# Investigation of the *OCA2* gene control regions and their possible role in normal pigment variation

Micaela Tanya Eisenberg

Supervisors:

Dr Thandiswa Ngcungcu

Dr Robyn Kerr

Mr Jorge da Rocha


A dissertation submitted to the Faculty of Health Science, University of the Witwatersrand, Johannesburg, in fulfilment of the requirements for the degree of Master of Science in Medicine.

Johannesburg, June 2020

# Declaration

I, Micaela Tanya Eisenberg, declare that this dissertation is my own unaided work. It is being submitted to the degree of Master of Science in Medicine to the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination to any other University.

.....

(Signature of Candidate)

19th day of June 2020 in Tel Aviv-Yafo, Israel.

# Abstract

Human pigmentation phenotypes form a diverse spectrum and the genetic basis of pigmentation is complex. The oculocutaneous albinism type II (*OCA2*) gene has been shown to have a role in both normal and abnormal human pigmentation. Variants in the regulatory regions of the *OCA2* gene have been suggested to influence its expression of the gene and by extension, contribute to the normal range of human pigment phenotypes. The promoter of the gene, previously identified, and an enhancer for *OCA2*, have been postulated as important determinants of variation in pigment phenotypes. The potential enhancer region occurs within intron 86 of the upstream neighbouring Hect Domain and RCC1-like Domain 2 (*HERC2*) gene which has no known pigment functions. This region has previously been characterised and determined to have enhancer-like properties where it is specifically active in melanocytes. Thus, it was hypothesised that normal variation in the promoter and enhancer regions of the *OCA2* gene may alter transcriptional regulation of the gene and this may contribute to the range of normal human pigment phenotypes. The aim of this study was to investigate the normal variation of the *OCA2* gene control region in black African individuals, which includes the associated promoter and this putative enhancer.

During this project, the putative enhancer region was initially interrogated to determine if this region interacts with the *OCA2* promoter, if it has enhancer-like properties and if these properties are exclusive to melanocytes. Hi-C, virtual chromatin conformation capture-on-chip and Chromatin Interaction Analysis by Paired-End Tag data were accessed to investigate possible interactions of the promoter and putative enhancer. Additionally, histone modification and chromatin conformation data from the Roadmap Epigenomics project were utilised to investigate enhancer-like properties of the putative enhancer region in melanocytes and other cell types. Following this, the normal variation within the *OCA2* regulatory regions was investigated in publicly available data from the 1000 Genomes Project (KGP) and the African Genome Variation Project (AGVP). Common variation in the promoter and enhancer were identified in African populations and annotated with bioinformatics tools to determine if these normal variants are possibly functional in modulating *OCA2* expression and impacting pigment variation in the normal population.

The putative enhancer region was shown to interact with the *OCA2* promoter and to have enhancer-like properties in melanocyte derived cancer cell lines and normal melanocytes. This

confirms that the region has enhancer-functionality in melanocytes. A total of 287 variants were identified in the two regulatory regions combined, which included variants that were shared between the KGP and AGVP datasets. The evidence generated from the bioinformatic tools was pooled to narrow down the collection of variants to a set of 7 variants from the promoter and 10 variants from the enhancer which had the strongest evidence for being functionally significant. rs12913832 and rs7495174, from the enhancer and promoter respectively, had been previously associated with pigment phenotypes. These variants had the strongest collected evidence for functionality and modulating the expression of *OCA2*, which could affect the pigment phenotype of individuals.

The purpose of this study was to characterise the normal variation in the regulatory regions of the *OCA2* gene. There are variants in these regulatory regions that are likely to be functional and that differ in frequency between African and non-African populations. These variants may be associated with altered pigment phenotypes that are common in African ancestry populations. Since the *OCA2* gene is involved in abnormal pigment disorders such as albinism, this work supports the search for causal mutations for associated disorders in the regulatory regions of the gene. An established catalogue of the normal variation has now been produced and could be used in future studies to help in identifying mutations that are truly pathogenic from those that are simply part of normal variation in the African population. This would prevent false positive associations of normal variation with albinism. Additionally, functional normal variants can be investigated to determine their contribution to normal pigment phenotypes in functional studies.

# Dedication

For Apollo

# Acknowledgements

To my supervisors: Thandiswa, Robyn and Jorge, thank you. Thank you for your support, encouragement and guidance throughout this process. Thank you for believing in me, none of this would have been possible without each of you. It has been a true privilege to be your student.

Thank you to Prof. Amanda Krause and Prof. Michéle Ramsay. Your guidance and support during a challenging period are truly appreciated.

Thank you to my family. To my parents, Simon and Melanie, thank you for always supporting and encouraging me at every turn. Thank you to my siblings: Gabriel, for pushing me to expand my horizons and Ariella, for your eye for alignment.

Thank you to my friends and colleagues at the Division of Human Genetics and the Sydney Brenner Institute for Molecular Bioscience. Thank you to my friends, both near and far.

**Thank you to the financial supporters of this project:**

National Research Foundation of South Africa (DST-NRF Innovation Master's Scholarship, grant number: 114312)

University of the Witwatersrand: Faculty of Health Science Faculty Research Committee Individual Research Grants (grant number: 001 254 8466101 5121105 000000 0000000000 5254) and Postgraduate Merit Award.

South African Society for Human Genetics for funding for registration and the Division of Human Genetics for travel funding to attend the South African Society for Human Genetics Biennial Congress.

# Presentations arising from this study

**South African Society for Human Genetics Biennial Congress, 3-6 August 2019, Cape Town, South Africa**

Poster presentation: Bioinformatic investigation of the *OCA2* gene control regions and their possible role in normal pigment variation

**Molecular Biosciences Research Thrust Postgraduate Research Day, 28 November 2019, University of the Witwatersrand, Johannesburg, South Africa**

Poster presentation: Bioinformatic investigation of the *OCA2* gene control regions and their possible role in normal pigment variation

# Table of contents

# List of figures

# List of tables

# List of abbreviations

| | |
|---|---|
| 3C | Chromosome Conformation Capture |
| 4C | Chromosome Conformation Capture-on-Chip |
| AGVP | African Genome Variation Project |
| AS | Angelman Syndrome |
| BOCA | Brown oculocutaneous albinism |
| Bp | Base pair |
| CADD | Combined Annotation Dependent Depletion |
| ChIA-PET | Chromatin Interaction Analysis by Paired-End Tag |
| ChIP-seq | Chromatin immunoprecipitation sequencing |
| CHS | Chediak-Higashi Syndrome |
| dbSNP | Single Nucleotide Polymorphism Database |
| *DCT* | Dopachrome tautomerase gene |
| DHI | Dihydroxyindole |
| DHICA | Dihydroxyindole carboxylic acid |
| DNA | Dioxyribonucleic acid |
| DNase-seq | DNase I hypersensitive site sequencing |
| DOPA | L-3,4-dihydroxyphenylalanine |
| ENCODE | Encyclopaedia of DNA Elements |
| eQTL | Expression quantitative trait loci |
| FAIRE | Formaldehyde-assisted identification of regulatory elements |
| GRCh37 | Genome Reference Consortium Human Build 37 |
| GTEx | Genotype-Tissue Expression project |
| GWAVA | Genome Wide Annotation of Variants |
| H3K4me1 | Monomethylation at lysine 4 on histone 3 |

| | |
|---|---|
| H3K4me2 | dimethylation of lysine 4 of histone 3 |
| H3K4me3 | Trimethylation of lysine 4 on histone 3 |
| H3K27ac | Acetylation of lysine 27 of histone 3 |
| H3K27me3 | Trimethylation of lysine 27 of histone 3 |
| HCT116 | Colorectal carcinoma cell line |
| HeLa S3 | Cervix carcinoma cell line subclone |
| *HERC2* | Hect Domain and RCC1-like Domain 2 gene |
| HPS | Hermansky-Pudlak Syndrome |
| HREC | Human Research Ethics Committee |
| K562 | Chronic myeloid leukemia cell line |
| Kb | Kilo base |
| KGP | 1000 Genomes Project |
| MAF | Minor allele frequency |
| Mb | Mega base |
| *MC1R* | Melanocortin 1 receptor gene |
| MCF7 | Michigan Cancer Foundation-7 breast cancer cell line |
| NB4 | Acute promyelocytic leukemia cell line |
| NHEK | Primary Normal Human Epidermal Keratinocytes |
| OA | Ocular albinism |
| OCA | Oculocutaneous albinism |
| OCA2 | Oculocutaneous albinism type II |
| *OCA2* | Oculocutaneous albinism type II gene |
| POL2 | RNA polymerase II |
| PWS | Prader-Willi Syndrome |
| RNA | Ribonucleic acid |
| RNA-seq | RNA sequencing |
| RPMI7951 | Malignant melanoma cell line |

| | |
|---|---|
| *SLC45A2* | Solute carrier family 45 member 2 gene |
| *SLC24A5* | Solute carrier family 24 member 5 gene |
| SNP | Single nucleotide polymorphism |
| SNV | Single nucleotide variant |
| ssDNA | Single stranded DNA |
| TF | Transcription factor |
| *TYRP1* | Tyrosinase related protein 1 gene |
| TSS | Transcription start site |
| *TYR* | Tyrosinase gene |
| UCSC | University of California, Santa Cruz |
| UTR | Untranslated region |
| UV | Ultraviolet |
| UVA | Ultraviolet A |
| UVB | Ultraviolet B |
| VCF | Variant Call File |
| VEP | Ensembl Variant Effect Predictor |
| WGS | Whole Genome Sequences |

**Population codes**

| | |
|---|---|
| ACB* | African Caribbeans in Barbados (KGP) |
| AMR* | Admixed Americans (KGP) |
| ASW* | Americans of African Ancestry in South West United States of America (KGP) |
| BAG | Baganda from Uganda (AGVP) |
| EAS* | East Asians (KGP) |
| EUR* | Europeans (KGP) |

| | |
|---|---|
| ESN* | Esan in Nigeria (KGP) |
| ETH | Ethiopian from Ethiopia (Amhara, Oromo, Somali, Wolayta, Gumuz ethnic groups) (AGVP) |
| GWD* | Gambian in the Western Divisions in the Gambia (KGP) |
| LWK* | Luhya in Webuye, Kenya (KGP) |
| MSL* | Mende in Sierra Leone (KGP) |
| SAS* | South Asians (KGP) |
| YRI* | Yoruba in Ibadan, Nigeria (KGP) |
| ZUL | Zulu from South Africa (AGVP) |

*****(The 1000 Genomes Project Consortium 2015)

# Chapter 1 - Introduction

Human pigmentation has long been a topic of research interest. Hair, skin and eye colours can be distinctive, and can vary widely. The colour of these features can be more common in certain populations than others, but a wide range of phenotypes within a population is also possible. It has been established that there is a relationship between pigment phenotype and genetics. Pigmentation of the hair, skin and eyes is a polygenic trait. The genetics of pigmentation are important to study to understand the underlying genetic control of normal pigmentation as well as abnormal pigment disorders. Additionally, by determining which genetic factors contribute to normal pigment variation, it can become possible to identify the cause of abnormal phenotypes by excluding normal variation in genetic studies. The genetics of pigmentation is important to research in Africa, considering that the variation in skin colour is wide and there are conditions of abnormal pigmentation which are common in African populations.

## 1.1 Melanin and pigment in the human body

Melanocytes are cells which are responsible for the production of pigment in humans. They originate from neural crest cells and are found in the hair, skin and iris (as reviewed by Mort et al. 2015). In the skin, melanocytes are located in the basal layer of the epidermis (Figure 1.1), while they occur in the proximal bulb of hair follicles (Tobin 2011). The sites where melanocytes occur in the hair and eye are shown in Appendix C. Melanocytes can also be found in organs where the cells do not have a pigment related function - such as in the inner ear, the nervous system and the heart (Steel and Barkway 1989). Thus, melanocytes can have different functions depending on which tissue they are located in.

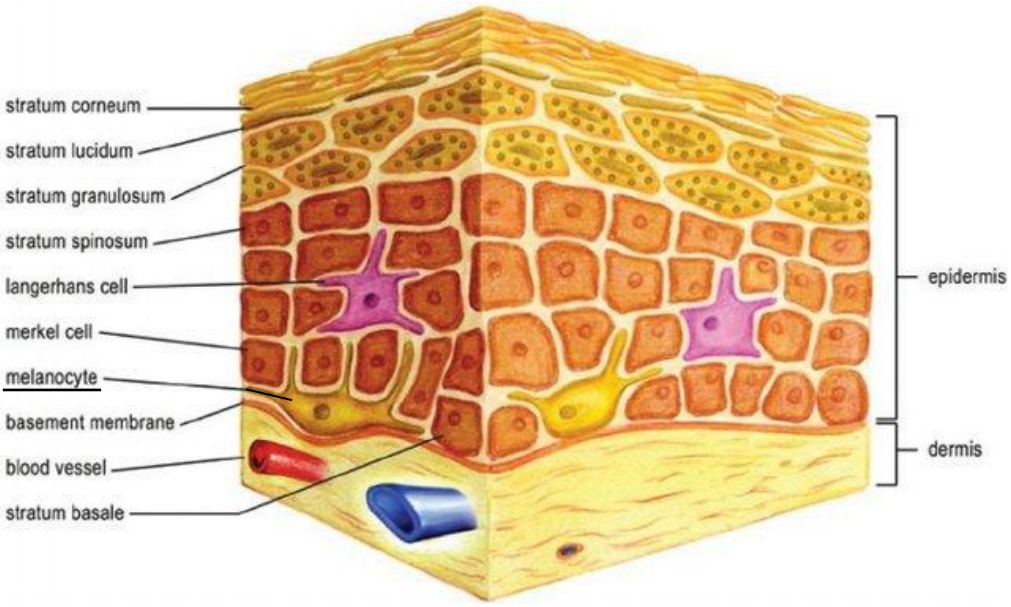



**Figure 1.1: The structure of the epidermis.** The position of the melanocytes in the epidermis is indicated by the underlined label (adapted from Farage et al. 2007).

Melanocytes which have pigment functions produce melanin, a protein that determines the colour of human skin, hair and eyes. It is produced in melanosome organelles which occur within the melanocyte cells. In the skin and hair, melanin containing vesicles that originate from the melanocyte plasma membrane are transported to the surrounding keratinocytes through the process of melanosome transport (Tadokoro et al. 2016). There are two major types of melanin which are responsible for pigment within the human body. These are eumelanin, which gives a black-brown colour when present, and pheomelanin, whose presence is associated with a yellow-red colour (Prota 1980a).

### 1.1.1 Normal variation of pigmentation

Melanin related pigmentation of hair, eye and skin colour varies extensively within and between human populations (Figure 1.2). Additionally, particular combinations of hair, skin and eye colours are typical of certain populations and can be rarer in others, such as blonde and red hair, which naturally occur only in European ancestry populations.

**Figure 1.2: A collection of individuals displaying a spectrum of naturally occurring human pigment phenotypes.** (adapted from Angélica Dass, Humanae Project, https://www.angelicadass.com/humanae-project).

Both eumelanin and pheomelanin are found in the hair as well as the skin, and contribute to the pigment of these features (Prota 1980a; Thody et al. 1991). A higher concentration of eumelanin is present in darker skin tones and a lower concentration of eumelanin corresponds to lighter skin tones (Thody et al. 1991). The concentration of pheomelanin may differ between skin tones, though pheomelanin content does not correlate strongly with the overall pigment of the skin (Thody et al. 1991). It has been determined that the concentration of eumelanin relative to the total amount of melanin is responsible for the tone of skin colour (Thody et al. 1991; Slominski et al. 2004). Therefore, variation in the amount of eumelanin present in the skin is a determinant in the variation of human skin tones (Thody et al. 1991).

The ratio of eumelanin to pheomelanin in hair determines the overall hair colour of the individual (as discussed in Ito and Wakamatsu 2011a). As seen in Figure 1.3, the presence of eumelanin in the hair gives rise to brown colours. In a spectrum of hair colour from black to blonde, the highest concentration of eumelanin results in darker brown or black hair, while the lowest concentration of eumelanin is present in blonde hair. Pheomelanin is present in similar amounts in all hair colours but in a smaller quantity than eumelanin. The exception is red hair, which has approximately equal quantities of eumelanin and pheomelanin (Ito et al. 2011). Therefore, the increased proportion of pheomelanin in red hair is responsible for its characteristic colour**.**

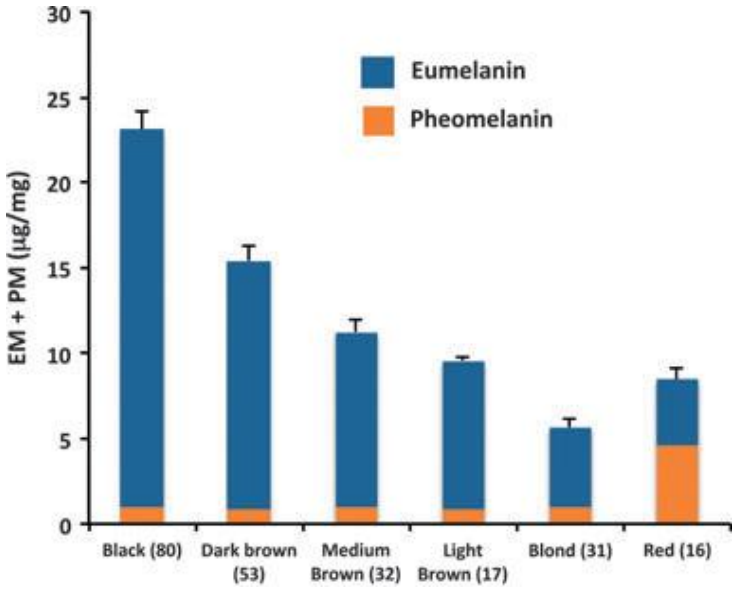



**Figure 1.3: The relative melanin content of 228 human hair samples of different colours.** The content of eumelanin and pheomelanin in these hair samples is indicated (Ito et al. 2011).

Eumelanin concentration also determines eye colour, where a larger amount of eumelanin in the iris results in brown eye colour and little eumelanin is present in blue eyes. However, other light eye colours (green, grey and hazel) and shades of blue are determined by factors such as the structure and the quantity of collagen in the iris. Therefore, the role of melanin in eye colour is restricted to blue or brown eye colour (as reviewed by Sturm and Larsson 2009).

### 1.1.2 Melanin biosynthesis pathway

Melanogenesis is a multistep melanin synthesis pathway that converts the amino acid tyrosine to melanin (Figure 1.4). The first step occurs when tyrosinase hydroxylates tyrosine to L-3,4-dihydroxyphenylalanine (DOPA). Thereafter, tyrosinase converts DOPA to DOPAquinone (Lerner et al. 1949). The pathway will then diverge depending on if eumelanin or pheomelanin is being synthesised. In the presence of cysteine, DOPAquinone will react with cysteine and be converted to cysteinylDOPA. This will then undergo oxidation and polymerisation to pheomelanin (Prota 1980b). Without cysteine, DOPAquinone will spontaneously undergo cyclisation to DOPAchrome, which is followed by a series of redox reactions to give rise to the intermediates dihydroxyindole (DHI) and dihydroxyindole carboxylic acid (DHICA). These will then polymerise to generate eumelanin (Raper 1926; Mason 1948).

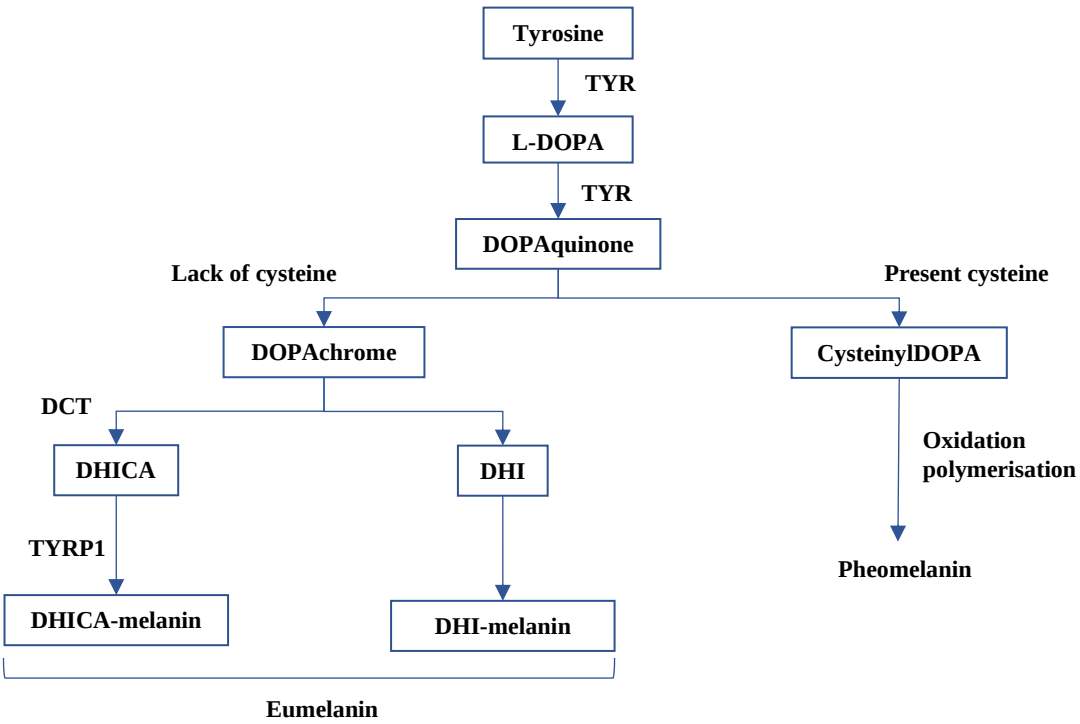



**Figure 1.4: A simplified representation of the melanin biosynthesis pathway.** The products of this pathway are eumelanin and pheomelanin (adapted from Cichorek et al. 2013)

### 1.1.3 The function of melanin

The sun emits ultraviolet (UV) radiation which is comprised primarily of ultraviolet A (UVA) and to a lesser extent, ultraviolet B (UVB). Exposure of the skin to UV radiation is a risk to skin health at the DNA level. UVA generates DNA damaging reactive oxygen species which cause single stranded DNA (ssDNA) breaks and crosslinking between DNA and proteins (Brenner and Hearing 2008). UVB is directly absorbed by the DNA which forms pyrimidine dimers, where neighbouring pyrimidine bases covalently bond together (Brenner and Hearing 2008). Normally, exposure to UV radiation induces apoptosis of melanocytes that have acquired damaging mutations, which prevents further growth of the damaged cells (Bivik et al. 2006). If this DNA damage is not repaired, damaged cells can escape apoptosis and form melanomas (Kobayashi et al. 1993; Bivik et al. 2006).

Melanin, specifically eumelanin, is functional in response to the normal irradiation of human skin by UV radiation. Eumelanin is photoprotective as it protects DNA in the skin against the mutagenic effects of UV radiation (Kobayashi et al. 1993). Eumelanin collects above the nucleus of individual epidermal keratinocytes, in the direct path of UV radiation which is bound for the nucleus (Montagna and Carlisle 1991). Eumelanin acts as a filter to scatter and absorb

the incident UV radiation, thereby preventing it from reaching the lower epidermis (Kaidbey et al. 1979). Moreover, eumelanin acts in a concentration dependent manner, where photoprotection is more effective in skin which has increased melanin content and is consequently darker in pigment (Kobayashi et al. 1993).

Following exposure to UVB radiation, human skin darkens in colour by increasing melanin production in a normal process called tanning. UV induced photodamage or the repair of this damage is a factor that stimulates melanogenesis in the skin (Gilchrest and Eller 1999). Melanin synthesis occurs three to five days after initial UVR exposure and occurs to protect the epidermal DNA against further UV injury. This occurs parallel with increased activity of tyrosinase for melanin production in the melanocytes (Eller and Gilchrest 2000).

In contrast, pheomelanin is not photoprotective (Thody et al. 1991). Pheomelanin produces a higher concentration of free radicals when exposed to UV radiation, compared to irradiated eumelanin (Thody et al. 1991). Additionally, UV exposed pheomelanin can become mutagenic and cause mutations in melanocytes. This has been suggested as the reason for freckle development and the susceptibility of lightly pigmented individuals to skin cancers (Harsanyl et al. 1980). Irradiation of pheomelanin with UVR can encourage the release of histamines and is associated with higher levels of apoptosis in cells (Cesarini 1988; Hill and Hill 2000). Therefore, UV irradiation of pheomelanin can contribute to skin damage (Thody et al. 1991).

## 1.2 Genes known to be involved in melanogenesis

Many genes form part of this complex molecular pathway, thus pigment biosynthesis is polygenic (Table 1.1). The genes which directly code for enzymes involved in melanin biosynthesis include tyrosinase (*TYR*) which codes for the enzyme that catalyses the initial steps of the pathway (Lerner et al. 1949). Tyrosinase related protein 1 (*TYRP1*) and dopachrome tautomerase (*DCT*, previously known as tyrosinase related protein 2, *TYRP2*) code for enzymes that are involved in later steps of the biosynthesis of eumelanin. DCT converts DOPAchrome to DHICA and TYRP1 catalyses the further conversion of DHICA to DHICA-melanin, which is a light brown eumelanin (Bouchard et al. 1994; Kobayashi et al. 1994).

Other genes encode proteins which are not directly involved in the pathway but have been implicated in pigment related functions by their association with pigment variation. The oculocutaneous albinism type II (*OCA2*) gene, which was previously known as the *P* gene, codes for a melanosomal membrane associated protein, though the function of the OCA2

protein is unclear (Oetting et al. 2005). This gene will be discussed in more detail in section 1.3.

Another important gene in this pathway is melanocortin 1 receptor (*MC1R*), which codes for a melanocyte surface G coupled receptor for α melanocyte stimulating hormone. When the MC1R protein is stimulated by its ligand, eumelanin synthesis is triggered via a signalling pathway. The DNA damage response in melanocytes is simultaneously activated against photo damage to prevent cancer formation (García-Borrón et al. 2014).

Solute channel proteins have also been linked to pigment phenotypes. The solute carrier family 45 member 2 gene (*SLC45A2*) codes for melanosome membrane associated transport protein (MATP) which is thought to modify the activity of tyrosinase by regulating the pH of the melanosome (Bin et al. 2015). Additionally, the solute carrier family 24 member 5 gene (*SLC24A5*) codes for a potassium-dependent sodium/calcium exchanger protein (NCKX5) and this exchange function contributes to melanogenesis (Ginger et al. 2008). Therefore, genes that are directly and indirectly involved in melanin synthesis have an impact on the individual pigment phenotype.

Melanogenesis during tanning leads to increased expression of genes involved in the melanin synthesis pathway (Suzuki et al. 2002). The mRNA levels of these genes were examined by in situ hybridisation and detection of the mRNA levels in control and UVB irradiated skin samples (Suzuki et al. 2002). The upregulated genes in UVB irradiated skin in the melanin synthesis pathway included *TYR*, *TYRP1*, *DCT*, *PMEL*, *OCA2* and *MITF* (Suzuki et al. 2002). Therefore, tanning, which is a normal bodily response, upregulates genes directly and indirectly involved in melanogenesis. This suggests that the relative level of expression of pigment genes can determine the pigment phenotype.

**Table 1.1: The more common pigment genes, their normal functions and associations with disease when disrupted.**

| Gene | Chromosomal location | Protein action | Associated disease | OMIM number | References |
|---|---|---|---|---|---|
| *TYR* | 11q14.3 | Catalyses the initial steps of melanin synthesis from tyrosine | OCA1 | 606933 | (Lerner et al. 1949; Tomita et al. 1989; Giebel et al. 1990) |
| *TYRP1* | 9p23 | Catalyses the conversion of DHICA to DHICA-melanin | OCA3 | 115501 | (Kobayashi et al. 1994; Boissy et al. 1996; Manga et al. 1997) |
| *DCT* | 13q32.1 | Catalyses the conversion of dopachrome to DHICA | - | 191275 | (Bouchard et al. 1994) |
| *OCA2* | 15q11.2-q12 | Transports tyrosine into the melanosome, processes the tyrosinase enzyme and regulates melanosomal pH | OCA2 | 611409 | (Rinchik et al. 1993; Rosemblat et al. 1998; Puri et al. 2000; Toyofuku et al. 2002; Chen et al. 2002) |
| *MC1R* | 16q24.3 | Receptor for α melanocyte stimulating hormone. Stimulates eumelanin synthesis and activates DNA damage responses | Melanoma risk | 155555 | (Abdel-Malek et al. 1995; Böhm et al. 2005; Kadekaro et al. 2005) |
| *SLC45A2* | 5p13.2 | Regulates the melanosomal pH | OCA4 | 606202 | (Newton et al. 2001; Bin et al. 2015) |
| *SLC24A5* | 15q21.1 | Potassium-dependent exchange of sodium and calcium across the melanosomal membrane | OCA6 | 609802 | (Ginger et al. 2008; Wei et al. 2013) |

OCA1 - oculocutaneous albinism type I, OCA2 - oculocutaneous albinism type II, OCA3 - oculocutaneous albinism type III, OCA4 - oculocutaneous albinism type IV, OCA6 - oculocutaneous albinism type VI. *TYR* - tyrosinase, *TYRP1* - tyrosinase related protein 1, *DCT* - dopachrome tautomerase, *OCA2* - oculocutaneous albinism type II, *MC1R* - melanocortin 1 receptor, *SLC45A2* - solute carrier family 45 member 2, *SLC24A5* - solute carrier family 24 member 5.

## 1.3 The oculocutaneous albinism type II (*OCA2*) gene

### Location and structure

The *OCA2* gene is located on chromosome 15. The gene lies in close proximity to the imprinted region that is involved in Prader-Willi and Angelman syndromes (Butler et al. 1996). The *OCA2* gene has 25 exons, of which 24 are coding - the first exon forms part of the 5' untranslated region. *OCA2* codes for a melanosomal membrane associated protein which has 12 transmembrane domains (Gardner et al. 1992; Rinchik et al. 1993; Lee et al. 1995).

### Function

Though the function of the OCA2 protein is unclear (Oetting et al. 2005), several possible functions have been proposed. The OCA2 protein has been suggested to transport small molecules like tyrosine and process tyrosinase in melanocytes (Rosemblat et al. 1998; Toyofuku et al. 2002; Chen et al. 2002). Other possible functions of the protein include regulation of the metabolism of glutathione and regulating the pH within the melanosome (Puri et al. 2000; Staleva et al. 2002).

Though its function has not been completely elucidated, inhibition of OCA2 functioning has been shown to cause changes in the morphology and number of melanosomes in the melanocytes (Park et al. 2015). This suggests that the OCA2 protein is involved in melanocyte structure and function. Despite its indirect role in melanogenesis, the *OCA2* gene has been linked to both normal and abnormal pigment phenotypes.

### Expression pattern

The *OCA2* gene has been implicated in both normal and abnormal pigment in humans from varying ancestries. The *OCA2* gene was initially linked with abnormal pigment owing to OCA2 albinism, which later resulted in the association of *OCA2* with normal brown eye and hair colour (Eiberg and Mohr 1996). Thereafter, *OCA2* was considered a pigment gene and has been consistently replicated to correlate with pigmentation phenotypes in various studies on several populations. The expression of the *OCA2* gene is mostly localised in the skin and retina as seen in Figure 1.5 (Uhlén et al. 2015). This correlates with the role of the gene in eye and skin colour.

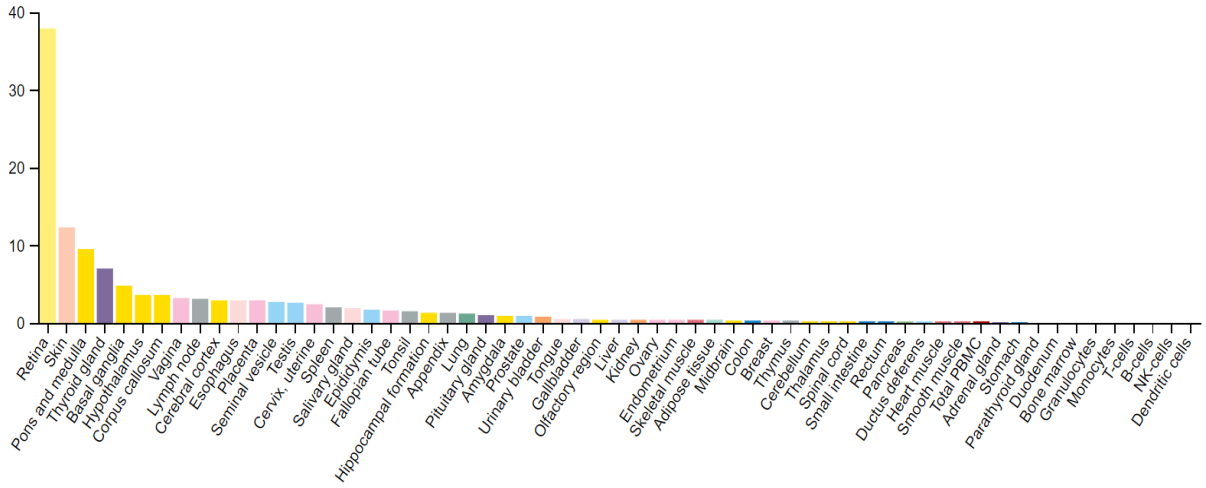



**Figure 1.5: The expression of the *OCA2* gene in various tissues from the Human Protein Atlas (Uhlén et al. 2015).** *OCA2* gene expression is highest in the retina and the skin. This normalised expression data is from a consensus set from the Human Protein Atlas, Genotype-Tissue Expression project and FANTOM5 project (https://www.proteinatlas.org/ENSG00000104044-OCA2/tissue).

## Normal variation

Variation within the *OCA2* gene has been investigated for its effect on normal human pigmentation using blue and nonblue iris colour as a proxy in European individuals (Duffy et al. 2007). The most significant associations with iris colour variation were from the single nucleotide polymorphisms (SNPs) rs7495174 T/C, rs4778241 A/C (formerly rs6497268) and rs4778138 A/G (formerly rs11855019) which are found in the *OCA2* 5' untranslated region (UTR). No other significant associations were identified to variants within the gene (Duffy et al. 2007). Thus, variants in *OCA2* regulatory regions were deemed to be more likely to influence *OCA2* expression than variants within the gene (Duffy et al. 2007). This points to regulatory regions as interesting to investigate for normal human pigmentation.

Additionally, the gene has been associated with normal skin pigment phenotypes in several African populations (Crawford et al. 2017). In RNA sequencing (RNA-seq) of melanocyte cultures from individuals of different ancestries, *OCA2* expression was increased in individuals of African ancestry. This suggests that *OCA2* expression correlates with darker pigmentation in many African individuals (Crawford et al. 2017). Several variants in the *OCA2* gene region that are in linkage disequilibrium were also correlated with increased expression of *OCA2* as well as darker pigmentation of the skin in the African individuals. The variant that was most significantly associated with normal differential skin pigmentation was rs1800404. This variant

is located in exon 10 of *OCA2* and has been previously associated with eye colour in European ancestry individuals (Eriksson et al. 2010; Crawford et al. 2017).

### 1.3.1 Abnormal pigmentation phenotypes due to abnormal *OCA2* expression

There are several pigmentation disorders which are linked to the presence of mutations in the different pigment genes.

With regards *OCA2*, the level of transcription has been previously demonstrated to be important in the phenotype of some pigment disorders. Duplications and deletions of the *OCA2* gene or the region containing *OCA2* have been shown to change the amount of melanin produced and result in abnormal pigmentation phenotypes.

Prader-Willi Syndrome (PWS) and Angelman Syndrome (AS) can be caused by large heterozygous deletions of the chromosome 15q11-q13 region which can include the *OCA2* gene (Nicholls 1993; Spritz et al. 1997). These individuals are hemizygous for *OCA2* and commonly present with hypopigmented features (Spritz et al. 1997) such as lighter skin and blonde hair when compared to their normally pigmented family members (Figure 1.6). Moreover, there have been documented cases of PWS and AS individuals from African, Chinese and European ancestries who are also affected with oculocutaneous albinism (Creel et al. 1986; Wallis and Beighton 1989; Fryburg et al. 1991). Thus, the hypopigmented phenotype of AS and PWS has been suggested to result from reduced expression of the remaining functional copy the *OCA2* gene in hemizygous individuals (Spritz et al. 1997).

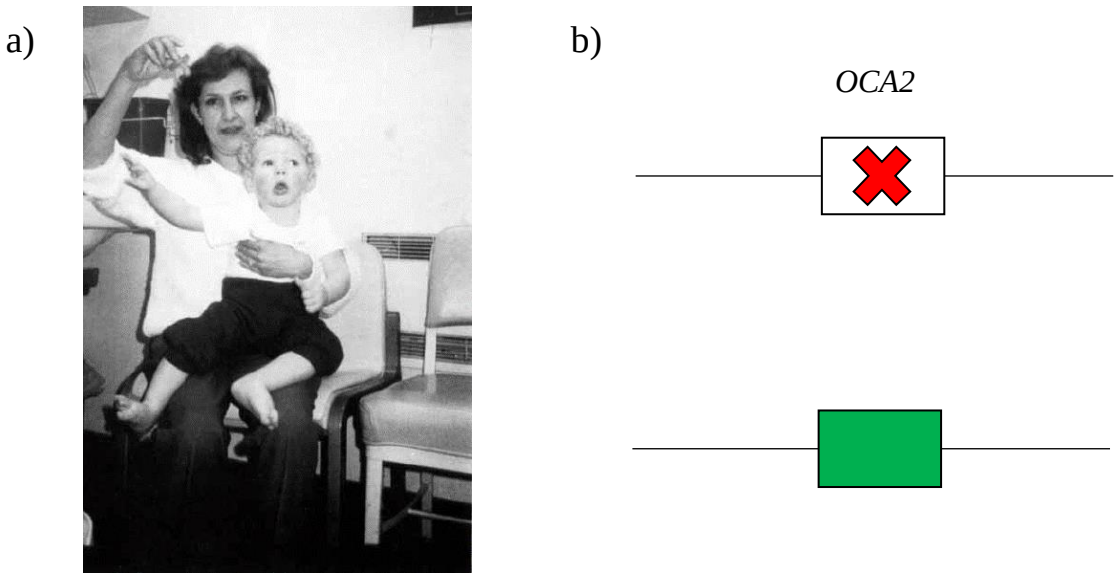



**Figure 1.6: The hypopigmentation phenotype is related to *OCA2* copy number in Angelman and Prader-Willi Syndromes.** a) A male Angelman Syndrome individual who is hypopigmented when compared to his normally pigmented mother (Fryburg et al. 1991). b) A schematic representation of the single copy of the *OCA2* gene in individuals with deletion-type Angelman or Prader-Willi Syndromes who have heterozygous deletions of the gene region and are thus hemizygous for *OCA2*.

Conversely, duplication of the *OCA2* region results in extra copies of the gene and hyperpigmented phenotypes (Akahoshi et al. 2001). A woman with interstitial chromosomal duplication of 15q11.2±q14 presented with a generalised hyperpigmented phenotype (Figure 1.7). Her cells were trisomic for *OCA2* and this was speculated to increase expression of *OCA2* in her skin (Akahoshi et al. 2001).

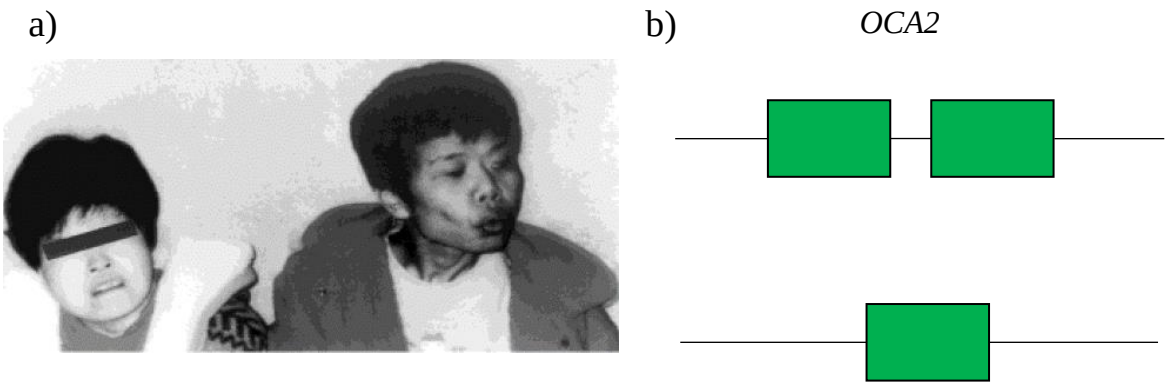



**Figure 1.7: The pigment phenotype and *OCA2* copy number in a case of hyperpigmentation.** a) A normally pigmented individual (left) alongside a hyperpigmented female (right) (Akahoshi et al. 2001). b) A schematic representation of the copy number of the *OCA2* gene in the hyperpigmented individual.

Furthermore, two patients have been described who exhibited pigmentary dysplasia for hyperpigmentation and were mosaic for an additional chromosome 15. In both cases, there were two normal copies of chromosome 15 alongside a copy of chromosome 15 which was quadruple for the *OCA2* region (Figure 1.8). Therefore, these individuals were hexasomic for *OCA2* (Akahoshi et al. 2004; Kraoua et al. 2011). A patient presented with pigmentary dysplasia of normally pigmented skin surrounded by hyperpigmented areas (Akahoshi et al. 2004), while the second patient had pigmentary dysplasia of hyperpigmented and hypopigmented patches (Kraoua et al. 2011). In both cases, the pigmentary dysplasia was thought to be the outcome of the abnormal *OCA2* copy number; where increased copy number led to increased pigmentation, decreased copy number resulted in decreased pigmentation and normal copy number was associated with normal pigmentation (Akahoshi et al. 2004; Kraoua et al. 2011).

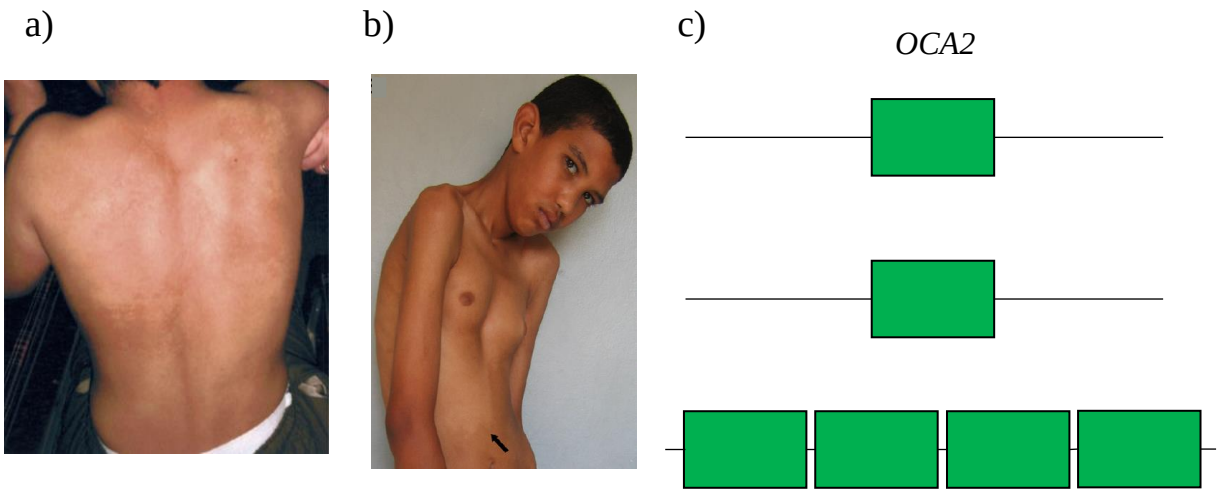



**Figure 1.8: The pigment phenotypes and *OCA2* copy number in two cases of pigmentary dysplasia.** a) A male with pigmentary dysplasia of normally pigmented skin surrounded by hyperpigmented areas (Akahoshi et al. 2004). b) A male with pigmentary dysplasia (Kraoua et al. 2011). The arrow indicates a region of pigmentary dysplasia. c) A schematic representation of the copy number of the *OCA2* gene in these individuals.

Another patient exhibited diffuse hypo- and hyperpigmentation on her trunk, as well as linear hypopigmentation on her legs (Qumsiyeh et al. 2003). Her cells had two additional copies of chromosome 15, each of which had a duplicated PW critical region, in addition to the normal number of chromosomes (Qumsiyeh et al. 2003). Therefore, her cells were hexasomic for *OCA2* (Figure 1.9) which is hypothesised to correlate with the hyperpigmentation of the patient's skin (Qumsiyeh et al. 2003).

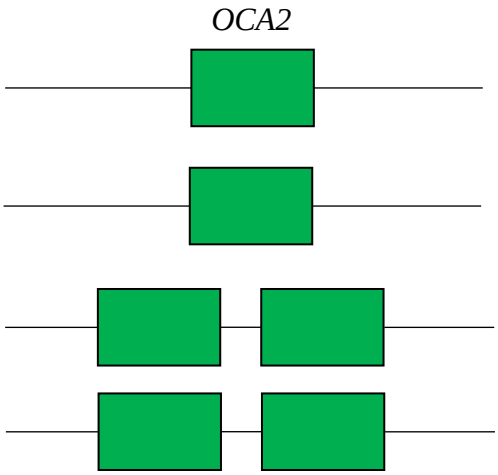



**Figure 1.9: A schematic representation of the copy number of the *OCA2* gene in the female with hyper- and hypopigmentation (Qumsiyeh et al. 2003).**

There is a knowledge gap in the effect of the *OCA2* gene expression on the individual pigmentation phenotype. The previously mentioned studies have indicated that abnormal gene dosage of *OCA2* seems to have direct phenotypic consequences on the level of pigmentation. An increase in the copy number of *OCA2* correlates with increased pigmentation while a decrease in copy number is associated with a decreased pigment phenotype. These cases of abnormal *OCA2* copy number do not demonstrate the effect of differential *OCA2* expression under normal circumstances, which may contribute to the range of normal pigment phenotypes. Instead, they point towards studying the normal expression of the gene in varying skin tones to determine if differences in *OCA2* expression is correlated with different skin colours.

## Albinism

Albinism describes of a group of heterogeneous inherited disorders of pigmentation, which are characterised by hypopigmentation and impaired vision. These disorders represent the extreme phenotype caused by aberrant pigment biosynthesis. There are several types of albinism, which differ in their presentation and which genes are affected. The mutated genes occur across the genome (Table 1.2). Albinism can be broadly divided into ocular and oculocutaneous albinism. Ocular albinism (OA) affects the pigmentation of the eyes only. Oculocutaneous albinism (OCA) results in hypopigmentation of the eyes, skin and hair. There are many different mutations which result in the same albinism phenotype and the different types of OCA present slightly differently to each other. OCA can also present as a feature of another syndrome, which is termed syndromic albinism, as seen in Chediak-Higashi syndrome (CHS) and Hermansky-

Pudlak syndrome (HPS). CHS is caused by mutations in the *LYST* gene (Dufourcq-Lagelouse et al. 1999). HPS has several subtypes that associate with specific genes, such as HPS1 which is linked to mutations in the *HPS1* gene (Oh et al. 1998).

**Table 1.2: The types of albinism and their associated genes (adapted from Kromberg and Manga, 2018).**

| | **Disorder** | **Gene** | **Chromosome Location** | **OMIM number** | **Reference** |
|---|---|---|---|---|---|
| Classic oculocutaneous albinism | OCA1 | *TYR* | 11q14.3 | 203100 | (Tomita et al. 1989; Giebel et al. 1990) |
| | OCA2 | *OCA2* | 15q11.2-q12 | 203200 | (Rinchik et al. 1993) |
| | OCA3 | *TYRP1* | 9p23 | 203290 | (Boissy et al. 1996; Manga et al. 1997) |
| | OCA4 | *SLC45A2* | 5p13.3 | 606574 | (Newton et al. 2001) |
| | OCA5 | *Unknown* | 4q24 | 615312 | (Kausar et al. 2013) |
| | OCA6 | *SLC24A5* | 15q21.1 | 113750 | (Wei et al. 2013) |
| | OCA7 | *C10orf11* | 10q22.2-q22.3 | 615179 | (Grønskov et al. 2013) |
| Syndromic albinism | CHS1 | *LYST* | 1q42.1-q42.2 | 214500 | (Dufourcq-Lagelouse et al. 1999) |
| | HPS1 | *HPS1* | 10q23.1-q23.3 | 203300 | (Oh et al. 1998) |
| Ocular albinism | OA1 | *GPR143* | Xp22.3 | 300500 | (Bassi et al. 1995) |

OCA1 - oculocutaneous albinism type I, OCA2 - oculocutaneous albinism type II, OCA3 - oculocutaneous albinism type III, OCA4 - oculocutaneous albinism type IV, OCA5 - oculocutaneous albinism type V, OCA6 - oculocutaneous albinism type VI, CHS1 - Chediak-Higashi syndrome type 1, HPS1 - Hermansky-Pudlak syndrome type 1, OA1- ocular albinism type 1. *TYR* - tyrosinase, *OCA2* - oculocutaneous albinism type II, *TYRP1* - tyrosinase related protein 1, *SLC45A2* - solute carrier family 45 member 2, *SLC24A5* - solute carrier family 24 member 5, *C10orf11* - leucine rich melanocyte differentiation associated, *LYST* - lysosomal trafficking regulator, *HPS1* - Hermansky-Pudlak syndrome 1, *GPR143* - G-protein coupled receptor 143.

## Oculocutaneous albinism type II (OCA2)

Oculocutaneous albinism type II (OCA2, MIM number: 203200) is the most prevalent form of albinism worldwide and the most common form of albinism in individuals of African ancestry

(Witkop et al. 1989). The prevalence of OCA2 is more frequent in African populations than in other populations (Table 1.3).

**Table 1.3: The prevalence of oculocutaneous albinism type II (OCA2) in various populations.**

| Population | Prevalence | Reference |
|---|---|---|
| Ibo, Nigeria | 1 in 1100 | (Okoro 1975) |
| Tanzania | 1 in 1400 | (Luande et al. 1985) |
| South African Black | 1 in 3900 | (Kromberg and Jenkins 1982) |
| African American | 1 in 15000 | (Witkop et al. 1972) |
| European ancestry, North America | 1 in 37000 | (Witkop et al. 1972) |

OCA2 is caused by mutations in the *OCA2* gene (Rinchik et al. 1993). The condition is characterised by hypopigmentation of the skin, eyes and hair owing to abnormal, insufficient melanin production (Figure 1.10). Eumelanin is virtually absent, however, OCA2 individuals can synthesise small amounts of pheomelanin, which is typically acquired in childhood and accumulates with age (Ramsay et al. 1992; Lee et al. 1994). Therefore, OCA2 individuals are at increased risk of acquiring skin cancer, due to the lack of photoprotective eumelanin (Kromberg et al. 1989; Kobayashi et al. 1993).

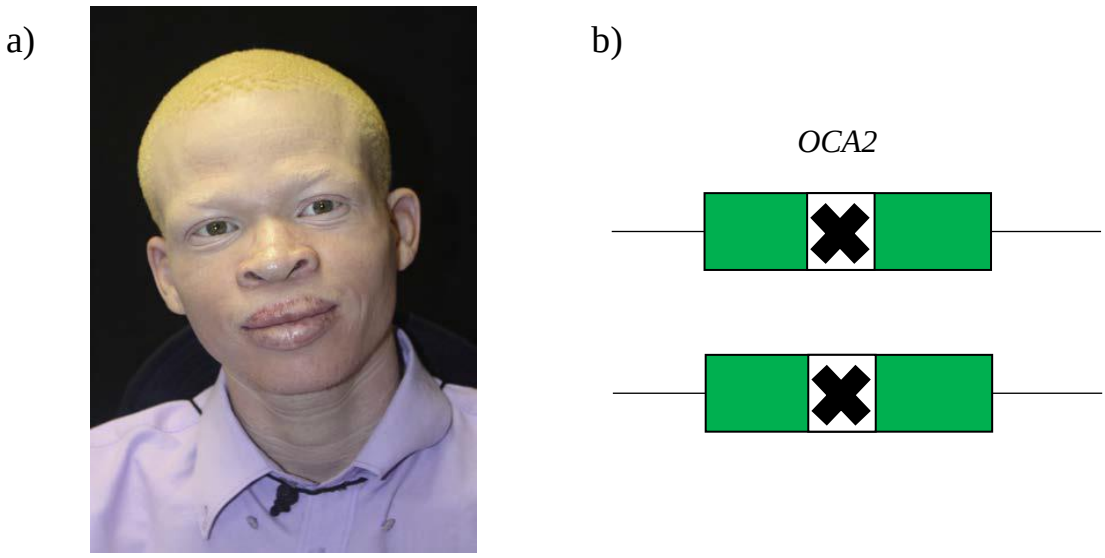



**Figure 1.10: The clinical phenotype of oculocutaneous albinism type II in individuals of African ancestry.** a) A male with classic OCA2 with creamy white skin, white/blonde hair, blue-grey/brown irides (Kromberg and Manga, 2018). b) A schematic representation of the homozygous 2.7 kb deletion of the *OCA2* gene.

A 2.7 kb intragenic deletion involving exon 7 of the *OCA2* gene has been commonly described in African OCA2 individuals (Durham-Pierre et al. 1994). This frameshift mutation results in a truncated, non-functional OCA2 gene product (Durham-Pierre et al. 1994). In African OCA2 individuals, this mutation accounted for 78% of *OCA2* mutations in Southern Africans and 77% in Tanzanians (Stevens et al. 1995; Spritz et al. 1995). Several other point mutations in the *OCA2* gene have been documented as causal for OCA2 (Oetting 2009) but no second common mutation has been described yet, in any population group.

**Brown oculocutaneous albinism (BOCA)**

Brown oculocutaneous albinism (BOCA, MIM number: 203200) is classified as a subtype of OCA2, as it was linked to the *OCA2* gene locus and has similar but milder features (Manga et al. 2001). This is an interesting phenotype of albinism. BOCA has mainly been described in individuals of African ancestry (King and Rich 1986). It has been observed that BOCA individuals have slightly darker overall pigmentation when compared to OCA2 affected individuals, as they have decreased but not abolished melanin production. However, BOCA individuals typically are hypopigmented when compared to a normally pigmented individual as seen in Figure 1.11 (King et al. 1980; Manga et al. 2001).

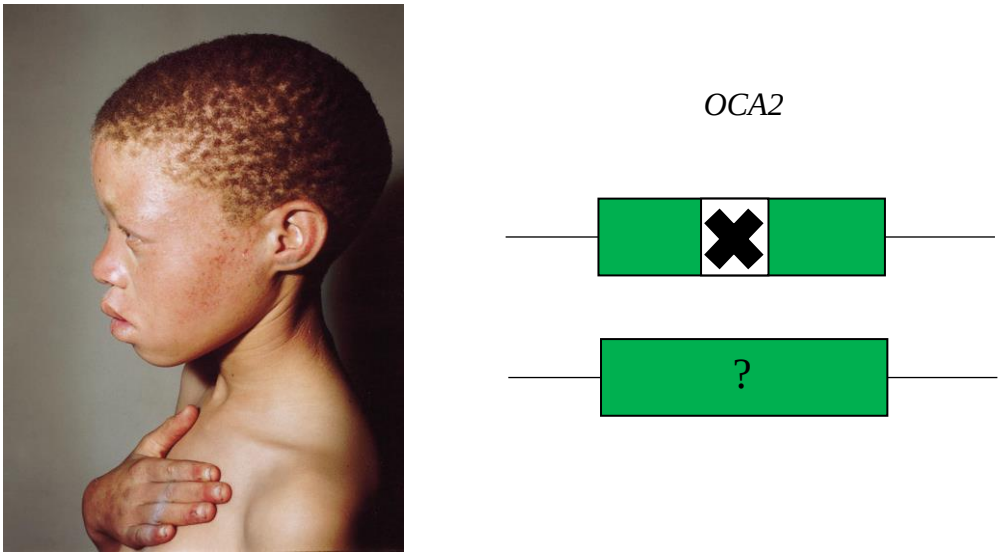



**Figure 1.11: Clinical phenotype of brown oculocutaneous albinism in individuals of African ancestry.** a) A male with BOCA with some level of pigmentation (Manga et al., 2001). b) A schematic representation of the heterozygous 2.7 kb deletion of the *OCA2* gene while the other mutation is unknown.

Since BOCA was linked to the *OCA2* gene locus (Manga et al., 2001), affected individuals were tested for the common Black OCA2 2.7 kb deletion mutation: 9/10 BOCA individuals were found to be heterozygous for the 2.7 kb deletion (Manga et al. 2001). Therefore, in addition to the 2.7 kb deletion, there is a second unknown mutation in all individuals which is causing the BOCA phenotype which has not yet been identified (Kerr et al. 2000).

The *OCA2* gene has long been of interest in the study of both normal and abnormal pigment phenotypes. It plays a role in normal pigment and can disrupt this when mutated, though the protein does not have a direct role in melanin synthesis pathway. The variation in the range of normal pigment phenotypes could be due to differential expression of *OCA2*, which would require investigation of the regulatory regions that are associated with the gene.

## 1.4 Gene regulation and characteristics of regulatory regions

Gene expression is controlled to be specific for tissue type and timing during development (Maston et al. 2006; Ong and Corces 2011).

Most protein coding genes are transcribed by RNA polymerase II (POL2) and the initiation of transcription is tightly controlled. The initiation of transcription occurs through the cooperation

of the promoter and enhancer regulatory regions which are instrumental in regulating the expression of coding genes as seen in Figure 1.12 (Maston et al. 2006).

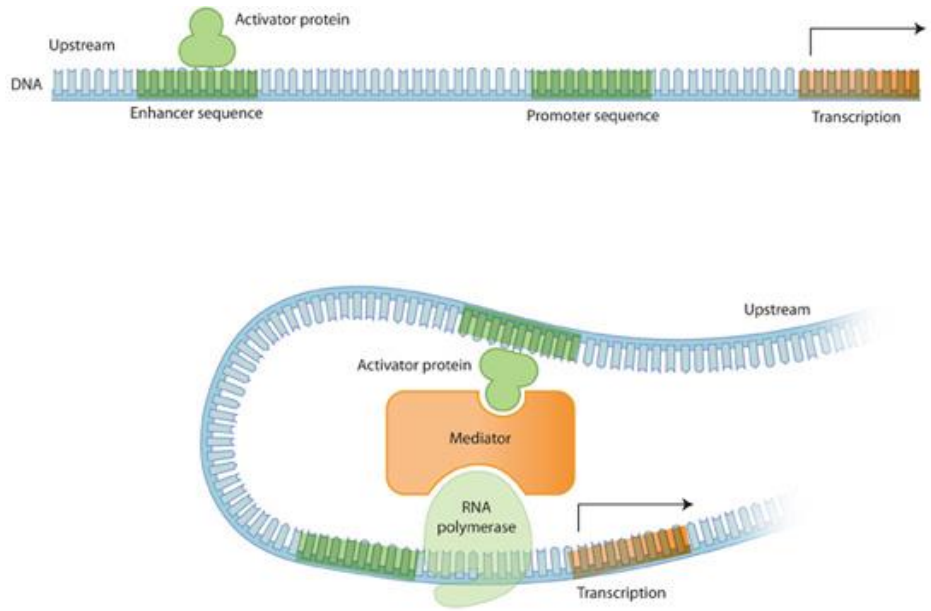



**Figure 1.12: The regulatory regions associated with genes in the human genome.** Most genes require the action of promoter and enhancer regions, which interact by chromatin looping. The promoter typically occurs immediately upstream of the transcription start site while the enhancer may be located further from the gene (O'Connor and Adams 2010).

### 1.4.1 Promoters

The promoter is a regulatory region that occurs immediately upstream of the gene and contains the transcription start site (TSS), which defines the direction of transcription for that gene. The promoter binds transcriptional machinery to initiate transcription which includes several transcription factors (TFs), which can be divided into general TFs and sequence specific activator proteins. Promoters contain consensus sequences for these transcription factors and the promoter is divided up into two main regions: the core promoter and the proximal promoter. The core promoter is where the general TFs bind and the initiation complex forms, which is close to the TSS. The proximal promoter is further upstream of the TSS but usually only a few hundred bases more than the core promoter. It contains consensus sequences for the activator proteins (Maston et al. 2006). These are general properties of promoters but promoters have been shown to vary widely in their composition and their tissue specificity (Mora et al. 2016).

TFs act in combination to enhance or repress gene expression and this adds another layer of control to the process (Maston et al. 2006). Once the TFs have bound to the promoter in a sequential fashion, they recruit POL2 to the region, therefore the transcription initiation machinery has assembled and transcription can proceed (Maston et al. 2006; Ong and Corces 2011).

### 1.4.2 Enhancers

Enhancers are another class of regulatory region which act to increase transcription of genes with which they interact. Enhancers contain binding sites for multiple TFs which usually cluster and the TFs cooperate to upregulate transcription (Ong and Corces 2011). The enhancer region occurs at a distance from the gene it regulates, whether it is in a neighbouring gene or on an entirely different chromosome (Ong and Corces 2011). The position of the enhancer relative to the gene it regulates is not fixed as the enhancer can occur upstream or downstream of the gene of interest (Maston et al. 2006). Enhancers are highly tissue specific and generally are only active in a small number of tissue types (Zhu et al. 2013).

Many genes require the involvement of an enhancer in transcription initiation before they can be expressed (Ong and Corces 2011). The promoter and enhancer mutually bind to activating TFs and the transcription machinery. In doing do, the chromatin forms a long-range loop which bypasses the intervening sequence (Figure 1.12). The two regions are thereby able to associate despite their relative distance from each other at the sequence level (Maston et al. 2006; Ong and Corces 2011). It has also been suggested that the initiation complex forms at the enhancer and thereafter interacts with the promoter to recruit POL2. Moreover, enhancers are modular as one promoter can interact with several enhancers at different times, in different tissues or in response to different stimuli (Maston et al. 2006).

### 1.4.3 Histone and chromatin modification of regulatory regions

The level of gene expression can be correlated with epigenetic modifications present in the gene region. The epigenetic modifications can be associated with accessibility of the chromatin or post-translational histone modifications as markers of gene regions; which have characteristic histone modification signatures and specific combinations of histone marks have functional relevance (Harmston and Lenhard 2013). These marks are established during development and later influence the transcription of that gene by interacting with TFs or with chromatin modifying enzymes to change the chromatin state (Ong and Corces 2011). These epigenetic

signals can be utilised to predict if a region of interest is a regulatory region and if so, which one.

Chromatin state is correlated with accessibility of the chromatin to the transcription initiation machinery; subsequently open chromatin is conducive for gene expression but closed chromatin is not. Two indicators of chromatin state include DNase hypersensitivity sites and trimethylation of lysine 27 of histone 3 (H3K27me3). DNase I binds to and cleaves DNA at sites, termed DNase hypersensitivity sites, where chromatin is in an open configuration. Open chromatin is typically associated with regulatory regions, such as promoters and enhancers, and genes that are being transcribed (Gross and Garrard 1988). Contrastingly, H3K27me3 is a repressive histone modification mark associated with closed chromatin, which suppresses transcription by inhibiting the activation of POL2 (Ferrari et al. 2014). These chromatin state indicators can be used to identify putative regulatory regions as they would be more likely to occur within regions of open chromatin marked by DNase hypersensitivity than H3K27me3.

Regulatory regions are defined by specific chromatin features. Promoters are defined by trimethylation of lysine 4 on histone 3 (H3K4me3) (Harmston and Lenhard 2013). They can be enriched for H3K27me3 and simultaneously depleted for H3K4me3 when they are inactive. However, should both of these marks occur in the same promoter, this could be an indication of a bivalent promoter which is repressed but is poised to be activated when required (Harmston and Lenhard 2013). Nucleosomes associated with enhancers are often not densely compacted and contain the H2A.Z and H3.3 histone variants which contribute to nucleosome plasticity. This allows for TFs and histone modifying enzymes to be recruited to the enhancer region (Ong and Corces 2011).

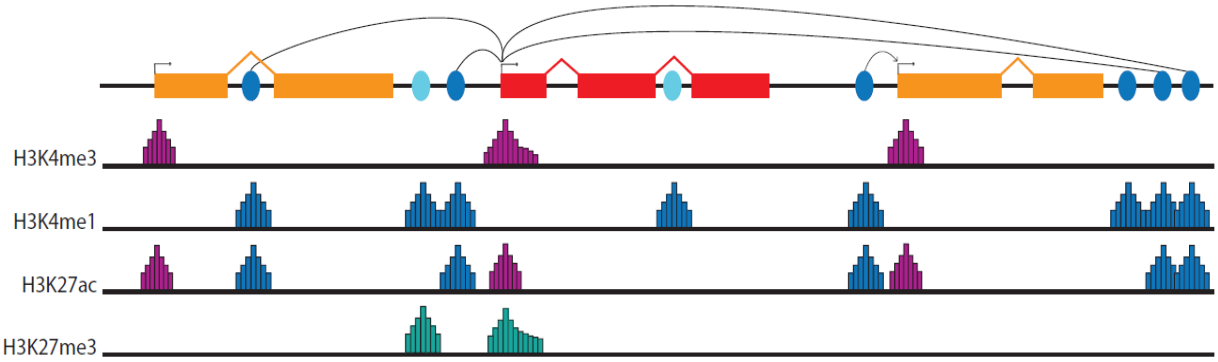



**Figure 1.13: The histone modifications that are typically associated with gene regulatory regions**. The various possible interactions between the enhancers and gene promoters are indicated. Promoters are enriched for H3K4me3, while active enhancers are associated with H3K4me1 and H3K27ac marks. Poised enhancers are marked by H3K4me1 and H3K27me3. The target gene is red in colour while adjacent genes are represented as orange rectangles. Enhancers are represented as blue ovals and poised enhancers are a lighter blue. Adapted from (Harmston and Lenhard 2013).

Additionally, histones within enhancer regions are enriched for specific modifications of the histone tails. These activating modifications are strongly correlated with gene expression and are specific to cell type (Ong and Corces 2011). Enhancers are enriched for monomethylation at lysine 4 on histone 3 (H3K4me1), dimethylation of lysine 4 of histone 3 (H3K4me2) and acetylation of lysine 27 of histone 3 (H3K27ac) as seen in Figure 1.13. Poised enhancers, which are currently inactive but can be activated at a later stage, are marked by H3K4me1 and H3K27me3, with the absence of H3K27ac. Replacing H3K27me3 with H3K27ac to is thought to occur during the activation of enhancers so H3K27ac is an indication of active enhancers (Ong and Corces 2012; Harmston and Lenhard 2013). Therefore, H3K4me1 and H3K27ac, coupled with DNase hypersensitivity and a lack of H3K27me3, are informative markers for identifying enhancer regions when interrogating epigenetic modification of gene regions.

### 1.4.4 Techniques to identify regulatory regions

Identifying regulatory regions in the human genome has been emphasised to understand their impact on gene expression within normal tissues as well as how changes in expression can affect disease phenotypes. While many techniques have been developed to characterise regulatory regions, specific attention will be paid to those that have been used to annotate regulatory regions in large scale projects such as the Encyclopedia of DNA Elements (ENCODE) and Roadmap Epigenomics Projects.

## Chromatin interaction

In order to determine which regions of the genome interact, it is important to determine the three-dimensional folding of chromatin first. This is accomplished by chromatin conformation capture (3C) techniques, which crosslink genome regions that are in close proximity to a single locus (Dekker et al. 2002). Once these regions have been identified, this will indicate which genome regions are close enough to interact with each other (Dekker et al. 2013). Since there are limitations on this method, it has been further developed for specific applications. Chromatin conformation capture-on-chip (4C) can generate indications of the genome-wide interactions of a single locus (Simonis et al. 2006) while Hi-C demonstrates an unbiased view of all the regions across the genome that interact with every other region (Lieberman-Aiden et al. 2009). Another technique, Chromatin Interaction Analysis by Paired-End Tag (ChIA-PET) determines which genome-wide chromatin regions interact when they are mediated with a protein of interest (Fullwood et al. 2009). These methods are depicted in Figure 1.14. Since these methods are based on finding regions that interact, it is possible to use them to find interactions of promoters and enhancers which are expected to interact during transcription. This can also apply to finding an enhancer region for a gene which has an already characterised promoter.

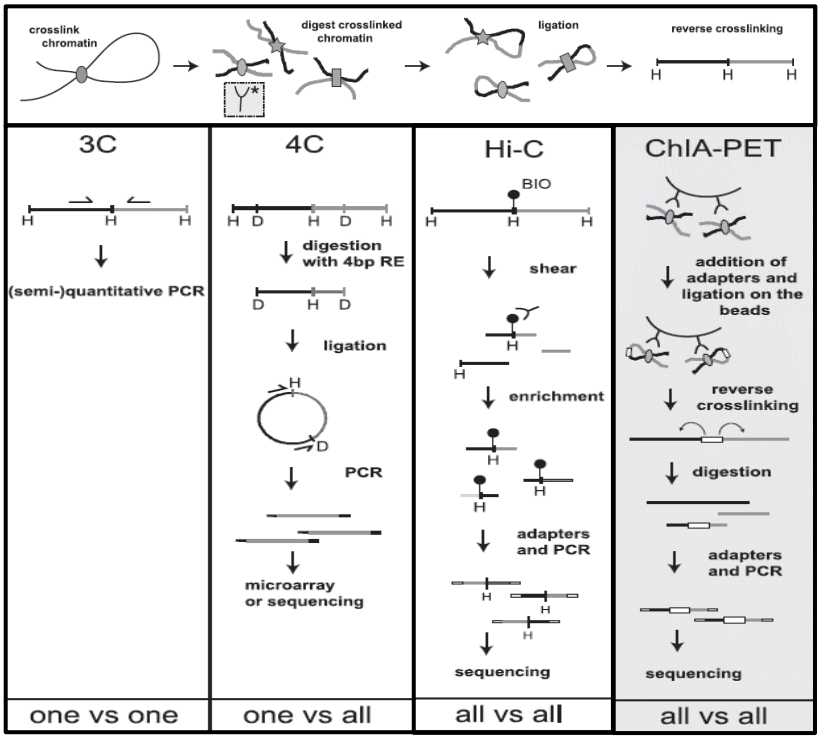



**Figure 1.14: Techniques that are used to investigate chromatin interaction.** The depicted techniques include chromatin conformation capture (3C), chromatin conformation capture-on-chip (4C), Hi-C and Chromatin Interaction Analysis by Paired-End Tag (ChIA-PET). The common steps between the techniques are indicated as well as the technique specific steps (adapted from Wit and Laat 2012)**.**

## Chromatin accessibility

Signatures of open chromatin typically coincide with regulatory regions and can be used to locate these regions. DNase I hypersensitive site sequencing (DNase-seq) can be used to identify DNase hypersensitivity sites which co-localise with a regulatory region of interest (Boyle et al. 2008). DNase-seq digests the genome with DNase I and then sequences the resulting fragments by high-throughput sequencing methods. This will determine the location of the hypersensitivity sites when those reads are aligned to the human genome reference sequence (Boyle et al. 2008). If a region of interest aligns with a DNase hypersensitivity site, it could have a regulatory function. Moreover, DNase hypersensitivity signals are higher at the centre and lower towards the ends of a regulatory region. This can be used as a contributing factor to identify regulatory regions like promoters and enhancers (Thurman et al. 2012).

**Histone modifications of regulatory regions**

These methods should be combined with a way of finding epigenetic signals that are characteristic of regulatory regions such as chromatin immunoprecipitation sequencing (ChIP-seq). This technique identifies regions of DNA that bind a specific protein of interest. The protein and DNA are crosslinked, the DNA is fragmented and antibodies specific to that protein are used to precipitate out the DNA-protein complex. The DNA is then isolated and sequenced to identify the genome-wide locations of binding sites for that protein (Barski et al. 2007; Johnson et al. 2007). For histone modifications that are associated with regulatory regions, it is possible to use antibodies that are specific for histones with those modifications (Bernstein et al. 2010). When considering DNase hypersensitivity and histone modification patterns together, a region of interest which has DNase hypersensitivity regions that are flanked by regulatory region specific histone modifications may have regulatory properties (Thurman et al. 2012).

A combination of data from these methods could be used to search for regulatory regions in the genome without having a prior idea of where the regulatory regions may occur. Additionally, they could be used to determine if a region of interest is a promoter or enhancer, if there is prior evidence to suggest that the region has a regulatory function, such as the case of searching for possible enhancers for the *OCA2* gene.

## 1.5 A putative enhancer for the *OCA2* gene

The rs12913832 SNP correlates strongly with skin, hair and eye colour (Sturm et al. 2008; Branicki et al. 2009). It is located within intron 86 of the Hect Domain and RCC1-like Domain 2 (*HERC2*) gene, which lies 12 kb upstream of *OCA2* (Figure 1.15). *HERC2* has a normal role in the DNA damage repair response by assisting the assembly of DNA repair proteins (Bekker-Jensen et al. 2010). The protein can act as a E3 ubiquitin ligase to target BRCA1 for degradation and activate other E3 ubiquitin ligases (Wu et al. 2010; Kühnle et al. 2011). Since rs12913832 seemed to correlate with pigment phenotypes but occurred in a gene which did not have pigment related functions, it was speculated that the region surrounding the SNP was an enhancer (Sturm et al. 2008).

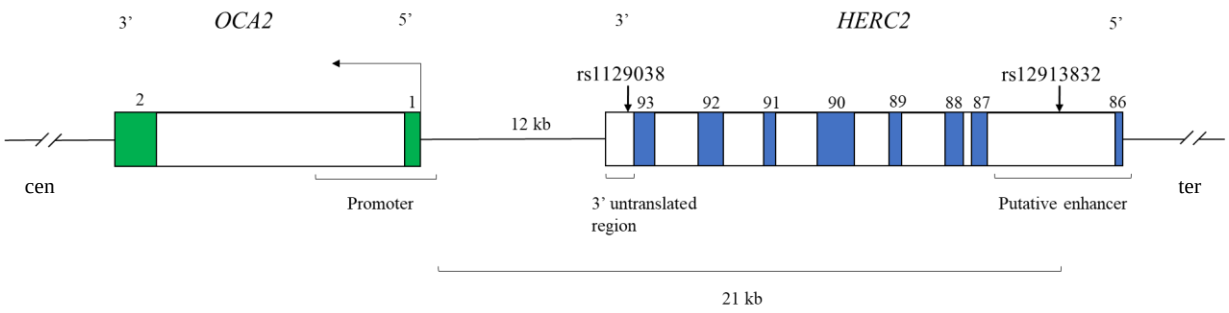



**Figure 1.15: The relative positions of the *HERC2* and *OCA2* genes on chromosome 15.** The 3' end of *HERC2* and the 5' end of *OCA2* are depicted. Both genes are transcribed off the reverse strand. The coloured blocks represent exons of the genes. The positions of the rs12913832 and rs1129038 SNPs, which are in complete linkage, are indicated. rs12913832 is located within intron 86 of the *HERC2* gene.

To confirm this speculation, Visser and colleagues characterised this region for enhancer-like properties in human epidermal melanocytes from darkly and lightly pigmented individuals (Visser et al. 2012). The MCF7 breast cancer cell line was utilised as a control in all tests. A formaldehyde-assisted identification of regulatory elements (FAIRE) test was performed to indicate if the region contained regulatory DNA based on differences in the ability of chromatin to cross-link between regulatory DNA regions and the bulk of the chromatin (Giresi et al. 2007). The FAIRE results indicated that the rs12913832 *HERC2* region had open chromatin structure in dark and light pigmented melanocyte cell lines, but there was closed chromatin structure in the same region in MCF7 (Visser et al. 2012). By CHIP-qPCR signs of active chromatin - including acetylated histone H3, H3K4me1 and H3K27ac - were detected in the region in the melanocytes and these signals were absent in MCF7 (Visser et al. 2012). Therefore, the region could be considered an enhancer. The region has also been confirmed as specific for activity in melanocytes by a luciferase reporter assay. There was no increased luciferase activity when the enhancer vector was transfected into HEK293 kidney tissue in comparison to an empty vector. However, in G361 melanoma cells the luciferase activity was increased. Moreover, this region is highly conserved in many animal species. Thus, these results suggest that the region surrounding rs12913832 is a melanocyte specific enhancer (Visser et al. 2012).

### 1.5.1 The function of the putative *OCA2* enhancer

This enhancer has primarily been studied for the effect of the rs12913832 alleles (A/G on the forward strand) on pigmentation. Darkly pigmented melanocytes which carried the rs12913832 A allele were associated with increased transcription of *OCA2* (Cook et al. 2009; Visser et al.

2012). The HLTF, LEF1, and MITF transcription factors bound to the region and assisted the formation of a long-range chromatin loop that linked the *OCA2* promoter to this enhancer region. Additionally, the darkly pigmented cell line had higher levels of the signatures of enhancers (Visser et al. 2012). Therefore, the A allele is associated with increased melanin production and overall pigmentation (Cook et al. 2009; Branicki et al. 2009). Conversely, lightly pigmented melanocytes with the G allele were associated with decreased transcription factor binding, loop formation and expression of *OCA2* (Cook et al. 2009; Visser et al. 2012). Thus, the G allele is associated with lower melanin production and overall pigmentation (Cook et al. 2009). Moreover, the melanin content of A/G heterozygotes was intermediate between A/A and G/G genotyped melanocytes (Cook et al. 2009). Therefore, transcriptional control of the *OCA2* gene results in differing levels of pigmentation in normal melanocytes. This is illustrated in Figure 1.16.

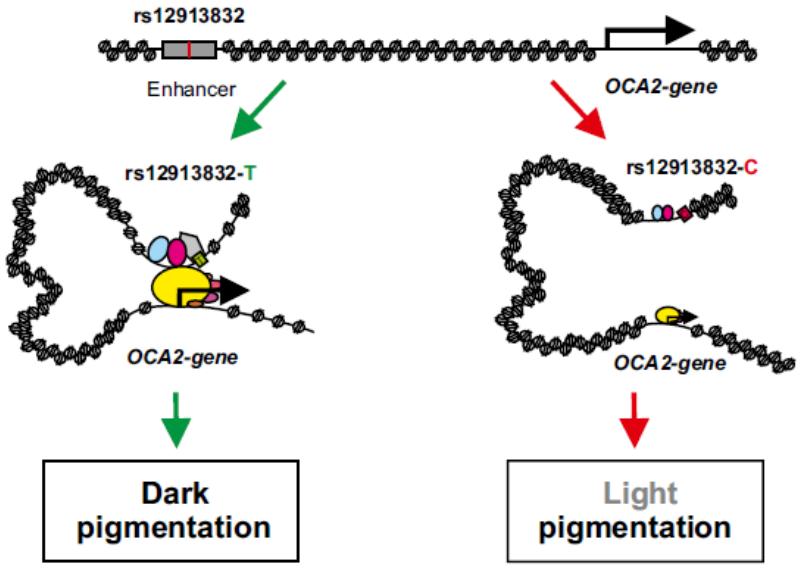



**Figure 1.16: Model of chromatin looping to facilitate the interactions between the *OCA2* promoter and the putative enhancer to generate different levels of pigmentation.** The grey pentagon represents HLTF, the pink oval represents MITF, the blue oval is LEF1 and the yellow oval represents the transcription machinery. The orientation of the chromatin features has been reversed and the depicted rs12913832 alleles are from the reverse strand (Visser et al. 2014).

Moreover, the genotype of this SNP affects eye colour in European populations where the A allele is associated with brown eyes and the G allele with blue eyes (Sturm et al. 2008; Branicki et al. 2009). This genotype alone has been indicated to explain 68.8% of normal eye colour

variation in Scandinavian individuals (Andersen et al. 2016). rs12913832 is also relevant to hair pigment where the G allele is frequent in blonde haired individuals (Branicki et al. 2009).

The previously mentioned studies focused on individuals of European ancestry, however, little data is available on the allele frequencies in African populations. The rs1129038 T allele, found within the 3' UTR of *HERC2* as seen in Figure 1.15, is in complete linkage disequilibrium with the rs12913832 G allele (Donnelly et al. 2012). Both variants were thought to occur near an *OCA2* regulatory element (Donnelly et al. 2012). Therefore, these SNPs were investigated as a blue eye colour associated TG haplotype. The haplotype was more frequent in samples of European and south west Asian ancestry but was less common in African populations (Donnelly et al. 2012).

In Table 1.4, the minor allele frequencies of these SNPs in the 1000 Genomes Project are presented. The frequency of alleles associated with darker skin pigmentation are higher in African and Asian populations while the alleles associated with lighter skin occurred at higher frequencies in European populations (The 1000 Genomes Project Consortium 2015).

**Table 1.4: The minor allele frequencies of rs12913832 and rs1129038 in a selection of populations from the 1000 Genomes Project (The 1000 Genomes Project Consortium 2015).**

| | | **rs12913832** | | **rs1129038** | |
|---|---|---|---|---|---|
| **Super Population** | **Population** | **Minor allele** | **Minor allele frequency** | **Minor allele** | **Minor allele frequency** |
| African | Luhya in Webuye, Kenya | G | 0 | T | 0.01 |
| | Yoruba in Ibadan, Nigeria | G | 0 | T | 0 |
| European | Finnish in Finland | A | 0.097 | C | 0.118 |
| | British in England and Scotland | A | 0.176 | C | 0.170 |
| Asian | Han Chinese in Beijing, China | G | 0 | T | 0 |
| | Japanese in Tokyo, Japan | G | 0 | T | 0 |

In a more recent study of Khoe and San individuals, from Nama and ǂKhomani ethnicities respectively, the frequency of the rs12913832 G allele in the Nama was 8% and in the ǂKhomani the frequency was 12% (Martin et al. 2017). Though the Khoe and San are known to have

lighter skin tones, this data still indicates that the rs12913832 G allele occurs at low frequencies in African populations. These data suggest that the rs12913832 A allele is associated with darker pigmentation and indicate a role of the enhancer on pigmentation in African populations, though this has not been well studied.

## 1.6 Tools to explore population level variation and predict functionality of variants

When investigating genetic variation at a population level, it is important to have a large number of samples available, to detect both the common and rare variation in that population. Rare variants will only be picked up with fine enough sequencing resolution and with sufficient individuals to determine that they are truly present.

Publicly available datasets of genetic information are a valuable resource that fulfil this requirement. However, there are a limited number of large datasets that include a variety of African populations. The 1000 Genomes Project (KGP) (The 1000 Genomes Project Consortium 2015) and the African Genome Variation Project (AGVP) (Gurdasani et al. 2015) have generated low coverage whole genome sequences for a large number of African individuals from the West, East and South of Africa. Data from these projects can be used to interrogate various genetic questions related to the African genome. The data generated from these projects can be used to improve associations in genome wide association studies by imputing them into the data to fill in missing genotypes, particularly those at low frequencies (Wood et al. 2013). Therefore, these resources are valuable to look for common variation linked to normal traits in African populations.

Additional publicly available datasets that can be valuable in the interrogation of regulatory regions are ENCODE and the Roadmap Epigenomics project. The purpose of ENCODE was to identify functional regions throughout the human genome, for many cell types which included immortal and normal cell lines. Chromatin interaction data from this project can be used to identify regulatory regions that interact such as promoters and enhancers (ENCODE Project Consortium 2012). The purpose of the Roadmap Epigenomics project was to establish profiles of normal epigenetic signals in many normal cell types (Romanoski et al. 2015). This data can be used to identify regulatory regions by locating DNase hypersensitivity sites that were determined by DNase-seq and the regulatory region specific histone modifications by ChIP-seq (Bernstein et al. 2010). The combination of data from these two projects are highly useful in

identifying regulatory regions which could apply to identifying a putative enhancer for the *OCA2* gene.

Once variation in the region of interest has been identified, it is important to perform functional annotation to determine if the variants are functionally relevant. There are many bioinformatics tools which can be used for this purpose. For coding variation, one may want to determine if the variant changes the structure and therefore the function of the protein of interest. However, for non-coding variation such as the variation in regulatory regions, the amino acid sequence of the protein may be unaffected but the expression of the protein may be changed. This could have several possible consequences. The variant could occur within a transcription factor binding site or a protein binding site and interrupt it when the alternative allele is present; the variant may occur within an expression quantitative trait locus; or the variant may interrupt alternate splicing and lead to intron retention or exon skipping (Mansur et al. 2018).

There are many bioinformatic tools that can be used to predict these functional effects. Major sources of annotations for variants include the Ensembl Variant Effect Predictor (McLaren et al. 2016), RegulomeDB (Boyle et al. 2012), HaploReg (Ward and Kellis 2016), and the Genotype-Tissue Expression project (GTEx Consortium 2015). Additional tools for annotations and prediction of functionality include atSNP (Zuo et al. 2015) and Genome Wide Annotation of Variants (Ritchie et al. 2014). These tools draw from large datasets of information about what sort of features the variant coincides with and this can be used to predict if the variant is functional. Many of these tools have overlapping information and some have been updated more recently than others. Therefore, it is preferable to use an assortment of tools as opposed to a single approach, as if they agree with each other it is stronger evidence. These tools will be discussed in further detail in Chapter 2.

## 1.7 Study rationale

Human pigmentation is a variant phenotype which has a clear genetic association. The *OCA2* gene plays a definite role in hair, skin and eye pigmentation. This is clearly illustrated by the impact that a lack of expression of the *OCA2* gene has on pigmentation (hypopigmentation), as well as the effect of supernumerary copy number (hyperpigmentation). African ancestry individuals are known to have a range of normal pigment phenotypes, particularly with reference to skin tone. This could be affected by the relative level of *OCA2* expression which may contribute to the measure of bodily pigmentation. Therefore, it is hypothesised that altered

transcriptional regulation of the gene could lead to normal pigment variation. Functional variants in regulatory regions can change the level of expression of the gene and this may contribute to a range of normal pigment phenotypes. Therefore, the regulatory regions of *OCA2* should be inspected for common functional variants that could potentially alter the expression of *OCA2*.

The regulatory regions of the gene could also possibly harbour causal mutations which may cause milder forms of albinism, such as BOCA. Before identifying causative mutations in these regions for *OCA2* related albinism, it would be necessary to understand the normal underlying genetic variation which contributes to the spectrum of normal pigment phenotypes. The African genome is highly variable with respect to the reference genome that was built on European sequence data. It is therefore necessary to characterise the normal variation in African genomes so it can serve as a baseline and thereby narrow the search for causal mutations in future research projects looking for abnormal pigment phenotypes.

In addition to an already identified promoter, a possible enhancer for the *OCA2* gene in the neighbouring *HERC2* gene provides another region to investigate for variants that may contribute to normal pigment. The *OCA2* enhancer and promoter can be important as variation in these regions may impact their function and by extension how much melanin is synthesised. Therefore, this variation may give rise to the variety of normal pigmentation phenotypes.

The *OCA2* gene control regions may contain variation that is important for the normal functioning of the gene and determining a range of normal pigment phenotypes. Additionally, improved understanding of normal variation can be informative regarding their function in normal pigment phenotypes in African populations.

## 1.8 Aim and study objectives

**Aim**

The aim of this study was to investigate the normal variation of the *OCA2* gene control regions.

**Objectives**

1. Define the putative enhancer region of *OCA2*.

2. Extract the putative promoter and enhancer regions for *OCA2* from the 1000 Genomes Project (KGP) and African Genome Variation Project (AGVP) variant call files (VCF).
3. Call variant frequencies for KGP and AGVP populations.
4. Perform functional annotation in the region to assess the potential impact of variants using the Variant Effect Predictor (VEP) and other bioinformatic tools.
5. Interpret variants with large frequency differences for particular relevance to potential regulatory effects.

# Chapter 2 - Materials and methods

## 2.1 Description of the KGP and AGVP datasets

The datasets used for this project are sourced from publicly available low coverage (4-8X) African whole genome sequences (WGS) from the 1000 Genomes Project (KGP) and the African Genome Variation Project (AGVP). All coordinates from these sequences are based on the GRCh37 build of the human genome. Low coverage sequencing is suited to identify common variants. Rare variants that occur with an allele frequency less than 0.5% may be sequencing artifacts and may require downstream analysis to differentiate between true and erroneous calls (Navon et al. 2013). Thus, these rare variants would need confirmation prior to actioning. Consequently, this study will focus on common variation only.

The KGP WGS data had an average coverage of 7.4X in all samples. There was a total of 666 African individuals in the KGP data. These individuals were from populations found on the African continent and from populations of African descent which reside outside of the continent. The populations were sampled in the Caribbean, the United States of America, Nigeria, the Gambia, Sierra Leone and Kenya. The African Caribbeans in Barbados (ACB) and Americans of African Ancestry in South West United States of America (ASW) populations are of African ancestry but reside outside of the African continent. They are likely to have admixed with other populations in their locality. Thus, these individuals may have different substructure to the continental Africans and this can be used to indicate the potential effect of admixture to change the allele frequencies of common variation (The 1000 Genomes Project Consortium 2015).

WGS from the AGVP populations for a total of 320 individuals were utilised. These individuals were from the Bagandan, Ethiopian and Zulu populations. For the purposes of this study, the AGVP population names were abbreviated as indicated in Table 2.1. All AGVP populations were sampled on the African continent. The generated sequences had an average coverage of 4X, additionally, the AGVP Ethiopian sequences had higher average coverage of 4-8X (Gurdasani et al. 2015).

The naming convention and sample sizes of the populations from these two datasets are described in Figure 2.1 and Table 2.1 below. Overall, data from a total of 986 individuals were available for this study. The summary of the workflow described in these methods to identify

and annotate functional variation within the regulatory regions of *OCA2* is outlined in Figure 2.2.

**Table 2.1: The names and sample sizes of the 1000 Genomes Project and African Genome Variation Project populations used in this study.**

| Dataset | Population code | Population description | Number of samples |
|---|---|---|---|
| 1000 Genomes Project | ACB | African Caribbeans in Barbados | 96 |
| | ASW | Americans of African Ancestry in South West United States of America | 66 |
| | ESN | Esan in Nigeria | 99 |
| | GWD | Gambian in the Western Divisions in the Gambia | 113 |
| | LWK | Luhya in Webuye, Kenya | 99 |
| | MSL | Mende in Sierra Leone | 85 |
| | YRI | Yoruba in Ibadan, Nigeria | 108 |
| African Genome Variation Project | BAG | Baganda from Uganda | 100 |
| | ZUL | Zulu from South Africa | 100 |
| | ETH | Ethiopian from Ethiopia (Amhara, Oromo, Somali, Wolayta, Gumuz ethnic groups) | 120 |
| | | Total | 986 |

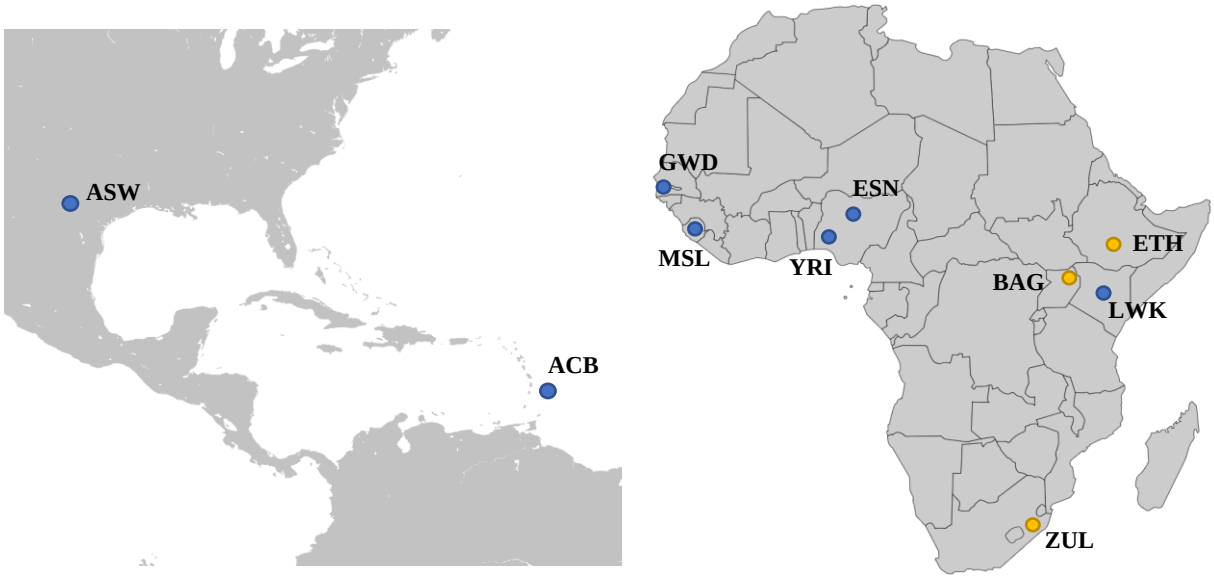



**Figure 2.1: Location of the KGP and AGVP populations.** These populations include the continental and diasporic African ancestry populations. The blue dots represent populations from KGP and the yellow are from AGVP.

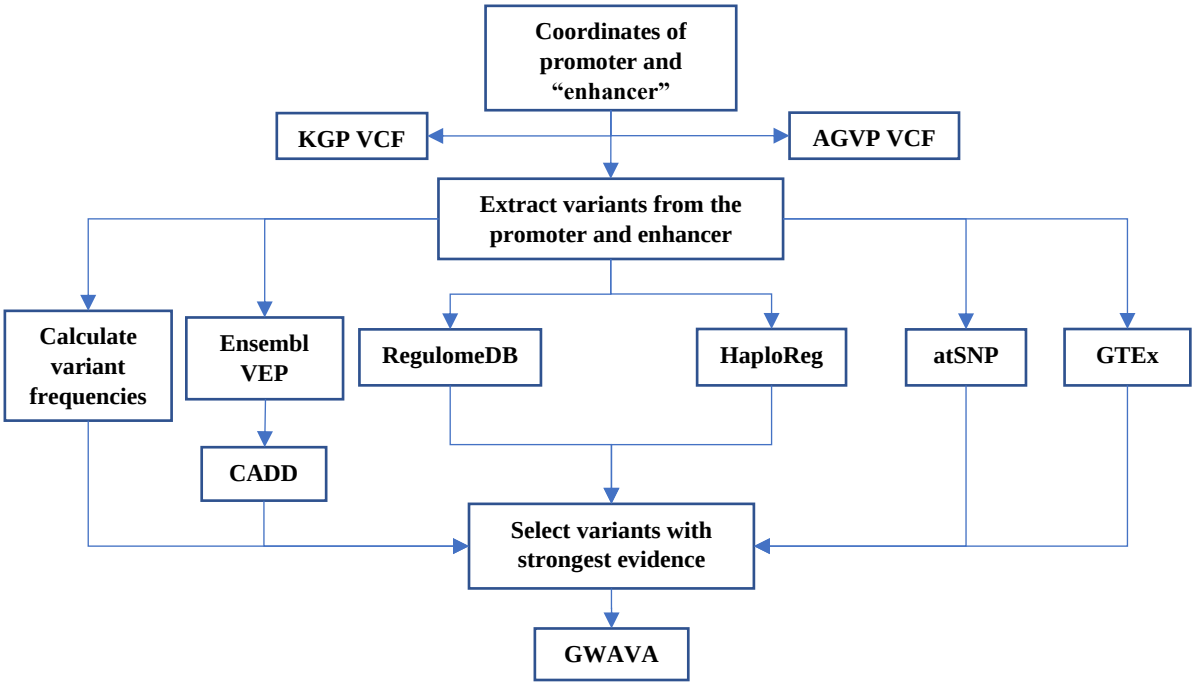



**Figure 2.2: Bioinformatics workflow to extract sequence information from the *OCA2* regulatory regions.** Variants were identified and annotated for potential functionality using several tools.

## 2.2 Identifying the putative promoter and enhancer regions for *OCA2*

The coordinates in this project are based on the GRCh37 build of the human genome. The *OCA2* promoter coordinates are chromosome 15: 28339081 – 28345199. These coordinates had been sourced from Ensembl build 37 (release 96) (Zerbino et al. 2018), where the region had previously been characterised in various cell types by projects such as the Encyclopaedia of DNA Elements and the Roadmap Epigenomics Project. The promoter and its associated flanking regions were included in the definition of the promoter region.

The coordinates for the potential enhancer region within *HERC2* were chosen to be chromosome 15: 28364618 – 28366618. This region lies 1 kb upstream and downstream of the rs12913832 SNP, which has been identified by literature to have an important role in human pigment phenotypes and has been the focus of work to identify if the region surrounding it is an enhancer (Visser et al. 2012). Enhancers are usually 50 bp-1.5 kb in length (as reviewed in Blackwood and Kadonaga 1998), thus if the region of interest was an enhancer, the chosen coordinates for the putative enhancer may overlap with this feature.

### 2.2.1 Determining the properties of the putative enhancer

The Encyclopaedia of DNA Elements (ENCODE) characterised functional regions in the non-coding DNA, in many cell types. This data is publicly available and can be used to identify possible regulatory regions. Chromatin Interaction Analysis by Paired-End Tag data, as accessed on the UCSC genome browser, was investigated for interactions between the *HERC2* and *OCA2* regions to see if the regions which had been defined as promoter and enhancer interacted as expected.

Chromosome conformation capture data was accessed for Hi-C (all interactions vs all interactions), virtual chromosome conformation capture-on-chip (4C - one target region vs all other regions) and Chromatin Interaction Analysis by Paired-End Tag (ChIA-PET – to detect interactions of different regions which mutually bind to a protein of interest, which was RNA polymerase II) for interaction of chromosomal regions between the promoter and potential enhancer. The Hi-C and virtual 4C data was viewed on the 3D Genome Browser (Wang et al. 2018) (http://promoter.bx.psu.edu/hi-c/). As there are no normal melanocyte cell lines in the Hi-C and virtual 4C dataset, available cell lines which were the most similar to melanocytes and originated from the skin were investigated to approximate interactions in melanocytes. The

selected cell lines were RPMI7951, melanocytes from a malignant melanoma cell line (Lajoie et al. 2015), and NHEK which is a normal keratinocyte cell line (Rao et al. 2014).

ChIA-PET data was interrogated in the UCSC genome browser for GRCh37 (http://genome.ucsc.edu/, accessed in November 2019). All of the cell lines from the experiment were investigated for interactions of RNA polymerase II (POL2), the polymerase which binds to the promoter to initiate transcription of a gene (Rosenbloom et al. 2013). Owing to the interaction of promoters and enhancers, POL2 binds to both regulatory regions to initiate transcription and can be used as an indicator of this interaction in chromatin interaction experiments (Li et al. 2012). ChIA-PET was performed on five cancer cell lines: myeloid leukemia from erythrocytes (K562), breast cancer (MCF7), cervix carcinoma (HeLa S3), colon cancer (HCT116) and acute promyelocytic leukemia (NB4) (Li et al. 2012). These cell lines included those with the same developmental lineage as melanocytes (MCF7 and HeLa S3) and others that are of a different developmental lineage to melanocytes (K562, HCT116 and NB4). The presence or absence of interactions between the *OCA2* promoter and enhancer would indicate if these interactions, and the activity of the putative enhancer are specific to melanocytes.

The Roadmap Epigenomics Project investigated normal epigenetic signatures in many cell types in healthy and diseased adult and foetal tissue (Romanoski et al. 2015). The epigenetic data from foetal foreskin melanocytes were accessed to identify if the putative enhancer region had enhancer-like characteristics in melanocytes. Enhancer specific signals include H3K4me1 (signal for all enhancers), H3K27ac (histone signal for active enhancers) and DNase hypersensitivity. Thereafter, data from other cell lines was investigated to determine if the putative enhancer activity was specific to melanocytes. Melanocytes originate from neural crest cells (Mort et al. 2015), therefore, cell lines from the same lineage such as foetal thymus (descends from neural crest cells (Müller et al. 2008)) and foetal brain (from the ectoderm cell lineage) were investigated. Additionally, colon smooth muscle (mesoderm cell lineage (De Santa Barbara et al. 2003), a different lineage to melanocytes which originate in the ectoderm) where *OCA2* gene is not expressed was used to as a comparison for enhancer specificity.

## 2.3 Extraction of variation data from the promoter and enhancer regions from KGP and AGVP variant call files (VCF)

The region of chromosome 15: 28324255 – 29324255 was extracted from the KGP and AGVP datasets. The region extends from exon 2 of *OCA2* and 1 Mb upstream of this starting point, which would include the previously characterised promoter for the *OCA2* gene as well as the *HERC2* gene which was suspected to harbour an enhancer for the *OCA2* gene. The data from this chromosome region was extracted from the individual population data into a variant call file (VCF).

The African KGP populations were used to generate a VCF for each population. All AGVP populations were used in this step and the data for all individuals was used to generate a single VCF for the AGVP data. This VCF was then split using sample identifiers according to the population, to generate a VCF for each AGVP population. This was performed individually for each population using PLINK version 1.9 (www.cog-genomics.org/plink/1.9/) (Chang et al. 2015). The generated data included all SNPs found within this region and their genotypes in all individuals for each population.

## 2.4 Variant frequency calling from the KGP and AGVP populations

PLINK version 1.9 was used to call the allele frequencies of all variants in each population from the VCFs (Chang et al. 2015). The generated frequency file included the reference and alternative alleles as well as the minor allele frequency.

The frequency files were generated separately for the promoter and enhancer regions of the *OCA2* gene based on their coordinates for each of the African KGP populations. These files were merged and filtered to remove nonpolymorphic variants. This list of alleles was maintained for reference. The joint files were filtered based on the minor allele frequency (MAF) in all of the African populations simultaneously, to determine which variants could be considered for further analysis and which were very common. A frequency threshold of 1% is considered common, while 5% and 10% indicate variants that are very common. The MAF was also compared to other continental populations in KGP. Frequency calling was then carried out in the promoter and enhancer regions for the combined AGVP data from all three populations to determine the overall allele frequencies of the variants. Following this, population specific VCFs were generated to calculate the population specific frequency for each variant in the

regulatory regions. The combined AGVP frequency files for the three populations were filtered by MAF using the same frequency thresholds as previously.

A list of variants identified in the promoter and enhancer regions were compared between the KGP and AGVP datasets. This would indicate which variants were common in both datasets and which other variants were only found in one of the datasets. This study focused on how common the variants were in all the African populations and how the frequencies differed between African populations. This was in addition to how the African variant allele frequencies compared to the frequencies seen the non-African continental KGP populations.

## 2.5 Functional annotation of variants

### 2.5.1 Variant Effect Predictor (VEP)

The set of variants extracted from the *OCA2* regulatory regions was non-coding. These variants were analysed using the Ensembl Variant Effect Predictor (VEP) version 94 (McLaren et al. 2016). VEP has a large database of information concerning known variation and their properties, such as location, variant nomenclature and population frequencies in larger studies such as KGP. The GRCh37 version of VEP was utilised for this step. Since some variants in the dataset did not have Single Nucleotide Polymorphism Database (dbSNP) identifiers in the VCFs, all variants were submitted to VEP to determine their existing identifiers and the corresponding frequencies of these variants in the KGP continental populations. The current version of dbSNP is build 153.

It is also possible to integrate annotation tools into VEP. Many tools for coding variation are available, such as SIFT (Kumar et al. 2009) and PolyPhen2 (Adzhubei et al. 2010), however they are not applicable for analysis of non-coding variation. A popular plugin that can be used to annotate non-coding variation in VEP is Combined Annotation Dependent Depletion (CADD). CADD can be used to predict deleterious effects of single nucleotide variation (SNV) and small insertions and deletions in coding and non-coding regions (Rentzsch et al. 2019). CADD compares all known variation in the human genome based on several layers of annotation such as regulatory effects and sequence conservation to generate scores that rank the variant for its deleteriousness. Scaled CADD scores are PHRED-like, which means that they are derived using $\log_{10}$, and rank all known SNVs in the human genome for easier interpretation of predicted potential deleterious effects of the variants (Table 2.2). The scaled score (CADD

version 1.3) was utilised to interpret potential functionality of all identified variants in the dataset.

**Table 2.2: A breakdown of CADD scaled scores (Rentzsch et al. 2019).**

| Scaled score | Classification | Scaled score meaning |
|---|---|---|
| <10 | Benign | Variant is in the bottom 90% of deleterious variants in the human genome |
| >10 | Likely benign | Variant is in the top 10% of deleterious variants in the human genome |
| >15 | Likely deleterious | Variant is in the top 5% of deleterious variants in the human genome |
| >20 | Deleterious | Variant is in the top 1% of deleterious variants in the human genome |
| >30 | Highly deleterious | Variant is in the top 0.1% of deleterious variants in the human genome |

### 2.5.2 RegulomeDB

RegulomeDB version 1.1 (http://regulomedb.org/) can be used to perform functional annotation for potential regulatory and epigenetic signals of particular variants in non-coding regions, as well as a prediction of how they might function if they occur in a regulatory element (Boyle et al. 2012). This version of RegulomeDB is based on dbSNP build 141 and contains ChIP-seq and DNase-seq data from ENCODE (ENCODE Project Consortium 2012), Position-Weight Matrix information which describes transcription factor (TF) binding from several sources such as JASPAR CORE (Bryne et al. 2008) and other experimental and predictive evidence for functionality. The RegulomeDB database scored variants by the available functional evidence to differentiate between variants which are more likely to be functional and those which are not.

As seen in Table 2.3, the smaller the score, the more likely it is that the variant occurs in a functional regulatory element. A score of 1 and its subcategories is likely to affect binding and are linked to expression of a gene target. A score of 2 and its subcategories are likely to affect binding while a score of 3 is less likely to affect binding. There is minimal binding evidence for scores of 4, 5 and 6 (Boyle et al. 2012).

**Table 2.3: A description of RegulomeDB scores by level of evidence (Boyle et al., 2012).**

| Score | Level of evidence |
|---|---|
| 1a | eQTL + TF binding + matched TF motif + matched DNase footprint + DNase peak |
| 1b | eQTL + TF binding + any motif + DNase footprint + DNase peak |
| 1c | eQTL + TF binding + matched TF motif + DNase peak |
| 1d | eQTL + TF binding + any motif + DNase peak |
| 1e | eQTL + TF binding + matched TF motif |
| 1f | eQTL + TF binding/DNase peak |
| 2a | TF binding + matched TF motif + matched DNase footprint + DNase peak |
| 2b | TF binding + any motif + DNase footprint + DNase peak |
| 2c | TF binding + matched TF motif + DNase peak |
| 3a | TF binding + any motif + DNase peak |
| 3b | TF binding + matched TF motif |
| 4 | TF binding + DNase peak |
| 5 | TF binding or DNase peak |
| 6 | Motif hit |

Since the most current version of RegulomeDB is based on build 141 of dbSNP, the website is unable to recognise variants that have been identified in more recent builds of dbSNP and this results in an error message. Therefore, the RegulomeDB database was downloaded and manually offline searched against the list of variants obtained from KGP and AGVP for the enhancer and promoter regions. For each variant appearing in RegulomeDB, the overall score was isolated as well as the evidence used to generate the score.

### 2.5.3 HaploReg

HaploReg version 4.1 (https://pubs.broadinstitute.org/mammals/haploreg/haploreg.php) can be used for functional annotation of variants which collates their regulatory information with other variants that occur in the same haplotype block (Ward and Kellis 2016). The functional information is from a variety of sources including conservation scores for the variant, minor allele frequencies from KGP phase 1 (The 1000 Genomes Project Consortium 2012) and epigenetic signals and chromatin state predictions for cell lines from Roadmap Epigenomics as

well as ENCODE. Additionally, there is information regarding protein binding sites from ENCODE ChIP-seq data, as well as TF binding sites and gene expression data from several sources. The variant information available from the most recent version of HaploReg is based on dbSNP build 141. Therefore, HaploReg will not provide information on variants from more recent versions of dbSNP.

To determine which enhancer and promoter variants had been annotated by HaploReg, their chromosomal coordinates were entered to find all variants from those regions which had been annotated by HaploReg. The tool pools the functional evidence for each variant into a summary table but the information for individual variants is also available. The annotations that were applicable for the melanocyte cell line were focused on, to see which variants may be functional in the regulatory elements in a cell type that expresses the *OCA2* gene.

### 2.5.4 atSNP

atSNP (http://atsnp.biostat.wisc.edu/) can be used to identify TF binding sites which overlap with variants in a statistically significant manner, for their reference and alternate alleles (Zuo et al. 2015). The binding sites are defined as gain of function or loss of function. Gain of function TF binding sites occur when the alternate allele for the variant creates a new binding site. The pairing of the TF and the variant's reference allele is not significant, but the alternate allele is significant. Additionally, loss of function TF binding sites are described where a TF binding site is present for the reference allele but it is lost when the alternate allele is substituted. Here the pairing of the reference allele and the TF binding site is significant, but the alternate allele is insignificant. Statistical significance for allele effects is defined by a combination of three p values where the cutoff for significance is at $p \leq 0.05$. The combination of p-values to define gain of function or loss of function effects of a variant are described in Table 2.4 below.

**Table 2.4: The p-values which define statistical significance for gain of function or loss of function effects (Zuo et al. 2015).**

| Effect | P-value SNP Impact | P-value Reference | P-value SNP |
|---|---|---|---|
| Gain of function | ≤0.05 | >0.05 | ≤0.05 |
| Loss of function | ≤0.05 | ≤0.05 | >0.05 |

A list of variants from the promoter and enhancer regions were entered to find TF binding sites which overlapped with the variant and had gain of function and loss of function effects. When stipulating that the parameters for either effect, the p-values matched those of the table above.

### 2.5.5 Genotype-Tissue Expression (GTEx) eQTL data

The Genotype-Tissue Expression project (https://gtexportal.org/home/) was undertaken to determine the normal tissue specific gene expression patterns in humans, as well as how genetic variation can influence these patterns of expression (GTEx Consortium 2015). It is possible to query the database for information about expression quantitative trait loci (eQTL), loci that can explain a level of the variation in gene expression which is caused by genetic variants. This may include searching for variants of interest to see if they influence the expression of any or specific genes and in which tissues this applies. GTEx was utilised in this study to identify if the regulatory region variants had potential influences on *OCA2* gene expression.

Though it is possible to browse the GTEx dataset on the dedicated website, it is only possible to search for information on individual variants. However, the dataset is downloadable and can be interrogated for multiple variants simultaneously manually and offline. Variants in the dataset are described by their position in GRCh37 coordinates and their possible alleles, therefore, it is not possible to use a variant's dbSNP identifier to search for applicable information. However, GTEx has made a variant look up table available for download. This table includes information about each variant in the dataset such as location, alleles and dbSNP identifier. This table was searched against the list of variants identified from the promoter and enhancer regions from KGP and AGVP to identify which of these variants appeared in the GTEx dataset and their applicable GTEx identifier for further interrogation of the data.

The variant identifiers were searched against the multi-tissue data for both regulatory regions to identify which variants appeared in the dataset, which gene they regulated and in which tissue the variant had the greatest effect on gene expression.

Data was available to interrogate single tissue cis eQTLs for specific tissue types. The single tissue data for skin was accessed, since melanin is produced by the skin and *OCA2* gene expression is increased in skin. Gene expression data in skin was accessed for sun exposed skin from the lower leg and non-sun exposed skin from the suprapubic region. This data was available in two file formats, which describe significant variant-gene pairs (variants that

significantly affect the expression of that gene in the tissue) and eGenes (genes that have more than one variant which regulate their expression). Both files were searched for the identified variants, to see if they had a significant relationship with expression of a gene and if they regulated *OCA2* gene expression specifically. Following this, the eGene and variant-gene pair files from both kinds of skin were interrogated for any variants which significantly affected *OCA2* gene expression in sun exposed and non-sun exposed skin.

## 2.6 Interpretation of variants

The evidence from the various tools was consolidated into an Excel spreadsheet. From this list, a set of variants were selected that were considered top hits which had the strongest functional evidence based on their annotations. The variants were excluded from further analysis if they had MAF ≤1% in all populations, none of the tools had annotated the variant and the variant had a benign or likely benign CADD score (<15). Additionally, variants were excluded if they had a MAF >1% or a likely deleterious or deleterious CADD score (>15), and they had not been annotated by the tools.

Furthermore, variants were included for further analysis if they satisfied one or more of the following criteria: variants had high CADD scores (>15), there was a frequency differential between African and non-African populations (common in Africa but not elsewhere or common elsewhere but not in Africa) and the variant had been annotated by at least one of the bioinformatic tools.

The variants that had some level of evidence from KGP and AGVP were consolidated into an Excel spreadsheet and the duplicate variants were removed by eye, leaving only single instances of individual variants. The evidence from the remaining variants were compared to determine which variants had the strongest evidence from the tools. This would imply some level of functionality and the variant may have a potential effect on expression of *OCA2*.

The top hits were then submitted to Genome Wide Annotation of Variants (GWAVA) (https://www.sanger.ac.uk/sanger/StatGen_Gwava) to predict their functionality (Ritchie et al. 2014). GWAVA sources its genomic and epigenomic annotation datasets mainly from ENCODE and uses them to discriminate between disease causing variation from the Human Gene Mutation Database (Stenson et al. 2009) and three sets of control SNVs from KGP phase 1 which had MAF >1%. The unmatched set of control variants was made up of a random

selection of SNVs from across the genome. The transcription start site (TSS) control set comprised of variants that were matched for distance to the nearest TSS at a genome-wide scale. The last set of control variants was the region set which contained all KGP variants found in the 1 kb region surrounding the Human Gene Mutation Database variants. A GWAVA score was generated for a variant of interest against each of the control datasets. These scores ranged from 0-1 and a variant with a score >0.5 was considered functional.

The tools and databases used in this study were pre-existing, where each has associated advantages and disadvantages. These tools were chosen on the basis of their ability to annotate non-coding variation and being publicly accessible. The annotation tools which were used in this study are summarised in Table 2.5.

**Table 2.5: Summary of annotation tools used in this study.**

| Tool | Data training sets or sources | Strengths | Weakness | Reference |
|---|---|---|---|---|
| Combined Annotation Dependent Depletion | The training set is made up of contrasted neutral variants which became fixed in humans since the divergence of apes and human ancestors are against simulated de novo variants which are free of selective pressure. A proportion of these variants are likely to be deleterious. | • Annotated coding and non-coding variants.<br>• Training variants were labelled objectively and systematically.<br>• Can be applied to any genomic feature which has coordinates in the reference assembly. This includes novel variants. | • Variants labelled as pathogenic or benign from the training set for each variant is not a prefect approximation. | (Rentzsch et al. 2019) |
| RegulomeDB | Data sources included ChIP-seq peaks for transcription factors, position weight matrices for transcription factor binding, DNase-seq footprints, detection of damaging variants by Polyphen-2 and literature from PubMed. | • Used experimentally derived evidence for association of coding and noncoding variants with functional regulatory regions.<br>• Applicable for functional annotation of variants identified in genome-wide association studies for disease. | • The list of variants was based on dbSNP 141 and had not been updated.<br>• Cannot annotate novel variants. | (Boyle et al. 2012) |
| HaploReg | Data sources included 1000 Genomes Project phase 1, ENCODE project, Roadmap Epigenomics Project, Genotype Tissue Expression project. | • Identification of functional variants which are in linkage disequilibrium with regulatory features. | • Annotations had not been updated regularly.<br>• Cannot annotate novel variants. | (Ward and Kellis 2016) |
| atSNP | Data sources included position weight matrices from the ENCODE project and the JASPAR database. | • Applicable to annotate large numbers of variants.<br>• Identified transcription factor binding sites which are generated or lost through the presence of the alternate allele of a variant. | • The list of variants is based on dbSNP 144.<br>• Cannot annotate novel variants. | (Zuo et al. 2015) |
| Genotype-Tissue Expression project | Association scores for variants with expression quantitative trait loci in various tissues. | • Includes expression data for various genes across tissue types.<br>• Can be used to link changes in gene expression to diseased and normal phenotypes. | • Tissues were sampled from post-mortem donors. Expression patterns may differ from living tissues. | (GTEx Consortium 2015) |

**Table 2.5 continued: Summary of annotation tools used in this study.**

| | | | | |
|---|---|---|---|---|
| Genome Wide Annotation of Variants | The training set was comprised of three sets of control SNVs: matched for distance from a transcription start site, variants within 1 kb of Human Gene Mutation Database variants and a random set of SNVs from around the genome. | • Possible to annotate coding and non-coding variation.<br>• The random forest algorithm has been modified to prevent class imbalance between different genomic features. | • Parameters for prioritisation of variants cannot be modified. | (Ritchie et al. 2014) |

# Chapter 3 - Results

This study aimed to characterise the normal variation within the *OCA2* gene control regions. Initially, it was necessary to confirm if a putative enhancer which had previously been identified had characteristics which are associated with enhancer regions. This included investigation into possible interactions between the promoter and putative enhancer of the *OCA2* gene using chromatin interaction data as well as characterisation of the putative enhancer for enhancer like histone modifications from publicly available datasets. This will show that the potential enhancer does have enhancer properties.

Thereafter, normal common variants within these regulatory regions could be identified from whole genome sequences of African individuals from the 1000 Genomes Project and the African Genome Variation Project. The variants were annotated for potential functionality in influencing expression of the *OCA2* gene using publicly available annotation tools. This will identify variants which had the strongest evidence for functionality in those regions and if they associated with differential pigmentation phenotypes.

## 3.1 Identifying the promoter and putative enhancer for *OCA2*

### 3.1.1 Interactions of the *OCA2* promoter and putative enhancer region

Chromatin interaction data was accessed to investigate if the *OCA2* promoter and the putative enhancer interact in the closest available cell lines to melanocytes. Hi-C and virtual chromosome conformation capture-on-chip (4C) data was accessed for melanocytes from a malignant melanoma cell line (RPMI7951), and a normal keratinocyte cell line (NHEK) via the 3D Genome Browser (accessed April 2019).

In the Hi-C data, *HERC2* and *OCA2* occurred in the same transcriptionally active domain. In RPMI7951, the heatmap indicated that there were interactions between the promoter region and the putative enhancer region, which appeared to be among the stronger interactions involving the promoter within this region (Figure 3.1). In NHEK, there were low levels of interactions between these two regions (Figure 3.2). The level of interaction between the regulatory regions that was seen in NHEK was relatively less than in RPMI7951.

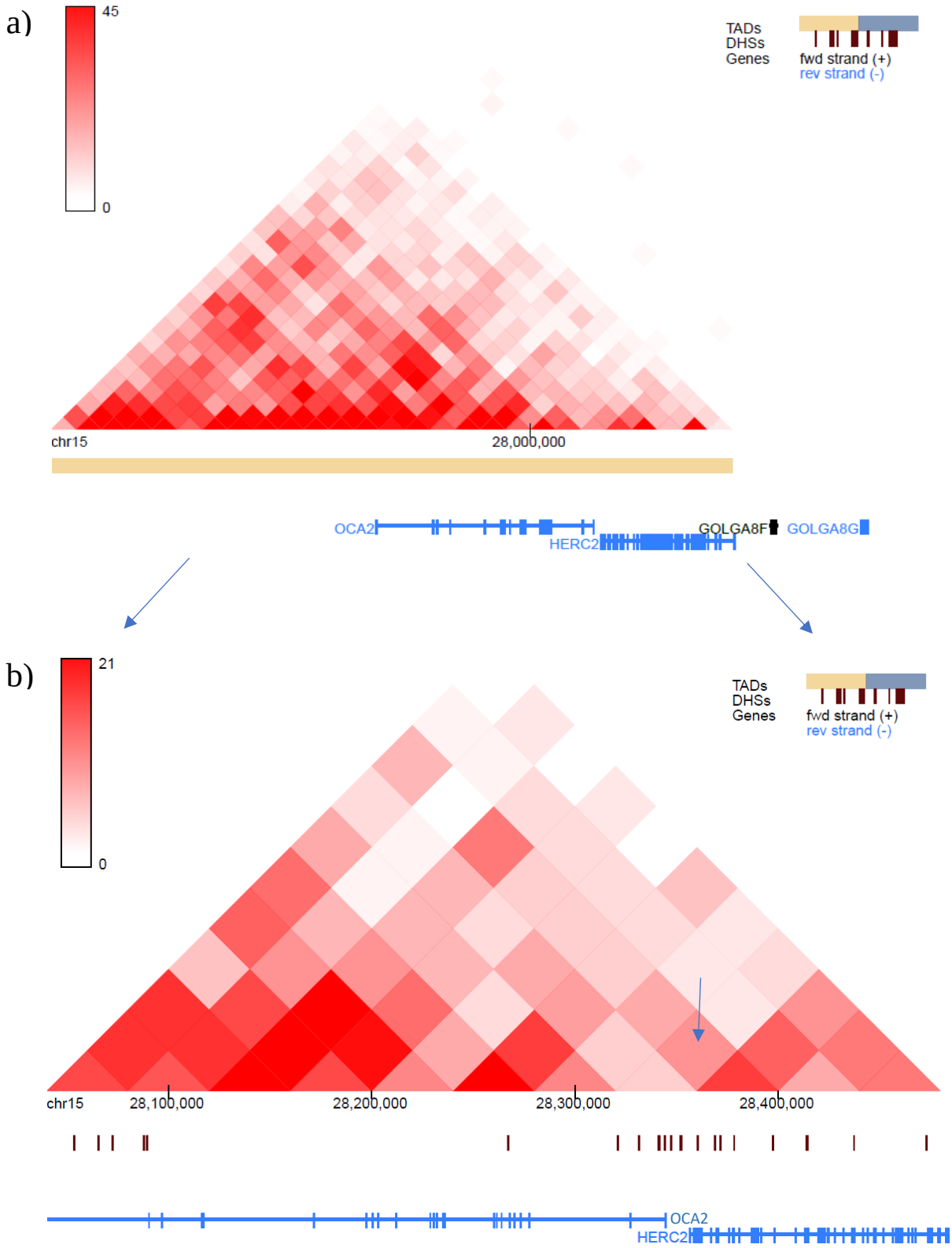



**Figure 3.1: Hi-C data for the *OCA2-HERC2* region in RPMI7951 melanocytes from a malignant melanoma cell line.** a) An overall view of the interactions in the *OCA2-HERC2* region on chromosome 15. b) A zoomed in view of the same region. The arrow indicates the point of interaction of the *OCA2* promoter and putative enhancer. Image generated using the 3D Genome Browser (Wang et al. 2018).

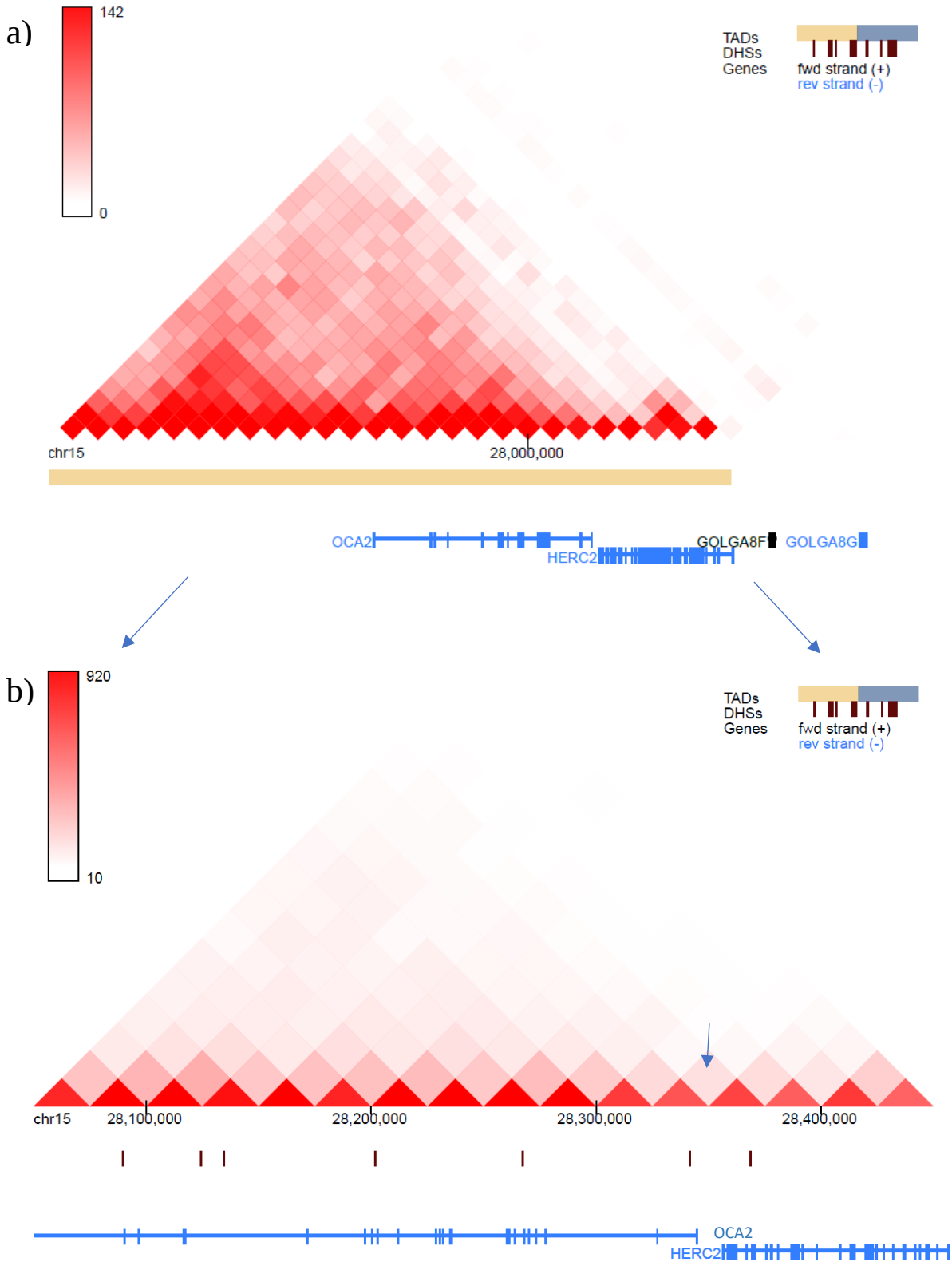



**Figure 3.2: Hi-C data for the *OCA2-HERC2* region in NHEK normal keratinocyte cells.** a) An overall view of the interactions in the *OCA2-HERC2* region on chromosome 15. b) A zoomed in view of the same region. The arrow indicates the point of intersection of the *OCA2* promoter and putative enhancer regions. Image generated using the 3D Genome Browser (Wang et al. 2018).

In the virtual chromosome conformation capture-on-chip (4C) data, the interaction between the rs12913832 region and the *OCA2* promoter was relatively stronger in RPMI7951 (Figure 3.3) than NHEK (Figure 3.4). In NHEK, rs12913832 had a stronger interaction with other regions in *HERC2* than the *OCA2* promoter. In RPMI7951, the interaction between the promoter and putative enhancer region was the strongest interaction in the region.

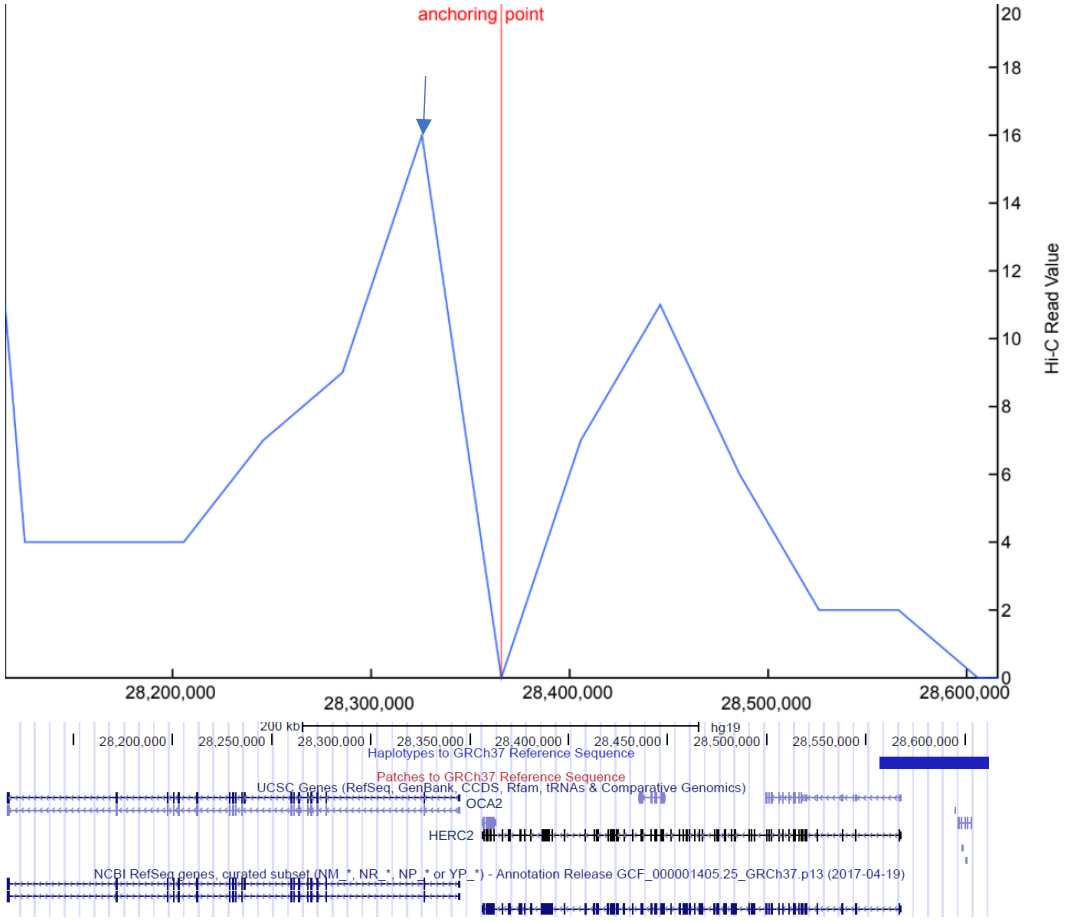



**Figure 3.3: Virtual chromatin conformation capture-on-chip data (4C) for the *OCA2-HERC2* region in RPMI7951 malignant melanoma cells with rs12913832 (enhancer region) as the anchoring point.** The arrow indicates the interactions of the rs12913832 region with the general location of the promoter region. The height of a peak indicates the read count for an interaction between the current locus and the anchor point. Image generated using the 3D Genome Browser (Wang et al. 2018). The corresponding aligned gene region as seen on the UCSC genome browser is below the 4C diagram (http://genome.ucsc.edu/).

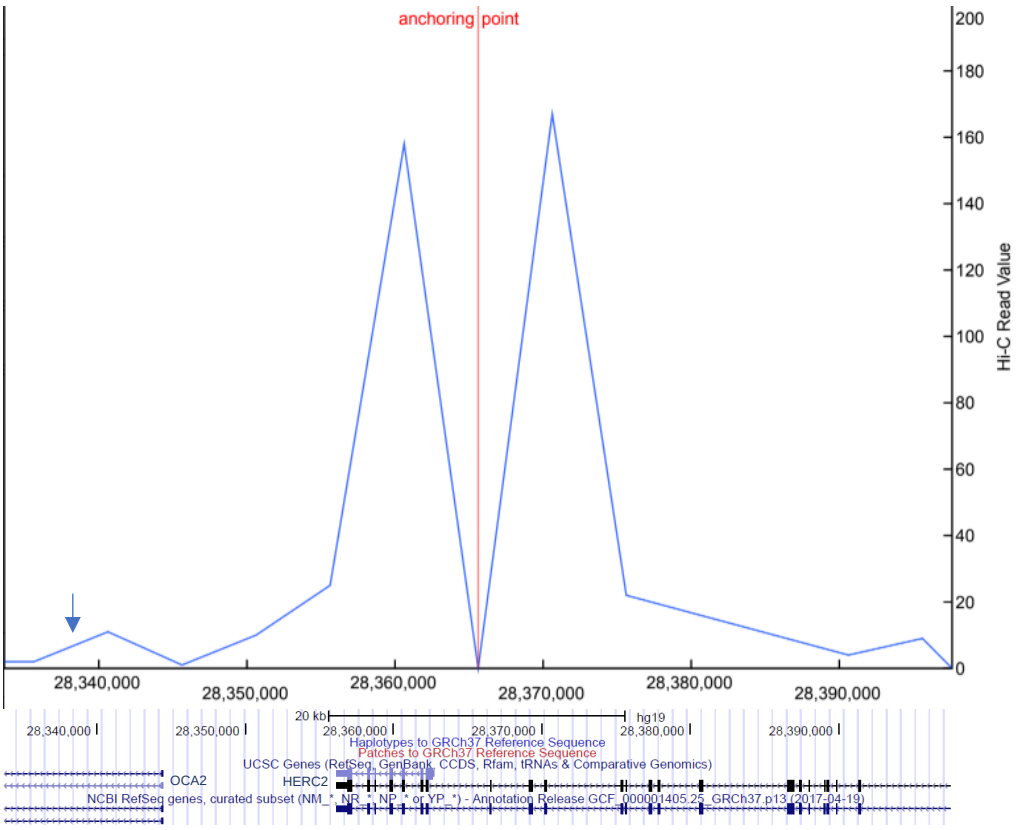



**Figure 3.4: Virtual chromatin conformation capture-on-chip (4C) data for the *OCA2-HERC2* region in NHEK normal keratinocytes with rs12913832 (enhancer region) as the anchoring point.** The arrow indicates the interactions of the rs12913832 region with the general location of the promoter region. The height of a peak indicates the read count for an interaction between the current locus and the anchor point. Image generated using the 3D Genome Browser (Wang et al. 2018). The corresponding aligned gene region as seen on the UCSC genome browser is below the 4C diagram (http://genome.ucsc.edu/).

In the Chromatin Interaction Analysis by Paired-End Tag (ChIA-PET) data, to determine if the interactions were present in cell lines of the same and different developmental lineages to melanocytes, all five cell lines were tested for RNA polymerase II (POL2) interactions in the *OCA2-HERC2* region (accessed November 2019). The K562 myeloid leukemia from erythrocytes cell line had a POL2 mediated interaction between the promoter region of *OCA2* and the 3' end of *HERC2* (Appendix D). This interaction did not align with the putative enhancer region in intron 86 of *HERC2*. Otherwise, no interactions between the *OCA2* promoter and *HERC2* were detected any of the other cell lines used in the ChIA-PET experiments, including MCF7 and HeLa S3 which, like melanocytes, were derived from ectodermal developmental lineages (Appendix D).

### 3.1.2 Histone modifications and chromatin state in the putative enhancer

The Roadmap Epigenomics project data was accessed to determine if the putative enhancer has enhancer-like epigenetic signatures and if these signatures are specific to melanocytes (accessed May 2019). The epigenetic signatures associated with enhancers include H3K4me1 (all enhancers) and H3K27ac (active enhancer). Additionally, to investigate the chromatin state of this region, DNase hypersensitivity sites were assessed for open chromatin and H3K27me3 which was associated with closed chromatin. The cell lines included were foetal foreskin melanocytes, foetal thymus, foetal brain and colon smooth muscle.

In melanocytes, the putative enhancer exhibited epigenetic markers of DNase hypersensitivity sites where the hypersensitivity signal was highest in the centre of the putative enhancer region. The histone modification profile of this region lacked H3K27me3 and was marked by the presence of H3K4me1 and H3K27ac histone modifications on the nucleosomes which flanked the DNase hypersensitivity site (Figure 3.5). This indicated that the region is marked by open chromatin and enhancer specific signals.

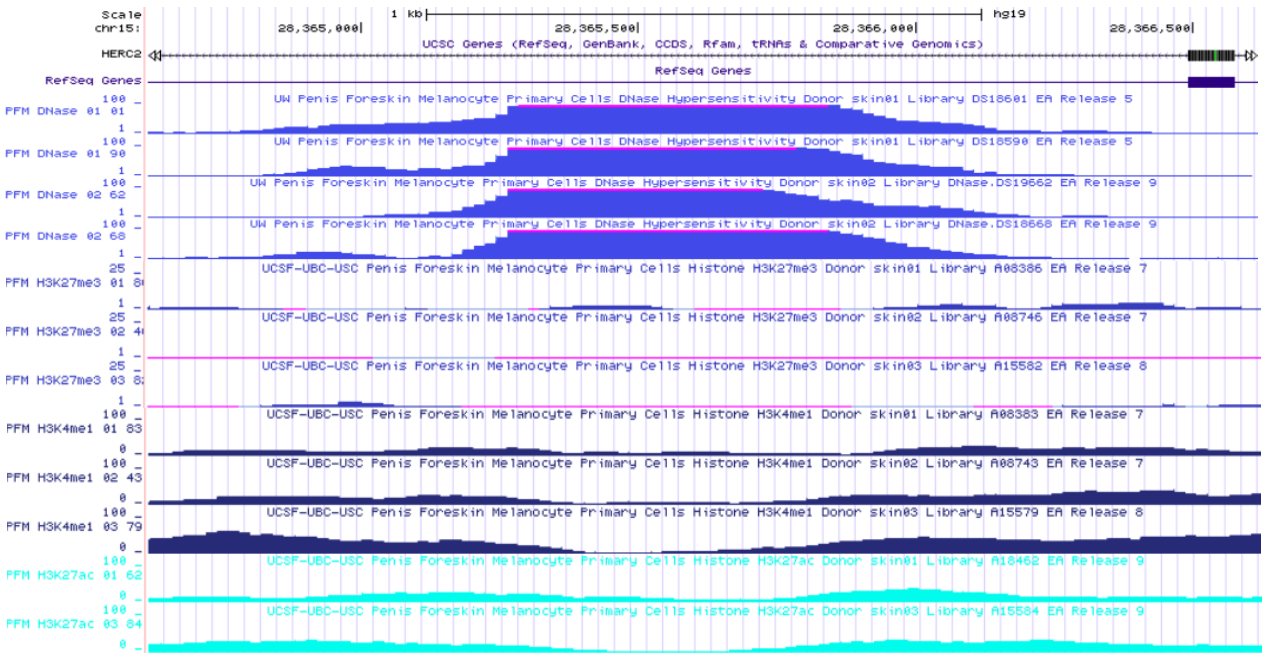



**Figure 3.5: A selection of enhancer specific signals in the foetal foreskin melanocyte cell line from the Roadmap Epigenomics Project for the coordinates (chromosome 15: 28364618 – 28366618).** This region represents the putative enhancer. A peak represents the presence of the histone modification signal while the absence of a peak correlates with the lack of the modification signal.

These properties are absent from the region in the comparative cell types (Appendix E). In the thymus, DNase hypersensitivity sites, H3K27me3, H3K4me1 and H3K27ac signals were absent from the region. In the brain, there was very little DNase hypersensitivity data, as well as low levels of H3K27me3 signals and no H3K4me1. There is no data available for H3K27ac signals in the foetal brain samples. In smooth colon muscle, there was no data for DNase hypersensitivity sites. However, there were low levels of H3K27me3 in the region, no H3K4me1 or H3K27ac signals in the region. The enhancer specific signals for the comparative cell lines can be seen in Appendix E. This indicated that the enhancer specific signatures were only present in melanocytes.

## 3.2 Frequencies of variants in the enhancer and promoter regions in the African KGP populations

When the KGP data was used to identify variants within the *OCA2* promoter region (chromosome 15: 28339081 – 28345199), 149 variants were located within this region. Sixty-eight of these variants were polymorphic in all the African populations, while 81 variants were nonpolymorphic in the African populations. This constituted the majority of variants identified by KGP in the promoter region. Twenty-four of these variants that were detected in the African populations had a MAF >1%, 15 variants had a MAF >5% and five had a MAF >10%.

Fifty-five variants were identified from the potential enhancer region of *OCA2* (chromosome 15: 28364618 – 28366618). Twenty-nine of these variants were polymorphic in the African populations. However, 26 variants were nonpolymorphic in the observed African populations. Seven variants had a MAF >1%, six variants had a MAF >5% and three variants had a MAF >10%. Very few variants in the enhancer region were common in African KGP populations.

## 3.3 Frequencies of variants from AGVP

In the promoter, 60 variants were identified from the AGVP data. All 60 variants were polymorphic, therefore all of the variants in this region were detected in the AGVP populations. Twenty-four of the variants had a MAF >1%, there were 13 variants with MAF >5% and nine variants with MAF >10%.

In the enhancer, 23 variants were identified in the AGVP data. Twenty-two variants were polymorphic and one variant was nonpolymorphic in all the AGVP populations. There were eight variants with MAF >1%, six variants with MAF >5% and four variants with MAF >10%.

There were 16 enhancer variants and 38 promoter variants present in both KGP and AGVP. As such, there were seven enhancer variants and 22 promoter variants that were specific to AGVP. These variants did not have corresponding allele frequencies from KGP. Three of the AGVP enhancer variants and a further nine promoter variants were novel.

## 3.4 Variants that were most likely to be functionally significant

The variants with the strongest evidence for functionality were selected based on the criteria established in Chapter 2. Following the exclusion of variants which did not have any evidence from the annotation tools, were rare and had a CADD score <15; a total of 137 variants were excluded. Variants which had no functional evidence from the bioinformatic tools but were common or had a CADD score >15 were initially retained for analysis. These 20 variants were subsequently removed owing to the lack of functional evidence. The remaining variants which had evidence were deduplicated, which resulted in the removal of 47 duplicates. There were 83 final variants left after selection based on evidence. This constituted of 28 enhancer variants and 56 promoter variants. Once the remaining variants were stratified based on the strength of their annotations, there were 10 variants in the enhancer and seven variants in the promoter with the strongest evidence for functionality (Figure 3.6). The approximate locations of these variants in the context of the *OCA2-HERC2* gene region are indicated in Figure 3.7.

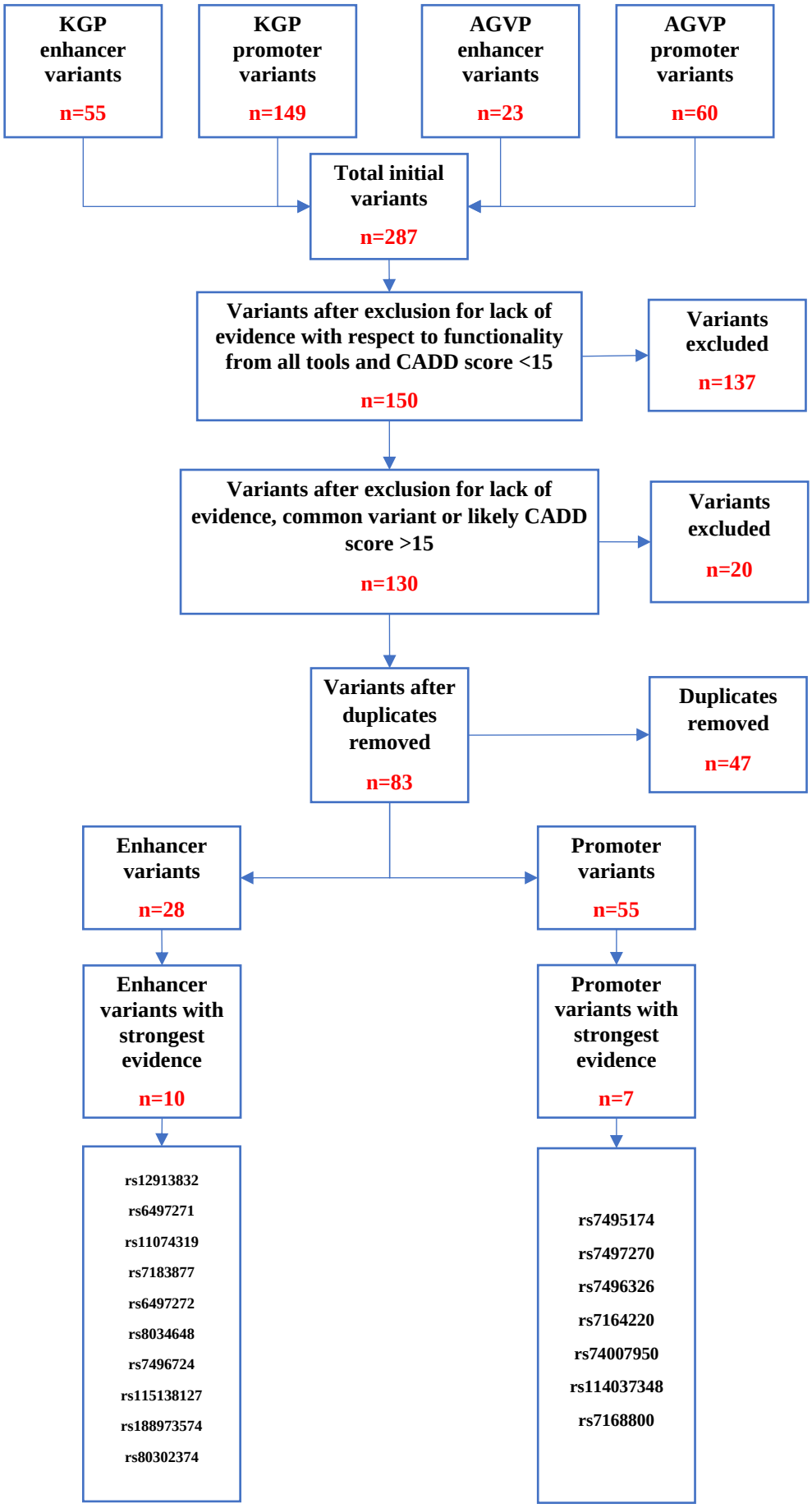



**Figure 3.6: Workflow of selecting the variants with the strongest evidence.**

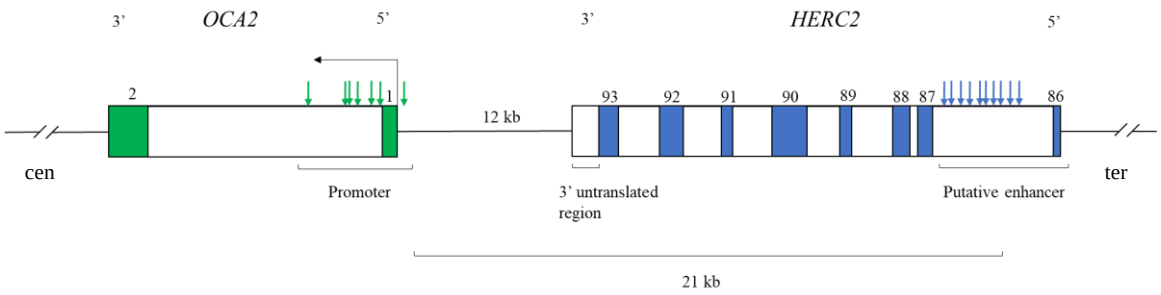



**Figure 3.7: The overall architecture of a section of the *OCA2* and *HERC2* genes, as well as the relative positions of the variants with the strongest evidence for functionality.** The gene exons are coloured blocks and the introns are colourless blocks. The green arrow are the promoter variants and the enhancer variants are indicated by blue arrows.

### 3.4.1 GWAVA

The top hits from the promoter and enhancer were submitted to GWAVA to predict their functionality (accessed July 2019). In the enhancer region, the rs12913832 variant had consistent scores ≥0.7, which is an indication that the variant is functional. rs6497271 had a score ≥0.5 with the TSS and unmatched classifiers, indicating functionality. Additionally, rs188973574 had a score of 0.51 with the unmatched classifier, indicating functionality. The rest of the variants had scores <0.5 with all classifiers, which is a prediction of non-functionality of these variants (Table 3.1).

**Table 3.1: The GWAVA scores for the enhancer variants with the strongest evidence for functionality.**

| Variant | Chromosome | Position | Region score | TSS score | Unmatched score |
|---|---|---|---|---|---|
| rs11074319 | 15 | 28364954 | 0.35 | 0.21 | 0.16 |
| rs7496724 | 15 | 28365088 | 0.28 | 0.23 | 0.13 |
| rs115138127 | 15 | 28365199 | 0.25 | 0.28 | 0.19 |
| rs6497271 | 15 | 28365431 | 0.42 | 0.5 | 0.56 |
| rs188973574 | 15 | 28365451 | 0.48 | 0.39 | 0.51 |
| rs12913832* | 15 | 28365618 | 0.75 | 0.7 | 0.74 |
| rs7183877 | 15 | 28365733 | 0.28 | 0.16 | 0.07 |
| rs80302374 | 15 | 28365867 | 0.31 | 0.2 | 0.16 |
| rs6497272 | 15 | 28366190 | 0.37 | 0.18 | 0.18 |
| rs8034648 | 15 | 28472307 | 0.32 | 0.09 | 0.03 |

*Variant with evidence for functionality.

The promoter variant scores were generally <0.5 according to the region and TSS classifiers, thus, they are predicted to be non-functional in this case. However, rs7497270 and rs7496326 had scores ≥0.5 under these same categories. All variants had scores >0.5 in the unmatched classifier. These cases indicate functionality of the variants (Table 3.2).

**Table 3.2: The GWAVA scores for the promoter variants with the strongest evidence for functionality.**

| Variant | Chromosome | Position | Region score | TSS score | Unmatched score |
|---|---|---|---|---|---|
| rs74007950* | 15 | 28339882 | 0.29 | 0.41 | 0.58 |
| rs7168800* | 15 | 28341575 | 0.37 | 0.4 | 0.73 |
| rs7164220* | 15 | 28341609 | 0.29 | 0.39 | 0.69 |
| rs114037348* | 15 | 28341863 | 0.29 | 0.39 | 0.68 |
| rs7495174* | 15 | 28344238 | 0.25 | 0.45 | 0.93 |
| rs7497270* | 15 | 28344328 | 0.46 | 0.64 | 0.95 |
| rs7496326* | 15 | 28344695 | 0.31 | 0.5 | 0.91 |

*Variant with evidence for functionality.

### 3.4.2 Collection of evidence for the variants that were most likely to be functional

Two variants from the set of likely functional variants were considered the most likely to be functional, based on the evidence collected from the bioinformatics tools. These variants were rs7495174 from the promoter and rs12913832 from the enhancer region. Additionally, these variants represented the most likely to be functional and phenotypically relevant in each of the regulatory regions for *OCA2*. The evidence for these variants will be presented below.

**rs7495174**

rs7495174 was located in the promoter and has substantial evidence to suggest that it may be functionally relevant. rs7495174 had previously been associated with hair colour. It has been associated with gene expression for more than one gene and cell type. rs7495174 has been predicted to influence *HERC2* expression in whole blood (p-value = $3.763 \times 10^{-22}$) as well as in lymphoblastoid cells. The variant has also been associated with *OCA2* expression in lung tissue. This variant had a RegulomeDB score of 3a and its CADD score was 8.401, which is not likely to be deleterious. The proteins that were predicted to bind to the region surrounding rs7495174 included CTCF, MAX, CMYC, ZBTB7A and ZNF263. It was also predicted that the alternate allele of this variant would change binding sites for the ERα-a, Esr2, Gm397, Mtf1, RAR, RXRA and T3R transcription factors. ChromHMM state predictions suggested that rs7495174 occurs in the vicinity of a transcription start site, which is within a promoter. The variant occurred within the H3K4me3, H3K4me1 and H3K27ac histone modifications which further suggest that it occurs in an active regulatory region in melanocytes. This variant was shown to colocalise with a DNase hypersensitivity site and was predicted to be functional by GWAVA when compared to random SNVs from across the genome.

**rs12913832**

rs12913832 occurred in the enhancer region and represented the variant in that region, has been associated with pigment phenotypes such as skin and eye colour. In the enhancer, rs12913832 was associated with *HERC2* expression where the strongest association was in whole blood (p-value = $1.36 \times 10^{-7}$). Was also indicated to affect the expression of *CHKB*. It had a RegulomeDB score of 5 and was indicated to occur in an enhancer region in HaploReg. CADD score of 12.86, likely benign. Colocalised with DNase hypersensitivity and histone modifications for active enhancers. ChromHMM also predicted that the variant occurred in an enhancer region.

Of the full list of variants identified in KGP and AGVP, the only variant annotated by atSNP was rs12913832, which was located in the enhancer (accessed May 2019). The tool identified a total of 95 significant pairings of TF binding motifs influenced by the variant. For gain of function effects, 71 TF binding sites were predicted to be established when the alternate allele was present. When stipulating loss of function, a further 24 TF binding sites were predicted to be disrupted by the presence of the alternate allele but were usually present with the reference allele. Some TFs which were predicted to be affected by a loss of function included MYC, ATF1, ESRRA and SOX10. Furthermore, TFs predicted to be affected by a gain of function by the alternate allele included TBX2, HES7, IRF and ETV7.

### 3.4.3 Allele frequencies

There were definite allele frequency differences in the promoter variants between the African and non-African populations. For all examined variants except rs7168800, the minor allele was generally common in all African populations. rs7495174, rs7497270, rs7496326 and rs7164220 were generally more common in EAS and SAS; while they were less common in EUR and AMR compared to the African populations. rs74007950 and rs114037348 were rare or absent in the non-African populations but were very common in the African populations. Though rs7168800 was common in a few of the African populations, it was largely rare in these populations. This variant was more common in the non-African populations (Figure 3.8)

For the enhancer variants, five of the variants were common in all populations. rs6497271, rs6497272, rs11074319 and rs7183877 were more common in the non-African populations. By comparison, rs8034648 was more common in the African populations. rs7496724 and rs115138127 were common in the African populations but rare in the non-African populations. rs188973574 and rs80302374 were common in a few African populations each but were absent from the non-African populations. The exception to this trend was rs12913832, which was absent in the majority of the African populations and EAS. The variant was common in ACB, ASW, GWD and the remaining non-African populations (Figure 3.9 and Figure 3.10).

### 3.4.4 Annotations that indicate functionality of variants

There were differing levels of annotation of the variants by the bioinformatics tools. The promoter variants were broadly annotated to be functional by GWAVA and indicated to occur in a promoter region by the ChromHMM states and their associated histone modifications. All variants except rs7496326 were indicated by HaploReg to change TF binding sites when the

alternative allele was present. rs7495174, rs7497270 and rs7496326 coincided with binding sites for proteins. All promoter variants were associated with regulation of gene expression in GTEx and/or other datasets. None of these variants were predicted to be deleterious by CADD. The variants that had the strongest evidence of functionality were rs7495174 and rs7497270, though rs7495174 had the additional evidence of having been associated with a pigmentation phenotype (Table 3.3).

The enhancer variants were all indicated to have enhancer specific histone modifications and have enhancer appropriate ChromHMM states. rs12913832, rs6497271 and rs188973574 were predicted to be functional by GWAVA. rs6497271 was the only variant predicted to be deleterious by CADD. None of the variants were predicted to occur in a protein binding site but all variants except rs12913832 and rs11074319 were predicted by HaploReg to change TF binding sites. Comparatively, rs12913832 was the only variant predicted by atSNP to change TF binding sites. Several variants were indicated to regulate gene expression. rs12913832, rs6497271, rs11074319, rs7183877 and rs6497272 were annotated by HaploReg to regulate *HERC2* expression. rs12913832 was additionally annotated for *CHKB* expression and rs6497272 was annotated for *GOLGA8G* expression. GTEx also suggested that rs12913832 regulated *HERC2* expression. Taking this information into account, rs12913832 had the strongest indication of functionality in the enhancer region (Table 3.4).

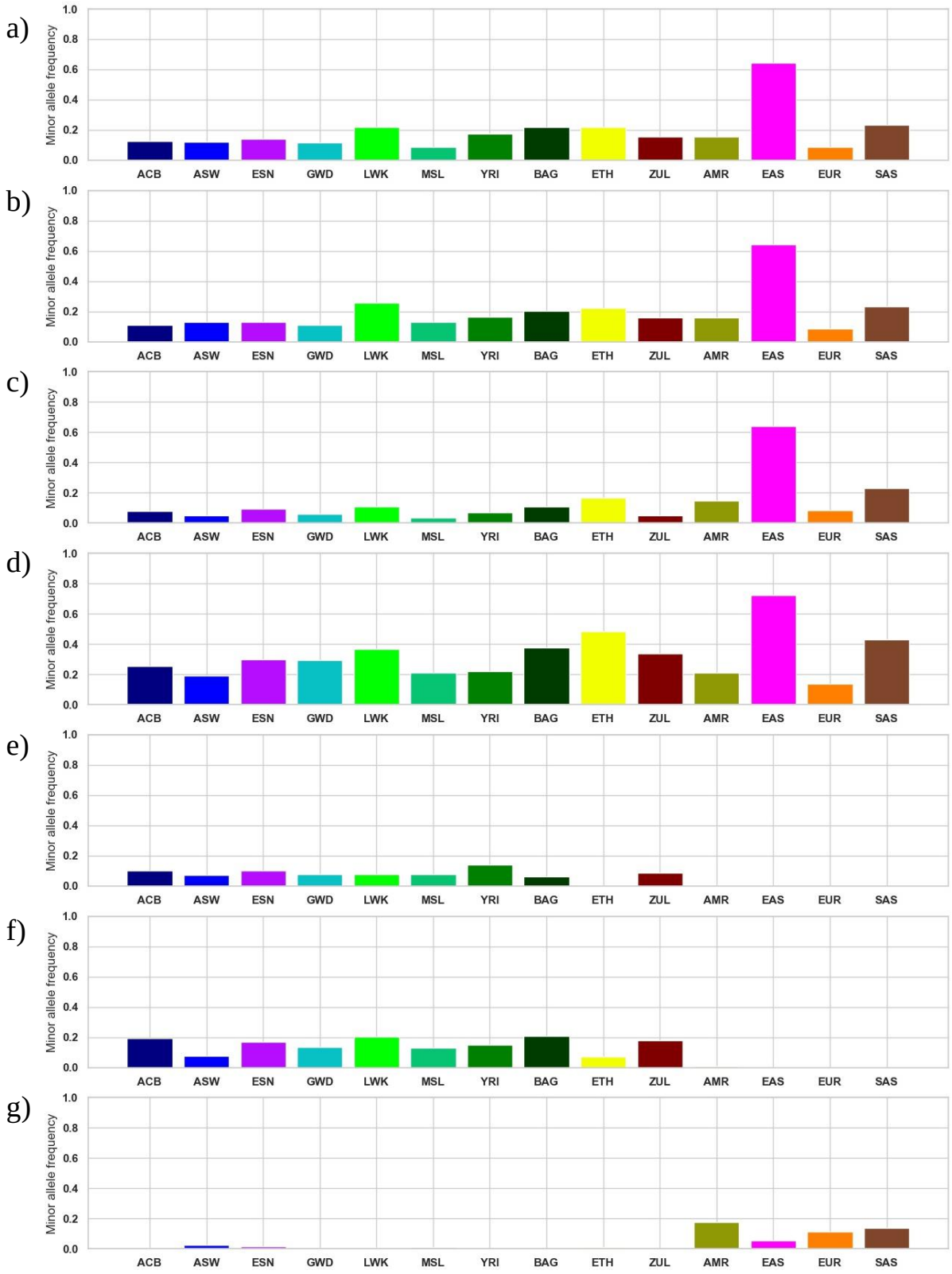



**Figure 3.8: The minor allele frequencies of the promoter variants with the strongest evidence in the examined African and non-African populations.** a) rs7495174 (G), b) rs7497270 (T), c) rs7496326 (T), d) rs7164220 (C), e) rs74007950 (T), f) rs114037348 (G), g) rs7168800 (A).

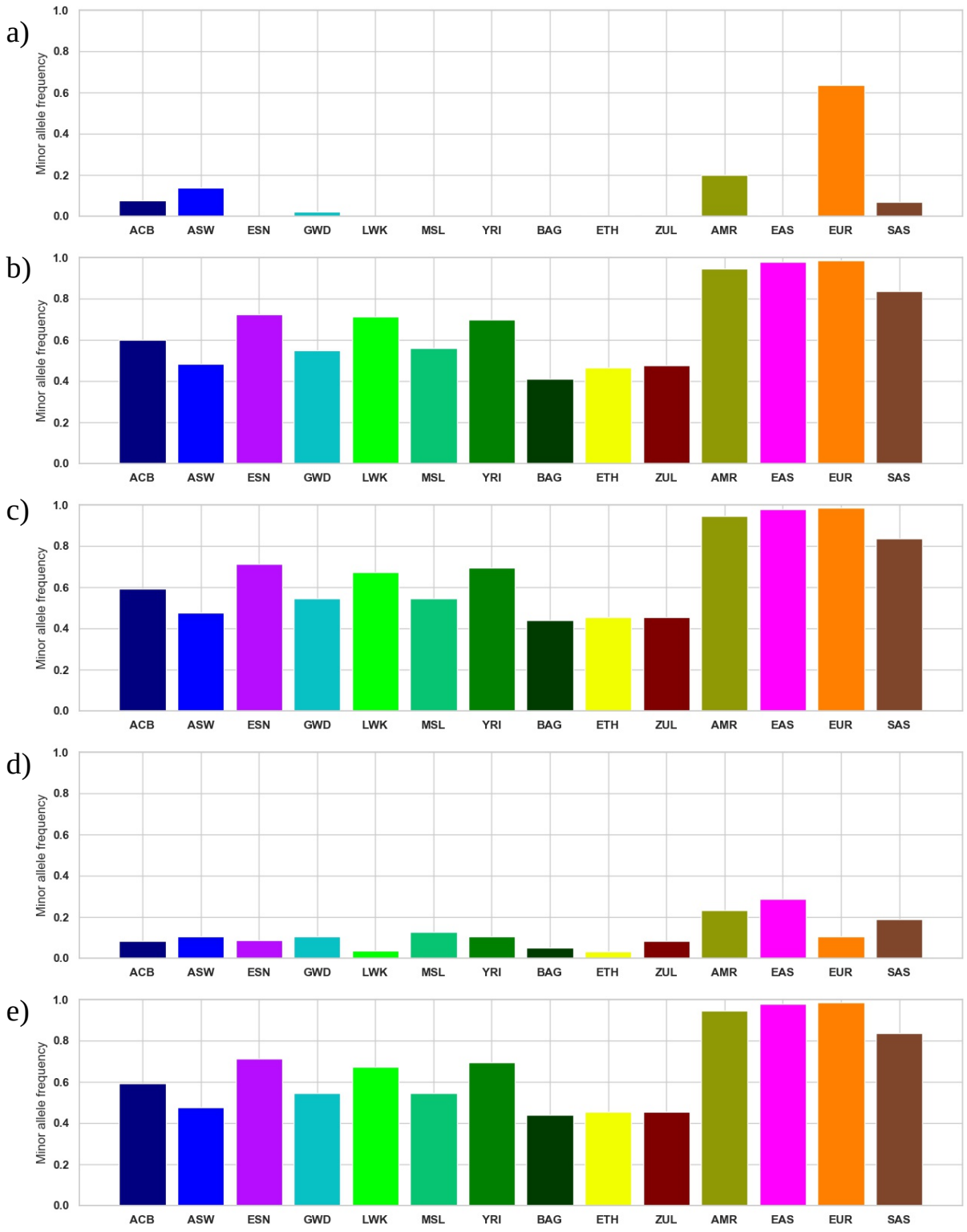



**Figure 3.9: The minor allele frequencies of the first five enhancer variants with the strongest evidence in the examined African and non-African populations.** a) rs12913832 (G), b) rs6497271 (G), c) rs11074319 (G), d) rs7183877 (A), e) rs6497272 (G).

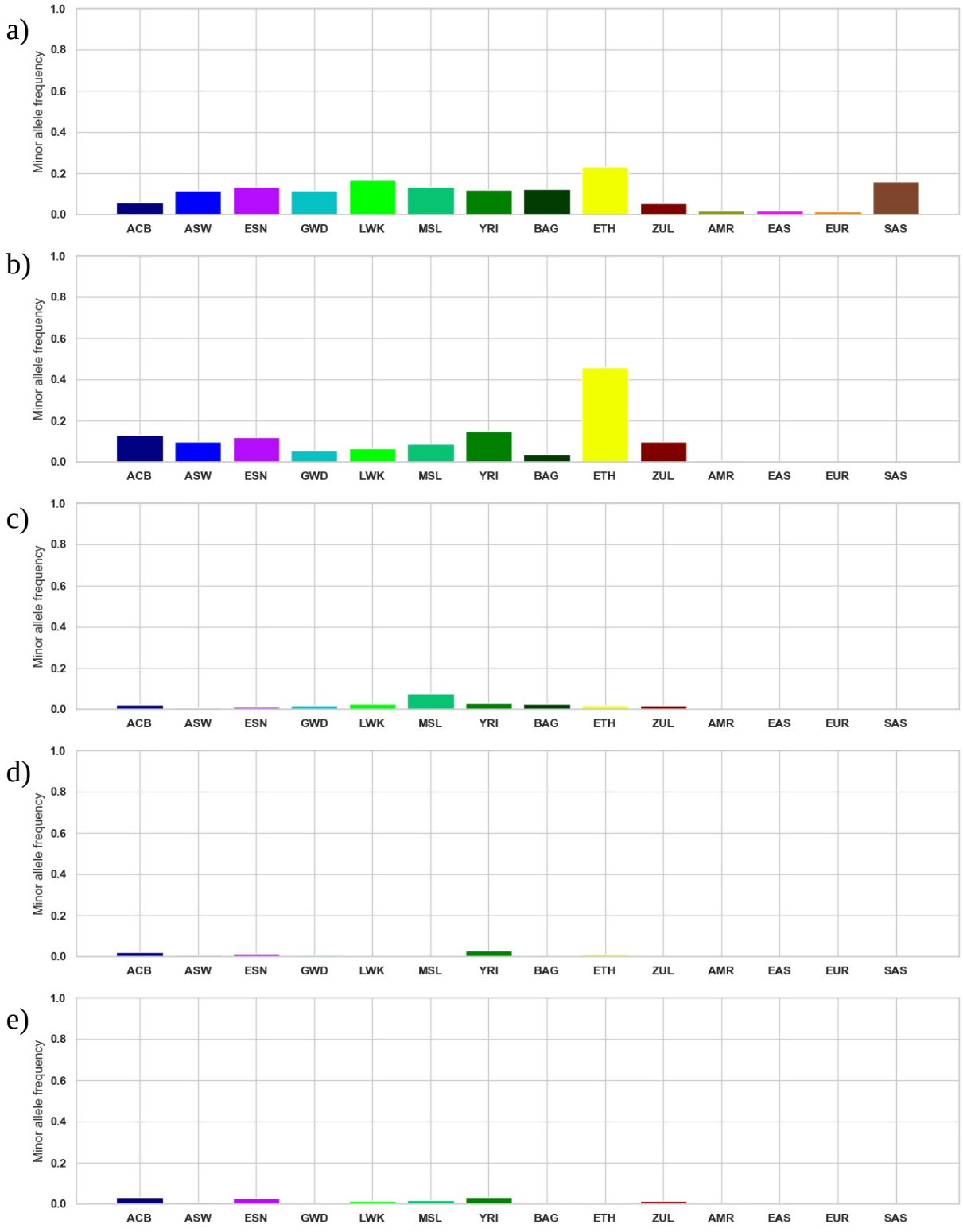



**Figure 3.10: The minor allele frequencies of the second five enhancer variants with the strongest evidence in the examined African and non-African populations.** a) rs8034648 (T), b) rs7496724 (G), c) rs115138127 (A), d) rs188973574 (C), e) rs80302374 (C).

**Table 3.3: The evidence associated with the promoter variants that are most likely to be functional.**

| | | | | | HaploReg | | | | | | | | | | atSNP | GTEx | GWAVA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Variant** | **Associated phenotype** | **CADD score** | **Predicted regulatory region consequence** | **RegulomeDB score** | **DNase hypersensitivity** | **H3K4me3** | **H3K4me1** | **H3K27ac** | **ChromHMM 15 state[1]** | **ChromHMM 25 state[2]** | **Protein binding[3]** | **Transcription factor binding sites changed[4]** | **Conservation prediction[5]** | **Genes regulated and significant tissue[6]** | **Transcription factor binding sites changed** | **Genes regulated in significant tissues** | **Functionality prediction** |
| rs7495174 | Hair colour | 8.401 | Yes | 3a | Yes | Yes | Yes | Yes | 1 | 2 or 3 | CTCF, MAX, CMYC, ZBTB7A, ZNF263 | ERα-a, Esr2, Gm397, Mtf1, RAR, RXRA, T3R | None | *HERC2* in whole blood and lymphoblastoid *OCA2* in lung | None | *HERC2* in whole blood | Functional (unmatched) |
| rs7497270 | None | 12.27 | Yes | 2b | Yes | Yes | Yes | Yes | 1 | 2 or 3 | POL2, MAX, E2F6, ZNF263 | EWSR1-FLI1, Hand1, SP1, TATA | None | *HERC2* in whole blood and lymphoblastoid | None | *HERC2* in whole blood | Functional (unmatched and TSS) |
| rs7496326 | None | 8.857 | Yes | 4 | Yes | Yes | Yes | Yes | 1 | 2 | ZBTB7A | None | None | *HERC2* in whole blood and lymphoblastoid | None | *HERC2* in whole blood | Functional (unmatched and TSS) |
| rs7164220 | None | None | Yes | 4 | Yes | Yes | Yes | Yes | 3 | 2 or 3 | None | EBF, SP1 | None | *HERC2* in whole blood | None | *HERC2* in whole blood | Functional (unmatched) |

**Table 3.3 continued: The evidence associated with the promoter variants that are most likely to be functional.**

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| rs74007950 | None | 10.63 | Yes | 4 | No | Yes | Yes | Yes | 3 | 3 | None | BHLHE40, CTCF, E2F, EWSR1-FLI1, MAZR, NERF1a, NRSF, Pbx3, RXRA, SMC3, TCF12, WT1, YY1, p300 | None | None | None | *RPL41P2* in brain cerebellum | Functional (unmatched) |
| rs114037348 | None | 8.518 | Yes | 4 | Yes | Yes | Yes | Yes | 1 or 3 | 2 or 3 | None | NF-κB | None | None | None | *OCA2* in tibial artery | Functional (unmatched) |
| rs7168800 | None | 8.995 | Yes | 4 | Yes | Yes | Yes | Yes | 1 or 3 | 2 or 3 | None | CHOP::C/EBPα, Myb | None | *HERC2* in lymphoblastoid | None | None | Functional (unmatched) |

[1]ChromHMM 15 state description (Ernst and Kellis 2012), full list of codes in Appendix F: 1 – Active transcription start site, 3 – Transcribed at gene 5' and 3'

[2]ChromHMM 25 state description (Ernst and Kellis 2015), full list of codes in Appendix F: 2 – Promoter upstream transcription start site, 3 - Promoter downstream transcription start site with DNase

[3](The ENCODE Project Consortium 2011)

[4](Kheradpour and Kellis 2014)

[5] Conservation prediction was by GERP (Cooper et al. 2005) and SiPhy (Garber et al. 2009)

[6](Hao et al. 2012; Lappalainen et al. 2013; Westra et al. 2013; Leslie et al. 2014; Li et al. 2014; GTEx Consortium 2015)

**Table 3.4: The evidence associated with the enhancer variants that are most likely to be functional.**

| | | | | | HaploReg | | | | | | | | | | atSNP | GTEx | GWAVA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Variant** | **Associated phenotype** | **CADD score** | **Predicted regulatory region consequence** | **RegulomeDB score** | **DNase hypersensitivity** | **H3K4me3** | **H3K4me1** | **H3K27ac** | **ChromHMM 15 state[1]** | **ChromHMM 25 state[2]** | **Protein binding[3]** | **Transcription factor binding sites changed[4]** | **Conservation prediction[5]** | **Genes regulated and significant tissue[6]** | **Transcription factor binding sites changed** | **Genes regulated in significant tissues** | **Functionality prediction** |
| rs12913832 | Hair, eye and skin colour | 12.86 | No | 5 | Yes | No | Yes | Yes | 7 | 11 | None | None | GERP and SiPhy | *HERC2* and *CHKB* *HERC2* in whole blood and lymphoblastoid | 95 sites | *HERC2* in whole blood | Functional (unmatched, TSS and region) |
| rs6497271 | None | 20.8 | No | 5 | Yes | No | Yes | Yes | 7 | 11 | None | BATF, CTCF, Foxa, HNF4, RXRA, Sin3Ak-20, Sox-13 | GERP and SiPhy | *HERC2* in lymphoblastoid | None | None | Functional (unmatched and TSS) |
| rs11074319 | None | 4.164 | Yes | 5 | Yes | No | Yes | Yes | 6 | 11 | None | None | None | *HERC2* in lymphoblastoid | None | None | Non-functional |
| rs7183877 | None | 4.62 | No | 5 | Yes | No | Yes | Yes | 7 | 11 | None | PLZF | None | *HERC2* in lymphoblastoid | None | None | Non-functional |

**Table 3.4 continued: The evidence associated with the enhancer variants that are most likely to be functional.**

| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| rs6497272 | None | 3.341 | No | 5 | Yes | No | Yes | Yes | 6 or 7 | 11 | None | CDP_7, Sin3Ak-20_disc4 | None | *HERC2* in brain cortex and lymphoblastoid *GOLGA8G* in LUAD | None | None | Non-functional |
| rs8034648 | None | None | Yes | 5 | No | No | Yes | Yes | 6 or 7 | 11 | None | SIX5, Spz1 | None | None | None | None | Non-functional |
| rs7496724 | None | 4.844 | No | 5 | Yes | No | Yes | Yes | 2 or 7 | 11 | None | Mrg1::Hoxa9, RXRA, XBP-1 | None | None | None | None | Non-functional |
| rs115138127 | None | 0.57 | No | 5 | Yes | No | Yes | Yes | 2 or 7 | 11 | None | MOVO-B | None | None | None | None | Non-functional |
| rs188973574 | None | 3.417 | No | 5 | Yes | No | Yes | Yes | 7 | 11 | None | Arid3a, Nkx2, STAT, Sox_3, THAP1 | GERP | None | None | None | Functional (unmatched) |
| rs80302374 | None | 0.121 | No | 5 | Yes | No | Yes | Yes | 7 | 11 | None | CDP, Obox6 | None | None | None | None | Non-functional |

[1]ChromHMM 15 state description (Ernst and Kellis 2012), full list of codes in Appendix F: 2 – Flanking active transcription start site, 6 – Genic enhancers, 7 - Enhancers

[2]ChromHMM 25 state description (Ernst and Kellis 2015), full list of codes in Appendix F: 11 - Transcription 3' preferential and enhancer

[3](The ENCODE Project Consortium 2011)

[4](Kheradpour and Kellis 2014)

[5] Conservation prediction was by GERP (Cooper et al. 2005) and SiPhy (Garber et al. 2009)

[6](Hao et al. 2012; Lappalainen et al. 2013; Westra et al. 2013; Leslie et al. 2014; Li et al. 2014; GTEx Consortium 2015)

## 3.5 Annotating the variants for regulatory effects

The general results that the bioinformatics tools generated for the collection of variants will be briefly elaborated upon below.

### 3.5.1 Variant Effect Predictor

When the full list of variants from KGP and AGVP for the two regions were submitted to VEP, all variants that had been previously identified were annotated with their appropriate dbSNP identifier (accessed April 2019). There were 12 (three in the enhancer, nine in the promoter) novel variants identified in AGVP that did not have an already established identifier. These variants were referred to by their position on chromosome 15. VEP also identified two variants which have previously been associated with pigment phenotypes. These variants have been discussed above.

Moreover, VEP generated CADD scores for all the variants that it was able. When a score of <15 is considered not likely to be pathogenic, the majority of variants identified were considered benign or likely benign. 3 variants in the enhancer region were predicted to be pathogenic or likely pathogenic. rs6497271 had a score of 20.8, which was predicted to be pathogenic, and was found in both KGP and AGVP. rs149950976 was only found in KGP and had a score of 15.76 which was considered likely pathogenic. rs769505604, which was from the AGVP data, had a score of 15.43 which was considered likely pathogenic. No variants in the promoter region for both KGP and AGVP were predicted to be pathogenic (Figure 3.11).

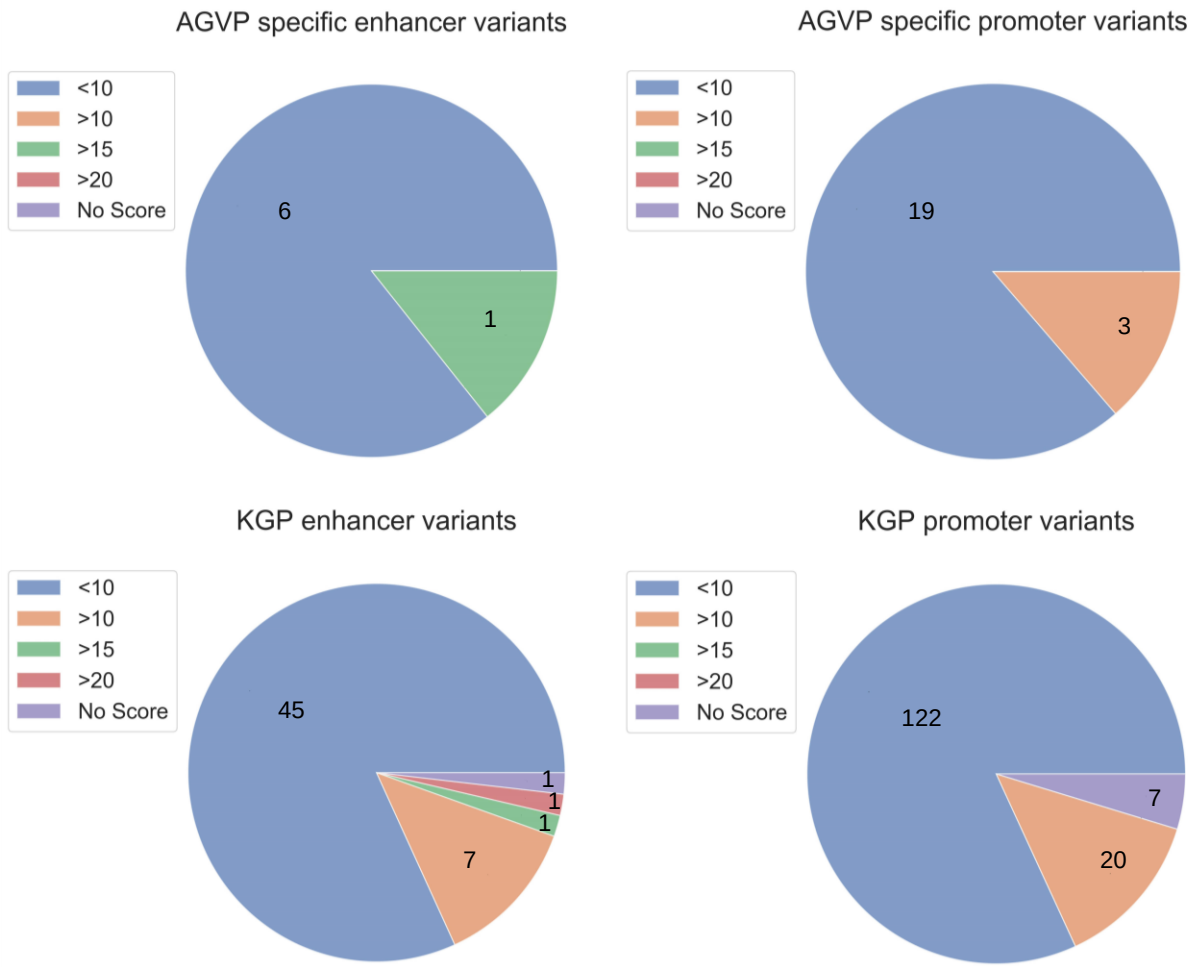



**Figure 3.11: CADD scores for the variants from KGP and AGVP promoter and enhancer by category of CADD score.** A score of <10 is benign, >10 is likely benign, >15 is likely deleterious and >20 is deleterious. The KGP variants included variants that were common between the KGP and AGVP datasets, or exclusive to KGP.

### 3.5.2 RegulomeDB

RegulomeDB (accessed May 2019) had annotations for a total of 83 variants from the dataset (Figure 3.12). Twenty-eight variants from the enhancer were present in the RegulomeDB data. Of these: 24 variants had a score of 5, 1 variant had a score of 6 and another 3 variants had a score of 7. This is considered low evidence for functionality as these variants had minimal evidence for binding.

There were 55 variants from the promoter region that were represented in RegulomeDB. Three variants had a score of 2a and eight variants were scored 2b, which are likely to affect binding. There were five variants with a score of 3a (less likely to affect binding), 30 variants with a score of 4 and nine variants which had a score of 5 (minimal binding evidence). Therefore, a

total of 16 promoter variants were predicted to have some measure of functionality by RegulomeDB.

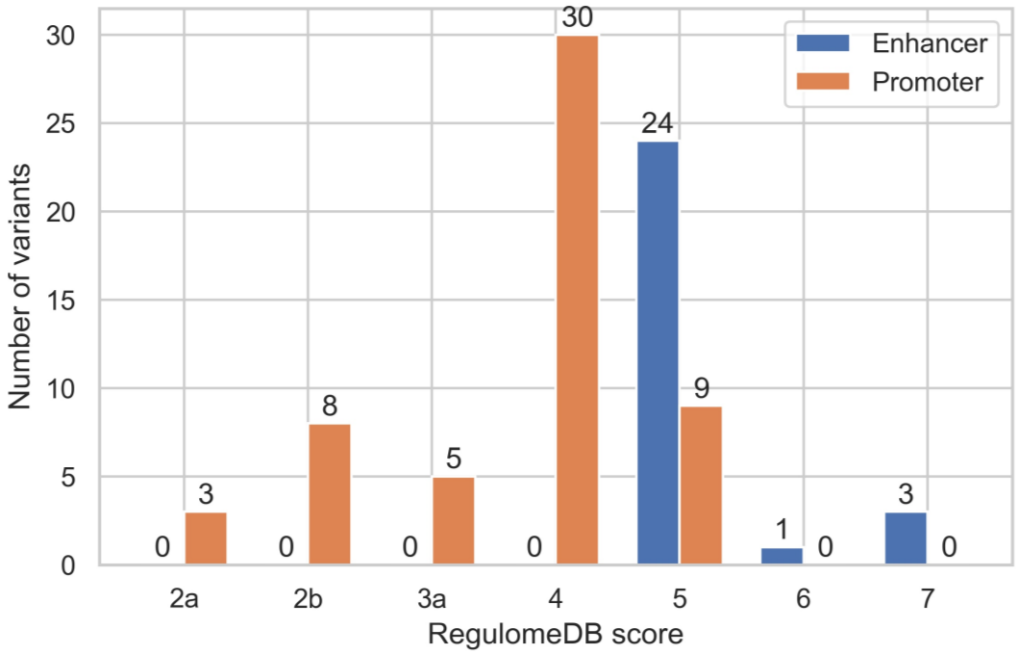



**Figure 3.12: The distribution of RegulomeDB scores in the *OCA2* enhancer and promoter regulatory regions.**

The AGVP variants that were represented in the RegulomeDB database were already identified in KGP. Therefore, no novel variants or ones that were identified in more recent projects could be annotated by RegulomeDB.

### 3.5.3 HaploReg

When the chromosomal coordinates for the *OCA2* promoter and enhancer regions were entered into HaploReg (accessed May 2019), it was able to source information for 80 variants between these two regions. In the enhancer region, 28 variants from KGP were identified. All of these variants were enriched for enhancer specific histone modifications and 20 variants coincided with DNase hypersensitivity sites in skin tissues (which included the Roadmap Epigenomics foetal penis foreskin melanocyte lines). Twenty-three variants changed a motif of a TF binding site and nine of the variants were predicted to be conserved by SiPhy and/or GERP. There were

several variants that were associated with gene expression, such that three variants had hits in GRASP QTL data and five variants had hits in eQTL databases.

There were 52 variants from KGP which were identified in the promoter region. There were four variants (rs143282714, rs187240702, rs186380226 and rs189130255) which appeared in the HaploReg and in KGP, but they did not have KGP population frequencies. These variants were not analysed, therefore, HaploReg data for 48 variants from the promoter was included for analysis. This resulted in a total of 78 variants which were annotated by HaploReg. Forty-six variants in this region were enriched for promoter histone modifications and all coincided with DNase hypersensitivity sites in various tissues. Eighteen of the variants were indicated to occur in a protein binding site and 42 variants changed a TF binding motif. One variant appeared in GRASP QTL data and five variants had hits in eQTL databases. None of the promoter variants were predicted to be conserved by SiPhy or GERP.

The variants that were annotated by HaploReg and RegulomeDB were compared to see how extensively they overlapped, as well as how many variants were annotated by a single tool. All 28 variants in the enhancer matched between the two tools. In the promoter, 48 variants appeared in both datasets. The 4 variants that did not have KGP frequencies were only identified in HaploReg (rs143282714, rs187240702, rs186380226 and rs189130255). Comparatively, seven variants were only identified by RegulomeDB (rs115656956, rs11857546, rs368326577, rs369180390, rs56140938, rs41307122 and rs373321235). Additionally, with particular reference to the enhancer variants, HaploReg appeared to have more evidence available to annotate these variants for potential functionality.

The only AGVP variants that were present in the HaploReg data had previously been identified in KGP. None of the novel AGVP variants or variants that had been identified by more recent projects such as gnomAD could be annotated by HaploReg.

### 3.5.4 GTEx

Seven variants from the *OCA2* regulatory regions were detected in the GTEx multi-tissue data for eQTL effects (accessed June 2019). In the promoter, six variants had associated eQTL information (Table 3.5), one of which was associated with *OCA2* expression (rs114037348). However, none of these associations were applicable to skin tissue.

**Table 3.5: Promoter variants with associated eQTL information from GTEx.**

| Variant | Associated gene | Tissue | p-value |
|---|---|---|---|
| rs74007950 | *RPL41P2* | Brain – Cerebellum | $2.282x10^{-5}$ |
| rs7164220 | *HERC2* | Whole blood | $6.4x10^{-11}$ |
| rs114037348* | *OCA2* | Artery - Tibial | $9.329x10^{-6}$ |
| rs7495174 | *HERC2* | Whole blood | $3.763x10^{-22}$ |
| rs7497270 | *HERC2* | Whole blood | $1.894x10^{-25}$ |
| rs7496326 | *HERC2* | Whole blood | $4.328x10^{-20}$ |

* Variant which was significantly associated with *OCA2* expression which occurs in an *OCA2* regulatory region.

When specifically looking at the single tissue cis-eQTL data in skin tissues, no variants from the promoter or enhancer region were significantly associated with gene expression in sun exposed or non-sun exposed skin. Since none of the variants of interest influenced any gene expression in skin tissue and especially since they did not affect *OCA2* expression, the search was broadened for any variants associated with *OCA2* expression in skin, both sun exposed and non-sun-exposed. GTEx identified rs1129038 in sun exposed skin and rs62007495 in non-sun exposed skin as influencing *OCA2* expression. Both variants were located at the 3' end of *HERC2*, therefore neither was located in the promoter or enhancer regions. When searching for any variants associated with *OCA2* expression in the multi-tissue data, there was a total of 22 variants associated with expression of the gene (Table 3.6). rs114037348 was once again the only variant from the *OCA2* regulatory regions which was identified. The other variants occurred in the *OCA2*, *HERC2*, and *GABRG3* genes as well as the *PDCD6IPP2* pseudogene.

**Table 3.6: Variants associated with *OCA2* expression in multi-tissue GTEx data and the gene in which they are located.**

| Variant | Gene | Tissue | p-value |
|---|---|---|---|
| rs13380044 | *GABRG3* | Artery - Tibial | $4.5x10^{-5}$ |
| rs7402428 | *GABRG3* | Oesophagus - Gastroesophageal Junction | $6.1x10^{-7}$ |
| rs17565953 | *OCA2* | Nerve - Tibial | $2.1x10^{-5}$ |
| rs8043100 | *OCA2* | Nerve - Tibial | $5.0x10^{-5}$ |
| rs78211154 | *OCA2* | Nerve - Tibial | $5.0x10^{-5}$ |
| rs1579937 | *OCA2* | Nerve - Tibial | $4.7x10^{-5}$ |
| rs1579936 | *OCA2* | Nerve - Tibial | $4.6x10^{-5}$ |
| rs2594915 | *OCA2* | Nerve - Tibial | $1.7x10^{-5}$ |
| rs2442127 | *OCA2* | Nerve - Tibial | $3.3x10^{-5}$ |
| rs2442128 | *OCA2* | Nerve - Tibial | $1.7x10^{-5}$ |
| rs2703959 | *OCA2* | Nerve - Tibial | $5.1x10^{-5}$ |
| rs1466096 | *OCA2* | Liver | $1.0x10^{-6}$ |
| rs140589534 | *OCA2* | Liver | $6.7x10^{-6}$ |
| rs72712677 | *OCA2* | Liver | $6.7x10^{-6}$ |
| rs7181038 | *OCA2* | Liver | $6.1x10^{-6}$ |
| rs114037348* | *OCA2* | Artery - Tibial | $9.329x10^{-6}$ |
| rs189052531 | *HERC2* | Adipose - Subcutaneous | $2.3x10^{-5}$ |
| rs191686746 | *HERC2* | Adipose - Subcutaneous | $2.3x10^{-5}$ |
| rs568265944 | *HERC2* | Adipose - Subcutaneous | $1.5x10^{-5}$ |
| rs537157566 | *HERC2* | Adipose - Subcutaneous | $1.5x10^{-5}$ |
| rs200476780 | *PDCD6IPP2* | Nerve - Tibial | $5.4x10^{-5}$ |
| rs34679140 | *PDCD6IPP2* | Nerve - Tibial | $5.1x10^{-5}$ |

*Variant significantly associated with *OCA2* expression which occurs in an *OCA2* regulatory region.

Following the use of the bioinformatic tools to annotate the variants of interest, it was of interest to establish which tools were the most useful in annotating non-coding variants based on the number of variants they annotated. There were 150 variants identified in KGP and AGVP which were not annotated by any of the bioinformatic tools. The tool with the most annotations was

RegulomeDB (83 variants) and then HaploReg (76 variants). GTEx annotated seven variants while atSNP was the least useful tool with a single annotation (Figure 3.13).

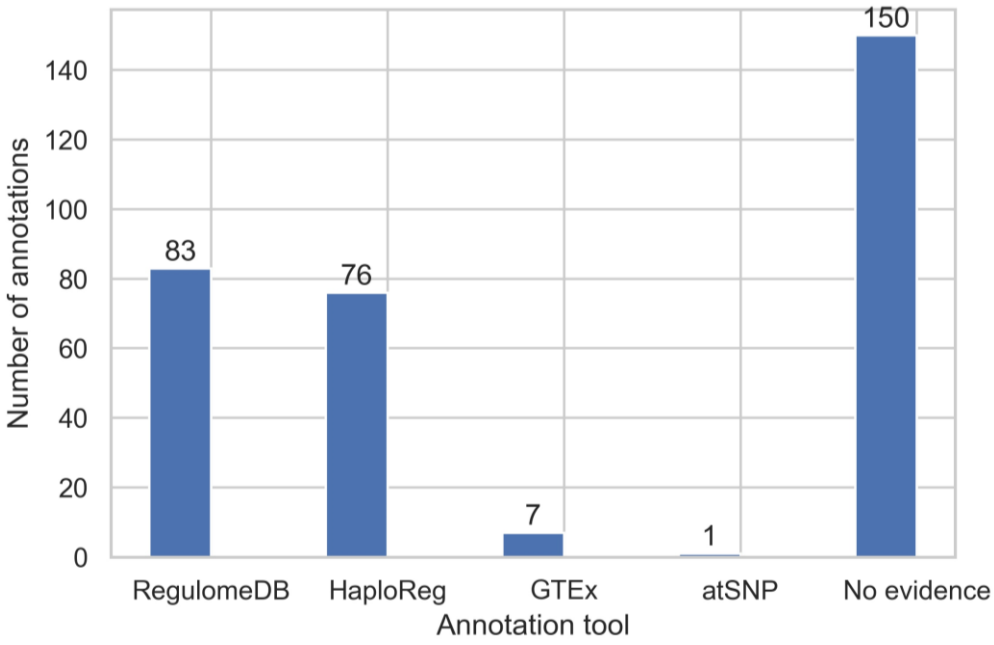



**Figure 3.13: Total number of annotations for the variants from the data set (KGP and AGVP promoter and enhancer) which were annotated by each of the tools.** Variants had no annotations or were annotated by: RegulomeDB, HaploReg, atSNP or GTEx.

# Chapter 4 – Discussion

The oculocutaneous albinism type II gene (*OCA2*) is an important pigment gene in melanogenesis as it contributes indirectly to the process of melanin production (Rosemblat et al. 1998; Toyofuku et al. 2002; Chen et al. 2002). In the case of *OCA2*, modulation of expression could have an influence on the pigmentation phenotype of the individual. Additionally, normal variation within the regulatory regions of the gene can contribute to gene expression of the *OCA2* gene and contribute to normal pigmentation. The normal variation can be eliminated from consideration when searching for pathogenic mutations for oculocutaneous albinism type II (OCA2) and its subtypes which are associated with the *OCA2* gene (Rinchik et al. 1993; Manga et al. 2001). We hypothesised that normal variation in the regulatory regions of the *OCA2* gene could influence the expression of the gene and this can be a contributing factor to the vast range of normal human pigment phenotypes. Therefore, the aim of this study was to investigate the normal variation of the *OCA2* gene control regions.

This study was motivated to characterise normal variation of the *OCA2* regulatory regions, to understand normal variation within the region and to prevent these variants from erroneously being attributed to a genetic condition. There have been cases where studies that reported on normal variation as pathogenic erroneously. For a study that looked at carrier testing for 448 severe recessive childhood genetic disorders, they found that 122 of 460 variants that had been cited by literature as disease causing were incorrectly attributed to the disease (Bell et al. 2011). Possible reasons included: the mutations were actually polymorphisms, sequencing errors or lack of evidence for pathogenicity (Bell et al. 2011). In another study on hypertrophic cardiomyopathy, five variants from the Human Genome Mutation Database had been classified as pathogenic but were misclassified. These variants had been reported to African or unspecified ancestry patients as causal variants for hypertrophic cardiomyopathy. These variants were significantly more common in the African-American population as compared to the European ancestry American population. Had this knowledge been available when the variants were associated with the disease, they probably would have been classified as benign (Manrai et al. 2016).

Normal variation is different in various populations, since pigment phenotypes can be uniquely associated with certain population ancestries, the combination of variation that leads to these normal phenotypes is likely to be unique to the population. The differences in what is

considered to be normal variation in different populations is important when looking at subtypes of OCA2 albinism such as brown oculocutaneous albinism (BOCA). An established baseline of which variants are normal in the population of interest is a logical and necessary step before trying to determine what variants are abnormal and are associated with a genetic condition.

This study used publicly available data to identify and characterise the putative enhancer for the *OCA2* gene for enhancer like properties. Chromatin interaction data was utilised to test for interactions of the *OCA2* promoter and the putative enhancer. Additionally, the Roadmap Epigenomics data was accessed to investigate if the putative enhancer was enriched for enhancer specific epigenetic marks in melanocytes. Publicly available whole genome sequence data generated by the 1000 Genomes Project (KGP) and African Genome Variation Project (AGVP) (Gurdasani et al. 2015; The 1000 Genomes Project Consortium 2015) was utilised to interrogate normal variation in the promoter and enhancer regions of the *OCA2* gene in African populations. Functional annotation of the variants was performed using bioinformatics tools to predict which variants are more likely to be functional in the regulatory regions and if they would influence the expression of the *OCA2* gene to determine a range of normal pigment phenotypes.

## 4.1 Identifying the putative enhancer for *OCA2*

A putative enhancer for *OCA2* has been identified in the neighbouring *HERC2* gene which surrounds the rs12913832 variant (Sturm et al. 2008; Visser et al. 2012). When it was characterised, it was suggested that the region has enhancer like properties in melanocytes (Visser et al. 2012). Data from other sources were used to determine if the region is indeed an active melanocyte specific enhancer, as this region could regulate the expression of the *OCA2* gene. Melanocytes were interrogated as the OCA2 protein is functional in this cell type to produce melanin. The properties of this region were investigated in other cell types from similar and more distant developmental lineages where the gene is not expected to be expressed. The data that was accessed included chromatin interaction (ChIA-PET), Hi-C and virtual chromatin conformation capture-on-chip (4C) (Li et al. 2012; Rao et al. 2014; Lajoie et al. 2015) as well as epigenetic signatures from various cell types from the Roadmap Epigenomics project (Romanoski et al. 2015).

### 4.1.1 Chromatin interaction

Transcription of protein coding genes is mediated by RNA polymerase II (POL2), which is marked by the assembly of the transcription machinery on the promoter of the gene. Many of these genes require the recruitment of an enhancer through chromatin looping for gene activation before they can be transcribed (Maston et al. 2006; Ong and Corces 2011). Thus, transcribed genes are usually marked by the interaction of the associated promoter and enhancer regions for the gene. To determine if a region of interest is an enhancer, chromatin interaction data generated by ChIA-PET can be used to determine if that region interacts with a known promoter.

Hi-C, virtual 4C (accessed April 2019) and ChIA-PET data (accessed November 2019) were utilised to test for chromatin interactions. Interactions between the putative enhancer and the *OCA2* promoter in cell types that were the most similar to melanocytes were interrogated, since no normal melanocyte cell lines were included in the project. In melanocytes from melanoma cell lines (RPMI7951) interactions between the promoter and putative enhancer regions are stronger than in normal keratinocytes (NHEK) in the Hi-C and 4C data (Figures 3.1-3.4). There were no interactions between the *OCA2* promoter and the putative enhancer regions in the ChIA-PET cell lines which included erythroleukemia cell lines (K562) and others (Appendix D). This was applicable for MCF7 breast cancer and HeLa S3 cervical cancer which, like melanocytes, were of the ectodermal developmental lineage. The same was true for K562 erythroleukemia, HCT116 colon cancer and NB4 acute promyelocytic leukemia; which were not of the of ectodermal development lineages. This indicated that the interactions between the promoter and enhancer regions were specific to melanocyte derived cell lines.

There are limitations in this data for this analysis. These cell lines are cancerous and are expected to have defective expression patterns. The studies that generated the data did not utilise any melanocyte lines, thus, it can only be inferred that these results are applicable to normal melanocytes. Additionally, the scale of values in the Hi-C and virtual 4C data were different between the RPMI7951 and NHEK cell lines. This was likely since the data was generated by groups at different institutes.

### 4.1.2 Histone modification and chromatin accessibility

Regulatory regions are defined by specific epigenetic signatures which can be used to predict if a region of interest has regulatory characteristics. H3K27ac and H3K4me1 are enhancer

associated histone modifications; where H3K4me1is associated with both active and inactive enhancers and H3K27ac identifies active enhancers. A lack of the repressive H3K27me3 mark is also characteristic of active enhancers. The region is usually also marked by DNase hypersensitivity sites which make it accessible to the transcription initiation machinery (Gross and Garrard 1988; Ong and Corces 2012; Harmston and Lenhard 2013; Ferrari et al. 2014).

Using the Roadmap Epigenomics data, in melanocyte cell lines, the putative enhancer region was enriched for the presence of H3K27ac and H3K4me1, as well as DNase hypersensitivity sites (accessed May 2019). The DNase hypersensitivity signals were highest towards the centre of this region and lower at the ends. The region was flanked by the enhancer specific histone modifications, which is a pattern of epigenetic signals associated with regulatory regions (Thurman et al. 2012). This suggests that the region of interest is an enhancer in melanocytes. The H3K27ac histone modification was present in the region which indicated that the putative enhancer region was active in melanocytes.

The H3K27ac and H3K4me1 histone modifications and DNase hypersensitivity signatures were absent in the comparative cell lines (thymus, brain and colon smooth muscle), which indicated that the enhancer is not active in those cell lines. Though the thymus and brain cell types are of a common developmental lineage with melanocytes, they did not have enhancer like characteristics. This would suggest that the presence and activity of the enhancer is melanocyte specific.

The coordinates for the enhancer were chosen to be 1 kb upstream and downstream of rs12913832, which has had the most evidence in previous literature to occur within an enhancer for the *OCA2* gene. Enhancers are typically 50 bp-1.5 kb in length (as reviewed in Blackwood and Kadonaga 1998), thus, the region investigated may overlap with the true enhancer region. These coordinates are not definitive coordinates for the enhancer which remain to be determined.

The region surrounding rs12913832 was marked by enhancer like characteristics and interacted with the *OCA2* promoter, therefore, it can be suggested that the region is an active enhancer in melanocytes. This supports previous findings by Visser and colleagues, that the region surrounding rs12913832 is a melanocyte specific active enhancer (Visser et al. 2012). There is a lack of other studies that have characterised this putative enhancer for enhancer-like properties to contest this conclusion.

## 4.2 Identifying variation in the *OCA2* promoter and enhancer from publicly available data

The promoter and enhancer coordinates for the *OCA2* gene were extracted from variant call files (VCF) from the 1000 Genomes Project (KGP) and the African Genome Variation Project (AGVP). The frequencies of the extracted variants were calculated per population and the overall frequencies of the variants were also obtained. The variants were also stratified by their minor allele frequencies (MAF) to determine which of the variants in the promoter and enhancer regions were common and could contribute to a range of pigmentation phenotypes in the African populations. A MAF of 1% was used as the threshold frequency for common variation.

In the KGP data, the majority of variants identified from the promoter were absent in the African populations. Twenty-four of these variants were common in Africa (16%), while 68 variants (46%) were polymorphic and 81 were nonpolymorphic (54%). In the enhancer region, seven variants were common in Africa (13%). There were 29 variants that were polymorphic (53%) and 26 variants that were nonpolymorphic (47%) in the African populations.

Comparatively, of the 60 promoter variants identified in AGVP, all of the variants were polymorphic and 24 of them were common (40%). Twenty-two of the enhancer variants were polymorphic (96%) and eight variants were common (35%). Most of the variants identified in AGVP were shared with KGP. In total, seven enhancer variants and 22 promoter variants were specific to AGVP, including novel variants or ones that were identified since the release of KGP. There were three AGVP enhancer variants and a further nine promoter variants that were novel. The remaining four enhancer variants and 13 promoter variants were identified in projects released after KGP, such as Trans-Omics for Precision Medicine (TOPMed) and the UK 10,000 Genomes Project (UK10K), while many of them were replicated in the Genome Aggregation Database (gnomAD) (The UK10K Consortium 2015; Taliun et al. 2019; Karczewski et al. 2019). Novel variants or ones that had been identified more recently were excluded from further analysis as no concrete evidence for their functionality could be obtained from the annotation tools, though some of the variants were common in African populations.

There are many variants within the regulatory regions of the *OCA2* gene. Of these, the majority of identified variants were not common in the African populations. In the KGP data, 54% of the promoter variants and 53% of the enhancer variants were nonpolymorphic. The KGP VCFs contained the location of variants from every population, even if they were private. In comparison, AGVP was focused on African data only, therefore a larger proportion of the

variants were polymorphic and common. While African populations share common variants with other worldwide populations, a number of common African variants are rare in other populations.

## 4.3 Functional annotation of identified variants

The identified KGP and AGVP variants were interrogated with various bioinformatic tools to annotate them and determine if they were likely to be functional in altering the expression of the *OCA2* gene.

### 4.3.1 RegulomeDB

RegulomeDB annotations were found for 55 promoter variants and 28 enhancer variants (accessed May 2019). The enhancer variants did not have strong evidence for functionality according to RegulomeDB. None of the variants had a score lower than 5, which indicated that the variants overlapped with a TF binding site or DNase hypersensitivity site. This still may be useful as regulatory regions often contain TF binding sites particular to their function and are marked by DNase hypersensitivity. Additional evidence would be required to demonstrate that these variants occurred in a a regulatory region. The promoter variants had more evidence for functionality, as they had a range of RegulomeDB scores from 2a to 5, correlating to stronger evidence to weak evidence for functionality. This correlates with the status of the promoter as a known regulatory region.

RegulomeDB is considered necessary and useful for annotating non-coding variants. However, the dataset has not been updated since 2012 (Boyle et al. 2012) and does not have access to more recent information which could further annotate the variants. This could result in some variants being scored for stronger evidence of functionality since experimental evidence from Visser and colleagues (2012) has emerged to indicate that the enhancer region is functional. Another limitation of RegulomeDB is that it does not have access to variants which have been more recently identified. Phase 1 of KGP was published in 2012 (The 1000 Genomes Project Consortium 2012) and was subsequently updated since then. Therefore, RegulomeDB cannot access variant information generated by KGP since 2012. When submitting a full list of variants to the website, an error occurs when variants from more recent versions of dbSNP were submitted. Additionally, the website will only return variants with allelic frequency >1% when the coordinates of a chromosomal region are queried. Therefore, the most efficient method of finding annotations for all variants of interest was to query the dataset manually offline.

### 4.3.2 HaploReg

HaploReg annotated 80 variants from the collective enhancer and promoter regions (accessed May 2019). The variants in both regions had evidence which suggested that the variants were associated with their respective regulatory region. HaploReg annotated approximately the same number of variants as RegulomeDB, but it seemed to have access to more information about the variants since it has been updated in 2015 (Ward and Kellis 2016). The evidence from the Roadmap Epigenomics project added annotations for the variants in melanocytes, this added value particularly in the enhancer region. This created a disagreement between HaploReg and RegulomeDB functionality scoring in the enhancer, as HaploReg indicated that the enhancer variants were more functional. owing to more recent set of information. HaploReg is still limited as it has not been updated since 2015. The dataset does not have access to variants from later phases of KGP after phase 1. The combination of HaploReg and RegulomeDB works well to annotate non-coding variants as opposed to using a single tool.

### 4.3.3 atSNP

Of all the variants submitted to atSNP, only rs12913832 variant was predicted to change transcription factor binding sites (accessed May 2019). This variant was predicted to change 95 TF binding sites, 71 gain of function sites and another 24 sites that were predicted to have loss of function effects. There was a disagreement between atSNP and HaploReg in terms of which TF binding sites were influenced by the variants of interest. HaploReg did not find any TF binding sites which rs12913832 influences, while atSNP did not identify any of the TF binding sites which HaploReg associated with the other variants. The tools seemed to use similar sources of TF binding information but it seems that HaploReg required more stringent p-values for significant association of a variants and a TF binding site (Ward and Kellis 2012; Zuo et al. 2015).

### 4.3.4 GTEx

A total of seven variants were found to have any associated eQTL information from GTEx, six variants from the promoter and one from the enhancer (accessed June 2019). There was little evidence in the GTEx dataset that the promoter and enhancer variants influenced *OCA2* expression. Only rs114037348, a promoter variant, was significantly associated with *OCA2* gene expression in tibial artery. The other variants had eQTL effects that were associated with other genes, which included *RPL41P2* and *HERC2*.

It would be expected that variants in regulatory regions of *OCA2* would be associated with expression of that gene in skin, however, no significant associations between *OCA2* regulatory region variants and gene expression were detected. The two variants that were identified to influence *OCA2* gene expression in skin occurred in the 3' UTR of *HERC2*, which is outside of the regulatory regions. When identifying variants that significantly influenced *OCA2* expression, in any tissue, several variants were detected. These variants occurred in the *OCA2*, *HERC2*, and *GABRG3* genes as well as the *PDCD6IPP2* pseudogene. However, none of the variants that were identified in the *OCA2* and *HERC2* genes were in the *OCA2* regulatory regions.

From this, it appears that *OCA2* expression is quite complicated, as several nearby genes may house regulatory regions for *OCA2*. This may apply to tissue specific expression of the gene as various nearby genes may house enhancers that control *OCA2* gene expression in other tissues. Promoters are generally able to interact with several enhancers and each of these can be specific in a small number of tissues (Maston et al. 2006; Zhu et al. 2013). In the Human Protein Atlas expression data consensus set for *OCA2*, the gene was mainly expressed in the retina and the skin, but it was expressed in other tissues at low levels such as in the pons and medulla as well as the thyroid gland (Uhlén et al. 2015). This could explain the significant influence of variants from other genes had on *OCA2* gene expression in different tissues.

A possible limitation of the GTEx data is that tissue samples from post-mortem individuals were utilised for the project (GTEx Consortium 2015). There may be differences in *OCA2* gene expression in living and post-mortem skin tissue. This would need further testing to confirm. Another issue with the GTEx website was the can only query one variant at a time and cannot search for batches of variants on the website. For efficiency, the datasets were downloaded and manually browsed offline.

Though the bioinformatic tools used for functional annotation were useful, each of them was limited in the number of variants that they were able to annotate. There were 150 variants (64%) that were not annotated by any of the tools. Additionally, variants which have been identified more recently will not have been tested in the large projects and will not have functional annotations. The most useful tools were RegulomeDB and HaploReg. These tools had a large amount of information available to annotate variants and had many of the variants of interest appeared in their databases. Should these databases be updated with the most recent

information, they would likely be of greater assistance in annotating non-coding variation, which is generally more difficult to annotate than coding variation.

## 4.4 Evaluation of the variants with the most evidence for functionality

The variants that were most likely to be functional were selected based on how well they were annotated. Variants that appeared in both KGP and AGVP were filtered while variants that did not have enough evidence were removed from further analysis. There is partial evidence to suggest that these variants may be functional, though some variants may have stronger evidence than others. The set of variants with the strongest evidence for functionality were comprised of seven variants from the promoter and 10 from the enhancer. These variants will be discussed in terms of how they have been described in literature and from this study data. Since 11 of these variants have not been described in the literature, only the variants that had evidence from the literature will be discussed in detail here as they have a stronger indication of functionality. The variants that will not be described from the enhancer are: rs11074319, rs6497272, rs8034648, rs7496724, rs115138127, rs188973574 and rs80302374. From the promoter, the variants that will not be discussed are: rs7496326, rs74007950, rs114037348 and rs7168800.

**Promoter variants with the most functional evidence**

Variant rs7495174 was the variant from the promoter with the strongest evidence of functionality. The variant coincided with epigenetic signatures of promoter, protein binding sites and changed several transcription factor (TF) binding sites. The variant also has been indicated to influence *HERC2* expression and *OCA2* expression in lung tissue (Table 3.3). rs7495174 has been significantly associated with several hair colour phenotypes (blonde vs non-blonde, brown vs non-brown, light vs dark, red vs non-red) which included a twin study in the Netherlands where the variant was associated with brown, blonde and light hair colours (Lin et al. 2015). The rs7495174 was deemed significant for blonde, brown and black hair colour in a study looking at hair colour in European ancestry populations (Han et al. 2008).

rs7495174 has also been associated with eye colour. A haplotype of rs7495174 with rs4778241 (formerly known as rs6497268) and rs4778138 (formerly known as rs11855019), was considered a significant predictor for blue or non-blue eye colour (Duffy et al. 2007; Kayser et al. 2008). The homozygous rs7495174(T)-rs6497268(G)-rs11855019(T) haplotype was more frequent in individuals with lighter pigmented skin, hair and eyes (Duffy et al. 2007); this haplotype now corresponds to rs4778138(A)-rs4778241(C)-rs7495174(A) using the dbSNP

forward strand orientation alleles. Though rs4778241 and rs4778138 occurred in the first intron of *OCA2*, neither occurred in the promoter region defined in this study, so their functional impact on *OCA2* expression was not considered. The rs7495174 variant may also influence skin pigmentation as lightly pigmented melanocyte cell lines carried an A allele while the G allele was present in the darkly pigmented melanocytes (Visser et al. 2012). The rs7495174 A allele is the major allele in all populations, in Africa where skin colour is typically darker, as well as in Europe where skin colour is generally lighter. The alternate G allele is common in all the KGP and AGVP populations though it is most common in East Asia (Figure 3.8). While the variant MAF may not necessarily correlate with worldwide skin pigmentation phenotypes, it has been significantly associated with hair colour and may be a marker of dark hair colour. This variant has evidence to indicate that it is functional and may influence *OCA2* expression.

rs7164220 was likely to be functional based on the annotations collected from various bioinformatic tools. It coincides with epigenetic signatures of promoters and its alternate allele alters TF binding sites. The variant has also been associated with expression of *HERC2* in whole blood (Table 3.3). In experimental evidence, both darkly pigmented and lightly pigmented melanocytes shared the C allele. This variant was predicted to be conserved in vertebrates and is in linkage disequilibrium (LD, $r^2$>0.8) with rs7495174 (Visser et al. 2012). The C allele of rs7164220 is the minor allele and is common in all populations, though it is most common in East Asia (Figure 3.8). rs7164220 is likely to be functional but it does not have not a clear association with differing pigmentation phenotypes.

rs7497270 overlapped with epigenetic signatures of promoters, protein binding sites and changed several TF binding sites. The variant also has been shown to influence *HERC2* expression in whole blood and lymphoblastoid cell lines (Table 3.3). In differently pigmented melanocytes, the genotype was homozygous C in lightly pigmented melanocytes but heterozygous CT in darkly pigmented melanocytes. This variant is also in LD ($r^2$>0.8) with rs7495174 (Visser et al. 2012). In the KGP and AGVP data, C is the major allele and is common in all populations, though the minor T allele is also common in all populations but more so in East Asia (Figure 3.8). As with rs7164220, rs7497270 is likely functional but does not associate clearly with different pigmentation phenotypes. This is likely since these variants are in LD with rs7495174, which has been associated with normal pigmentation phenotypes.

**Enhancer variants with the most functional evidence**

rs12913832 was the variant in the enhancer region which had the most evidence of functionality. The variant was predicted to be marked by H3K27ac and H3K4me1 which are characteristic epigenetic signatures of enhancers. It was also predicted to be conserved across species and to alter 95 TF binding sites. The variant was also predicted to influence *HERC2* expression in lymphoblastoid cell lines and whole blood as well as choline kinase beta (*CHKB*) expression (Table 3.4). In the literature, rs12913832 has been associated with hair, skin and eye pigment phenotypes and was the main reason why the surrounding region was thought to be an enhancer. rs12913832 has been associated with skin colour, freckling and tanning ability in Polish individuals; where the G allele was more frequent in paler individuals, who tan poorly, were sunburned badly and had freckles (Zaorska et al. 2019). This replicates the findings of Visser and colleagues where the G allele was associated with light pigmentation and A was associated with a dark pigmentation phenotype (Visser et al. 2012). The frequency of these alleles was consistent in KGP and AGVP, such that the G allele is highly common in European populations (63.6% in EUR KGP) but is absent in African populations (Figure 3.9). In South Asian individuals, the variant associated with eye colour, where a GG genotype was associated with green eye colour, which also suggests that lighter eye colour is modulated by other variants in South Asians as compared to Europeans where it is a determinant of blue eye colour (Sturm et al. 2008; Edwards et al. 2016; Jonnalagadda et al. 2019). The predictive ability of rs12913832 for pigmentation phenotypes has been demonstrated by its inclusion in a set of informative variants for forensic prediction of eye and hair colour in several assays (Walsh et al. 2012, 2013). This variant has a wealth of information which indicate its functionality and it may modulate the enhancer function to determine expression of the *OCA2* gene, though this has not been indicated in eQTL experiments.

rs6497271 was predicted to occur within an enhancer region, to overlap with protein binding sites and alter several TF binding sites. The variant has also been predicted to influence expression of *HERC2* in lymphoblastoid cell lines and to be conserved across different species (Table 3.4). This variant has been associated with skin colour in European ancestry populations (Visser et al. 2012) and occurred in a lymphoid enhancer-binding factor 1 (LEF1) binding site which is further evidence of its functionality (Sturm et al. 2008). LEF1 is a transcription factor that has been indicated to interact with microphthalmia-associated transcription factor (MITF) in activating the dopachrome tautomerase (*DCT*) gene, which functions in the melanogenesis pathway (Yasumoto et al. 2002). Further evidence indicated that rs6497271 occurred in a SOX2

consensus motif which modulated MITF levels in melanocytes (Crawford et al. 2017). The ancestral A allele was detected in haplotypes from South Asians and Australo-Melanesians, where these haplotypes were associated with dark pigmentation. These haplotypes were similar or identical to those in Africans, which could suggest that the haplotypes including rs6497271 originated in Africa and were carried by the human ancestors that settled in Asia. The derived G allele was associated with light skin pigmentation and was most common in Europeans and San populations (Crawford et al. 2017). The G allele of rs6497271 is common in all KGP and AGVP populations, though it occurs at a higher frequency in non-African populations, echoing the pigmentation phenotype association seen by Crawford and colleagues (Figure 3.9).

rs7183877 occurred within enhancer specific epigenetic signatures and altered a TF binding site. This variant was also associated with *HERC2* expression in lymphoblastoid cell lines. These aspects seem to indicate that the variant is functional, though it was predicted to be non-functional by GWAVA (Table 3.4). GWAVA did not utilise experimental evidence from the Roadmap Epigenomics project for its prediction of functionality (Ritchie et al. 2014), thus the variant was predicted to be non-functional. This was despite the experimental evidence from the Roadmap Epigenomics project and literature to suggest that rs7183877 was functional in an enhancer region. In the literature, rs7183877 was associated with hair colour in European ancestry populations. The variant has also been termed an eye colour predictor (Eriksson et al. 2010), and it has improved the prediction of green-hazel eye colour alongside other variants that are predictive for eye colour (Han et al. 2008; Ruiz et al. 2013). Both lightly and darkly pigmented melanocyte samples had a G allele at this variant but it was more enriched in melanocytes than in MCF7 breast cancer cells (Visser et al. 2012). The G allele is the major allele in KGP and AGVP populations, though the A allele is still common in all populations. The A allele is most common in non-African KGP populations (Figure 3.9). From this collected evidence, rs7183877 seemed to be associated with pigmentation phenotypes and is likely to be functional.

The approach of identifying regulatory regions through gene expression patterns and enhancer characteristics has been seen in the association of the sulfite oxidase (*SUOX*) gene with vitiligo in skin (Qi et al. 2018). This study used gene expression data from GTEx to identify tissue specific expression of the gene and followed up by checking for histone modifications for regulatory regions, which indicated that the region of interest was an enhancer. A functional CRISPR interference study confirmed that this region influenced *SUOX* expression. This indicates that a bioinformatic predictive approach can be useful in detecting regulatory regions.

Overall, the variants mentioned had the strongest supporting evidence from the annotation tools as well as from literature and were the most likely to be functional in the promoter and enhancer regions, though the supporting evidence varied. Some variants had stronger still evidence of functionality and were more likely to influence *OCA2* gene expression and therefore potentially change pigmentation. The variants with the strongest evidence were rs12913832 from the enhancer and rs7495174 from the promoter. These variants are likely to influence *OCA2* expression and their alleles seem to associate with different pigment phenotypes in different populations. This can point to the role of regulatory regions of pigment genes in determining a normal pigment phenotype as well as to abnormal pigment phenotypes such as those of albinism when gene expression is decreased. The regulatory regions of pigment genes such as *OCA2* can then be considered a good place to search for causal variants for albinism and interesting subtypes such as BOCA.

**Conclusions**

A putative enhancer region for the *OCA2* gene has previously been identified. It was confirmed using publicly available datasets that the region was marked by characteristics of enhancer regions and was active in melanocytes specifically. Within this enhancer region as well as the promoter, there were variants in these regulatory regions which were likely to be functional for pigmentation. Variants in these regions occurred at different frequencies between African and non-African populations, such that different alleles for the same variants could associate with differential pigment phenotypes (such as one with light pigment and another with dark pigment). The variants which had the strongest evidence from the collection of annotations could be predicted to be potentially functional and to modulate the action of their respective regulatory region. This meant that they could influence *OCA2* gene expression. Of these, rs7495174 and rs12913832 had the strongest evidence of functionality in the promoter and enhancer respectively. Though novel variants have been identified in individuals of African ancestry in the *OCA2* regulatory regions, more information is still required to annotate these variants or variants that have not yet been assessed by expression based methods.

**Limitations of the study**

- A limitation of this study was the low coverage of the whole genome sequences in the KGP and AGVP data. Low coverage sequencing can identify common variation, but cannot reliably identify rare variation, both of which are important when investigating

normal variation. No clinical or laboratory analysis of expression was available for these samples.

- The annotation tools are largely limited as they have not been updated recently and with the most recent version of dbSNP, therefore many variants cannot be annotated. This is particularly important when investigating variation in the African genome, as it is highly variable and novel variants may be found. However, these novel variants will not be annotated by the standard tools. Therefore, their functions may require functional studies to be confirmed.

**Future direction**

- The bioinformatics tools utilised in this study can be incorporated into an annotation pipeline for non-coding variation for applications such as whole genome sequencing or genome wide association studies.
- The characterisation of normal variation in the *OCA2* regulatory regions will allow for further interrogation of causal mutations for OCA2 albinism and its subtypes such as BOCA. This provides a basis for removal of false positives and could assist in identifying the true causal mutations for the disorders.
- Functional studies can be performed to determine if the variants truly influence *OCA2* expression and contribute to the normal range of pigment phenotypes.
- In addition, this study has reiterated that the African genome is highly variable and one African population cannot be used as a proxy for the rest. The normal variation will not be constant between African populations; therefore, it would be necessary to identify variation in many African populations in the event of searching for population specific causal mutations for albinism.

# References


Abdel-Malek Z, Swope VB, Suzuki I, et al (1995) Mitogenic and melanogenic stimulation of normal human melanocytes by melanotropic peptides. Proc Natl Acad Sci U S A 92:1789–1793

Adzhubei IA, Schmidt S, Peshkin L, et al (2010) A method and server for predicting damaging missense mutations. Nat Methods 7:248–249. https://doi.org/10.1038/nmeth0410-248

Akahoshi K, Fukai K, Kato A, et al (2001) Duplication of 15q11.2–q14, including the P gene, in a woman with generalized skin hyperpigmentation. Am J Med Genet 104:299–302

Akahoshi K, Spritz RA, Fukai K, et al (2004) Mosaic supernumerary inv dup(15) chromosome with four copies of the P gene in a boy with pigmentary dysplasia. Am J Med Genet A 126A:290–292. https://doi.org/10.1002/ajmg.a.20580

Andersen JD, Pietroni C, Johansen P, et al (2016) Importance of nonsynonymous OCA2 variants in human eye color prediction. Mol Genet Genomic Med 4:420–430. https://doi.org/10.1002/mgg3.213

Barski A, Cuddapah S, Cui K, et al (2007) High-Resolution Profiling of Histone Methylations in the Human Genome. Cell 129:823–837. https://doi.org/10.1016/j.cell.2007.05.009

Bassi MT, Schiaffino MV, Renieri A, et al (1995) Cloning of the gene for ocular albinism type 1 from the distal short arm of the X chromosome. Nat Genet 10:13–19. https://doi.org/10.1038/ng0595-13

Bekker-Jensen S, Danielsen JR, Fugger K, et al (2010) HERC2 coordinates ubiquitin-dependent assembly of DNA repair factors on damaged chromosomes. Nat Cell Biol 12:80–86. https://doi.org/10.1038/ncb2008

Bell CJ, Dinwiddie DL, Miller NA, et al (2011) Carrier Testing for Severe Childhood Recessive Diseases by Next-Generation Sequencing. Sci Transl Med 3:65ra4. https://doi.org/10.1126/scitranslmed.3001756

Bernstein BE, Stamatoyannopoulos JA, Costello JF, et al (2010) The NIH Roadmap Epigenomics Mapping Consortium. Nat Biotechnol 28:1045–1048. https://doi.org/10.1038/nbt1010-1045

Bin B-H, Bhin J, Yang SH, et al (2015) Membrane-Associated Transporter Protein (MATP) Regulates Melanosomal pH and Influences Tyrosinase Activity. PLoS ONE 10:. https://doi.org/10.1371/journal.pone.0129273

Bivik CA, Larsson PK, Kågedal KM, et al (2006) UVA/B-Induced Apoptosis in Human Melanocytes Involves Translocation of Cathepsins and Bcl-2 Family Members. J Invest Dermatol 126:1119–1127. https://doi.org/10.1038/sj.jid.5700124

Blackwood EM, Kadonaga JT (1998) Going the Distance: A Current View of Enhancer Action. Science 281:60–63. https://doi.org/10.1126/science.281.5373.60

Böhm M, Wolff I, Scholzen TE, et al (2005) alpha-Melanocyte-stimulating hormone protects from ultraviolet radiation-induced apoptosis and DNA damage. J Biol Chem 280:5795–5802. https://doi.org/10.1074/jbc.M406334200

Boissy RE, Zhao H, Oetting WS, et al (1996) Mutation in and lack of expression of tyrosinase-related protein-1 (TRP-1) in melanocytes from an individual with brown oculocutaneous albinism: a new subtype of albinism classified as "OCA3". Am J Hum Genet 58:1145–1156

Bouchard B, Del Marmol V, Jackson IJ, et al (1994) Molecular characterization of a human tyrosinase-related-protein-2 cDNA. Patterns of expression in melanocytic cells. Eur J Biochem 219:127–134. https://doi.org/10.1111/j.1432-1033.1994.tb19922.x

Boyle AP, Davis S, Shulha HP, et al (2008) High-Resolution Mapping and Characterization of Open Chromatin across the Genome. Cell 132:311–322. https://doi.org/10.1016/j.cell.2007.12.014

Boyle AP, Hong EL, Hariharan M, et al (2012) Annotation of functional variation in personal genomes using RegulomeDB. Genome Res 22:1790–1797

Branicki W, Brudnik U, Wojas-Pelc A (2009) Interactions Between HERC2, OCA2 and MC1R May Influence Human Pigmentation Phenotype. Ann Hum Genet 73:160–170

Brenner M, Hearing VJ (2008) The Protective Role of Melanin Against UV Damage in Human Skin. Photochem Photobiol 84:539–549. https://doi.org/10.1111/j.1751-1097.2007.00226.x

Bryne JC, Valen E, Tang M-HE, et al (2008) JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. Nucleic Acids Res 36:D102–D106. https://doi.org/10.1093/nar/gkm955

Butler MG, Christian SL, Kubota T, Ledbetter DH (1996) A 5-Year-Old White Girl With Prader-Willi Syndrome and a Submicroscopic Deletion of Chromosome 15q11q13. Am J Med Genet 65:137–141. https://doi.org/10.1002/(SICI)1096-8628(19961016)65:2<137::AID-AJMG11>3.0.CO;2-R

Cesarini J-P (1988) Photo-Induced Events in the Human Melanocytic System: Photoaggression and Photoprotection. Pigment Cell Res 1:223–233. https://doi.org/10.1111/j.1600-0749.1988.tb00420.x

Chang CC, Chow CC, Tellier LC, et al (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience 4:. https://doi.org/10.1186/s13742-015-0047-8

Chen K, Manga P, Orlow SJ (2002) Pink-eyed Dilution Protein Controls the Processing of Tyrosinase. Mol Biol Cell 13:1953–1964. https://doi.org/10.1091/mbc.02-02-0022.

Cichorek M, Wachulska M, Stasiewicz A, Tymińska A (2013) Skin melanocytes: biology and development. Adv Dermatol Allergol Dermatol Alergol 30:30–41. https://doi.org/10.5114/pdia.2013.33376

Cook AL, Chen W, Thurber AE, et al (2009) Analysis of Cultured Human Melanocytes Based on Polymorphisms within the SLC45A2/MATP, SLC24A5/NCKX5, and OCA2/P Loci. J Invest Dermatol 129:392–405

Cooper GM, Stone EA, Asimenos G, et al (2005) Distribution and intensity of constraint in mammalian genomic sequence. Genome Res 15:901–913. https://doi.org/10.1101/gr.3577405

Crawford NG, Kelly DE, Hansen MEB, et al (2017) Loci associated with skin pigmentation identified in African populations. Science 358:eaan8433. https://doi.org/10.1126/science.aan8433

Creel DJ, Bendel CM, Wiesner GL, et al (1986) Abnormalities of the Central Visual Pathways in Prader–Willi Syndrome Associated with Hypopigmentation. N Engl J Med 314:1606–1609. https://doi.org/10.1056/NEJM198606193142503

De Santa Barbara P, Van Den Brink GR, Roberts DJ (2003) Development and differentiation of the intestinal epithelium. Cell Mol Life Sci 60:1322–1332. https://doi.org/10.1007/s00018-003-2289-3

Dekker J, Marti-Renom MA, Mirny LA (2013) Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nat Rev Genet 14:390–403. https://doi.org/10.1038/nrg3454

Dekker J, Rippe K, Dekker M, Kleckner N (2002) Capturing Chromosome Conformation. Science 295:1306–1311. https://doi.org/10.1126/science.1067799

Donnelly MP, Paschou P, Grigorenko E, et al (2012) A global view of the OCA2-HERC2 region and pigmentation. Hum Genet 131:683–696

Duffy DL, Montgomery GW, Chen W, et al (2007) A Three–Single-Nucleotide Polymorphism Haplotype in Intron 1 of OCA2 Explains Most Human Eye-Color Variation. Am J Hum Genet 80:241–252

Dufourcq-Lagelouse R, Lambert N, Duval M, et al (1999) Chediak-Higashi syndrome associated with maternal uniparental isodisomy of chromosome 1. Eur J Hum Genet 7:633–637. https://doi.org/10.1038/sj.ejhg.5200355

Durham-Pierre D, Gardner JM, Nakatsu Y, et al (1994) African origin of an intragenic deletion of the human P gene in tyrosinase positive oculocutaneous albinism. Nat Genet 7:176–179

Edwards M, Cha D, Krithika S, et al (2016) Iris pigmentation as a quantitative trait: Variation in populations of European, East Asian and South Asian ancestry and association with candidate gene polymorphisms. Pigment Cell Melanoma Res 29:141–162. https://doi.org/10.1111/pcmr.12435

Eiberg H, Mohr J (1996) Assignment of genes coding for brown eye colour (BEY2) and brown hair colour (HCL3) on chromosome 15q. Eur J Hum Genet EJHG 4:237–241

Eller MS, Gilchrest BA (2000) Tanning as Part of the Eukaryotic SOS Response. Pigment Cell Res 13:94–97. https://doi.org/10.1034/j.1600-0749.13.s8.17.x

ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74

Eriksson N, Macpherson JM, Tung JY, et al (2010) Web-based, participant-driven studies yield novel genetic associations for common traits. PLoS Genet 6:e1000993. https://doi.org/10.1371/journal.pgen.1000993

Ernst J, Kellis M (2012) ChromHMM: automating chromatin state discovery and characterization. Nat Methods 9:215–216. https://doi.org/10.1038/nmeth.1906

Ernst J, Kellis M (2015) Large-scale epigenome imputation improves data quality and disease variant enrichment. Nat Biotechnol 33:364–376. https://doi.org/10.1038/nbt.3157

Farage MA, Miller KW, Elsner P, Maibach HI (2007) Structural Characteristics of the Aging Skin: A Review. Cutan Ocul Toxicol 26:343–357. https://doi.org/10.1080/15569520701622951

Ferrari KJ, Scelfo A, Jammula S, et al (2014) Polycomb-Dependent H3K27me1 and H3K27me2 Regulate Active Transcription and Enhancer Fidelity. Mol Cell 53:49–62. https://doi.org/10.1016/j.molcel.2013.10.030

Fryburg JS, Breg WR, Lindgren V (1991) Diagnosis of Angelman syndrome in infants. Am J Med Genet 38:58–64. https://doi.org/10.1002/ajmg.1320380114

Fullwood MJ, Liu MH, Pan YF, et al (2009) An Oestrogen Receptor α-bound Human Chromatin Interactome. Nature 462:58–64. https://doi.org/10.1038/nature08497

Garber M, Guttman M, Clamp M, et al (2009) Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics 25:i54–i62. https://doi.org/10.1093/bioinformatics/btp190

García-Borrón JC, Abdel-Malek Z, Jiménez-Cervantes C (2014) MC1R, the cAMP pathway and the response to solar UV: Extending the horizon beyond pigmentation. Pigment Cell Melanoma Res 27:699–720. https://doi.org/10.1111/pcmr.12257

Gardner JM, Nakatsu Y, Gondo Y, et al (1992) The mouse pink-eyed dilution gene: association with human Prader-Willi and Angelman syndromes. Science 257:1121–1124. https://doi.org/10.1126/science.257.5073.1121

Giebel LB, Strunk KM, King RA, et al (1990) A frequent tyrosinase gene mutation in classic, tyrosinase-negative (type IA) oculocutaneous albinism. Proc Natl Acad Sci 87:3255–3258

Gilchrest BA, Eller MS (1999) DNA Photodamage Stimulates Melanogenesis and Other Photoprotective Responses. J Investig Dermatol Symp Proc 4:35–40. https://doi.org/10.1038/sj.jidsp.5640178

Ginger RS, Askew SE, Ogborne RM, et al (2008) SLC24A5 Encodes a trans-Golgi Network Protein with Potassium-dependent Sodium-Calcium Exchange Activity That Regulates Human Epidermal Melanogenesis. J Biol Chem 283:5486–5495. https://doi.org/10.1074/jbc.M707521200

Giresi PG, Kim J, McDaniell RM, et al (2007) FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome Res 17:877–885. https://doi.org/10.1101/gr.5533506

Grønskov K, Dooley CM, Østergaard E, et al (2013) Mutations in C10orf11, a Melanocyte-Differentiation Gene, Cause Autosomal-Recessive Albinism. Am J Hum Genet 92:415–421. https://doi.org/10.1016/j.ajhg.2013.01.006

Gross DS, Garrard WT (1988) Nuclease Hypersensitive Sites in Chromatin. Annu Rev Biochem 57:159–197. https://doi.org/10.1146/annurev.bi.57.070188.001111

GTEx Consortium (2015) The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science 348:648–660. https://doi.org/10.1126/science.1262110

Gurdasani D, Carstensen T, Tekola-Ayele F, et al (2015) The African Genome Variation Project shapes medical genetics in Africa. Nature 517:327–332. https://doi.org/10.1038/nature13997

Han J, Kraft P, Nan H, et al (2008) A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation. Plos Genet 4:e1000074. https://doi.org/10.1371/journal.pgen.1000074

Hao K, Bossé Y, Nickle DC, et al (2012) Lung eQTLs to Help Reveal the Molecular Underpinnings of Asthma. PLoS Genet 8:. https://doi.org/10.1371/journal.pgen.1003029

Harmston N, Lenhard B (2013) Chromatin and epigenetic features of long-range gene regulation. Nucleic Acids Res 41:7185–7199. https://doi.org/10.1093/nar/gkt499

Harsanyl ZP, Post PW, Brinkmann JP, et al (1980) Mutagenicity of melanin from human red hair. Experientia 36:291–292. https://doi.org/10.1007/BF01952282

Hill HZ, Hill GJ (2000) UVA, Pheomelanin and the Carcinogenesis of Melanoma. Pigment Cell Res 13:140–144. https://doi.org/10.1034/j.1600-0749.13.s8.25.x

Ito S, Nakanishi Y, Valenzuela RK, et al (2011) Usefulness of alkaline hydrogen peroxide oxidation to analyze eumelanin and pheomelanin in various tissue samples: application to chemical analysis of human hair melanins. Pigment Cell Melanoma Res 24:605–613. https://doi.org/10.1111/j.1755-148X.2011.00864.x

Ito S, Wakamatsu K (2011) Diversity of human hair pigmentation as studied by chemical analysis of eumelanin and pheomelanin. J Eur Acad Dermatol Venereol 25:1369–1380. https://doi.org/10.1111/j.1468-3083.2011.04278.x

Jing R, Dong X, Li K, et al (2014) Two distinct phenotypes in pigmented cells of different embryonic origins in eyes of pale ear mice. Exp Eye Res 119:35–43. https://doi.org/10.1016/j.exer.2013.12.007

Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-Wide Mapping of in Vivo Protein-DNA Interactions. Science 316:1497–1502. https://doi.org/10.1126/science.1141319

Jonnalagadda M, Faizan MA, Ozarkar S, et al (2019) A Genome-Wide Association Study of Skin and Iris Pigmentation among Individuals of South Asian Ancestry. Genome Biol Evol 11:1066–1076. https://doi.org/10.1093/gbe/evz057

Kadekaro AL, Kavanagh R, Kanto H, et al (2005) alpha-Melanocortin and endothelin-1 activate antiapoptotic pathways and reduce DNA damage in human melanocytes. Cancer Res 65:4292–4299. https://doi.org/10.1158/0008-5472.CAN-04-4535

Kaidbey KH, Agin PP, Sayre RM, Kligman AM (1979) Photoprotection by melanin—a comparison of black and Caucasian skin. J Am Acad Dermatol 1:249–260. https://doi.org/10.1016/S0190-9622(79)70018-1

Karczewski KJ, Francioli LC, Tiao G, et al (2019) Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. bioRxiv 531210. https://doi.org/10.1101/531210

Kausar T, Bhatti MA, Ali M, et al (2013) OCA5, a novel locus for non-syndromic oculocutaneous albinism, maps to chromosome 4q24. Clin Genet 84:91–93. https://doi.org/10.1111/cge.12019

Kayser M, Liu F, Janssens ACJW, et al (2008) Three genome-wide association studies and a linkage analysis identify HERC2 as a human iris color gene. Am J Hum Genet 82:411–423. https://doi.org/10.1016/j.ajhg.2007.10.003

Kerr R, Stevens G, Manga P, et al (2000) Identification of P gene mutations in individuals with oculocutaneous albinism in sub-Saharan Africa. Hum Mutat 15:166–172

Kheradpour P, Kellis M (2014) Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. Nucleic Acids Res 42:2976–2987. https://doi.org/10.1093/nar/gkt1249

King RA, Creel D, Arvenka J, et al (1980) Albinism in Nigeria with delineation of new recessive oculocutaneous type. Clin Genet 17:259–270

King RA, Rich SS (1986) Segregation analysis of brown oculocutaneous albinism. Clin Genet 29:496–501

Kobayashi N, Muramatsu T, Yamashina Y, et al (1993) Melanin Reduces Ultraviolet-Induced DNA Damage Formation and Killing Rate in Cultured Human Melanoma Cells. J Invest Dermatol 101:685–689. https://doi.org/10.1111/1523-1747.ep12371676

Kobayashi T, Urabe K, Winder A, et al (1994) Tyrosinase related protein 1 (TRP1) functions as a DHICA oxidase in melanin biosynthesis. EMBO J 13:5818–5825

Kraoua L, Chaabouni M, Ewers E, et al (2011) Hexasomy of the Prader–Willi/Angelman critical region, including the OCA2 gene, in a patient with pigmentary dysplasia: Case report. Eur J Med Genet 54:e446–e450. https://doi.org/10.1016/j.ejmg.2011.04.007

Kromberg JGR, Castle D, Zwane EM, Jenkins T (1989) Albinism and skin cancer in Southern Africa. Clin Genet 36:43–52. https://doi.org/10.1111/j.1399-0004.1989.tb03365.x

Kromberg JGR, Jenkins T (1982) Prevalence of albinism in the South African negro. S Afr Med J 61:383–386

Kühnle S, Kogel U, Glockzin S, et al (2011) Physical and Functional Interaction of the HECT Ubiquitin-protein Ligases E6AP and HERC2. J Biol Chem 286:19410–19416. https://doi.org/10.1074/jbc.M110.205211

Kumar P, Henikoff S, Ng PC (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 4:1073–1081. https://doi.org/10.1038/nprot.2009.86

Lajoie BR, Dekker J, Kaplan N (2015) The Hitchhiker's Guide to Hi-C Analysis: Practical guidelines. Methods San Diego Calif 72:65–75. https://doi.org/10.1016/j.ymeth.2014.10.031

Lappalainen T, Sammeth M, Friedländer MR, et al (2013) Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501:506–511. https://doi.org/10.1038/nature12531

Lee ST, Nicholls RD, Bundey S, et al (1994) Mutations of the P gene in oculocutaneous albinism, ocular albinism, and Prader-Willi syndrome plus albinism. N Engl J Med 330:529–534. https://doi.org/10.1056/NEJM199402243300803

Lee S-T, Nicholls RD, Jong MTC, et al (1995) Organization and sequence of the human P gene and identification of a new family of transport proteins. Genomics 26:354–363. https://doi.org/10.1016/0888-7543(95)80220-G

Lerner AB, Fitzpatrick TB, Calkins E, Summerson WH (1949) Mammalian Tyrosinase: Preparation and Properties. J Biol Chem 178:185–195

Leslie R, O'Donnell CJ, Johnson AD (2014) GRASP: analysis of genotype–phenotype results from 1390 genome-wide association studies and corresponding open access database. Bioinformatics 30:i185–i194. https://doi.org/10.1093/bioinformatics/btu273

Li G, Ruan X, Auerbach RK, et al (2012) Extensive Promoter-centered Chromatin Interactions Provide a Topological Basis for Transcription Regulation. Cell 148:84–98. https://doi.org/10.1016/j.cell.2011.12.014

Li Q, Stram A, Chen C, et al (2014) Expression QTL-based analyses reveal candidate causal genes and loci across five tumor types. Hum Mol Genet 23:5294–5302. https://doi.org/10.1093/hmg/ddu228

Lieberman-Aiden E, van Berkum NL, Williams L, et al (2009) Comprehensive mapping of long range interactions reveals folding principles of the human genome. Science 326:289–293. https://doi.org/10.1126/science.1181369

Lin BD, Mbarek H, Willemsen G, et al (2015) Heritability and Genome-Wide Association Studies for Hair Color in a Dutch Twin Family Based Sample. Genes 6:559–576. https://doi.org/10.3390/genes6030559

Luande J, Henschke CI, Mohammed N (1985) The Tanzanian human albino skin. Natural history. Cancer 55:1823–1828. https://doi.org/10.1002/1097-0142(19850415)55:8<1823::aid-cncr2820550830>3.0.co;2-x

Manga P, Kromberg JG, Box NF, et al (1997) Rufous oculocutaneous albinism in southern African Blacks is caused by mutations in the TYRP1 gene. Am J Hum Genet 61:1095–1101

Manga P, Kromberg JGR, Turner A, et al (2001) In Southern Africa, Brown Oculocutaneous Albinism (BOCA) Maps to the OCA2 Locus on Chromosome 15q: P-Gene Mutations Identified. Am J Hum Genet 68:782–787

Manrai AK, Funke BH, Rehm HL, et al (2016) Genetic Misdiagnoses and the Potential for Health Disparities. N Engl J Med 375:655–665. https://doi.org/10.1056/NEJMsa1507092

Mansur YA, Rojano E, Ranea JAG, Perkins JR (2018) Chapter 7 - Analyzing the Effects of Genetic Variation in Noncoding Genomic Regions. In: Deigner H-P, Kohl M (eds) Precision Medicine. Academic Press, pp 119–144

Martin AR, Lin M, Granka JM, et al (2017) An Unexpectedly Complex Architecture for Skin Pigmentation in Africans. Cell 171:1340-1353.e14

Mason HS (1948) The chemistry of melanin; mechanism of the oxidation of dihydroxyphenylalanine by tyrosinase. J Biol Chem 172:83–99

Maston GA, Evans SK, Green MR (2006) Transcriptional Regulatory Elements in the Human Genome. Annu Rev Genomics Hum Genet 7:29–59

McLaren W, Gil L, Hunt SE, et al (2016) The Ensembl Variant Effect Predictor. Genome Biol 17:122

Montagna W, Carlisle K (1991) The architecture of black and white facial skin. J Am Acad Dermatol 24:929–937. https://doi.org/10.1016/0190-9622(91)70148-U

Mora A, Sandve GK, Gabrielsen OS, Eskeland R (2016) In the loop: promoter–enhancer interactions and bioinformatics. Brief Bioinform 17:980–995. https://doi.org/10.1093/bib/bbv097

Mort RL, Jackson IJ, Patton EE (2015) The melanocyte lineage in development and disease. Development 142:620–632. https://doi.org/10.1242/dev.106567

Müller SM, Stolt CC, Terszowski G, et al (2008) Neural Crest Origin of Perivascular Mesenchyme in the Adult Thymus. J Immunol 180:5344–5351. https://doi.org/10.4049/jimmunol.180.8.5344

Navon O, Sul JH, Han B, et al (2013) Rare Variant Association Testing Under Low-Coverage Sequencing. Genetics 194:769–779. https://doi.org/10.1534/genetics.113.150169

Newton JM, Cohen-Barak O, Hagiwara N, et al (2001) Mutations in the Human Orthologue of the Mouse underwhite Gene (uw) Underlie a New Form of Oculocutaneous Albinism, OCA4. Am J Hum Genet 69:981–988

Nicholls RD (1993) Genomic imprinting and uniparental disomy in Angelman and Prader-Willi syndromes: a review. Am J Med Genet 46:16–25

O'Connor CM, Adams JU (2010) How Do Cells Decode Genetic Information into Functional Proteins? In: Essentials of Cell Biology. NPG Education, Cambridge, MA.

Oetting WS (2009) Albinism Database. http://www.ifpcs.org/albinism/. Accessed 17 Apr 2018

Oetting WS, Garrett SS, Brott M, King RA (2005) P gene mutations associated with oculocutaneous albinism type II (OCA2). Hum Mutat 25:323–323

Oh J, Ho L, Ala-Mello S, et al (1998) Mutation Analysis of Patients with Hermansky-Pudlak Syndrome: A Frameshift Hot Spot in the HPS Gene and Apparent Locus Heterogeneity. Am J Hum Genet 62:593–598. https://doi.org/10.1086/301757

Ohnemus U, Uenalan M, Inzunza J, et al (2006) The Hair Follicle as an Estrogen Target and Source. Endocr Rev 27:677–706. https://doi.org/10.1210/er.2006-0020

Okoro AN (1975) Albinism in Nigeria. Br J Dermatol 92:485–492. https://doi.org/10.1111/j.1365-2133.1975.tb03116.x

Ong C-T, Corces VG (2011) Enhancer function: new insights into the regulation of tissue-specific gene expression. Nat Rev Genet 12:283–293. https://doi.org/10.1038/nrg2957

Ong C-T, Corces VG (2012) Enhancers: emerging roles in cell fate specification. EMBO Rep 13:423–430. https://doi.org/10.1038/embor.2012.52

Park S, Morya VK, Nguyen DH, et al (2015) Unrevealing the role of P-protein on melanosome biology and structure, using siRNA-mediated down regulation of OCA2. Mol Cell Biochem 403:61–71

Prota G (1980a) Recent Advances in the Chemistry of Melanogenesis in Mammals. J Invest Dermatol 75:122–127. https://doi.org/10.1111/1523-1747.ep12521344

Prota G (1980b) Cysteine and Glutathione in Mammalian Pigmentation. In: Cavallini D, Gaull GE, Zappia V (eds) Natural Sulfur Compounds: Novel Biochemical and Structural Aspects. Springer US, Boston, MA, pp 391–397

Puri N, Gardner JM, Brilliant MH (2000) Aberrant pH of Melanosomes in Pink-Eyed Dilution (p) Mutant Melanocytes. J Invest Dermatol 115:607–613. https://doi.org/10.1046/j.1523-1747.2000.00108.x

Qi Z, Xie S, Chen R, et al (2018) Tissue-specific Gene Expression Prediction Associates Vitiligo with SUOX through an Active Enhancer. bioRxiv 337196. https://doi.org/10.1101/337196

Qumsiyeh MB, Rafi SK, Sarri C, et al (2003) Double supernumerary isodicentric chromosomes derived from 15 resulting in partial hexasomy. Am J Med Genet A 116A:356–359. https://doi.org/10.1002/ajmg.a.10050

Ramsay M, Colman MA, Stevens G, et al (1992) The tyrosinase-positive oculocutaneous albinism locus maps to chromosome 15q11.2-q12. Am J Hum Genet 51:879–884

Rao SSP, Huntley MH, Durand NC, et al (2014) A three-dimensional map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 159:1665–1680. https://doi.org/10.1016/j.cell.2014.11.021

Raper HS (1926) The Tyrosinase-Tyrosine Reaction. Biochem J 20:735–742

Rentzsch P, Witten D, Cooper GM, et al (2019) CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res 47:D886–D894. https://doi.org/10.1093/nar/gky1016

Rinchik EM, Bultman SJ, Horsthemke B, et al (1993) A gene for the mouse pink-eyed dilution locus and for human type II oculocutaneous albinism. Nature 361:72–76

Ritchie GRS, Dunham I, Zeggini E, Flicek P (2014) Functional annotation of non-coding sequence variants. Nat Methods 11:294–296. https://doi.org/10.1038/nmeth.2832

Romanoski CE, Glass CK, Stunnenberg HG, et al (2015) Epigenomics: Roadmap for regulation. Nature 518:314–316. https://doi.org/10.1038/518314a

Rosemblat S, Sviderskaya EV, Easty DJ, et al (1998) Melanosomal Defects in Melanocytes from Mice Lacking Expression of the Pink-Eyed Dilution Gene: Correction by Culture in the Presence of Excess Tyrosine. Exp Cell Res 239:344–352. https://doi.org/10.1006/excr.1997.3901

Rosenbloom KR, Sloan CA, Malladi VS, et al (2013) ENCODE Data in the UCSC Genome Browser: year 5 update. Nucleic Acids Res 41:D56–D63. https://doi.org/10.1093/nar/gks1172

Ruiz Y, Phillips C, Gomez-Tato A, et al (2013) Further development of forensic eye color predictive tests. Forensic Sci Int Genet 7:28–40. https://doi.org/10.1016/j.fsigen.2012.05.009

Simonis M, Klous P, Splinter E, et al (2006) Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4C). Nat Genet 38:1348–1354. https://doi.org/10.1038/ng1896

Slominski A, Tobin DJ, Shibahara S, Wortsman J (2004) Melanin Pigmentation in Mammalian Skin and Its Hormonal Regulation. Physiol Rev 84:1155–1228. https://doi.org/10.1152/physrev.00044.2003

Song N, Lang RA (2008) Chapter 9 - Eye Development Using Mouse Genetics. In: Tsonis PA (ed) Animal Models in Eye Research. Academic Press, London, pp 120–133

Spritz RA, Bailin TU, Nicholls RD, et al (1997) Hypopigmentation in the Prader-Willi syndrome correlates with P gene deletion but not with haplotype of the hemizygous P allele. Am J Med Genet 71:57–62

Spritz RA, Fukai K, Holmes SA, Luande J (1995) Frequent intragenic deletion of the P gene in Tanzanian patients with type II oculocutaneous albinism (OCA2). Am J Hum Genet 56:1320–1323

Staleva L, Manga P, Orlow SJ (2002) Pink-eyed Dilution Protein Modulates Arsenic Sensitivity and Intracellular Glutathione Metabolism. Mol Biol Cell 13:4206–4220. https://doi.org/10.1091/mbc.E02-05-0282

Steel KP, Barkway C (1989) Another role for melanocytes: their importance for normal stria vascularis development in the mammalian inner ear. Development 107:453–463

Stenson PD, Mort M, Ball EV, et al (2009) The Human Gene Mutation Database: 2008 update. Genome Med 1:13. https://doi.org/10.1186/gm13

Stevens G, van Beukering J, Jenkins T, Ramsay M (1995) An intragenic deletion of the P gene is the common mutation causing tyrosinase-positive oculocutaneous albinism in southern African Negroids. Am J Hum Genet 56:586–591

Sturm RA, Duffy DL, Zhao ZZ, et al (2008) A Single SNP in an Evolutionary Conserved Region within Intron 86 of the HERC2 Gene Determines Human Blue-Brown Eye Color. Am J Hum Genet 82:424–431

Sturm RA, Larsson M (2009) Genetics of human iris colour and patterns. Pigment Cell Melanoma Res 22:544–562. https://doi.org/10.1111/j.1755-148X.2009.00606.x

Suzuki I, Kato T, Motokawa T, et al (2002) Increase of Pro-opiomelanocortin mRNA Prior to Tyrosinase, Tyrosinase-Related Protein 1, Dopachrome Tautomerase, Pmel-17/gp100, and P-Protein mRNA in Human Skin After Ultraviolet B Irradiation. J Invest Dermatol 118:73–78

Tadokoro R, Murai H, Sakai K, et al (2016) Melanosome transfer to keratinocyte in the chicken embryonic skin is mediated by vesicle release associated with Rho-regulated membrane blebbing. Sci Rep 6:38277. https://doi.org/10.1038/srep38277

Taliun D, Harris DN, Kessler MD, et al (2019) Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. bioRxiv 563866. https://doi.org/10.1101/563866

The 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. Nature 526:68–74

The 1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491:56–65. https://doi.org/10.1038/nature11632

The ENCODE Project Consortium (2011) A User's Guide to the Encyclopedia of DNA Elements (ENCODE). PLoS Biol 9:e1001046. https://doi.org/10.1371/journal.pbio.1001046

The UK10K Consortium (2015) The UK10K project identifies rare variants in health and disease. Nature 526:82–90. https://doi.org/10.1038/nature14962

Thody AJ, Higgins EM, Wakamatsu K, et al (1991) Pheomelanin as well as Eumelanin Is Present in Human Epidermis. J Invest Dermatol 97:340–344. https://doi.org/10.1111/1523-1747.ep12480680

Thurman RE, Rynes E, Humbert R, et al (2012) The accessible chromatin landscape of the human genome. Nature 489:75–82. https://doi.org/10.1038/nature11232

Tobin DJ (2011) The cell biology of human hair follicle pigmentation. Pigment Cell Melanoma Res 24:75–88. https://doi.org/10.1111/j.1755-148X.2010.00803.x

Tomita Y, Takeda A, Okinaga S, et al (1989) Human oculocutaneous albinism caused by single base insertion in the tyrosinase gene. Biochem Biophys Res Commun 164:990–996. https://doi.org/10.1016/0006-291X(89)91767-1

Toyofuku K, Valencia JC, Kushimoto T, et al (2002) The Etiology of Oculocutaneous Albinism (OCA) Type II: The Pink Protein Modulates the Processing and Transport of Tyrosinase. Pigment Cell Res 15:217–224. https://doi.org/10.1034/j.1600-0749.2002.02007.x

Uhlén M, Fagerberg L, Hallström BM, et al (2015) Tissue-based map of the human proteome. Science 347:1260419. https://doi.org/10.1126/science.1260419

Visser M, Kayser M, Grosveld F, Palstra R-J (2014) Genetic variation in regulatory DNA elements: the case of OCA2 transcriptional regulation. Pigment Cell Melanoma Res 27:169–177. https://doi.org/10.1111/pcmr.12210

Visser M, Kayser M, Palstra R-J (2012) HERC2 rs12913832 modulates human pigmentation by attenuating chromatin-loop formation between a long-range enhancer and the OCA2 promoter. Genome Res 22:446–455

Wallis CE, Beighton PH (1989) Synchrony of oculocutaneous albinism, the Prader-Willi syndrome, and a normal karyotype. J Med Genet 26:337–339

Walsh S, Liu F, Wollstein A, et al (2013) The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA. Forensic Sci Int Genet 7:98–115. https://doi.org/10.1016/j.fsigen.2012.07.005

Walsh S, Wollstein A, Liu F, et al (2012) DNA-based eye colour prediction across Europe with the IrisPlex system. Forensic Sci Int Genet 6:330–340. https://doi.org/10.1016/j.fsigen.2011.07.009

Wang Y, Song F, Zhang B, et al (2018) The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. Genome Biol 19:151. https://doi.org/10.1186/s13059-018-1519-9

Ward LD, Kellis M (2016) HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. Nucleic Acids Res 44:D877–D881. https://doi.org/10.1093/nar/gkv1340

Ward LD, Kellis M (2012) HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res 40:D930–D934. https://doi.org/10.1093/nar/gkr917

Wei A-H, Zang D-J, Zhang Z, et al (2013) Exome sequencing identifies SLC24A5 as a candidate gene for nonsyndromic oculocutaneous albinism. J Invest Dermatol 133:1834–1840. https://doi.org/10.1038/jid.2013.49

Westra H-J, Peters MJ, Esko T, et al (2013) Systematic identification of trans-eQTLs as putative drivers of known disease associations. Nat Genet 45:1238–1243. https://doi.org/10.1038/ng.2756

Wit E de, Laat W de (2012) A decade of 3C technologies: insights into nuclear organization. Genes Dev 26:11–24. https://doi.org/10.1101/gad.179804.111

Witkop CJ, Niswander JD, Bergsma DR, et al (1972) Tyrosinase positive oculocutaneous albinism among the Zuni and the Brandywine triracial isolate: Biochemical and clinical characteristics and fertility. Am J Phys Anthropol 36:397–405. https://doi.org/10.1002/ajpa.1330360311

Witkop CJ, Quevedo WC, Fitzpatrick FC, King RA (1989) Albinism. In: Scriver CR, Beaudet AL, Sly WS, Valle D (eds) The metabolic basis of inherited disease, 6th edn. McGraw-Hill, New York, pp 2905–2947

Wood AR, Perry JRB, Tanaka T, et al (2013) Imputation of Variants from the 1000 Genomes Project Modestly Improves Known Associations and Can Identify Low-frequency Variant - Phenotype Associations Undetected by HapMap Based Imputation. PLOS ONE 8:e64343. https://doi.org/10.1371/journal.pone.0064343

Wu W, Sato K, Koike A, et al (2010) HERC2 Is an E3 Ligase That Targets BRCA1 for Degradation. Cancer Res 70:6384–6392. https://doi.org/10.1158/0008-5472.CAN-10-1304

Yasumoto K, Takeda K, Saito H, et al (2002) Microphthalmia-associated transcription factor interacts with LEF-1, a mediator of Wnt signaling. EMBO J 21:2703–2714. https://doi.org/10.1093/emboj/21.11.2703

Zaorska K, Zawierucha P, Nowicki M (2019) Prediction of skin color, tanning and freckling from DNA in Polish population: linear regression, random forest and neural network approaches. Hum Genet 138:635–647. https://doi.org/10.1007/s00439-019-02012-w

Zerbino DR, Achuthan P, Akanni W, et al (2018) Ensembl 2018. Nucleic Acids Res 46:D754–D761. https://doi.org/10.1093/nar/gkx1098

Zhu J, Adli M, Zou JY, et al (2013) Genome-wide Chromatin State Transitions Associated with Developmental and Environmental Cues. Cell 152:642–654. https://doi.org/10.1016/j.cell.2012.12.033

Zuo C, Shin S, Keleş S (2015) atSNP: transcription factor binding affinity testing for regulatory SNP detection. Bioinformatics 31:3353–3355. https://doi.org/10.1093/bioinformatics/btv328

# Appendices

## Appendix A: Ethics certificate

UNIVERSITY OF THE WITWATERSRAND JOHANNESBURG

R14/49 Ms M Eisenberg

**HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)**
**CLEARANCE CERTIFICATE NO. M180743**

| | |
|---|---|
| **NAME: (Principal Investigator)** | Ms M Eisenberg |
| **DEPARTMENT:** | School of Pathology<br>Division of Human Genetics<br>National Health Laboratory Service |
| **PROJECT TITLE:** | Investigation of the OCA2 gene control regions and their possible role in the aetiology of brown oculotaneous albinism (BOCA) in Black South Africans |
| **DATE CONSIDERED:** | 27/07/2018 |
| **DECISION:** | Approved unconditionally |
| **CONDITIONS:** | |
| **SUPERVISOR:** | Drs T Ngcungcu and R Kerr |
| **APPROVED BY:** | Dr CB Penny, Chairperson, HREC (Medical) |
| **DATE OF APPROVAL:** | 21/12/2018 |

**This clearance certificate is valid for 5 years from date of approval. Extension may be applied for.**

**DECLARATION OF INVESTIGATORS**

To be completed in duplicate and **ONE COPY** returned to the Research Office Secretary on 3rd floor, Phillip V Tobias Building, Parktown, University of the Witwatersrand, Johannesburg.
I/We fully understand the conditions under which I am/we are authorised to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions. Should any departure be contemplated from the research protocol as approved, I/we undertake to resubmit to the Committee. I agree to submit a yearly progress report. When a funder requires annual re-certification, the application date will be one year after the date of the meeting when the study was initially reviewed. In this case, the study was initially reviewed in July and will therefore reports and re-certification will be due early in the month of July each year. Unreported changes to the application may invalidate the clearance given by the HREC (Medical).

Principal Investigator Signature — Date: 21/12/2018

**PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES**

# Appendix B: Plagiarism documentation


**PLAGIARISM DECLARATION TO BE SIGNED BY ALL HIGHER DEGREE STUDENTS**

SENATE PLAGIARISM POLICY: APPENDIX ONE

I Micaela Tanya Eisenberg (Student number: 726264) am a student registered for the degree of MSc(Med) in the academic year 2019.

I hereby declare the following:

- I am aware that plagiarism (the use of someone else's work without their permission and/or without acknowledging the original source) is wrong.
- I confirm that the work submitted for assessment for the above degree is my own unaided work except where I have explicitly indicated otherwise.
- I have followed the required conventions in referencing the thoughts and ideas of others.
- I understand that the University of the Witwatersrand may take disciplinary action against me if there is a belief that this is not my own unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.
- I have included as an appendix a report from "Turnitin" (or other approved plagiarism detection) software indicating the level of plagiarism in my research document.

Signature: MEisenberg Date: 21/01/2020

726264:Micaela_Eisenberg_726264_dissertation_final.pdf

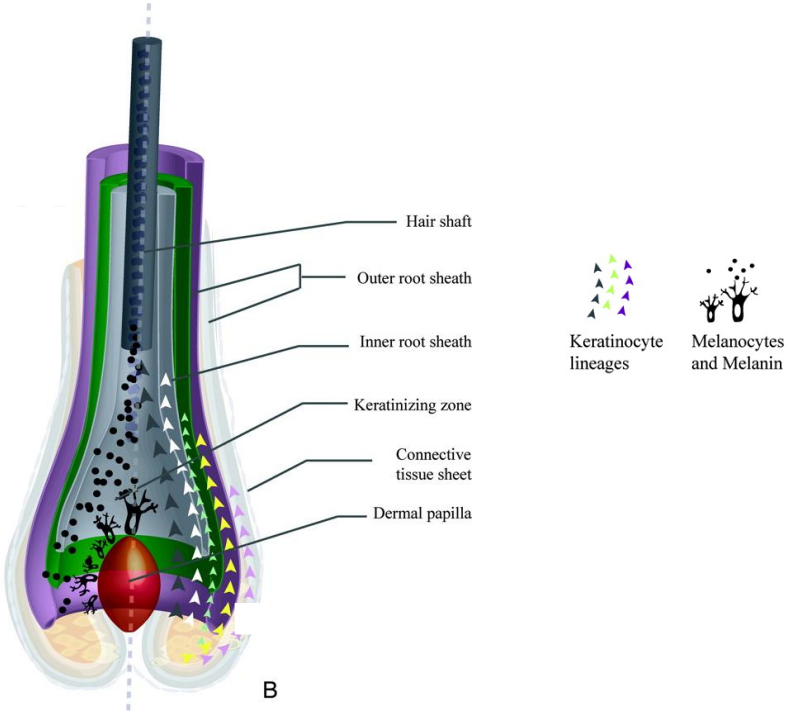
ORIGINALITY REPORT

| 15% | 8% | 9% | 10% |
|---|---|---|---|
| SIMILARITY INDEX | INTERNET SOURCES | PUBLICATIONS | STUDENT PAPERS |

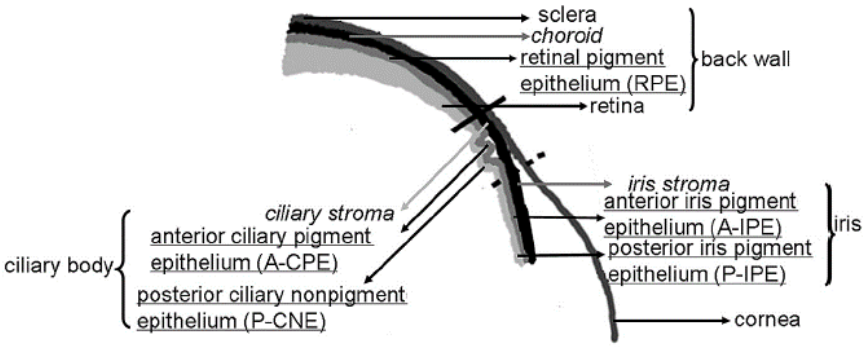
PRIMARY SOURCES

| | | |
|---|---|---|
| 1 | www.nature.com<br>Internet Source | <1% |
| 2 | Submitted to University of Leicester<br>Student Paper | <1% |
| 3 | Submitted to Middle East Technical University<br>Student Paper | <1% |
| 4 | hdl.handle.net<br>Internet Source | <1% |
| 5 | Submitted to University College London<br>Student Paper | <1% |
| 6 | www.ncbi.nlm.nih.gov<br>Internet Source | <1% |
| 7 | Submitted to University of Witwatersrand<br>Student Paper | <1% |
| 8 | genomebiology.biomedcentral.com<br>Internet Source | <1% |
| 9 | Submitted to University of Stellenbosch, South Africa | <1% |

# Appendix C: The site of melanocytes in the hair and eye

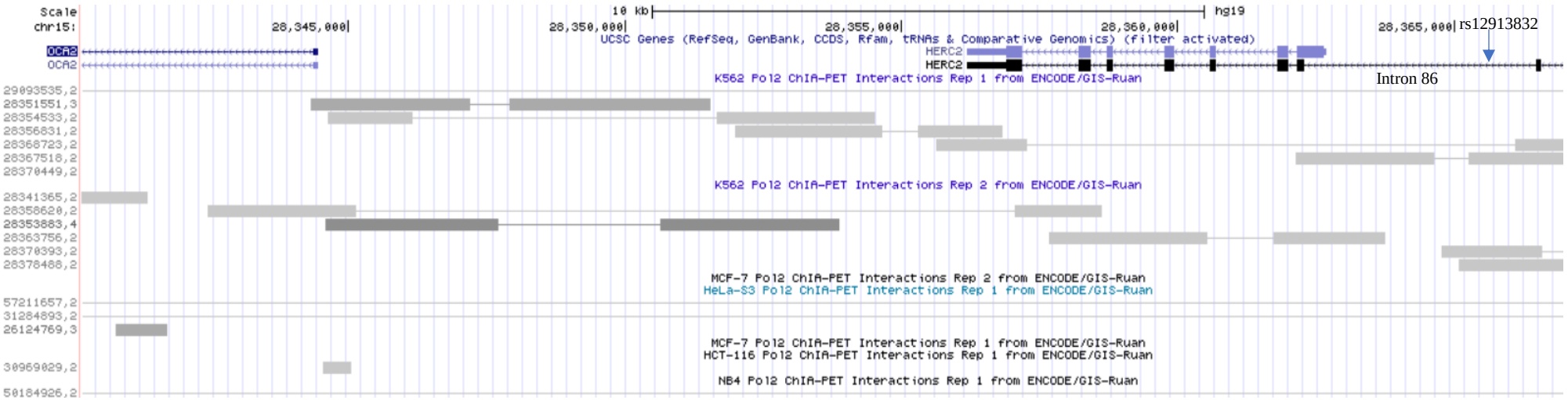



**Appendix figure 1: The location of melanocytes in the hair bulb.** Melanocytes are found in the central region of the bulb, at the root of the hair, surrounding the dermal papilla (adapted from Ohnemus et al. 2006).



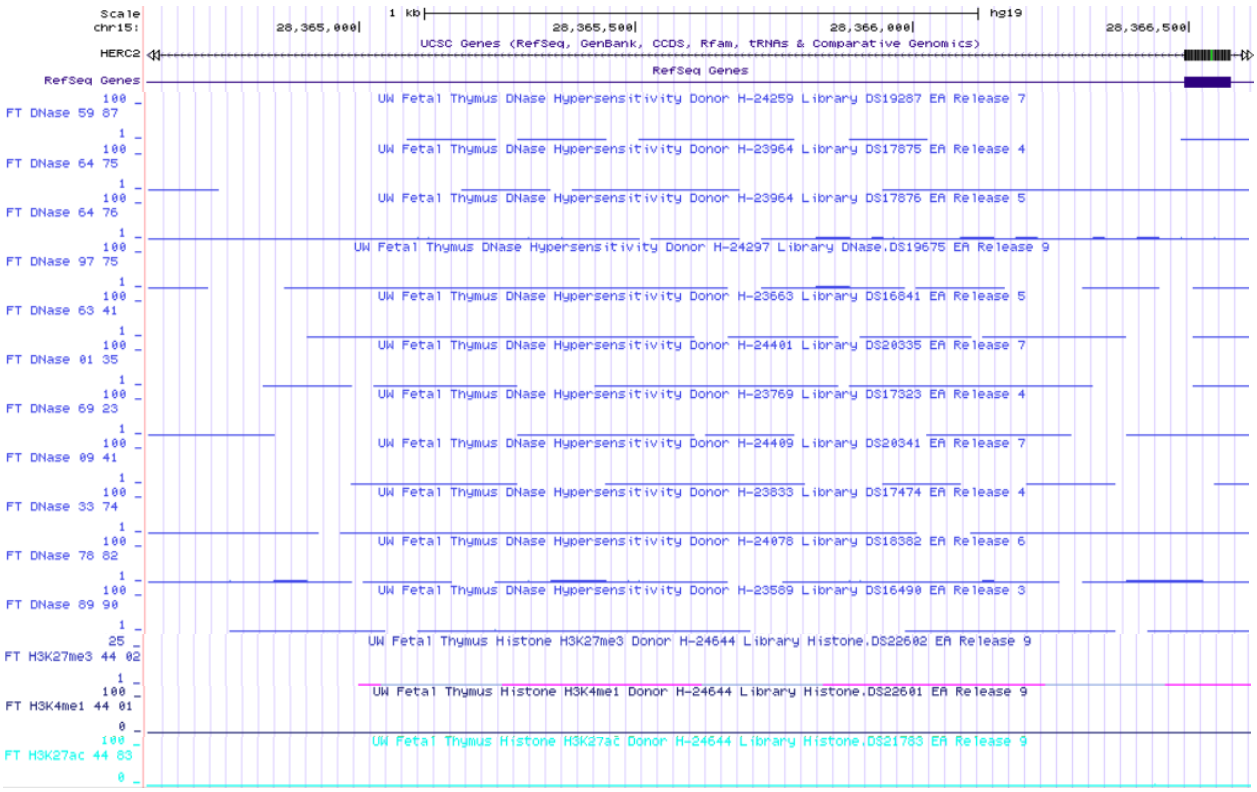

**Appendix figure 2: The structures of the anterior mouse eye which contain pigmented cells.** Melanocytes which are derived from neural crest cells are represented by the labels in italics: located in the ciliary stroma, iris stroma, and choroid. The underlined locations contain pigmented cells derived from the early eye structure (Jing et al. 2014). The organisation of the mouse eye is similar to that of humans and has been extensively used as a model for human eye research (Song and Lang 2008).

# Appendix D: Chromatin Interaction Analysis by Paired-End Tag interactions for RNA polymerase II

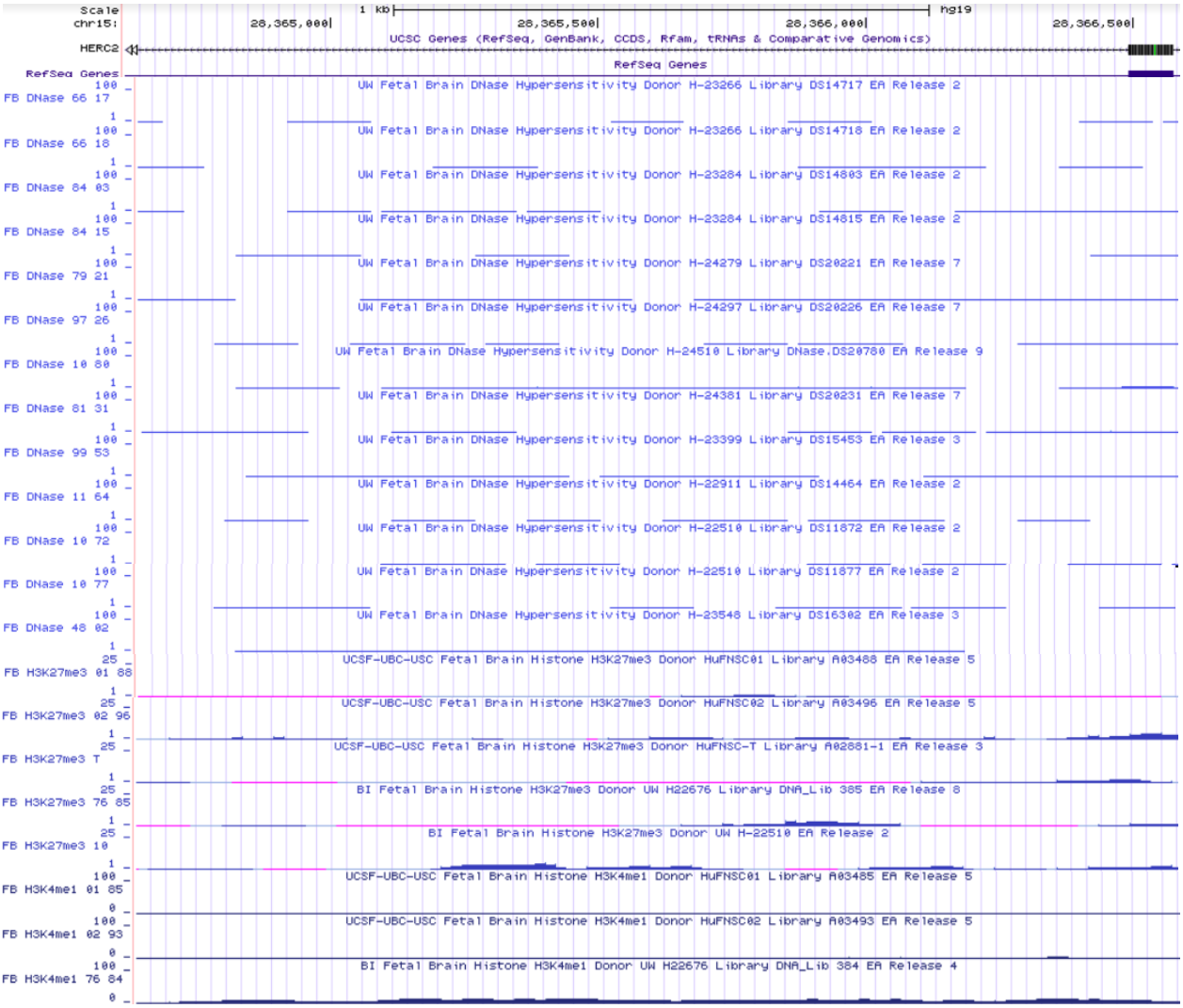



**Appendix figure 3: ChIA-PET data in the *OCA2-HERC2* region for all cell lines (K562 myeloid leukemia from erythrocytes, MCF7 breast cancer, HeLa S3 cervix carcinoma, HCT116 colon cancer and NB4 acute promyelocytic leukemia), when tested for interactions between regions that bind to RNA polymerase II (POL2).** The arrow indicates the approximate location of the rs12913832 SNP which is contained in the putative enhancer region. Image generated using the UCSC genome browser GRCh37 build (http://genome.ucsc.edu/).

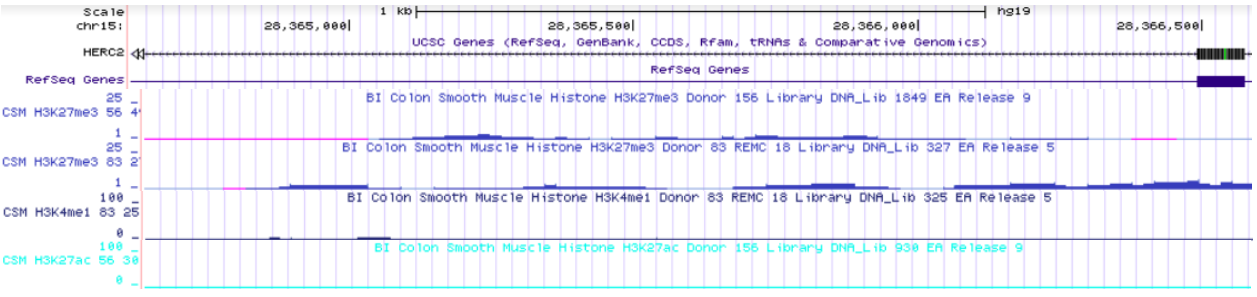
# Appendix E: Histone modification and chromatin accessibility data in the putative enhancer region for control cell types from the Roadmap Epigenomics project




**Appendix figure 4: A selection of enhancer specific signals in the foetal thymus cell line from the Roadmap Epigenomics Project for the coordinates (chromosome 15: 28364618 – 28366618).** This region represents the putative enhancer.




**Appendix figure 5: A selection of enhancer specific signals in the foetal brain cell line from the Roadmap Epigenomics Project for the coordinates (chromosome 15: 28364618 – 28366618).** This region represents the putative enhancer.




**Appendix figure 6: A selection of enhancer specific signals in the colon smooth muscle cell line from the Roadmap Epigenomics Project for the coordinates (chromosome 15: 28364618 – 28366618).** This region represents the putative enhancer.

# Appendix F: Description of the ChromHMM states

## Appendix table 1: A description of the ChromHMM 15 states (Ernst and Kellis 2012).

| State number | Mnemonic | Description |
|---|---|---|
| 1 | TssA | Active transcription start site |
| 2 | TssAFlnk | Flanking active transcription start site |
| 3 | TxFlnk | Transcribed at gene 5’ and 3’ |
| 4 | Tx | Strong transcription |
| 5 | TxWk | Weak transcription |
| 6 | EnhG | Genic enhancers |
| 7 | Enh | Enhancers |
| 8 | ZNF/Rpts | ZNF genes and repeats |
| 9 | Het | Heterochromatin |
| 10 | TssBiv | Bivalent/Poised transcription start site |
| 11 | BivFlnk | Flanking bivalent transcription start site or enhancer |
| 12 | EnhBiv | Bivalent enhancer |
| 13 | ReprPC | Repressed PolyComb |
| 14 | ReprPCWk | Weak repressed PolyComb |
| 15 | Quies | Quiescent/Low |

## Appendix table 2: A description of the ChromHmm 25 states (Ernst and Kellis 2015).

| State number | Mnemonic | Description |
|---|---|---|
| 1 | TssA | Active transcription start site |
| 2 | PromU | Promoter upstream transcription start site |
| 3 | PromD1 | Promoter downstream transcription start site with DNase |
| 4 | PromD2 | Promoter downstream transcription start site |
| 5 | Tx5’ | Transcription 5’ preferential |

| | | |
|---|---|---|
| 6 | Tx | Strong transcription |
| 7 | Tx3’ | Transcription 3’ preferential |
| 8 | TxWk | Weak transcription |
| 9 | TxReg | Transcribed and regulatory (promoter or enhancer) |
| 10 | TxEnh5’ | Transcription 5’ preferential and enhancer |
| 11 | TxEnh3’ | Transcription 3’ preferential and enhancer |
| 12 | TxEnhW | Transcribed and weak enhancer |
| 13 | EnhA1 | Active enhancer 1 |
| 14 | EnhA2 | Active enhancer 2 |
| 15 | EnhAF | Active enhancer flank |
| 16 | EnhW1 | Weak enhancer 1 |
| 17 | EnhW2 | Weak enhancer 2 |
| 18 | EnhAc | Primary H3K27ac possible Enhancer |
| 19 | DNase | DNase only |
| 20 | ZNF/Rpts | ZNF genes and repeats |
| 21 | Het | Heterochromatin |
| 22 | PromP | Poised promoter |
| 23 | PromBiv | Bivalent promoter |
| 24 | ReprPC | Repressed PolyComb |
| 25 | Quies | Quiescent/Low |

## Appendix G: Scripts

**Extracting a list of variants for the region of interest from population files**

#the coordinates for the region of interest were contained in a tab delimited text file (file2) in the format:

#chromosome_number starting_coordinate ending_coordinate

#accessing a bfile (file1) for a specific KGP population and recoding it into a VCF (file3) for a particular range of coordinates contained in a text file (file2)

#activate PLINK

./plink \

#stipulate the input file (file1 or the path for the file) as a bfile (.bed, .bam or .fam file format)

--bfile file1 \

#extract a list of SNPs as stipulated in file2

--extract range file2 \

#output the data as a VCF

--recode vcf \

#stipulate the name of the file to be generated

--out file3

#recoding the bfile that had all AGVP individuals into population specific VCFs (after splitting the individual IDs into individual text files), can also be done to (--keep)

#activate plink

./plink \

#stipulate the input file (filename) as a bfile (.bed, .bam or .fam file format)

--bfile filename \

#keep the individuals specified in the text file with the appropriate individual identifiers

--keep population.txt \

#output the data as a VCF

--recode vcf \

#stipulate the name of the file to be generated

--out populationfile

#make vcfs for the specific regions (using chromosome number and coordinates) from the population specific files

#activate plink

./plink \

#stipulate the input file (filename) as a bfile (.bed, .bam or .fam file format)

--bfile filename \

#keep the individuals specified in the text file with the appropriate individual identifiers

```
--keep population.txt \

#specify the chromosome number

--chr “chromosome number” \

#stipulate the starting coordinate as a base pair

--from-bp “starting coordinate” \

#stipulate the ending coordinate as a base pair

--to-bp “ending coordinate” \

#output the data as a VCF

--recode vcf \

#stipulate the name of the file to be generated

--out populationfile



#generating a VCF for a specific region from a VCF of a broad region

#activate plink

./plink \

#stipulate the input VCF file

--vcf  file1.vcf \

#specify the chromosome number

--chr “chromosome number” \
```

#stipulate the starting coordinate as a base pair

--from-bp “starting coordinate” \

#stipulate the ending coordinate as a base pair

--to-bp “ending coordinate” \

#output the data as a VCF

--recode vcf \

#stipulate the name of the file to be generated

--out file2

**Calling variant frequencies and filtering frequency files for variants with specific allele frequencies**

#call allele frequencies for a VCF (--vcf stipulates a VCF as the input file, --freq)

#activate plink

./plink \

#stipulate the input VCF file

--vcf file1.vcf  \

#call allele frequencies for the SNPs in the file

--freq  \

#stipulate the name of the file to be generated

```
--out file2



#calling frequencies for a specific region, using the coordinates of the region

#activate plink

./plink \

#stipulate the input VCF file

--vcf file1.vcf  \

#call allele frequencies for the SNPs in the file

--freq \

#specify the chromosome number

--chr “chromosome number” \

#stipulate the starting coordinate as a base pair

--from-bp “starting coordinate” \

#stipulate the ending coordinate as a base pair

--to-bp “ending coordinate” \

#stipulate the name of the file to be generated

--out file2



#merging the frequency files into a joint frequency file
```

#heading structure from the frequency file is retained, with population names added to the headings (population_headingname)

#paste the data from file2 adjacent to file1 without changing any information

paste file1 file 2 \

#output to file

> file3

#removing the repetitive CHR and SNP columns from the input joint frequency file (file1)

#repetitive columns are specified columns ($7 is CHR, $8 is SNP, $13 is CHR etc)

# the repetitive columns are replaced by a blank space

#print the rest of the information as is ($0)

#output to new file (file2)

awk '{ $7=$8=$13=$14=$19=$20=$25=$26=$31=$32=$37=$38="  "; print $0 }' file1 > file2

#add a new column to the joint frequency file which is the sum of all MAF values per variant from each population ($5 is MAF for population 1, $9 is MAF for population 2 etc) in the joint frequency file

#output to a new file (file2)

awk '{ print $0 "\t" $5 + $9 + $13 + $17 + $21 + $25 + $29 }' file1 > file2

#remove the lines where the sum of MAF equals 0 (the variant was not polymorphic in any of the populations), represented by column 31 ($31).

#output to a new file (file2)

awk '$31 != 0' file1 > file2

#remove the lines where the sum of MAF does not equal 0 (the variant was polymorphic in at least one population), represented by column 31 ($31)

#output to a new file (file2)

awk '$31 == 0' file1 > file2

#to filter all the MAF columns in the joint frequency file (file1) simultaneously above a certain frequency threshold (as a decimal) ($5 is MAF for population 1, $9 is MAF for population 2 etc).

#output to a new file (file2)

awk '$5 > “value” && $9 > “value” && $13 > “value” && $17 > “value” && $21 > “value” && $25 > “value” && $29 > “value” ' file1 > file2

**Generating a list of known variants using their IDs from VEP**

#extract a list of variants from a VEP results file, including variant ID and position information.

#the input file repeats variant information for all transcripts they occur on

#print tab separated columns containing entered variant ID, location and dbSNP ID ($1 is the entered variant ID, $2 is the location, $20 is the existing variant ID) from the specified input file (file1)

awk '{ print $1 "\t" $2 "\t" $20 }' file1 \

#following this, remove all identical lines and generate a list of unique lines of information

| uniq \

#output the unique lines to a new file (file2)

> file2

#extracting a list of existing variants (which is now $3) as a single column.

#input file is specified (file1)

#output to a new file (file2)

#manually add in chromosomal coordinates for unknown variants

awk '{ print $3 }' file1 > file2

**Finding different and common variants between datasets contained in two different files, generated using previous steps**

Lists of variants must be sorted before comparing

#activate sorting tool

./sort \

#stipulate input file to be sorted (file1)

file1 \

#output to a new file (file2)

> file2

#find differences between the lists of variants (variants that are present in one of the datasets only)

#activate diff tool

./diff \

#stipulate input files to compare (file1 and file2)

file1 file2 \

#output list of differences to new file (file3)

> file3

#find common variants in both lists

#activate join tool

./join \

#stipulate input files to compare (file1 and file2)

file1 file2 \

#output list of shared variants to new file (file3)

> file3

**Searching large datasets for variants of interest**

#activate grep tool for searching

./grep \

#search for terms that match exactly

-w \

#search in a specified file (file2) using terms (variants) defined in an input file (file1)

-f file1 file2 \

#output the lines containing the matching variants to a new file (file3)

> file3