multiplex 1, and the individual amplicons that comprise it. There did not appear to be increased non-specificity in the PCR multiplex 1 lane (Fig 15, lane 8) however.

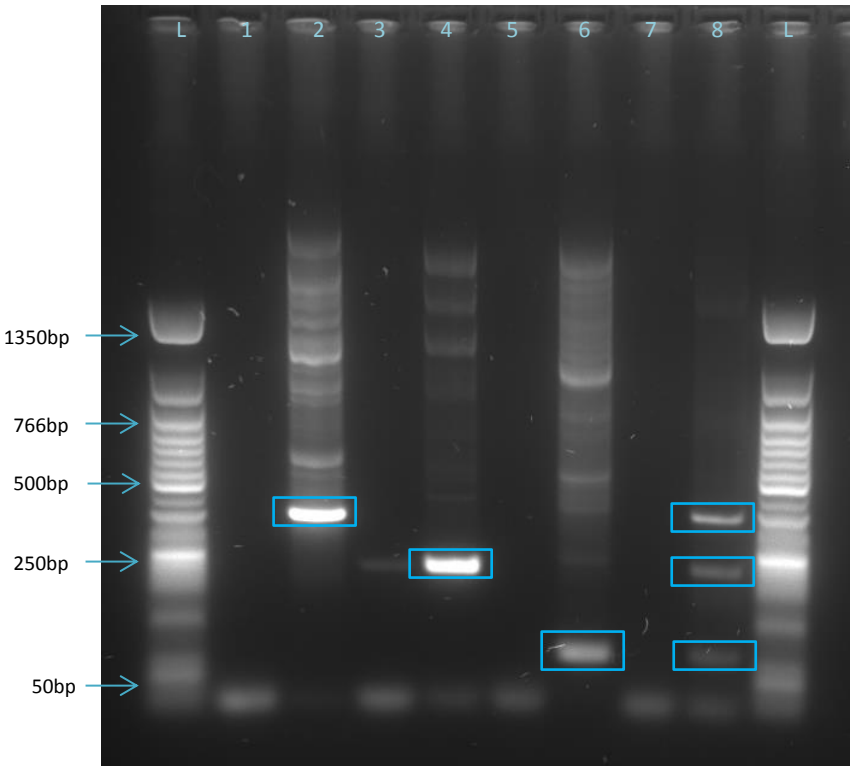

Figure 15: 2% TAE Agarose gel showing the PCR products of the individual three amplicons of *HERC2* rs1129038 (380bp, lane 2), *OCA2* rs1800414 (258bp, lane 4) and *SLC24A5* rs1426654 (112bp, lane 6) as well as PCR Multiplex 1 in lane 8 (comprising of those three amplicons). Gel electrophoresed at 120V for 1h, visualised using GelRed (Biotium, CA. U.S.A).

Since the PCR multiplex had not shown non-specificity using a 40 cycle reaction, all for of the PCR Mutiplexes were attempted using the 40 cycle PCR reaction. As can be seen in Figure 16, PCR Multiplexes 1 and 2 showed their relevant amplicons (3 and four amplicons respectively), with little non-specificity. Multiplex 3 and 4 showed higher non-specificity, and highly unequal PCR band brightness'.

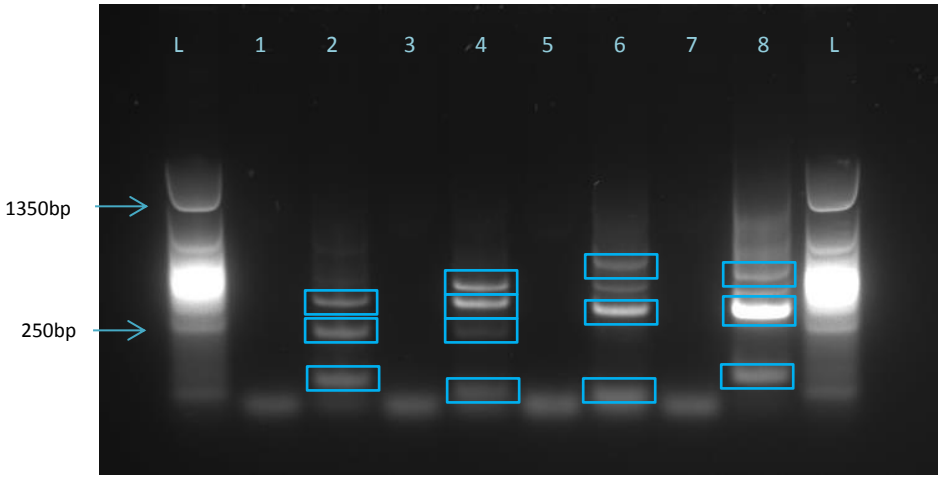

Figure 16: 2% TAE Agarose gel showing the PCR products of the four multiplexes: PCR Multiplex 1 (lane 2), PCR Multiplex 2 (lane 4), PCR Multiplex 3 (lane 6) and PCR Multiplex 4 (lane 8). Lanes 1, 3, 5 and 7 are NTC reactions. Gel electrophoresed at 120V for 1h, visualised using GelRed (Biotium, CA. U.S.A).

Figure 17 shows the four PCR multiples performed under varying conditions: using the same primer ratios seen in Figure R16, but with a 50µl reaction volume (lanes 1, 5, 9 and 14); the NTC reaction (lanes 2, 6, 10 and 13); using different primer ratios, with 25µl reaction volume (lanes 3, 7 and 11); and using different primer ratios, with 50µl reaction volume (lanes 4, 8, 12 and 14). Multiplex 4 only had a double reaction volume variant performed. The products in lane 1, 5, 11 and 14 were considered to be the best.

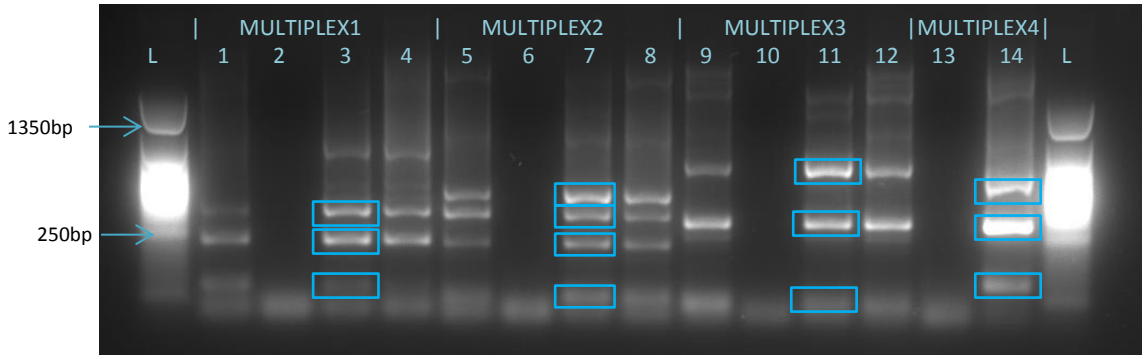

Figure 17: 2% TAE Agarose gel showing the PCR products of the four multiplexes under varying reaction conditions: PCR Multiplex 1 (lane 1-4), PCR Multiplex 2 (lane 5-8), PCR Multiplex 3 (lane 9-12) and PCR Multiplex 4 (lane 13-14). Lanes 2, 6, 10 and 13 are NTC reactions. Gel electrophoresed at 120V for 1h, visualised using GelRed (Biotium, CA. U.S.A).

Figure 18 shows the shows a PCR reaction for PCR Multiplex 2, with the ratio of the *PROCR* rs2069945 primers in the primer mix increased relative to the ratios seen in lane 5 of Fig. 17. This change was made to increase the brightness of the *PROCR* rs2069945 amplicon (95bp) relative to the other three amplicons.

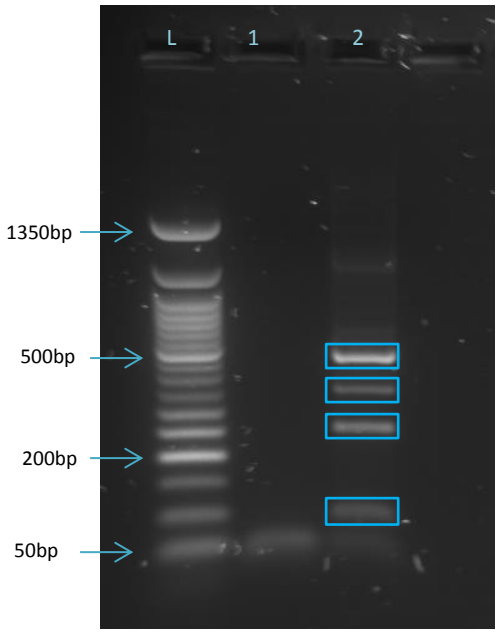

Figure 18: 2% TAE Agarose gel showing the PCR products Multiplex 2, with an altered primer ratio to attempt to increase the brightness if the smallest amplicon. Gel electrophoresed at 120V for 1h, visualised using Sybasafe (Thermofisher, MA, USA).

3.5 SNaPshot® optimisation

The SNaPshot® reactions were performed and optimised using the products of PCR Multiplex 1 and 2 (SNaPshot® Multiplex 1) and PCR Multiplex 3 and 4 (SNaPshot® Multiplex 2). Initial reactions revealed two primary issues: 1) the peaks of the SNaPshot® products in general were low; 2) certain products were exceedingly high relative to the other peaks present.

Improving the low general peak height was attempted using multiple different approaches: altering the primer ratios in the PCRs, to increase the PCR amplicon concentrations of the SNaPshot® products that were exceedingly low (as seen in section 3.4); increasing the amount of PCR product in the initial rSap and ExoI cleanup step from 5µl to 7.5µl, to increase the template amplicons for the SNaPshot® reaction; increasing the volume of each PCR multiplex product added to the relevant SNaPshot® multiplex from 1µl to 2µl; to increase the template amplicons for the SNaPshot® reaction; and to increase the amount of SNaPshot® reaction product undergoing capillary electrophoresis from 1µl to 3µl to maximise the SNaPshot® reaction product and thereby increase the peak heights in the electropherograms. Decreasing the disproportionately high peak heights was primarily attempted by altering the primer ratios in the SNaPshot® reaction such that products with high peak heights had a decreased primer concentration.

Figure 19 shows an example of a singleplex SNaPshot result for *HERC2* rs1129038 that had a heterozygous result, as can be seen by the blue and green peaks.

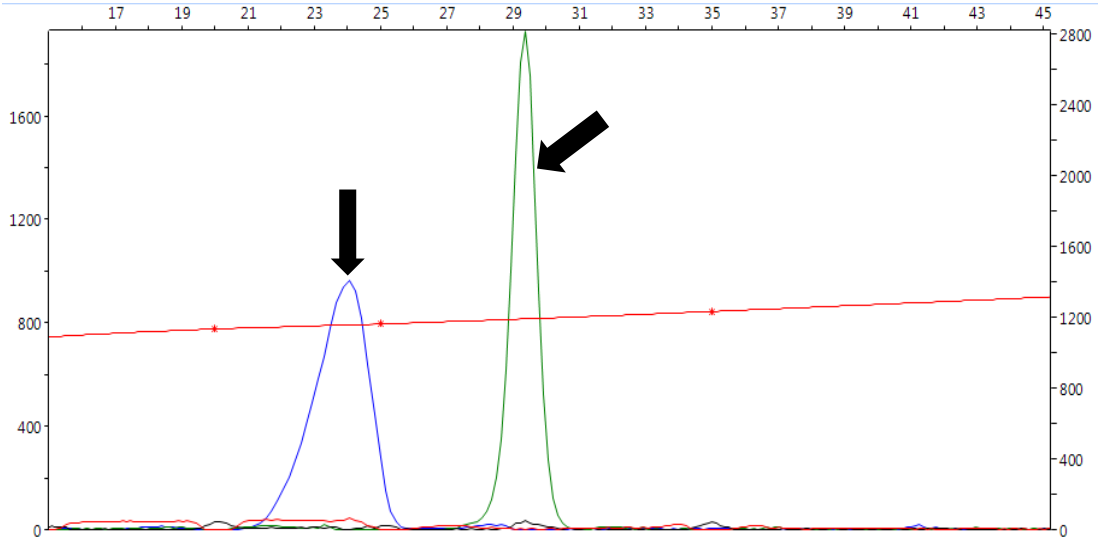

Figure 19: electropherogram of a SNaPshot® *HERC2* rs1129038 singleplex reaction. The arrows indicate the two peaks present, a heterozygous result.

Figure 20 shows an example of a singleplex SNaPshot result for SLC24A5 rs1426654 that had a homozygous result, as can be seen by the singular blue peak.

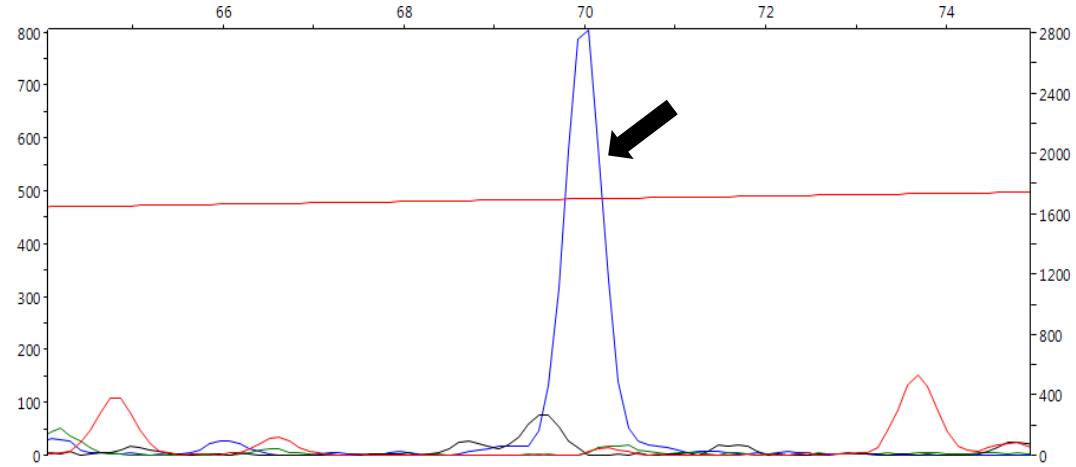

Figure 20: electropherogram of a SNaPshot® *SLC24A5* rs1426654 singleplex reaction. The arrows indicate the singular peak present, a homozygous result.

The scope of the minor dissertation project and time constraints did not allow for the SNaPshot® Multiplex reactions to be completely optimised. The SNaPshot® Multiplex results that have most optimised as far as constraints have allowed (using the best PCR products in R17 and 18) can be seen in Figures 21-24 (SNaPshot® Multiplex 1) and Figures 25-26 (SNaPshot® Multiplex 2)

Figure 21 shows the first portion of SNaPshot® Multiplex 1, consisting of the SNaPshot® amplicons of *HERC2* rs1129038 and *OCA2* rs1800407. One of each of the expected peaks is present for both *HERC2* rs1129038 and *OCA2* rs1800407, indicating that the second peak for each has dropped out. More optimisation of the SNaPshot® primer ratios is clearly required for these amplicons.

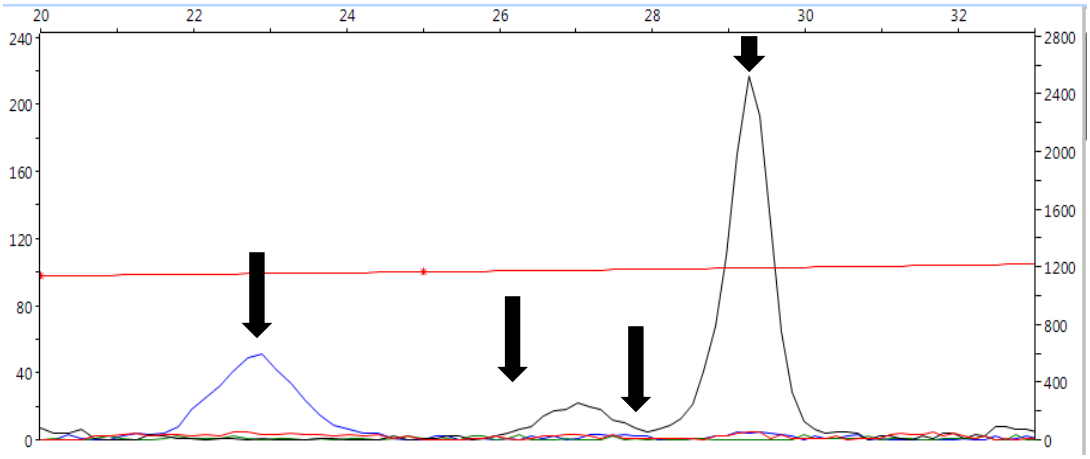

Figure 21: partial electropherogram of SNaPshot® Multiplex 1, with arrows indicating peaks (or where the peaks usually are previously were present) of the following: *HERC2* rs1129038, *HERC2* rs1129038, *OCA2* rs1800407 and *OCA2* rs1800407 respectively

Figure 22 shows the second portion of SNaPshot® Multiplex 1, consisting of the SNaPshot® amplicons of *SLC45A2* rs26722, *OCA2* rs1800414 and *OCA2* rs1800401. All of the peaks for the three SNPs shown here are present, though all (especially the peaks for *OCA2* rs1800401) are low.

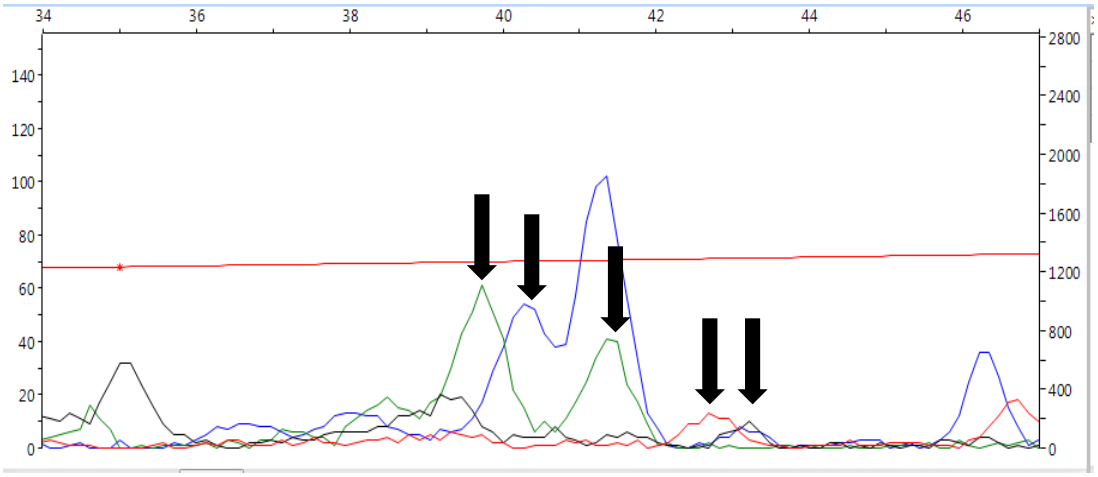

Figure 22: partial electropherogram of SNaPshot® Multiplex 1, with arrows indicating peaks (or where the peaks usually are previously were present of the following: *SLC45A2* rs26722, *OCA2* rs1800414, *OCA2* rs1800414, *OCA2* rs1800401 and *OCA2* rs1800401 respectively.

Figure 23 shows the third portion of SNaPshot® Multiplex 1, consisting of the SNaPshot® amplicons of *PROCR* rs2069945. Both expected peaks for *PROCR* rs2069945 are present, indicating a heterozygous result. The electropherogram surrounding the two peaks is relatively messy.

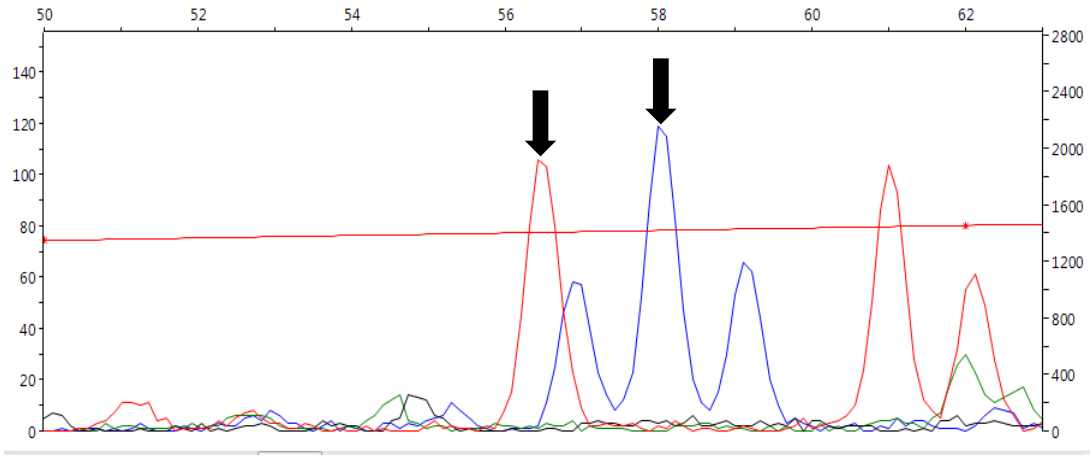

Figure 23: partial electropherogram of SNaPshot® Multiplex 1, with arrows indicating peaks (or where the peaks usually are previously were present of the following: *PROCR* rs2069945 and *PROCR* rs2069945.

Figure 24 shows the fourth portion of SNaPshot® Multiplex 1, consisting of the SNaPshot® amplicon of *SLC24A5* rs1426654. There is a single, high RFU peak present, indicating a homozygous result.

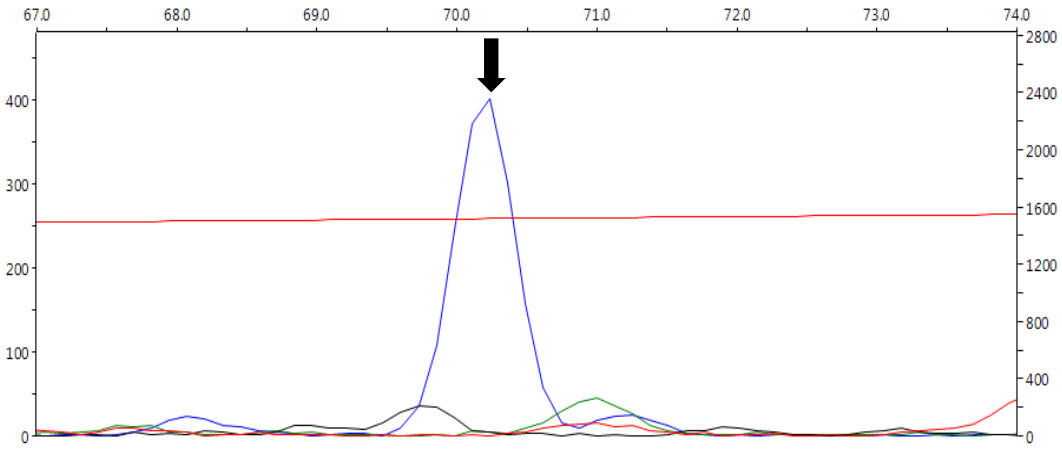

Figure 24: partial electropherogram of SNaPshot® Multiplex 1, with arrows indicating peaks (or where the peaks usually are previously were present of the following: *SLC24A5* rs1426654.

Figure 25 shows the first portion of SNaPshot® Multiplex 2, consisting of the SNaPshot® amplicons of *ASIP* rs6058017, *TYR* rs1126809 and *AFG3L1* rs4785763. Only the expected peak for *AFG3L1* rs4785763 is present, indicating the need for further optimisation of the SNaPshot® primer ratios is required.

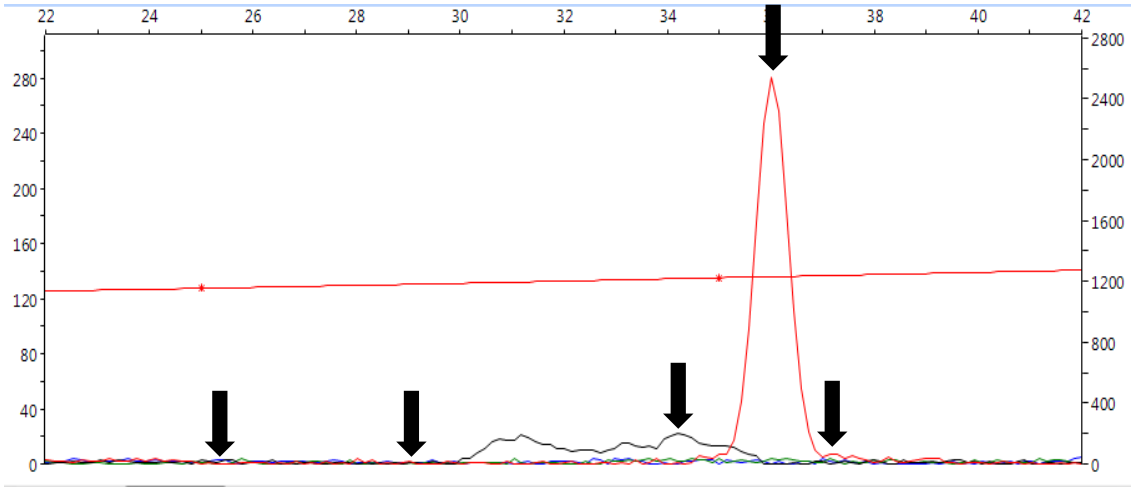

Figure 25: partial electropherogram of SNaPshot® Multiplex 2, with arrows indicating peaks (or where the peaks usually are previously were present of the following: *ASIP* rs6058017, *ASIP* rs6058017, *TYR* rs1126809, *AFG3L1* rs4785763 and *TYR* rs1126809 respectively.

Figure 26 shows the second portion of SNaPshot® Multiplex 2, consisting of the SNaPshot® amplicons of *SLC45A2* rs16891982, *ADAM17* rs1524668 and *TLR1* rs4540055. While all expected peaks were present, the electropherogram is messy, potentially indicating the need for an improved clean-up protocol.

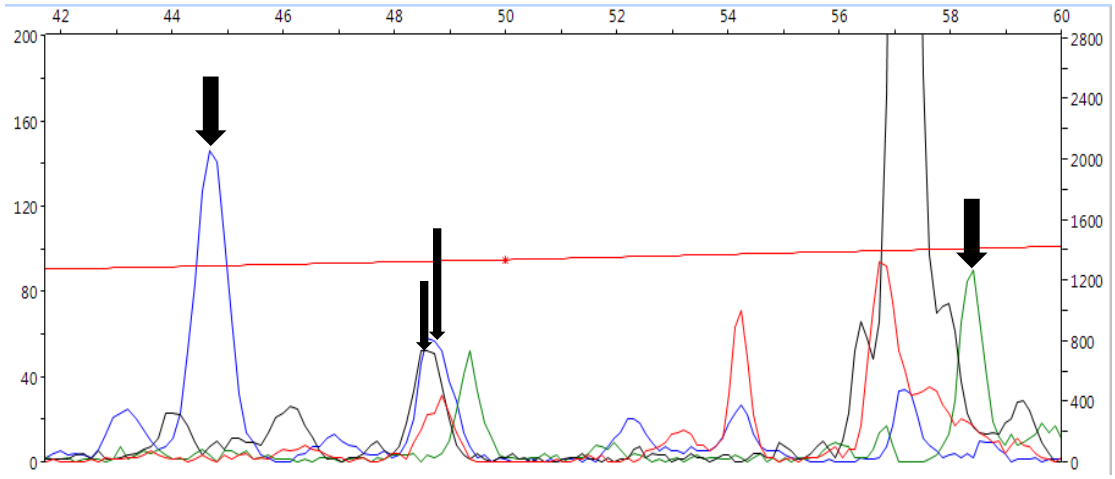

Figure 26: partial electropherogram of SNaPshot® Multiplex 2, with arrows indicating peaks (or where the peaks usually are previously were present) of the following: *SLC45A2* rs16891982, *ADAM17* rs1524668 and *TLR1* rs4540055 respectively.

It is clear that further optimisation is required for both SNaPshot® Multiplex reactions, but that work (and the genotyping of the cohort),will be continued in future projects.

Chapter 4: Discussion

4.1 Identification of the deceased

Salt River Mortuary is one of seventeen mortuaries in the Western Cape, and investigates 3500 – 4000 cases per year. Among these cases, approximately 5-10% of deceased individuals remain unidentified (in-house data, unpublished). Some of these individuals are beyond recognition; for example, those that are burnt, decomposed or skeletonised. Attempts to identify these individuals have been made through techniques such as anthropological methods, DNA profiling, analysis of dental records and the use of soft biometric identifications such as scars, marks and tattoos [71] but the majority of these individuals have remained unidentified. As such, there is a strong motivation to explore alternative methods to identify these individuals.

One such method is molecular phenotyping, which provides a possible additional test to predict externally visible characteristic (EVCs) from the DNA of a donor [72]. The estimation of appearance that this sort of test provides is not intended to represent the characteristics of the individual precisely, but rather gives an idea of what the individual may have looked like. This may narrow done the search criteria by excluding EVCs outside the possibilities predicted by the test, or provide an investigative lead, which could be verified by conventional DNA profiling or other identification techniques [73].

Many studies regarding molecular phenotyping have been conducted in populations of European and Asian Ancestry [17,18,24,29,30,50,52,74,75], however, few have been conducted in admixed populations, such as South Africans. Some studies in admixed populations have even shown that some traditionally used AIMs were not informative in their populations [60]. Thus, there is a need to explore the informative value of AIMs in the South African context, and particularly in a post-mortem forensic context. This study entailed the development of an assay to sequence 13 AIMs and apply it to one case example, for which ethics approval was given, to demonstrate a proof of concept.

4.2 Qualitative assessment of the pigmentation phenotypes of the case study:

The sequencing results of the 13 amplicons provided the data from which to interpret the genotypes for the respective AIMs (see section 3.2). However, these genotypes needed to be compared to the literature to estimate the ancestry and externally visible characteristics of this individual.

The results of the genotyping were somewhat dispersed between phenotypes, for example, for eye colour and skin tone, the association of the relevant alleles were equally balanced between lighter and darker pigmentation phenotypes. However, a qualitative assessment was performed, using the assumption that each of the SNPs had an equal propensity to affect the pigmentation features it was associated with. An assessment performed in this fashion estimated the following information with regards to the decedent: the majority of the AIM alleles were associated with darker hair, so a hair colour ranging from medium to dark brown would be expected. The alleles of the AIMs were equally balanced with regards to associations with lighter and darker eye colour, thus a medium toned eye colour (likely green, hazel or light brown) could be estimated.

The alleles were also equally balanced between being associated with lighter and darker skin colour. However, studies that utilise "skin colour" as a metric for pigmentation often have broad categories for this variable, such as the nine point skin tone scale used by Gravlee & Dressler (2005), therefore at best one could assume an "intermediate" skin tone, therefore neither very light, nor very dark [76]. The AIM alleles that have associations with MI, favour a higher MI. Unlike skin tone classifications, MI is an exact measurement of skin pigmentation, so an association with higher MI does not necessarily specify a large difference in skin tone. For example, a significant difference in MI within a population may result from as small a difference as 2 MI units [77].

Lastly, the STR profiling results provided the additional information that the decedent was female. Overall, given these genotypes and their previous associations, the female decedent was anticipated to be from Mixed Ancestry, have light to dark brown hair, hazel eyes and an intermediate skin tone, but tending towards the darker range of such a category (given the associations with higher MI). This interpretation was made by the researchers, who were blind to the results of the forensic autopsy. After the genetic analysis was complete, the results from this study were compared to the notes of the forensic pathologist who performed the autopsy, who classified the neonate as a "Coloured" female. The interpretations were in agreement with the assessment of the pathologist, which suggests that the AIMs utilised in this study have potential relevance within the South African population. However, the information from the genetic analysis provided far richer detail with regards to the externally visible appearance of the neonate. This is important as the identification of many deceased individuals is challenging due to post-mortem changes and decomposition, and in the case of neonates, due to scavenging [78]. As such, there are many instances where the sex, skin tone, eye colour and ancestry of an individual cannot be discerned at autopsy.

In the context of the neonate used for the case study, the phenotypic predictions provided by the assay would not necessarily bring us closer to identifying the decedent, mainly because the decedent

was partly recognisable, but also so young and did not have any 'differentiating' features yet. In a context in which the decedent was badly decayed, burnt or skeletonised however, a prediction such as the one offered using our assay could provide valuable information regarding appearance. One must however take into account the demographics of the area in which the neonate was found. The predicted phenotypes of the decedent were consistent with the assessment of the pathologists that the decedent was 'Coloured'. The predicted EVCs could be shared by a number of individuals even in a general context (given the broadness of the predicted phenotypes), this is of greater certainty in the specific context of Cape Town, in which 42.4% of the population share the designation of 'Coloured' (www.statssa.gov.za). Perhaps in another province where 'Coloured' are not as common, this would have been more useful information.

This does not mean that the assay would not be of value in other contexts however. If, for example, the decedent was older, even a broad characterisation of their appearance could narrow down the search for their identity (such as through searching missing persons cases). Additionally, a predicted phenotype that is rare in the geographical context that a decedent is found would be very informative (for example, a decedent with blond hair and blue eyes in a primarily African community), as it would elicit fewer comparisons to potential identities. An important aspect to consider is that no form of evidence should ever be utilised in isolation. A broad phenotypic assessment in conjunction with anthropological information, personal belongings etc. could create a relatively complete image regarding a person's identity.

4.3 The need for a low cost, objective prediction model

During the interpretation of genotypes, it was assumed that each allele had equal contribution towards the eye, skin and hair colour (as relevant to each SNP) of the individual. This was a limitation, and it was noted that the exact magnitude of the effect of the alleles of each AIM on pigmentation could not be determined without modelling the independent effect of each allele in a greater sample size.

Further, it was noted that some of the alleles utilised had previously been associated with broad, subjective categories, such as 'darker skin tone', and not an objective measurement such as MI. While models have been developed to predict appearance based on large numbers of SNPs, these assays rely on technology such as NGS and genotyping chips, and these resources are not readily available in developing countries. Some assays have been developed that utilises smaller numbers of SNPs, but these often focus on one or two pigmentation phenotypes, such as the HIrisPlex assay

which predicts eye and hair colour [12]. Furthermore, the informative value of these SNPs which show association in homogenous countries, has been questioned in admixed populations [60]. As such, to further explore the value of these 13 AIMs in the South African population, the relationship between genotypes and objective MI was investigated.

To this end, 104 South African individuals were recruited for this research study, which added to the existing cohort of 195 South African individuals who had also been recruited for study by other researchers in our group, and who had given consent to be used for related research. Therefore, the recruitment of individuals for the total cohort is ongoing. From each participant collected thus far, who gave informed consent, demographic information was obtained, along with an objective MI reading and a saliva mouth rinse sample.

Statistical analysis of the demographic data revealed that there were significantly different average Melanin Indices between the different population groups. However, as can be seen in Figure 9, there is an overlap in the MI values of all of the different population groups (as defined by census population categories). This provides evidence that, within South Africa, the use of population groups as a proxy for describing pigmentation phenotypes may not be informative, since each population group encompasses such a large range of potential MI values, despite significant differences in average Melanin Index. This suggests a movement towards more objective descriptions of pigmentation (such as MI) would be helpful in general forensic usage. Precise objective measures such as MI have the added benefit in that they cannot be interpreted erroneously by different individuals because the measure is strictly defined.

To genotype the 13 AIMs in the larger cohort, it was decided that singleplex PCR and Sanger sequencing was not the most feasible method in terms of cost and time; as such, a multiplex PCR and multiplex SNaPshot assay was designed. Four multiplex PCR assays were successfully optimised in this regard and SNaPshot primers were designed. Unfortunately, the two multiplex SNaPshot PCR assays were not fully optimised within the scope of this project, but since this project is part of a larger ongoing project, this work will be continued in the future.

4.3.1 Design and Optimisation of Genotyping Assays

The design of a functional genotyping assay of any sort is dependent on multiple factors, including the stringency of the bioinformatic primer design, and thorough optimisation of the assays. Both the design and optimisation aspects of this project encountered various impediments in the course of the research, some of which will be discussed below.

One of the difficulties that arose during optimisation was due to the use of forensic samples, in that there are multiple types (buccal, blood etc.) that may not amplify equally well in PCR assays. The samples collected from living controls for the optimisation aspect of this study were oral mouth rinse samples, to minimise invasiveness. Furthermore, oral mouth rinse samples had been utilised (and had worked well) in the pilot study. However, when the forensic post-mortem blood sample was utilised in the case example, the assay had to be re-optimised. An example of this can be seen in Figures 2 and 3 and 4, which are representative of saliva, saliva and blood samples respectively. The PCR assays that had been completely optimised using saliva mouth rinse samples had to be re-optimised, due to a proliferation of non-specific amplicons of a greater size than the expected amplicons, when using the blood sample. This problem may occur due to a difference in quality between the two samples. This concern may occur in future in the post-mortem context, especially in cases where the individual is burnt or decomposed or skeletonised. The quality of the DNA sample may be compromised compared to the saline mouth rinse and blood sample used in this research.

Blood samples are generally of a higher quality than saliva or buccal samples, allowing for the amplification of larger fragments which could not be amplified during optimisation with lower quality DNA samples (the DNA may be sheared into smaller fragments in a lower quality sample)[79]. This was previously noted by Heathfield and Reid (2017) who found that buccal samples collected post-mortem were significantly more degraded than blood samples from the same individual ($p < 0.001$) [80]. While the use of a higher quality sample would be preferred, in the forensic context one is often limited with regards to the type of sample one will receive or obtain.

Another problematic aspect was the consignment of the PCR primers into PCR multiplexes that would provide sufficient PCR products for all the AIMs involved, but could also have their products combined into functional SNaPshot® multiplex assays. A manifestation of this can be seen in contrasting the results in sections 3.4 and 3.5. Figures 10-18 show that the PCR multiplexes were optimised to the point where all amplicons in a multiplex showed both bright individual bands and relatively equal intensity of the amplicons within a multiplex. Despite this, it can be seen in Figures 21-26 that all the SNaPshot® amplicons within each multiplex were not of equal RFU, to the point that some amplicons were not present at all. Attempts to correct this included changing the ratio of SNaPshot® primers used in each multiplex, increasing the ratio of the primers of those SNaPshot® amplicons that were of low RFU or absent. Furthermore, the ratios of the PCR multiplex products added to the SNaPshot® multiplex reactions were altered to the same end. These changes did lead

to improvement with regards to the inequality of product RFU, but not sufficiently. However, the timeframe and scope of the project did not allow for the SNaPshot® multiplex reactions to be fully optimised, despite full optimisation of the PCR multiplex reactions. One of the primary remaining issues with the optimisation of the SNaPshot® assays was the low RFU of some SNaPshot® products compared to others within the same multiplex. In future, changes that could be attempted are: increasing the number of multiplexes, since this would decrease the number of amplicons in each multiplex, thereby reducing competition for reagents; redesigning the SNaPshot primers that do not perform well in the reaction; although the ability to redesign SNaPshot® primers may be limited due to their nature[81]. The optimisation of the SNaPshot® assays could be part of the subject for a future research project however.

4.4 Ethical and legal considerations

The primary ethical concern as far as research on human subjects is concerned has always been consent. Consent with regards to healthy, living subjects is relatively simple: for example, all the demographics, data and DNA samples received for this study were given by individuals you had given full, written, informed consent to participate in this study. Consent with regards to the deceased can be considered a more nebulous topic however. The neonate decedent that was used as the subject for the case study of this research had been unidentified for over 30 days, so the rights of the decedent had been given over to governmental jurisdiction. Usually consent would need to be given by the next of kin for samples from the decedent to be utilised in research. However, in this case, ethical clearance from the Human Research Committee of UCT, as well as permission from both the Forensic Pathology Services and Salt-River Mortuary itself, was sufficient. Furthermore, given that the sample used for this study was taken in the normal course of the autopsy of the decedent, and that the assay performed was only predicting EVCs (i.e. phenotypes that would be visible to the naked eye in a public setting), the potential benefits of performing the study were considered to outweigh such ethical ambiguity that could be considered to be present.

In terms of the legal considerations regarding the research, the Criminal Law (Forensic Procedures) Amendment Act, 37 of 2013, dictates that the South African Police force and associated departments cannot perform DNA analysis that relies on coding regions of the genome. The assay we have created would be considered to fall foul of these restrictions. However, the assays performed in this study with regards to the decedent were utilised to the ends of research only. Since the determined information regarding the decedent was not returned to the South African police Service (SAPS), and would not be utilised in any medico-legal proceedings, the restrictions that constrain the SAPS in these situations did not apply to this research.

4.5 Strengths and Limitations of the study

There are several limitations to this research and the conclusions drawn thereof. The first is that the cohort self-reported all of the demographic information that we obtained. This may result in discrepancies between self-reported ethnic groups and their ancestry, for example. Another limitation is that the qualitative assessment of the appearance of the decedent could only estimate the pigmentation phenotypes in terms of broad categories. This is not so much a limitation of this research in particular, so much as a limitation of the field of molecular phenotyping as it stands at present. Further, the association studies that were used to provide the information to perform the qualitative assessment were not performed in South Africa, therefore it cannot be known whether the associations found therein are present or show the same magnitude within the South African population. The associations of the 13 AIMs (and the magnitude of the effects of each AIM on the various pigmentation phenotypes) cannot be accurately measured until a larger cohort study and modelling is performed.

However, this study showed that 10 of the 13 AIMs utilised in the assay may have predictive capabilities regarding pigmentation within the South African context, but this proof of concept would need to be verified in a statistically significant cohort. This provides the impetus for further research regarding them in our context. Additionally, while the predicted phenotypes were of broad categories, they nevertheless provided information regarding the decedent beyond that which could be determined by pathologists. This information, in conjunction with other identifying evidence, could prove useful in medicolegal cases regarding unrecognisable individuals.

4.6 Conclusion

AIMs can be useful in the forensic context because they can be used to estimate ancestry and, when the AIMs are situated in the vicinity of pigmentation associated genes, can be used to estimate appearance. The estimation of appearance in particular would be useful to help identify the donors of DNA on crime scenes and in our context, unrecognizable decedents.

This study designed an assay to sequence 13 AIMs to estimate the appearance of an identified deceased neonate to demonstrate a proof of concept. This led to the development of a second research question, and to this end, an attempt to optimise a multiplex SNaPshot® PCR assay was made, to genotype these 13 AIMs in a larger cohort and model associations of pigmentation phenotypes such as MI.

The primary objective was completed, allowing a qualitative prediction of appearance, which also concurred with information provided by pathologists. However, assessment will need to be supported by further research performed using larger scale genotyping studies of the 13 AIMs in the South African population, such as completing optimisation of the SNaPshot® assay designed here and genotyping of the 385 individuals whose DNA, MI and demographics could be collected in the scope of the research. However, accuracy of the qualitative study nevertheless suggests that the 13 AIMs examined in this research have relevance within the South African context. These 13 AIMs could provide useful in future with regards to aiding the identification of unrecognizable decedents through the prediction of pigmentation phenotypes, and therefore, appearance.

References

Referenced in the style of the Forensic Science International journal.

[1] R. Kosoy, R. Nassir, C. Tian, P.A. White, L.M. Butler, G. Silva, et al., Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America, Hum.Mutat. 30 (2009) 69-78.

[2] M.D. Shriver, E.J. Parra, S. Dios, C. Bonilla, H. Norton, C. Jovel, et al., Skin pigmentation, biogeographical ancestry and admixture mapping, Hum.Genet. 112 (2003) 387-399.

[3] S. Beleza, N.A. Johnson, S.I. Candille, D.M. Absher, M.A. Coram, J. Lopes, et al., Genetic architecture of skin and eye color in an African-European admixed population, PLoS Genet. 9 (2013) e1003372.

[4] World Health Organization, International Commission on Non-Ionizing Radiation Protection, Global solar UV index: a practical guide, (2002).

[5] R. McKenzie, Application of a simple model to calculate latitudinal and hemispheric differences in ultraviolet radiation, Weather and climate. 11 (1991) 3-14.

[6] J.K. Wagner, E.J. Parra, H. L Norton, C. Jovel, M.D. Shriver, Skin responses to ultraviolet radiation: effects of constitutive pigmentation, sex, and ancestry, Pigment cell research. 15 (2002) 385-390.

[7] J.H. Relethford, Hemispheric difference in human skin color, Am.J.Phys.Anthropol. 104 (1997) 449-457.

[8] A.L. Cook, W. Chen, A.E. Thurber, D.J. Smit, A.G. Smith, T.G. Bladen, et al., Analysis of cultured human melanocytes based on polymorphisms within the SLC45A2/MATP, SLC24A5/NCKX5, and OCA2/P loci, J.Invest.Dermatol. 129 (2009) 392-405.

[9] G. Lucotte, G. Mercier, F. Diéterlen, I. Yuasa, A decreasing gradient of 374F allele frequencies in the skin pigmentation gene SLC45A2, from the north of West Europe to North Africa, Biochem.Genet. 48 (2010) 26-33.

[10] S.M. Callegari-Jacques, D. Grattapaglia, F.M. Salzano, S.P. Salamoni, S.G. Crossetti, M.E. Ferreira, et al., Historical genetics: spatiotemporal analysis of the formation of the Brazilian population, Am.J.Hum.Biol. 15 (2003) 824-834.

[11] R. Nassir, R. Kosoy, C. Tian, P.A. White, L.M. Butler, G. Silva, et al., An ancestry informative marker set for determining continental origin: validation and extension using human genome diversity panels, BMC genetics. 10 (2009) 1.

[12] S. Walsh, F. Liu, A. Wollstein, L. Kovatsi, A. Ralf, A. Kosiniak-Kamysz, et al., The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA, Forensic Science International: Genetics. 7 (2013) 98-115.

[13] J. Graf, R. Hodgson, A. Van Daal, Single nucleotide polymorphisms in the MATP gene are associated with normal human pigmentation variation, Hum.Mutat. 25 (2005) 278-284.

[14] K. Nakayama, S. Fukamachi, H. Kimura, Y. Koda, A. Soemantri, T. Ishida, Distinctive distribution of AIM1 polymorphism among major human populations with different skin color, J.Hum.Genet. 47 (2002) 92-94.

[15] I. Yuasa, K. Umetsu, S. Harihara, A. Kido, A. Miyoshi, N. Saitou, et al., Distribution of the F374 Allele of the SLC45A2 (MATP) Gene and Founder-Haplotype Analysis, Ann.Hum.Genet. 70 (2006) 802-811.

[16] F. Liu, K. van Duijn, J.R. Vingerling, A. Hofman, A.G. Uitterlinden, A.C.J. Janssens, et al., Eye color and the prediction of complex phenotypes from genotypes, Current Biology. 19 (2009) R192-R193.

[17] H. Nan, P. Kraft, D.J. Hunter, J. Han, Genetic variants in pigmentation genes, pigmentary phenotypes, and risk of skin cancer in Caucasians, International journal of cancer. 125 (2009) 909-917.

[18] K. Vongpaisarnsin, K. Vongpaisarnsin, Eye colour single nucleotide polymorphisms (SNPs) variants in Thai population, Forensic Science International: Genetics Supplement Series. 4 (2013) e198-e199.

[19] de Cerqueira, Caio Cesar Silva, T. Hünemeier, J. Gomez-Valdés, V. Ramallo, C.D. Volasko-Krause, A.A.L. Barbosa, et al., Implications of the admixture process in skin color molecular assessment, PloS one. 9 (2014) e96886.

[20] E.E. Quillen, M. Bauchet, A.W. Bigham, M.E. Delgado-Burbano, F.X. Faust, Y.C. Klimentidis, et al., OPRM1 and EGFR contribute to skin pigmentation differences between Indigenous Americans and Europeans, Hum.Genet. 131 (2012) 1073-1080.

[21] D.F. Durso, S.P. Bydlowski, M.H. Hutz, G. Suarez-Kurtz, T.R. Magalhães, S.D.J. Pena, Association of genetic variants with self-assessed color categories in Brazilians, PloS one. 9 (2014) e83926.

[22] M. Soejima, Y. Koda, Population differences of two coding SNPs in pigmentation-related genes SLC24A5 and SLC45A2, Int.J.Legal Med. 121 (2007) 36-39.

[23] T.K. Leite, R.M. Fonseca, N.M. De França, E.J. Parra, R.W. Pereira, Genomic ancestry, self-reported “color” and quantitative measures of skin pigmentation in Brazilian admixed siblings, PLoS One. 6 (2011) e27162.

[24] K. Eaton, M. Edwards, S. Krithika, G. Cook, H. Norton, E.J. Parra, Association study confirms the role of two OCA2 polymorphisms in normal skin pigmentation variation in East Asian populations, Am.J.Hum.Biol. 27 (2015) 520-525.

[25] F. de Araújo Lima, F. de Toledo Gonçalves, C. Fridman, SLC24A5 and ASIP as phenotypic predictors in Brazilian population for forensic purposes, Leg.Med. 17 (2015) 261-266.

[26] P.A. Kanetsky, J. Swoyer, S. Panossian, R. Holmes, D. Guerry, T.R. Rebbeck, A polymorphism in the agouti signaling protein gene is associated with human pigmentation, The American Journal of Human Genetics. 70 (2002) 770-775.

[27] C. Zeigler-Johnson, S. Panossian, S.M. Gueye, M. Jalloh, D. Ofori-Adjei, P.A. Kanetsky, Population differences in the frequency of the agouti signaling protein g. 8818a> G polymorphism, Pigment cell research. 17 (2004) 185-187.

[28] C. Bonilla, L. Boxill, S.A. Mc Donald, T. Williams, N. Sylvester, E.J. Parra, et al., The 8818G allele of the agouti signaling protein (ASIP) gene is ancestral and is associated with darker skin color in African Americans, Hum.Genet. 116 (2005) 402-406.

[29] M. Edwards, A. Bigham, J. Tan, S. Li, A. Gozdzik, K. Ross, et al., Association of the OCA2 polymorphism His615Arg with melanin content in east Asian populations: further evidence of convergent evolution of skin pigmentation, PLoS Genet. 6 (2010) e1000867.

[30] W. Branicki, F. Liu, K. van Duijn, J. Draus-Barini, E. Pośpiech, S. Walsh, et al., Model-based prediction of human hair color using DNA variants, Hum.Genet. 129 (2011) 443-454.

[31] T.R. Rebbeck, P.A. Kanetsky, A.H. Walker, R. Holmes, A.C. Halpern, L.M. Schuchter, et al., P gene as an inherited biomarker of human eye color, Cancer Epidemiol.Biomarkers Prev. 11 (2002) 782-784.

[32] S. Choudhry, N.E. Coyle, H. Tang, K. Salari, D. Lind, S.L. Clark, et al., Population stratification confounds genetic association studies among Latinos, Hum.Genet. 118 (2006) 652-664.

[33] J. Beuten, I. Halder, S.P. Fowler, H.H. Gőring, R. Duggirala, R. Arya, et al., Wide disparity in genetic admixture among Mexican Americans from San Antonio, TX, Ann.Hum.Genet. 75 (2011) 529-538.

[34] A. Ruiz-Linares, K. Adhikari, V. Acuña-Alonzo, M. Quinto-Sanchez, C. Jaramillo, W. Arias, et al., Admixture in Latin America: geographic structure, phenotypic diversity and self-perception of ancestry based on 7,342 individuals, PLoS Genet. 10 (2014) e1004572.

[35] T. Juárez-Cedillo, J. Zuñiga, V. Acuña-Alonzo, N. Pérez-Hernández, J.M. Rodríguez-Pérez, R. Barquera, et al., Genetic admixture and diversity estimations in the Mexican Mestizo population from Mexico City using 15 STR polymorphic markers, Forensic Science International: Genetics. 2 (2008) e37-e39.

[36] N. Godinho, C. Gontijo, M. Diniz, G. Falcão-Alencar, G. Dalton, C. Amorim, et al., Regional patterns of genetic admixture in South America, Forensic Science International: Genetics Supplement Series. 1 (2008) 329-330.

[37] R. Rodrigues de Moura, A.V.C. Coelho, V. de Queiroz Balbino, S. Crovella, L.A.C. Brandão, Meta-analysis of Brazilian genetic admixture and comparison with other Latin America countries, Am.J.Hum.Biol. 27 (2015) 674-680.

[38] S. Avena, M. Via, E. Ziv, E.J. Pérez-Stable, C.R. Gignoux, C. Dejean, et al., Heterogeneity in genetic admixture across different regions of Argentina, PloS one. 7 (2012) e34695.

[39] S. Beleza, J. Campos, J. Lopes, I.I. Araújo, A.H. Almada, A.C. e Silva, et al., The admixture structure and genetic variation of the archipelago of Cape Verde and its implications for admixture mapping studies, PloS one. 7 (2012) e51103.

[40] S. Wang, N. Ray, W. Rojas, M.V. Parra, G. Bedoya, C. Gallo, et al., Geographic patterns of genome admixture in Latin American Mestizos, PLoS Genet. 4 (2008) e1000037.

[41] A. Lucassen, K. Ehlers, P.J. Grobler, A.L. Shezi, Allele frequency data of 15 autosomal STR loci in four major population groups of South Africa, Int.J.Legal Med. 128 (2014) 275-276.

[42] N. Patterson, D.C. Petersen, R.E. van der Ross, H. Sudoyo, R.H. Glashoff, S. Marzuki, et al., Genetic structure of a unique admixed population: implications for medical research, Hum.Mol.Genet. 19 (2010) 411-419.

[43] E. de Wit, W. Delport, C.E. Rugamika, A. Meintjes, M. Möller, P.D. van Helden, et al., Genome-wide analysis of the structure of the South African Coloured Population in the Western Cape, Hum.Genet. 128 (2010) 145-153.

[44] L.B. Barreiro, O. Neyrolles, C.L. Babb, L. Tailleux, H. Quach, K. McElreavey, et al., Promoter variation in the DC-SIGN–encoding gene CD209 is associated with tuberculosis, PLoS Med. 3 (2006) e20.

[45] M. Sinha, E.K. Larkin, R.C. Elston, S. Redline, Self-reported race and genetic admixture, N.Engl.J.Med. 354 (2006) 421-422.

[46] I. Halder, M. Shriver, M. Thomas, J.R. Fernandez, T. Frudakis, A panel of ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: utility and applications, Hum.Mutat. 29 (2008) 648-658.

[47] Y.C. Klimentidis, G.F. Miller, M.D. Shriver, Genetic admixture, self-reported ethnicity, self-estimated admixture, and skin pigmentation among Hispanics and Native Americans, Am.J.Phys.Anthropol. 138 (2009) 375-383.

[48] M.I. McCarthy, J.N. Hirschhorn, Genome-wide association studies: potential next steps on a genetic journey, Hum.Mol.Genet. 17 (2008) R156-65.

[49] W. Branicki, U. Brudnik, A. Wojas-Pelc, Interactions between HERC2, OCA2 and MC1R may influence human pigmentation phenotype, Ann.Hum.Genet. 73 (2009) 160-170.

[50] P. Sulem, D.F. Gudbjartsson, S.N. Stacey, A. Helgason, T. Rafnar, M. Jakobsdottir, et al., Two newly identified genetic determinants of pigmentation in Europeans, Nat.Genet. 40 (2008) 835-837.

[51] R.K. Valenzuela, M.S. Henderson, M.H. Walsh, N. Garrison, J.T. Kelch, O. Cohen-Barak, et al., Predicting phenotype from genotype: normal pigmentation, J.Forensic Sci. 55 (2010) 315-322.

[52] H. Eiberg, J. Troelsen, M. Nielsen, A. Mikkelsen, J. Mengel-From, K.W. Kjaer, et al., 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 (2008) 177-187.

[53] N. Fracasso, E. Andrade, C. Andrade, L. Zanão, M. Silva, L. Marano, et al., Association of SNPs from the SLC45A2 gene with human pigmentation traits in Brazil, Forensic Science International: Genetics Supplement Series. 4 (2013) e342-e343.

[54] B. Marcheco-Teruel, E.J. Parra, E. Fuentes-Smith, A. Salas, H.N. Buttenschøn, D. Demontis, et al., Cuba: exploring the history of admixture and the genetic basis of pigmentation using autosomal and uniparental markers, PLoS Genet. 10 (2014) e1004488.

[55] C. Fridman, Cardena, Mari Maki Síria Godoy, F. de Araújo Lima, F. de Toledo Gonçalves, Is it possible to use Forensic DNA phenotyping in Brazilian population? Forensic Science International: Genetics Supplement Series. 5 (2015) e378-e380.

[56] U. Soundararajan, L. Yun, M. Shi, K.K. Kidd, Minimal SNP overlap among multiple panels of ancestry informative markers argues for more international collaboration, Forensic Science International: Genetics. 23 (2016) 25-32.

[57] C. Phillips, W. Parson, B. Lundsberg, C. Santos, A. Freire-Aradas, M. Torres, et al., Building a forensic ancestry panel from the ground up: The EUROFORGEN Global AIM-SNP set, Forensic Science International: Genetics. 11 (2014) 13-25.

[58] M. Daya, L. Van Der Merwe, U. Galal, M. Möller, M. Salie, E.R. Chimusa, et al., A panel of ancestry informative markers for the complex five-way admixed South African Coloured population, PloS one. 8 (2013) e82224.

[59] H. Tsai, S. Choudhry, M. Naqvi, W. Rodriguez-Cintron, E.G. Burchard, E. Ziv, Comparison of three methods to estimate genetic ancestry and control for stratification in genetic association studies among admixed populations, Hum.Genet. 118 (2005) 424-433.

[60] de Cerqueira, Caio Cesar Silva, T. Hünemeier, J. Gomez-Valdés, V. Ramallo, C.D. Volasko-Krause, A.A.L. Barbosa, et al., Implications of the admixture process in skin color molecular assessment, PloS one. 9 (2014) e96886.

[61] M. de la Puente, C. Santos, M. Fondevila, L. Manzo, Á. Carracedo, M. Lareu, et al., The Global AIMs Nano set: A 31-plex SNaPshot assay of ancestry-informative SNPs, Forensic Science International: Genetics. 22 (2016) 81-88.

[62] K.B. Mullis, F.A. Faloona, Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction, Methods Enzymol. 155 (1987) 335-350.

[63] T. Koressaar, M. Remm, Enhancements and modifications of primer design program Primer3, Bioinformatics. 23 (2007) 1289-1291.

[64] P.M. Vallone, J.M. Butler, AutoDimer: a screening tool for primer-dimer and hairpin structures, BioTechniques. 37 (2004) 226-231.

[65] J. Ye, G. Coulouris, I. Zaretskaya, I. Cutcutache, S. Rozen, T.L. Madden, Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction, BMC Bioinformatics. 13 (2012) 1.

[66] Thermo Fisher Scientific, BigDye™ Terminator v3.1 Cycle Sequencing Kit User Guide, (2016).

[67] E. Werle, C. Schneider, M. Renner, M. Volker, W. Fiehn, Convenient single-step, one tube purification of PCR products for direct sequencing, Nucleic Acids Res. 22 (1994) 4354-4355.

[68] M.S. Hayney, P. Dimanlig, J.J. Lipsky, G.A. Poland, Utility of a “swish and spit” technique for the collection of buccal cells for TAP haplotype determination, 70 (1995) 951-954.

[69] M.S. Hayney, G.A. Poland, J.J. Lipsky, A Noninvasive ‘Swish and Spit ‘Method for Collecting Nucleated Cells for HLA Typing by PCR in Population Studies, Hum.Hered. 46 (1996) 108-111.

[70] T. Zayats, T.L. Young, D.A. Mackey, F. Malecaze, P. Calvas, J.A. Guggenheim, Quality of DNA extracted from mouthwashes, PLoS One. 4 (2009) e6165.

[71] J. Lee, A.K. Jain, R. Jin, Scars, marks and tattoos (SMT): Soft biometric for suspect and victim identification, (2008) 1-8.

[72] M. Kayser, Forensic DNA phenotyping: predicting human appearance from crime scene material for investigative purposes, Forensic Science International: Genetics. 18 (2015) 33-48.

[73] M. Kayser, P.M. Schneider, DNA-based prediction of human externally visible characteristics in forensics: motivations, scientific challenges, and ethical considerations, Forensic Science International: Genetics. 3 (2009) 154-161.

[74] N. Murray, H.L. Norton, E.J. Parra, Distribution of two OCA2 polymorphisms associated with pigmentation in East-Asian populations, Human Genome Variation. 2 (2015).

[75] I. Yuasa, K. Umetsu, S. Harihara, A. Miyoshi, N. Saitou, K.S. Park, et al., OCA2* 481Thr, a hypofunctional allele in pigmentation, is characteristic of northeastern Asian populations, J.Hum.Genet. 52 (2007) 690-693.

[76] C.C. Gravlee, W.W. Dressler, Skin pigmentation, self-perceived color, and arterial blood pressure in Puerto Rico, Am.J.Hum.Biol. 17 (2005) 195-206.

[77] T.W. Lim, M.H. Lee, A study of skin color by melanin index according to sex, age, site and skin phototype in Koreans, Annals of Dermatology. 14 (2002) 71-76.

[78] L. du Toit-Prinsloo, C. Pickles, Z. Smith, J. Jordaan, G. Saayman, The medico-legal investigation of abandoned fetuses and newborns—a review of cases admitted to the Pretoria Medico-Legal Laboratory, South Africa, Int.J.Legal Med. 130 (2016) 569-574.

[79] T. Zayats, T.L. Young, D.A. Mackey, F. Malecaze, P. Calvas, J.A. Guggenheim, Quality of DNA extracted from mouthwashes, PLoS One. 4 (2009) e6165.

[80] L.J. Heathfield, K.M. Reid, Success rate of forensic DNA profiling from cotton swabs and blood samples obtained from deceased infants at Salt River Mortuary, The SAPS Forensic Services 4th Annual Conference (2017).

[81] B. Mehta, R. Daniel, C. Phillips, D. McNevin, Forensically relevant SNaPshot® assays for human DNA SNP analysis: a review, Int.J.Legal Med. (2017) 1-17.

Appendices

Appendix 1: Ethical clearance for study from Human Research Ethics Committee of UCT

UNIVERSITY OF CAPE TOWN
Faculty of Health Sciences
Human Research Ethics Committee

Room E52-24 Old Main Building
Groote Schuur Hospital
Observatory 7925
Telephone [021] 406 6492
Email: sumayah.ariefdien@uct.ac.za
Website: www.health.uct.ac.za/fhs/research/humanethics/forms

03 May 2016

HREC REF: 158/2016

Ms L Heathfield
Forensic Medicine and Toxicology
Pathology
Entrance 3, Level 1
Falmouth Building-FHS

Dear Ms Heathfield

PROJECT TITLE: THE ASSESSMENT OF 18 MOLECULAR MARKERS FOR SKIN COLOUR IN SOUTH AFRICA: (MPhil candidate-Mr G Pharo) sub-study linked to 317/2015

Thank you for your response letter dated 26 April 2016, addressing the issues raised by the Human Research Ethics Committee (HREC).

It is a pleasure to inform you that the HREC has **formally approved** the above-mentioned study.

Approval is granted for one year until the 30 May 2017.

Please submit a progress form, using the standardised Annual Report Form if the study continues beyond the approval period. Please submit a Standard Closure form if the study is completed within the approval period.
(Forms can be found on our website: www.health.uct.ac.za/fhs/research/humanethics/forms)

Please quote the HREC REF in all your correspondence.

We acknowledge that the student, Gavin Pharo will also be involved in this study.

Please note that the ongoing ethical conduct of the study remains the responsibility of the principal investigator.

Please note that for all studies approved by the HREC, the principal investigator **must** obtain appropriate institutional approval before the research may occur.

Yours sincerely

PROFESSOR M BLOCKMAN
CHAIRPERSON, FHS HUMAN RESEARCH ETHICS COMMITTEE

Federal Wide Assurance Number: FWA00001637.

HREC 158/2016

Institutional Review Board (IRB) number: IRB00001938
This serves to confirm that the University of Cape Town Human Research Ethics Committee complies to the Ethics Standards for Clinical Research with a new drug in patients, based on the Medical Research Council (MRC-SA), Food and Drug Administration (FDA-USA), International Convention on Harmonisation Good Clinical Practice (ICH GCP), South African Good Clinical Practice Guidelines (DoH 2006), based on the Association of the British Pharmaceutical Industry Guidelines (ABPI),and Declaration of Helsinki (2013) guidelines.
The Human Research Ethics Committee granting this approval is in compliance with the ICH Harmonised Tripartite Guidelines E6: Note for Guidance on Good Clinical Practice (CPMP/ICH/135/95) and FDA Code Federal Regulation Part 50, 56 and 312.

HREC 158/2016

Appendix 2: PARTICIPANT QUESTIONNAIRE:

PARTICIPANT QUESTIONNAIRE:

Participant reference number: _____

Have you participated in this study previously? If no, please continue to question 1.
☐ Yes ☐ No

Have you used self-tanner, tanning beds or skin-lightening treatments? If yes, please inform the researcher.

☐ Yes ☐ No

1. Sex ☐ Male ☐ Female

2. Age ☐ 18 – 39 ☐ 40 – 60 ☐ 60+

3. Self-reported population group according to South African Census categories

☐African Black ☐Coloured ☐Indian/Asian ☐White

4. Self-reported eye colour

☐ Blue ☐ Grey ☐ Green ☐ Hazel ☐ Brown

5. Self-reported hair colour

☐ Black ☐ Brown ☐ Blond ☐ Red ☐ Grey

6. Ancestral origin (if known)

☐ European – N W S E

☐Africa – N W S E

☐Asian – N W S E

☐ Middle Eastern

☐ Other

Specify: _____

7. Your self-reported ethnicity (In reference to cultural self-identification)

_____

8. Father's ethnicity and/or ancestral origin

_____

9. Mother's ethnicity and/or ancestral origin

_____

10. Parental (father's side) grandparent's ethnicity and/or ancestral origin

Grandfather: _____

Grandmother: _____

11. Maternal (mother's side) grandparent's ethnicity and/or ancestral origin

Grandfather: _____

Grandmother: _____

12. Recorded Melanin Index and Colour Reflectance Values

Recorded Melanin Index		Red Reflectance Value	Blue Reflectance Value	Green Reflectance Value
Right Inner Fore-arm				
Right inner Arm Above Elbow				
Left Inner Fore-arm				
Left inner Arm Above Elbow				
Calculated Average				
Forehead				

Appendix 3: Internal and External Primer Sequences

Table A: External primer set sequences

AIM SNP PRIMER	Sequence (5'-3')
ADAM17 rs1524668 F	AGAATGTGACAGTTGAGTACGG
ADAM17 rs1524668 R	AGGAGGTGTTCAGTCGGTTG
AFG3L1 rs4785763 F	GCAGTTTGGGGTGAGTGAG
AFG3L1 rs4785763 R	GTTTTGGGTGAGATTCCGCA
ASIP rs6058017 F	AGTCTGGATGGGGATGGAGG
ASIP rs6058017 R	GCGAAGGGACCGAGAACTTT
HERC2 rs1129038 F	AGTCAGTCTCTCCACTCCCTC
HERC2 rs1129038 R	CCTGAGTCCTACACCTGTTTC
OCA2 rs1800401 F	ATCTCAAGCCTCCCTGACTG
OCA2 rs1800401 R	CCCCTACTCACTGTTCATTGTC
OCA2 rs1800407 F	GCACCTGAGAATGGAACCT
OCA2 rs1800407 R	TGGCTTGTACTCTCTCTGTGT
OCA2 rs1800414 F	GGCTCTGAAACCTTCCCAT
OCA2 rs1800414 R	CGTGATTCCAGTTGCGTAG
SLC24A5 rs1426654 F	CCCTTGGATTGTCTCAGGATG
SLC24A5 rs1426654 R	TGAGTAAGCAAGTATAAGGAGCAA
SLC45A2 rs26722 F	TGGCTTCATCTTCCCTGGTT
SLC45A2 rs26722 R	GACCCTCCATTGTCATCAGA
SLC45A2 rs16891982 F	TCCAAGTTGTGCTAGACCAGA
SLC45A2 rs16891982 R	CGAAAGAGGAGTCGAGGTTG
TYR rs1126809 F	GCATTCTGGAGGTTCAAAACT
TYR rs1126809 R	GCATTGGCTTCTGGATAAACTTC
PROCR rs2069945 F	GCAAACCTTGGCTCTGCTAC
PROCR rs2069945 R	CCTTTCCCCAGTGGCTTAAT
TLR1 rs4540055 F	TGTGCCTCTGATCACTTTTGAATAC
TLR1 rs4540055 R	CCTAGCCAACTCCAGAGTTCAT

Table B: Internal primer set sequences

AIM SNaPshot PRIMER	Sequence (5'-3')
HERC2 rs1129038 S-R	ACCAGGCAGCAGAGC
OCA2 rs1800407 S-F	AAAAAAAAAAAACCCACACCCGTCCC
OCA2 rs1800414 S-R	AAAAAAAAAAAAAGGTTTCTCTCTTACAGC
OCA2 rs1800401 S-R	AAAAAAAAAAAAAAAAAAAATGTCCATCAGCATC
SLC45A2 rs26722 S-R	AAAAAAAAAAAAATGGAATGTACGAGTATGGTTCTATC
SLC24A5 rs1426654 S-R	AAAAAAAAAAAAAAAAAAAAAAAAAATCTCAGGATGTTGCAGGC
PROCR rs2069945 S-R	AAAGGTGCCACGTCGTGAAAGTCTGACAATATTATTAACCCCAGTCTAC ATG
ASIP rs6058017 S-R	ACCCGAAGCCCTGCC
AFG3L1 rs4785763 S-R	CTACAAACTCCTCATAGGTAGACTTC
TYR rs1126809 S-R	AAAAAAAAAAAAAAATGAAGAGGACGGTGCCTT
SLC45A2 rs16891982 S-F	AAAAAAAAAATTATGTTATATCTTACACGGAGTTGATGCA
ADAM17 rs1524668 S-F	AAAAAAAAAAAAAAAAAAAAAAAAAAAAAGAACATCCAGTCACCATA
TLR1 rs4540055 S-R	AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAACGAA GCAGTGATCAGCAC

Appendix 4: Sequencing Electropherograms

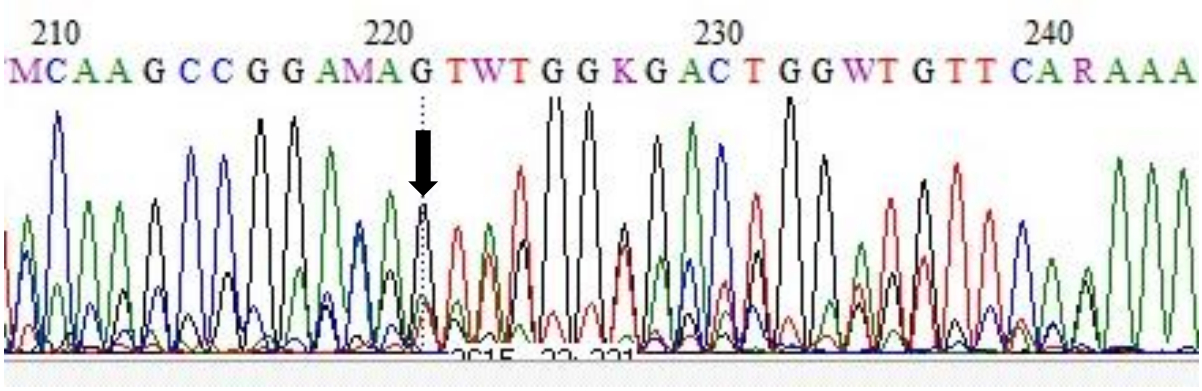

Figure A: electropherogram showing the sequenced region surrounding *ADAM17* rs1524668 A>C. The peak corresponding to rs1524668 is indicated by a black arrow. The SNP identity is G/G, but since the sequencing was performed using the reverse primer.

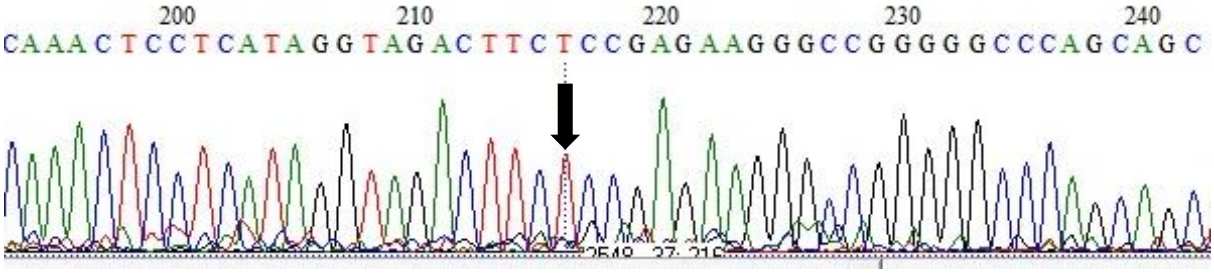

Figure B: electropherogram showing the sequenced region surrounding *AFG3L1* rs4785763 A>C. The peak corresponding to rs4785763 is indicated by a black arrow. SNP identity is T/T, but since the sequencing was performed using the reverse primer, the true identity is A/A.

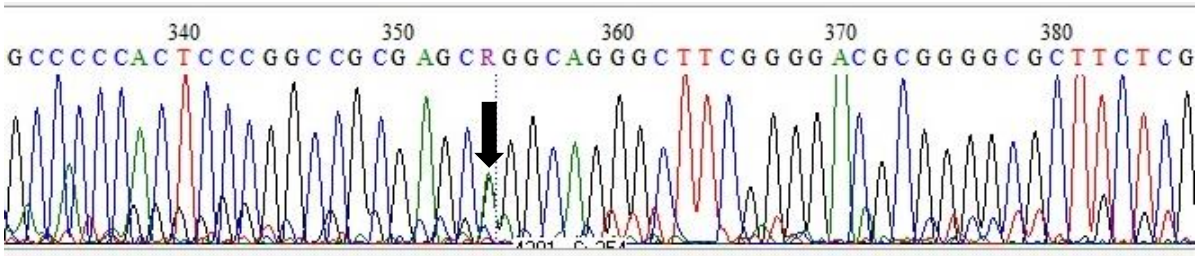

Figure C: electropherogram showing the sequenced region surrounding *ASIP* rs6058017 G>A. The peak corresponding to rs6058017 is indicated by a black arrow. The SNP identity is R, or G/A.

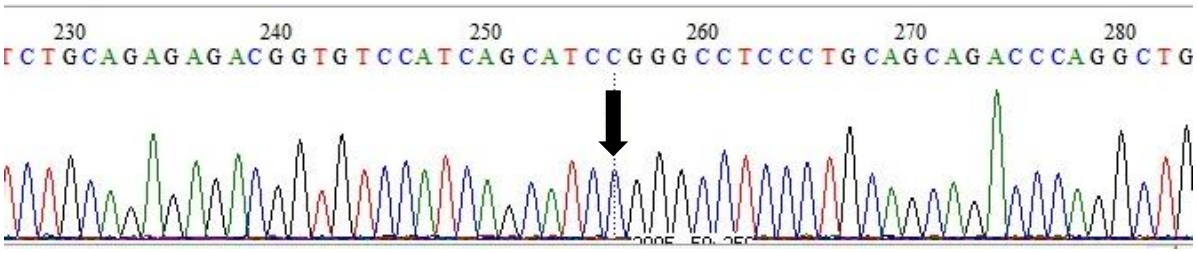

Figure D: electropherogram showing the sequenced region surrounding *OCA2* rs1800401 C>T. The peak corresponding to rs1800401 is indicated by a black arrow. The SNP identity is C/C, but since the sequencing was performed using the reverse primer, the true identity is G/G.

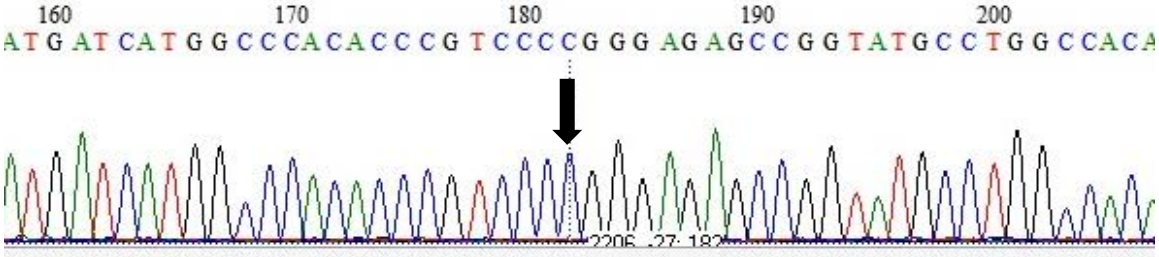

Figure E: electropherogram showing the sequenced region surrounding *OCA2* rs1800407 G>A. The peak corresponding to rs1800407 is indicated by a black arrow. The SNP identity is C/C.

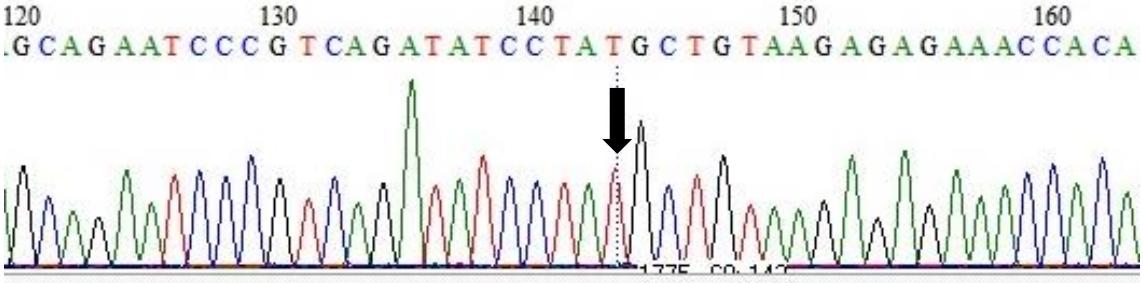

Figure F: electropherogram showing the sequenced region surrounding *OCA2* rs1800414 A>G. The peak corresponding to rs1800414 is indicated by a black arrow. The SNP identity is T/T.

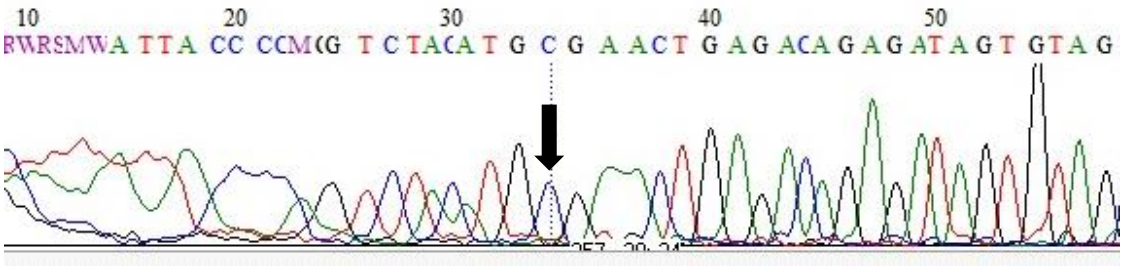

Figure G: electropherogram showing the sequenced region surrounding *PROCR* rs2069945 A/C/G. The peak corresponding to rs2069945 is indicated by a black arrow. The SNP identity is C/C, but since the sequencing was performed using the reverse primer, the true identity is G/G. The sequencing prior to the SNP appears to be very messy, but region surrounding the SNP can still be discerned. The identity of the SNP may be in question however.

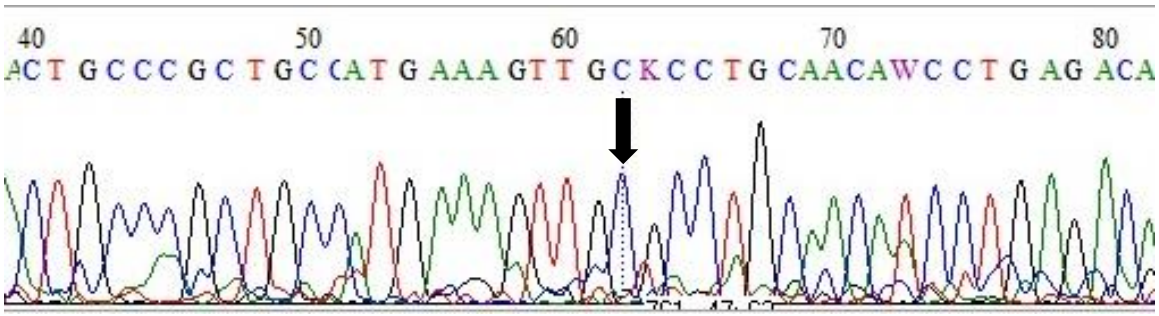

Figure H: electropherogram showing the sequenced region surrounding *SLC24A5* rs1426654 G>A. The peak corresponding to rs1426654 is indicated by a black arrow. The SNP identity is C/C, but since the sequencing was performed using the reverse primer, the true identity is G/G.

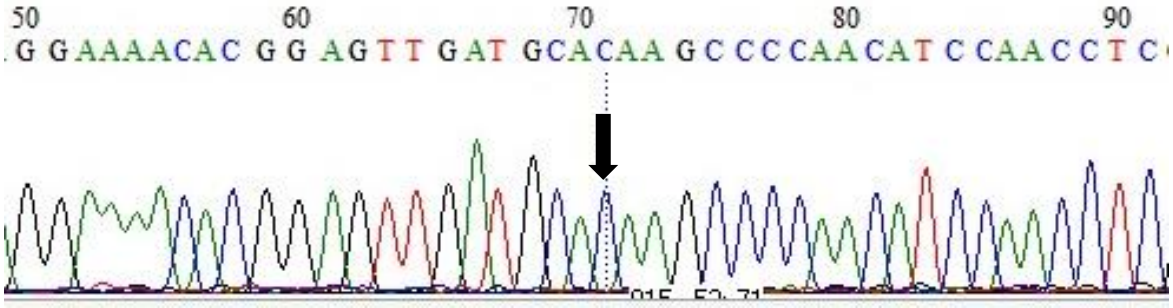

Figure I: electropherogram showing the sequenced region surrounding *SLC45A2* rs16891982 C>G. The peak corresponding to rs16891982 is indicated by a black arrow. The SNP identity is C/C.

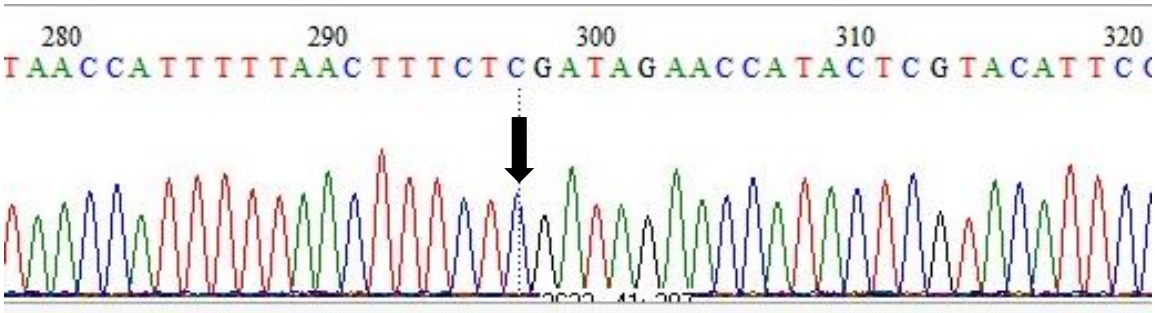

Figure J: electropherogram showing the sequenced region surrounding *SLC45A2* rs26722 G>A. The peak corresponding to rs26722 is indicated by a black arrow. The SNP identity is C/C.