THE DEVELOPMENT OF A SINGLE NUCLEOTIDE POLYMORPHISM DATABASE FOR FORENSIC IDENTIFICATION OF SPECIFIED PHYSICAL TRAITS

By

ALECIA GERALDINE NAIDU

A thesis submitted in fulfillment of the requirements for the degree of *Magister Scientiae* at the South African National Bioinformatics Institute (SANBI), University of the Western Cape.

November 2009

Supervisor: Professor Vladimir Bajic

RESPICE PROSPICE
UNIVERSITY *of the*
WESTERN CAPE

SANBI

KEYWORDS

Bioinformatics

Forensics

Pigmentation

Height

Single Nucleotide Polymorphism

Population

Text-Mining

MySQL

Forensic SNP Phenotype Database

Web user interface

ABSTRACT

The development of a single nucleotide polymorphism database for forensic identification of specified physical traits

A. G. Naidu

Magister Scientiae in Bioinformatics, Thesis,

South African National Bioinformatics Institute, University of the Western Cape

Many Single Nucleotide Polymorphisms (SNPs) found in coding or regulatory regions within the human genome lead to phenotypic differences that make prediction of physical appearance, based on genetic analysis, potentially useful in forensic investigations. Complex traits such as pigmentation can be predicted from the genome sequence, provided that genes with strong effects on the trait exist and are known. Phenotypic traits may also be associated with variations in gene expression due to the presence of SNPs in promoter regions. In this project, the identification of genes associated with these physical traits of potential forensic relevance have been collated from the literature using a text mining platform and hand curation. The SNPs associated with these genes have been acquired from public SNP repositories such as the International HapMap project, dbSNP and Ensembl. Characterization of different population groups based on the SNPs has been performed and the results and data stored in a MySQL database. This database contains SNP genotyping data with respect to physical phenotypic differences of forensic interest. The potential forensic-relevance of the SNP information contained in this database has been verified through *in silico* SNP analysis aimed at establishing possible relationships between SNP occurrence and phenotype. The software used for this analysis is MATCH™. Data management and access has been enhanced by the use of a functional web-based front-end which enables the users to extract and display SNP information without running complex Structured Query Language (SQL) statements from the command

line. This Forensic SNP Phenotype resource can be accessed at http://forensic.sanbi.ac.za/alecia_forensics/Index.html

DECLARATION

I declare that *The Development of a Single Nucleotide Polymorphism Database for Forensic Identification of Specified Physical Traits* is my own work, that it has not been submitted for any degree or examination in any other university, and that all the sources I have used or quoted have been indicated and acknowledged by complete references.

Alecia Geraldine Naidu

March 2010

Signed: A.G.N

ACKNOWLEDGEMENT

I would wish to thank my supervisor Prof. Vlad Bajic for the support and guidance through my Masters project. I'd also wish to thank Dr. Adam Dawe for his support during the data collection, analysis as well as the writing up of this thesis. To Allan Kamau, thank you for assisting me with the database design as well as the web interface deployment. I would like to thank Frederick Kamanu for aiding me with SNP extraction and much appreciation goes to Dr. Sebastian Schmeier for offering assistance in understanding transcription factor mapping. For the server allocation and configuration, much appreciation goes to Peter van Heusden and Dale Gibbs.

I give all thanks and glory to my Lord and Savior to whom I owe everything. A special thanks to my parents for their immense and unwavering support and love during my studies.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SQL STATEMENTS

ABBREVIATIONS

ASF	Apache Software Foundation
ASIP	Agouti Signaling Protein
cAMP	cyclic Adenosine mono-Phosphate
CSV	Comma Separated Values
DBMS	Database Management System
DES	Dragon Exploration System
DNA	Deoxyribonucleic Acid
ERD	Entity Relationship Diagram
FK	Foreign Key
GL	Gene List
HA	Human Anatomy
HTML	Hyper Text Markup Language
HGP	Human Genes and Proteins
INDEL	Insertion or Deletion
JDBC	Java Database Connectivity
JSP	JavaServer Pages
MATP	Membrane Associated Transporter Protein
MC1R	Melanocortin 1 Receptor
NCBI	National Centre for Biotechnology Information
OCA2	Oculocutaneous Albinism II
OMIM	Online Mendelian Inheritance in Man
PCR	Polymerase Chain Reaction
PK	Primary Key
SANBI	South African National Bioinformatics Institute
SLC24A5	Solute Carrier Family 24-member 5
SNP	Single Nucleotide Polymorphism
SQL	Structured Query Language
STR	Short Tandem Repeat
TFBS	Transcription Factor Binding Site
TSV	Tab Separated Values

TYR	Tyrosinase
UV	Ultraviolet
VDR	Vitamin D Receptor
XML	Extensive Markup Language

CHAPTER 1: INTRODUCTION

1.1 Rationale

Skin colouring, height and facial features are some of the key physical descriptors of a person, and are encoded in the genome. This is evidenced by the striking similarity observed in identical twins (van Daal, 2008). There has been a sharp increase in the number of genetic markers employed in the fight against crime, hence putting the focus onto the genes associated with notable physical traits. The genetic analysis of these genes is offering an attractive prospect for forensic investigations (Walsh 2004; Branicki *et al*., 2007).

With the completion of the Human Genome Project (Collins *et al.*, 2003), many research laboratories are currently engaged in gene and genetic marker identification. At the forefront is the need to characterize, among many other complex physical traits, the markers for eye, hair and skin colour as well as baldness. These markers include Single Nucleotide Polymorphisms (SNPs) and Short Tandem Repeats (STR). There is still on-going research in SNP physical characteristics prediction in forensics.

Much research has been done on the association between the SNPs and physical traits (van Daal 2008; Liu *et al*., 2009). However, this data has not been comprehensively integrated into one single resource which could be accessed by investigators wanting to research this area in various populations. The creation of such a resource would obviate the need to painstakingly mine through different literature sources and various biological databases for this data. Automation through use of a database would also ensure fast retrieval of SNP-related data for predicting physical traits. At this point it's important to point out that the identification of complex traits based on SNPs will almost always be likelihoods rather than certainties.

The forensic science community avoids, where possible, focusing on markers which are predictive of disease status or susceptibility, and hence aims at restricting genetic typing of coding loci to only those which are predictive of an observable physical trait. One way of curating the genetic variations that are associated with specified phenotypes is the development of databases to relate those variations to the trait (Budowle and van Daal 2008). Such a database would be created in two phases. The first phase, which is the staging database, would ensure that new SNPs that have been curated in the public repositories (dbsnp, Ensembl and Hapmap) can be automatically retrieved from flat files and uploaded onto this database, cleaned and further processed. The second and final phase of the database would contain data that has been processed through the staging database and is ready for querying via a dynamic user interface. Forensic geneticists would be able to extract, through querying the database, potential genetic markers for predicting physical traits.

1.2 Aims and Objectives

Given the above rationale, this project aims to develop a curated database of forensic biomarkers associated with genes that can be used to predict the hair, skin or eye colouring phenotypes as well as the height of the sample source by:

- Identification of genes associated with physical traits of potential forensic relevance. This will be determined from the literature using a text mining platform and manual curation.
- Acquisition of SNPs associated with these genes from the International HapMap project, dbSNP and Ensembl.
- Storage of the data in a MySQL database
- Development of a user-friendly front-end web interface that enhances data retrieval and analysis.
- Characterization of different population groups based on the SNPs

CHAPTER 2: LITERATURE REVIEW

2.1 Forensics: Background and Genetics

Forensic science aims at solving legal issues both in criminal law as well as in civil cases. Major applications of forensic analysis include disputes on kinship, semen detection to resolve rape cases, autopsies for human identification following accident investigations and insurance company fraud investigations. As a specialty, forensic genetics is the application of molecular genetics techniques to resolve these legal cases. In crime scene investigations, forensic genetic analysis is typically applied to determine whether the DNA material obtained matches a target profile (Jobling and Gill 2004; Morling 2004).

Simple observation of family member's resemblance makes clear that phenotypic traits have a genetic basis and should therefore be applicable in identity based-testing and prediction from an individual's DNA. This has, for a long time, been the ambition of many forensic researchers throughout the genetic age (Graham 2008).

As a matter of routine, forensic scientists use STRs, found in the non-coding regions of the human genome, to uniquely identify an individual from the trace DNA elements recovered from the crime scene. An STR is a DNA sequence containing a variable number of repeated short (2-6 base pairs) sequences such as (AATG)*n* found on non-coding region of the human genome. The letter *n* denotes the number of times that the given sequence has been repeated. When STRs are used, estimating the probability that any given two DNA profiles could match by chance is one of the most controversial issues in the forensic use of DNA evidence. In order to draw conclusive inferences from forensic DNA analysis, a sample size greater than 100 individuals of a given population is required. STRs have a very high mutation rate and are highly variable among individuals. STRs also occur in non-coding regions of the genome, as opposed to SNPs which are also present in coding regions and may be responsible for different phenotypic traits. This makes SNPs more informative from a forensic angle. Hence, in the near future forensic investigators will be able to use SNPs to establish the probability that a crime suspect has a certain phenotype. This could be done using trace amounts of DNA material left at a crime scene, and by focusing at the most apparent descriptors of an individual's physical appearance. These include height, colouring and facial features. Resolution of these characteristics at the crime scene would provide crucial probative information and would obviate the need to have a suspect in hand (van Daal 2008; Bianchi and Lio 2007).

Single Nucleotide Polymorphisms (SNPs) are finding use as new markers of interest to the forensic community. Some of the advantages of applying SNPs in the forensic genetics include their low mutation rate as well as their abundance in the human genome. This also makes them very important while dealing with degraded DNA samples. SNPs also offer the possibility of automating sample analysis using high-throughput technologies such as Polymerase Chain Reaction (PCR).

Due to the fact that all genetic differences can be used as markers, DNA has become the most fundamental identification factor. Even with the fact that any two individual human beings share a 99% identity in their genetic constitution, there still exist millions of genetic differences between these individuals. Figure 1 illustrates the major developments of the forensic technology and the increased use of DNA analyses in forensic identification (Bianchi and Lio 2007).

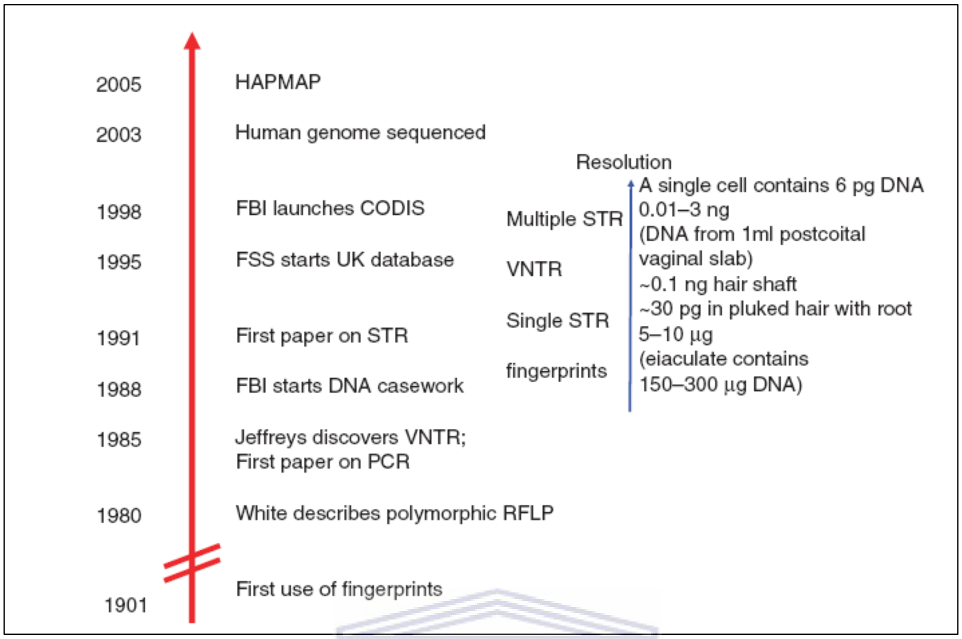


Figure 1: Time scales of major events in forensic DNA typing (left) and examples of resolution power of the different techniques(right) (Bianchi and Lio 2007).

2.2 Physical Traits of Forensic Interest

Scientific evidence shows that the human genome is highly complex, and this leads to a number of polygenic physical traits, such as skin and hair colour being controlled by the interaction of numerous genes together with environmental factors. However, only a limited number of DNA analysis methods available can predict physical traits. These methods are not guaranteed to be accurate. While some characteristics are significantly affected by the environment, others are largely determined as a result of genetic make-up, hence they are highly heritable (Jobling and Gill 2004; Pulker *et al.*, 2007).

SNPs that are phenotype informative can be used to establish a high probability that an individual has a particular phenotypic trait, such as skin color, hair color, or eye color and this could give forensic experts an investigative lead to identify the perpetrator of a crime. Colouring, height and facial features are some of the most common descriptors of an individual's appearance. Since the genetic basis of hair, skin and eye colour is well understood, to date most work on phenotype SNPs has concentrated on these traits. Numerous studies have been carried out on *melanocortin 1 receptor* (*MC1R*) which is the first human gene shown to be associated with normal pigment variation (Budowle and van Daal, 2008).

2.2.1 Pigmentation

Pigmentation is the only relevant physical trait that has undergone serious investigation in forensic genetics (Jobling and Gill 2004). The type and amount of melanin determines human pigmentation, functions of which include photoprotection (van Daal, 2008). Pigmentation traits are highly heritable with a score of 60–90% (Pulker *et al*., 2007).

The Melanocortin 1 Receptor (MC1R) gene is the best studied human pigmentation gene. Its product, melanin, is synthesized from tyrosine as either black/brown eumelanin or yellow/red pheomelanin. This pathway occurs within specialized post-Golgi lysosomal organelles known as melanosomes found in melanocytes (Sturm and Frudakis, 2004). Stimulation of MC1R by alpha-Melanocyte Stimulating Hormone (α-MSH) leads to an increase in cyclic Adenosine mono-Phosphate (cAMP) and the production of eumelanin pigment. Antagonism of this interaction results in a cAMP decrease and production of pheomelanin pigment. Tyrosinase can catalyse the hydroxylation of tyrosine to 3, 4-dihydroxyphenylalanine (DOPA), and the oxidation of DOPA to DOPAquinone (Fig. 2). The eumelanin pigment (black or brown) is derived from the DOPAchrome metabolism, whereas the pheomelanin pigment (red or yellow) is derived from the metabolism of 5-S-cysteinylDOPA. These pigments are eventually deposited over the nucleus of the keratinocytes, giving rise to pigmentation phenotype of hair and skin colour (Parra 2007; Sturm *et al*., 2001).

Several MC1R polymorphisms have been identified that show a strong association with red hair and fair skin. Most individuals with red hair are compound heterozygotes or homozygous for one MC1R polymorphism (Branicki *et al.* 2007).

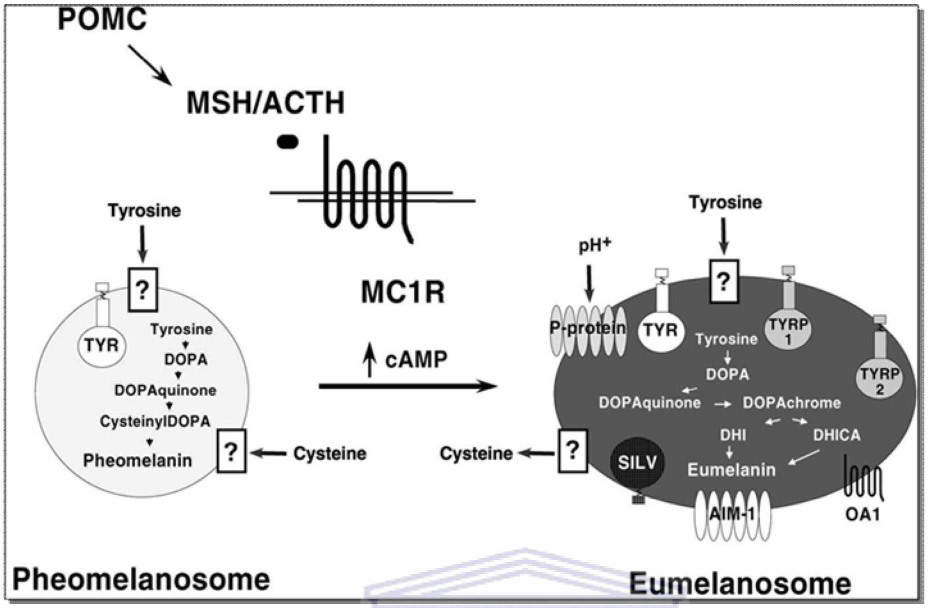


Figure 2: Metabolic pathway for the systhesis of eumelanin and pheomelanin (Sturm *et al*., 2001)

According to Duffy *et al* (2007) while no single gene has been determined for eye colour, the Oculocutaneous Albinism II (OCA2) gene has been found to be a significant contributor. This gene accounts for roughly 74% of the total variation in people's eye color. The effects of *OCA2* expression and its correlation to human pigmentation have been linked to three SNPs that occur near its loci. Forensic genetics studies show that a large number of polymorphisms within a large number of genes are associated with iris colors, suggesting the complexity of the genetics of iris colour pigmentation (Frudakis *et al*., 2003).

The evolution of new traits has enabled humans to adapt to challenging environmental conditions. Human skin color which is regulated by the expression of melanin is an environmental adaptation to different levels of exposure to ultraviolet (UV) rays. This has been observed with people indigenous to Northern Europe having pale skin, compared to people indigenous to Africa who have dark skin (figure 3). Melanin production is increased at low altitudes to protect against continual exposure to irradiation. A decrease in melanin activates the Vitamin D pathway, hence offering a variety of health benefits, which include protection against rickets (osteomalacia) for light skinned individuals (Anno *et al*., 2008). UV light strength varies with latitude and therefore melanin levels correlate with geographical location (Parra 2007).

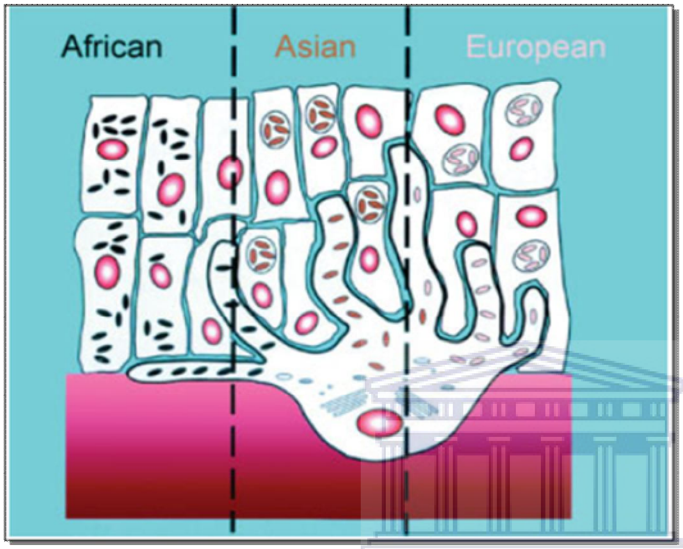


Figure 3: Pigmentation differences among the different population groups highlighting the gradient of melanosome size and number in dark (African), intermediate (Asian), and light skin (European) (Barsh 2003).

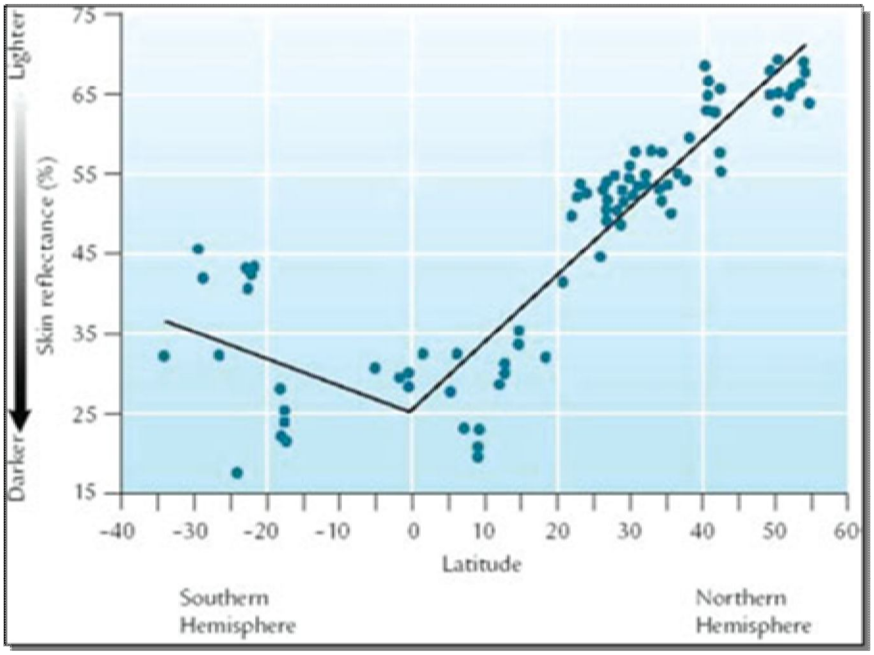


Figure 4: Relationship of skin reflectance with latitude. This depicts a trend whereby lighter skin is observed with increasing latitude and darker skin with decreasing latitude as one moves from the equator (Parra 2007).

In the equatorial and tropical areas, primarily sub-Saharan Africa, South Asia, Australia and Melanesia, skin pigmentation tends to be darker as compared to the areas located far from the equator. Based on reflectometry, a correlation of skin pigmentation and latitude (figure 4), has been scientifically established. The ultraviolet ray intensity explains the relationship between pigmentation and latitude. UVR intensity is greater at the equator but shows a progressive decrease with an increase in latitude (Parra 2007).

Considerable scientific evidence links immunity and melanization, from a genetic, biochemical and functional angle. Recent findings regarding Attractin, a protein with functions in regulating both melanization and immunity has further strengthened this link. Through a mechanism that is still under research, Attractin mediates the inhibitory reaction of agouti and agouti related protein against alpha-melanocyte stimulating hormone (α-MSH). α-MSH binds to the melanocortin-1 receptor (MC1R), whose activation is necessary for the biosynthesis of eumelanin. Attractin is found in large numbers in the cells of the immune system and especially the activated T-cells. Once released in to the serum, Attractin promotes the macrophage dispersal. The immunoregulatory functions of α-MSH have been well documented in previous studies (Mackintosh 2001).

Certain pigmentation genes such as MC1R are characterized by alleles that are more prominent in certain populations. For example, MC1R has been extensively researched in people with red hair and Caucasians. Variations in Agouti Signaling Protein (ASIP) and OCA2 may play a shared role in shaping light and dark pigmentation on a global scale, whereas genes like Solute Carrier Family 24-member 5 (SLC24A5), Membrane Associated Transporter Protein (MATP), and Tyrosinase (TYR) have a predominant role in the development of light skin in Europeans but not in East Asians (Norton *et al*., 2007).

2.2.2 Height and Stature

Height which is a complex and highly heritable trait, is associated with a number of genes which are less well understood. Many hormones, growth factors and transcription factors mediate bone elongation which plays a role in height. The main hormonal regulator is growth hormone, which mediates bone growth through the action of insulin-like growth factor1. The Vitamin D Receptor (VDR) gene which codes for the Vitamin D receptor also plays a role in height through bone elongation. Polymorphisms that show associations to adult height have previously been identified in a number of genes including Aromatase (CYP19), VDR and estrogen receptor α (ERα) (Dahlgren *et al*., 2008, Budowle and van Daal, 2008 , van Daal, 2008).

2.3 Single Nucleotide Polymorphisms

A total of 99.9% of an individual’s DNA sequences is identical to that of another person. Of the 0.1% difference, over 80% are SNPs. An SNP is a single base substitution of one nucleotide with another, where both versions are observed in the general population at a frequency greater than 1%. In an entire human genome there are approximately 10-30,000,000 potential SNPs. The abundance of SNPs means that they could potentially play a role in the future of differentiating individuals from one another (Butler, 2005). A point mutation is a single base substitution whereby one base within an organism's genome is replaced with another (Freese, 1959). Nonsense point mutations code for a stop codon which ultimately truncates a protein during the translation process. Missense mutations code for amino acids that are different from the ones generated in non-mutants. Silent mutations, which code for the same or a different amino acid, do not lead to a functional change in the final protein (Salisbury *et al*. 2003). SNPs may arise from point mutations when their frequency in the population is greater than 1% (Wang *et al*., 1998).

DNA is comprised of only four chemical entities, i.e. Adenine, Guanine, Cytosine, and Thymine (Butler, 2005). The information available for growth and maintenance within the body is determined by the order and sequence of these bases. Complimentary bases pair up with each other to form base pairs. The base 'A' pairs with 'T', while 'G' pairs with 'C'. Each base is also attached to a pentose sugar molecule and also to a phosphate molecule, forming a nucleotide. Nucleotides occur in two long strands that form a double helix structure.

DNA variation is exhibited in the form of different alleles which are the various possibilities at a particular gene locus. Both STRs and sequence variations such as SNPs exist in human populations (figure 5) (Butler 2005). These forms of variation enable forensic DNA identification because many different alleles can exist in both coding and non-coding regions of the genome. When multiple unlinked markers are used in forensic DNA testing, high powers of discrimination can be achieved.

An example of an SNP is observed when a given individual "W" has the sequence GAACCT while individual "Z" has sequence GAGCCT. The third base 'A' has been replaced by a 'G'. The polymorphism in this instance can be expressed as an allele A/G provided that the frequency is above the 1% threshold in the population for each allele.

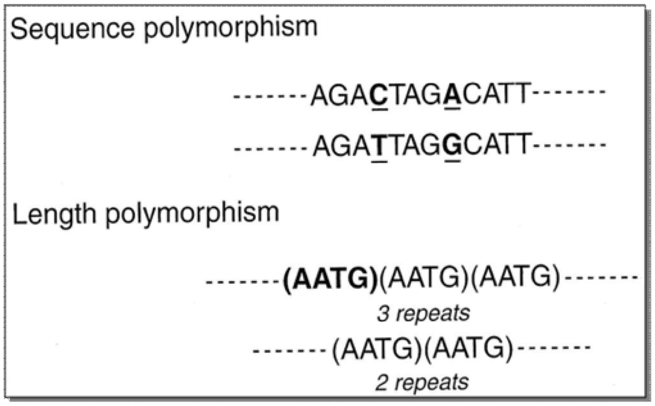

Sequence polymorphism
-------AGACTAGAACATT-------
-------AGATTAGGCATT-------
Length polymorphism
-------(AATG)(AATG)(AATG)-------
3 repeats
-------(AATG)(AATG)-------
2 repeats

Figure 5: An example of Sequence polymorphism (top) and a Length polymorphism (bottom) (Butler 2005).

Although for most human SNPs, only two DNA base variants have been observed, SNPs with three or four alleles still exist. SNPs are less informative in forensic genetics than STRs, which are much more polymorphic. Since SNPs are not as polymorphic as STRs, a larger number of SNPs is needed to obtain the same level of information as that obtained from STRs. Approximately, 40-60 SNPs have an equivalent discriminatory power of 13-15 STR loci. SNP typing is usually done on small fragments of DNA which are roughly 40-50 base pairs in length (Butler *et al*., 2007).

DNA found at a crime scene is usually degraded as a result of exposure to elements within non-physiological conditions. It could also be decomposed through the action of microorganisms. Since a smaller target region is needed for DNA typing such as PCR, a higher recovery of information from degraded DNA samples is theoretically possible with SNPs. Due to the majority of SNPs being biallelic (having two alleles, e.g., C and T) there are only three genotypes (CC, CT, and TT) (figure 6). This highlights the importance of SNP use in crime and in identification cases since only a single nucleotide needs to be measured instead of an array of consecutive nucleotides (Butler *et al*., 2007).

In paternity and especially in immigration cases, SNPs are more informative since mutation rates in general are much lower than those of STRs (Butler 2005). This implies that they are more likely to be fixed in a population hence making them very stable in terms of inheritance. (Morling 2004; Butler 2005; Budowle and van Daal 2008).

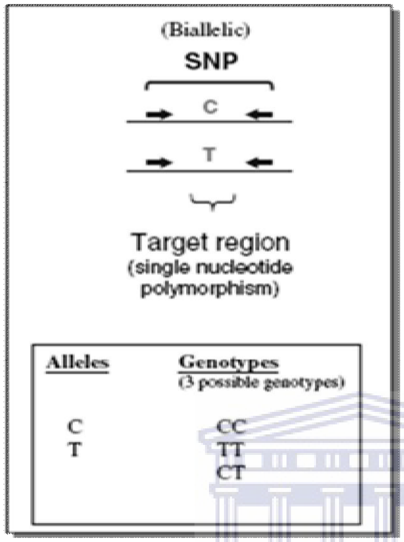


Figure 6: The number of possible allele combinations for a biallelic SNP and relative size compared to STRs of the target region with SNPs (Butler *et al*., 2007).

Human phenotypic traits such as pigmentation and height do not occur as a result of a single variation in the genome or even a single gene (Pulker *et al*., 2007). The polygenic nature of these complex traits means they are most likely influenced by a number of SNPs occurring in a variety of genes that are expressed together in specific combinations. Other factors such as DNA methylation and copy number variations may also contribute to these phenotypes (van Daal 2008). DNA analysis at the crime scene has the potential to provide investigators with all the information needed for suspect identification (Budowle and van Daal, 2008; Butler *et al*., 2007; Jobling and Gill, 2004).

Genetic variation at the level of SNPs may be attributed to most genetic diversity among individuals. This accounts for differences in physical appearance such as colouring, height and other morphological attributes. Mutations in the *MATP* gene have been shown to cause occulocutaneous albinism type 4, while other polymorphisms have been involved in normal pigmentation variation (Graf *et al.*, 2007). Previously, SNPs that affect coding regions and change protein structure and function, had been thought to explain most of the variation in phenotypes. SNPs that occur in non-coding sequences are most often without consequences (Lercher and Hurst 2002). However, recent evidence suggests that variation in non-coding sequences is equally important. If the changed base pair is in the promoter sequence of a gene, the expression of that gene may change. Also, if the SNP occurs in the splice site of an intron, it may interfere with correct splicing of the transcribed pre-mRNA (Kim and Borevitz 2006).

2.4 SNP Data Resources and Tools

A wide variety of tools and databases for SNP discovery, assay design, tagging, and functional analysis are publicly available and free to the scientific community. This project makes use of the International Hapmap project, dbSNP andEnsembl as described below. These three resources have shared data that can be accessed from any single point through hyperlinks.

2.4.1 The International HapMap Project

The main goal of the International HapMap Project is to determine the common patterns of DNA sequence variation within the human genome and to freely avail it to the general public. An international consortium has been put in place to create a map of these patterns across the genome by determining the genotypes of one million or more sequence variants, their frequencies and the level of association, in DNA samples from populations with ancestry across parts of Africa, Asia and Europe. The HapMap Web site at http://www.hapmap.org and to reflect the international nature of the project, the site is available in the languages of the countries that participated in the project, namely English, French, Chinese, Japanese, and Yoruba. SNP data in

Hapmap has been acquired from eleven ethnic groups in an international effort initiated and coordinated by the HapMap Consortium (Barnes 2006; Thorisson *et al*., 2005).

2.4.2 dbSNP

dbSNP serves as a central, public repository for genetic variation. Once the variations have been identified and catalogued in the database, additional laboratory experiments can be carried out by geneticists using that information. dbSNP links variations to other resources within the National Center for Biotechnology Information (NCBI). The data in dbSNP is available freely in a variety of forms. dbSNP build 129 was used to extract the SNP data for this project. SNP data in dbSNP is gathered from assays submitted by individual researchers as well as research groups (figure 7). (Sherry *et al*., 1999; Sherry *et al*., 2001).

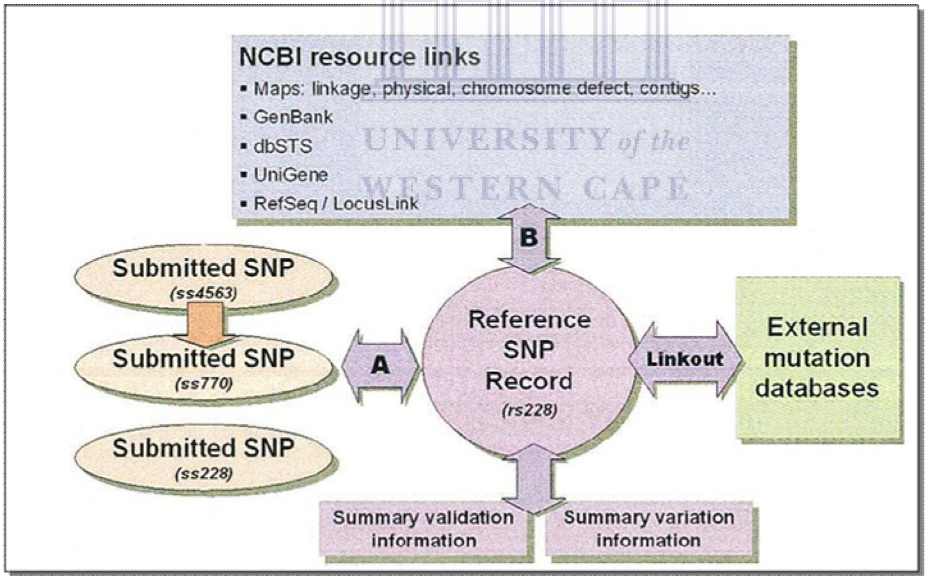


Figure 7: An overview of dbSNP resources at the NCBI (Sherry *et al*., 1999).

2.4.3 Ensembl

The Ensembl (http://www.ensembl.org/) database project aims at providing a Bioinformatics framework to organize and manage data around the sequences of large genomes. It is a complete source of stable automatic annotation of the human genome sequence, and is available as either an interactive web site or as flat files. Ensembl is a joint scientific project between the European Bioinformatics Institute and the Wellcome Trust Sanger Institute, that annotate and display genome information. SNP data used by this project was extracted from Ensembl version 53. Other SNP resources include the UCSC genome browser system (Karolchik *et al.*, 2007) and the NCBI genome resources (Birney *et al.*, 2004; Hubbard *et al.*, 2002; Wheeler *et al.*, 2001).

2.4.4 Other Forensic SNP Resources

Numerous SNP tools and databases have been created for SNP-phenotype prediction and SNP related information. However, none exists for providing information regarding SNPs associated with genes of forensic interest. The Combined DNA Index System (CODIS) is the FBI-funded computer-based system which allows investigators to search and locate DNA profiles such as those for multiple offenders (Phillips 2008). SNPforId (http://www.snpforid.org) is a web-based tool for the query and visualization of the SNP allele frequency data generated by the SNPforID consortium. The web interface allows visitors to review the allele frequencies of the studied markers from all the available populations used by SNPforID to validate global SNP variability (Amigo *et al*., 2008).

The Forensic SNP resource (http://www.cstl.nist.gov/div831/strbase/SNP.htm.) aims at providing general information on SNP markers that may be of interest in human identification applications. Most of these markers come from The SNP Consortium (TSC) efforts. MitoMap (http://www.mitomap.org) is a database of polymorphisms and mutations of the mitochondrial DNA. Others include Forensic Science Service (http://www.forensic.gov.uk/forensic_t/index.htm), European DNA Profiling group (EDNAP) (http://www.rechtsmedizin.uni-mainz.de/Remedneu/ednap/ednap.htm) and The European Network of Forensic Science Institutes (ENFSI) (http://www.enfsi.org/).

2.5 The Bioinformatics Approach

Bioinformatics and forensic DNA analysis are inherently interdisciplinary in that they draw their techniques from statistics and computer science (Bianchi and Lio 2007). Bioinformatics tools offer a computerized platform for the quick discovery of novel polymorphisms and the dissemination of genetic data and research findings (Walsh 2004).

2.5.1 Biomedical Text and Data Mining

Knowledge Discovery in Databases (KDD) (figure 8) refers to the overall process of discovering useful knowledge from data, while data mining refers to a particular step in this process. (Fayyed *et al*., 1996).

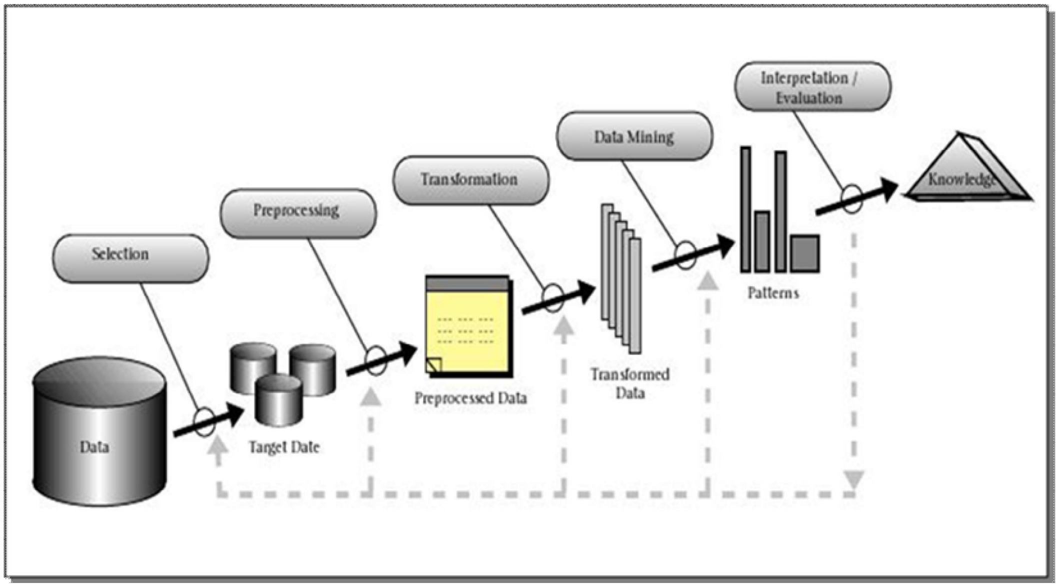


Figure 8: Steps involved in knowledge discovery (Fayyed *et al*., 1996).

The first step involves selecting a data source on which discovery is to be made. Once the data has been identified, it has to be preprocessed. This usually involves steps such as deciding how to represent missing data. Preprocessed data is further refined through transformation and then the actual mining is carried out. Different patterns that are observed in this step are later analysed and presented as new knowledge (Fayyed *et al*, 1996).

With the frequently overwhelming increase in the volume of scientific literature, text-mining is becoming increasingly important in extracting and summarizing information. Text-mining also ensures convenience and efficiency and can also be applied in visualization of important potential associations between different concepts in clear graphical representations (Kaur *et al*., 2009; Sagar *et al*., 2008).

The goal of biomedical text mining is to allow researchers identify needed information more accurately. Text mining also facilitates establishment of relationships hidden by the large amounts of available data, and hence move the burden of information overload from the researcher to the computer by applying algorithmic, statistical and data analysis methods to the huge amount of biomedical knowledge that is currently in the literature (Cohen and Hersh 2005).

The Dragon Explorer System (DES), http://apps.sanbi.ac.za/des/index.php is a web based, text-mining tool for post-processing large numbers of PubMed queries. It uses a dictionary based text-mining approach to extract information from text documents (Sagar *et al*., 2008). DES also links together the extracted information and forms new hypotheses which could be later explored or validated using experimentation.

Text mining aims at discovering unknown information. One of the biggest questions in genomics which could be answered by the use of text mining is the discovery of protein to protein interactions. The key lies in not looking for direct occurrence of pairs but rather for articles that mention individual protein names, keeping track of related names and then finding the same set of words in other articles (Hearst 2003).

Text mining mainly involves the process of structuring the input text. This structuring involves the parsing, addition of some features, removal of other and the subsequent entry into a database. The final step involves the derivation of patterns within the structured data and the evaluation and interpretation of the output. Many current literature mining approaches rely on statistical analysis of word occurrences which are calculated over the whole PubMed database and the results are weighted associations between biological entities. The underlying principle of these global strategies is that if two biological entities commonly occur together or appear in similar contexts, then they could have some biological relationship (Krallinger *et al*., 2008).

2.5.2 Relational Databases

A database is an ordered collection of data, which is normally stored in one or more associated files. The data is stored in tables, where cross reference between those tables is possible. The existence of relationships among these tables leads to the database being called a relational database (Kofler 2005).

A relational database is comprised of multiple tables each pertaining to a specific topic. Each table contains a field, or a combination of fields, where the data value uniquely identifies the record. This field is referred to as the Primary Key and is a unique identifier for a given record**.** Hence, the tables are related to each other by a common field. A foreign key is a column in a table whose values must be listed as a primary key in another table. Foreign keys are an extremely important part of ensuring referential integrity. The relational database model allows files to be related by means of a common field. A one-to-many relationship exists when the primary record can have many related records. A one-to-many relationship is created if only one of the related fields is a primary key or has a unique constraint. In a many-to-many relationship, a record in one table can have many matching records in a different table, and vice versa. Such relationships are defined by using a third table, called a joining table, whose primary key consists of the foreign keys from both tables

(Forta 2004; Hernandez 2003).

Database normalization, (figure 9) is a technique that reduces the occurrence of data anomalies and poor data integrity. It is governed by a set of rules called normal forms and involves breaking down large tables into smaller ones in order to eliminate redundant data. It is also carried out to avoid problems with inserting, updating, or deleting data (Hernandez 2003).

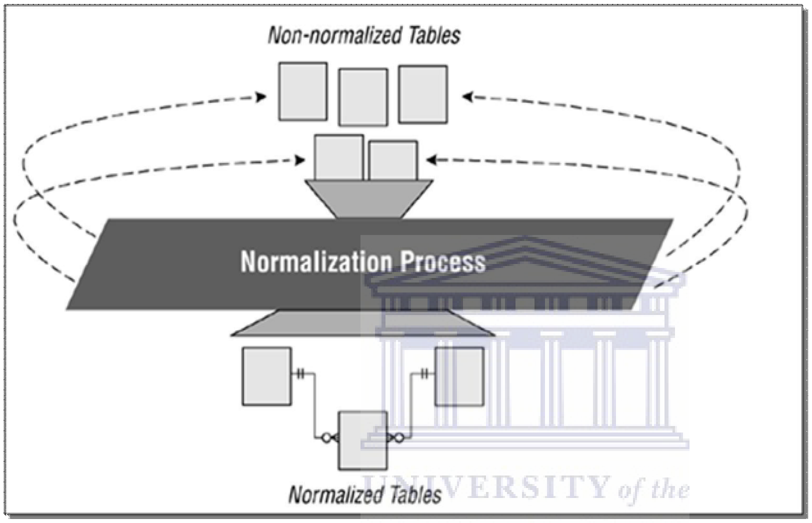


Figure 9: A graphical representation of the general normalization process (Hernandez 2003).

Structured Query Language (SQL) is a database programming language that is used to create, modify, maintain, and query relational databases. SQL is relatively easy to learn and allows people to quickly learn how to perform queries on a relational database. MySQL is a relational database system which includes the programs for managing relational databases. The tasks of a relational database system include: secure storage of data; processing of commands for querying, analyzing, sorting existing data (Kofler 2005; Delisle 2006)

CHAPTER 3: RESEARCH DESIGN AND METHODOLOGY

Figure 10 summarizes all the main steps in the development of the Forensic SNP Phenotype Database. All these steps have been thoroughly discussed throughout the chapter.

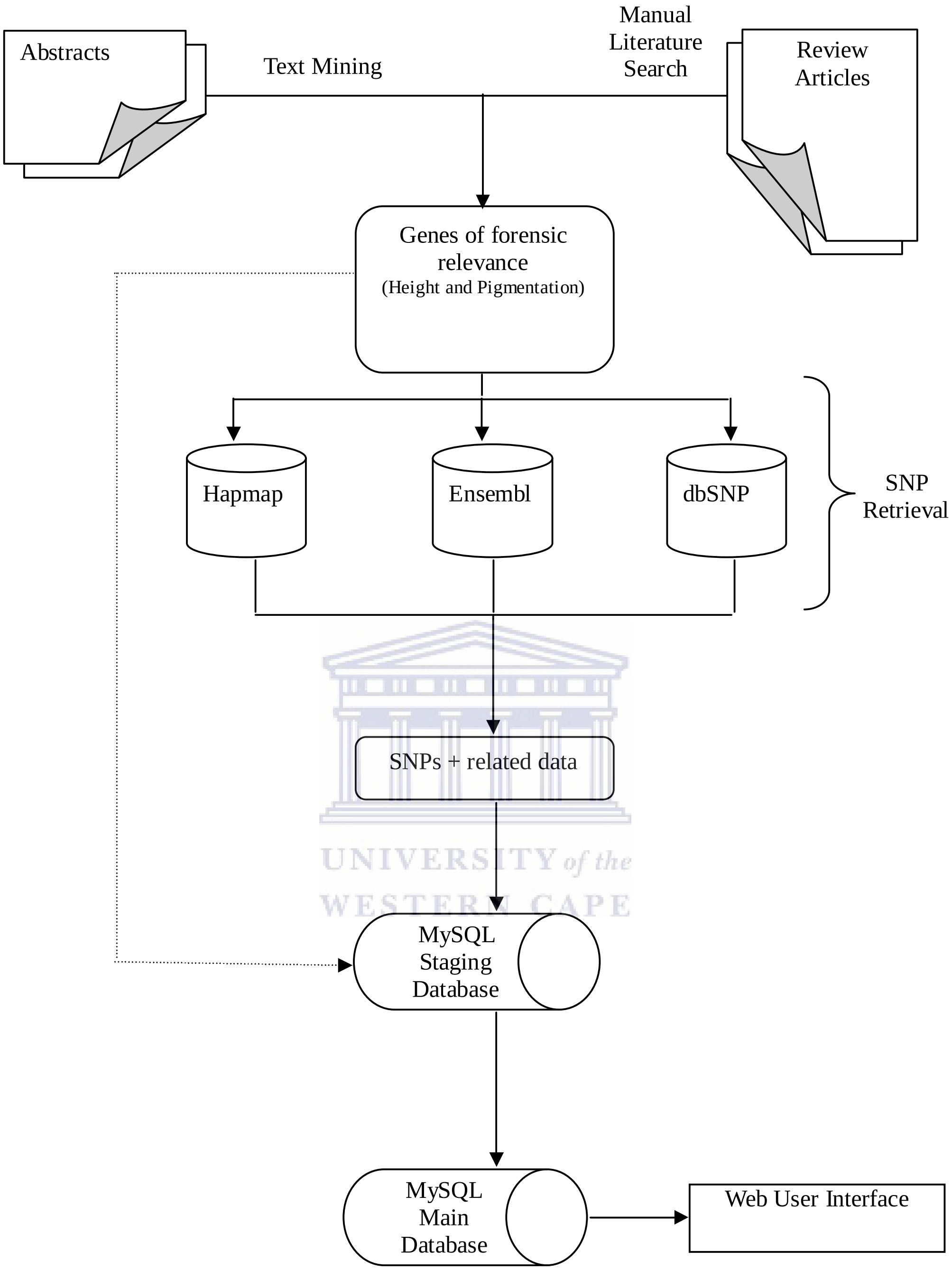


Figure 10: An overview of the research methodology

3.1 Gene Selection

Physical traits are observable characteristics determined by specific genes or a set of genes. While a few traits are due to only one gene (and its alleles), most human genetic traits occur as a result of interactions between several genes (Barsh 2003). A non-redundant list of genes associated with physical traits, compiled from the literature and text-mining, forms the basis of the forensic SNPs database. The physical traits, pigmentation and height, were selected as an initial focus due to their being currently the most well studied in the literature with regard to SNPs. All the SNPs linked the genes associated with these traits were then retrieved using manual literature searching and text-mining.

3.1.1 Manual Literature Curation

Manually searching the literature for relevant data is the gold standard for textual information retrieval. In searching for genes associated with physical traits, specific PubMed search queries were mainly directed at finding review papers in forensic journals. Specifically, forensic, skin, eye, and hair pigmentation keywords were used. This was done to ascertain the number of physical trait genes already mentioned in forensic-specific journals. By choosing the most recent review papers, this would capture the most updated gene-phenotype information. A list of defined genes, mined directly from the literature by hand for height and pigmentation physical traits, was compiled.

3.1.2 Text-Mining

The aim of text-mining in this project was to obtain a gene list of genes associated with physical traits of interest namely, pigmentation and height in a fast and efficient way by inspecting gene co-occurrences in abstracts from PubMed.

The Dragon Explorer System (DES) uses a dictionary based text-mining approach for extracting potentially relevant information from text documents. The dictionaries contain numerous variants of names and symbols customary, compiled from the literature and published databases, for specific types of entities (Sagar *et al*., 2008).

These PubMed searches were ran to include as many abstracts which may contain information on genes associated with phenotypic traits (specific relevance to forensics or anthropological), instead of conducting a very specific search and limiting the search to forensic keywords and lowering the chance of detecting any potential genes for physical traits of forensic interest.

Two PubMed search queries, downloaded on 24th of January 2009 and 5th of February 2009 for query one and query 2 respectively, were performed and the abstracts which were formatted in XML format were retrieved.

Query 1: (Physical trait OR physical traits) AND (gene OR genes) AND human

Query 2: gene AND (facial OR physical trait OR physical attribute* OR physical phenotype OR physical appearance OR physical characteristic) NOT disease NOT Disorder NOT therapy NOT cancer

The gene list obtained from the manual literature review and the one from the automated text mining procedure were integrated into a final nonredundant gene list. This is the gene list which was used for the SNP retrieval. From this integrated gene list, the gene descriptions such as gene symbol, approved gene name, and Entrez gene ID were matched against the HUGO Gene Nomenclature Committee (HGNC) official gene symbols. The HGNC is part of the Human Genome Organisation (HUGO), which approves unique gene names and symbols for every human gene (Povey *et al*., 2001). These symbols are universal amongst scientific resources. This ensures that these gene symbols can be used in public bioinformatics resources such as Ensembl, HapMap and NCBI. All the genes without a HGNC symbol they were excluded. The composition of the final gene list was presented in a table (Appendix 1).

3.2 SNP Retrieval

SNPs were extracted for the gene list of interest using chromosome co-ordinates from the BioMart interface of the Ensembl gene annotation platform. These co-ordinates were then extended by 10,000 base pairs both upstream and downstream of the gene. This was done to ensure that SNPs that may potentially be located in the regulatory regions of the gene were retrieved. The gene chromosomal coordinates were used as input and SNPs were retrieved from dbSNP build 129, HapMap Release 27, and Ensembl version 53. This was done to ensure that as many SNPs as possible were retrieved and hence the SNP list was as comprehensive as possible.

3.3 Phenotype data

Phenotype information was retrieved from BioMart using the Ensembl variation 56 database and the Online Mendelian Inheritance in Man database (OMIM) at the NCBI. This was done through extraction of the literature-confirmed SNPs, which are associated with normal pigmentation and height traits. A final list of specific traits associated with phenotypes was compiled.

3.4 Population and other SNP Related Data

An important aspect of this research project is to enable forensic experts characterize the SNPs by their predominant populations. Population information was acquired from the Hapmap (Appendix 4) and also from the dbSNP (Appendix 3). For a comprehensive SNP database, additional SNP related data was compiled. The gene coordinates (Appendix 2) shows the chromosomal coordinates for each of the gene in the gene list. These coordinates extend up to 10,000 bases both upstream and downstream of the gene. SNP descriptions were acquired from dbSNP (Appendix 5).

3.5 Database Construction

In order for this resource to be of use to the forensic community, the database has been presented as a collection of Single Nucleotide Polymorphism (SNP) genotyping information. A user friendly web interface has also been implemented and is available at http://forensic.sanbi.ac.za/alecia_forensics/Index.html

3.5.1 Database Design

The Forensic SNP Phenotype Database consists of 11 tables with each representing a different entity within the database model. Unique identifiers or Primary Keys have been assigned to each table having been constructed from the table name with the prefix of 'id_'. The 'UNIQUE' constraint solely identifies each record in a database table and provides a guarantee for uniqueness for a column or set of columns. These keys exclusively identify each table throughout the database and have also been used to establish relationships between the various entities. Relationships ensure that the database is easily accessed, managed and updated through a series of logical Structured Query Language (SQL) statements. Primary keys are of a serial nature, set as an "auto increment" feature, so as to facilitate addition of extra fields at a later stage.

The gene table (Item 1), has been designed to store the Entrez gene id, a unique identifier assigned to each gene that is curated by the NCBI. Due to the importance and nature of this field within the database, it cannot be empty and has hence been assigned the 'NOT NULL' attribute. The only other field within this table is the 'id_gene' primary key field. The entrez_id field has been tagged as unique by assigning a unique index constraint on it.

```
CREATE TABLE gene (
        id_gene INT UNSIGNED NOT NULL AUTO_INCREMENT,
        entrez_id INT NOT NULL,
        PRIMARY KEY pk_gene (id_gene)
);
CREATE UNIQUE INDEX gene___unique ON gene(entrez_id);
```

Item 1: An SQL statement for creating the gene table

The gene_description table carries the Entrez gene ID, the official gene symbol, the full name of the gene and the gene's alias terms of reference as documented by the Human Genome Organization (HUGO) consortium. With the exception of the table's primary key and Entrez gene fields, which are numerical and hence have been assigned 'Integer' attribute, the rest of the fields are textual and all carry a variable character (VARCHAR) attribute. Item 2 shows the SQL statement used to create the gene_description table and assign its unique indices.

```
CREATE TABLE gene_description (
        id_gene_description INT NOT NULL AUTO_INCREMENT,
        entrez_id INT NOT NULL,
        gene_symbol VARCHAR(45),
        gene_name VARCHAR(250),
        aliases VARCHAR(250),
        PRIMARY KEY pk_gene_description (id_gene_description)
);
CREATE UNIQUE INDEX gene_description___unique ON
gene_description(entrez_id,gene_symbol,gene_name);
```

Item 2: An SQL for creating the gene_description table

The gene coordinate mappings (Item 3), as curated by the Ensembl genome server, have been stored in the gene_coordinate table. Other than the Primary Key, the other fields in the table include the chromosome number, the gene start loci as well as the gene ending loci.

```
CREATE TABLE gene_coordinates (
        id_gene_coordinates INT NOT NULL AUTO_INCREMENT,
        entrez_id INT NOT NULL,
        chromosome VARCHAR(20),
        start_pos INT,
        stop_pos INT,
        PRIMARY KEY pk_gene_coordinates (id_gene_coordinates)
);
CREATE UNIQUE INDEX gene_coordinates___unique ON gene_coordinates
(entrez_id,chromosome,start_pos,stop_pos);
```

Item 3: An SQL statement for creating the gene_coordinates table

The trait table (Item 4) contains data regarding height and pigmentation, which are the two broad phenotypic traits of forensic interest.

```
CREATE TABLE trait (
        id_trait INT NOT NULL AUTO_INCREMENT,
        trait VARCHAR (45) NOT NULL,
        PRIMARY KEY pk_trait(id_trait)
);
CREATE UNIQUE INDEX trait__unique ON trait(trait);
```

Item 4: An SQL statement for creating the trait table

The gene_trait_general (Item 5) table is a joining table for the gene and the trait tables. The only fields marked for this table are the Primary Key, the Entrez gene ID and the physical characteristic (Height or Pigmentation). The two broad traits pigmentation and height were mapped to every gene in the genelist. All the genes are either associated with height or pigmentation, with the exception of one gene, PPARD, which codes for both pigmentation and height. This confirms that there exists a many-many relationship between the gene and the trait tables.

```
CREATE TABLE gene_trait_general (
            id_gene_trait_general INT NOT NULL AUTO_INCREMENT,
            entrez_id INT NOT NULL,
            trait VARCHAR(255),
            PRIMARY KEY pk_gene_trait_general (entrez_id,trait)
);
CREATE UNIQUE INDEX gene_trait_general__unique ON
gene_trait_general(id_gene_trait_general);
```

Item 5: An SQL statement for creating the gene_trait_general table

The high resolution trait table (Item 6) has been created to store specific height and pigmentation features such as blond hair. The literature confirmed phenotype attributes were extracted from the Ensembl database and the OMIM repository at the NCBI.

```
  CREATE TABLE high_res_trait (
         id integer INT NOT NULL AUTO_INCREMENT,
         trait VARCHAR(255),
);

 CREATE UNIQUE INDEX high_res_trait__unique ON high_res_trait(trait);
```

Item 6: An SQL statement for creating the high resolution trait table.

All the Single Nucleotide Polymorphism identifiers have been stored in the ‘snp’_table. These identifiers which have been collected from the dbSNP, the HapMap Project and Ensembl are all composed of a unique number that has an ‘rs’ prefix. Item 7 shows the SQL statement that was used to create the SNP table.

```
CREATE TABLE snp (
        id_snp INT NOT NULL AUTO_INCREMENT,
        rs_id VARCHAR(45) NOT NULL,
        PRIMARY KEY pk_snp (id_snp)
);
CREATE UNIQUE INDEX snp__unique ON snp(rs_id);
```

Item 7: An SQL statement for creating the SNP table

The SNP_gene table (Item 8) maps each SNP to a given gene. Both the entrez_id and the rs_id fields were set as the primary key field for this table so as to facilitate its role as a mapping table for the gene and the SNP tables.

```
CREATE TABLE snp_gene (
        id_snp_gene INT NOT NULL AUTO_INCREMENT,
        entrez_id INT NOT NULL,
        rs_id VARCHAR(45),
        PRIMARY KEY pk_snp_gene (entrez_id,rs_id)
);
CREATE UNIQUE INDEX snp_gene__unique ON snp_gene(id_snp_gene);
```

Item 8: An SQL statement for creating the snp_gene table

The SNP_description table (Item 9) contains all the main features of an SNP.

```
CREATE TABLE snp_description (
	id_snp_description INT NOT NULL AUTO_INCREMENT,
	rs_id VARCHAR(45) NOT NULL,
	allele VARCHAR(45),
	chrm_pos VARCHAR(45),
	snp_class VARCHAR(45),
	fxn_class VARCHAR(255),
	flanking_sequence VARCHAR(255),
	validation VARCHAR(255),
	PRIMARY KEY pk_snp_description (id_snp_description)
);
CREATE UNIQUE INDEX snp_description__unique ON snp_description(rs_id);
```

Item 9: An SQL statement for creating the SNP_description table

The pop_and_freq_dbsnp table (Item 10) contains the dbSNP genotyping information for any given SNP. The pop_and_freq_hapmap was modeled on this table and contains the genotyping information for SNPs acquired from the HapMap repository.

```
CREATE TABLE pop_and_freq_dbsnp (
	id_pop_and_freq_dbsnp INT NOT NULL AUTO_INCREMENT,
	rs_id VARCHAR(45) NOT NULL,
	observed_allele VARCHAR(10),
	observed_allele_freq VARCHAR(10),
	observed_allele2 VARCHAR(10),
	observed_allele2_freq VARCHAR(10),
	pop_class VARCHAR(255),
	gen_pop VARCHAR (255),
	PRIMARY KEY pk_pop_and_freq_dbsnp(id_pop_and_freq_dbsnp)
);

CREATE UNIQUE INDEX pop_and_freq_dbsnp__unique ON
pop_and_freq_dbsnp(rs_id,observed_allele,observed_allele_freq,observed_allele2,obs
erved_allele2_freq,pop_class,gen_pop);
```

Item 10: An SQL statement for creating the pop_and_freq_dbsnp table

As Highlighted for the case of gene_description table in Item 11, a unique index can be assigned to several fields, all at once, within a table hence virtually coupling their occurrence within the table.

```
CREATE UNIQUE INDEX gene___unique ON gene(entrez_id);

CREATE UNIQUE INDEX gene_description___unique ON
gene_description(entrez_id,gene_symbol,gene_name);
```

Item 11: An SQL statement for adding a unique constraint on key fields

A staging database was created prior to populating the main database with data. This was carried out to facilitate pre-processing and the transformation of data. The staging database does not have any constraints on the tables, or foreign key constraints, hence no established relationships among the various entities.

After retrieving the data from their various sources in various formats such as XML and Tab Separated Values (TSV), the data sets were transformed into Comma Separate Values (CSV) format. This data was imported into the staging database by means of SQL scripts. A staging database had to be used for processing data for this set of tables. If data needs to be added at a later stage, raw data can be imported to a staging database rather than altering each table separately.

All of the raw datasets were first imported into a staging database in CSV format before being transformed and imported into the “main database”. Transformation involved the removal of duplicated and erroneous datasets (Item 12). The main database ensures that data integrity is maintained, since table constraints were enforced and foreign keys were established.

```
SELECT a.* FROM SNP_gene a LEFT JOIN SNP b ON b.rs_id=a.rs_id WHERE
b.rs_id IS NULL;

DELETE FROM SNP_gene USING (SELECT a.* FROM SNP_gene a LEFT JOIN SNP
b ON b.rs_id=a.rs_id WHERE b.rs_id IS NULL)a WHERE SNP_gene.rs_id=a.rs_id
AND SNP_gene.entrez_id=a.entrez_id;
```

Item 12: SQL illustrating the removal of erroneous entries from the SNP_gene table which dont map to the SNP parent table

The ‘INSERT’ SQL statement was used to insert rows to a database table. This could be done either to a single complete row, a single partial row or to insert the result of a query. When large datasets were directly imported from an input file to the database, the LOAD statement was used, as highlighted in Item 13.

```
LOAD DATA INFILE 'c:\\Documents and
Settings\\ALECIA\\Desktop\\forensic\\load_data\\pop_and_freq_hapmap\\table_pop_and_
freq_hapmap.csv' INTO TABLE pop_and_freq_hapmap_staging

FIELDS TERMINATED BY ',' ENCLOSED BY '"' ESCAPED BY '\\'
LINES TERMINATED BY '\r\n'
IGNORE 1 LINES
(rs_id,ref_allele,ref_allele_freq,other_allele,other_allele_freq,pop);
```

Item 13: An SQL statement illustrating how the raw dataset were loaded from a CSV file into a staging database table

3.5.2 Populating tables and setting up relationships

The data contained in the staging table was used to populate the “main database” through a series of SQL statements such as the one illustrated in Item 14.

```
INSERT INTO SNP_gene (entrez_id,rs_id)
SELECT DISTINCT entrez_id,rs_id
FROM SNP_gene_staging;
```

Item 14: SQL statement used to populate the SNP_gene table using data from the SNP_gene table of the staging database

Foreign Keys have been described throughout the database. These keys may occur individually or as a group of fields and they point to another key in a given table (usually a Primary Key). This propagates their role in maintaining referential integrity within the database or among a set of tables through relationships. Three types of relationships have been established in this database. These are one-to-one, one-to-many and the many-to-many relationship.

3.5.3 Developing the Graphical User Interface

The Forensic SNP Phenotype Database implements a 3-tier architecture comprising of a database tier at the bottom (figure 11), the application tier in the middle and the client tier on top. A virtual machine named *forensic* was setup at the South African National Bioinformatics Institute (SANBI). Postgresql, Tomcat, Ant and Java were installed on this virtual machine to facilitate development of a dynamic web front end database application.

To achieve a dynamic interaction between the user and the forensic database, Tomcat which is a servlet container developed by the Apache Software Foundation (ASF) implements the Java Servlet and the JavaServer Pages (JSP) specifications which perform various tasks including database access via the Java Database Connectivity (JDBC) driver.

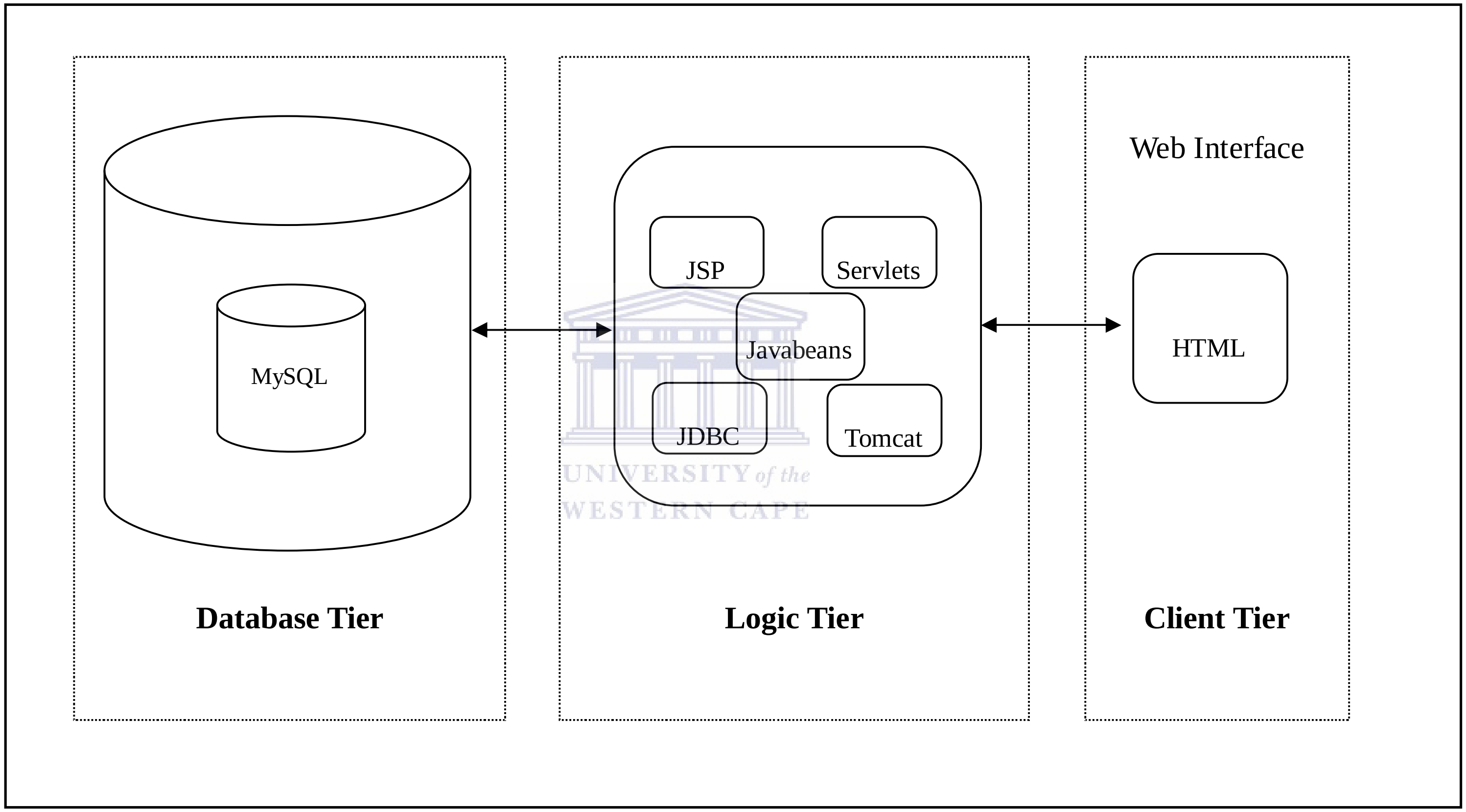


Figure 11: 3-tier architecture for the Forensic SNP Phenotype Database

CHAPTER 4: RESULTS AND DISCUSSION

4.1 Manual Literature Search

A total of 27 pigmentation and 51 height genes were retrieved from searching review journal papers.

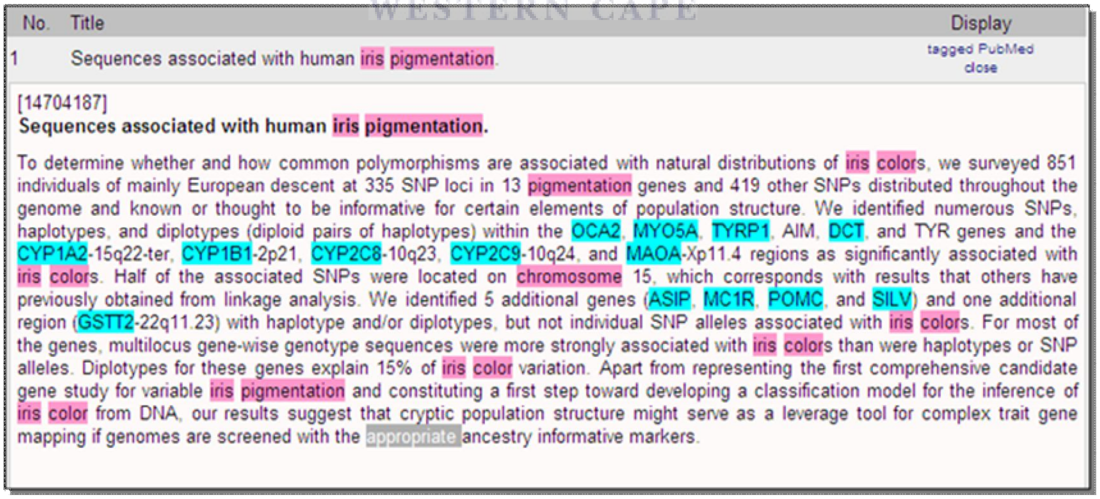
4.2 Text-Mining

From query 1 in the methods, 1257 abstracts were downloaded on the 24th of January 2009. Query 2 yielded a total of 3998 abstracts on the 5th of February 2009. A total of 28 pigmentation and 37 height genes were retrieved from text-mining.

The list of abstracts contained in each query was processed by the DES and the abstracts supporting the linkage of genes and proteins associated with the term 'Pigmentation' and 'Height', were read to determine the validity. DES maps the entities from the dictionaries to the Pubmed abstracts submitted and any entity found in the abstract is highlighted with a specific colour, based on the dictionary. This facilitates identification. The Human Genes and Proteins (HGP) dictionary is colour-coded blue while the Human Anatomy (HA) is pink in colour. An example is shown in Figure 12 where, CYP1A2 (in blue) indicates that this gene has one abstract that contains the entity "pigmentation". Figure 13 shows a list of genes and proteins that potentially have an association with the entity pigmentation, since they occur within the same abstract. The nature of the association and the relevance thereof needs to be determined by manual curation i.e. physical reading of the abstracts.

Once the entity list is displayed, the genes and proteins had to be manually verified through reading of the abstracts and checking for their relevance, figure 13 highlights a sample entity list. Genes associated with any disorders or diseases or which did not mention any association with normal height phenotype were excluded. Only genes which were associated with normal height and pigmentation phenotype were selected.

Once the relevant height and pigmentation genes were filtered, they were integrated with the manually searched literature gene list. These then had to be cleaned and standardized to give a final gene list whose unique identifiers are their HGNC symbols and Entrez gene ids. The cleaning and standardization is a prerequisite since it has been established that when filtering genes, the text-miner for example will count IGF1 and IGF-1 as two different (entities) genes. The two in fact refer to one and the same gene, Insulin-like Growth Factor 1, a putative height gene. This owes to the fact that authors of manuscripts use different synonyms for the same gene.



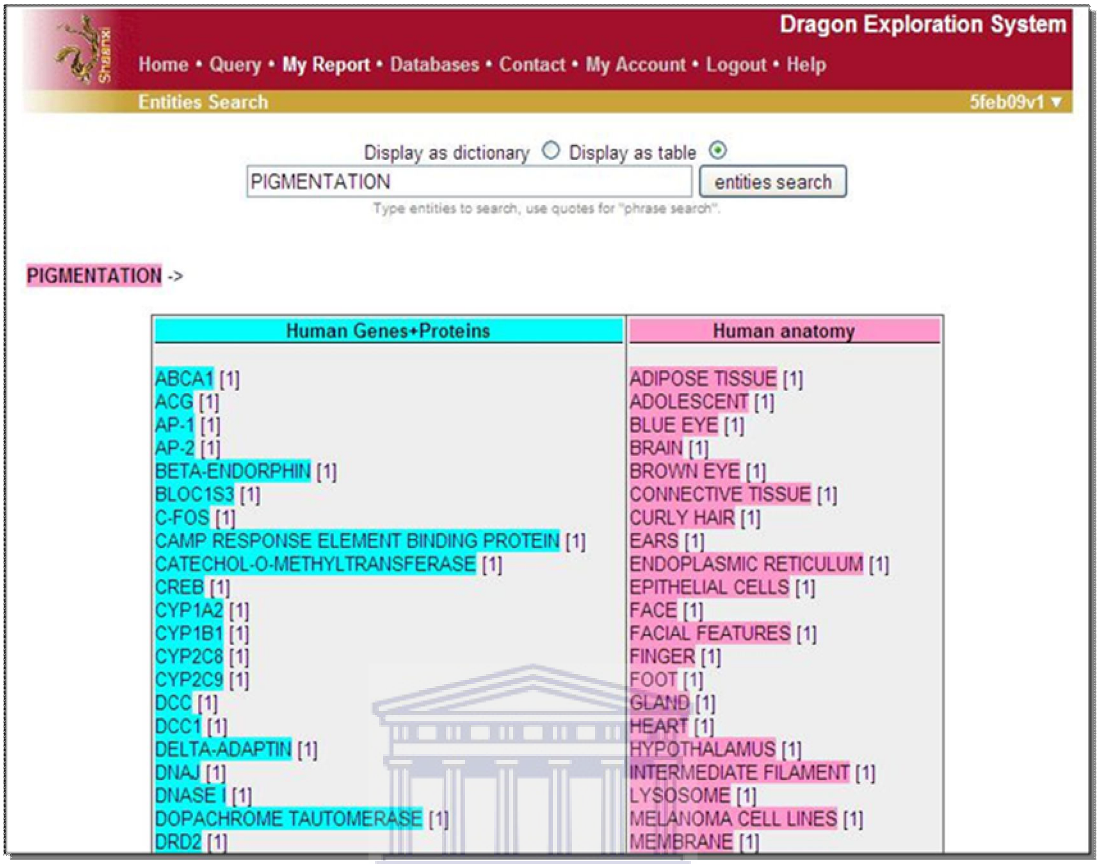
No.	Title	Display
1	Sequences associated with human iris pigmentation.	tagged PubMed close

[14704187]

Sequences associated with human iris pigmentation.

To determine whether and how common polymorphisms are associated with natural distributions of iris colors, we surveyed 851 individuals of mainly European descent at 335 SNP loci in 13 pigmentation genes and 419 other SNPs distributed throughout the genome and known or thought to be informative for certain elements of population structure. We identified numerous SNPs, haplotypes, and diplotypes (diploid pairs of haplotypes) within the OCA2, MYO5A, TYRP1, AIM, DCT, and TYR genes and the CYP1A2-15q22-ter, CYP1B1-2p21, CYP2C8-10q23, CYP2C9-10q24, and MAOA-Xp11.4 regions as significantly associated with iris colors. Half of the associated SNPs were located on chromosome 15, which corresponds with results that others have previously obtained from linkage analysis. We identified 5 additional genes (ASIP, MC1R, POMC, and SILV) and one additional region (GSTT2-22q11.23) with haplotype and/or diplotypes, but not individual SNP alleles associated with iris colors. For most of the genes, multilocus gene-wise genotype sequences were more strongly associated with iris colors than were haplotypes or SNP alleles. Diplotypes for these genes explain 15% of iris color variation. Apart from representing the first comprehensive candidate gene study for variable iris pigmentation and constituting a first step toward developing a classification model for the inference of iris color from DNA, our results suggest that cryptic population structure might serve as a leverage tool for complex trait gene mapping if genomes are screened with the appropriate ancestry informative markers.

Figure 12: An abstract highlighting the pigmentation entity and related entities from the HGP dictionary


Dragon Exploration System
Home • Query • My Report • Databases • Contact • My Account • Logout • Help
Entities Search
5feb09v1 ▼
entities search

Display as dictionary ○ Display as table ●

PIGMENTATION

Type entities to search, use quotes for "phrase search".

PIGMENTATION ->

Human Genes+Proteins	Human anatomy
ABCA1 [1]	ADIPOSE TISSUE [1]
ACG [1]	ADOLESCENT [1]
AP-1 [1]	BLUE EYE [1]
AP-2 [1]	BRAIN [1]
BETA-ENDORPHIN [1]	BROWN EYE [1]
BLOC1S3 [1]	CONNECTIVE TISSUE [1]
C-FOS [1]	CURLY HAIR [1]
CAMP RESPONSE ELEMENT BINDING PROTEIN [1]	EARS [1]
CATECHOL-O-METHYLTRANSFERASE [1]	ENDOPLASMIC RETICULUM [1]
CREB [1]	EPITHELIAL CELLS [1]
CYP1A2 [1]	FACE [1]
CYP1B1 [1]	FACIAL FEATURES [1]
CYP2C8 [1]	FINGER [1]
CYP2C9 [1]	FOOT [1]
DCC [1]	GLAND [1]
DCC1 [1]	HEART [1]
DELTA-ADAPTIN [1]	HYPOTHALAMUS [1]
DNAJ [1]	INTERMEDIATE FILAMENT [1]
DNASE I [1]	LYSOSOME [1]
DOPACHROME TAUTOMERASE [1]	MELANOMA CELL LINES [1]
DRD2 [1]	MEMBRANE [1]

Figure 13: Screenshot highlighting the list of genes (in blue) resulting from pigmentation entity for query2.

Out of a total of 1257 abstracts which were submitted to the Dragon Text-miner from query one, 1113 abstracts had entities from either of the dictionaries found in these abstracts. The dictionaries are the HGP and the HA. A total of 81.1% of the entities were found in the HA dictionary and 18.9 % stemmed from the HA dictionary. Similarly, 3627 abstracts out of a total of 3998, which constitutes 86.3% of the entities, were found in the HGP dictionary. Only 13.7 % were located in the HA dictionary.

It was interesting to note that abstracts retrieved in Query 1 provided no extra pigmentation genes from the genes found in Query 2. However there were 3 extra height-related genes which were not picked up in Query 2 and the literature-searched genes. These genes are the SET binding Factor II (SBF2), Filamin beta (FLNB) and Lamin A (LMNA). Overall, the manual literature searching, together with the text-mining complemented each other resulting in a complete gene list for physical traits of forensic relevance.

Text-mining results are not 100% conclusive and still require hand curation to eliminate non-specific entities and false positives. Acronyms, abbreviations, previous names and aliases are sometimes incorrectly identified as genes or proteins (by the Human Gene Protein dictionary) and may thus need to be well checked for relevance and validity as gene or protein entities.

Entity association is based on the co-occurrence of the entities within the abstracts or sentences. If the two entities are repeatedly mentioned together, there are chances that they are linked directly or indirectly or a relationship exists between the two entities (Sagar *et al.*, 2008). For example "MC1R" (a gene from the HGP dictionary) almost always occurs with the entity "pigmentation" and "hair" and "red" (all from the HA dictionary). Upon manual curation, it is easily verified that the MC1R gene plays a role in red hair colour.

A clear example of false positive identification is with regard to the Aromatase (CYP19A1), Pubmed ID 11549641. In a text mining mediated search for putative height genes, the Aromatase alias name CYP19 is included in the results as an independent and separate entity.

Text-mining is an automated, fast and efficient means of extracting specific information from a large number of published articles in a less laborious manner and in a short space of time. In this thesis, the use of text-mining was utilized in acquiring a set of forensic relevant genes associated with physical traits, namely pigmentation and height.

4.3 Database Schema

The MySQL Database Management System (DBMS) facilitates the storage and retrieval of data as well as well as multiple access, offer security, data integrity and support the support for various applications.

4.4 Gathering database statistics

Since different population groups are known to have different genetic determinants of pigmentation (Liu *et al.*, 2009), a similar approach was employed for height to characterize SNPs as genetic determinants for height in different populations. Pigmentation genes such as MC1R are characterized by alleles that are prominent in certain populations (Branicki *et al*., 2007). This justifies the inclusion of the SNP allele frequencies in the Forensic SNP Phenotype Database. These polymorphisms can therefore be used to differentiate small differences both in a population and among different populations. Raw SQL statements were ran to generate database statistics, plot graphs and ascertain the relevance of the data. However, the end-users will only be able to interact with the database through a web user interface.

4.4.1 Analysing SNP Allele Frequency Data

The Forensic SNP Phenotype Database contains information regarding the SNP allele frequencies between various populations. The patterns observed in allele frequencies could infer phenotypic differences such as hair and skin colour among the various population groups. This could be used to deduce the population specificity of SNPs for forensic identification.

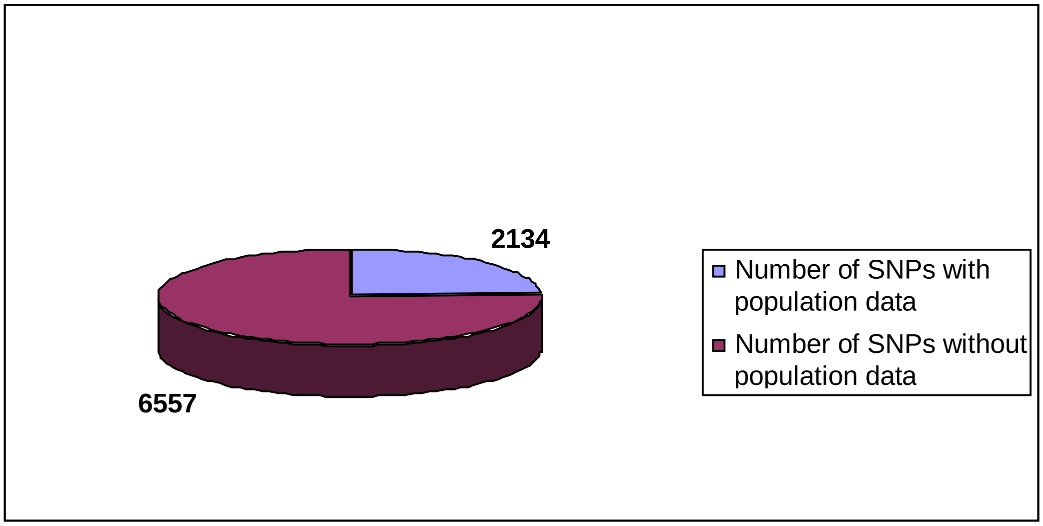


Figure 14: A pie chart illustrating the occurrence of population data for SNPs. A total of 65574 SNPs lack population data while 21341 SNPs have population data.

Out of a total of 86915 SNPs, 65574 SNPs constituting 75.4% did not have population information. This implies that 21341 SNPs, which translates to 24.6%, have not been validated for population data in dbSNP (figure 14). This result was obtained by running the SQL query statement in Item 15. This presents the opportunity for forensic geneticists to determine the SNP genotypes and allele frequencies with respect to certain population groups of interest. In a different study (van Daal, 2008), notable physical differences, such as hair colour, have been observed between different population groups where allele frequencies tend to differ significantly. In the ASIP gene, the frequency of the 'G' allele (g8818G) is higher than that of 'A' (g8818A) among the Aborigines as compared to the Australians of Caucasian origin. Genotype studies have also shown that there is a decreased level in the amount of ASIP mRNA suggesting that the SNP, which occurs in the 3' UTR, affects mRNA stability leading to less amount of pheomelanin, hence a resultant dark skin and hair colour. The ASIP gene also antagonizes MC1R, a key regulator of melanin production.

```
SELECT count(a.*) FROM SNP a LEFT JOIN tmp_rs_id b
ON b.rs_id=a.rs_id
WHERE b.rs_id IS NULL;
```

Item 15: SQL query statement for retrieving SNP population data count

4.4.2 Forensic SNP Phenotype Database Populations

There are a total of 22 unique populations contained in the Forensic SNP Phenotype Database, 11 of which are from Hapmap and 11 from dbSNP. Table 1 lists all the Hapmap populations and also provides their specific ethnic and geographic information. dbSNP does not have a classification for race and ethnic groups but allows filtering of SNP data using field population classes. These classes are based on geographical location (Table 2) and the genotype data contained is also available at the International Hapmap Project's site (NCBI, 2003).

Table 1: A list of populations as defined by the HapMap Project

Population Group	Description
ASW	African ancestry in Southwest USA
CEU	Utah residents with Northern and Western European ancestry from the CEPH collection
CHB	Han Chinese in Beijing, China
CHD	Chinese in Metropolitan Denver, Colorado
GIH	Gujarati Indians in Houston, Texas

JPT	Japanese in Tokyo, Japan
LWK	Luhya in Webuye, Kenya
MEX	Mexican ancestry in Los Angeles, California
MKK	Maasai in Kinyawa, Kenya
TSI	Toscans in Italy
YRI	Yoruba in Ibadan, Nigeria (West Africa)

Table 2: A list of populations as curated by dbSNP at the NCBI

Population Group	**Description**
CENTRAL ASIA	Samples from Russia and satellite Republics, Nations bordering Indian Ocean between East Asia and Persian Gulf regions.
CENTRAL/SOUTH AFRICA	Nations south of Equator, Madagascar & neighboring Island Nations.
CENTRAL/SOUTH AMERICA	Samples from Mainland Central and South America, Island Nations of western Atlantic, Gulf of Mexico and Eastern Pacific.
EAST ASIA	Samples from Eastern and South Eastern Mainland Asia, Northern Pacific Island Nations.
EUROPE	Samples from Europe north and west of Caucasus Mountains, Scandinavia, Atlantic Islands.
MULTI-NATIONAL	Samples that were designed to maximize measures of heterogeneity or sample human diversity in a global fashion. Examples OEFNER|GLOBAL and CEPH repository.
NORTH AMERICA	All samples north of Tropic of Cancer. This would include defined samples of U.S. Caucasians, African Americans and Hispanics and NCBI|NIHPDR.
NORTH/EAST AFRICA & MIDDLE	Samples collected from North Africa (including Sahara desert), East Africa (south to Equator), Levant,

Population Group	Description
EAST	and Persian Gulf.
PACIFIC	Samples from Australia, New Zealand, Central and Southern Pacific Islands, Southeast Asian Peninsular/Island Nations.
UNKNOWN	Samples with unknown geographic province that are not global in nature.
WEST AFRICA	Sub-Saharan Nations bordering Atlantic north of Congo River, and Central/Southern Atlantic Island Nations.

4.4.3 The Count of SNPs for Each of the Populations

The Forensic SNP Phenotype Database can be queried for the number of SNPs found in each population by looking at any given gene. By examining two population groups which have notable physical differences such as height or skin colour, forensic geneticists may be able to identify possible SNP causing those differences. These SNPs could later be confirmed in a laboratory experiment. This could not only enable characterization of SNPs by populations but could also be used to identify the ethnicity of a criminal, through DNA left at a crime scene.

```
SELECT COUNT(*),a.gen_pop FROM (SELECT DISTINCT
rs_id,gen_pop FROM pop_and_freq_dbsnp a)a GROUP BY a.gen_pop;
```

Item 16: An SQL statement for determining the number of SNPs per dbSNP population

In order to determine the population with the highest number of SNPs, the SQL statement in Item 16 was executed. The East Asia population, which is composed of individuals from East and South Eastern Mainland Asia as well as the Northern Pacific Island Nations, was found to have the highest number of SNPs at 10,021. The least number of SNPs was found in the Pacific population with a total of 4 SNPs (figure 15). This low occurrence of SNPs could be due to a measurement bias. When a similar query (item 17) was run for the HapMap populations, the trend that emerged is that ASW, individuals with African ancestry, had the most number of SNPs at 7,375 while JPT, Japanese individuals, had the least at 5,908 SNPs (figure 16). The observed trend is in keeping with studies (Tishkoff and Williams 2002), which indicate that when nuclear DNA markers are analysed, Africa is the worlds most genetically diverse region. The Recent African Origin (RAO) model predicts that all genetic lineages originate from a common African ancestor and that non-African population groups have a subset of the genetic variation that is found in all modern African populations.

```
SELECT count(*) FROM (SELECT DISTINCT a.rs_id FROM
pop_and_freq_hapmap a)a;
```

Item 17: An SQL statement for determining the number of SNPs per HapMap population

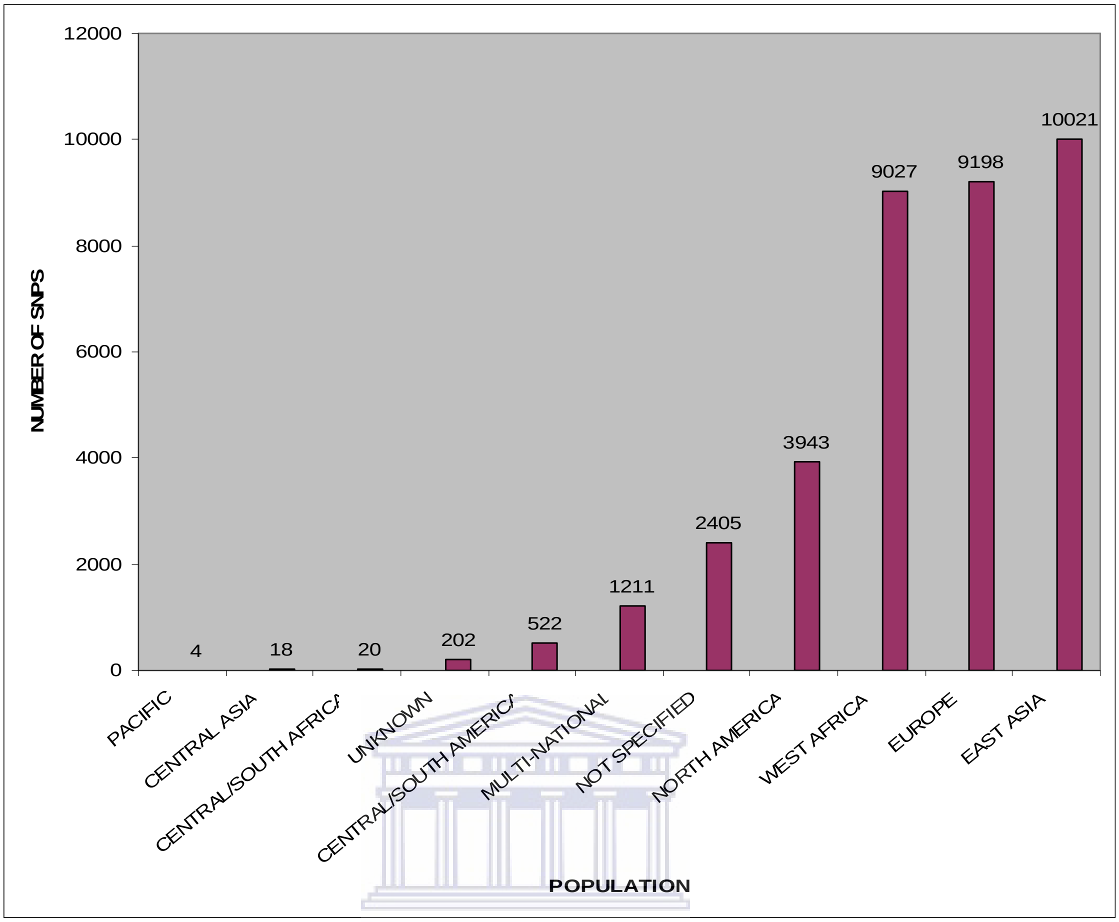


Figure 15: A bar graph showing the number of SNPs in each dbSNP population with the number of SNPs represented on the y-axis and the each population group on the x-axis.

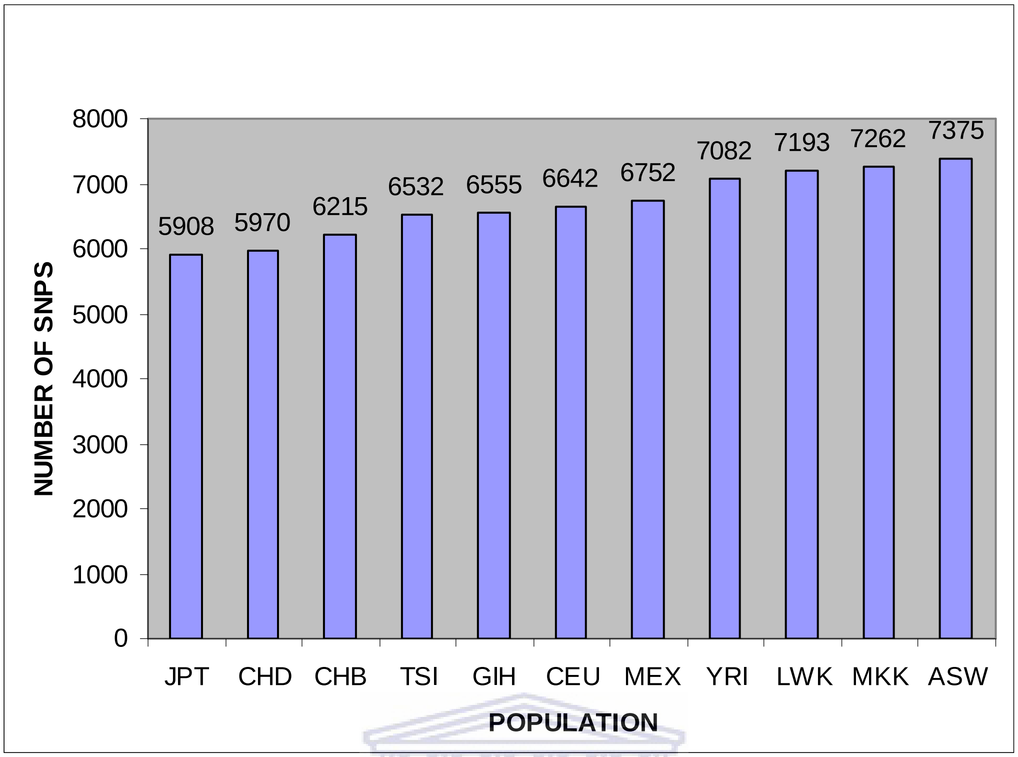


Figure 16: A bar graph showing the number of SNPs in each HapMap population with the number of SNPs represented on the y-axis and the each population group on the x-axis.

4.4.4 The Count of SNPs for Each Gene per Population

```
SELECT COUNT(*)AS
count_of_snps,a.pop,a.SNP_source,b.gene_symbol,a.entrez_id,b.gene_name
FROM gene_SNP_pop a
JOIN gene_description b ON b.entrez_id=a.entrez_id GROUP BY
a.entrez_id,a.SNP_source,a.pop,b.gene_symbol,b.gene_name ORDER BY a.pop;
```

Item 18: An SQL statement showing the count of each gene for each population

Table 3, which is a sample file, shows how individual genes contribute to a set of SNPs for a given physical trait in a specific population. This might be of use to a forensic geneticist or scientist who may want to investigate the SNPs in a group of genes found in a specific population. This could also be done to determine the SNP allele frequency for evaluation. The data was obtained by running the SQL statement in item 18.

Table 3: Enumeration of SNPs per gene for each population

SNP COUNT	POPULATION GROUP	POPULATION DATA SOURCE	GENE SYMBOL
12	EUROPE	dbsnp	MC1R
76	EUROPE	dbsnp	HERC2
346	EUROPE	dbsnp	OCA2
22	EUROPE	dbsnp	GDF5
83	EUROPE	dbsnp	HMGA2
28	CEU	hapmap	MC1R
58	CEU	hapmap	HERC2
189	CEU	hapmap	OCA2
13	CEU	hapmap	GDF5
75	CEU	hapmap	HMGA2

By looking at the number of SNPs found in each gene, their distribution was analysed across all the 108 genes (Appendix 6). For example, the MC1R, which is a well studied pigmentation gene involved in forensic physical trait predictions, has a total of seventy five 75 SNPs. The distribution of these 75 SNPs among 7 populations namely ASW, YRI, LWK, CEU, MKK, MEX and TSI was found to be 30, 29, 28, 28, 27, 27 and 27 respectively. From the population distribution pattern of SNPs in this gene alone (table 4), it can be noted clearly that no single population has a complete set all the 75 variants. The population with the highest number is ASW with a total of 30 SNPs. This means that some SNPs could be population specific i.e may be present in some population groups and not in others. Such SNPs, provided they are play a functional or regulatory role, could be population specific. Such SNPs could explain, in terms of their low mutation rate and becoming fixed in certain populations, why certain populations have certain traits that are not present in other populations. Forensic geneticists can now utilize this data from this database to elucidate the SNPs occurring in genes at the population level for population specific traits.

Table 4: The number of SNPs for each population for the MC1R gene

POPULATION	SNP COUNT
ASW	30
YRI	29
CEU	28
LWK	28
MEX	27
MKK	27
TSI	27
GIH	25
JPT	23
CHB	21
NOT SPECIFIED	20
CHD	19

EAST ASIA	19
EUR	12
EUROPE	12
MULTINATIONAL	10
CENTRAL SOUTH AMERICA	9
NORTH AMERICA	8
WEST AFRICAN	8
CENTRAL/SOUTH AFRICA	1

Table 5: A table highlighting the distribution of height and pigmentation associated SNPs across all the population groups found in dbsnp and hapmap

COUNT OF SNPS	POPULATION	SNP SOURCE	TRAIT
5462	ASW	hapmap	height
2037	ASW	hapmap	pigmentation
16	CENTRAL ASIA	dbsnp	height
2	CENTRAL ASIA	dbsnp	pigmentation
17	CENTRAL/SOUTH AFRICA	dbsnp	height
3	CENTRAL/SOUTH AFRICA	dbsnp	pigmentation
333	CENTRAL/SOUTH AMERICA	dbsnp	height
189	CENTRAL/SOUTH AMERICA	dbsnp	pigmentation
4952	CEU	hapmap	height
1801	CEU	hapmap	pigmentation
4654	CHB	hapmap	height
1646	CHB	hapmap	pigmentation
4452	CHD	hapmap	height
1594	CHD	hapmap	pigmentation
7345	EAST ASIA	dbsnp	height
2693	EAST ASIA	dbsnp	pigmentation
6784	EUROPE	dbsnp	height
2431	EUROPE	dbsnp	pigmentation
4835	GIH	hapmap	height
1803	GIH	hapmap	pigmentation
4394	JPT	hapmap	height
1591	JPT	hapmap	pigmentation
5330	LWK	hapmap	height
1981	LWK	hapmap	pigmentation

5013	MEX	hapmap	height
1825	MEX	hapmap	pigmentation
5356	MKK	hapmap	height
2021	MKK	hapmap	pigmentation
840	MULTI-NATIONAL	dbsnp	height
354	MULTI-NATIONAL	dbsnp	pigmentation
2899	NORTH AMERICA	dbsnp	height
1052	NORTH AMERICA	dbsnp	pigmentation
1422	NOT SPECIFIED	dbsnp	height
1044	NOT SPECIFIED	dbsnp	pigmentation
4	PACIFIC	dbsnp	pigmentation
4857	TSI	hapmap	height
1776	TSI	hapmap	pigmentation
184	UNKNOWN	dbsnp	height
18	UNKNOWN	dbsnp	pigmentation
6693	WEST AFRICA	dbsnp	height
2345	WEST AFRICA	dbsnp	pigmentation
5239	YRI	hapmap	height
1956	YRI	hapmap	pigmentation

Table 5 shows the numbers of SNPs both for pigmentation and height that are found in each population.

4.4.5 Functional Classes of SNPs

With 65,475 SNPs (Appendix 7), the most common functional SNP class occurs in the intronic gene regions. A total of 14,020 SNPs did not have any functional class assigned to them. The second highest class was the “utr-3”, three prime untranslated region of a gene, which is a regulatory region. The SNPs in this region might affect transcription in regulatory sequences thus affecting gene expression in human pigmentation (Duffy 2007).

From this it could be implied that SNPs associated with the specified physical traits (height and pigmentation) are most likely to occur in coding regions but also, though also to a lesser extent, in regulatory region. SNPs occurring in promoter, coding and intronic regions are associated with phenotype variation. Specifically, SNPs occurring within the promoter regions of genes have been documented to enhance traits such as blue and brown eye color variation, as well as variation in hair colour. The functional classes of SNPs being discussed are classified as per dbSNP which computes a functional context for SNPs by basing the class on the relationship between a variation and any local gene features.

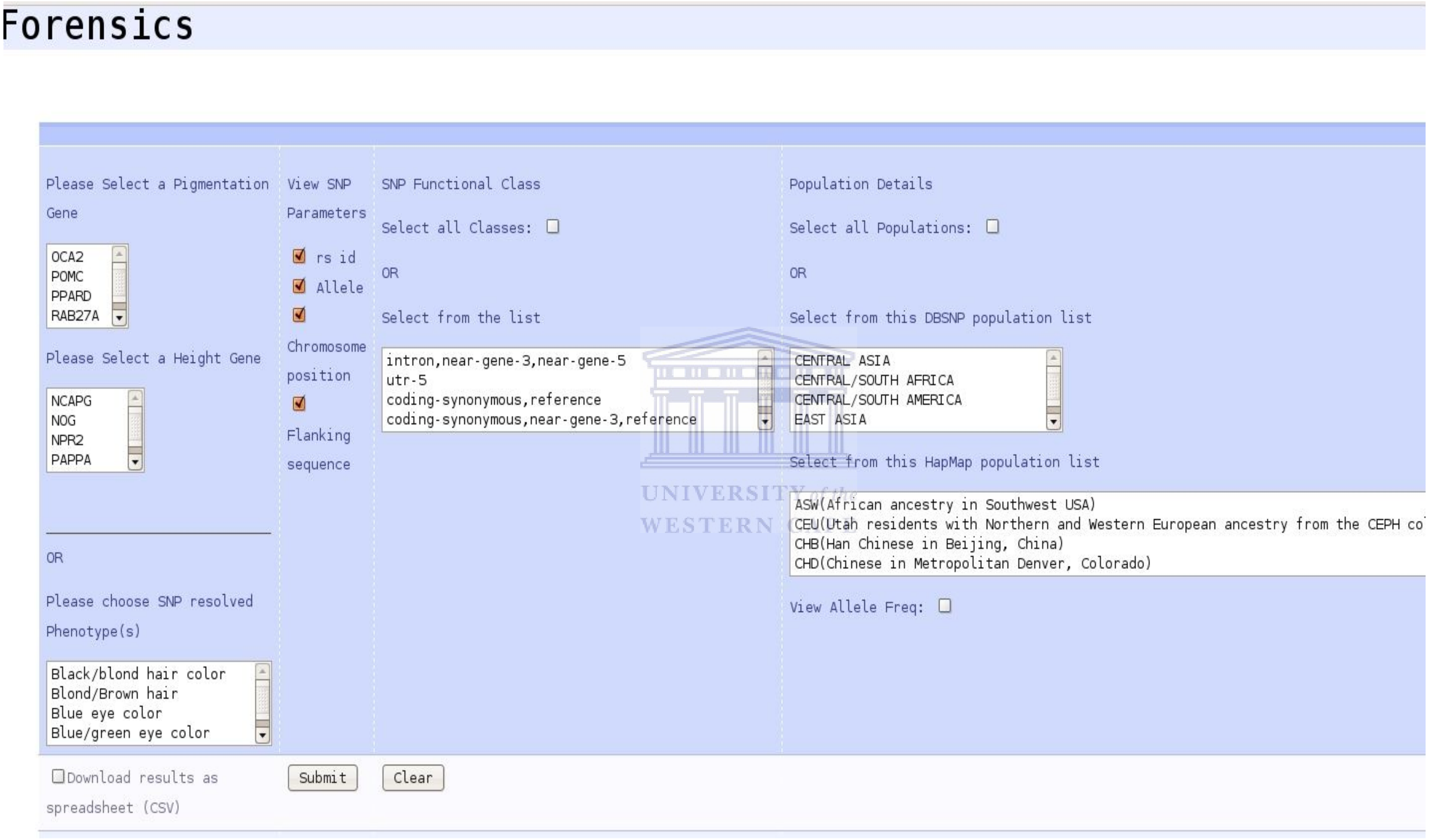
In the Forensic SNP Phenotype Database, 71,232 SNPs are bi-allelic, such as A/T, and the remaining 15,681 are of other types such as insertions or deletions such as -/T. This shows that there are insertions and deletions (indels) that may possibly play a minor role in certain complex trait phenotypes. Previous studies have been able to demonstrate that insertions within the MC1R gene potentially play a role in red hair colour (Branicki *et al*., 2007).

4.5 The Web User Interface

A web user interface has been developed aimed at granting users access to the Forensic SNP Phenotype Database. This SNP resource can be accessed at http://forensic.sanbi.ac.za/alecia_forensics/Index.html. Figure 17 illustrates the Forensic SNP Phenotype Database web homepage.

Database users can navigate through the various functions of the forensic SNP repository through a web browser which acts as a client in this 3-tier model. It enables the display of HTML resources, issues HTML requests and processes the responses issues by both the user and also those received from the MySQL back-end. This interaction is mediated by the Tomcat server through implementation of standard protocols. One of the main advantages of a 3-tier architecture with reference to the client is that the client is independent of the operating system and does not need extra software to function optimally. Although there are minimal differences between the performances of different browsers, more or less they operate similarly.

The middle tier for the Forensic SNP Phenotype Database executes the application logic which enhances the smooth navigation of the database. This involves processing the inputs it receives from the browser as well as interacting with the database. It is made up of the Tomcat web server, JSP and the Java scripting language engine. Since the entire request requires the output after running the program and interacting with the databases, the web server calls the scripting engine to perform those activities. Due to the fact that the program that executes the application's logic has been embedded into static HTML users can smoothly interact with the SQL back-end straight through their browser.



Forensics

Please Select a Pigmentation Gene

OCA2
POMC
PPARD
RAB27A

Please Select a Height Gene

NCAPG
NOG
NPR2
PAPPA

OR

Please choose SNP resolved Phenotype(s)

Black/blond hair color
Blond/Brown hair
Blue eye color
Blue/green eye color

View SNP Parameters

☑ rs id
☑ Allele
☑ Chromosome position
☑ Flanking sequence

SNP Functional Class

Select all Classes: □

OR

Select from the list

intron,near-gene-3,near-gene-5
utr-5
coding-synonymous,reference
coding-synonymous,near-gene-3,reference

Population Details

Select all Populations: □

OR

Select from this DBSNP population list

CENTRAL ASIA
CENTRAL/SOUTH AFRICA
CENTRAL/SOUTH AMERICA
EAST ASIA

Select from this HapMap population list

ASW(African ancestry in Southwest USA)
CEU(Utah residents with Northern and Western European ancestry from the CEPH co
CHB(Han Chinese in Beijing, China)
CHD(Chinese in Metropolitan Denver, Colorado)

View Allele Freq: □

□ Download results as spreadsheet (CSV)

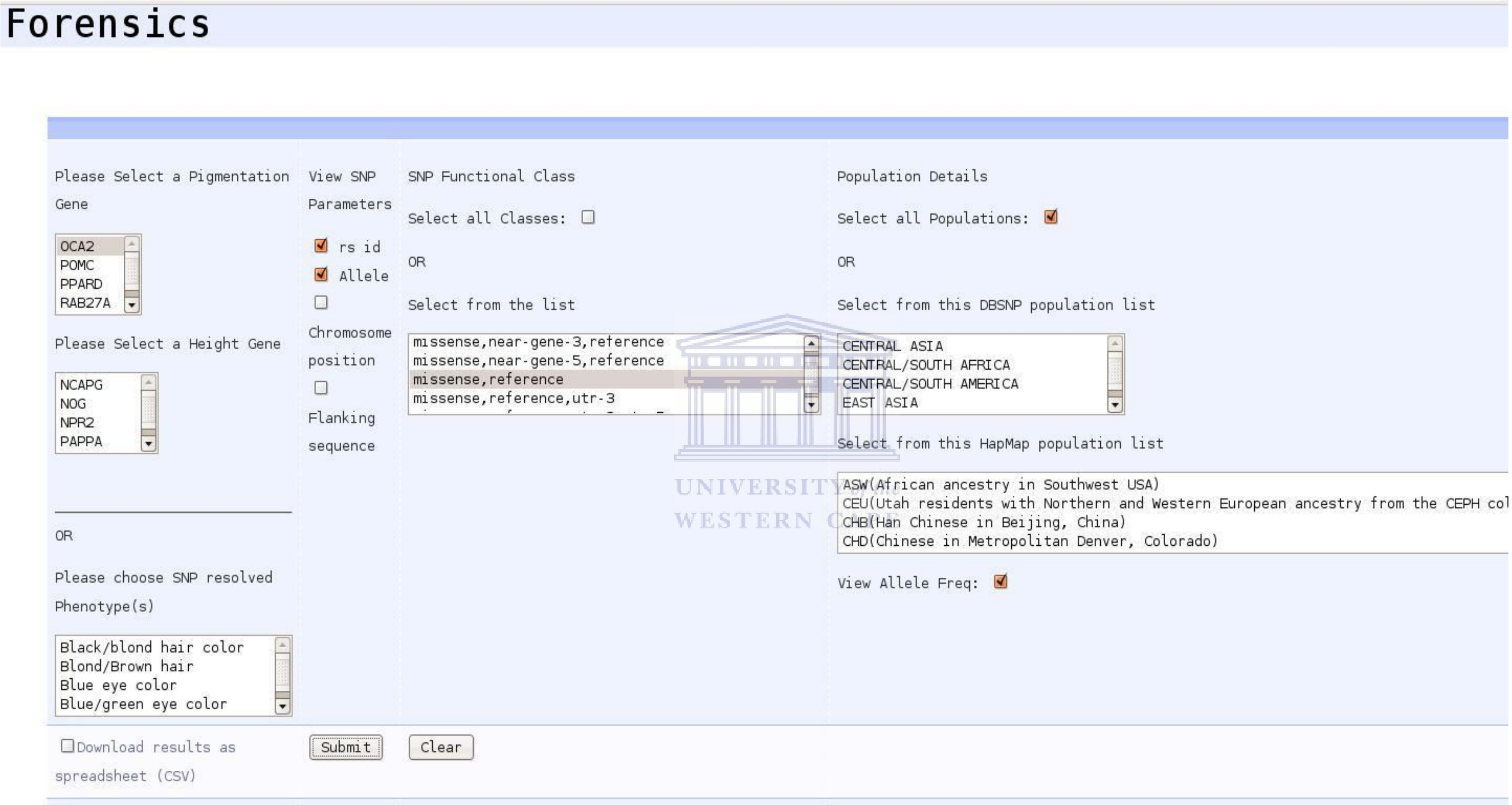
Figure 17: Screenshot of the Forensic SNP Phenotype Database homepage.

4.5.1 Navigating through the web interface

Extraction of information from the Forensic SNP Phenotype Database has been implemented through a series of filters. This enables users to interact with the database through different query parameters (figure 18). The first step in extraction of SNPs involves the selection of either a gene, that has documented forensic relevance, or a physical trait that has an underlying forensic importance. Gene selection is mediated through list-box selection options. There are two categories of genes namely the height genes and the pigmentation genes. Height and pigmentation traits have been selected since they are visually recordable and largely aid in categorizing population groups. User can also query the Forensic SNP Phenotype Database by directly selecting a phenotype such as 'blue eye color', from a list-box. This will display all the SNPs that are associated with the selected gene or phenotype.

To aid in answering specific questions, researchers can apply other functionalities of the web interface to retrieve specific datasets. For instance, the results could be refined through a series of filters which enable the user to display only specific information. Once the filters are placed, a number of attributes can be selected for the resulting SNP data. For example in “View SNP Parameters” users may select SNP characteristics such as the alleles for that SNP, the rs id and the chromosome position of an SNP in the human reference genome. Choosing the flanking sequence option for these SNPs could aid researchers on primer design in experimental SNP typing assays and high throughput SNP genotyping.

Since one of the main objectives of the Forensic SNP Phenotype Database is the characterization of SNPs with regard to different geographic population groups, various population groups have been availed on the web user interface. In this manner, researchers can choose the population groups for which they want to characterize the SNPs. These populations groups are based on those available at the HapMap project website and the dbSNP repository. The allele frequency data originates from two databases; dbSNP and HapMap. In the results output, the origin of the allele frequency data has been labeled, so users can discern between the two types of data. This is because SNPs from HapMap are 'true' SNPs while those from dbSNP have in-dels and other polymorphisms as well. Filtering of the results based on the user input parameters is facilitated by the logic tier in the 3-tier model. It also enables query formation and results retrieval from the database. Figure 19 illustrates how a user can query for all the missense and reference SNPs occurring in the OCA2 gene. The query targets all dbSNP and HapMap population groups. The user also wants to display only the rs ID, the allele and the allele frequency information for those SNPs.



Forensics

Please Select a Pigmentation Gene

OCA2
POMC
PPARD
RAB27A

Please Select a Height Gene

NCAPG
NOG
NPR2
PAPPA

OR

Please choose SNP resolved Phenotype(s)

Black/blond hair color
Blond/Brown hair
Blue eye color
Blue/green eye color

View SNP Parameters

☑ rs id
☑ Allele
□ Chromosome position
□ Flanking sequence

SNP Functional Class

Select all Classes: □

OR

Select from the list

missense,near-gene-3,reference
missense,near-gene-5,reference
missense,reference
missense,reference,utr-3

Population Details

Select all Populations: ☑

OR

Select from this DBSNP population list

CENTRAL ASIA
CENTRAL/SOUTH AFRICA
CENTRAL/SOUTH AMERICA
EAST ASIA

Select from this HapMap population list

ASW(African ancestry in Southwest USA)
CEU(Utah residents with Northern and Western European ancestry from the CEPH col
CHB(Han Chinese in Beijing, China)
CHD(Chinese in Metropolitan Denver, Colorado)

View Allele Freq: ☑

□ Download results as spreadsheet (CSV)

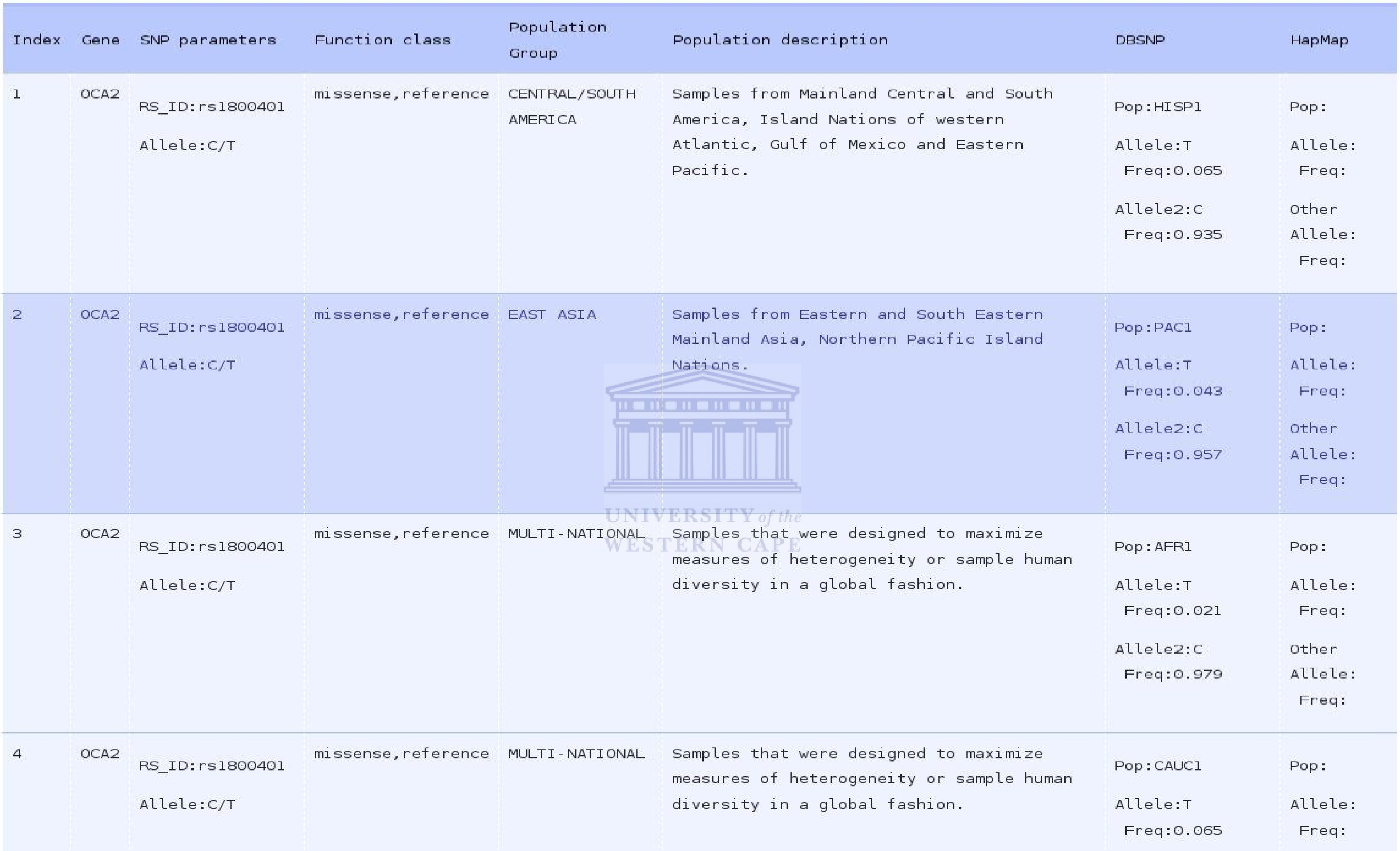
Figure 18: A Screenshot illustrating user query page of the Forensic SNP Phenotype Database. The actual query is available at (*http://forensic.sanbi.ac.za/alecia_forensics/Index.html*).

4.6 Results display

The query results can be viewed in two output formats, HyperText Markup Language (HTML) and Comma Separated Values (CSV) file. The HTML page result-display (figure 19) allows quick viewing. It is also the preferable mode if the resultant query returns only a few SNPs. The in-depth results can be downloaded in CSV format which allows for parsing using any scripting language such as perl, or for further statistical or computational analysis or applications.

In designing the user interface results display, two other types of filtering checks were put into place (in addition to the gene and phenotype filters), which influence how the results can be viewed or how the user may wish to display the results. The "SNP Functional Class" criteria allows for results to be filtered by the SNP functional classes, thereby refining the output. The "Population Details" filter also caters for SNPs to be refined by populations of interest. If the user selects the "View Allele Frequency" option, population details will be displayed together with the corresponding allele frequency.

With regard to SNP functional classes, it is important to note that not all SNPs in the Forensic SNP phenotype database have been assigned functional classes. Out of the 86,915 SNPs, 14,020 do not have functional classes. This is due to the unavailability of this data from dbSNP. Figure 19 illustrates the results whereby a user can query for all the missense and reference SNPs occurring in the OCA2 gene. The query had targeted all dbSNP and HapMap population groups. The user also wanted to display only the rs ID, the allele and the allele frequency information for those SNPs.



Index	Gene	SNP parameters	Function class	Population Group	Population description	DBSNP	HapMap
1	OCA2	RS_ID:rs1800401 Allele:C/T	missense,reference	CENTRAL/SOUTH AMERICA	Samples from Mainland Central and South America, Island Nations of western Atlantic, Gulf of Mexico and Eastern Pacific.	Pop:HISP1 Allele:T Freq:0.065 Allele2:C Freq:0.935	Pop: Allele: Freq: Other Allele: Freq:
2	OCA2	RS_ID:rs1800401 Allele:C/T	missense,reference	EAST ASIA	Samples from Eastern and South Eastern Mainland Asia, Northern Pacific Island Nations.	Pop:PAC1 Allele:T Freq:0.043 Allele2:C Freq:0.957	Pop: Allele: Freq: Other Allele: Freq:
3	OCA2	RS_ID:rs1800401 Allele:C/T	missense,reference	MULTI-NATIONAL	Samples that were designed to maximize measures of heterogeneity or sample human diversity in a global fashion.	Pop:AFR1 Allele:T Freq:0.021 Allele2:C Freq:0.979	Pop: Allele: Freq: Other Allele: Freq:
4	OCA2	RS_ID:rs1800401 Allele:C/T	missense,reference	MULTI-NATIONAL	Samples that were designed to maximize measures of heterogeneity or sample human diversity in a global fashion.	Pop:CAUC1 Allele:T Freq:0.065	Pop: Allele: Freq:

Figure 19: Screenshot illustrating the results page for the given query in figure 18.

4.7 Applied Examples

The Forensic SNP Phenotype Database aims to understand or discover the genetic determinants of human physical variation within the forensic context. A worked example was carried out to demonstrate the potential forensic-relevance of the SNP information contained in the Forensic SNP Phenotype Database. This has been done through *in silico* SNP analysis aimed at establishing possible relationships between SNP occurrence and phenotype. Both the possibility of SNPs having an effect on the phenotype through alternation of transcription binding sites and also through amino acid alterations has been tested.

4.7.1 SNPs and Transcription Factor Binding Sites (TFBSs)

Analysis and identification of SNPs in promoter regions aims at investigating whether their presence affects phenotype. SNPs occurring in the gene promoter region potentially affects the phenotype by altering gene expression level. Through this analysis, the significance and relevance of the data contained in the Forensic SNP Phenotype Database was established.

Firstly the promoter sequences (3200bp) for the MC1R gene was extracted and run through the MATCH™ software so as to predict TFBSs which might occur in this region (Kel *et al*., 2003). After a scan with MATCH™ the SNPs tagged to this promoter region were extracted from NCBI dbSNP 129. The TFBS and SNPs were then mapped to the promoter sequence to determine if SNPs do occur in any of the binding sites. This was done visually by comparing absolute chromosome positions. Table 6 shows SNPs occuring in the binding sites of these transcription factors.

Table 6: Table showing SNPs disrupting TFBSs

SNP	Chromosome Position	Transcription Factor
rs3212362	88512845	TBX5
rs3212360	88512718	ZF5 and Muscle initiator sequence

Once these SNPs were determined, the promoter sequence input file was altered to incorporate them. The mutated promoter sequence was then processed in MATCH™ for the second time to determine if these SNPs introduce any new binding sites. One new site was found for the GR transcription factor, occurring just after the binding site for TBX5.

It was also found that the SNP may increase the probability of TBX5 binding to the binding site as the matrix probability was noted to be slightly higher from 0.975 to 0.99. The consensus sequence also changes from a G:10 to A:25, showing a higher affinity for 'A' in the binding matrix than 'G'. With the other two transcription factors, there appears to be not much change to the matrix probability and the consensus sequence scores with the occurrence of SNPs in the TFBSs.

From this analysis, it may be concluded that the SNP increases the probability of binding of the transcription factor TBX5 to the TFBS. This is based on the higher matrix match probability as well as the analysis of the binding matrix and the consensus binding site. As a result of this one SNP, the TBX5 may now bind more strongly to the binding site and this may affect the regulation of gene expression. Hence, less or more gene product may occur which may explain possible variations in skin or hair phenotypes.

These SNPs disrupting TFBSs may be candidate markers in the field of forensics for predicting physical traits and these could be further verified experimentally through SNP genotyping methods. Further information regarding these SNPs and the populations where they occur has been provided in the Forensic SNP Phenotype Database. This makes it easy to assess potential candidate forensic markers in physical trait prediction.

4.7.2 SNPs and Amino Acids substitution

A total of 29 SNPs were extracted for the MC1R gene. Out of these, only 5 were found to be non-synonymous and missense. These SNPs were again selected since they have an effect on the translation of a protein and might be able to explain the phenotypic difference among individuals. For each of those 5 SNPs, the amino acid residue change was noted. In 4 out of 5 cases, the amino acid change was from a non-polar to another non-polar amino acid. Only one SNP had a basic to a non-polar change. It is likely that, in general, a basic to non-polar amino acid change would be more likely to have a potential functional effect while a non-polar to non-polar would perhaps be less likely. This could be true assuming that other variables, such as size of the side chain, do not play a role.

Worked examples were illustrated to identify possible functional variants, both amino acid and non-coding regulatory regions, which may alter the structure or activity of a gene or gene product and could be responsible for various observable physical traits.

The aim of the Forensic SNP Phenotype Database is to help forensic scientists identify potential SNPs that could be used to predict certain physical traits, using trace DNA material left at a crime scene. Instead of mining other SNP databases for relevant SNP markers, this database provides a comprehensive source of relevant pigmentation and height SNP data complete with their associated genes. Different population groups have different genetic determinants and since the Forensic SNP Phenotype Database has well characterized population data it may be possible for researchers to get population specific alleles for each specific trait (pigmentation and height).

This database can be used as a starting point for carrying out SNP genotyping so as to enable population specific markers to be laboratory confirmed. Many studies have tested SNPs on physical traits and validated them but no integrative tool exists for this SNP information. Pigmentation and height traits were selected since they are widely studied in forensics.

The Forensic SNP Phenotype Database can be used as a tool to analyse and assess correlations between genetic variants, the majority of which are SNPs, and the phenotypic differences in traits on a population level. This database also caters for the phenotypes which may be associated with the indels. Identifying SNPs across geographically and ethnically diverse human populations is important since the greater variation of SNP frequencies between populations the more informative those SNPs are and hence the higher their level of importance.

CONCLUSION

The Forensic SNP Phenotype Database has been created with the view to discover and understand the genetic determinants of human physical variation within the forensic context. This database presents SNP data which can be used as genetic markers for the prediction of certain physical traits for forensic identification. Application of SNPs to infer an individual's phenotype can be potentially instrumental in narrowing down a pool of criminal suspects in the preliminary stages of a criminal investigation. An SNP-based forensic identification tool can provide intelligence to forensic investigators when STR profiling has failed, does not match the profile in question or there are no available eye witnesses. As more literature becomes available in understanding genetic variants associated with complex traits, the Forensic SNP Phenotype Database can be extended to accommodate the SNPs associated with these traits. This would enhance its application as a starting point for forensic related SNP analysis, in phenotype prediction.

REFERENCES

Amigo, J., Phillips, C., Lareu, M., & Carracedo, A. 2008, "The SNPforID browser: an online tool for query and display of frequency data from the SNPforID project", *Int.J.Legal Med.*, vol. 122, no. 5, pp. 435-440.

Anno, S., Abe, T., & Yamamoto, T. 2008, "Interactions between SNP alleles at multiple loci contribute to skin color differences between caucasoid and mongoloid subjects", *Int.J.Biol Sci*, vol. 4, no. 2, pp. 81-86.

Barnes, M. R. 2006, "Navigating the HapMap", *Brief.Bioinform.*, vol. 7, no. 3, pp. 211-224.

Barsh, G. S. 2003, "What controls variation in human skin color?", *PLoS Biol*, vol. 1, no. 1, p. E27.

Bianchi, L. & Lio, P. 2007, "Forensic DNA and bioinformatics", *Brief.Bioinform.*, vol. 8, no. 2, pp. 117-128.

Birney, E., Andrews, T. D., Bevan, P., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cuff, J., Curwen, V., Cutts, T., Down, T., Eyras, E., Fernandez-Suarez, X. M., Gane, P., Gibbins, B., Gilbert, J., Hammond, M., Hotz, H. R., Iyer, V., Jekosch, K., Kahari, A., Kasprzyk, A., Keefe, D., Keenan, S., Lehvaslaiho, H., McVicker, G., Melsopp, C., Meidl, P., Mongin, E., Pettett, R., Potter, S., Proctor, G., Rae, M., Searle, S., Slater, G., Smedley, D., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Storey, R., Ureta-Vidal, A., Woodwark, K. C., Cameron, G., Durbin, R., Cox, A., Hubbard, T., & Clamp, M. 2004, "An overview of Ensembl", *Genome Res*, vol. 14, no. 5, pp. 925-928.

Branicki, W., Brudnik, U., Kupiec, T., Wolanska-Nowak, P., & Wojas-Pelc, A. 2007, "Determination of phenotype associated SNPs in the MC1R gene", *J.Forensic Sci*, vol. 52, no. 2, pp. 349-354.

Budowle, B. & van Daal, A. 2008, "Forensically relevant SNP classes", *Biotechniques,* vol. 44, no. 5, pp. 603-8, 610.

Butler, J. M.; Cobler, M. D.; Vallone, P. M. (2007), "STRs vs. SNPs: thoughts on the future of forensic DNA testing", Forensic Sci Med Pathol, vol. 3, pp. 200-205.

Cohen, A. M. & Hersh, W. R. 2005, "A survey of current work in biomedical text mining", *Brief.Bioinform.*, vol. 6, no. 1, pp. 57-71.

Collins, F. S., Morgan, M., & Patrinos, A. 2003, "The Human Genome Project: lessons from large-scale biology", *Science*, vol. 300, no. 5617, pp. 286-290.

Corte-Real, F. (2004), 'Forensic DNA databases.', *Forensic Sci Int* **146 Suppl**, S143--S144.

Delisle, M (2006). *Creating your MySQL Database: Practical Design Tips and Techniques*. Birmingham: Packt Publishing. 105.

Duffy, D. L., Montgomery, G. W., Chen, W., Zhao, Z. Z., Le, L., James, M. R., Hayward, N. K., Martin, N. G., & Sturm, R. A. 2007, "A three-single-nucleotide polymorphism haplotype in intron 1 of OCA2 explains most human eye-color variation", *Am.J.Hum.Genet.*, vol. 80, no. 2, pp. 241-252.

Fayyed, U, Piatetsky-Shapiro, G, and Smyth, P (1996) *From data Mining to Knowledge Discovery in Databases*. American Association for Artificial Intelligence

Forta, B. (2004) *Sams Teach Yourself SQL in 10 Minutes*, *3rd Edition*, Sams Publishing.256

Freese, E. 1959, "The difference between spontaneous and base-analogue induced mutations of phage T4", *Proc.Natl.Acad.Sci U.S.A*, vol. 45, no. 4, pp. 622-633.

Frudakis, T., Thomas, M., Gaskin, Z., Venkateswarlu, K., Chandra, K. S., Ginjupalli, S., Gunturi, S., Natrajan, S., Ponnuswamy, V. K., & Ponnuswamy, K. N. 2003, "Sequences associated with human iris pigmentation", *Genetics*, vol. 165, no. 4, pp. 2071-2083.

Graf, J., Voisey, J., Hughes, I., & van Daal, A. 2007, "Promoter polymorphisms in the MATP (SLC45A2) gene are associated with normal human skin color variation", *Hum.Mutat.*, vol. 28, no. 7, pp. 710-717.

Graham, E. A. 2008, "DNA reviews: predicting phenotype", *Forensic Sci Med.Pathol.*, vol. 4, no. 3, pp. 196-199.

Hearst, M, 2003, *What is Text Mining?*, University of California Berkeley, viewed 10 May 2009, <http://people.ischool.berkeley.edu/~hearst/text-mining.html>.

Hernandez, M.J 2003, *Database Design for Mere Mortals: A Hands-On Guide to Relational Database Design*, 2nd Edition, Addison-Wesley.672

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., Durbin, R., Eyras, E., Gilbert, J., Hammond, M., Huminiecki, L., Kasprzyk, A., Lehvaslaiho, H., Lijnzaad, P., Melsopp, C., Mongin, E., Pettett, R., Pocock, M., Potter, S., Rust, A., Schmidt, E., Searle, S., Slater, G., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Stupka, E., Ureta-Vidal, A., Vastrik, I., & Clamp, M. 2002, "The Ensembl genome database project", *Nucleic Acids Res*, vol. 30, no. 1, pp. 38-41.

Jobling, M. A. & Gill, P. 2004, "Encoded evidence: DNA in forensic analysis", *Nat.Rev.Genet.*, vol. 5, no. 10, pp. 739-751.

Karolchik, D., Hinrichs, A. S., & Kent, W. J. 2007, "The UCSC Genome Browser", *Curr.Protoc.Bioinformatics*, vol. Chapter 1, p. Unit.

Kaur, M., Radovanovic, A., Essack, M., Schaefer, U., Maqungo, M., Kibler, T., Schmeier, S., Christoffels, A., Narasimhan, K., Choolani, M., & Bajic, V. B. 2009, "Database for exploration of functional context of genes implicated in ovarian cancer", *Nucleic Acids Res*, vol. 37, no. Database issue, p. D820-D823.

Kel, A. E., Gossling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O. V., & Wingender, E. 2003, "MATCH: A tool for searching transcription factor binding sites in DNA sequences", *Nucleic Acids Res*, vol. 31, no. 13, pp. 3576-3579.

Kim, S. K. & Borevitz, J. 2006, "Mining the HapMap to dissect complex traits", *Genome Biol*, vol. 7, no. 3, p. 310.

Kofler, M (2005). *The Definitive MySQL 5*. 3rd ed. New York: Apress. 785

Lao, O., de Gruijter, J. M., van Duijn, K., Navarro, A., & Kayser, M. 2007, "Signatures of positive selection in genes associated with human skin pigmentation as revealed from analyses of single nucleotide polymorphisms", *Ann.Hum.Genet.*, vol. 71, no. Pt 3, pp. 354-369.

Lercher, M. J. & Hurst, L. D. 2002, "Human SNP variability and mutation rate are higher in regions of high recombination", *Trends Genet.*, vol. 18, no. 7, pp. 337-340.

Liu, F., van Duijn, K., Vingerling, J. R., Hofman, A., Uitterlinden, A. G., Janssens, A. C., & Kayser, M. 2009, "Eye color and the prediction of complex phenotypes from genotypes", *Curr.Biol*, vol. 19, no. 5, p. R192-R193.

Mackintosh, J. A. 2001, "The antimicrobial properties of melanocytes, melanosomes and melanin and the evolution of black skin", *J.Theor.Biol*, vol. 211, no. 2, pp. 101-113.

Miller, C. T., Beleza, S., Pollen, A. A., Schluter, D., Kittles, R. A., Shriver, M. D., & Kingsley, D. M. 2007, "cis-Regulatory changes in Kit ligand expression and parallel evolution of pigmentation in sticklebacks and humans", *Cell*, vol. 131, no. 6, pp. 1179-1189.

Morling, N. 2004, "Forensic genetics", *Lancet*, vol. 364 Suppl 1, p. s10-s11.

NCBI, 2003, The NCBI Handbook, National Centre for Biotechnology Information, Maryland, viewed 04 May 2009, <http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=handbook.table.ch5.ch5-t5>

Norton, H. L., Kittles, R. A., Parra, E., McKeigue, P., Mao, X., Cheng, K., Canfield, V. A., Bradley, D. G., McEvoy, B., & Shriver, M. D. 2007, "Genetic evidence for the convergent evolution of light skin in Europeans and East Asians", *Mol.Biol Evol.*, vol. 24, no. 3, pp. 710-722.

Parra, E. J. 2007, "Human pigmentation variation: evolution, genetic basis, and implications for public health", *Am.J.Phys.Anthropol.*, vol. Suppl 45, pp. 85-105.

Phillips, M. L. 2008, "Crime scene genetics: Transforming forensic science through molecular technologies", *Bioscience*, vol. 58, no. 6, pp. 484-489.

Povey, S., Lovering, R., Bruford, E., Wright, M., Lush, M., & Wain, H. 2001, "The HUGO Gene Nomenclature Committee (HGNC)", *Hum.Genet.*, vol. 109, no. 6, pp. 678-680.

Pulker, H., Lareu, M. V., Phillips, C., & Carracedo, A. 2007, "Finding genes that underlie physical traits of forensic interest using genetic tools", *Forensic Sci Int.Genet.*, vol. 1, no. 2, pp. 100-104.

Sagar, S., Kaur, M., Dawe, A., Seshadri, S. V., Christoffels, A., Schaefer, U., Radovanovic, A., & Bajic, V. B. 2008, "DDESC: Dragon database for exploration of sodium channels in human", *BMC Genomics*, vol. 9, p. 622.

Salisbury, B. A., Pungliya, M., Choi, J. Y., Jiang, R., Sun, X. J., & Stephens, J. C. 2003, "SNP and haplotype variation in the human genome", *Mutat.Res*, vol. 526, no. 1-2, pp. 53-61.

Sherry, S. T.; Ward, M. & Sirotkin, K. (1999), 'dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation.', *Genome Res* **9**(8), 677--679.

Sherry, S. T.; Ward, M. H.; Kholodov, M.; Baker, J.; Phan, L.; Smigielski, E. M. & Sirotkin, K. (2001), 'dbSNP: the NCBI database of genetic variation.', *Nucleic Acids Res* **29**(1), 308--311.

Sturm, R. A., Teasdale, R. D., & Box, N. F. 2001, "Human pigmentation genes: identification, structure and consequences of polymorphic variation", *Gene*, vol. 277, no. 1-2, pp. 49-62.

Sturm, R. A. & Frudakis, T. N. 2004, "Eye colour: portals into pigmentation genes and ancestry", *Trends Genet.*, vol. 20, no. 8, pp. 327-332.

Thorisson, G. A., Smith, A. V., Krishnan, L., & Stein, L. D. 2005, "The International HapMap Project Web site", *Genome Res*, vol. 15, no. 11, pp. 1592-1593.

Tishkoff, S. A. & Williams, S. M. 2002, "Genetic analysis of African populations:

human evolution and complex disease", *Nat.Rev.Genet.*, vol. 3, no. 8, pp. 611-621.

Walsh, S. J. 2004, "Recent advances in forensic genetics", *Expert Rev.Mol.Diagn.*, vol. 4, no. 1, pp. 31-40.

Wang, D. G., Fan, J. B., Siao, C. J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., Kruglyak, L., Stein, L., Hsie, L., Topaloglou, T., Hubbell, E., Robinson, E., Mittmann, M., Morris, M. S., Shen, N., Kilburn, D., Rioux, J., Nusbaum, C., Rozen, S., Hudson, T. J., Lipshutz, R., Chee, M., & Lander, E. S. 1998, "Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome", *Science*, vol. 280, no. 5366, pp. 1077-1082.

Wheeler, D. L., Church, D. M., Lash, A. E., Leipe, D. D., Madden, T. L., Pontius, J. U., Schuler, G. D., Schriml, L. M., Tatusova, T. A., Wagner, L., & Rapp, B. A. 2001, "Database resources of the National Center for Biotechnology Information", *Nucleic Acids Res*, vol. 29, no. 1, pp. 11-16.

APPENDICES

Appendix 1: Nonredundant Gene List

Official Gene Symbol	Official Gene Name
ACAN	aggrecan
ADAMTSL3	ADAMTS-like 3
AHSG	alpha-2-HS-glycoprotein
ANAPC13	anaphase promoting complex subunit 13
ANKFN1	ankyrin-repeat and fibronectin type III domain containing 1
AP3B1	adaptor-related protein complex 3, beta 1 subunit
AP3D1	adaptor-related protein complex 3, delta 1 subunit
ASIP	agouti signaling protein, nonagouti homolog (mouse)
ATXN3	ataxin 3
BMP2	bone morphogenetic protein 2
BMP6	bone morphogenetic protein 6
C6orf106	chromosome 6 open reading frame 106
C6orf173	chromosome 6 open reading frame 173
CABLES1	Cdk5 and Abl enzyme substrate 1
CDK6	cyclin-dependent kinase 6
CHCHD7	coiled-coil-helix-coiled-coil-helix domain containing 7
COL11A2	collagen, type XI, alpha 2
COL1A2	collagen, type I, alpha 2
CREB1	cAMP responsive element binding protein 1
CYP19A1	cytochrome P450, family 19, subfamily A, polypeptide 1
CYP1A2	cytochrome P450, family 1, subfamily A, polypeptide 2
CYP1B1	cytochrome P450, family 1, subfamily B, polypeptide 1
CYP2C8	cytochrome P450, family 2, subfamily C, polypeptide 8
CYP2C9	cytochrome P450, family 2, subfamily C, polypeptide 9
DCT	dopachrome tautomerase (dopachrome delta-isomerase, tyrosine-related protein 2)
DLEU7	deleted in lymphocytic leukemia, 7
DNM3	dynamin 3
DOT1L	DOT1-like, histone H3 methyltransferase (S. cerevisiae)
DRD2	dopamine receptor D2
DYM	dymeclin
EFEMP1	EGF-containing fibulin-like extracellular matrix protein 1
EGFR	epidermal growth factor receptor (erythroblastic leukemia viral (v-erb-b) oncogene homolog, avian)

ESR1	estrogen receptor 1
F2RL1	coagulation factor II (thrombin) receptor-like 1
FBLN5	fibulin 5
FBN1	fibrillin 1
FGF2	fibroblast growth factor 2 (basic)
FLNB	filamin B, beta (actin binding protein 278)
FOXN1	forkhead box N1
FUBP3	far upstream element (FUSE) binding protein 3
GDF5	growth differentiation factor 5
GH1	growth hormone 1
GHR	growth hormone receptor
GNA12	guanine nucleotide binding protein (G protein) alpha 12
GPR126	G protein-coupled receptor 126
GPR143	G protein-coupled receptor 143
GSTT2	glutathione S-transferase theta 2
HERC2	hect domain and RLD 2
HHIP	hedgehog interacting protein
HIST1H1D	histone cluster 1, H1d
HMGA1	high mobility group AT-hook 1
HMGA2	high mobility group AT-hook 2
HPS1	Hermansky-Pudlak syndrome 1
HPS6	Hermansky-Pudlak syndrome 6
IGF1	insulin-like growth factor 1 (somatomedin C)
IGF2	insulin-like growth factor 2 (somatomedin A)
IGFBP3	insulin-like growth factor binding protein 3
IHH	Indian hedgehog homolog (Drosophila)
IQCH	IQ motif containing H
KITLG	KIT ligand
LCORL	ligand dependent nuclear receptor corepressor-like
LEP	leptin
LIN28B	lin-28 homolog B (C. elegans)
LMNA	lamin A/C
LRP5	low density lipoprotein receptor-related protein 5
MAOA	monoamine oxidase A
MC1R	melanocortin 1 receptor (alpha melanocyte stimulating hormone receptor)
MITF	microphthalmia-associated transcription factor
MYO5A	myosin VA (heavy chain 12, myoxin)
NCAPG	non-SMC condensin I complex, subunit G
NOG	noggin
NPR2	natriuretic peptide receptor B/guanylate cyclase B (atrionatriuretic peptide receptor B)

OCA2	oculocutaneous albinism II
PAPPA	pregnancy-associated plasma protein A, pappalysin 1
PLAG1	pleiomorphic adenoma gene 1
POMC	proopiomelanocortin
PPARD	peroxisome proliferator-activated receptor delta
PPARG	peroxisome proliferator-activated receptor gamma
PTCH1	patched homolog 1 (Drosophila)
PXMP3	peroxisomal membrane protein 3, 35kDa
RAB27A	RAB27A, member RAS oncogene family
RNF135	ring finger protein 135
ROR2	receptor tyrosine kinase-like orphan receptor 2
SBF2	SET binding factor 2
SCMH1	sex comb on midleg homolog 1 (Drosophila)
SDR16C5	short chain dehydrogenase/reductase family 16C, member 5
SH3GL3	SH3-domain GRB2-like 3
SHOX	short stature homeobox
SILV	silver homolog (mouse)
SLC24A4	solute carrier family 24 (sodium/potassium/calcium exchanger), member 4
SLC24A5	solute carrier family 24, member 5
SLC45A2	solute carrier family 45, member 2
SOCS2	suppressor of cytokine signaling 2
SPAG17	sperm associated antigen 17
TBX2	T-box 2
TNFRSF11A	tumor necrosis factor receptor superfamily, member 11a, NFKB activator
TRIP11	thyroid hormone receptor interactor 11
TSEN15	tRNA splicing endonuclease 15 homolog (S. cerevisiae)
TXK	TXK tyrosine kinase
TYR	tyrosinase (oculocutaneous albinism IA)
TYRP1	tyrosinase-related protein 1
UQCC	ubiquinol-cytochrome c reductase complex chaperone
USP9Y	ubiquitin specific peptidase 9, Y-linked (fat facets-like, Drosophila)
VDR	vitamin D (1,25- dihydroxyvitamin D3) receptor
WDR60	WD repeat domain 60
ZBTB38	zinc finger and BTB domain containing 38
ZNF462	zinc finger protein 462
ZNF678	zinc finger protein 678

Appendix 2: Sample Chromosomal Gene Co-ordinates

Entrez_ Id	Chrom	Gene_Start	Gene_Stop
176	15	87147709	87218716
57188	15	82113842	82499598
197	3	187813544	187821799
25847	3	135679240	135687519
162282	17	51585835	51915006
8546	5	77333909	77626284
8943	19	2051993	2102556
434	20	32311832	32320809
4287	14	91598883	91642707
650	20	6696745	6708910
654	6	7672010	7826960
64771	6	34663050	34772603
387103	6	126702946	126711447
91768	18	18969725	19092614
1021	7	92072175	92301148
79145	8	57286869	57293728
1302	c6_COX	33199165	33228982
1278	7	93861809	93898480
1385	2	208102917	208171806
1588	15	49287546	49418099
1544	15	72828237	72835994
1545	2	38148250	38156796
1558	10	96786519	96819244
1559	10	96688405	96739133
1638	13	93889844	93929924
220107	13	50295451	50316076
26052	1	170077261	170648480
84444	19	2115148	2181007
1813	11	112785528	112851103
54808	18	44824170	45241077
2202	2	55946607	56004436
1956	7	55054219	55242524
2099	6	152053324	152466099
2150	5	76150610	76166895
10516	14	91405511	91490499
2200	15	46487797	46725210
2247	4	123967313	124038840
2317	3	57969167	58133018

Appendix 3: Sample Population data from dbSNP

rs_id	Observed allele	Observed allele freq	Observed allele2	Observed allele2 freq	Gen pop
rs1001522	G	0.623	A	0.377	EUROPE
rs1001522	G	0.167	A	0.833	EAST ASIA
rs1001522	G	0.211	A	0.789	EAST ASIA
rs1001522	G	0.3	A	0.7	WEST AFRICA
rs10016563	C	1			EAST ASIA
rs10016563	T	0.446	C	0.554	EUROPE
rs10016563	T	0.045	C	0.955	EAST ASIA
rs10016563	T	0.182	C	0.818	WEST AFRICA
rs10016563	T	0.521	C	0.479	NORTH AMERICA
rs10016563	T	0.217	C	0.783	NORTH AMERICA
rs10016563	T	0.021	C	0.979	NORTH AMERICA
rs10027991	G	0.917	A	0.083	NORTH AMERICA
rs10027991	G	0.795	A	0.205	NORTH AMERICA
rs10027991	G	0.979	A	0.021	NORTH AMERICA
rs10028204	G	0.958	A	0.042	EUROPE
rs10028204	G	0.911	A	0.089	EAST ASIA
rs10028204	G	0.933	A	0.067	EAST ASIA
rs10028204	G	0.817	A	0.183	WEST AFRICA
rs10032674	T	1			EUROPE
rs10032674	T	1			EAST ASIA
rs10032674	T	1			EAST ASIA
rs10032674	T	1			WEST AFRICA
rs10033663	A	0.478	C	0.522	NORTH AMERICA
rs10033663	A	0.526	C	0.474	NORTH AMERICA
rs1003366	A	0.833	C	0.167	NORTH AMERICA

3					
rs1003484	G	0.72	A	0.28	MULTI-NATIONAL
rs1003484	G	0.633	A	0.367	WEST AFRICA
rs1003484	G	0.763	A	0.237	NORTH AMERICA
rs1003484	G	0.556	A	0.444	NORTH AMERICA
rs1003484	G	0.409	A	0.591	NORTH AMERICA
rs1003484	G	0.587	A	0.413	NOT SPECIFIED
rs1003484	G	0.826	A	0.174	NOT SPECIFIED
rs1003484	G	0.625	A	0.375	MULTI-NATIONAL
rs1003484	G	0.694	A	0.306	MULTI-NATIONAL
rs1003484	G	0.565	A	0.435	MULTI-NATIONAL
rs1003484	G	0.717	A	0.283	CENTRAL/SOUTH AMERICA
rs1003484	G	0.5	A	0.5	EAST ASIA
rs1003484	G	0.625	A	0.375	EUROPE
rs1003484	G	0.444	A	0.556	EAST ASIA
rs1003484	G	0.511	A	0.489	EAST ASIA
rs1003484	G	0.608	A	0.392	WEST AFRICA

Appendix 4: Sample Population data from HapMap

Rs id	Ref Allele	Ref Allele Freq	Other Allele	Other Allele Freq	Pop
rs10039112	T	0.596	G	0.404	ASW
rs10039112	T	0.649	G	0.351	CEU
rs10039112	T	0.732	G	0.268	CHB
rs10039112	T	0.729	G	0.271	CHD
rs10039112	T	0.711	G	0.289	GIH
rs10039112	T	0.738	G	0.262	JPT
rs10039112	T	0.386	G	0.614	LWK
rs10039112	T	0.745	G	0.255	MEX
rs10039112	T	0.35	G	0.65	MKK
rs10039112	T	0.695	G	0.305	TSI
rs10039112	T	0.491	G	0.509	YRI
rs10039931	C	0.489	A	0.511	ASW
rs10039931	C	0.45	A	0.55	CEU
rs10039931	C	0.64	A	0.36	CHB
rs10039931	C	0.671	A	0.329	CHD
rs10039931	C	0.608	A	0.392	GIH
rs10039931	C	0.64	A	0.36	JPT
rs10039931	C	0.343	A	0.657	LWK
rs10039931	C	0.436	A	0.564	MEX
rs10039931	C	0.269	A	0.731	MKK
rs10039931	C	0.532	A	0.468	TSI
rs10039931	C	0.407	A	0.593	YRI
rs10043750	C	0.745	T	0.255	ASW
rs10043750	C	0.64	T	0.36	CEU
rs10043750	C	0.72	T	0.28	CHB
rs10043750	C	0.714	T	0.286	CHD
rs10043750	C	0.723	T	0.277	GIH
rs10043750	C	0.738	T	0.262	JPT
rs10043750	C	0.578	T	0.422	LWK
rs10043750	C	0.766	T	0.234	MEX
rs10043750	C	0.5	T	0.5	MKK
rs10043750	C	0.688	T	0.312	TSI
rs10043750	C	0.644	T	0.356	YRI
rs10054454	T	0.915	C	0.085	ASW
rs10054454	T	0.667	C	0.333	CEU
rs10054454	T	0.841	C	0.159	CHB
rs10054454	T	0.793	C	0.207	CHD
rs10054454	T	0.759	C	0.241	GIH

Appendix 5: Sample SNP descriptions

rs id	allele	chrom pos	SNP class	fxn class	flanking sequence	validation
rs10080749	A/G	6:34326142	SNP		gtgtcctccagctctggatccactg**[A/G**]acaaaccataaaagcaagacctgaa	no-info
rs10080956	A/G	6:142741586	SNP	intron	tcccacacaactcaaactattttct**[A/G**]tagagttactaccaaaatcccaatg	no-info
rs10081143	C/T	6:35476760	SNP	intron	tcatgccaagcagtgccaagcgcca**[C/T]**gtctgctccccagatacctgtgttt	by-hapmap
rs10081480	G/T	8:78074978	SNP	utr-5	gagaactgcgctttagcggcgctgc[g/t]gaaggcactggatggccaaacaacc	by-cluster,by-frequency, by-hapmap
rs10082581	C/G	11:67859347	SNP	intron	gcccaggctagtctcaaacttctgg**[C/G**]ctcaagtgaacctcccaccttggcc	by-cluster
rs10082583	A/G	11:67859369	SNP	intron	tggcctcaagtgaacctcccacctt**[A/G]**gcctcccaaagtgttgggattacag	by-cluster
rs10082622	C/T	11:67859368	SNP	intron	ctggcctcaagtgaacctcccacct**[C/T]**ggcctcccaaagtgttgggattaca	by-cluster
rs10082655	A/C	11:67859339	SNP	intron	gccatgttgcccaggctagtctcaa**[A/C]**cttctggcctcaagtgaacctccca	by-cluster
rs10082656	A/T	11:67859357	SNP	intron	gtctcaaacttctggcctcaagtga**[A/T]**cctcccaccttggcctcccaaagtg	by-cluster
rs10082674	C/T	11:9777601	SNP	intron	cccaaactggggttaacgatcaggt**[C/T**]tactggaagtggaagagaagatgca	by-hapmap
rs10083198	C/T	12:46582231	SNP	intron	tatgtccattgtcctggtataacca**[C/T]**gcatggcatcccctttcactctcag	by-2hit-2allele,by-cluster,by-frequency
rs10083447	A/G	14:91534179	SNP	intron	gcaaaggtggggtaggggttaattc**[A/G]**caggaatattgtctctactagccta	by-2hit-2allele,by-frequency,by-hapmap
rs10083510	C/T	14:91532249	SNP	intron	aaactccatacccattaaatagtaa**[C/T]**tcccctttcccttgaaaactaccat	by-2hit-2allele,by-frequency,by-hapmap
rs10083569	C/T	15:82225224	SNP	intron	agcatttccaataagtgaccttata**[C/T]**gtgaaatctgtgtttaattactaat	by-2hit-2allele

Appendix 6: Number of Snps per Gene

COUNT	ENTREZ_ID	GENE_SYMBOL
1048	2099	ESR1
781	81846	SBF2
777	1956	EGFR
770	26052	DNM3
687	4920	ROR2
659	4948	OCA2
541	162282	ANKFN1
484	57188	ADAMTSL3
480	2690	GHR
470	1588	CYP19A1
466	2200	FBN1
465	5069	PAPPA
410	54808	DYM
396	4644	MYO5A
394	5468	PPARG
360	123041	SLC24A4
352	4286	MITF
341	2317	FLNB
328	654	BMP6
326	1302	COL11A2
322	3479	IGF1
301	1021	CDK6
295	8546	AP3B1
286	200162	SPAG17
279	7421	VDR
257	64799	IQCH
252	22955	SCMH1
247	2768	GNA12
245	5467	PPARD
242	254251	LCORL
241	2247	FGF2
241	7299	TYR
226	6457	SH3GL3
211	1559	CYP2C9
200	4041	LRP5
200	176	ACAN
199	57211	GPR126
190	10516	FBLN5
187	1278	COL1A2
185	58499	ZNF462

184	8091	HMGA2
183	1558	CYP2C8
180	8924	HERC2
175	4287	ATXN3
166	55245	UQCC
163	339500	ZNF678
162	4254	KITLG
161	9321	TRIP11
159	253461	ZBTB38
158	1813	DRD2
154	5873	RAB27A
149	55112	WDR60
148	3481	IGF2
145	8792	TNFRSF11A
143	51151	SLC45A2
142	91768	CABLES1
139	64399	HHIP
138	3952	LEP
136	3257	HPS1
133	64771	C6orf106
127	389421	LIN28B
123	1545	CYP1B1
121	2150	F2RL1
114	7294	TXK
99	1638	DCT
96	5727	PTCH1
93	1544	CYP1A2
91	4128	MAOA
90	2202	EFEMP1
81	84444	DOT1L
80	3486	IGFBP3
75	4157	MC1R
75	1385	CREB1
73	8943	AP3D1
72	64151	NCAPG
68	8939	FUBP3
68	4000	LMNA
67	6490	SILV
65	4935	GPR143
65	5828	PXMP3
63	7306	TYRP1
57	650	BMP2
56	2688	GH1
52	197	AHSG
50	3549	IHH
50	2953	GSTT2

48	8835	SOCS2
47	6909	TBX2
47	220107	DLEU7
47	8200	GDF5
46	283652	SLC24A5
44	5443	POMC
43	116461	TSEN15
43	8456	FOXN1
42	5324	PLAG1
41	4882	NPR2
38	25847	ANAPC13
38	195814	SDR16C5
37	3007	HIST1H1D
34	84282	RNF135
32	434	ASIP
31	8287	USP9Y
30	9241	NOG
29	3159	HMGA1
20	6473	SHOX
20	387103	C6orf173
16	79145	CHCHD7
10	79803	HPS6

Appendix 7: SNP Functional Classes

65475	intron
14020	
1857	utr-3
1165	missense
1113	near-gene-5
967	near-gene-3
914	coding-synonymous
307	intron
231	intron
187	intron
180	utr-5
79	frameshift
52	near-gene-3
48	Intron, reference
43	near-gene-3
35	intron
29	Missense, reference
26	coding-synonymous, reference
20	Intron, utr-5
18	nonsense
18	Missense, utr-3
17	near-gene-5
16	coding-synonymous, reference
11	near-gene-5
10	Missense, reference
9	Intron, near-gene-5
6	coding-synonymous, reference
6	Intron, near-gene-3, reference
6	splice-3
5	frameshift
5	near-gene-3
5	splice-5
4	coding-synonymous, near-gene-3, reference
3	Frameshift, reference
3	coding-synonymous, reference, reference
3	coding-synonymous, utr-5
2	coding-synonymous, utr-3
2	Intron, utr-3
2	Missense, utr-3, utr-5

2	Frameshift, urt-3
2	Frameshift, reference
2	Missense, urt-5
1	coding-synonymous, reference
1	Intron
1	coding-synonymous, reference
1	Intron, urt-3
1	Frameshift, near-gene-3, reference
1	Intron, utr-5
1	Intron, urt-3, utr-5
1	Frameshift, utr-5
1	Intron, near-gene-5, reference