IDENTIFICATION OF SINGLE NUCLEOTIDE POLYMORPHISMS WITHIN THE *OCT2* GENE IN THE SOUTH AFRICAN BLACK POPULATION

Nina Claire Wilson

Supervisor: Demetra Mavri-Damelin

Co-supervisor: Ananyo Choudhury

A Dissertation submitted to the Faculty of Science, University of the Witwatersrand, in partial fulfilment of the requirements for the degree of Master of Science

Johannesburg, 2016

DECLARATION

I, Nina Claire Wilson (550108), am a student registered for the Degree of Master of Science in the academic year 2016, at the University of the Witwatersrand, Johannesburg.

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 examination for the above degree is my own unaided work except where explicitly indicated otherwise.
- I have not submitted this work before for any other degree or examination at any other University.
- 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.

Signature: Wilson ____________________ 25th day of August 2016

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Dr Demetra Mavri-Damelin, for her continuous support and guidance throughout this study; my co-supervisor, Dr Ananyo Choudhury, for his assistance with the bioinformatics work and for the access and retrieval of data from the various studies used; Professor Rob Veale for his cell culture facilities; my laboratory colleagues at the Functional Genetics Research Laboratory for their knowledge and assistance; the National Health Laboratory Service (NHLS) for donating the 10 genomic DNA samples; the National Research Foundation for funding this study and for the block grant; the University of the Witwatersrand for the Post-Graduate Merit Award and lastly my family and close friends for their encouragement, patience and constant support.

ABSTRACT

The Organic Cation Transporter 2 (*OCT2*) gene is responsible for facilitating the transport of cationic compounds, which include both endogenous substrates and clinical drugs. Single nucleotide polymorphisms (SNPs) within this gene were extensively explored in the South African black population as little research has been conducted on these individuals so far. We sequenced the *OCT2* promoter region of 10 DNA samples from the South African black population and identified four SNPs and one INDEL. We performed a luciferase assay to determine their effects on gene expression and we found two variants (rs59695691 and rs138765638) that showed a statistically significant change in luciferase expression suggesting that they may be associated with a change in *OCT2* regulatory function. We also indentified thirteen SNPs and two INDELs within the *OCT2* promoter region, and nine SNPs within the *OCT2* coding region through analysing various South African population studies. These variations could affect both gene expression and protein function. These findings help contribute to filling the gap pertaining to OCT variation in South African populations.

RESEARCH OUTPUTS

National and International Conferences:

Poster Presentation: Wilson, N.C., Choudhury, A., Mavri-Damelin, D. Identification of single nucleotide polymorphisms in the organic cation transporter 2 (*OCT2*) promoter region in the South African black population. The 16th Southern African Society of Human Genetics Congress, Pretoria, South Africa, 2015.

Poster Presentation: Wilson, N.C., Choudhury, A., Mavri-Damelin, D. Identification of single nucleotide polymorphisms in the organic cation transporter 2 (*OCT2*) promoter region in the South African black population. Young Researchers Forum by the Southern African Society of Human Genetics, Pretoria, South Africa, 2015.

Internal Conferences:

Poster Presentation: Wilson, N.C., Choudhury, A., Mavri-Damelin, D. Identification of single nucleotide polymorphisms in the organic cation transporter 2 (*OCT2*) promoter region in the South African black population. Molecular Biosciences Research Thrust Postgraduate Day, Johannesburg, 2015.

Poster Presentation: Wilson, N.C., Choudhury, A., Mavri-Damelin, D. Identification of single nucleotide polymorphisms in the organic cation transporter 2 (*OCT2*) promoter region in the South African black population. 7th Cross-Faculty Graduate Symposium, Johannesburg, 2016.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

bp	Base pairs
dbSNP	Database of Single Nucleotide Polymorphisms
DC	Dicarboxylate
DMEM	Dulbecco’s Modified Eagle Medium
DNA	Deoxyribonucleic Acid
E-box	Enhancer Box
E. coli	*Escherichia coli*
FBS	Fetal Bovine Serum
FWD	Forward
HEK	Human Embryonic Kidney
HiFi	High Fidelity
HS	Human sample
INDEL	Insertion and Deletion
kb	Kilobase
MAF	Minor Allele Frequency
MRC-5	Medical Research Council cell strain 5
mRNA	Messenger RNA
MSA	Multiple sequence alignment
NCBI	National Center for Biotechnology Information
NEB	New England Biolabs
NHLS	National Health Laboratory Service
OA$^+$	Organic Anion
OAT	Organic Anion Transporter
OC$^-$	Organic Cation
OCT	Organic Cation Transporter
OCTN	Organic Cations and Zwitterions
PBS	Phosphate Buffered Saline
PCR	Polymerase Chain Reaction
PolyPhen	Polymorphism Phenotyping
REV	Reverse
RNA	Ribonucleic Acid
TAE	Tris-Acetate-EDTA

Taq	*Thermus aquaticus*
TMD	Transmembrane Domain
TSS	Transcription Start Site
SDS	Sodium Dodecyl Sulfate
SIFT	Sorting Intolerant From Tolerant
SLC	Solute Carrier Superfamily
SNP	Single Nucleotide Polymorphism
Stats SA	Statistics South Africa
USF-1	Upstream stimulatory factor-1
UTR	Untranslated Region
VEP	Variant Effect Predictor
WT	Wild Type
ZI	Zwitterion

1. Introduction

1.1 SLC22 Family of transporters

Various endogenous compounds, drugs, toxins and environmental products are absorbed, distributed throughout and excreted from the human body by broad-specificity transporters that belong to the solute carrier superfamily 22 (SLC22) (Koepsell, Lips & Volk, 2007). This family includes transporters for organic cations (OCTs), organic anions (OATs) as well as organic cations and zwitterions (OCTNs) together, in which their main roles all occur in the kidney and the liver. The OCTs are members of the *SLC22A* family that contains the three subtypes of OCTs called OCT1 (*SLC22A1*), OCT2 (*SLC22A2*) and OCT3 (*SLC22A3*) (Koepsell, Lips & Volk, 2007).

All OCT proteins have been predicted to share a 12-transmembrane domain (TMD) structure (Figure 1). Between TMD1 and TMD2 the structure comprises of a large extracellular hydrophilic loop which has glycosylation sites and is responsible for playing a role in oligomerisation (Brast *et al*., 2011; Keller *et al*., 2011). Between TMD6 and TMD7, the structure comprises of a large intracellular loop that has phosphorylation sites which are important for protein kinases to regulate the SLC22 transporters (Koepsell, Lips & Volk, 2007).

Due to these conserved structural features, the OCTs share high sequence similarity. The cationic transporters from different species are between 551 and 557 amino acids in length and between OCT1 and OCT2 there are 285 (51 %) common amino acids; among OCT1, 2 and 3 there are 188 (34 %) identical amino acids and between all OCTs and OCTNs there are 92 (17 %) conserved amino acids (Burckhardt & Wolff, 2000). Of these conserved amino acids there are 4 cysteines and 13 prolines, which likely contribute to the formation of the protein’s secondary structure (Burckhardt and Wolff, 2000). Maintaining the secondary structure may occur through the conserved charged amino acids either by the binding of charged substrates or through salt bridges. These include 3 aspartic acids, 6 glutamic acids, and 7 arginines (Burckhardt & Wolff, 2000).

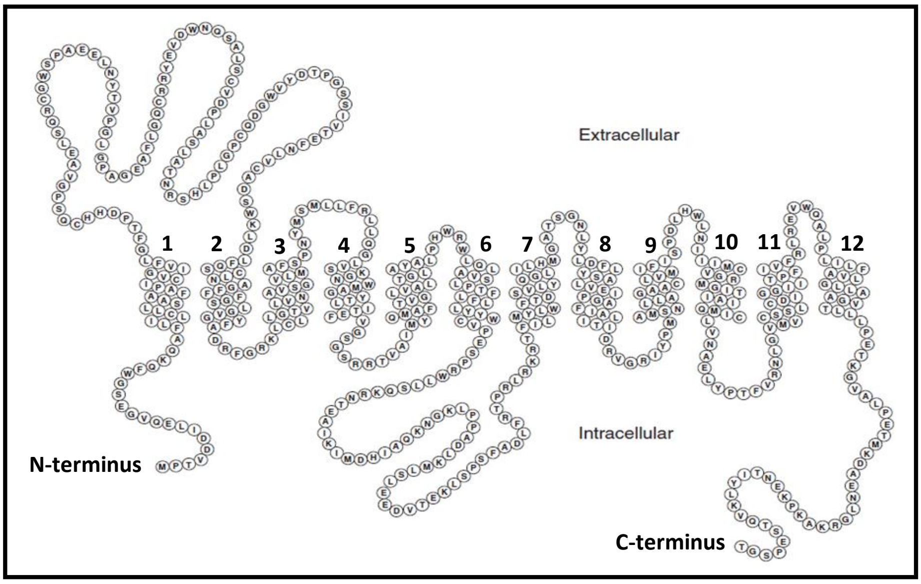


Figure 1: The 12-transmembrane domain (TMD) predicted structure of OCT proteins. Between TMD1 and TMD2 a large glycosylated extracellular loop is found, and between TMD6 and TMD7 a large intracellular loop that contains phosphorylation sites is found. The N- and C-terminus are located intracellularly. (Modified from Volk, 2014).

1.2 OCT2 (*SLC22A2*)

This study focuses on the human *OCT2* gene that is encoded by *SLC22A2*, which is situated on the reverse strand of chromosome 6q26. The gene is 45 kilobases (kb) in length and consists of 11 coding exons (Koehler *et al*., 1997; Gründemann & Schömig, 2000). The Ensembl genome browser (GRCh38.p5) shows that *OCT2* has six transcripts: the longest transcript of 3737 bp produces a protein of 334 amino acids, the second longest transcript of 2597 bp produces the longest protein of 555 amino acids, which is used for this study, and the other four transcripts are not translated into proteins (Yates *et al*., 2015). The expression of this transporter mainly occurs at the basolateral membrane in the renal proximal tubule cells of the kidney (Figure 2), but expression also occurs in the placenta, spleen, lung, inner ear, small intestine, thymus and the brain (Gorboulev *et al*., 1997; Motohashi *et al*., 2002; Koepsell, Schmitt & Gorboulev, 2003). OCT2 facilitates the transport of numerous cationic drugs from the circulation of the body into the renal epithelial cells of the kidney, where they are finally excreted into the urine (Gorboulev *et al*., 1997; Motohashi *et al*., 2002). Transportation of organic cations occurs in both directions across the plasma membrane via substrate concentration gradients and the membrane potential (Koepsell, 2011). Examples of cationic compounds that are transported by OCT2 are shown in Table 1.

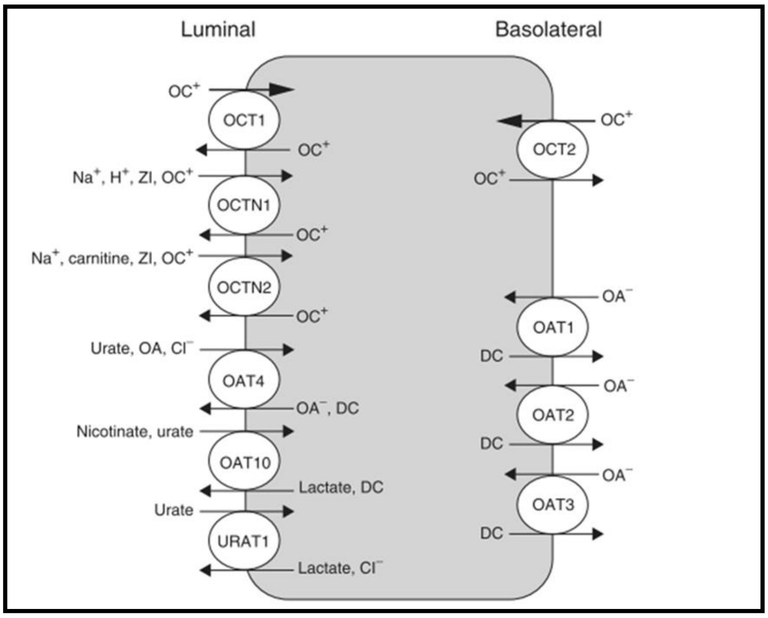


Figure 2: **The location of the different SLC22 transporters present in the human renal proximal tubule cell.** *OCT2* is expressed in the basolateral membrane. The difference in the line thickness with regards to OCT1 and OCT2 indicates the direction to which organic cation uptake is favoured. Abbreviations: DC, dicarboxylate; OA^{+}, organic anion; OC$^{–}$, organic cation; ZI, zwitterion. (Volk, 2014).

Table 1: Examples of cationic compounds transported by the human *OCT2* gene.

Class	Compound	Drug category
Drugs	Metformin[b]	Antidiabetic
	Phenformin[c]	Antidiabetic
	Cisplatin[b]	Anticancer
	Oxaliplatin[d]	Anticancer
	Quinine[b]	Antimalarial
	Lamivudine[a]	Antiretroviral
	Cimetidine[b]	Antihistamine, histamine H$_2$ receptor
	Amantadine[b]	Antiparkinsonian agent, antiviral
	Acetylsalicylate[c]	NSAIDs*
	Salicylate[c]	NSAIDs*
	Creatinine[b]	Metabolite
Neurotransmitters	Acetylcholine[b]	Endogenous
	Dopamine[b]	Endogenous
	Histamine[b]	Endogenous
	Epinephrine[b]	Endogenous
	Norepinephrine[b]	Endogenous
	Seratonin[b]	Endogenous
	Choline[c]	Endogenous, Dietary supplement

[a] Jung *et al*., 2008. Reviewed in [b] Koepsell, Lips & Volk, 2007; [c] Fujita *et al*., 2006; [d] Yonezawa *et al*., 2006.

*Non-steroidal anti-inflammatory drugs

Orthologs of *OCT2* have been cloned from human (Gorboulev *et al.*, 1997), pig (Gründemann *et al.*, 1997), mouse (Mooslehner & Allen, 1999) and rabbit (Zhang, Evans & Wright, 2003) since its original isolation from a rat renal library in 1996 (Okuda *et al.*, 1996).

1.3 *OCT2* promoter and transcription factors

A promoter, found upstream of the translation start codon (ATG), contributes to the regulation of gene expression and is the site to which ribonucleic acid (RNA) polymerase binds prior to the initiation of transcription. The promoter is usually classified as the region that is 1 kb from the transcription start site (TSS) (Zhang, 1998). Transcription factor binding sites, which can be either positive or negative regulators of gene expression, are located within this region and can enhance or diminish the expression of the gene.

The upstream stimulatory factor 1 (USF-1), a member belonging to the family of basic helix-loop-helix-leucine zipper transcription factors, is known to bind to the enhancer box (E-box) region of the *OCT2* promoter (Corre & Galibert, 2005). The *OCT2* promoter is furthermore known to contain a CCAAT box where a number of different transcription factors such as nuclear factor-Y (Mantovani, 1999) and CCAAT/enhancer-binding proteins can bind (Ramji & Foka, 2002). Figure 3 shows a graphical representation of the transcription factor binding region of the *OCT2* promoter. It was observed in a study by Asaka *et al.* (2007) that an E-box mutation resulted in a decrease in the promoter activity of *OCT2* while an overexpression of USF-1 resulted in enhanced promoter activity (Asaka *et al.*, 2007).

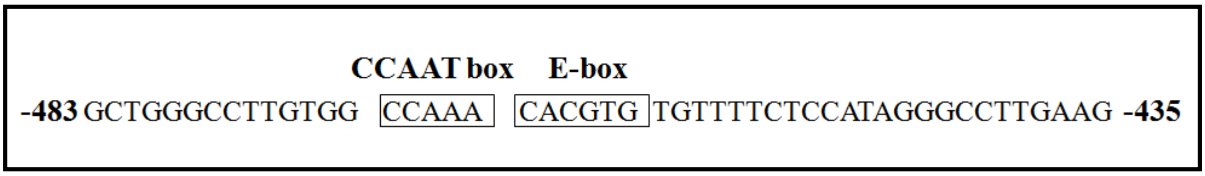

CCAAT box
E-box
-483 GCTGGGCCTTGTGG CCAAA CACGTG TGTTTTCTCCATAGGGCCTTGAAG -435

Figure 3: The transcription factor binding region of the *OCT2* promoter spanning from -483 to -435. Numbering is relative to the translation start site. This sequence indicates the location of the CCAAT box and the E-box regions that are necessary for transcription initiation. USF-1 binds to the E-box. (Modified from Asaka *et al.*, 2007).

Genetic imprinting regulates organ specific expression of *OCT2*. Aoki *et al.* (2008) showed in their study that deoxyribonucleic acid (DNA) methylation regulates the expression of *OCT2* in the kidneys. Liver cells showed hypermethylated levels whereas kidney cells showed hypomethylated levels at all CpG sites in the promoter, and there was a low level of methylation at the CpG site present in the E-box. Methylation, in the *OCT2* promoter or the E-box specifically, repressed the binding of USF-1 and hence reduced the transcriptional activity of *OCT2* (Aoki *et al*., 2008). The study also established that although the methylation status of CpG sites is low in the kidneys, there were differences between different individuals observed in the promoter region.

Posttranscriptional regulation and transport activity of OCT2 is controlled by several signalling pathways, which involve protein phosphorylation caused by protein kinase A, protein kinase C, phosphatidylinositol-3-kinase, and calcium/calmodulin complexes (Ciarimboli & Schlatter, 2005; Koepsell, Lips & Volk, 2007).

1.4 The role of OCT2 in the treatment of human diseases

Since a number of the drugs administered for the treatment of various diseases (Table 1) rely on OCT2 transport, OCT2 is a target for the effective treatment of various human diseases some of which are cancer, type 2 diabetes and Parkinson’s disease. Mutations, deletions, and single nucleotide polymorphisms (SNPs) within the *OCT2* gene may influence the way a person responds to treatment for various diseases (Houtsma, Guchelaar & Gelderblom, 2010).

Cancer is a disease that globally affects both developed and developing countries including South Africa. Each year approximately 14 million people globally and more than 100 000 South Africans are diagnosed with cancer (http://www.cansa.org.za/south-african-cancer-statistics/ - accessed 15/03/2016). Cisplatin, a substrate of OCT2, is a platinum-based drug commonly used worldwide as an anticancer agent for various solid tumors (Ciarimboli *et al*., 2005; Yonezawa *et al*., 2005) including: bladder, endometrial, ovarian, cervical, testicular, lung, as well as head and neck cancer (Perez, 1998; Go & Adjei, 1999; Zhang *et al*., 2006). One primary mechanism of action of cisplatin involves the formation of cisplatin-DNA adducts, which causes DNA damage within the rapidly proliferating cancer cells and results in apoptosis (Johnson *et al*., 1989). The involvement of OCT2 in the renal circulation leads to an accumulation of cisplatin within the kidney and unfortunately causes severe nephrotoxicity, which is one of the major side effects of cisplatin therapy and therefore limits its clinical application despite its effectiveness (Yonezawa *et al*., 2005).

Before the 1990s diabetes was considered a rare disease in Sub-Saharan Africa, however, it has now emerged that there has been a rise in the burden in Sub-Saharan Africa as well as globally (Mbanya *et al*., 2010; Levitt, 2008). Peer *et al*. (2012) demonstrated that there has been an increase in the prevalence of diabetes in urban black South Africans compared to 20 years ago and this was the first study to show the significant rise in diabetes in South Africa (Peer *et al*., 2012). It has been shown that people who suffer from diabetes show a lower expression of OCT2 (Ciarimboli *et al*., 2005). Metformin, transported by OCT1 and OCT2, is one of the most commonly prescribed drugs used for the treatment of type 2 diabetes because it improves insulin sensitivity (Bailey & Turner, 1996; Kirpichnikov, McFarlane & Sowers, 2002). Metformin is also favoured because it has beneficial effects such as lowering lipid levels, and reduces hypoglycaemia and cardiovascular risks whilst displaying very few adverse side effects (Kirpichnikov, McFarlane & Sowers, 2002).

Human OCT2 messenger RNA (mRNA) has also been identified in cells of the central nervous system (Gorboulev *et al*., 1997; Busch *et al*., 1998) and thus dopamine, a neurotransmitter that is transported in the brain, is another substrate of OCT2. Parkinson’s disease is associated with dopamine depletion and since amantadine, an anti-Parkinsonian drug, interacts with OCT2, it suggests that OCT2 may be a target to treat Parkinson’s disease (Busch *et al*., 1998).

Genetic variation in the *OCT2* gene may affect the transport of drugs administered for the treatment of different diseases and so it has been suggested that in the near future SNPs will be important genetic markers in individual-based diagnosis and treatment (Zienolddiny & Skaug, 2012). Therefore, identifying genetic polymorphisms in different populations can lead to more specialised and effective medical treatments.

1.5 Genetic polymorphisms in *OCT2*

Genetic polymorphism is a term used to describe the stable coexistence of two or more distinct genotypes for a given trait in a population where the less common genotype has a frequency of at least 1 % in the population (Meyer, 2000). SNPs are the most common DNA sequence variations found within a population even though more than 99 % of the human DNA sequence is the same. For every kb of DNA sequence present in the human genome there is approximately 1 SNP (Frazer *et al.*, 2007).

Genes encoding drug transporters, such as *OCT2,* in which genetic polymorphisms may occur, may contribute to the variation in the pharmacokinetic / pharmacodynamic profiles of clinically essential drugs (Takane *et al.*, 2008). Genetic modifications to *OCT2*, either in the coding or non-coding regions, can change the expression level or activity of OCT2. As a result, an increase or decrease in the levels of drug substrates for this particular transporter could consequently affect drug disposition, efficacy and toxicity (Evans & Relling, 1999; Ieiri *et al.*, 2006). Clinical studies, *in vitro* and *in vivo* animal experiments, have shown that changes to *OCT2* expression levels can contribute to the individual variation in pharmacokinetics.

1.5.1 Coding region polymorphisms

Genetic polymorphisms in the *OCT2* gene coding region have been characterised in various ethnic populations including Caucasian, African-American and Japanese (Leabman *et al.*, 2002; Fukushima-Uesaka *et al.*, 2004). Four nonsynonymous SNPs identified in the *OCT2* gene in the African-American population resulted in altered OCT2 transport function. The less frequent variants (M165I, R400C and K432Q) tended to result in more significant and deleterious functional changes of OCT2 transport, whereas the most frequent nonsynonymous variant (A270S) displayed more subtle effects on the functional transport of OCT2 (Leabman *et al.*, 2002). Nonsynonymous SNPs that result in impaired OCT2 function may also potentially decrease the clearance of norepinephrine or serotonin, which are neurotransmitters transported by OCT2 and are released in the brain. This may lead to changes in mood related behaviour (Bacq *et al.*, 2011).

1.5.2 Non-coding region polymorphisms

Ogasawara *et al*. (2008) identified a functionally significant genetic polymorphism in the *OCT2* promoter region in Japanese nephrectomised patients. The deletion of AAG at position -578 to -576 (-578_-576delAAG) was shown to significantly reduce the activity of the *OCT2* promoter and heterozygotes of this deletion were shown to have lower OCT2 mRNA levels, which were found to be insignificant (Ogasawara *et al.*, 2008). This variation is annotated as (-318_-316delAAG) and has a dbSNP ID rs138765638 according to the Ensembl genome browser. Important to note is that there is a discrepancy between published data and the

Ensembl genome browser as to where the TSS begins for the *OCT2* gene. Therefore for this study all the SNPs identified are numbered relative to the translation start site where the A from ATG is denoted as +1 and the nucleotide immediately upstream of the A is denoted as −1.

One example, identified in the Chinese population is that of a SNP in the promoter region −1283 T > C, which was associated with decreased luciferase activity compared to the wild type (WT) indicating that the transcriptional activity of the *OCT2* gene decreased. The presence of this SNP, however, showed no change in the renal clearance of metformin (Wang, 2007). This may be due to the study being conducted in healthy subjects and therefore another study should be conducted in diabetic patients to compare results.

SNPs found in both the 5′ flanking region and 5′ UTR may influence the binding and interaction of transcription factors that regulate transcription of the OCT2 gene (Hayashi, Watanabe & Kawajiri, 1991; Sharma, Mount & Karrow, 2008). The USF-1 transcription factor as well as other transcription factors found in these regions within the promoter may have a stronger affinity for the promoter and therefore increase the binding of RNA polymerase to the transcription start site. This will enhance transcription of the *OCT2* gene and consequently produce more protein. Alternatively, these transcription factors may have a weaker affinity for the promoter and therefore decrease the binding of RNA polymerase. This will reduce transcription of the *OCT2* gene and consequently produce less protein. It is also possible that these SNPs may have no influence on transcription factor binding.

The 5′ UTR SNPs may not just regulate transcription, but may also alter the translation efficiency and regulation of the OCT2 mRNA into protein. The 5′ UTR is found to have a high GC content, which allows secondary structural features to be formed. These structural features include hairpin loops (Kozak, 1991) and G-quadruplexes (Sen & Gilbert, 1988) which can control translation initiation and regulation by influencing the recruitment of ribosomes to the translation start site (Gingras, Raught & Sonenberg, 1999). Hairpin loops form stable structures when the average free energy is less than − 50 kcal/mol which enables these structures to inhibit translation (Araujo *et al*., 2012). If a SNP is found in this region, it may prevent the hairpin loop from forming or it may change the average free energy causing an unstable structure and therefore it may promote translation.

1.5.3 G-quadruplexes

G-quadruplexes are not commonly found within the human genome but they do occur more often than expected in gene promoters (Huppert & Balasubramanian, 2007). G-quadruplexes are four-stranded helical structures rich in guanine nucleic acid sequences that are stabilised by G-quartets (Sen & Gilbert, 1988). These G-quartets are planar arrays of four guanines held together by Hoogsteen hydrogen bonds (Gellert, Lipsett & Davies, 1962). These structures are expected to occur more easily in RNA than in DNA because RNA is mostly single-stranded and therefore there is no competition for complementary base-pairing during the folding process compared to DNA which is double-stranded.

Huppert *et al*. (2008) used bioinformatics approaches to determine that there are approximately 2000 genes found in the human genome that have the potential to contain G-quadruplexes in the 5′ UTR which show an overrepresentation of G-quadruplex motifs at the 5′ end of the 5′ UTRs (Huppert *et al*., 2008). G-quadruplexes have been identified in the promoter regions of genes such as the chicken ß-Globin, the human *BCL-2*, the human *VEGF*, *HIF*-1R and the oncogenes *c-myc* and *c-kit* (Howell *et al*., 1995; Dai *et al*., 2006; Sun *et al*., 2005; de Armond *et al*., 2005; Siddiqui-Jain *et al*., 2002; Rankin *et al*., 2005). Both oncogenes have been shown to control the activity of transcription. It is possible that G-quadruplexes may form within the *OCT2* promoter region. The formation of these G-quadruplexes could subsequently alter the level of transcription regulation as shown in Figure 4.

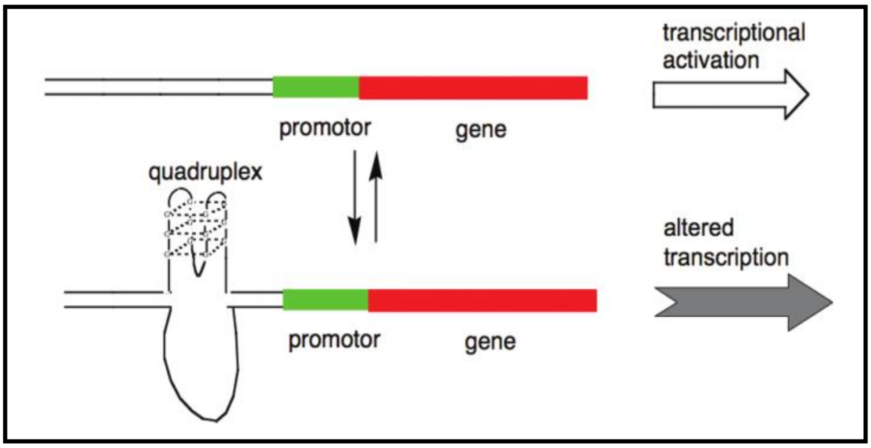


Figure 4: A schematic diagram showing transcription regulation *via* formation of a G-quadruplex in the promoter region of a gene. (Huppert & Balasubramanian, 2007).

Since G-quadruplexes can play such an important role in the genome it is very important to accurately predict the formation and location of G-quadruplexes in order to study their biological functions. If SNPs are identified within a G-quadruplex-forming sequence, it may disrupt the formation of the G-quadruplex structure in a gene and consequently affect its regulation of transcription or translation. Prediction of G-quadruplexes can be identified using software programs, such as QGRS Mapper, but since G-quadruplexes have only recently become a hot topic in research there are few resources available for one to use. QGRS Mapper can analyse both genomic and RNA sequences and uses a specific algorithm search sequence: G_x-N_{y1}-G_x-N_{y2}-G_x- N_{y3}-G_x, where x is ≥ 2 and corresponds to the number of guanine tetrads in the G-quadruplex; y1-y3 corresponds to the length of the loops connecting the guanine tetrads and N corresponds to any of the five bases (adenine, guanine, cytosine, thymine and uracil) (Kikin, D'Antonio & Bagga, 2006).

1.6 African genetic diversity

Due to the high level of genetic diversity within the African continent, there has been a continuous increase in research in this area (Tishkoff *et al.*, 2009; Ramsay, 2012). The genetic diversity among the African populations is greater when compared to other populations such as the Caucasians or Asians (Rosenberg *et al.*, 2002) and this diversity found among the African populations has been poorly studied in the past (Hardy *et al.*, 2008; Tishkoff *et al.*, 2009).

Recent large-scale genome sequencing projects are gradually bridging this gap and providing a more comprehensive view of the African genomic diversity. The 1000 Genomes Project contains whole genome sequence data for more than 500 individuals from five African populations: Esan in Nigeria; Luhya in Webuye, Kenya; Gambian in Western Division, The Gambia; Mende in Sierra Leone and Yoruba in Ibadan, Nigeria (1000 Genome Project Consortium, 2015). In addition, it includes more than 150 genomes from two populations with African ancestry: Africans from southwest United States and Caribbean population in Barbados (1000 Genome Project Consortium, 2015). More information on these populations is shown in Table 5. Similarly, the African Genome Variation Project has generated 300 whole genome sequences from Eastern and Southern African populations (Gurdasani *et al.*, 2014). More such data are expected from ongoing projects like the Southern African Human Genome Programme (Pepper, 2011), the H3Africa Project (http://h3africa.org/), and the TrypanoGEN study (http://www.trypanogen.net/). These datasets are providing us with the opportunity to estimate and compare genetic variation among African populations and to also study genomic regions of functional significance with greater precision.

The southern Africa region (according to the United Nations Geoscheme) constitutes five countries: Botswana, Lesotho, Namibia, Swaziland and South Africa (http://millenniumindicators.un.org/unsd/methods/m49/m49regin.htm - accessed 17/08/2016). This region is the home to a predominant number of Bantu-speakers that expanded southwards from Nigeria and Cameroon nearly 5000 years ago (Blench, 2006; Campbell & Tishkoff, 2010), reaching South Africa about 1500 years ago (Ehret, 1998).

Statistics South Africa (Stats SA) estimates that the mid-year population of South Africa for 2015 was nearly 55 million, where more than 44 million people (80.5 %) are from the African population; more than 4.8 million (8.8 %) from the coloured population; more than 4.5 million (8.3 %) from the white population; and more than 1.3 million (2.4 %) from the Indian/Asian population (https://www.statssa.gov.za/ publications/P0302/P03022015.pdf - accessed 04/02/2016). The African Bantu-speaking population was the main focus of this study, and it includes the following ethnic groups: Zulu, Xhosa, Sotho, Venda, Tswana, Tsonga, Swazi and Ndebele. SNPs in the OCT2 gene promoter region have yet to be characterised well in the southern African populations to date and were therefore the main focus of this study.

1.7 SNPs identified within the *OCT2* promoter region in the South African black population

A preliminary study conducted in our laboratory identified three heterozygous SNPs within the *OCT2* promoter region from sequencing of just 10 DNA samples (from South African black individuals). One SNP was found in the 5′ flanking sequence (−289 G/A) and two SNPs were found in the 5′ untranslated region (UTR) (−246 C/T and −195 G/A). The location of each SNP with regards to the transcription factor binding region of the *OCT2* promoter region is shown in Figure 5.

According to the Ensembl genome browser (GRCh38.p5) the *OCT2* promoter region has 33 5′ UTR variants of which 30 are classified as SNPs and there are 291 upstream gene variants of which 275 are SNPs (Yates *et al*., 2015). The study presented here involved investigating a longer *OCT2* promoter region since data from Ensembl showed a great number of potentially significant SNPs located further upstream from the transcription factor binding region of the *OCT2* promoter. Therefore the genomic DNA from the 10 black South African individuals used in the preliminary study were sequenced again in this study to not only confirm the presence of the three SNPs already identified, but to also look for any new or novel SNPs within the expanded *OCT2* promoter region.

```
-799  tctaggacacaaagatagtggcttggacacacctgcctgcatttacacttgacctgtctg
-739  cgacgtaaacactttcctctttccctccagatgggttaaggggaaggacacttcagggtt
-679  gaaacgcaggaataccagattggagcaaacacttttaaaagcagagttataaaatctgg
-619  acaacatcaaaacaagcagccccagcatgcatcccgacggctcttgttgttggttggaga
-559  atgagcccagcagtcaggcttgcaacccacttcgaatctggaccagggttctgacacgga
-499  tcctggttcacatcacgctgggccttgtggccaaacacgtgtgttttctccatagggcct
-439  tgaagaaaagctggcggtgcgcatgagataggagtatattaagttcctggctgctcgggg
-379  cactacgggaagattactgggctgtgatatgggccagcactcagattccctgcggtggga
-319  cacagagggcgggttgtttgtgctgctggcgtggagcaccgacaagcctgtggagaacca
-259  GTTATAATAAACACGACAGGCATCCTGGGAGTGAGCTCAGGGCATTTGGGAAGTGCAGAA
-199  GGACATGCACCCCCGCTGGAGGGGTGCACCTTTGAAGTCAGCTGGACCAAGGAAAGGCCC
-139  TGCCCTGAAGGCTGGTCACTTGCAGAGGTAAACTCCCCTCTTTGACTTCTGGCCAGGGTT
 -79  TGTGCTGAGCTGGCTGCAGCCGCTCTCAGCCTCGCTCCGGGCACGTCGGGCAGCCTCGGG
 -19  CCCTCCTGCCTGCAGGATCATGCCCACCACCGTGGACGATGTCCTGGAGCATGGAGGGGA
```

KEY:

Primers	E-box	Translation start site	5′ Flanking sequence
CCAAT box	SNPs	Translated sequence	5′ UTR

Figure 5: The various elements located within the *OCT2* gene sequence. The primers highlighted in yellow amplified a 693 bp product of the *OCT2* promoter region. The bold and underlined text represents the transcription factor binding region of the *OCT2* promoter region spanning from −483 to −435. The three SNPs highlighted in blue are those identified in the preliminary study at positions −289, −246 and −195, respectively. Numbering is relative to the translation start site. This sequence is located on the reverse strand.

1.8 In summary

To date there is little information available regarding SNPs in the *OCT2* gene and the effect of these SNPs on *OCT2* gene expression in the South African population. Thus, the intention of this study was to identify these SNPs and determine the effect that they may have on *OCT2* gene expression. This was done by sequencing the promoter region of 10 DNA samples from the South African black population and using the luciferase reporter system to determine their effects on gene expression. Furthermore, data for the *OCT2* promoter and coding regions was extracted from various South African sequencing studies that focused on the Bantu-speaking population. These results were compared to other African and non-African populations. Bioinformatic tools were then used to analyse the various SNPs identified, as well as to predict transcription factor binding sites and G-quadruplex formation in the *OCT2* promoter region where SNPs were identified in order to determine whether the SNPs may impact *OCT2* gene regulation.

2. AIM AND OBJECTIVES

Aim

To identify, functionally characterise and analyse SNPs located within the *OCT2* gene in the South African black population.

Objectives

2.1 SNP identification and functional analysis

2.1.1 To sequence and identify SNPs within the *OCT2* promoter region from 10 genomic DNA samples.

2.1.2 To functionally characterise each SNP found within the 10 genomic DNA samples using the luciferase reporter assay.

2.2 Bioinformatic analysis

2.2.1 To identify SNPs found within the *OCT2* promoter and coding regions.

2.2.2 To predict transcription factor binding sites within the *OCT2* promoter region.

2.2.3 To predict the formation of G-quadruplexes within the *OCT2* promoter region.

3. MATERIALS AND METHODS

The recipes for all the reagents used in this study can be found in Appendix A.

3.1 SNP identification and functional analysis

3.1.1 Subjects

Genomic DNA that was extracted from blood samples of 10 unrelated black South Africans was donated by the National Health Laboratory Service (NHLS) for use in this study. The Human Research Ethics Committee of the University of the Witwatersrand approved this study (clearance certificate numbers: M10745 and M120640).

3.1.2 Primer design

Primers were designed to amplify across the *OCT2* gene promoter and were checked for non-target binding to other regions in the genome using the UCSC In-Silico PCR genome browser available at: https://genome.ucsc.edu/cgi-bin/hgPcr (Kent *et al*., 2002). Primer set one, designed with the aid of the OligoAnalyzer 3.1 tool from Integrated DNA Technologies, targeted a larger region of the *OCT2* promoter compared to the preliminary study, which isolated a smaller region of the *OCT2* promoter. The larger *OCT2* promoter region was amplified from all human samples 1−10 (HS1−10). The promoter region was extended because the Ensembl genome browser showed a great number of SNPs upstream of the transcription factor binding region of the *OCT2* promoter and these may have significance for *OCT2* expression.

Primer set two was designed to incorporate two different restriction enzyme sites onto the 5′ ends of the primers in order to clone the *OCT2* promoter region into a pGL4.10 luciferase reporter vector. The forward primer was designed to contain the *Nhe*I restriction site and the reverse primer was designed to contain the *Eco*RV (*Eco*321) restriction site. The NEBcutter V2.0 program was used to ensure that these two restriction sites were not present in the *OCT2* sequence amplified. Both PCR primer sets, listed in Table 2, were designed with similar GC contents and melting temperatures and were synthesised by Inqaba Biotechnical Industries (Pty) Ltd, South Africa.

Table 2: The different primer sets used to target the *OCT2* promoter region.

Primer name	Sequence (5′ to 3′)	T_m (°C)	GC content (%)	Length (bp)
OCT2pFWD2	CCC CTG ATG TGT GAG AGC AG	64.50	60.00	20
OCT2pREV2	CGT AGC GCC TAC ACT GTC TTG	64.52	57.14	21
OCT2FNhe	ATA AGC TAG CCC CCT GAT GTG TGA GAG CAG	70.10	53.33	30
OCT2REco	GAG CGA TAT CCG TAG CGC CTA CAC TGT CTT G	71.25	54.84	31

FWD: Forward, REV: Reverse.

OCT2pFWD2 and OCT2pREV2 = Primer set one with an amplicon size of 1698 bp.

OCT2FNhe and OCT2REco = Primer set two with an amplicon size of 1718 bp.

The underlined region is the recognition site for the respective restriction enzymes used.

3.1.3 DNA extraction

Primers were optimised on genomic DNA that was isolated from Human Embryonic Kidney (HEK293) cells using the phenol-chloroform DNA extraction protocol. HEK293 cells were cultured in Dulbecco’s Modified Eagle Medium (DMEM):F12 in the ratio 3:1 with 10 % fetal bovine serum (FBS) at 37 °C in a humidified atmosphere. A confluent 100 mm dish of HEK293 cells was washed three times with 10 ml of cold 1 x PBS. The cells were scraped in 4 ml of cold 1 x PBS and 1 ml transferred into four 1.5 ml microfuge tubes. The cells were centrifuged at 5 000 rpm for 5 min at 4 °C to pellet the cells and the PBS was removed from all four tubes. The cell pellet from one tube was re-suspended in 50 µl of cold 1 x PBS and the three remaining tubes were stored at −70 °C for future use. The cells were lysed and the RNA was removed by adding 450 µl of genomic DNA lysis buffer together with 1 µl of 10 mg/ml stock RNase A (20 µg/ml final concentration) and this was incubated for 1 h at 37 °C in a water-bath. The protein was removed by adding 2.37 µl of 21.2 mg/ml stock Proteinase-K (100 µg/ml final concentration) and incubated at 50 °C in a water-bath for a further 2 h with swirling every 20 min. Half the volume of buffered phenol was added to the tube, vortexed for 3 s, then half the volume of chloroform: isoamyl alcohol (24:1) added, vortexed for 3 s and then this was centrifuged at 13 000 rpm for 25 min at RT. The upper aqueous phase was carefully removed and placed into a new tube and the previous step repeated on the aqueous phase. The new upper aqueous phase was transferred into a clean 1.5 ml microfuge tube and an equal volume of chloroform: isoamyl alcohol (24:1) was added, vortexed for 3 s and centrifuged at 13 000 rpm for 25 min at 4 °C. This step was repeated on the new upper aqueous phase in a clean 1.5 ml microfuge tube. A 0.1 volume of 3 M sodium acetate and 2 volumes of

100 % ethanol was added to the final aqueous phase and the tube inverted 10 times to mix and then allowed to stand overnight at −20 °C. The DNA was collected into a pellet by centrifugation at 13 000 rpm for 15 min at 4 °C. The DNA pellet was then washed twice with 70 % ethanol and collected by centrifugation at 13 000 rpm for 5 min at 4 °C. Excess ethanol was discarded and the pellet was air dried. The DNA pellet was re-suspended in 50 µl of autoclaved Milli-Q water. The quality and the yield of the DNA isolated was determined using a NanoDrop 1000 spectrophotometer and then DNA was stored at −20 °C. DNA was also resolved on a 1 % (Tris-Acetate-EDTA (TAE) buffer) agarose gel stained with 1 µl of 10 mg/ml stock ethidium bromide (200 ng/ml final concentration) to look at integrity.

3.1.4 Polymerase chain reaction (PCR)

The primer sets were first tested on HEK293 genomic DNA using KAPA *Thermus aquaticus* (*Taq*) DNA polymerase from KAPA Biosystems in order to prevent wastage of the valuable human DNA samples and then optimised using KAPA high-fidelity (HiFi) DNA polymerase from KAPA Biosystems. KAPA HiFi DNA polymerase was preferred for amplifying OCT2 from the genomic DNA from HS1−10 because it has a 3′→5′ exonuclease (proofreading) activity which KAPA *Taq* does not, and as such offers error rates approximately 100 times lower than *Taq* DNA polymerase.

3.1.4.1 KAPA *Taq* DNA polymerase:

The PCR reactions were performed in a 20 µl volume, containing 75−200 ng of genomic DNA, 1 x KAPA *Taq* ReadyMix and 0.4 µM of each primer. Cycling consisted of an initial 2 min activation step at 95 °C, followed by a total of 35 cycles using the following conditions: 95 °C denaturation for 30 s, primer annealing at 60 °C (primer set one) or 65 °C (primer set two) for 30 s, and primer extension at 72 °C for 2 min, and 2 min of final extension at 72 °C and a 4 °C holding step.

3.1.4.2 KAPA HiFi DNA polymerase:

The PCR reactions were performed in a 20 µl volume, containing 75 ng of genomic DNA from HS1−10, 1 x KAPA HiFi HotStart ReadyMix and 0.3 µM of each primer. Cycling consisted of an initial 3 min activation step at 95 °C, followed by a total of 30 cycles using the following conditions: 98 °C denaturation for 20 s, primer annealing at 62 °C (primer set one) or 72 °C (primer set two) for 15 s, and primer extension at 72 °C for 1 min 30 s, and 1 min 30 s of final extension at 72 °C and a 4 °C holding step.

In each PCR reaction a negative control, which contained no template DNA, was used to detect any possible non-specific amplification and contamination. All of the HS1−10 PCR products were separated by 1 % agarose gels (prepared as previously described) and consequently purified using the Thermo Scientific™ GeneJET PCR purification kit to remove any primers, dNTPs, unincorporated labelled nucleotides, enzymes and salts, prior to downstream uses.

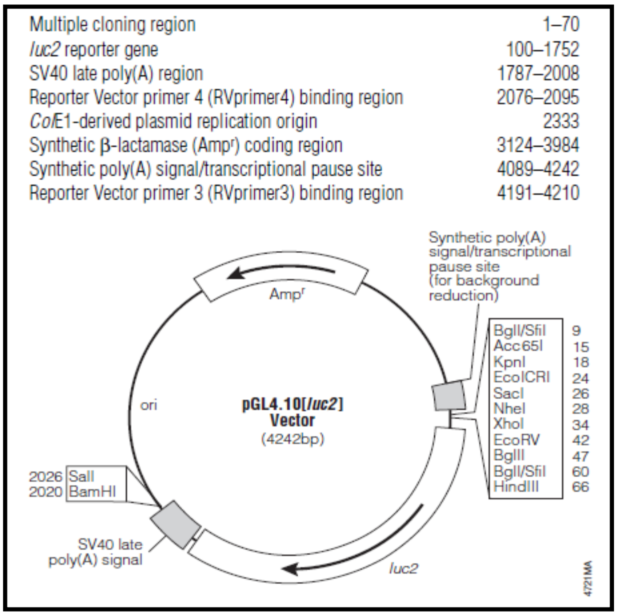
3.1.5 Preparation of chemically-competent cells

Chemically-competent cells were made using ampicillin sensitive *Escherichia coli* (*E. coli)* JM109 bacteria because *E. coli* can be easily manipulated and are easy to grow in the laboratory. An *E. coli* colony was grown in 5 ml LB broth overnight at 37 °C with shaking at 180 rpm. This 5 ml was then added to a 250 ml Erlenmeyer flask containing 50 ml LB broth and incubated at 37 °C with shaking at 180 rpm until an absorbance reading of 0.5−0.6 was observed on the NanoDrop 1000 spectrophotometer. All steps were performed on ice from this point onward. 15 ml of LB broth containing the bacteria was poured into a 50 ml tube and centrifuged at 3000 rpm (1630 g with swing-out rotor at 162 mm) for 5 min at 4 °C. The supernatant was discarded and another 15 ml was added and repeated as before. This was done until all the LB broth containing the bacteria was used. The bacterial pellet was resuspended gently in 15 ml of ice cold 0.1 M $MgCl_2$ and incubated on ice for 30 min. The tube was then centrifuged for 5 min at 4 °C at 1630 g. The supernatant was discarded and the pellet was resuspended in 2 ml of 0.1 M $CaCl_2$ with 15 % glycerol. The chemically-competent bacteria were aliquoted into pre-chilled 1.5 ml microfuge tubes and stored at −70 °C.

3.1.6 Cloning of the OCT2 promoter into the pGL4.10 vector

3.1.6.1 Preparing the pGL4.10[*luc2*] Vector

The pGL4 Luciferase Reporter Vectors can provide a mechanism for evaluating regulation of mammalian gene expression. In this study the pGL4.10[*luc2*] Vector, as shown in Figure 6, was used. It encodes the luciferase reporter gene *luc2* (*Photinus pyralis*) and is intended for high expression and reduced incongruous transcription. This vector contains an ampicillin resistance gene which acts as a selectable marker. It does not contain a promoter but it contains a multiple cloning region to allow for the cloning of a promoter of choice, in this study it will be the *OCT2* promoter. This vector was chosen because it will allow for the functional characterisation of the SNP samples using the luciferase assay and therefore give an indication if SNPs found within the *OCT2* promoter region affect the ability of the promoter to drive gene expression.



Multiple cloning region	1–70
luc2 reporter gene	100–1752
SV40 late poly(A) region	1787–2008
Reporter Vector primer 4 (RVprimer4) binding region	2076–2095
*Col*E1-derived plasmid replication origin	2333
Synthetic β-lactamase (Amp^r) coding region	3124–3984
Synthetic poly(A) signal/transcriptional pause site	4089–4242
Reporter Vector primer 3 (RVprimer3) binding region	4191–4210

Figure 6: Features list and map for the pGL4.10[*luc2*] Vector

A basic transformation protocol was followed as per the instructions provided with the KAPA Rapid Ligation System where 10 ng of the pGL4.10 vector DNA was transformed into 50 µl of competent cells. Briefly, 5µl of the DNA was incubated with the competent cells for 30 min on ice, followed by heat shock at 42 °C for 90 s, followed by an additional 2 min incubation on ice. Cells were then incubated at 37 °C on a shaker at 180 rpm for 1 h and then plated onto LB agar plates containing 100 µg/ml of ampicillin and left at 37 °C overnight. A negative control containing no DNA was used to detect any possible contamination and show that the *E.coli* are susceptible to the ampicillin. The next day, a colony was then picked, added to 3 ml of LB broth containing 100 µg/ml of ampicillin and incubated at 37 °C overnight on a shaker at 180 rpm. The vector DNA was then isolated using the Zyppy™ plasmid miniprep kit (Zymo Research Inc) according to the manufacturer's instructions to ensure that the pGL4.10 plasmid DNA was separated from *E. coli* DNA. The quality and the yield of the vector DNA isolated was checked using a NanoDrop 1000 spectrophotometer and then stored at −20 °C. The DNA was also resolved on a 1 % agarose gel and prepared as previously described.

3.1.6.2 Digestion of the pGL4.10 vector and the PCR products

The purified pGL4.10 vector and the purified HS1−10 PCR products were digested using the *Nhe*I and *Eco*RV restriction endonucleases (Thermo Scientific™) so that the PCR products can be ligated and cloned into the pGL4.10 vector. *Nhe*I produces a sticky-end (G/CTAGC) DNA fragment while *Eco*RV produces a blunt-end (GAT/ATC) DNA fragment, thereby allowing for directional cloning.

A single digestion of the purified pGL4.10 vector was first performed with each enzyme to ensure that the enzymes were in working order. A negative control containing no DNA was used to detect any contamination. 1 µl of either enzyme was used to restrict 1 µg of pGL4.10 vector DNA in the buffer supplied with the enzyme. Once it was confirmed that the enzymes were working a double digestion was done on the pGL4.10 vector and the HS1−10 PCR products. 1 µl of both enzymes were used with 1 µg of pGL4.10 vector DNA or 1 µg of HS1−10 PCR product. Each tube was mixed gently, very briefly centrifuged and then incubated at 37 °C for 16 h. *Nhe*I was inactivated at 65 °C for 20 min and *Eco*RV was inactivated at 80 °C for 20 min. The digested samples were then separated by 1 % agarose gels prepared as previously described.

The double-digested pGL4.10 vector DNA and HS1−10 PCR products were all purified using the Thermo Scientific™ GeneJET PCR purification kit according to the manufacturer's instructions. The DNA samples were eluted in 30 µl of elution buffer and quantified using a NanoDrop 1000 spectrophotometer.

3.1.6.3 Ligation

The Thermo Scientific™ T4 DNA ligation protocol was followed where a 3:1 molar ratio of double-digested insert DNA (HS1−10) to double-digested vector DNA was used in a 20 µl reaction, and this was calculated to give 60.75 ng of insert DNA and 50 ng of vector DNA. A negative control containing only digested vector DNA was used to determine if there was any re-ligation of the vector or if any undigested vector remained following the restriction digest. Each sample was vortexed and centrifuged briefly before being incubated at 22 °C for 2 h and then left overnight at 4 °C for ligation of DNA to proceed. The *OCT2* promoter from HS1−10 was therefore ligated into the pGL4.10 vector and hence pGL4.10-OCT2 promoter recombinant vectors were created.

3.1.7 Transformation

The basic transformation protocol was followed as per the KAPA rapid ligation system where 5 µl of ligation mix was transformed into 50 µl of *E. coli* chemically-competent cells. The pGL4.10-OCT2 promoter recombinant vector as well as the negative control which contained no insert DNA was transformed into *E.coli* as previously described. A positive control, which contained 3 ng of undigested pGL4.10 vector and a negative control which contained only competent cells, were also included. These controls were used to ensure that the competent cells were still viable, no contamination had occurred and that cells were still susceptible to ampicillin. The antibiotic ampicillin was added to the LB agar plates to select only for *E. coli* cells that had been transformed with the pGL4.10 vectors.

3.1.8 Isolating the pGL4.10-OCT2 promoter

3.1.8.1 Colony PCR

A colony PCR for each HS1−10 was done to determine which *E. coli* cells had been transformed with the pGL4.10-OCT2 promoter recombinant vectors and which cells had been transformed with the empty pGL4.10 vector. This involved selecting a colony from the plate and culturing it in 50 µl of LB broth with ampicillin at 37 °C for 1 h. 2 µl were then used in a colony PCR. A third primer set, listed in Table 3, was designed to bind to either side of the multiple cloning region of the pGL4.10 vector. These primers were used for each colony PCR. If the vector was empty then the PCR would amplify a region of 243 bp. If the vector contained the *OCT2* promoter DNA insert then the PCR would amplify a region of 243 bp plus the size of the *OCT2* promoter fragment. Therefore since primer set two yielded a 1718 bp PCR product, a colony PCR should show a size of 1961 bp on the agarose gel. Multiple colonies per human DNA sample were analysed by colony PCR.

The colony PCR reactions were performed in a 20 µl volume, containing 2 µl of the colony grown in 50 µl of LB broth for 1 h at 37 °C, 1 x KAPA *Taq* ReadyMix and 0.4 µM of each primer from primer set three. Cycling consisted of an initial 2 min activation step at 95 °C, followed by a total of 25 cycles using the following conditions: 95 °C denaturation for 30 s, primer annealing at 57 °C (primer set three), and primer extension at 72 °C for 2 min, and 2 min of final extension at 72 °C and a 4 °C holding step. These were then resolved on 1% agarose gels prepared as previously described to identify which colonies contained the recombinant vectors.

Table 3: Primer set three used for a colony PCR.

Primer name	Sequence (5′ to 3′)	T_m (°C)	GC content (%)	Length (bp)
pGL4.10 FWD	CTA GCA AAA TAG GCT GTC CC	60.40	50.00	20
pGL4.10 REV	TTC ATG GCT TTG TGC AGC T	58.01	47.37	19

FWD: Forward, REV: Reverse.

pGL4.10 FWD and pGL4.10 REV = Primer set three with an amplicon size of 1961 bp if the OCT2 promoter insert is present or 243 bp if it is not present.

3.1.8.2 pGL4.10-OCT2 promoter isolation

The pGL4.10 vectors which contained the *OCT2* promoter DNA insert from HS1−10 were isolated using the alkaline lysis treatment method in order to separate the vector DNA from *E. coli* DNA.

After identification of the colonies of interest by PCR, the remainder of 50 µl was used to inoculate 3 ml of LB broth containing 100 µg/ml of ampicillin and incubated at 37 °C overnight on a shaker at 180 rpm. The next day the bacteria was harvested by centrifuging 2 ml of the bacterial culture in a clean 2 ml microfuge tube at 12 000 rpm for 1 min. The remaining 1 ml was made into a glycerol stock. After centrifugation, the supernatant was discarded and the bacterial pellet was resuspended and vortexed in 110 µl of 25 mM Tris-HCl (pH 8.0) and 10 mM EDTA, to which 100 µl of 0.4 M NaOH and 2 % SDS was added and mixed gently by inverting a couple of times at RT. 120 µl of cold 5 M potassium acetate (pH 5.5) was added and mixed thoroughly by inverting. This mixture was then incubated at RT for 3 min and then centrifuged at 12 000 rpm for 4 min. The supernatant was transferred into a new 1.5 ml microfuge tube and 200 µl of isopropanol was added and mixed thoroughly by inverting a couple of times. The mixture was then incubated at RT for 1 min and then centrifuged at 13 000 rpm for 1 min. The isopropanol was then removed from the DNA pellet that remained behind and 500 µl of cold 70 % ethanol was added and mixed by inverting a couple of times. This mixture was then centrifuged at 13 000 rpm for 1 min, the excess ethanol was discarded and the DNA pellet air dried for 10 min to evaporate any residual ethanol. The DNA pellet was re-suspended in 50 µl of autoclaved Milli-Q water and the quality and the yield of the DNA isolated were checked using a NanoDrop 1000 spectrophotometer. The DNA was then resolved on a 1 % agarose gel, prepared as previously described, and DNA stored at −20 °C. The purified vector for each HS1−10 was sent for sequencing.

3.1.9 Sequencing and data analysis

The *OCT2* promoter within the pGL4.10 vector for each HS1−10 was sequenced at Inqaba Biotechnical Industries (Pty) Ltd. The Sanger DNA sequencing method was performed on these samples using the BigDye$^{®}$ Terminator v3.1 Cycle Sequencing on an ABI 3500XL sequencer. The sequencing used the pGL4.10FWD (listed in Table 3) and OCT2pREV2 (listed in Table 2) primers and the results returned were analysed using the FinchTV software. SNPs in the *OCT2* promoter region were identified and the SNPs identified in the preliminary study were confirmed. The identification of SNPs was done through a multiple sequence alignment (MSA) on Clustal Omega available at: https://www.ebi.ac.uk/Tools/msa/clustalo/ where a comparison was made to the WT *OCT2* sequence that was obtained through the Ensembl database (Yates *et al*., 2015). Since the SNPs identified in the preliminary study were heterozygous, either the WT or the mutant strand could be cloned into the vector. For HS7 and HS9 the cloning of the WT occurred and therefore a site-directed mutagenesis was performed on these two samples to introduce the SNP of interest. For all other SNP containing samples, the SNP-containing allele was successfully cloned into the pGL4.10 vector.

3.1.10 Site-directed mutagenesis

Site-directed mutagenesis was performed on a pGL4.10 vector that contained the WT *OCT2* promoter, in order to introduce the SNP of interest as found in the HS7 and HS9 samples. Primers, listed in Table 4, were designed directly adjacent to each other on opposite strands of the vector DNA so that the entire plasmid of nearly 6 kb could be amplified. The forward primer was designed to include the SNP of interest, which was introduced at the 5′ end of the primer by substitution of the WT nucleotide for the mutant nucleotide. Site-directed mutagenesis was followed as per the instructions on the KAPA HiFi DNA polymerase site-directed mutagenesis protocol.

Briefly, primers were first 5′ phosphorylated using T4 polynucleotide kinase (New England Biolabs (NEB)) to allow for subsequent ligation of the PCR products following amplification. The PCR amplification generates large numbers of linear duplex molecules, which significantly outnumber the original WT template. The mutagenesis PCR reactions were performed in a 50 µl volume, containing 1 ng of the vector containing the WT *OCT2* promoter DNA insert, 1 x KAPA HiFi HotStart ReadyMix and 0.3 µM of each primer. PCR mixtures consisting of only the primers and nuclease free water were incubated at 95 °C for 10 min to melt any secondary structures that may form between the primers. Cycling consisted of an initial 3 min activation step at 95 °C, followed by a total of 16 cycles using the following conditions: 98 °C denaturation for 20 s, primer annealing at 66 °C for HS7 and 65 °C for HS9 for 15 s, and primer extension at 72 °C for 3 min, and 5 min of final extension at 72 °C and a 4 °C holding step. A negative control without KAPA HiFi DNA polymerase was included to detect whether there was any parental DNA that was carried over to the next step. All of the PCR products were separated by 1 % agarose gels, prepared as previously described. The residual methylated template and any hemi-methylated DNA was digested using the methylation-specific *Dpn*I restriction endonuclease (NEB). T4 DNA ligase (Thermo Scientific™) was then used to circularise the linear mutated DNA and then the ligation mixture was used to transform competent *E. coli* cells. A colony PCR was then performed using primer set three (listed in Table 3) where multiple colonies from the mutated HS7 and HS9 DNA samples were analysed. Colonies with positive bands of 5960 bp were incubated overnight at 37 °C in LB broth containing 100 µg/ml of ampicillin. Alkaline lysis was performed and the *OCT2* promoter within the vector was sequenced at Inqaba Biotechnical Industries (Pty) Ltd using the pGL4.10FWD (Table 3) and OCT2pREV2 (Table 2) primers. The SNPs of interest were confirmed through the analysis of the sequences. We successfully incorporated the SNPs found in HS7 and HS9, however, unfortunately, despite many attempts including temperature modifications and primer optimisation for HS7, a secondary mutation occurred at the beginning of where the forward primer binds. Therefore we were unable to functionally characterise the SNP identified in this sample.

Table 4: The different primers designed for site-directed mutagenesis.

Primer name	Sequence (5′ to 3′)	T_m (°C)	GC content (%)	Length (bp)
HS7FWD	ATA AAC ATG ACA GGC ATC CTG GG	62.77	47.83	23
HS7REV	TAT AAC TGG TTC TCC ACA GGC TTG	62.86	45.83	24
HS9FWD	AGA AGG ACG TGC ACC CCC G	66.64	68.42	19
HS9REV	GCA CTT CCC AAA TGC CCT GAG CTC	67.98	58.33	24

FWD: Forward, REV: Reverse

The underlined nucleotide is the mutant nucleotide that was substituted into the sample.

3.1.11 Functional characterisation of each SNP using the luciferase reporter assay

3.1.11.1 Luciferase activity measurement

The luciferase reporter assay was employed in the Medical Research Council cell strain 5 (MRC-5), a human fetal lung fibroblast cell line that is easily transfectable. This cell line was maintained in a humidified atmosphere of 5 % CO_2 and 95 % air at 37 °C and cultured in

DMEM supplemented with 10 % FBS with a 1 % penicillin-streptomycin antibiotic. After the culture had reached 70 % confluence, the cells were trypsinised and seeded onto 96-well culture plates at a density of 12 000 cells per well. After 24 h when the wells had reached approximately 80 % confluency, 200 ng of the vectors containing the different SNPs, were transfected into the MRC-5 cells in quadruplicate using the Turbofect transfection reagent by Thermo Scientific™. An empty (promoterless) pGL4.10 vector was used as a negative control. After transfection, the cells were incubated for 24 h at 37 °C. After 24 h the cell culture plates were left to stand at RT for 15 min and then the luciferase assay was performed by adding 100 µl of the Steady Glo® luciferase assay substrate (Promega) to each well. After 5 min at RT the samples in each well were transferred onto a 96-well white plate. The luciferase activity, in the form of light, was measured three times using the Glomax 96 microplate luminometer (Promega) to ensure accuracy and validity. The expression of luciferase was determined for each vector and these were compared to the WT *OCT2* promoter. Statistical significance was analysed using a paired t-test and $p < 0.05$ was considered to be statistically significant. Therefore, the effect that these SNPs may have on *OCT2* gene expression in the South African black population can be inferred from these results.

3.2 Bioinformatic analysis

3.2.1 Identification of SNPs within the *OCT2* promoter and coding regions

SNPs were identified within the *OCT2* promoter and coding regions in various studies, as listed in Table 5, and compared to data obtained from the 1000 Genomes Project. These studies were conducted on individuals within the South African black population and can also be classified as Bantu-speaker individuals. Where the ethnic group of the individuals in the studies was known it has been stated. The different ethnic groups included Zulu, Sotho and Xhosa. SNPs were analysed using various bioinformatic tools available online.

With the assistance of my co-supervisor, Dr Ananyo Choudhury, data from studies which have not been made available to the public as yet, and included individuals from different ethnic groups within the South African black population, were made available for use in this study. These studies included data from 40 Bantu-speaking individuals (Carstens, N., *et al*. 2016, personal communication) and 15 Xhosa, Zulu and Sotho individuals (Ramsay, M., *et al*. 2016, personal communication). Data which included sequencing the whole genomes of 100 Zulu individuals (Gurdasani *et al*., 2014) has been published, however the three studies do not specifically analyse the *OCT2* gene. Raw data in various file formats that was specific to the chromosome locations of the *OCT2* gene promoter and coding regions was made available for analysis. The information in the files was combined into one and analysed using bioinformatic tools that are available online. These tools included the wANNOVAR tool available at: http://wannovar.usc.edu/ (Wang, Li & Hakonarson, 2010; Chang & Wang, 2012), the Ensembl Genome Browser available at: http://www.ensembl.org/, and the National Center for Biotechnology Information (NCBI) dbSNP database available at: http://www.ncbi.nlm.nih.gov/projects/SNP/. SNPs were identified through the results obtained and these were summarised in tables listed in the results section. A study by Jacobs *et al*. had already identified SNPs in 96 Xhosa individuals within the *OCT2* gene, so therefore this data was added to the tables with the results from the other studies (Jacobs *et al*., 2015).

The *OCT2* promoter region analysis compared minor allele frequencies (MAF) values from the SNPs identified in the four studies mentioned as well as from this study of the 10 South African black individuals. The *OCT2* coding region only compared MAF values from the SNPs identified in the four studies since this study did not sequence the *OCT2* coding regions. These MAF values from the different studies were then compared to the MAF values obtained from the 1000 Genomes Project, which included the African population and overall MAF values. The overall MAF values excluding the African population was calculated to determine the frequencies within the non-African populations (European, American, South Asian and East Asian). This allowed us to determine how similar or dissimilar the South African black population is in comparison to others.

Table 5: The studies used for the analysis and comparison of the SNPs identified in the *OCT2* promoter and coding regions.

Source	Population	Description	*n*	Type of sequencing	Reference
Present study	South African black population	South African	10	1700 bp Promoter (Sanger)	
Carstens and colleagues	Bantu-speaking	South African	40	Exome (HC)	Carstens *et al*., 2016 (Personal communication)
Ramsay and colleagues	Xhosa, Zulu and Sotho	South African	15	WGS (HC)	Ramsay *et al*., 2016 (Personal communication)
Gurdasani and colleagues	Zulu	South African	100	WGS (LC)	Gurdasani *et al*., 2014
Jacobs and colleagues	Xhosa	South African	96	500 bp Promoter and WGS (Sanger)	Jacobs *et al*., 2015
1000 Genomes Project – Phase 3	AFR	African	661	WGS (LC)	1000 Genomes Project Consortium, 2015
	ACB	African Caribbean in Barbados	96		
	ASW	African Ancestry in Southwest US	61		
	ESN	Esan in Nigeria	99		
	LWK	Luhya in Webuye, Kenya	99		
	GWD	Gambian in Western Division, The Gambia	113		
	MSL	Mende in Sierra Leone	85		
	YRI	Yoruba in Ibadan, Nigeria	108		

n – number of subjects; **WGS** – Whole genome sequencing; **LC** – Low coverage; **HC** – High coverage.

The possible effect of the *OCT2* promoter SNPs on *OCT2* regulation were predicted using the RegulomeDB database version 1.1. available at: http://www.regulomedb.org/ (Boyle *et al*., 2012). This database uses a total of 962 experimental data sets from the ENCODE project, published literature, and public datasets, as well as manual annotations and computational predictions to annotate SNPs and identify their presumed regulatory potential in the human genome (Boyle *et al*., 2012). The SNPs we identified were submitted to the database, using the dbSNP IDs or the chromosome co-ordinates for novel SNPs. A RegulomeDB score ranking from 1 to 6 was given; the numbers represent the supporting datatypes, such as DNase sensitivity and chromatin states, that the SNP has been identified to contain. The lower the score the higher the chance that there is a possible regulatory effect for that particular SNP.

The amino acid substitution consequences for any missense SNPs identified within the *OCT2* coding region were predicted using the Variant Effect Predictor (VEP) tool available from Ensembl at: http://www.ensembl.org/Tools/VEP. The Sorting Intolerant From Tolerant (SIFT) and Polymorphism Phenotyping version 2 (PolyPhen-2) prediction scores from the VEP tool were noted and it was from these scores that the predicted consequence of each missense SNP was made. The SIFT program uses sequence homology to assign a score; scores less than 0.05 are predicted to be damaging or deleterious and scores greater than 0.05 are predicted to be tolerated (Ng & Henikoff, 2003). The PolyPhen-2 tool uses protein structure to predict the consequence of each missense SNP where a score of 0 indicates that the SNP is probably damaging, 1 indicates possibly damaging, 2 indicates benign and 3 indicates unknown (Adzhubei *et al*., 2013).

3.2.2 Prediction of transcription factor binding sites within the *OCT2* promoter region

The potential transcription factor binding sites within the *OCT2* promoter were predicted using the ALGGEN PROMO program version 3.0.2 available at: http://alggen.lsi.upc.es/cgibin/promo_v3/promo/promoinit.cgi?dirDB=TF_8.3 (Messeguer *et al*., 2002; Farré *et al*., 2003). This program uses data from version 8.3 of the TRANSFAC database. Homo sapiens were chosen for the current factor species and the current site species. The WT or the mutant allele with a 15 bp flanking sequence added onto each side was submitted as the query sequence and a 5 % maximum matrix dissimilarity rate (95 % similarity rate) was used. Both the forward and reverse strands can be searched for matches although we submitted sequences that were on the reverse strand. This program allowed us to identify whether the putative transcription factor binding sites would be affected by the presence of SNPs or INDELs.

3.2.3 Prediction of the formation of G-quadruplexes within the *OCT2* promoter region

G-quadruplex formation within the *OCT2* promoter region was predicted using the QGRS Mapper software program available at: http://bioinformatics.ramapo.edu/QGRS/index.php (Kikin, D'Antonio & Bagga, 2006). The *OCT2* promoter sequence with each SNP and INDEL separately was compared to the *OCT2* promoter WT sequence to determine if there was a difference in the formation of G-quadruplexes. The default settings of this program were used for each SNP prediction: maximum QGRS length of 30, minimum G-Group size of 2, and the loop size from 0 to 36. There was no information indicating which strand of DNA to use for the prediction so the reverse strand was used since the *OCT2* transcript is found on this strand.

4. RESULTS

4.1 SNP identification and functional analysis

A genomic DNA extraction from HEK293 cells was carried out (Figure 7) because DNA was required to optimise the primers designed to amplify the *OCT2* promoter region. A single high molecular weight band with no smearing of the DNA indicated that the DNA was intact and not fragmented or degraded.

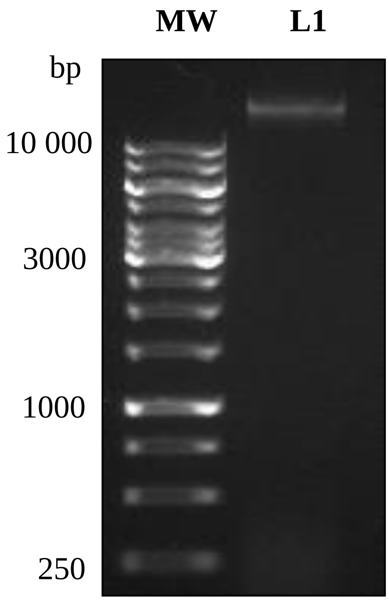


Figure 7: Genomic DNA extractions from HEK293 cells. Molecular weight (MW) – GeneRuler™ 1kb DNA ladder; L1: 300 ng HEK293 genomic DNA. The single band found above 10 000 bp represents intact genomic DNA. A 1 % agarose gel electrophoresis is shown.

Primer set one was optimised using KAPA HiFi DNA polymerase on HEK293 genomic DNA at annealing temperatures ranging from 60 – 64 °C as shown in Figure 8. Sharp bright bands were observed for all annealing temperatures at the expected size of 1698 bp indicating that the correct product was successfully amplified. It was decided that 62 °C would be the annealing temperature for amplifying the *OCT2* promoter region in the human samples since non-specific amplication at approximately 500 bp was observed for 60 °C. Primer dimers were observed for all annealing temperatures and a therefore the Thermo Scientific™ GeneJET PCR purification kit was used in order to send a clean PCR product for sequencing (Figure 9).

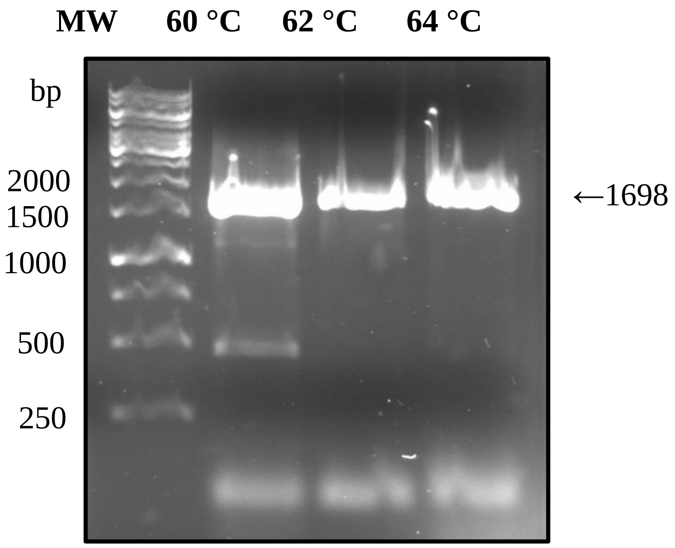


Figure 8: Primer set one optimisation on HEK293 genomic DNA. The 1698 bp *OCT2* promoter region was amplified at various annealing temperatures ranging from 60 – 64 °C. 75 ng of genomic DNA was used in each reaction using KAPA HiFi DNA polymerase. A 1 % agarose gel electrophoresis is shown. Molecular weight (MW) – GeneRuler™ 1kb DNA ladder. Primer dimers can be seen towards the bottom of each lane.

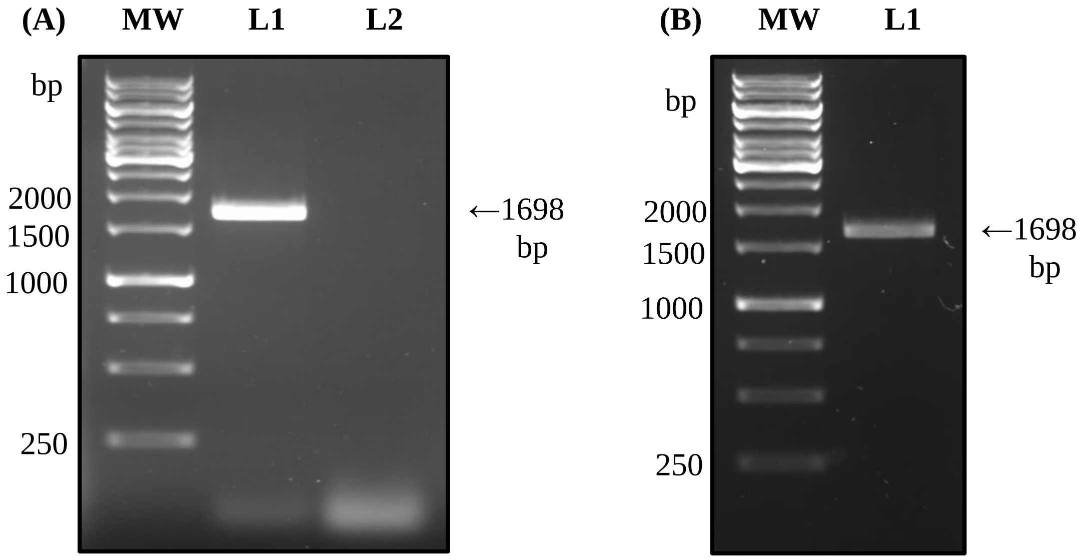


Figure 9: PCR products of HS1 using primer set one. PCR with an annealing temperature of 62 °C was used to amplify the 1698 bp *OCT2* promoter region using KAPA HiFi DNA polymerase. **(A) Before PCR purification.** L1: 75 ng genomic DNA; L2: negative control. Primer dimers were observed in both lanes. **(B) After PCR purification.** L1: 75 ng genomic DNA. The primer dimers were removed during purification. A 1 % agarose gel electrophoresis is shown. Molecular weight (MW) – GeneRuler™ 1kb DNA ladder.

The purified PCR product of HS1 was sequenced using the OCT2pFWD2 and OCT2pREV2 primers. The OCT2pREV2 primer returned good quality clear sequencing results whereas the OCT2pFWD2 primer did not. It was thus decided that the *OCT2* promoter would be cloned into a pGL4.10 vector and the *OCT2* insert within the vector would be sequenced using vector specific primers.

A second primer set (OCT2FNhe and OCT2REco primers) was designed to contain two different restriction endonuclease sites at their 5′ ends respectively in order to clone the promoter into a pGL4.10 luciferase reporter vector. This primer set was optimised using Kapa *Taq* DNA polymerase using HEK293 genomic DNA. An annealing temperature of 65 °C was first tried and a successful PCR product of 1718 bp was observed using 200 ng of genomic DNA (Figure 10A).

Since some of the human samples had low DNA concentrations, we also determined whether we could amplify the promoter region from lower DNA concentrations and we found that even with 75 ng of DNA a good amount of PCR product was obtained (Figure. 10B).

Primer set two was then optimised using KAPA HiFi DNA polymerase at annealing temperatures ranging from 66 – 74 °C (Figure 11). PCR products were observed for annealing temperatures ranging from 70 – 74 °C and we decided that 72 °C would be the annealing temperature for amplifying the *OCT2* promoter region in the human samples.

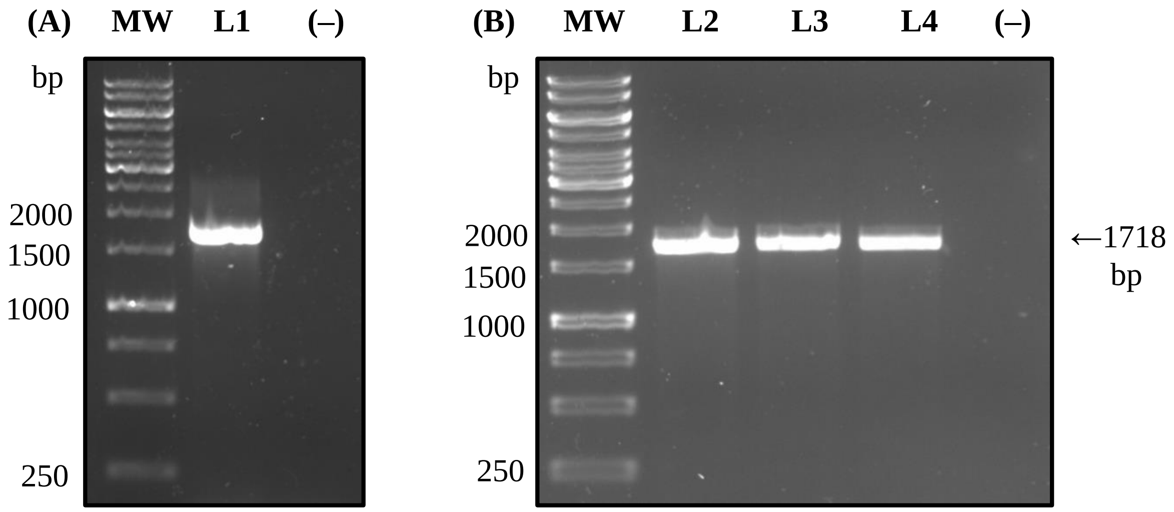


Figure 10: Different amounts of HEK293 genomic DNA amplified by PCR. The *OCT2* promoter region was amplified using primer set two at an annealing temperature of 65 °C. KAPA *Taq* DNA polymerase was used. Molecular weight (MW) – GeneRuler™ 1kb DNA ladder; L1 and L2 – 200 ng genomic DNA; L3 – 100 ng genomic DNA; L4 – 75 ng genomic DNA; (–) indicates the negative control. A distinct bright band at 1718 bp was observed for all the different amounts of DNA used. 1 % agarose gel electrophoresis is shown.

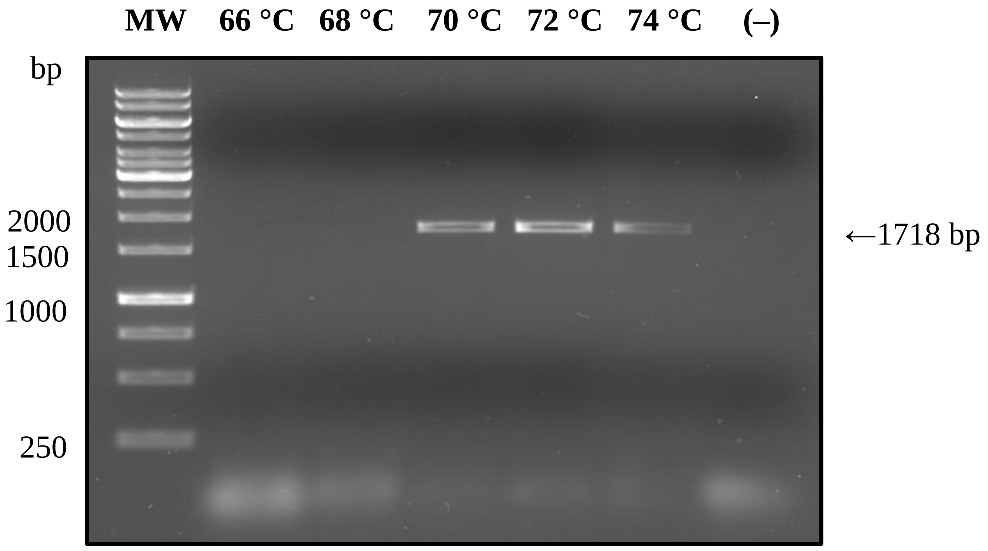


Figure 11: PCR optimisation of primer set two using HEK293 genomic DNA. The *OCT2* promoter region was amplified at various annealing temperatures ranging from 66 – 74 °C using KAPA HiFi DNA polymerase. Molecular weight (MW) – GeneRuler™ 1kb DNA ladder, (–) indicates the negative control. 1718 bp PCR products are clearly observed for annealing temperatures of 70 – 74 °C. A 1 % agarose gel electrophoresis is shown.

Following the optimisation of primer set two, each HS was amplified by PCR using 75 ng of genomic DNA at the chosen annealing temperature of 72 °C. HS1 was amplified first where at least three reactions per sample were performed to ensure enough DNA was recovered after the PCR purification step (Figure 12). The rest of the HS were also amplified and purified in the same way (gel images are not shown).

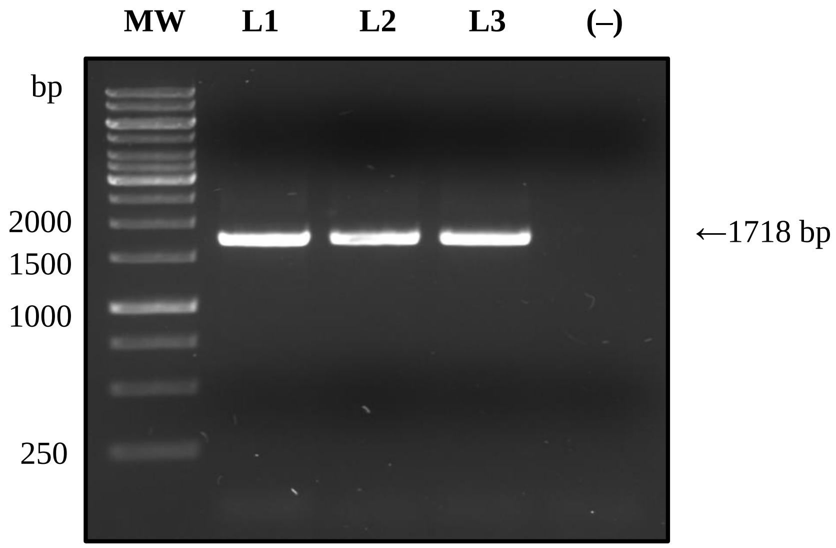


Figure 12: PCR amplicons of HS1 with primer set two. The *OCT2* promoter region was amplified using primer set two at an annealing temperature of 72 °C. KAPA HiFi DNA polymerase was used. Molecular weight (MW) – GeneRuler™ 1kb DNA ladder; L1 – L3: 75 ng of genomic DNA; (–) indicates the negative control. A distinct bright band at 1718 bp was observed for all the lanes. A 1 % agarose gel electrophoresis is shown.

Single digestions on the circular pGL4.10 vector were done to confirm that each restriction endonuclease was in working order. Linear DNA fragments were observed for both *Nhe*I and *Eco*RV (Figure 13). Since each restriction enzyme was working and no contamination was observed for the negative control, a double digestion was done on both the circular pGL4.10 vector and the PCR products of each HS1−10. A double digestion of HS1 was done first (Figure 14) and the rest of the double digestions are not shown. It was observed that the 6 kb undigested pGL4.10 vector migrated further in the gel compared to the digested linear pGL4.10 vector, as expected from the conformation of circular plasmids which causes them to migrate further.

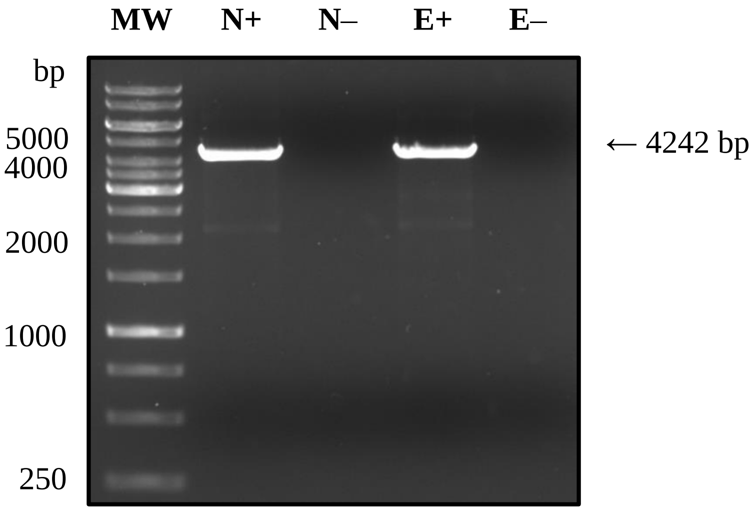


Figure 13: A single digestion of the purified pGL4.10 vector using *Nhe*I and *Eco*RV restriction endonucleases. Molecular weight (MW) – GeneRuler™ 1kb DNA ladder. N+: *Nhe*I positive control; N–: *Nhe*I no DNA negative control; E+: *Eco*RV positive control; E–: *Eco*RV no DNA negative control. Both positive controls show a distinct bright linear pGL4.10 vector fragment at 4242 bp indicating that each resrtriction enzyme is working. A 1 % agarose gel electrophoresis is shown.

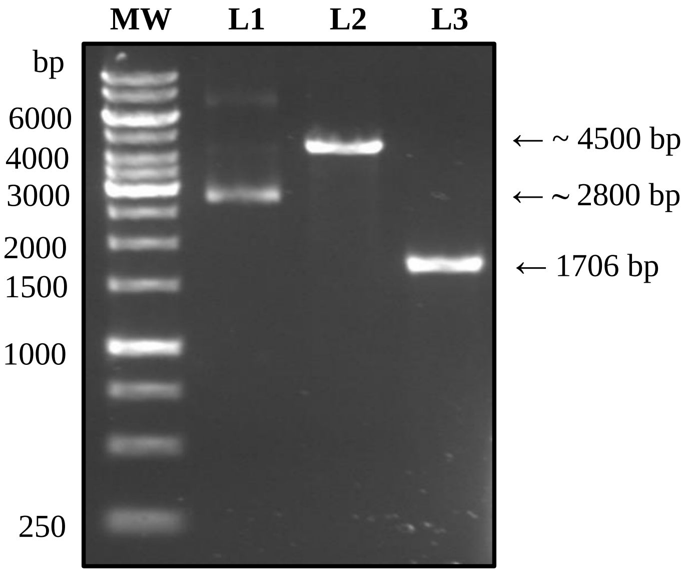


Figure 14: A double digestion of the purified pGL4.10 vector and HS1 PCR product using *Nhe*I and *Eco*RV restriction endonucleases. Molecular weight (MW) – GeneRuler™ 1kb DNA ladder. L1: circular undigested pGL4.10 vector at approximately 2800 bp; L2: linear double digested pGL4.10 vector at approximately 4500 bp; L3: double digested PCR product of HS1 at 1706 bp. A 1 % agarose gel electrophoresis is shown.

The double digested vector and PCR products were purified, quantified and then ligated together using T4 DNA ligase. The ligation mixture was transformed into *E. coli* chemically-competent cells and spread onto LB agar plates containing the ampicillin antibiotic. A few colonies were observed on the negative control which contained only digested vector DNA, indicating that there was either re-ligation of the vector or that some undigested vector remained following the restriction digest. The transformation was however successful because slightly more colonies were observed for HS1, HS7 and HS9 compared to the no insert DNA negative control (Figure 15). No contamination occurred as there were no colonies observed on the plate which contained no vector or insert DNA and this also showed the *E.coli* were susceptible to ampicillin.

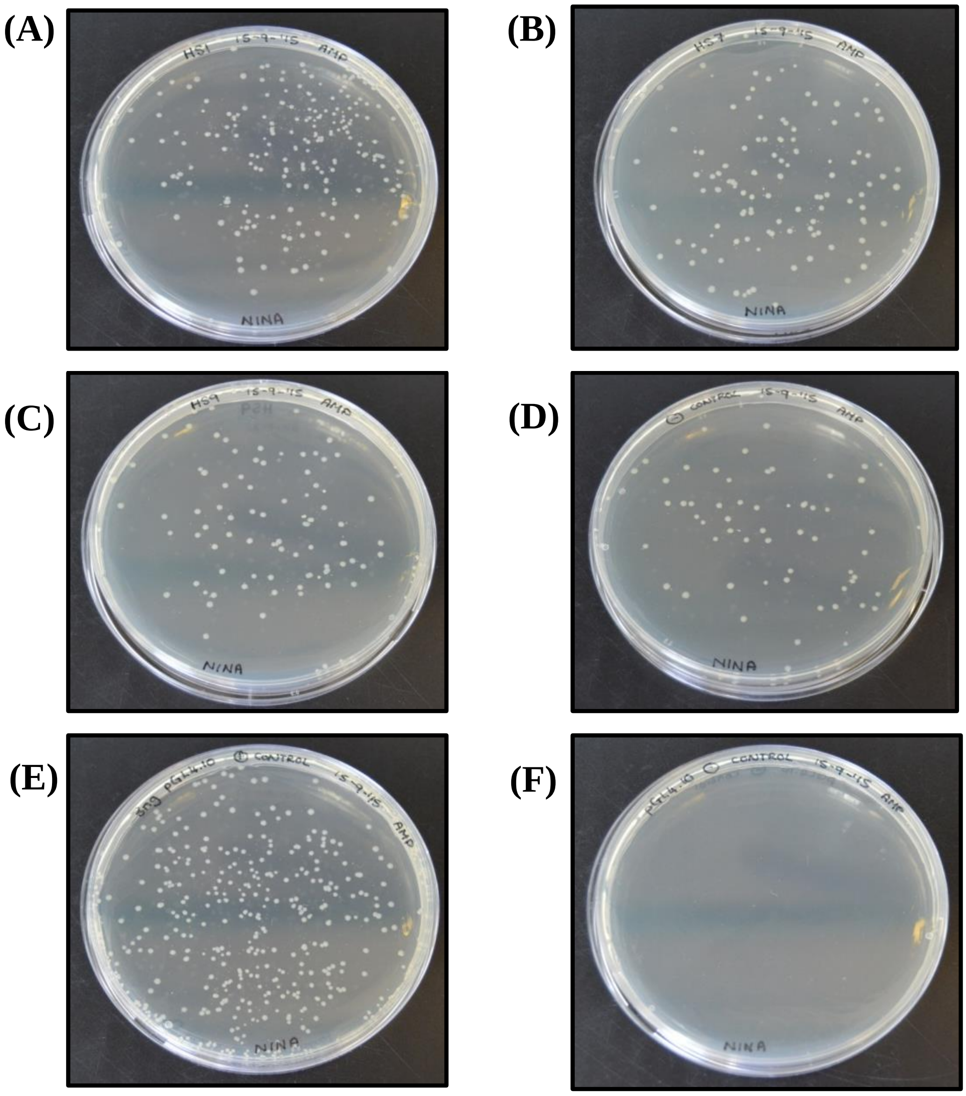


Figure 15: *E. coli* JM109 competent cells transformed with the pGL4.10 vector that contain *OCT2* promoter inserts from different samples. (A) HS1 transformed cells; (B) HS7 transformed cells; (C) HS9 transformed cells; (D) Negative control which contains only digested pGL4.10 vector; (E) 3 ng of undigested pGL4.10 vector; (F) Negative control which contains no pGL4.10 vector and no DNA insert. 200 µl of cell suspension was spread onto each LB agar plate which contains 100 µg/ml of ampicillin antibiotic.

Using primer set three, a colony PCR on HS1 (Figure 16) was done to differentiate between which *E. coli* colonies were transformed with the pGL4.10-OCT2 promoter recombinant vector and which colonies were transformed with the empty pGL4.10 vector. Colonies which contained the *OCT2* promoter insert were expected to give a PCR product of 1961 bp and those colonies which were empty and had no *OCT2* promoter insert were expected to give a band of

243 bp. A colony PCR was also done for the remaining samples but these results are not shown.

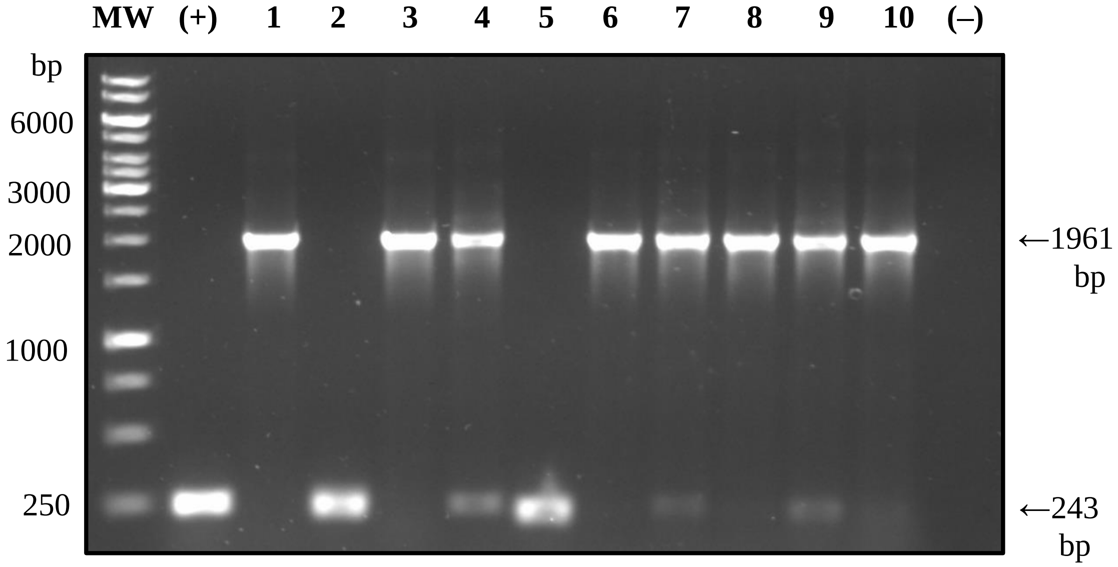


Figure 16: A colony PCR for HS1 using primer set three. Molecular weight (MW) – GeneRuler™ 1kb DNA ladder. (+): Positive control which contains the empty pGL4.10 vector; (–) indicates the negative control which contains no template DNA; 1–10: the number of colonies chosen for the colony PCR. Colonies 1, 3, 4, 6 – 10 show a bright distinct band at 1961 bp indicating that these colonies contain the *OCT2* promoter insert. Colonies 2 and 5 were the same as the positive control which indicate that these colonies do not contain the *OCT2* promoter insert at 243 bp. A 1 % agarose gel electrophoresis is shown.

Colonies which contained the *OCT2* inserts within the vector were sequenced using vector specific primers and SNPs were identified through a MSA on Clustal Omega where a comparison was made to the WT *OCT2* sequence. Since we had previously identified that the SNPs found in HS1, HS7 and HS9 are all heterozygous, either the WT or the mutant *OCT2* promoter sequence could be cloned into the pGL4.10 vector. The mutant sequence for HS1 was successfully cloned into the vector but the WT sequences for HS7 and HS9 were cloned, despite sequencing many different colonies, therefore a site-directed mutagenesis was performed on these two samples to introduce the SNP.

The vector containing the WT *OCT2* DNA insert from each sample was amplified by PCR, using the primers listed in Table 4, and then treated with the *Dpn*I restriction endonuclease to remove any parental DNA from being carried over into the subsequent steps (Figure 17). No parental DNA was carried over as no DNA bands were observed in the negative control lanes. The positive control lanes do not show bright DNA bands as low amounts of DNA was used to prevent the PCR from increasing the chance of any secondary mutations from occurring. An empty linear pGL4.10 vector is 4242 bp in size but the pGL4.10 vector containing the *OCT2* promoter insert of 1718 bp will have a total size of 5960 bp.

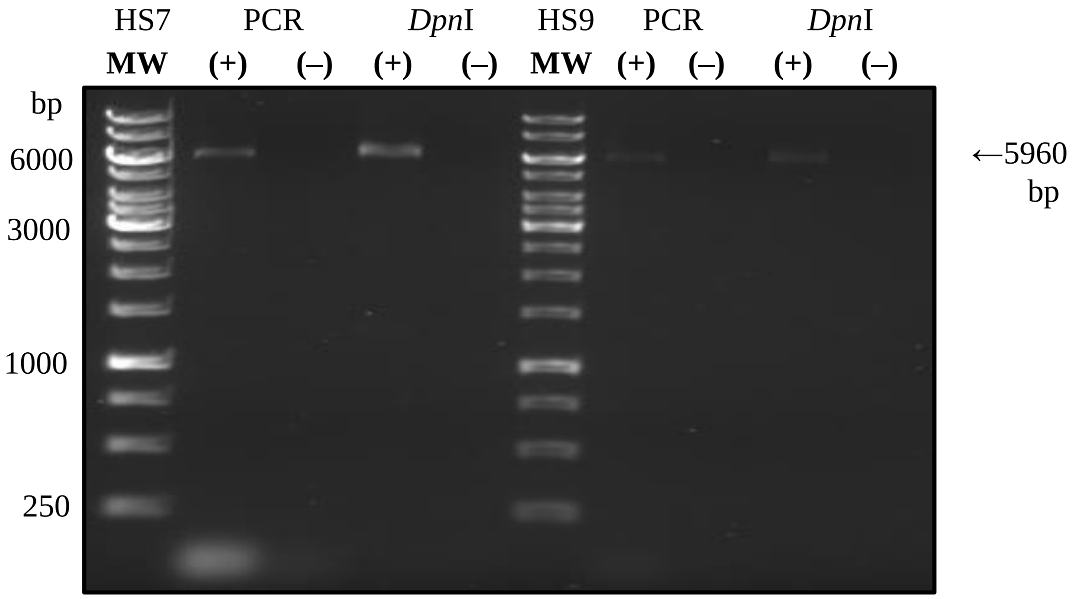


Figure 17: A site-directed mutagenesis PCR and *Dpn*I treatment for HS7 and HS9. HS7 is on the left of the gel and HS9 is on the right side of the gel. Molecular weight (MW) – GeneRuler™ 1kb DNA ladder. (PCR+): Positive control; (PCR–): Negative control which contains no KAPA HiFi DNA polymerase; (*Dpn*I +): Treatment of the PCR+ with *Dpn*I endonuclease. (*Dpn*I –): Treatment of the PCR– with *Dpn*I endonuclease. Each PCR+ produced a linear pGL4.10 vector containing the *OCT2* promoter insert at 5960 bp. A 1 % agarose gel electrophoresis is shown.

The transformation of *E.coli* with DNA modified by site-directed mutagenesis was successful because colonies were observed for the HS7 and HS9 plates and no colonies were observed on the negative control plates (Figure 18). No contamination occurred as there were no colonies observed on the plate which contained no vector or DNA insert.

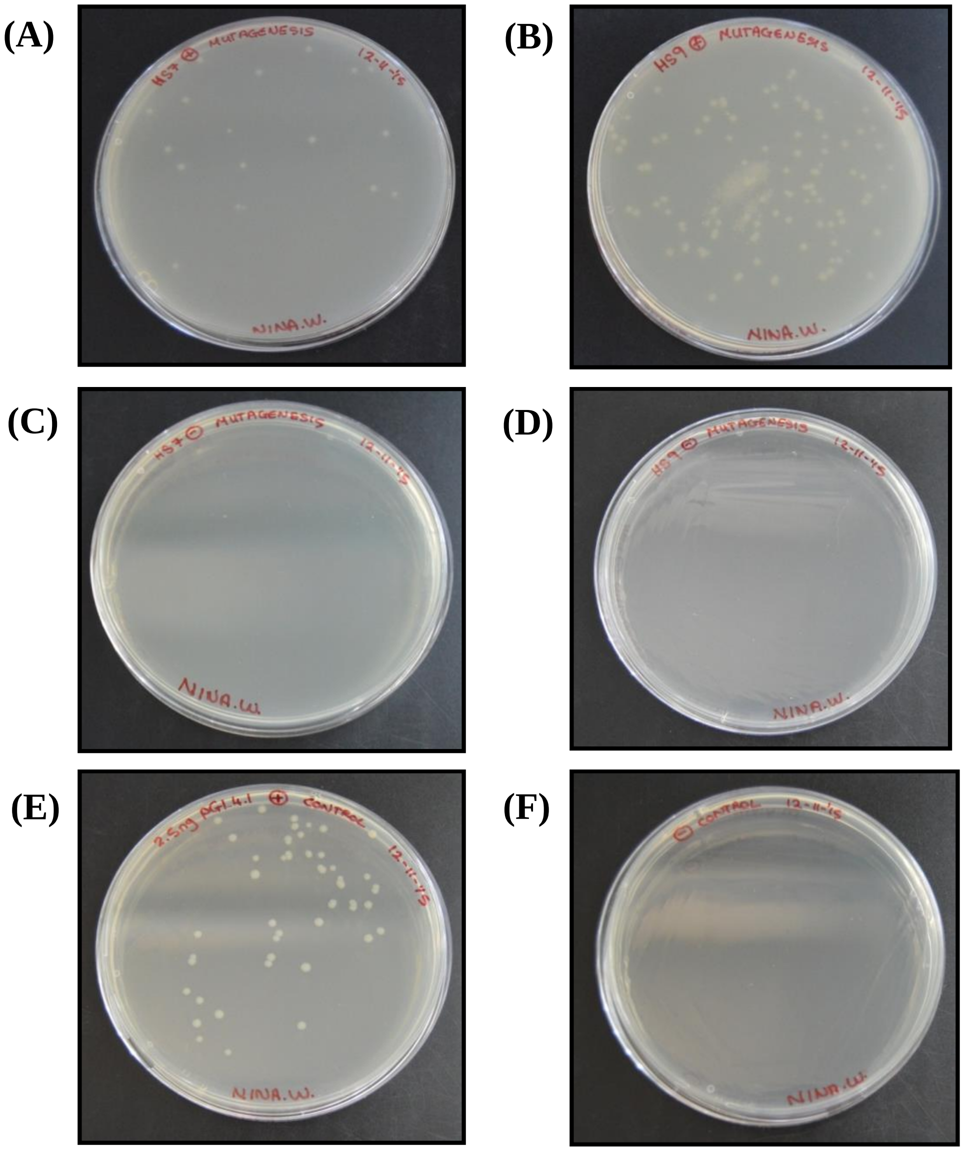

(A)
HS7 ⊕ MUTAGENESIS 12-11-15
NINA.W.
(B)
HS9 ⊕ MUTAGENESIS 12-11-15
NINA.W.
(C)
HS7 ⊖ MUTAGENESIS 12-11-15
NINA.W.
(D)
HS9 ⊖ MUTAGENESIS 12-11-15
NINA.W.
(E)
2.5ng pGL4.1 ⊕ CONTROL 12-11-15
NINA.W.
(F)
⊖ CONTROL 12-11-15
NINA.W.

Figure 18: *E. coli* JM109 competent cells transformed with the pGL4.10 vector containing *OCT2* promoter inserts that have undergone site-directed mutagenesis.
A) Mutated HS7 transformed cells; (B) Mutated HS9 transformed cells; (C) Mutated HS7 negative control; (D) Mutated HS9 negative control; (E) 2.5 ng of undigested pGL4.10 vector; (F) Negative control which contains no pGL4.10 vector and no DNA insert. 200 µl of cell suspension was spread onto each LB agar plate which contains 100 µg/ml of ampicillin antibiotic.

A few colonies from each transformation went through a colony PCR and those that contained the *OCT2* promoter region had their vectors purified and then sequenced. The sequencing revealed that the site-directed mutagenesis was successful for HS9 but not for HS7. Therefore, the site-directed mutagenesis PCR annealing temperature for HS7 was optimised using temperatures ranging from 63−68 °C where the lower and higher temperatures resulted in primer tandem repeats. At 66 °C the *OCT2* promoter sequence with the SNP of interest was successfully identified, however there was always a missing adenine base at the start of where the forward primer would bind. This is shown in Figure 19 where TAAT is changed to TAT. This is possibly due to the primers used for the HS7 site-directed mutagenesis being too similar and therefore the functional characterisation of the rs55920607 SNP identified in HS7 was not possible in this study.

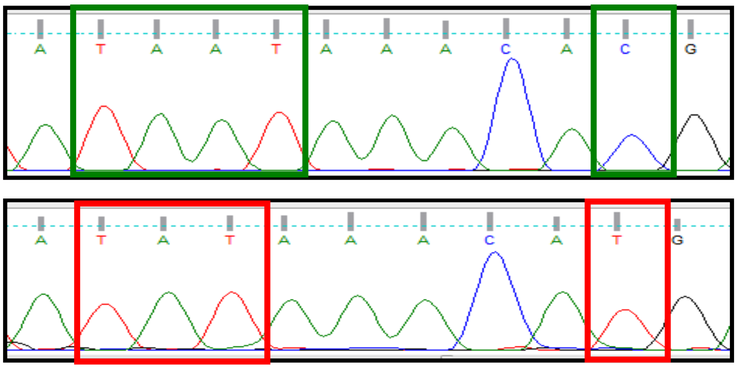


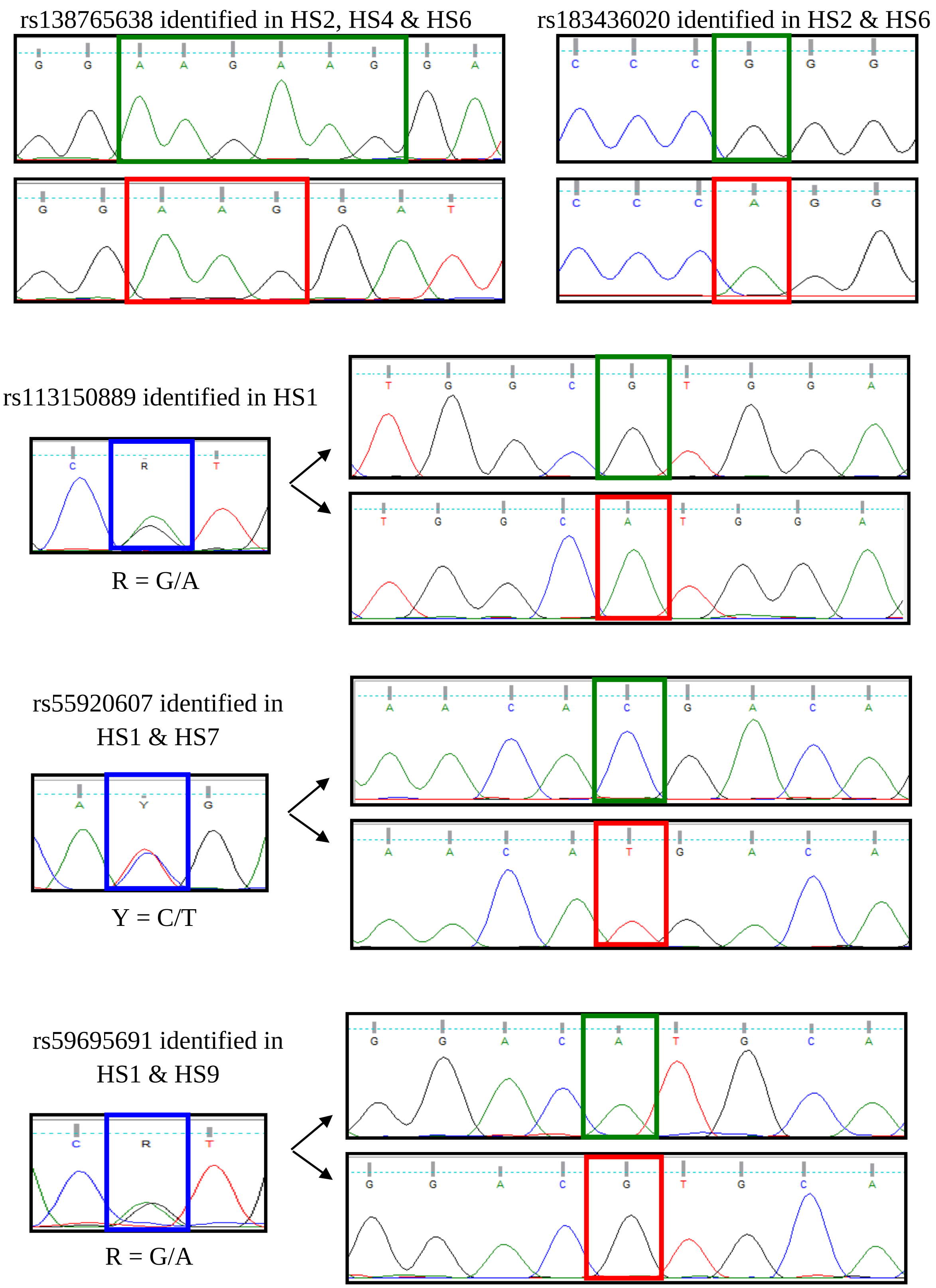
Figure 19: HS7 site-directed mutagenesis chromatograms viewed on FinchTV. The WT is boxed in green and the deletion or SNP of interest is boxed in red.

Through a MSA of HS1−10 (Appendix B) we identified four SNPs as well as one INDEL present in the 10 DNA samples investigated. These variations are listed in Table 6 and the chromatograms of these variations are shown in Figure 20. Their positions on the reverse strand of chromosome 6 are shown in Figure 21. Two SNPs were identified in the 5′ UTR and the other two SNPs as well as one INDEL were identified in the 5′ flanking region. It appeared in this study that we had identified a TTCA deletion (rs66512417) in all 10 DNA samples, however it was eventually determined that this deletion was actually WT and that the insertion of these bases is mutant, therefore none of the 10 samples had this INDEL present.

Table 6: The SNPs and INDEL identified within the 10 individuals from the South African black population.

dbSNP ID	Position from TSS*	Nucleotide change#	HS 1	HS 2	HS 3	HS 4	HS 5	HS 6	HS 7	HS 8	HS 9	HS 10
rs138765638	-965 to -963	AAG (del)		✓		✓		✓				
rs183436020	-924	G>A		✓				✓				
rs113150889	-289	G>A	✓									
rs55920607	-246	C>T	✓						✓			
rs59695691	-195	A>G	✓								✓	

*TSS: Translation start site, # on reverse strand



rs138765638 identified in HS2, HS4 & HS6

rs183436020 identified in HS2 & HS6

rs113150889 identified in HS1

R = G/A

rs55920607 identified in HS1 & HS7

Y = C/T

rs59695691 identified in HS1 & HS9

R = G/A

Figure 20: Chromatograms viewed on FinchTV for the SNPs and INDEL identified in the 10 DNA samples. The four SNPs and the INDEL are boxed in different colours: blue indicates the heterozygous SNPs identified in the preliminary study, green indicates WT, and red indicates mutant.

```
-1459   cttgcatttagatgcaacaacatcatcctaaagagatgcccctgatgtgtgagagcagaa
-1399   aggggtgatccctttcttccttatcctaaggctcacggccaacacccctataacaaaaga
-1339   caggttaacaagagaaaagcatgacaaatttatttgatcacgttttacatgacacaggag
-1279   ccttcattcagaatgaagacccacagatacagggaaaactggatggcagtatagaactgt
-1219   agttggacaaaaagggcagcagcccatgttcgcaggctgaggggaaacccagcaaggcc
-1159   tgtctgttcagatccgtcttggcccctctgtgcagcactccttcctccaggcaccgggga
-1099   caagactcctctggaatgcgggtctggatttctttacggcccactgttacacagaaaggc
-1039   agcggggaagttacagtggtatttctaggctttctggctggctttggggagaaaagagtc
 -979   tggtttccacgagctgctttgaggaagaaggattctcagttctatggcttgccccggggg
 -919   agaatgatgggtgagagaagagacaggagggcaggagaaggtcagagagagagactttgc
 -859   ttctgaggcctccaccttggggcatggatttctgagccccaacagaccttgacagaaaaa
 -799   tctaggacacaaagatagtggcttggacacacctgcctgcatttacacttgacctgtctg
 -739   cgacgtaaacactttcctctttccctccagatgggttaaggggaaggacacttcagggtt
 -679   gaaacgcaggaataccagattggagcaaacacttttaaaagcagagttataaaatctgg
 -619   acaacatcaaaacaagcagccccagcatgcatcccgacggctcttgttgttggttggaga
 -559   atgagcccagcagtcaggcttgcaacccacttcgaatctggaccagggttctgacacgga
 -499   tcctggttcacatcacgctgggccttgtggccaaacacgtgtgttttctccatagggcct
 -439   tgaagaaaagctggcggtgcgcatgagataggagtatattaagttcctggctgctcgggg
 -379   cactacgggaagattactgggctgtgatatgggccagcactcagattccctgcggtggga
 -319   cacagagggcgggttgtttgtgctgctggcgtggagcaccgacaagcctgtggagaacca
 -259   GTTATAATAAACACGACAGGCATCCTGGGAGTGAGCTCAGGGCATTTGGGAAGTGCAGAA
 -199   GGACATGCACCCCCGCTGGAGGGGTGCACCTTTGAAGTCAGCTGGACCAAGGAAAGGCCC
 -139   TGCCCTGAAGGCTGGTCACTTGCAGAGGTAAACTCCCCTCTTTGACTTCTGGCCAGGGTT
  -79   TGTGCTGAGCTGGCTGCAGCCGCTCTCAGCCTCGCTCCGGGCACGTCGGGCAGCCTCGGG
  -19   CCCTCCTGCCTGCAGGATCATGCCCACCACCGTGGACGATGTCCTGGAGCATGGAGGGGA
  +42   GTTTCACTTTTTCCAGAAGCAAATGTTTTTCCTCTTGGCTCTGCTCTCGGCTACCTTCGC
 +102   GCCCATCTACGTGGGCATCGTCTTCCTGGGCTTCACCCCTGACCACCGCTGCCGGAGCCC
 +162   CGGAGTGGCCGAGCTGAGTCTGCGCTGCGGCTGGAGTCCTGCAGAGGAACTGAACTACAC
 +222   GGTGCCGGGCCCAGGACCTGCGGGCGAAGCCTCCCCAAGACAGTGTAGGCGCTACGAGGT
```

KEY:

Primers	E-box	Translation start site	5′ Flanking sequence
CCAAT box	SNPs	Translated sequence	5′ UTR

Figure 21: The SNPs and INDEL identified in the upstream region of the *OCT2* gene sequence. The primers amplify a 1698 bp product of the *OCT2* promoter region. The bold and underlined text represents the transcription factor binding region of the *OCT2* promoter spanning from −483 to −435. Numbering is relative to the translation start site. The SNPs and the INDEL highlighted in blue are those identified within the 10 South African black individuals used in this study, and are shown on the reverse strand of chromosome 6.

A luciferase assay was performed to determine whether the SNPs and INDEL may affect transcription from the *OCT2* promoter (Figure 22). The sample which contained three SNPs (-195, -246 and -289) appeared to be associated with a decrease in luciferase activity in comparison to the WT. The sample containing both the -924 SNP and the AAG deletion appeared to show little change in luciferase activity in comparison to the WT. The luciferase activity of the sample containing only the AAG deletion was significantly increased whereas the sample with the -195 SNP was significantly decreased.

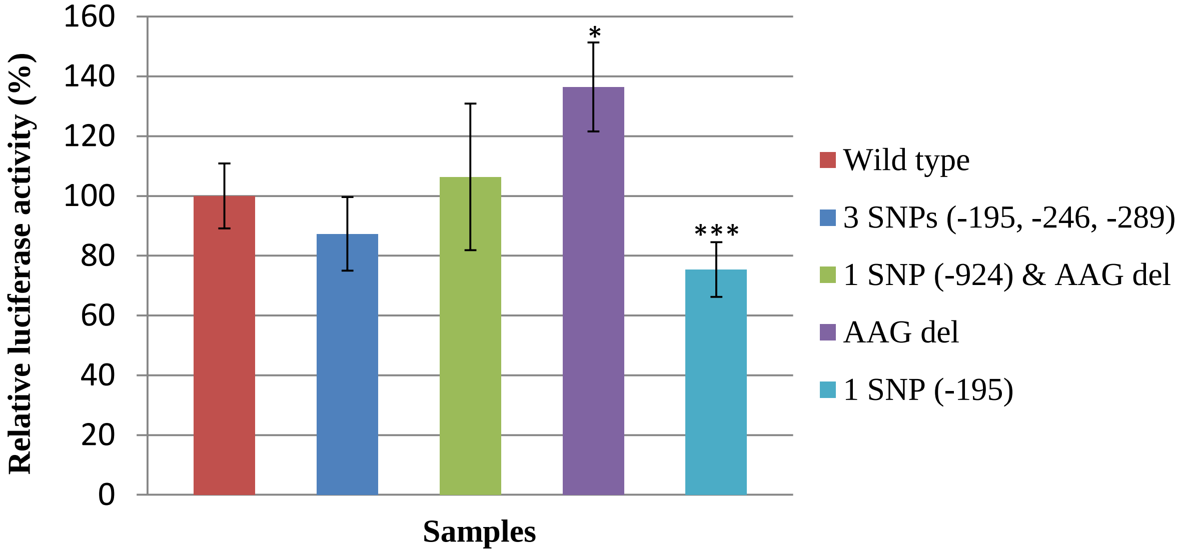


Figure 22: Transcriptional activity of the *OCT2* promoter inserts transfected in MRC-5 cells. Each sample value represents the mean ± standard deviation of quadruplicate experiments. Four different samples within the pGL4.10 luciferase vector were compared to the wild type sample. A paired t-test with $p < 0.05$ was considered to be statistically significant.

4.2 Bioinformatic analysis

4.2.1A Identification of SNPs within the *OCT2* promoter region

A total of thirteen SNPs, including one novel SNP, and two INDELs were identified within the *OCT2* promoter region from five studies conducted on individuals within the South African black population. These results are summarised in Table 7.

We identified eight SNPs (rs59695691, rs55920607, rs60249401, rs80154852, rs146811048, rs148965379, rs316023, rs3127573) and two INDELs (rs138765638 and rs66512417) that commonly occurred between the African and South African black populations. These variations: rs59695691, rs55920607, rs80154852, rs138765638, rs146811048, rs148965379 and rs3127573, showed notably higher allele frequencies in the South African black populations whereas rs316023 and rs66512417 showed notably higher allele frequencies in the non-African populations. Interestingly, we identified one SNP rs183436020 that had not been reported in any Southern African populations to date, yet we identified it to be present within the 10 DNA samples that we analysed in this study. Four SNPs (rs113150889, rs113384645, rs527961348 and 6:160680864) that were identified in the South African black populations have not been observed in any of the 1000 Genomes Project populations.

The RegulomeDB database was used to predict the effect of the *OCT2* promoter SNPs and INDELs on *OCT2* regulation. All of the SNPs and INDELs identified in the *OCT2* promoter region, except for rs3127573, gave a RegulomeDB score of 5. No data was found for the rs3127573 SNP.

Table 7: SNPs and INDELs identified within the *OCT2* promoter region on the forward strand of chromosome 6

dbSNP ID	WT allele	Minor allele	MAF Black SA [a]	MAF SA Bantu [b]	MAF SA Z,S,X [c]	MAF SA Zulu [d]	MAF SA Xhosa [e]	MAF AFR [f]	MAF all excl AFR	MAF all [g]	Nucleotide Position (GRCh38)	Nucleotide Position (GRCh37)	Position from ATG	Location
			n=10	*n*=40	*n*=15	*n*=100	*n*=96							
rs59695691	T	C	0.20	0.025	0.087	0.085	0.263	0.023	< 0.01	0.006	160,258,952	160,679,984	-195	Upstream / 5′ UTR / intronic
rs55920607	G	A	0.20	0.063	0.065	0.09	0.084	0.045	< 0.01	0.014	160,259,003	160,680,035	-246	Upstream / 5′ UTR / intronic
rs113150889	C	T	0.10	0	0.022	0	−	−	−	−	160,259,046	160,680,078	-289	Upstream / intronic
rs60249401	C	T	0	0.025	0	0.015	−	0.020	< 0.01	0.007	160,259,181	160,680,213	-424	Upstream / intronic
rs183436020	C	T	0.20	0	0	0	−	0.002	0	< 0.001	160,259,681	160,680,713	-924	Upstream / 5′ UTR
rs138765638	CTT	−	0.30	−	−	0.21	−	0.138	0.084	0.1	160,259,707-160,259,709	160,680,739-160,680,741	-953 to -955	Upstream / 5′ UTR
rs113384645	C	T	0	0.032	0.065	0.045	−	−	−	−	160,259,735	160,680,767	-978	Upstream / 5′ UTR
Novel	A	G	0	0	0.021	0	−	−	−	−	160,259,832	160,680,864	-1075	Upstream / 5′ UTR
rs80154852	G	A	0	0	0.065	0.095	−	0.057	< 0.01	0.016	160,259,902	160,680,934	-1146	Upstream / 5′ UTR
rs146811048	G	A	0	0.10	0.043	0.03	−	0.023	< 0.01	0.006	160,259,946	160,680,978	-1189	Upstream / 5′ UTR
rs148965379	A	T	0	0.091	0.043	0.03	−	0.023	< 0.01	0.006	160,259,979	160,681,011	-1222	Upstream / 5′ UTR
rs66512417	−	TGAA	0	−	−	0.045	−	0.014	0.090	0.071	160,260,027-160,260,030	160,681,059-160,681,062	-1270 to -1273	Upstream / 5′ UTR
rs527961348	A	G	0	0	0.021	0	−	−	−	−	160,260,223	160,681,255	-1466	Upstream / 5′ UTR
rs316023	T	C	0	0.261	0.109	0.16	−	0.305	0.469	0.420	160,260,282	160,681,314	-1525	Upstream / 5′ UTR
rs3127573	A	G	0	0.44	0.174	0.21	−	0.138	0.086	0.101	160,260,361	160,681,393	-1604	Upstream / 5′ UTR

a – Present study; **b** – Bantu-speaking individuals, (Carstens, N., *et al*. 2016, personal communication); **c** – Xhosa, Zulu and Sotho individuals, (Ramsay, M., *et al*.2016, personal communication); **d** – Zulu individuals, (Gurdasani *et al*., 2014); **e** – Xhosa individuals, (Jacobs *et al*., 2015); **f** – African populations from the 1000 Genomes Project Phase 3; **g** − All individuals from the 1000 Genomes Project Phase 3: African, American, East Asian, European and the South Asian. **(−)** indicates that either there is no information available or the variation could not be identified, ***n*** = number of individuals, **WT** = Wild type, **MAF** = Minor Allele Frequency.

4.2.1B Identification of SNPs within the *OCT2* coding regions

A total of nine SNPs were identified within the *OCT2* coding region from four studies conducted on individuals within the South African black population. These results are summarised in Table 8. Four of the SNPs were non-synonymous missense SNPs (Ser270Ala, Arg400Cys, Lys432Gln and Ile552Asn) and five were synonymous SNPs (Val94Val, Thr130Thr, Ser133Ser, Gly370Gly and Val502Val). Two synonymous SNPs (Val94Val and Gly370Gly) that were identified in the South African black populations have not been observed in any of the 1000 Genomes Project populations. Three SNPs (Ser133Ser, Arg400Cys and Ile552Asn) commonly occurred between only the African and South African black populations whereas two SNPs (Thr130Thr and Val502Val) showed notably higher allele frequencies in the African population when compared to the South African black populations.

The VEP tool was used to predict the possible functional consequences of the missense SNPs identified in the *OCT2* coding region. Three of the four missense SNPs (Ser270Ala, Arg400Cys, Lys432Gln) were predicted to be deleterious or possibly damaging, and one SNP (Ile552Asn) was predicted to be tolerated and benign according to the SIFT and PolyPhen scores.

Table 8: SNPs identified within the *OCT2* coding regions on the forward strand of chromosome 6

dbSNP ID	WT allele	Minor allele	MAF SA[a] Bantu	MAF SA[b] Z,S,X	MAF SA[c] Zulu	MAF SA[d] Xhosa	MAF AFR[e]	MAF all excl AFR	MAF all[f]	Position (GRCh38)	Position (GRCh37)	Functional consequence	Amino Acid change
			***n*=40**	***n*=15**	***n*=100**	***n*=96**							
rs772717144*	C	T	0	0	0.005	0.011	−	−	−	160,258,476	160,679,508	synonymous	Val94Val
rs624249	C	A	0.188	0.152	0.11	0.128	0.266	0.264	0.264	160,258,368	160,679,400	synonymous	Thr130Thr
rs112210325	C	A	0	0.043	0	0.012	0.002	0	0.001	160,258,359	160,679,391	synonymous	Ser133Ser
rs316019	C	A	0.20	0.043	0.09	0.149	0.185	0.118	0.137	160,249,250	160,670,282	missense	Ser270Ala[g]
rs58264151	G	A	0.013	0	0.005	0	−	−	−	160,243,741	160,664,773	synonymous	Gly370Gly
rs8177516	G	A	0.013	0.022	0.045	0.052	0.013	0	< 0.01	160,243,653	160,664,685	missense	Arg400Cys[g]
rs8177517	T	G	0.038	0	0.015	0.011	0.039	< 0.01	0.010	160,242,388	160,663,420	missense	Lys432Gln[g]
rs316003	C	T	0.40	0.20	0.425	0.333	0.576	0.207	0.307	160,224,800	160,645,832	synonymous	Val502Val
rs139045661	A	T	0.063	0.043	0.015	0.016	0.012	0	< 0.01	160,217,445	160,638,477	missense	Ile552Asn

a – Bantu-speaking individuals, (Carstens, N., *et al*. (2016) Personal communication); **b** – Xhosa, Zulu and Sotho individuals, (Ramsay, M., *et al*. (2016) Personal communication); **c** – Zulu individuals, (Gurdasani *et al*., 2014); **d** – Xhosa individuals, (Jacobs *et al*., 2015); **e** – African populations from the 1000 Genomes Project Phase 3; **f** − All individuals from the 1000 Genomes Project Phase 3: African, American, East Asian, European and the South Asian; **g** − These missense SNPs have been functionally characterised by Leabman *et al*., (2002), and have also been predicted to be deleterious or damaging using the VEP tool; **(−)** indicates that there is no information available, ***n*** = number of individuals, **WT** = Wild type, **MAF** = Minor Allele Frequency, *****The incorrect dbSNP ID was used in the Xhosa individuals study and it has been corrected in this study.

4.2.2 Prediction of transcription factor binding sites within the *OCT2* promoter region

The ALGGEN PROMO program was used to determine the potential transcription factor binding sites within the *OCT2* promoter and to determine whether these sites might change should a SNP or INDEL be present. The results are summarised in Table 9.

We identified four SNPs (rs60249401, rs527961348, rs316023, rs3127573) and one INDEL (rs138765638) that had no transcription factor binding site or had lost one or more transcription factor binding sites only when the mutant alleles were present and therefore fewer or no transcription factors were predicted to bind at these locations. Six SNPs (rs55920607, rs183436020, rs113384645, rs80154852, rs146811048, rs148965379) and one INDEL (rs66512417) were predicted to have transcription factor binding sites only when the mutant alleles were present and therefore transcription factors were predicted to bind at these locations. Two SNPs (rs59695691 and rs113150889) had different transcription factor binding sites present when the mutant alleles were present or not and therefore different transcription factors were predicted to bind at these locations.

Two SNPs (rs527961348 and 6:160680864) were predicted to have different transcription factor binding sites when the mutant alleles were present, however the same transcription factors (IRF-1 and RXR-alpha) were still predicted to bind at these locations for their respective SNPs. One INDEL (rs138765638) was predicted to lose the STAT4 transcription factor when the deletion occurred, however the rest of the transcription factors (c-Ets-2, c-Ets-1 and Elk-1) remained the same because the transcription factor binding sites remained the same.

Table 9: The putative transcription factor binding sites predicted within the *OCT2* promoter region using the ALGGEN PROMO program

dbSNP ID	Position from ATG	Flanking sequence on reverse strand (5′−3′)*	TF at Wild type	Wild type TF binding site #	TF at Mutant	Mutant TF binding site #
rs59695691	-195	gaagtgcagaaggacA/**G**tgcacccccgctgga	TFII-I	GGAC**A**T	HIF-1	AC**G**TGCACC
rs55920607	-246	cagttataataaacaC/**T**gacaggcatcctggg	–	–	XBP-1	A**T**GACA
rs113150889	-289	gtttgtgctgctggcG/**A**tggagcaccgacaag	ENKTF-1 AhR:Arnt	TGGC**G**TGG GCTGGC**G**TGG	YY1	**A**TGG
rs60249401	-424	tgaagaaaagctggc**G/A**gtgcgcatgagatag	ENKTF-1	TGGC**G**GTG	–	–
rs183436020	-924	tctatggcttgcccc**G/A**ggggagaatgatggg	–	–	PPAR-alpha: RXR-alpha EBF	TTGCCCC**A**GGG GCCCC**A**GGGGA
rs138765638	-953 to -955	cgagctgctttgaggAAG/-aaggattctcagttc	c-Ets-2 c-Ets-1 Elk-1 STAT4	TTTGAGG**AA** GAGG**AAG** TTGAGG**AAG** GG**AAG**A	c-Ets-2 c-Ets-1 Elk-1	TTTGAGG**AA** GAGG**AAG** TTGAGG**AAG**
rs113384645	-978	gggagaaaagagtctG/**A**gtttccacgagctgc	–	–	NF-AT1	CT**A**GTTTCCA
Novel	-1075	tctggaatgcgggtc**T/C**ggatttctttacggc	RXR-alpha	GGGTC**T**G	RXR-alpha	GGGTC**C**G
rs80154852	-1146	cctgtctgttcagat**C/T**cgtcttggcccctct	–	–	GR-beta	AGAT**T**
rs146811048	-1189	gcagcagcccatgttC/**T**gcaggctgaggggaa	–	–	C/EBPbeta	T**T**GC
rs148965379	-1222	atggcagtatagaacT/**A**gtagttggacaaaaa	–	–	PR B PR A	AAC**A**GTA AAC**A**GTA
rs66512417	-1270 to -1273	acacaggagccttca**-/TTCA**gaatgaagacccaca	–	–	GR-beta	TCA**TT**
rs527961348	-1466	gaaagacaaatttccT/**C**gtttgtcttgcattt	IRF-1 c-Ets-1 GR-alpha	TTTCC**T**GTT TTTCC**T**G CC**T**GT	IRF-1	TTTCC**C**GTT
rs316023	-1525	caggccaccttctct**A/G**cttggcaggatcacg	**-**	**-**	–	–
rs3127573	-1604	ctgctgactatccaa**T/C**agaaaaaagaaggag	GR-beta TFIID NF-Y	AA**T**AG **T**AGAAAA TATCCAA**T**	–	–

*Wild type/Mutant variation, #The bold and underlined regions indicate either the wild type or mutant sequence in the binding site, **(-)** indicates that no TF was predicted to bind to the OCT2 promoter when the specific allele was present, **TF** = Transcription factor.

4.2.3 Prediction of the formation of G-quadruplexes within the *OCT2* promoter region

The formation of G-quadruplexes within the *OCT2* promoter region was predicted using the QGRS Mapper software program. The *OCT2* promoter WT sequence revealed that a predicted total number of 17 G-quadruplexes and 117 overlapping G-quadruplexes would form. The sequences of the G-quadruplexes predicted to form are listed in Table 10 and the locations within the *OCT2* promoter region are shown in Figure 23.

The G-scoring system determines the likelihood of a sequence of nucleotides to form a stable G-quadruplex. The higher G-scoring sequences make better candidates for G-quadruplexes. The shorter loops tend to be more common than the longer loops, and usually G-quadruplexes have loops that are roughly equal in size and become more stable with the a greater number of guanine tetrads (Kikin, D'Antonio & Bagga, 2006).

Of the thirteen SNPs and two INDELs identified within the *OCT2* promoter region, only three SNPs were located within a G-quadruplex sequence. The novel SNP occurred at position 625, one SNP (rs183436020) occurred at position 776, and the other SNP (rs59595691) occurred at position 1505 as shown in Figure 23. Of these three SNPs only one of them (rs183436020) decreased the number of G-quadruplexes including overlaps to 112. None of the other SNPs or INDELs revealed a change to the G-quadruplex predicted formation.

Table 10: The G-quadruplex sequences found within the WT *OCT2* promoter region.

Position*	Length	G-Quadruplex sequence#	G-Score
108	29	GGAGAAGGAAAATCGGGTTAACTTTCTGG	15
515	24	GGCTGAGGGGAAAACCCAGCAAGG	9
613	27	GGAATGCGGGTCTGGATTTCTTTACGG	15
678	30	GGTATTTCTAGGCTTTCTGGCTGGCTTTGG	19
766	25	GGCTTGCCCCGGGGGAGAATGATGG	14
806	16	GGAGGGCAGGAGAAGG	19
847	21	GGCCTCCACCTTGGGGCATGG	11
993	25	GGGTTAAGGGGAAGGACACTTCAGG	17
1180	27	GGACCAGGGTTCTGACACGGATCCTGG	17
1309	20	GGCTGCTCGGGGCACTACGG	15
1374	19	GGTGGGACACAGAGGGCGG	16
1501	24	GGACATGCACCCCCGCTGGAGGGG	6
1544	28	GGACCAAGGAAAGGCCCTGCCCTGAAGG	12
1726	14	GGAGCATGGAGGGG	16
1921	23	GGTGCCGGGCCCAGGACCTGCGG	19
1968	20	GGCGCTACGAGGTGGACTGG	14
2057	22	GGCCCCTGCCGGGACGGCTGGG	16

*****Position in Figure 23, #Sequence on the reverse strand.

```
0001 TTTTAAGAAA ATAAGAAGCC TGCAGCTGCG GCAAGGAGAA AGACCCCGTG GGTCGCTCCT TTTAGACTAC GCTATTACAC CTGCTGACTA TCCAATAGAA
0101 AAAAGAAGGA GAAGGAAAAT CGGGTTAACT TTCTGGTATT TAGTTTGCAG ATATGTCAGC AGGCCACCTT CTCTACTTGG CAGGATCACG ATGTATTAAG
0201 AATCCTTTTG TCCACCAGGA AAGACAAATT TCCTGTTTGT CTTGCATTTA GATGCAACAA CATCATCCTA AAGAGATGCC CCTGATGTGT GAGAGCAGAA
0301 AGGGGTGATC CCTTTCTTCC TTATCCTAAG GCTCACGGCC AACACCCCTA TAACAAAAGA CAGGTTAACA AGAGAAAAGC ATGACAAATT TATTTGATCA
0401 CGTTTTACAT GACACAGGAG CCTTCATTCA GAATGAAGAC CCACAGATAC AGGGAAAACT GGATGGCAGT ATAGAACTGT AGTTGGACAA AAAGGGCAGC
0501 AGCCCATGTT CGCAGGCTGA GGGGAAAACC CAGCAAGGCC TGTCTGTTCA GATCCGTCTT GGCCCCTCTG TGCAGCACTC CTTCCTCCAG GCACCGGGGA
0601 CAAGACTCCT CTGGAATGCG GGTCTGGATT TCTTTACGGC CCACTGTTAC ACAGAAAGGC AGCGGGGAAG TTACAGTGGT ATTTCTAGGC TTTCTGGCTG
0701 GCTTTGGGGA GAAAAGAGTC TGGTTTCCAC GAGCTGCTTT GAGGAAGAAG GATTCTCAGT TCTATGGCTT GCCCCGGGGG AGAATGATGG GTGAGAGAAG
0801 AGACAGGAGG GCAGGAGAAG GTCAGAGAGA GAGACTTTGC TTCTGAGGCC TCCACCTTGG GGCATGGATT TCTGAGCCCC AACAGACCTT GACAGAAAAA
0901 TCTAGGACAC AAAGATAGTG GCTTGGACAC ACCTGCCTGC ATTTACACTT GACCTGTCTG CGACGTAAAC ACTTTCCTCT TTCCCTCCAG ATGGGTTAAG
1001 GGGAAGGACA CTTCAGGGTT GAAACGCAGG AATACCAGAT TGGAGCAAAC ACTTTTTAAA AGCAGAGTTA TAAAATCTGG ACAACATCAA AACAAGCAGC
1101 CCCAGCATGC ATCCCGACGG CTCTTGTTGT TGGTTGGAGA ATGAGCCCAG CAGTCAGGCT TGCAACCCAC TTCGAATCTG GACCAGGGTT CTGACACGGA
1201 TCCTGGTTCA CATCACGCTG GGCCTTGTGG CCAAACACGT GTGTTTTCTC CATAGGGCCT TGAAGAAAAG CTGGCGGTGC GCATGAGATA GGAGTATATT
1301 AAGTTCCTGG CTGCTCGGGG CACTACGGGA AGATTACTGG GCTGTGATAT GGGCCAGCAC TCAGATTCCC TGCGGTGGGA CACAGAGGGC GGGTTGTTTG
1401 TGCTGCTGGC GTGGAGCACC GACAAGCCTG TGGAGAACCA GTTATAATAA ACACGACAGG CATCCTGGGA GTGAGCTCAG GGCATTTGGG AAGTGCAGAA
1501 GGACATGCAC CCCCGCTGGA GGGGTGCACC TTTGAAGTCA GCTGGACCAA GGAAAGGCCC TGCCCTGAAG GCTGGTCACT TGCAGAGGTA AACTCCCCTC
1601 TTTGACTTCT GGCCAGGGTT TGTGCTGAGC TGGCTGCAGC CGCTCTCAGC CTCGCTCCGG GCACGTCGGG CAGCCTCGGG CCCTCCTGCC TGCAGGATCA
1701 TGCCCACCAC CGTGGACGAT GTCCTGGAGC ATGGAGGGGA GTTTCACTTT TTCCAGAAGC AAATGTTTTT CCTCTTGGCT CTGCTCTCGG CTACCTTCGC
1801 GCCCATCTAC GTGGGCATCG TCTTCCTGGG CTTCACCCCT GACCACCGCT GCCGGAGCCC CGGAGTGGCC GAGCTGAGTC TGCGCTGCGG CTGGAGTCCT
1901 GCAGAGGAAC TGAACTACAC GGTGCCGGGC CCAGGACCTG CGGGCGAAGC CTCCCCAAGA CAGTGTAGGC GCTACGAGGT GGACTGGAAC CAGAGCACCT
2001 TCGACTGCGT GGACCCCCTG GCCAGCCTGG ACACCAACAG GAGCCGCCTG CCACTGGGCC CCTGCCGGGA CGGCTGGGTG TACGAGACGC CTGGCTCGTC
2101 CATCGTCACC GAG
```

Figure 23: The *OCT2* promoter region showing the location of the predicted G-quadruplex structures. The SNPs and INDELs are highlighted in green and the translation start site is highlighted in purple. The G-quadruplex sequences are highlighted in yellow. A total of 17 G-quadruplexes were predicted in this region using the QGRS Mapper tool. Three SNPs (at positions 625, 776, and 1505) occur within a predicted G-quadruplex structure.

5. DISCUSSION

5.1 SNP identification and functional analysis

The results from the luciferase assay revealed that one SNP and one INDEL identified in the DNA samples investigated showed a significantly different ability to drive the expression of luciferase in comparison to the WT *OCT2* promoter sequence. The INDEL, the AAG deletion (rs138765638), showed a statistically significant increase in luciferase expression whereas the sample that contained the SNP at position -195 relative to the translation start site (rs59695691) showed a statistically significant decrease in luciferase expression. This suggests that these two variants may be associated with a respective increase and decrease in *OCT2* regulatory function. This implies that they may therefore also affect the regulation and expression of the OCT2 protein.

The sequences that contained more than one variant, as shown in Table 6, showed no statistically significant change in luciferase expression. The sequences containing both the -924 SNP (rs183436020) and the AAG deletion (rs138765638) showed no statistically significant difference in luciferase expression, which could possibly be because one variant may have a positive effect and the other variant a negative effect on luciferase expression thereby causing no overall change to occur. It is also possible that when combinations of variants occur together in a sample, such as (rs59695691, rs55920607 and rs113150889), that no change in luciferase expression is observed or that the individual variants in general do not effect luciferase expression. This highlights the importance of functionally characterising each individual variant and in the combinations that they appear in human DNA.

For the luciferase assay experiments, it is possible and likely that different cell types have a different array of transcription factors available, and these could affect the transcriptional regulation of *OCT2*. Since *OCT2* gene control is not well understood, different factors could be controlling expression in different tissue types so our results indicate a general trend in change in expression but we cannot say with certainty if this will be seen in every cell type that the *OCT2* gene is expressed in.

Since the two variants identified may be associated with a respective change in the regulation or expression of the OCT2 protein, they can subsequently alter drug uptake or transport in the body which can influence a person's response to treatment for various diseases.

When a person's genetic make-up enhances expression of a particular gene such as *OCT2*, a higher dose of medication may be required for effective treatment because the person's body may process the medication quicker compared to somebody who has a different genetic make-up. Conversely, a lower dose may be required if it is found that it is transported into cells more efficiently.

Alternatively, when a person's genetic make-up reduces expression of a particular gene such as *OCT2*, a lower dose of medication may be required for effective treatment. This may be due to the person's body processing the medication at a slower rate, which can cause the medication to remain in the bloodstream for longer periods and therefore may cause unpleasant side effects. Conversely, a higher dose may be required to increase the amount of drug transported into cells.

Since one of the SNPs (rs59695691) was identified to be heterozygous, the effect that it may have on *OCT2* gene expression and consequently protein expression, drug uptake and drug transport may not be as great as if it were a homozygous mutant SNP because one WT allele is still present for the *OCT2* gene. If the SNP was homozygous mutant then the effect may be greater for *OCT2* gene expression because no WT alleles are present. Nevertheless, it may still have an effect on gene expression. The preliminary study was able to genotype the three SNPs identified because the PCR products were sequenced. The genotypes of the additional variations (rs183436020 and rs138765638) identified in the longer *OCT2* promoter region were not determined in this study because the *OCT2* promoter insert within the vector was sequenced. This however was not important because one cannot determine the effect of a heterozygous SNP in a luciferase assay since only one of the two strands of DNA is cloned into the vector. The effect of heterozygous SNPs versus homozygous mutant SNPs on *OCT2* expression can be determined using mRNA studies.

5.2 Bioinformatic analysis

5.2.1 Identification of SNPs within the *OCT2* promoter and coding regions

Although most of the SNPs and INDELs commonly occurred between the African and South African populations (except for the four promoter and two synonymous SNPs that were not observed within the 1000 Genomes Project populations), the allele frequencies were found to vary notably between these two populations, which highlights the importance of Southern Africa centric evaluation of genetic variants that may show potential phenotypic significance.

We performed Sanger sequencing in 10 DNA samples, and whilst the possibility of finding a rare allele in this small sample number is extremely low, however, our findings from the other studies with sample numbers of 15, 40 and 100, increases the chance of finding a rare allele. We could therefore compare results from all the studies to find alleles that are most common to the South African black population and those alleles that may be less common or rare to this population. This can clearly be seen for the rs772717144 SNP that was found in the Xhosa and Zulu population studies, which had sample numbers of 96 and100. The small sample number of 10 should be increased to at least 100 in future studies.

A RegulomeDB score of 5 indicates that there is minimal binding evidence and that the variations with this score can possibly affect transcription factor binding or the DNase peak within a regulatory region. The only SNP that did not have a score, rs3127573, was due to there not being a common SNP in the uploaded genomic region. Since all of the other SNPs and INDELs were predicted to have a RegulomeDB score of 5, it suggests that rs3127573 would also most likely have a score of 5 and may possibly also affect *OCT2* regulation.

The VEP tool was used to predict the possible functional consequences of the missense SNPs identified in the *OCT2* coding region. The three missense SNPs (Ser270Ala, Arg400Cys, Lys432Gln) that were predicted to be deleterious or possibly damaging have already been characterised by Leabman *et al* in the African-American population to alter OCT2 transporter function (Leabman *et al*., 2002). The other missense SNP (Ile552Asn) was predicted to be tolerated and benign and this suggests that there is no possible functional consequence on OCT2 transport, however, since this SNP was only observed within the African and South African black population studies, the functional characterisation of this SNP must still be determined. Since these four missense SNPs commonly occur within the South African black population, their significance and characterisation should be investigated further within this population to determine their effect on OCT2 transport and drug uptake.

Another mechanism that could alter gene regulation, as previously mentioned, is DNA methylation, which has been found to modulate the expression of *OCT2* in the kidneys. In the *OCT2* promoter region, the methylation status of CpG sites is low but differences in the methylation status between different individuals has been observed (Aoki *et al.,* 2008). The presence of CpG sites occurring within the *OCT2* promoter region was not examined in this study. It is possible that these CpG sites could be modified by the presence of SNPs which could affect the regulation of the expression of the *OCT2* gene and therefore should be investigated in future studies.

5.2.2 Prediction of transcription factor binding sites within the *OCT2* promoter region

Whilst transcription factor binding sites may only bind one or two proteins, the transcription factors are often complexed to other proteins in order to be active, therefore if a SNP occurs within a transcription factor binding site then it may be possible that the particular transcription factor will not bind. This may cause any other proteins that are complexed with the transcription factor to not be activated and this can subsequently cause downstream effects on transcriptional regulation. On the contrary, a different transcription factor may bind when a SNP is present and therefore the different proteins that might be complexed with it can cause different downstream pathways to be regulated. It is also possible that the same transcription factor will bind to its site regardless of presence of polymorphisms and therefore no regulatory change would occur.

The ALGGEN PROMO program was used to determine the potential transcription factor binding sites within the *OCT2* promoter and to determine whether these sites might change should a SNP or INDEL be present. Most of the SNPs identified within the *OCT2* promoter region were predicted to change the transcription factor binding sites and thus resulted in a loss, gain or change in the transcription factors that were predicted to bind at *OCT2* promoter. These variations may therefore have a possible effect on *OCT2* regulation and OCT2 protein expression and therefore should be investigated further. A few variations had the same transcription factors predicted to bind to the *OCT2* promoter when the mutant alleles were present or not and this suggests that they may have no possible effect on *OCT2* regulation.

5.2.3 Prediction of the formation of G-quadruplexes within the *OCT2* promoter region

The formation of G-quadruplexes within the *OCT2* promoter region was predicted using the QGRS Mapper software program. Of the three SNPs that were shown to occur within a G-quadruplex predicted structure, only the one SNP (rs183436020) changed the overlapping G-quadruplex number from 117 to 112, which suggests that should the G-quadruplexes that are affected from forming be involved in controlling the activity of transcription, then the level of transcription may be altered. This SNP may therefore also control translation initiation and regulation because G-quadruplexes can influence the recruitment of ribosomes to the translation start site. This SNP should therefore be further investigated.

5.3 Study limitations and future studies

The sample size of this study was very small when compared to the other studies analysed and that of the 1000 Genomes Project so therefore it would be better to use a larger sample size in future studies. The promoter of the *OCT2* gene was the only region sequenced in the 10 samples analysed in this study and it would be beneficial in future studies to analyse the coding regions or the full genome sequences as well to give a complete picture and more precise account of the genetic variation that occurs within the South African black populations. It would also be of great interest in future studies to functionally characterise the other promoter SNPs that were identified as unique in the South African black population.

6. CONCLUSION

In this study we identified thirteen SNPs and two INDELs within the *OCT2* promoter region, and nine SNPs within the *OCT2* coding region through analysing various South African population studies. These variations could affect both gene expression and protein function. Many of the SNPs and INDELs, which commonly occurred between the South African black populations and African populations, were found to have allele frequencies that varied notably between the two populations. These observations highlight the importance of Southern Africa centric evaluation of genetic variants that may show potential phenotypic significance.

We also identified four SNPs and one INDEL within the *OCT2* promoter region from sequencing just 10 DNA samples where one SNP (rs59695691) and one INDEL (rs138765638) showed a statistically significant decrease and increase in luciferase expression. These variants should therefore be further investigated. Should these variants be found to be clinically functionally significant then this may enable the use of alternative or the development of new drug treatments since OCT2 transports cationic drugs used in the treatment of diseases such as cancer, type 2 diabetes & Parkinson's disease. These treatments can then target the specific SNPs or INDELs and can lead to the design of more efficient clinical trials for black South Africans suffering from various diseases in the future. The findings of this study therefore contribute to filling the gap pertaining to OCT variation in South African populations, and may impact greatly on healthcare provided across Africa.

7. REFERENCES

Adzhubei, I., Jordan, D.M. & Sunyaev, S.R. (2013) Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet*. **7**: 7–20.

Aoki, M., Terada, T., Kajiwara, M., Ogasawara, K., Ikai, I., Ogawa, O., Katsura, T. & Inui, K. (2008) Kidney-specific expression of human organic cation transporter 2 (OCT2/ SLC22A2) is regulated by DNA methylation. *Am. J. Physiol. Renal Physiol.* **295**: F165–F170.

Araujo, P.R., Yoon, K., Ko, D., Smith, A.D., Qiao, M., Suresh, U., Burns, S.C. & Penalva, L.O.F. (2012) Before it gets started: regulating translation at the 5′ UTR. *Comp. Funct. Genomics.* **2012**: 475731–465738.

Asaka, J., Terada, T., Ogasawara, K., Katsura, T. & Inui, K. (2007) Characterization of the basal promoter element of human organic cation transporter 2 gene. *J. Pharmacol. Exp. Ther.* **321**: 684–689.

Bacq, A., Balasse, L., Biala, G., Guiard, B., Gardier, A.M., Schinkel, A., Louis, F., Vialou, V., Martres, M.-P., Chevarin, C., Hamon, M., Giros, B. & Gautron, S. (2011) Organic cation transporter 2 controls brain norepinephrine and serotonin clearance and antidepressant response. *Mol. Psychiatry*. **17**: 926–938.

Bailey, C.J. & Turner, R.C. (1996) Metformin. *N. Engl. J. Med*. **334**: 574–579.

Blench, R. (2006) Archaeology, Language and the African Past. New York: Rowman & Littlefield Publishers Inc.

Boyle, A.P., Hong, E.L., Hariharan, M., Cheng, Y., Schaub, M.A., Kasowski, M., Karczewski, K.J., Park, J. *et al*. (2012) Annotation of functional variation in personal genomes using RegulomeDB. *Genome Research.* **22**(9): 1790–1797.

Brast, S., Grabner, A., Sucic, S., Sitte, H.H., Hermann, E., Pavenstädt, H., Schlatter, E. & Ciarimboli, G. (2011) The cysteines of the extracellular loop are crucial for trafficking of human organic cation transporter 2 to the plasma membrane and are involved in oligomerization. *FASEB J*. **26**: 976–986.

Burckhardt, G. & Wolff, N.A. (2000) Structure of renal organic anion and cation transporters. *Am. J. Physiol. Renal Physiol*. **278**: F853–66.

Busch, A.E., Karbach, U., Miska, D., Gorboulev, V., Akhoundova, A., Volk, C., Arndt, P., Ulzheimer, J.C., Sonders, M.S., Baumann, C., Waldegger, S., Lang, F. & Koepsell H. (1998) Human neurons express the polyspecific cation transporter hOCT2, which translocates monoamine neutrotransmitters, amantadine, and memantine. *Mol. Pharmacol.* **54**: 342–352.

Campbell, M.C. & Tishkoff, S.A. (2010) The evolution of human genetic and phenotypic variation in Africa. *Curr Biol.* **20**: R166–173.

Chang, X. & Wang, K. (2012) wANNOVAR: annotating genetic variants for personal genomes via the web. *J Med Genet*. **49**(7): 433–436.

Ciarimboli, G. & Schlatter, E. (2005) Regulation of organic cation transport. *Pflugers Arch. Eur. J. Physiol.* **449**: 423–441.

Ciarimboli, G., Ludwig, T., Lang, D., Pavenstadt, H., Koepsell, H., Piechota, H.J., Haier, J., Jaehde, U., Zisowsky, J. & Schlatter, E. (2005) Cisplatin nephrotoxicity is critically mediated via the human organic cation transporter 2. *Am. J. Pathol.* **167**: 1477–1484.

Corre, S. & Galibert, M.D. (2005) Upstream stimulating factors: highly versatile stress-responsive transcription factors. *Pigment Cell Res.* **18**: 337–348.

Dai, J., Dexheimer, T.S., Chen, D., Carver, M., Ambrus, A., Jones, R.A. & Yang, D. (2006) An intramolecular G-quadruplex structure with mixed parallel/antiparallel G-strands formed in the human BCL-2 promoter region in solution. *J. Am. Chem. Soc.* **128**: 1096–1098.

de Armond, R., Wood, S., Sun, D., Hurley, L.H. & Ebbinghaus, S.W. (2005) Evidence for the presence of a Guanine quadruplex forming region within a polypurine tract of the Hypoxia Inducible Factor 1R promoter. *Biochemistry*. **44**: 16341–16350.

Ehret, C. (1998) An African Classical Age: Eastern and Southern Africa in World History, 1000 B.C. to A.D. 400. USA: The University Press of Virginia.

Evans, W.E., & Relling, M.V. (1999) Pharmacogenomics: translating functional genomics into rational therapeutics. *Science.* **286**: 487–491.

Farré, D., Roset, R., Huerta, M., Adsuara, J.E., Roselló, L., Albà, M.M. & Messeguer, X. (2003) Identification of patterns in biological sequences at the ALGGEN server: PROMO and MALGEN. *Nucleic Acids Res*. **31**(13): 3651–3653.

Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P. *et al*. (2007) International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature*. **449**: 851–861.

Fujita, T., Urban, T.J., Leabman, M.K., Fujita, K. & Giacomini, K.M. (2006) Transport of drugs in the kidney by the human organic cation transporter, OCT2 and its genetic variants. *J Pharm Sci.* **95**(1): 25–36.

Fukushima-Uesaka, H., Maekawa, K., Ozawa, S., Komamura, K., Ueno, K., Shibakawa, M., Kamakura, S., Kitakaze, M., Tomoike, H., Saito, Y. & Sawada, J. (2004) SNP communication fourteen novel single nucleotide polymorphisms in the SLC22A2 gene encoding human organic cation transporter (OCT2). *Drug Metab. Pharmacokin*. **19**: 239–244.

Gellert, M., Lipsett, M.N. & Davies, D.R. (1962) Helix formation by guanylic acid. *Proc. Natl. Acad. Sci. USA* **48**: 2013–2018.

Gingras, A.C., Raught, B. & Sonenberg, N. (1999) eIF4 initiation factors: Effectors of mRNA recruitment to ribosomes and regulators of translation. *Annu. Rev. Biochem.* **68**: 913–963.

Go, R.S. & Adjei, A.A. (1999) Review of the comparative pharmacology and clinical activity of cisplatin and carboplatin. *J. Clin. Oncol.* **17**: 409– 422.

Gorboulev, V., Ulzheimer, J.C., Akhoundova, A., Ulzheimer-Teuber, I., Karbach, U., Quester, S., Baumann, C., Lang, F., Busch, A.E. & Koepsell, H. (1997) Cloning and characterization of two human polyspecific organic cation transporters. *DNA Cell Biol.* **16**: 871– 881.

Gründemann, D. & Schömig, E. (2000) Gene structures of the human non-neuronal monoamine transporters EMT and OCT2. *Hum. Genet.* **106**: 627–635.

Gründemann, D., Babin-Ebell, J., Martel, F., Örding, N., Schmidt, A. & Schömig, E. (1997) Primary structure and functional expression of the apical organic cation transporter from kidney epithelial LLC- PK1 cells. *J Biol. Chem.* **272**: 10408–10413.

Gurdasani, D., Carstensen, T., Tekola-Ayele, F., Pagani, L., Tachmazidou, I., Hatzikotoulas, K., Karthikeyan, S., Iles, L., Pollard, M.O., Choudhury, A., *et al*. (2014) The African Genome Variation Project shapes medical genetics in Africa. *North.* **517**: 327–332.

Hardy, B-J., Séguin, B., Ramesar, R., Singer, P.A. & Daar, A.S. (2008) South Africa: From species cradle to genomic applications. *Nat. Rev. Genet.* **9**: S19–S23.

Hayashi, S., Watanabe, J. & Kawajiri, K. (1991) Genetic polymorphisms in the 5′-flanking region change transcriptional regulation of the human cytochrome P450IIE1 gene. *J. Biochem.* **110**: 559–565.

Houtsma, D., Guchelaar, H. & Gelderblom, H. (2010) Pharmacogenetics in Oncology: A Promising Field. *Curr. Pharm. Des.* **16**(2): 155–163.

Howell, R.M., Woodford, K.J., Weitzmann, M.N. & Usdin, K. (1995) The chicken b-Globin gene promoter forms a novel ‘cinched’ tetrahelical structure. *J. Biol. Chem.* **271**: 5208–5214.

Huppert, J.L. & Balasubramanian, S. (2007) G-quadruplexes in promoters throughout the human genome. *Nucleic Acids Res.* **35**(2): 406–413.

Huppert, J.L., Bugaut, A., Kumari, S. & Balasubramanian, S. (2008) G-quadruplexes: the beginning and end of UTRs. *Nucleic Acids Res*. **36**: 6260–6268.

Ieiri, I., Takane, H., Hirota, T., Otsubo, K. & Higuchi, S. (2006) Genetic polymorphisms of drug transporters: pharmacokinetic and pharmacodynamic consequences in pharmacotherapy. *Expert Opin. Drug Metab. Toxicol*. **2**: 651–674.

Jacobs, C., Pearce, B., Du Plessis, M., Hoosain, N. & Benjeddou, M. (2015) Single nucleotide polymorphisms of the SLC22A2 gene within the Xhosa population of South Africa. *Drug Metab. Pharmacokinet.* **30**: 457–460.

Johnson, N.P., Butour, J.L., Villani, G., Wimmer, F.L., Defais, M., Pierson, V. & Brabec, V. (1989) Metal antitumor compounds: the mechanism of action of platinum complexes. *Prog. Clin. Biochem. Med.* **10**: 1–24.

Jung, N., Lehmann, C., Rubbert, A., Knispel, M., Hartmann, P., van Lunzen, J., Stellbrink, H.J., Faetkenheuer, G. & Taubert, D. (2008) Relevance of the organic cation transporters 1 and 2 for antiretroviral drug therapy in human immunodeficiency virus infection. *Drug Metab Dispos*. **36**(8): 1616–1623.

Keller, T., Egenberger, B., Gorboulev, V., Bernhard, F., Uzelac, Z., Gorbunov, D., Wirth, C., Koppatz, S., Dötsch, V., Hunte, C., Sitte, H.H. & Koepsell, H. (2011) The large extracellular loop of organic cation transporter 1 influences substrate affinity and is pivotal for oligomerization. *J. Biol. Chem.* **286**: 37874–37886.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M. & Haussler, D. (2002) UCSC Genome Browser: The human genome browser at UCSC. *Genome Res.* **12**(6):996–1006.

Kikin, O., D'Antonio, L. & Bagga, P.S. (2006) QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences. *Nucleic Acids Res*. **34**: W676–W682.

Kirpichnikov, D., McFarlane, S.I. & Sowers, J.R. (2002) Metformin: an update. *Ann. Intern. Med*. **137**: 25–33.

Koehler, M.R., Wissinger, B., Gorboulev, V., Koepsell, H. & Schmid, M. (1997) The two human organic cation transporter genes SLC22A1 and SLC22A2 are located on chromosome 6q26. *Cytogenet. Cell Genet.* **79**: 198–200.

Koepsell, H. (2011) Substrate recognition and translocation by polyspecific organic cation transporters. *Biol. Chem*. **392**: 95–101.

Koepsell, H., Lips, K. & Volk, C. (2007) Polyspecific organic cation transporters: structure, function, physiological roles, and biopharmaceutical implications. *Pharm. Res.* **24**: 1227–1251.

Koepsell, H., Schmitt, B.M. & Gorboulev, V. (2003) Organic cation transporters. *Rev. Physiol. Biochem. Pharmacol.* **150**: 36–90.

Kozak, M. (1991) Structural features in eukaryotic mRNAs that modulate the initiation of translation. *J. Biol. Chem*. **266**: 19867–19870.

Leabman, M. K., Huang, C. C., Kawamoto, M., Johns, S. J., Stryke, D., Ferrin, T. E., DeYoung, J., Taylor, T., Clark, A.G., Herskowitz, I. & Giacomini, K.M. (2002) Polymorphisms in a human kidney xenobiotic transporter, OCT2, exhibit altered function. *Pharmacogenet. Genom.* **12**(5): 395–405.

Levitt, N.S. (2008) Diabetes in Africa: epidemiology, management and healthcare challenges. *Heart*. **94**: 1376–1382.

Mantovani, R. (1999) The molecular biology of the CCAAT-binding factor NF-Y. *Gene (Amst)* **239**: 15–27.

Mbanya, J.C., Motala, A.A., Sobngwi, E., Assah, F.K. & Enoru, S.T. (2010) Diabetes in sub-Saharan Africa. *Lancet*. **375**: 2254–2266.

Messeguer, X., Escudero, R., Farré, D., Nuñez, O., Martínez, J. & Albà, M.M. (2002) PROMO: detection of known transcription regulatory elements using species-tailored searches. *Bioinformatics*, **18**(2), 333–334.

Meyer, U.A. (2000) Pharmacogenetics and adverse drug reactions. *Lancet.* **356**, 1667–1671.

Mooslehner, K.A. & Allen, N.D. (1999) Cloning of the mouse organic cation transporter 2 gene, Slc22a2, from an enhancer-trap transgene integration locus. *Mamm. Genome.* **10**: 218–224.

Motohashi, H., Sakurai, Y., Saito, H., Masuda, S., Urakami, Y., Goto, M., Fukatsu, A., Ogawa, O. & Inui, K. (2002) Gene expression levels and immunolocalization of organic ion transporters in the human kidney. *J. Am. Soc. Nephrol.* **13**: 866–874.

Ng, P.C. & Henikoff, S. (2003) SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res*. **31**(13): 3812–3814.

Ogasawara, K., Terada, T., Motohashi, H., Asaka, J., Aoki, M., Katsura, T., Kamba, T., Ogawa, O. & Inui, K. (2008) Analysis of regulatory polymorphisms in organic ion transporter genes (SLC22A) in the kidney. *J. Hum. Genet.* **53**: 607–614.

Okuda, M., Saito, H., Urakami, Y., Takano, M. & Inui, K.I. (1996) cDNA cloning and functional expression of a novel rat kidney organic cation transporter, OCT2. *Biochem. Biophys. Res. Commun.* **224**: 500–507.

Peer, N., Steyn, K., Lombard, C., Lambert, E. V., Vythilingum, B., & Levitt, N.S. (2012) Rising Diabetes Prevalence among Urban-Dwelling Black South Africans. *PLoS ONE.* **7**(9): 1–9.

Pepper, M.S. (2011) Launch of the Southern African Human Genome Programme. *S. Afr. Med. J.* **101**(5): 287–288.

Perez, R.P. (1998) Cellular and molecular determinants of cisplatin resistance. *Eur. J. Cancer.* **34**: 1535–1542.

Ramji, D.P. & Foka, P. (2002) CCAAT/enhancer-binding proteins: structure, function and regulation. *Biochem. J.* **365**: 561–575.

Ramsay, M. (2012) Africa: continent of genome contrasts with implications for biomedical research and health. *FEBS Lett*. **586**: 2813–2819.

Rankin, S., Reszka, A.P., Huppert, J., Zloh, M., Parkinson, G.N., Todd, A.K., Ladame, S., Balasubramanian, S. & Neidle, S. (2005) Putative DNA quadruplex formation within the human c-kit oncogene. *J. Am. Chem. Soc.* **127**: 10584–10589.

Rosenberg, N.A., Pritchard, J.K., Weber, J.L., Cann, H.M., Kidd, K.K., Zhivotovsky, L.A. & Feldman, M.W. (2002) Genetic structure of human populations. *Science*. **298**: 2381–2385.

Sen, D. & Gilbert, W. (1988) Formation of parallel four-stranded complexes by guanine-rich motifs in DNA and its implications for meiosis. *Nature*, **334**: 364–366.

Sharma, B.S., Mount, J. & Karrow, N.A. (2008) Functional characterization of a single nucleotide polymorphism in the 5' UTR region of the bovine toll-like receptor 4 gene. *Dev. Biol.* **132**: 331–336.

Siddiqui-Jain, A., Grand, C.L., Bearss, D.J. & Hurley, L.H. (2002) Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. *Proc. Natl. Acad. Sci. U.S.A.* **99**: 11593–11598.

Sun, D., Guo, K., Rusche, J.J. & Hurley, L.H. (2005) Facilitation of a structural transition in the polypurine/polypyrimidine tract within the proximal promoter region of the human VEGF gene by the presence of potassium and G-quadruplex-interactive agents. *Nucleic Acids. Res*. **33**: 6070–6080.

Takane, H., Shikata, E., Otsubo, K., Higuchi, S. & Ieiri, I. (2008) Polymorphism in human organic cation transporters and metformin action. *Pharmacogenomics.* **9**: 415–422.

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

Tishkoff, S.A., Reed, F.A., Friedlaender, F.R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J.B., Awomoyi, A.A., Bodo, J-M., Doumbo, O., *et al.* (2009) The genetic structure and history of Africans and African Americans. *Science*. **324**: 1035–1044.

United Nations Statistics Division - Standard Country and Area Codes Classifications (M49). http://millenniumindicators.un.org/unsd/methods/m49/m49regin.htm

Volk, C. (2014) OCTs, OATs, and OCTNs: structure and function of the polyspecific organic ion transporters of the SLC22 family. *WIREs Membr. Transp. Signal*. **3**: 1–13.

Wang, K., Li, M. & Hakonarson, H. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. **38**: e164.

Wang, Z. (2007) Genetic polymorphism of human organic cation transporter subtype 2 and genotype-phenotype relationship in Chinese population. PhD thesis, The Chinese University of Hong Kong.

Yates, A., Akanni, W., Amode, M.R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L., *et al*. (2015) Ensembl 2016. *Nucleic Acids Res.* **44**: D710–D716.

Yonezawa, A., Masuda, S., Nishihara, K., Yano, I., Katsura, T. & Inui, K. (2005) Association between tubular toxicity of cisplatin and expression of organic cation transporter rOCT2 (Slc22a2) in the rat. *Biochem Pharmacol.* **70**: 1823–1831.

Yonezawa, A., Masuda, S., Yokoo, S., Katsura, T. & Inui, K. (2006) Cisplatin and oxaliplatin, but not carboplatin and nedaplatin, are substrates for human organic cation transporters (SLC22A1–3 and multidrug and toxin extrusion family). *J. Pharmacol. Exp. Ther.* **319**: 879–886.

Zhang, M.Q. (1998) Identification of human gene core promoters in silico. *Genome Res*. **8**(3): 319–326.

Zhang, S., Lovejoy, K.S., Shima, J.E., Lagpacan, L.L., Shu, Y., Lapuk, A., Chen, Y., Komori, T. *et al*. (2006) Organic cation transporters are determinants of oxaliplatin cytotoxicity. *Cancer Res.* **66**: 8847–8857.

Zhang, X., Evans, K.K. & Wright, S.H. (2003) Molecular cloning of rabbit organic cation transporter rbOCT2 and functional comparisons with rbOCT1. *Am. J. Physiol. Renal Physiol.* **283**: F124– F133.

Zienolddiny, S. & Skaug, V. (2012) Single nucleotide polymorphisms as susceptibility, prognostic, and therapeutic markers of nonsmall cell lung cancer. *Lung cancer: Targets and Therapy*. **3**: 1–14.

APPENDIX A

Reagent recipes

<u>1 X PBS pH 7.2 (1 L):</u>

8 g NaCl

0.2 g KCl

2.90 g $Na_2HPO_4 \cdot 12H_2O$

0.24 g KH_2PO_4

The volume was adjusted to 1 L using distilled water and then autoclaved.

<u>0.5 M EDTA pH 8.0 (100 ml):</u>

18.2 g Na_2EDTA

Adjusted to pH 8.0 using NaOH pellets

The volume was adjusted to 100 ml using distilled water and then autoclaved.

<u>50 X TAE buffer (100 ml):</u>

24.2 g Trizma base

5.71 ml glacial acetic acid

10 ml 0.5 M EDTA at pH 8.0

The volume was adjusted to 100 ml using distilled water and then autoclaved.

This 50 X TAE buffer was diluted to make 1 X TAE buffer using distilled water.

<u>Lysis buffer (50 ml):</u>

10 mM Tris – 0.061 g Trizma base

100 mM EDTA – 1.46 g Na_2EDTA

2 % Sodium dodecyl sulphate (SDS) – 1.0 g SDS

The volume was adjusted to 50 ml using distilled water and then autoclaved.

<u>1 % Agarose gel:</u>

0.5 g agarose powder dissolved in 50 ml of 1 X TAE buffer

1 µl of 10 mg/ml stock Ethidium bromide (200 ng/ml final concentration).

6 X DNA loading dye:

3 ml glycerol

25 mg bromophenol blue

The volume was adjusted to 10 ml using distilled water.

DNA ladder mixture (100 µl):

50 µl of GeneRuler^{TM} 1kb DNA ladder

16.7 µl of 6 X DNA loading dye

33.3 µl autoclaved Milli-Q water

25 mM Tris-HCl, pH 8.0 and 10 mM EDTA (10 ml):

100 mM Tris (pH 8.0) – 0.12 g Trizma Base where the pH was adjusted using HCl. The volume was adjusted to 10 ml using distilled water and then autoclaved.

40 mM EDTA – 0.5 M EDTA was diluted using distilled water.

Equal volumes (2.5 ml) of each and double the volume (5 ml) of distilled water were mixed together to achieve the final concentration required.

0.4 M NaOH and 2 % SDS (12 ml):

0.8 M NaOH – 0.192 g NaOH, volume was adjusted to 6 ml using distilled water and then autoclaved.

4 % SDS – 0.32 g SDS, volume was adjusted to 6 ml using distilled water.

Equal volumes of each were mixed together to achieve the final concentration required.

3 M sodium acetate (10 ml):

2.46 g $NaC_2H_3O_2$, volume was adjusted to 10 ml using distilled water and then autoclaved.

5 M Potassium Acetate (pH 5.5) (10 ml):

2.941 g potassium acetate, volume was adjusted to 6 ml using distilled water.

1.15 ml of glacial acetic acid was added to the 6 ml above and the solution was adjusted to pH 5.5. Once the pH was achieved the volume was adjusted to 10 ml using distilled water and then autoclaved.

70 % ethanol (10 ml):

7 ml of 100 % ethanol mixed with 3 ml of distilled water.

LB Agar (50 ml):

1.75 g LB agar powder was dissolved in 50 ml of distilled water and then autoclaved.

LB Broth (50 ml):

1.0 g LB broth powder was dissolved in 50 ml of distilled water and then autoclaved.

SOC Broth (50 ml):

49.25 ml of autoclaved LB broth, 0.5 ml of 250 mM KCl and 0.25 ml of 2 M $MgCl_2$.

1 M KCl (10 ml):

0.74 g KCl salt was dissolved in 10 ml of distilled water and then autoclaved.

This stock solution was diluted using distilled water to make 250 mM KCl.

2 M $MgCl_2$ (10 ml):

1.90 g $MgCl_2$ powder was dissolved in 10 ml of distilled water and then autoclaved.

This stock solution was diluted using distilled water to make 0.1 M $MgCl_2$.

1 M $CaCl_2$ (10 ml):

1.11 g $CaCl_2$ salt was dissolved in 10 ml of distilled water and then autoclaved.

This stock solution was diluted using distilled water to make 0.1 M $CaCl_2$.

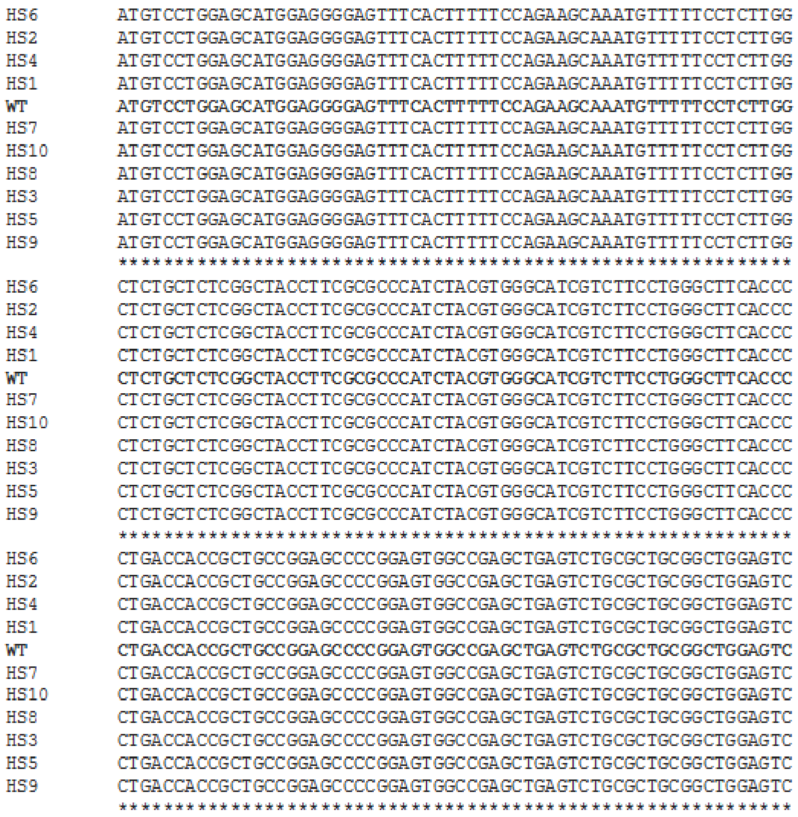
The complete MSA for HS1-10 APPENDIX B

```
HS6        TAGCCCCCTGATGTGTGAGAGCAGAAAGGGGTGATCCCTTTCTTCCTTATCCTAAGGCTC
HS2        TAGCCCCCTGATGTGTGAGAGCAGAAAGGGGTGATCCCTTTCTTCCTTATCCTAAGGCTC
HS4        TAGCCCCCTGATGTGTGAGAGCAGAAAGGGGTGATCCCTTTCTTCCTTATCCTAAGGCTC
HS1        TAGCCCCCTGATGTGTGAGAGCAGAAAGGGGTGATCCCTTTCTTCCTTATCCTAAGGCTC
WT         tagccccctgatgtgtgagagcagaaaggggtgatccctttcttccttatcctaaggctc
HS7        TAGCCCCCTGATGTGTGAGAGCAGAAAGGGGTGATCCCTTTCTTCCTTATCCTAAGGCTC
HS10       TAGCCCCCTGATGTGTGAGAGCAGAAAGGGGTGATCCCTTTCTTCCTTATCCTAAGGCTC
HS8        TAGCCCCCTGATGTGTGAGAGCAGAAAGGGGTGATCCCTTTCTTCCTTATCCTAAGGCTC
HS3        TAGCCCCCTGATGTGTGAGAGCAGAAAGGGGTGATCCCTTTCTTCCTTATCCTAAGGCTC
HS5        TAGCCCCCTGATGTGTGAGAGCAGAAAGGGGTGATCCCTTTCTTCCTTATCCTAAGGCTC
HS9        TAGCCCCCTGATGTGTGAGAGCAGAAAGGGGTGATCCCTTTCTTCCTTATCCTAAGGCTC
           ************************************************************

HS6        ACGGCCAACACCCCTATAACAAAAGACAGGTTAACAAGAGAAAAGCATGACAAATTTATT
HS2        ACGGCCAACACCCCTATAACAAAAGACAGGTTAACAAGAGAAAAGCATGACAAATTTATT
HS4        ACGGCCAACACCCCTATAACAAAAGACAGGTTAACAAGAGAAAAGCATGACAAATTTATT
HS1        ACGGCCAACACCCCTATAACAAAAGACAGGTTAACAAGAGAAAAGCATGACAAATTTATT
WT         acggccaacacccctataacaaaagacaggttaacaagagaaaagcatgacaaatttatt
HS7        ACGGCCAACACCCCTATAACAAAAGACAGGTTAACAAGAGAAAAGCATGACAAATTTATT
HS10       ACGGCCAACACCCCTATAACAAAAGACAGGTTAACAAGAGAAAAGCATGACAAATTTATT
HS8        ACGGCCAACACCCCTATAACAAAAGACAGGTTAACAAGAGAAAAGCATGACAAATTTATT
HS3        ACGGCCAACACCCCTATAACAAAAGACAGGTTAACAAGAGAAAAGCATGACAAATTTATT
HS5        ACGGCCAACACCCCTATAACAAAAGACAGGTTAACAAGAGAAAAGCATGACAAATTTATT
HS9        ACGGCCAACACCCCTATAACAAAAGACAGGTTAACAAGAGAAAAGCATGACAAATTTATT
           ************************************************************

HS6        TGATCACGTTTTACATGACACAGGAGCCTTCAGAATGAAGACCCACAGATACAGGGAAAA
HS2        TGATCACGTTTTACATGACACAGGAGCCTTCAGAATGAAGACCCACAGATACAGGGAAAA
HS4        TGATCACGTTTTACATGACACAGGAGCCTTCAGAATGAAGACCCACAGATACAGGGAAAA
HS1        TGATCACGTTTTACATGACACAGGAGCCTTCAGAATGAAGACCCACAGATACAGGGAAAA
WT         tgatcacgttttacatgacacaggagccttcagaatgaagacccacagatacagggaaaa
HS7        TGATCACGTTTTACATGACACAGGAGCCTTCAGAATGAAGACCCACAGATACAGGGAAAA
HS10       TGATCACGTTTTACATGACACAGGAGCCTTCAGAATGAAGACCCACAGATACAGGGAAAA
HS8        TGATCACGTTTTACATGACACAGGAGCCTTCAGAATGAAGACCCACAGATACAGGGAAAA
HS3        TGATCACGTTTTACATGACACAGGAGCCTTCAGAATGAAGACCCACAGATACAGGGAAAA
HS5        TGATCACGTTTTACATGACACAGGAGCCTTCAGAATGAAGACCCACAGATACAGGGAAAA
HS9        TGATCACGTTTTACATGACACAGGAGCCTTCAGAATGAAGACCCACAGATACAGGGAAAA
           ************************************************************

HS6        CTGGATGGCAGTATAGAACTGTAGTTGGACAAAAAGGGCAGCAGCCCATGTTCGCAGGCT
HS2        CTGGATGGCAGTATAGAACTGTAGTTGGACAAAAAGGGCAGCAGCCCATGTTCGCAGGCT
HS4        CTGGATGGCAGTATAGAACTGTAGTTGGACAAAAAGGGCAGCAGCCCATGTTCGCAGGCT
HS1        CTGGATGGCAGTATAGAACTGTAGTTGGACAAAAAGGGCAGCAGCCCATGTTCGCAGGCT
WT         ctggatggcagtatagaactgtagttggacaaaaagggcagcagcccatgttcgcaggct
HS7        CTGGATGGCAGTATAGAACTGTAGTTGGACAAAAAGGGCAGCAGCCCATGTTCGCAGGCT
HS10       CTGGATGGCAGTATAGAACTGTAGTTGGACAAAAAGGGCAGCAGCCCATGTTCGCAGGCT
HS8        CTGGATGGCAGTATAGAACTGTAGTTGGACAAAAAGGGCAGCAGCCCATGTTCGCAGGCT
HS3        CTGGATGGCAGTATAGAACTGTAGTTGGACAAAAAGGGCAGCAGCCCATGTTCGCAGGCT
HS5        CTGGATGGCAGTATAGAACTGTAGTTGGACAAAAAGGGCAGCAGCCCATGTTCGCAGGCT
HS9        CTGGATGGCAGTATAGAACTGTAGTTGGACAAAAAGGGCAGCAGCCCATGTTCGCAGGCT
           ************************************************************

HS6        GAGGGGAAAACCCAGCAAGGCCTGTCTGTTCAGATCCGTCTTGGCCCCTCTGTGCAGCAC
HS2        GAGGGGAAAACCCAGCAAGGCCTGTCTGTTCAGATCCGTCTTGGCCCCTCTGTGCAGCAC
HS4        GAGGGGAAAACCCAGCAAGGCCTGTCTGTTCAGATCCGTCTTGGCCCCTCTGTGCAGCAC
HS1        GAGGGGAAAACCCAGCAAGGCCTGTCTGTTCAGATCCGTCTTGGCCCCTCTGTGCAGCAC
WT         gaggggaaaacccagcaaggcctgtctgttcagatccgtcttggcccctctgtgcagcac
HS7        GAGGGGAAAACCCAGCAAGGCCTGTCTGTTCAGATCCGTCTTGGCCCCTCTGTGCAGCAC
HS10       GAGGGGAAAACCCAGCAAGGCCTGTCTGTTCAGATCCGTCTTGGCCCCTCTGTGCAGCAC
HS8        GAGGGGAAAACCCAGCAAGGCCTGTCTGTTCAGATCCGTCTTGGCCCCTCTGTGCAGCAC
HS3        GAGGGGAAAACCCAGCAAGGCCTGTCTGTTCAGATCCGTCTTGGCCCCTCTGTGCAGCAC
HS5        GAGGGGAAAACCCAGCAAGGCCTGTCTGTTCAGATCCGTCTTGGCCCCTCTGTGCAGCAC
HS9        GAGGGGAAAACCCAGCAAGGCCTGTCTGTTCAGATCCGTCTTGGCCCCTCTGTGCAGCAC
           ************************************************************

HS6        TCCTTCCTCCAGGCACCGGGGACAAGACTCCTCTGGAATGCGGGTCTGGATTTCTTTACG
HS2        TCCTTCCTCCAGGCACCGGGGACAAGACTCCTCTGGAATGCGGGTCTGGATTTCTTTACG
HS4        TCCTTCCTCCAGGCACCGGGGACAAGACTCCTCTGGAATGCGGGTCTGGATTTCTTTACG
HS1        TCCTTCCTCCAGGCACCGGGGACAAGACTCCTCTGGAATGCGGGTCTGGATTTCTTTACG
WT         tccttcctccaggcaccggggacaagactcctctggaatgcgggtctggatttctttacg
HS7        TCCTTCCTCCAGGCACCGGGGACAAGACTCCTCTGGAATGCGGGTCTGGATTTCTTTACG
HS10       TCCTTCCTCCAGGCACCGGGGACAAGACTCCTCTGGAATGCGGGTCTGGATTTCTTTACG
HS8        TCCTTCCTCCAGGCACCGGGGACAAGACTCCTCTGGAATGCGGGTCTGGATTTCTTTACG
HS3        TCCTTCCTCCAGGCACCGGGGACAAGACTCCTCTGGAATGCGGGTCTGGATTTCTTTACG
HS5        TCCTTCCTCCAGGCACCGGGGACAAGACTCCTCTGGAATGCGGGTCTGGATTTCTTTACG
HS9        TCCTTCCTCCAGGCACCGGGGACAAGACTCCTCTGGAATGCGGGTCTGGATTTCTTTACG
           ************************************************************

HS6        GCCCACTGTTACACAGAAAGGCAGCGGGGAAGTTACAGTGGTATTTCTAGGCTTTCTGGC
HS2        GCCCACTGTTACACAGAAAGGCAGCGGGGAAGTTACAGTGGTATTTCTAGGCTTTCTGGC
HS4        GCCCACTGTTACACAGAAAGGCAGCGGGGAAGTTACAGTGGTATTTCTAGGCTTTCTGGC
HS1        GCCCACTGTTACACAGAAAGGCAGCGGGGAAGTTACAGTGGTATTTCTAGGCTTTCTGGC
WT         gcccactgttacacagaaaggcagcggggaagttacagtggtatttctaggctttctggc
HS7        GCCCACTGTTACACAGAAAGGCAGCGGGGAAGTTACAGTGGTATTTCTAGGCTTTCTGGC
HS10       GCCCACTGTTACACAGAAAGGCAGCGGGGAAGTTACAGTGGTATTTCTAGGCTTTCTGGC
HS8        GCCCACTGTTACACAGAAAGGCAGCGGGGAAGTTACAGTGGTATTTCTAGGCTTTCTGGC
HS3        GCCCACTGTTACACAGAAAGGCAGCGGGGAAGTTACAGTGGTATTTCTAGGCTTTCTGGC
HS5        GCCCACTGTTACACAGAAAGGCAGCGGGGAAGTTACAGTGGTATTTCTAGGCTTTCTGGC
HS9        GCCCACTGTTACACAGAAAGGCAGCGGGGAAGTTACAGTGGTATTTCTAGGCTTTCTGGC
           ************************************************************

HS6        TGGCTTTGGGGAGAAAAGAGTCTGGTTTCCACGAGCTGCTTTGAGG---AAGGATTCTCA
HS2        TGGCTTTGGGGAGAAAAGAGTCTGGTTTCCACGAGCTGCTTTGAGG---AAGGATTCTCA
HS4        TGGCTTTGGGGAGAAAAGAGTCTGGTTTCCACGAGCTGCTTTGAGG---AAGGATTCTCA
HS1        TGGCTTTGGGGAGAAAAGAGTCTGGTTTCCACGAGCTGCTTTGAGGAAGAAGGATTCTCA
WT         tggctttggggagaaaagagtctggtttccacgagctgctttgaggaagaaggattctca
HS7        TGGCTTTGGGGAGAAAAGAGTCTGGTTTCCACGAGCTGCTTTGAGGAAGAAGGATTCTCA
HS10       TGGCTTTGGGGAGAAAAGAGTCTGGTTTCCACGAGCTGCTTTGAGGAAGAAGGATTCTCA
HS8        TGGCTTTGGGGAGAAAAGAGTCTGGTTTCCACGAGCTGCTTTGAGGAAGAAGGATTCTCA
HS3        TGGCTTTGGGGAGAAAAGAGTCTGGTTTCCACGAGCTGCTTTGAGGAAGAAGGATTCTCA
HS5        TGGCTTTGGGGAGAAAAGAGTCTGGTTTCCACGAGCTGCTTTGAGGAAGAAGGATTCTCA
HS9        TGGCTTTGGGGAGAAAAGAGTCTGGTTTCCACGAGCTGCTTTGAGGAAGAAGGATTCTCA
           **********************************************   ***********
```

```
HS6         GTTCTATGGCTTGCCCCAGGGGAGAATGATGGGTGAGAGAAGAGACAGGAGGGCAGGAGA
HS2         GTTCTATGGCTTGCCCCAGGGGAGAATGATGGGTGAGAGAAGAGACAGGAGGGCAGGAGA
HS4         GTTCTATGGCTTGCCCCGGGGGAGAATGATGGGTGAGAGAAGAGACAGGAGGGCAGGAGA
HS1         GTTCTATGGCTTGCCCCGGGGGAGAATGATGGGTGAGAGAAGAGACAGGAGGGCAGGAGA
WT          gttctatggcttgccccgggggagaatgatgggtgagagaagagacaggagggcaggaga
HS7         GTTCTATGGCTTGCCCCGGGGGAGAATGATGGGTGAGAGAAGAGACAGGAGGGCAGGAGA
HS10        GTTCTATGGCTTGCCCCGGGGGAGAATGATGGGTGAGAGAAGAGACAGGAGGGCAGGAGA
HS8         GTTCTATGGCTTGCCCCGGGGGAGAATGATGGGTGAGAGAAGAGACAGGAGGGCAGGAGA
HS3         GTTCTATGGCTTGCCCCGGGGGAGAATGATGGGTGAGAGAAGAGACAGGAGGGCAGGAGA
HS5         GTTCTATGGCTTGCCCCGGGGGAGAATGATGGGTGAGAGAAGAGACAGGAGGGCAGGAGA
HS9         GTTCTATGGCTTGCCCCGGGGGAGAATGATGGGTGAGAGAAGAGACAGGAGGGCAGGAGA
            ***************** ******************************************

HS6         AGGTCAGAGAGAGAGACTTTGCTTCTGAGGCCTCCACCTTGGGGCATGGATTTCTGAGCC
HS2         AGGTCAGAGAGAGAGACTTTGCTTCTGAGGCCTCCACCTTGGGGCATGGATTTCTGAGCC
HS4         AGGTCAGAGAGAGAGACTTTGCTTCTGAGGCCTCCACCTTGGGGCATGGATTTCTGAGCC
HS1         AGGTCAGAGAGAGAGACTTTGCTTCTGAGGCCTCCACCTTGGGGCATGGATTTCTGAGCC
WT          aggtcagagagagagactttgcttctgaggcctccaccttggggcatggatttctgagcc
HS7         AGGTCAGAGAGAGAGACTTTGCTTCTGAGGCCTCCACCTTGGGGCATGGATTTCTGAGCC
HS10        AGGTCAGAGAGAGAGACTTTGCTTCTGAGGCCTCCACCTTGGGGCATGGATTTCTGAGCC
HS8         AGGTCAGAGAGAGAGACTTTGCTTCTGAGGCCTCCACCTTGGGGCATGGATTTCTGAGCC
HS3         AGGTCAGAGAGAGAGACTTTGCTTCTGAGGCCTCCACCTTGGGGCATGGATTTCTGAGCC
HS5         AGGTCAGAGAGAGAGACTTTGCTTCTGAGGCCTCCACCTTGGGGCATGGATTTCTGAGCC
HS9         AGGTCAGAGAGAGAGACTTTGCTTCTGAGGCCTCCACCTTGGGGCATGGATTTCTGAGCC
            ************************************************************

HS6         CCAACAGACCTTGACAGAAAAATCTAGGACACAAAGATAGTGGCTTGGACACACCTGCCT
HS2         CCAACAGACCTTGACAGAAAAATCTAGGACACAAAGATAGTGGCTTGGACACACCTGCCT
HS4         CCAACAGACCTTGACAGAAAAATCTAGGACACAAAGATAGTGGCTTGGACACACCTGCCT
HS1         CCAACAGACCTTGACAGAAAAATCTAGGACACAAAGATAGTGGCTTGGACACACCTGCCT
WT          ccaacagaccttgacagaaaaatctaggacacaaagatagtggcttggacacacctgcct
HS7         CCAACAGACCTTGACAGAAAAATCTAGGACACAAAGATAGTGGCTTGGACACACCTGCCT
HS10        CCAACAGACCTTGACAGAAAAATCTAGGACACAAAGATAGTGGCTTGGACACACCTGCCT
HS8         CCAACAGACCTTGACAGAAAAATCTAGGACACAAAGATAGTGGCTTGGACACACCTGCCT
HS3         CCAACAGACCTTGACAGAAAAATCTAGGACACAAAGATAGTGGCTTGGACACACCTGCCT
HS5         CCAACAGACCTTGACAGAAAAATCTAGGACACAAAGATAGTGGCTTGGACACACCTGCCT
HS9         CCAACAGACCTTGACAGAAAAATCTAGGACACAAAGATAGTGGCTTGGACACACCTGCCT
            ************************************************************

HS6         GCATTTACACTTGACCTGTCTGCGACGTAAACACTTTCCTCTTTCCCTCCAGATGGGTTA
HS2         GCATTTACACTTGACCTGTCTGCGACGTAAACACTTTCCTCTTTCCCTCCAGATGGGTTA
HS4         GCATTTACACTTGACCTGTCTGCGACGTAAACACTTTCCTCTTTCCCTCCAGATGGGTTA
HS1         GCATTTACACTTGACCTGTCTGCGACGTAAACACTTTCCTCTTTCCCTCCAGATGGGTTA
WT          gcatttacacttgacctgtctgcgacgtaaacactttcctctttccctccagatgggtta
HS7         GCATTTACACTTGACCTGTCTGCGACGTAAACACTTTCCTCTTTCCCTCCAGATGGGTTA
HS10        GCATTTACACTTGACCTGTCTGCGACGTAAACACTTTCCTCTTTCCCTCCAGATGGGTTA
HS8         GCATTTACACTTGACCTGTCTGCGACGTAAACACTTTCCTCTTTCCCTCCAGATGGGTTA
HS3         GCATTTACACTTGACCTGTCTGCGACGTAAACACTTTCCTCTTTCCCTCCAGATGGGTTA
HS5         GCATTTACACTTGACCTGTCTGCGACGTAAACACTTTCCTCTTTCCCTCCAGATGGGTTA
HS9         GCATTTACACTTGACCTGTCTGCGACGTAAACACTTTCCTCTTTCCCTCCAGATGGGTTA
            ************************************************************

HS6         AGGGGAAGGACACTTCAGGGTTGAAACGCAGGAATACCAGATTGGAGCAAACACTTTTTA
HS2         AGGGGAAGGACACTTCAGGGTTGAAACGCAGGAATACCAGATTGGAGCAAACACTTTTTA
HS4         AGGGGAAGGACACTTCAGGGTTGAAACGCAGGAATACCAGATTGGAGCAAACACTTTTTA
HS1         AGGGGAAGGACACTTCAGGGTTGAAACGCAGGAATACCAGATTGGAGCAAACACTTTTTA
WT          aggggaaggacacttcagggttgaaacgcaggaataccagattggagcaaacactttttta
HS7         AGGGGAAGGACACTTCAGGGTTGAAACGCAGGAATACCAGATTGGAGCAAACACTTTTTA
HS10        AGGGGAAGGACACTTCAGGGTTGAAACGCAGGAATACCAGATTGGAGCAAACACTTTTTA
HS8         AGGGGAAGGACACTTCAGGGTTGAAACGCAGGAATACCAGATTGGAGCAAACACTTTTTA
HS3         AGGGGAAGGACACTTCAGGGTTGAAACGCAGGAATACCAGATTGGAGCAAACACTTTTTA
HS5         AGGGGAAGGACACTTCAGGGTTGAAACGCAGGAATACCAGATTGGAGCAAACACTTTTTA
HS9         AGGGGAAGGACACTTCAGGGTTGAAACGCAGGAATACCAGATTGGAGCAAACACTTTTTA
            ************************************************************

HS6         AAAGCAGAGTTATAAAATCTGGACAACATCAAAACAAGCAGCCCCAGCATGCATCCCGAC
HS2         AAAGCAGAGTTATAAAATCTGGACAACATCAAAACAAGCAGCCCCAGCATGCATCCCGAC
HS4         AAAGCAGAGTTATAAAATCTGGACAACATCAAAACAAGCAGCCCCAGCATGCATCCCGAC
HS1         AAAGCAGAGTTATAAAATCTGGACAACATCAAAACAAGCAGCCCCAGCATGCATCCCGAC
WT          aaagcagagttataaaatctggacaacatcaaaacaagcagccccagcatgcatcccgac
HS7         AAAGCAGAGTTATAAAATCTGGACAACATCAAAACAAGCAGCCCCAGCATGCATCCCGAC
HS10        AAAGCAGAGTTATAAAATCTGGACAACATCAAAACAAGCAGCCCCAGCATGCATCCCGAC
HS8         AAAGCAGAGTTATAAAATCTGGACAACATCAAAACAAGCAGCCCCAGCATGCATCCCGAC
HS3         AAAGCAGAGTTATAAAATCTGGACAACATCAAAACAAGCAGCCCCAGCATGCATCCCGAC
HS5         AAAGCAGAGTTATAAAATCTGGACAACATCAAAACAAGCAGCCCCAGCATGCATCCCGAC
HS9         AAAGCAGAGTTATAAAATCTGGACAACATCAAAACAAGCAGCCCCAGCATGCATCCCGAC
            ************************************************************

HS6         GGCTCTTGTTGTTGGTTGGAGAATGAGCCCAGCAGTCAGGCTTGCAACCCACTTCGAATC
HS2         GGCTCTTGTTGTTGGTTGGAGAATGAGCCCAGCAGTCAGGCTTGCAACCCACTTCGAATC
HS4         GGCTCTTGTTGTTGGTTGGAGAATGAGCCCAGCAGTCAGGCTTGCAACCCACTTCGAATC
HS1         GGCTCTTGTTGTTGGTTGGAGAATGAGCCCAGCAGTCAGGCTTGCAACCCACTTCGAATC
WT          ggctcttgttgttggttggagaatgagcccagcagtcaggcttgcaacccacttcgaatc
HS7         GGCTCTTGTTGTTGGTTGGAGAATGAGCCCAGCAGTCAGGCTTGCAACCCACTTCGAATC
HS10        GGCTCTTGTTGTTGGTTGGAGAATGAGCCCAGCAGTCAGGCTTGCAACCCACTTCGAATC
HS8         GGCTCTTGTTGTTGGTTGGAGAATGAGCCCAGCAGTCAGGCTTGCAACCCACTTCGAATC
HS3         GGCTCTTGTTGTTGGTTGGAGAATGAGCCCAGCAGTCAGGCTTGCAACCCACTTCGAATC
HS5         GGCTCTTGTTGTTGGTTGGAGAATGAGCCCAGCAGTCAGGCTTGCAACCCACTTCGAATC
HS9         GGCTCTTGTTGTTGGTTGGAGAATGAGCCCAGCAGTCAGGCTTGCAACCCACTTCGAATC
            ************************************************************

HS6         TGGACCAGGGTTCTGACACGGATCCTGGTTCACATCACGCTGGGCCTTGTGGCCAAACAC
HS2         TGGACCAGGGTTCTGACACGGATCCTGGTTCACATCACGCTGGGCCTTGTGGCCAAACAC
HS4         TGGACCAGGGTTCTGACACGGATCCTGGTTCACATCACGCTGGGCCTTGTGGCCAAACAC
HS1         TGGACCAGGGTTCTGACACGGATCCTGGTTCACATCACGCTGGGCCTTGTGGCCAAACAC
WT          tggaccagggttctgacacggatcctggttcacatcacgctgggccttgtggccaaacac
HS7         TGGACCAGGGTTCTGACACGGATCCTGGTTCACATCACGCTGGGCCTTGTGGCCAAACAC
HS10        TGGACCAGGGTTCTGACACGGATCCTGGTTCACATCACGCTGGGCCTTGTGGCCAAACAC
HS8         TGGACCAGGGTTCTGACACGGATCCTGGTTCACATCACGCTGGGCCTTGTGGCCAAACAC
HS3         TGGACCAGGGTTCTGACACGGATCCTGGTTCACATCACGCTGGGCCTTGTGGCCAAACAC
HS5         TGGACCAGGGTTCTGACACGGATCCTGGTTCACATCACGCTGGGCCTTGTGGCCAAACAC
HS9         TGGACCAGGGTTCTGACACGGATCCTGGTTCACATCACGCTGGGCCTTGTGGCCAAACAC
            ************************************************************
```

```
HS6         GTGTGTTTTCTCCATAGGGCCTTGAAGAAAAGCTGGCGGTGCGCATGAGATAGGAGTATA
HS2         GTGTGTTTTCTCCATAGGGCCTTGAAGAAAAGCTGGCGGTGCGCATGAGATAGGAGTATA
HS4         GTGTGTTTTCTCCATAGGGCCTTGAAGAAAAGCTGGCGGTGCGCATGAGATAGGAGTATA
HS1         GTGTGTTTTCTCCATAGGGCCTTGAAGAAAAGCTGGCGGTGCGCATGAGATAGGAGTATA
WT          gtgtgttttctccatagggccttgaagaaaagctggcggtgcgcatgagataggagtata
HS7         GTGTGTTTTCTCCATAGGGCCTTGAAGAAAAGCTGGCGGTGCGCATGAGATAGGAGTATA
HS10        GTGTGTTTTCTCCATAGGGCCTTGAAGAAAAGCTGGCGGTGCGCATGAGATAGGAGTATA
HS8         GTGTGTTTTCTCCATAGGGCCTTGAAGAAAAGCTGGCGGTGCGCATGAGATAGGAGTATA
HS3         GTGTGTTTTCTCCATAGGGCCTTGAAGAAAAGCTGGCGGTGCGCATGAGATAGGAGTATA
HS5         GTGTGTTTTCTCCATAGGGCCTTGAAGAAAAGCTGGCGGTGCGCATGAGATAGGAGTATA
HS9         GTGTGTTTTCTCCATAGGGCCTTGAAGAAAAGCTGGCGGTGCGCATGAGATAGGAGTATA
            ************************************************************
HS6         TTAAGTTCCTGGCTGCTCGGGGCACTACGGGAAGATTACTGGGCTGTGATATGGGCCAGC
HS2         TTAAGTTCCTGGCTGCTCGGGGCACTACGGGAAGATTACTGGGCTGTGATATGGGCCAGC
HS4         TTAAGTTCCTGGCTGCTCGGGGCACTACGGGAAGATTACTGGGCTGTGATATGGGCCAGC
HS1         TTAAGTTCCTGGCTGCTCGGGGCACTACGGGAAGATTACTGGGCTGTGATATGGGCCAGC
WT          ttaagttcctggctgctcggggcactacgggaagattactgggctgtgatatgggccagc
HS7         TTAAGTTCCTGGCTGCTCGGGGCACTACGGGAAGATTACTGGGCTGTGATATGGGCCAGC
HS10        TTAAGTTCCTGGCTGCTCGGGGCACTACGGGAAGATTACTGGGCTGTGATATGGGCCAGC
HS8         TTAAGTTCCTGGCTGCTCGGGGCACTACGGGAAGATTACTGGGCTGTGATATGGGCCAGC
HS3         TTAAGTTCCTGGCTGCTCGGGGCACTACGGGAAGATTACTGGGCTGTGATATGGGCCAGC
HS5         TTAAGTTCCTGGCTGCTCGGGGCACTACGGGAAGATTACTGGGCTGTGATATGGGCCAGC
HS9         TTAAGTTCCTGGCTGCTCGGGGCACTACGGGAAGATTACTGGGCTGTGATATGGGCCAGC
            ************************************************************
HS6         ACTCAGATTCCCTGCGGTGGGACACAGAGGGCGGGTTGTTTGTGCTGCTGGCGTGGAGCA
HS2         ACTCAGATTCCCTGCGGTGGGACACAGAGGGCGGGTTGTTTGTGCTGCTGGCGTGGAGCA
HS4         ACTCAGATTCCCTGCGGTGGGACACAGAGGGCGGGTTGTTTGTGCTGCTGGCGTGGAGCA
HS1         ACTCAGATTCCCTGCGGTGGGACACAGAGGGCGGGTTGTTTGTGCTGCTGGCATGGAGCA
WT          actcagattccctgcggtgggacacagagggcgggttgtttgtgctgctggcgtggagca
HS7         ACTCAGATTCCCTGCGGTGGGACACAGAGGGCGGGTTGTTTGTGCTGCTGGCGTGGAGCA
HS10        ACTCAGATTCCCTGCGGTGGGACACAGAGGGCGGGTTGTTTGTGCTGCTGGCGTGGAGCA
HS8         ACTCAGATTCCCTGCGGTGGGACACAGAGGGCGGGTTGTTTGTGCTGCTGGCGTGGAGCA
HS3         ACTCAGATTCCCTGCGGTGGGACACAGAGGGCGGGTTGTTTGTGCTGCTGGCGTGGAGCA
HS5         ACTCAGATTCCCTGCGGTGGGACACAGAGGGCGGGTTGTTTGTGCTGCTGGCGTGGAGCA
HS9         ACTCAGATTCCCTGCGGTGGGACACAGAGGGCGGGTTGTTTGTGCTGCTGGCGTGGAGCA
            **************************************************** *******
HS6         CCGACAAGCCTGTGGAGAACCAGTTATAATAAACACGACAGGCATCCTGGGAGTGAGCTC
HS2         CCGACAAGCCTGTGGAGAACCAGTTATAATAAACACGACAGGCATCCTGGGAGTGAGCTC
HS4         CCGACAAGCCTGTGGAGAACCAGTTATAATAAACACGACAGGCATCCTGGGAGTGAGCTC
HS1         CCGACAAGCCTGTGGAGAACCAGTTATAATAAACATGACAGGCATCCTGGGAGTGAGCTC
WT          ccgacaagcctgtggagaaccaGTTATAATAAACACGACAGGCATCCTGGGAGTGAGCTC
HS7         CCGACAAGCCTGTGGAGAACCAGTTATAATAAACATGACAGGCATCCTGGGAGTGAGCTC
HS10        CCGACAAGCCTGTGGAGAACCAGTTATAATAAACACGACAGGCATCCTGGGAGTGAGCTC
HS8         CCGACAAGCCTGTGGAGAACCAGTTATAATAAACACGACAGGCATCCTGGGAGTGAGCTC
HS3         CCGACAAGCCTGTGGAGAACCAGTTATAATAAACACGACAGGCATCCTGGGAGTGAGCTC
HS5         CCGACAAGCCTGTGGAGAACCAGTTATAATAAACACGACAGGCATCCTGGGAGTGAGCTC
HS9         CCGACAAGCCTGTGGAGAACCAGTTATAATAAACACGACAGGCATCCTGGGAGTGAGCTC
            *********************************** ************************
```

```
HS6         AGGGCATTTGGGAAGTGCAGAAGGACATGCACCCCCGCTGGAGGGGTGCACCTTTGAAGT
HS2         AGGGCATTTGGGAAGTGCAGAAGGACATGCACCCCCGCTGGAGGGGTGCACCTTTGAAGT
HS4         AGGGCATTTGGGAAGTGCAGAAGGACATGCACCCCCGCTGGAGGGGTGCACCTTTGAAGT
HS1         AGGGCATTTGGGAAGTGCAGAAGGACGTGCACCCCCGCTGGAGGGGTGCACCTTTGAAGT
WT          AGGGCATTTGGGAAGTGCAGAAGGACATGCACCCCCGCTGGAGGGGTGCACCTTTGAAGT
HS7         AGGGCATTTGGGAAGTGCAGAAGGACATGCACCCCCGCTGGAGGGGTGCACCTTTGAAGT
HS10        AGGGCATTTGGGAAGTGCAGAAGGACATGCACCCCCGCTGGAGGGGTGCACCTTTGAAGT
HS8         AGGGCATTTGGGAAGTGCAGAAGGACATGCACCCCCGCTGGAGGGGTGCACCTTTGAAGT
HS3         AGGGCATTTGGGAAGTGCAGAAGGACATGCACCCCCGCTGGAGGGGTGCACCTTTGAAGT
HS5         AGGGCATTTGGGAAGTGCAGAAGGACATGCACCCCCGCTGGAGGGGTGCACCTTTGAAGT
HS9         AGGGCATTTGGGAAGTGCAGAAGGACGTGCACCCCCGCTGGAGGGGTGCACCTTTGAAGT
            ************************** *********************************
HS6         CAGCTGGACCAAGGAAAGGCCCTGCCCTGAAGGCTGGTCACTTGCAGAGGTAAACTCCCC
HS2         CAGCTGGACCAAGGAAAGGCCCTGCCCTGAAGGCTGGTCACTTGCAGAGGTAAACTCCCC
HS4         CAGCTGGACCAAGGAAAGGCCCTGCCCTGAAGGCTGGTCACTTGCAGAGGTAAACTCCCC
HS1         CAGCTGGACCAAGGAAAGGCCCTGCCCTGAAGGCTGGTCACTTGCAGAGGTAAACTCCCC
WT          CAGCTGGACCAAGGAAAGGCCCTGCCCTGAAGGCTGGTCACTTGCAGAGGTAAACTCCCC
HS7         CAGCTGGACCAAGGAAAGGCCCTGCCCTGAAGGCTGGTCACTTGCAGAGGTAAACTCCCC
HS10        CAGCTGGACCAAGGAAAGGCCCTGCCCTGAAGGCTGGTCACTTGCAGAGGTAAACTCCCC
HS8         CAGCTGGACCAAGGAAAGGCCCTGCCCTGAAGGCTGGTCACTTGCAGAGGTAAACTCCCC
HS3         CAGCTGGACCAAGGAAAGGCCCTGCCCTGAAGGCTGGTCACTTGCAGAGGTAAACTCCCC
HS5         CAGCTGGACCAAGGAAAGGCCCTGCCCTGAAGGCTGGTCACTTGCAGAGGTAAACTCCCC
HS9         CAGCTGGACCAAGGAAAGGCCCTGCCCTGAAGGCTGGTCACTTGCAGAGGTAAACTCCCC
            ************************************************************
HS6         TCTTTGACTTCTGGCCAGGGTTTGTGCTGAGCTGGCTGCAGCCGCTCTCAGCCTCGCTCC
HS2         TCTTTGACTTCTGGCCAGGGTTTGTGCTGAGCTGGCTGCAGCCGCTCTCAGCCTCGCTCC
HS4         TCTTTGACTTCTGGCCAGGGTTTGTGCTGAGCTGGCTGCAGCCGCTCTCAGCCTCGCTCC
HS1         TCTTTGACTTCTGGCCAGGGTTTGTGCTGAGCTGGCTGCAGCCGCTCTCAGCCTCGCTCC
WT          TCTTTGACTTCTGGCCAGGGTTTGTGCTGAGCTGGCTGCAGCCGCTCTCAGCCTCGCTCC
HS7         TCTTTGACTTCTGGCCAGGGTTTGTGCTGAGCTGGCTGCAGCCGCTCTCAGCCTCGCTCC
HS10        TCTTTGACTTCTGGCCAGGGTTTGTGCTGAGCTGGCTGCAGCCGCTCTCAGCCTCGCTCC
HS8         TCTTTGACTTCTGGCCAGGGTTTGTGCTGAGCTGGCTGCAGCCGCTCTCAGCCTCGCTCC
HS3         TCTTTGACTTCTGGCCAGGGTTTGTGCTGAGCTGGCTGCAGCCGCTCTCAGCCTCGCTCC
HS5         TCTTTGACTTCTGGCCAGGGTTTGTGCTGAGCTGGCTGCAGCCGCTCTCAGCCTCGCTCC
HS9         TCTTTGACTTCTGGCCAGGGTTTGTGCTGAGCTGGCTGCAGCCGCTCTCAGCCTCGCTCC
            ************************************************************
HS6         GGGCACGTCGGGCAGCCTCGGGCCCTCCTGCCTGCAGGATCATGCCCACCACCGTGGACG
HS2         GGGCACGTCGGGCAGCCTCGGGCCCTCCTGCCTGCAGGATCATGCCCACCACCGTGGACG
HS4         GGGCACGTCGGGCAGCCTCGGGCCCTCCTGCCTGCAGGATCATGCCCACCACCGTGGACG
HS1         GGGCACGTCGGGCAGCCTCGGGCCCTCCTGCCTGCAGGATCATGCCCACCACCGTGGACG
WT          GGGCACGTCGGGCAGCCTCGGGCCCTCCTGCCTGCAGGATCATGCCCACCACCGTGGACG
HS7         GGGCACGTCGGGCAGCCTCGGGCCCTCCTGCCTGCAGGATCATGCCCACCACCGTGGACG
HS10        GGGCACGTCGGGCAGCCTCGGGCCCTCCTGCCTGCAGGATCATGCCCACCACCGTGGACG
HS8         GGGCACGTCGGGCAGCCTCGGGCCCTCCTGCCTGCAGGATCATGCCCACCACCGTGGACG
HS3         GGGCACGTCGGGCAGCCTCGGGCCCTCCTGCCTGCAGGATCATGCCCACCACCGTGGACG
HS5         GGGCACGTCGGGCAGCCTCGGGCCCTCCTGCCTGCAGGATCATGCCCACCACCGTGGACG
HS9         GGGCACGTCGGGCAGCCTCGGGCCCTCCTGCCTGCAGGATCATGCCCACCACCGTGGACG
            ************************************************************
```



```
HS6       ATGTCCTGGAGCATGGAGGGGAGTTTCACTTTTTCCAGAAGCAAATGTTTTTCCTCTTGG
HS2       ATGTCCTGGAGCATGGAGGGGAGTTTCACTTTTTCCAGAAGCAAATGTTTTTCCTCTTGG
HS4       ATGTCCTGGAGCATGGAGGGGAGTTTCACTTTTTCCAGAAGCAAATGTTTTTCCTCTTGG
HS1       ATGTCCTGGAGCATGGAGGGGAGTTTCACTTTTTCCAGAAGCAAATGTTTTTCCTCTTGG
WT        ATGTCCTGGAGCATGGAGGGGAGTTTCACTTTTTCCAGAAGCAAATGTTTTTCCTCTTGG
HS7       ATGTCCTGGAGCATGGAGGGGAGTTTCACTTTTTCCAGAAGCAAATGTTTTTCCTCTTGG
HS10      ATGTCCTGGAGCATGGAGGGGAGTTTCACTTTTTCCAGAAGCAAATGTTTTTCCTCTTGG
HS8       ATGTCCTGGAGCATGGAGGGGAGTTTCACTTTTTCCAGAAGCAAATGTTTTTCCTCTTGG
HS3       ATGTCCTGGAGCATGGAGGGGAGTTTCACTTTTTCCAGAAGCAAATGTTTTTCCTCTTGG
HS5       ATGTCCTGGAGCATGGAGGGGAGTTTCACTTTTTCCAGAAGCAAATGTTTTTCCTCTTGG
HS9       ATGTCCTGGAGCATGGAGGGGAGTTTCACTTTTTCCAGAAGCAAATGTTTTTCCTCTTGG
          ************************************************************
HS6       CTCTGCTCTCGGCTACCTTCGCGCCCATCTACGTGGGCATCGTCTTCCTGGGCTTCACCC
HS2       CTCTGCTCTCGGCTACCTTCGCGCCCATCTACGTGGGCATCGTCTTCCTGGGCTTCACCC
HS4       CTCTGCTCTCGGCTACCTTCGCGCCCATCTACGTGGGCATCGTCTTCCTGGGCTTCACCC
HS1       CTCTGCTCTCGGCTACCTTCGCGCCCATCTACGTGGGCATCGTCTTCCTGGGCTTCACCC
WT        CTCTGCTCTCGGCTACCTTCGCGCCCATCTACGTGGGCATCGTCTTCCTGGGCTTCACCC
HS7       CTCTGCTCTCGGCTACCTTCGCGCCCATCTACGTGGGCATCGTCTTCCTGGGCTTCACCC
HS10      CTCTGCTCTCGGCTACCTTCGCGCCCATCTACGTGGGCATCGTCTTCCTGGGCTTCACCC
HS8       CTCTGCTCTCGGCTACCTTCGCGCCCATCTACGTGGGCATCGTCTTCCTGGGCTTCACCC
HS3       CTCTGCTCTCGGCTACCTTCGCGCCCATCTACGTGGGCATCGTCTTCCTGGGCTTCACCC
HS5       CTCTGCTCTCGGCTACCTTCGCGCCCATCTACGTGGGCATCGTCTTCCTGGGCTTCACCC
HS9       CTCTGCTCTCGGCTACCTTCGCGCCCATCTACGTGGGCATCGTCTTCCTGGGCTTCACCC
          ************************************************************
HS6       CTGACCACCGCTGCCGGAGCCCCGGAGTGGCCGAGCTGAGTCTGCGCTGCGGCTGGAGTC
HS2       CTGACCACCGCTGCCGGAGCCCCGGAGTGGCCGAGCTGAGTCTGCGCTGCGGCTGGAGTC
HS4       CTGACCACCGCTGCCGGAGCCCCGGAGTGGCCGAGCTGAGTCTGCGCTGCGGCTGGAGTC
HS1       CTGACCACCGCTGCCGGAGCCCCGGAGTGGCCGAGCTGAGTCTGCGCTGCGGCTGGAGTC
WT        CTGACCACCGCTGCCGGAGCCCCGGAGTGGCCGAGCTGAGTCTGCGCTGCGGCTGGAGTC
HS7       CTGACCACCGCTGCCGGAGCCCCGGAGTGGCCGAGCTGAGTCTGCGCTGCGGCTGGAGTC
HS10      CTGACCACCGCTGCCGGAGCCCCGGAGTGGCCGAGCTGAGTCTGCGCTGCGGCTGGAGTC
HS8       CTGACCACCGCTGCCGGAGCCCCGGAGTGGCCGAGCTGAGTCTGCGCTGCGGCTGGAGTC
HS3       CTGACCACCGCTGCCGGAGCCCCGGAGTGGCCGAGCTGAGTCTGCGCTGCGGCTGGAGTC
HS5       CTGACCACCGCTGCCGGAGCCCCGGAGTGGCCGAGCTGAGTCTGCGCTGCGGCTGGAGTC
HS9       CTGACCACCGCTGCCGGAGCCCCGGAGTGGCCGAGCTGAGTCTGCGCTGCGGCTGGAGTC
          ************************************************************
```