UNIVERSITY of the
WESTERN CAPE

SANBI
South African National
Bioinformatics Institute

Integrating Regulatory and Methylome data for the discovery of clear cell Renal Cell Carcinoma (ccRCC) Variants

A thesis submitted in partial fulfilment of the requirements for the degree of Magister Scientiae in the Department of The South African National Bioinformatics Institute (SANBI), University of the Western Cape.

Tracey Calvert-Joshua

Student number: **3005444**

Supervisor: Dr Nicki Tiffin

Co-supervisor: Dr Junaid Gamieldien

Date: 24 November 2015

DAAD
South Africa

NRF National Research Foundation

KEYWORDS

clear cell renal cell carcinoma

non-coding DNA

whole genome sequencing

somatic mutations

ENCODE Project

transcription factor binding site disruptions

aberrant methylation

gene dysregulation

allele frequency

African population

ABSTRACT

Integrating Regulatory and Methylome data for the discovery of clear cell Renal Cell Carcinoma (ccRCC) Variants

T Calvert-Joshua

MSc Bioinformatics Thesis, Department of the South African National Bioinformatics Institute, University of the Western Cape

Kidney cancers, of which clear cell renal cell carcinoma comprises an estimated 70%, have been placed amongst the top ten most common cancers in both males and females. With a mortality rate that exceeds 40%, kidney cancer is considered the most lethal cancer of the genitourinary system. Despite advances in its treatment, the mortality- and incidence rates across all stages of the disease have continued to climb. Since the release of the Human Genome Project in the early 2000's, most genetics studies have focused on the protein coding region of the human genome, which accounts for a mere 2% of the entire genome. It has been suggested that diverting our focus to the other 98% of the genome, which was previously dismissed as non-functional "junk DNA", could possibly contribute significantly to our understanding of the underlying mechanisms of complex diseases.

In this study a whole genome sequencing somatic mutation dataset from the International Cancer Genome Consortium was used. The non-coding somatic mutations within the promoter, intronic, 5-prime untranslated and 3-prime untranslated regions of clear cell renal cell carcinoma-implicated genes were extracted and submitted to RegulomDB for their functional annotation.

As expected, most of the variants were located within the intronic regions and only a small subset of identified variants was predicted to be deleterious. Although the variants all belonged to a selected subset of kidney cancer-associated genes, the genes frequently mutated in the non-coding regions were not the same genes that were frequently mutated in the whole exome studies (where the focus is on the

coding sequences). This indicates that with whole genome sequencing studies a new set of genes/variants previously unassociated with the clear cell renal cell carcinoma could be identified. In addition, most of the non-coding somatic variants fell within multiple transcription factor binding sites. Since many of these variants were also deleterious (as predicted by RegulomDB), this suggests that mutations in the non-coding regions could contribute to disease due to their role in transcription factor binding site disruptions and their subsequent impact on transcriptional regulation. The substantial overlap between the genes with the most aberrantly methylated variants and the genes with the most transcription factor binding site disruptions, signifies a potential link between differential methylation and transcription factor binding site affinities. In contrast to the upregulated DNA methylation generally seen in promoter methylation studies, all of the significant hits in this study were hypomethylated, with the subsequent up-regulation of the genes of interest, suggesting that in the clear cell renal cell carcinoma, aberrant methylation may play a role in activating proto-oncogenes, rather than the silencing of genes. When a cross-analysis was carried out between the gene expression patterns and the transcription factor binding site disruptions, the non-coding somatic variants and differential methylation profiles, the genes affected again showed a clear overlap. Interestingly, most of the variants were not present in the 1000genomes data and thus represent novel mutations, which possibly occurred as a result of genomic instability. However, identifying novel variants are always promising, since they epitomise the possibility of developing pioneering ways to target diseases. The numerous detrimental effects a single non-coding mutation can have on other genomic processes have been demonstrated in this study and therefore validate the inclusion of non-coding regions of the genome in genetic studies in order to study complex multifactorial diseases.

24 November 2015

DECLARATION

I declare that *Integrating Regulatory and Methylome data for the discovery of clear cell Renal Cell Carcinoma (ccRCC) Variants* is my own work, that it has not been submitted for any degree or examination in any other university, and that all the sources I have used or quoted have been indicated and acknowledged by complete references.

Full name: Tracey Lynn Calvert-Joshua

Date: 24 November 2015

Signed.....

ACKNOWLEDGEMENTS

Abba Father, Lord Jesus and the ever-present Holy Spirit, all the words and languages in the world wouldn't suffice to articulate the majesty and captivating beauty of You. I will delight myself in You unceasingly, because all that You are overwhelms me.

To my husband Larry Joshua, you love me every day in so many ways by all the sacrifices you make for me- without murmuring, by continuing to treat me so preciously, even after 11 years. You are the most selfless, patient man I know. Thank you for being my sunshine and serenity. I love you.

Dr Nicki Tiffin, from the get-go I was impressed by the grace of your demeanour. You have been an inexhaustible source of knowledge, which you transfer to your mentees with both patience and humility. Thank you for letting me be a beneficiary of your expertise.

Thank you SANBI for allowing me to make the transition to a more computer-based mind-set and for all the individual members of SANBI, such as the likes of Dr Jean-Baka Entfellner, Hocine Bendou and Darlington Mapiye and Dr Junaid Gamieldien for your individual contributions.

Lastly, thank you to DAAD/NRF for lifting a mountain off my shoulders by investing financially into my career.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS AND ACRONYMS

DNA	deoxyribonucleic acid
RCC	renal cell carcinoma
ccRCC	clear cell renal cell carcinoma
GLOBOCAN	Global Burden of Cancer
LOH	Loss of heterozygosity
VHL	von Hippel-Lindau
BHD	Birt-Hogg-Dube
GE	gene expression
mRNA	messenger ribonucleic acid
pre-mRNA	pre-messenger RNAs
poly A	poly adenine
CDS	coding sequence
5'UTR	five prime untranslated region
3'UTR	three prime untranslated region
TSS	transcription start site
TFs	transcription factors
E. coli	*Escherichia coli*
TFBSs	Transcription factor binding sites
SNPs	Single Nucleotide Polymorphisms
Bps	base pairs
rSNPs	regulatory SNPs
AML	acute myeloid leukaemia
WGS	whole genome sequencing
WES	whole exome sequencing

m5C	5-methyl-cytosine
DNMTs	DNA methyltransferases
CpG	cytosine is immediately followed by a guanine
CGIs	CpG islands
β-value	Beta value
MAF	minor allele frequency
AF	allele frequency
NHGRI	National Human Genome Research Institute
ChIP	chromatin immunoprecipitation
FAIRE	Formaldehyde-Assisted Isolation of Regulatory Elements
NCBI	National Centre for Biotechnology Information
eQTLs	Expression Quantitative Trait Loci
SNVs	single nucleotide variants
GWAS	genome-wide association studies
DHSs	DNase hypersensitive sites
qPCR	quantitative Polymerase Chain Reaction
OMIM	Online Mendelian Inheritance in Man
IntOGen	Integrative Onco Genomics (31)
ICGC	International Cancer Genome Consortium
COSMIC	Catalogue of Somatic Mutations in Cancer
TCGA	The Cancer Genome Atlas
RefSeq	Reference Sequence project
HGNC Committee	Human Genome Organization Gene Nomenclature
EMBL-EBI Bioinformatics Institute	European Molecular Biology Library-European
UCSC	University of California at Santa Cruz

RNAP II	RNA polymerase 2
RNAP III	RNA polymerase 3
Bed	Browser Extensible data
BWA	Burrows-Wheeler Aligner
SVs	somatic variants
GO	Gene Ontology
GOI	genes of interest
RPKM	Reads Per Kilobase of transcript per Million
STDEV	standard deviation
KEGG	Kyoto Encyclopedia of Genes and Genomes
MCL	Markov Cluster Algorithm
MNPs	multi-nucleotide polymorphisms
GEO	Gene Expression Omnibus
PPARG	peroxisome proliferator activated receptor gamma
EBF1	Early B-cell factor 1
CEBPB	CCAAT/enhancer binding protein beta
CTCF	CCCTC-binding factor
UBC	ubiquitin C
UPS	ubiquitin-proteasome pathway
EGFR	epidermal growth factor receptor
DOCK2	Dedicator of cytokinesis 2 gene
CUBN	cubilin
PRPF8	pre-mRNA processing factor 8 gene
MITF	microphthalmia-associated transcription factor
PTPRD	protein tyrosine phosphatase receptor type delta
RUNX1	runt-related transcription factor 1

CHAPTER 1

1. INTRODUCTION AND LITERATURE REVIEW

1.1. Cancer Development

Cancer is a group of genetic diseases that develop as a result of germline mutations, epigenetic changes (which are non-inherited, modified gene functioning) or somatic mutations within normal cells (You and Jones, 2012). These mutations confer a selective growth advantage upon the newly transformed cells, which they, in turn, impart on their progeny during cell division (Ringo, 2004). Collectively these cells proliferate uncontrollably to form a tumour (Ringo, 2004). The six well-recognized features that cancer cells share are: uncontrolled proliferation, resistance to apoptosis (cell death), induction of angiogenesis (formation of new blood-vessels), circumvention of growth suppressors, tissue invasion and metastasis (Hanahan and Weinberg, 2011). In general, mutations in two classes of genes; tumour suppressor genes and proto-oncogenes, lead to cancer formation.

1.1.1. Tumour suppressor genes and oncogenes

Tumour suppressor genes are responsible for halting the cell cycle when the necessary cell signals which indicate that the integrity of the deoxyribonucleic acid (DNA) has been preserved, are not returned (Chau and Wang, 2003). Additionally, they promote DNA repair or apoptosis, depending on the extent of DNA damage during DNA replication. When tumour suppressor genes or the DNA sequences regulating these genes, undergo mutations, they may become underexpressed or inactivated due to loss-of-function mutations (Chau and Wang, 2003). Due to their inability to impede cell proliferation, they indirectly promote cancer development. By contrast, proto-oncogenes function in cell proliferation for growth, healing and tissue regeneration. They can however be converted to oncogenes through gain-of function point mutations, insertions, deletions, gene amplification events and chromosomal translocations (Chau and Wang, 2003). Oncogenes allow proteins to be expressed at higher-than-normal levels, resulting in tumourigenesis (Lodish et al., 2000a). After

several mutations the cell may develop the capacity to metastasize (Weinberg, 2007). Yet, despite the metastasis, cancers are always named after their site of origin called their primary tumour (Weinberg, 2007). Kidney cancers therefore originate within the kidneys as a result of normal kidney cells being transformed into malignant kidney cells.

1.2. Functions of the kidneys

The kidneys form part of the renal system which, besides the kidneys, is also comprised of the ureters, the bladder and the urethra (Mifflin and Shunker, 2005) (See Figure 1 for the anatomy of the kidney and the renal system). The major roles of the kidneys are the filtering and removal of wastes from the blood and the maintenance of electrolyte- and fluid balance within the body (Mifflin and Shunker, 2005). Over and above this, they also regulate the body's blood pressure and play a role in red blood cell synthesis, bone metabolism and maintaining the body's pH balance (Mifflin and Shunker, 2005).

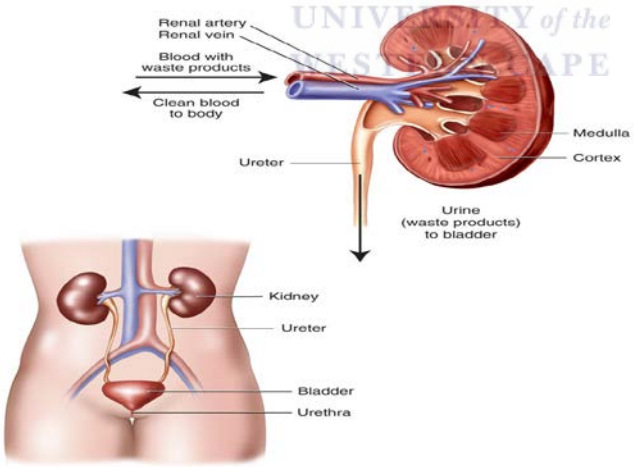


Figure 1: The bean-shaped structure of the kidney (top right) and the main organs forming the renal system (bottom) (Health Central, 2012).

1.3. Anatomy of the Kidneys and the Origin of Renal Cell Carcinoma

These bean-shaped organs are about the size of a fist. They are located in the upper abdominal cavity against the posterior muscular wall; one on either side of the vertebral column (Mifflin and Shunker, 2005). The functional units of the kidneys are the nephrons (~1 million per kidney). Each nephron contains a glomerulus, a collecting duct and the proximal- and distal convoluted tubules (See Figure 2. for the anatomy of the kidney and the nephron). The glomeruli filter waste and toxins from the body, forming a filtrate that generally also contains some useful chemicals (Mifflin and Shunker, 2005). The tubules are then responsible for reabsorbing essential water and chemicals from the filtrate back into the bloodstream. This leaves behind the urine which moves to the bladder via the collecting duct (Mifflin and Shunker, 2005). The proximal convoluted tubules, which lead from the glomerular structure, are the sites of origin of renal cell carcinoma (RCC); the most common type of kidney cancer (Zeng et al., 2014).

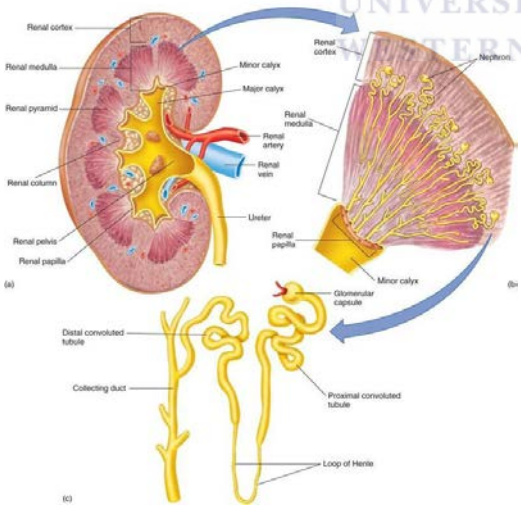


Figure 2: The anatomy of the kidney and the structure of the nephron which contains the collecting duct, the glomerulus and the tubules (McGraw-Hills Company, 2011).

1.4. Epidemiology of kidney cancers

Kidney cancers have been placed among the top ten most common cancers in both males and females (Ljungberg et al., 2011). With a mortality rate that exceeds 40% (Girgis *et al*., 2014), they are considered the most lethal cancers of the genitourinary system (Zeng *et al.*, 2013). Although kidney cancer affects all age groups, it is most commonly diagnosed in patients over 55 years of age (Tijani et al., 2012). Moreover, it occurs nearly twice as often in males than in females, which could be attributed to the prevalence of smoking and occupational exposures in males. Its subtype, RCC, accounts for more than 90% of all renal malignancies (Ljungberg et al., 2011). Yet, RCC is itself a heterogeneous group of cancers with the major subtype, clear cell renal cell carcinoma (ccRCC) accounting for approximately 70% of RCC cases (Zeng *et al.*, 2013). The other two major molecular subtypes, chromophobe- and papillary RCC, account for 5% and 10% of cases, respectively (Zeng *et al.*, 2013).

Kidney cancers are generally diagnosed rarely in Africans and Asians, but it is not yet clear whether this is due to a more efficient screening system in developed countries or to a predisposition to developing the disease in Caucasians (Tijani et al., 2012). Nevertheless, even in developed countries such as the United States, where individuals have access to the same level of healthcare, the renal cell cancer incidence rate differs among ethnic and racial groups, mirroring the rates of their countries of origin. This suggest that there might still be a genetic component that predisposes individuals to developing renal cell carcinoma (Chow et al., 2010).

Despite the higher incidence rates in Northern America and other developed countries, they have a much lower reported mortality rate. By contrast, in underdeveloped regions such as in Southern Africa, the mortality rate is almost equivalent to the incidence rate for kidney cancer. (See Figure 3 for the mortality and incidence rates in various parts of the world) (Global Burden of Cancer [GLOBOCAN], 2012).

However, more than 70% of RCC cases are diagnosed incidentally, when a patient is being screened for unrelated symptoms (Tijani et al., 2012). This could explain the lower mortality rate in developed countries, since an early diagnosis of the disease has been linked to a more favourable prognosis and in general, access to medical care is better in developed countries. In underdeveloped countries, patients are often only diagnosed during the advanced stages of the disease, resulting in a dismal prognosis (Tijani et al., 2012). Although lack of medical care has been suggested as a major contributor to the deficiency in proactive and targeted screening in underdeveloped countries, it has also been reported that an early diagnosis may be made in these regions, but due to cultural influences, many individuals choose to turn to traditional healers for help and only return when the cancer has metastasized (Tijani et al., 2012).

One cannot, however, rule out the asymptomatic nature of the disease as a primary reason for individuals not feeling prompted to seek medical assistance.

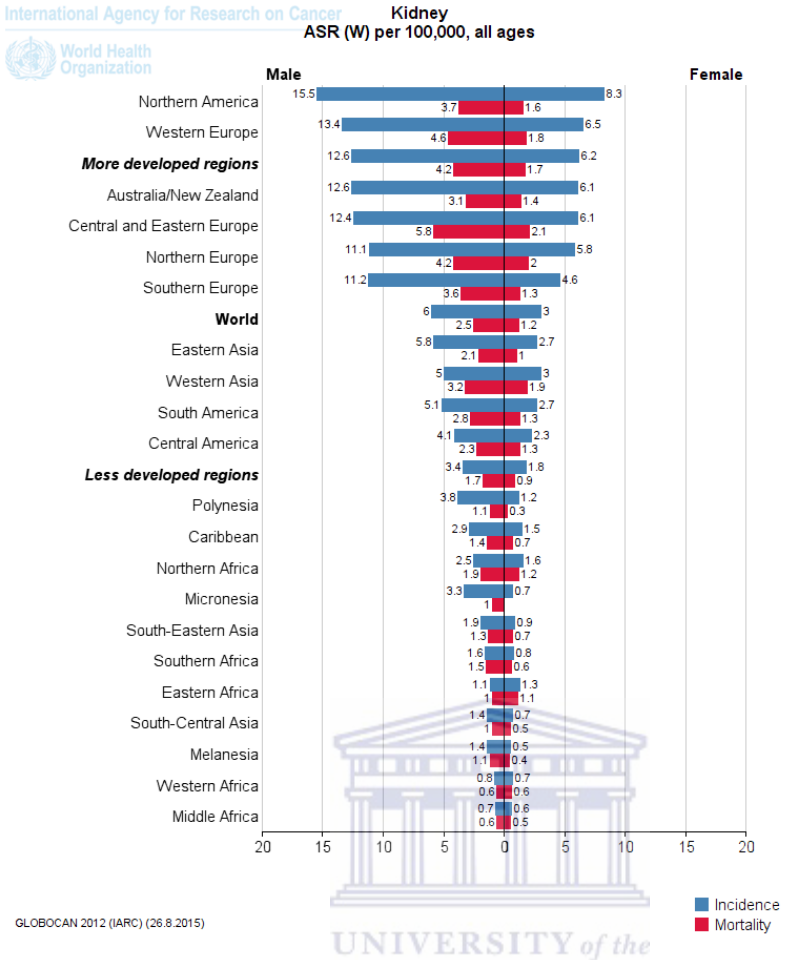


Figure 3: Age-standardized incidence and mortality rates for kidney cancer according to gender in various parts of the world. The incidence of kidney cancer is usually much higher in males than in females. Although the incidence and mortality rate in the more developed countries are higher compared to the less developed countries, the mortality rate is much lower than the corresponding incidence rate. In contrast, in the less developed countries, patients are much more likely to succumb to the disease (GLOBOCAN 2012).

1.5. Symptoms and Prognosis

RCC is usually asymptomatic, until the cancer has metastasized to distant organs (White et al., 2014). A "classical triad" of symptoms, namely: haematuria (blood in the urine), abdominal masses, and flank pain, which occur in unison, can sometimes be observed. Yet, this accounts for a mere 10% of individuals with kidney cancer (Tijani et al., 2012). The lack of defined symptoms often makes it a very difficult disease to diagnose. This is unfortunate, since an early diagnosis of RCC is

associated with a >80% chance of survival, five years after the diagnosis. However, more than 30% of patients are diagnosed at the metastatic stage when the survival rate drops to below 10%, as shown in Figure 4 (White et al., 2014).

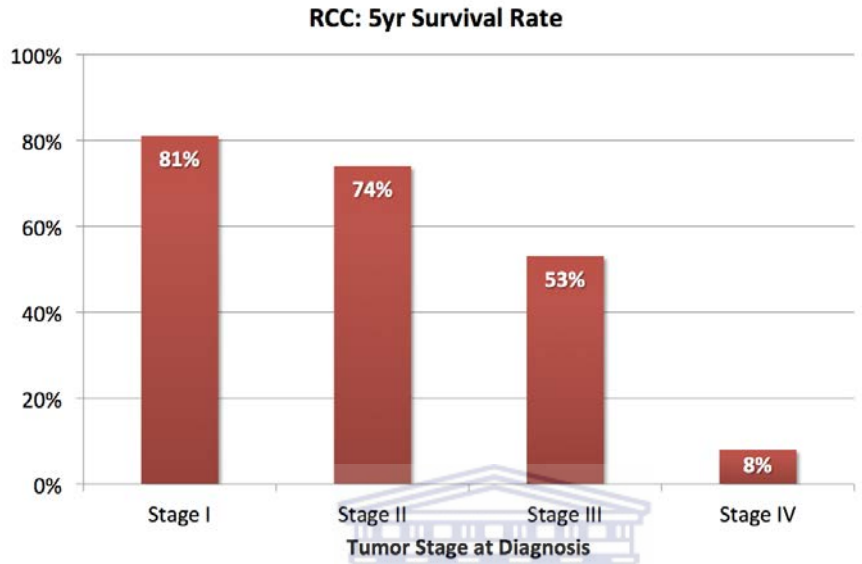


Figure 4: RCC 5-year survival rate, classified according to the Stage of the tumour (American Cancer Society, 2014). During the early stages of the disease (Stage I and even Stage II), the patient has an above 70% chance of surviving RCC, but during the metastatic stage (Stage 4) prognosis drops dismally to just 8%.

Metastatic RCC is complicated because it is highly resistant to radio- and other forms of systemic therapy, making surgical intervention the mainstay of treatment (Tijani et al., 2012). However patients with an advanced stage of the disease are generally too moribund and unfit for the recommended surgical intervention (Tijani et al., 2012). Furthermore, even after a partial or complete nephrectomy is carried out on patients declared fit for surgery, the prognosis is still very poor (Ljungberg et al., 2011). Approximately 20% of individuals have a relapse after surgery and subsequently develop metastatic RCC (Ljungberg et al., 2011). While kidney cancer does not necessarily discriminate against gender, race or age, certain risk factors can predispose an individual to developing the disease.

1.6. Risk Factors

Some of the most common risk factors are smoking, obesity and hypertension (Salehipoor et al., 2012). Nevertheless, taken together these account for only 49% of the cases. More than half of the cases remain unexplained and further risk factors need to be explored (Salehipoor et al., 2012). Currently, very little is known about the genetics of the disease development and progression and there are no serum biomarkers to accurately diagnose RCC (White et al., 2014). Cancer is however a genetic disease, thus a unique combination of mutational events can generally be observed for each sub-type of kidney cancer.

1.7. Cancer Genetics of clear cell renal cell carcinoma

Loss of heterozygosity (LOH) in the chromosome 3p arm is a consistent feature in 90% of ccRCC tumours, the primary subtype of RCC (Chau and Wang, 2003). LOH refers to the loss of the second functional copy of an allele in a heterozygous cell, making the cell homozygous for the mutated gene. The mutant gene is then often rendered non-functional, which may result in disease (Chau and Wang, 2003). Due to the size of the absent portion of the chromosome in ccRCC, it often encompasses all of the four most commonly mutated genes in ccRCC, namely: PBRM1, SETD2, BAP1 and VHL, as shown in Figure 5 (Gerlinger *et al.*, 2013). Furthermore, the LOH is often also associated with gains in the chromosome 5q arm, which harbours a number of proposed oncogenes (Girgis *et al.*, 2012). Of the commonly mutated genes in RCC, the von Hippel-Lindau (VHL) tumour suppressor gene is detected most frequently, accounting for approximately 60% of sporadic ccRCC (Zeng *et al.*, 2013). Similarly, the Birt-Hogg-Dube (BHD) gene has been associated with chromophobe RCC, while the c-MET oncogene has been linked to papillary RCC (Linehan et al., 2004). Recent microarray studies have also demonstrated the dynamic link between copy number aberrations, with a greater metastatic risk for patients with ccRCC and a subsequently poorer clinical prognosis (Girgis *et al.*, 2014). Still, few chromosomal abnormalities have been documented in RCCs and the pathogenesis of this disease

has of yet not been fully elucidated (White *et al.*, 2014). Furthermore, despite an increase in incidental diagnosis and improvements in screening techniques and treatment, there has been an increase in the disease-specific mortality rate within the last two decades (Dall'Oglio et al., 2011).

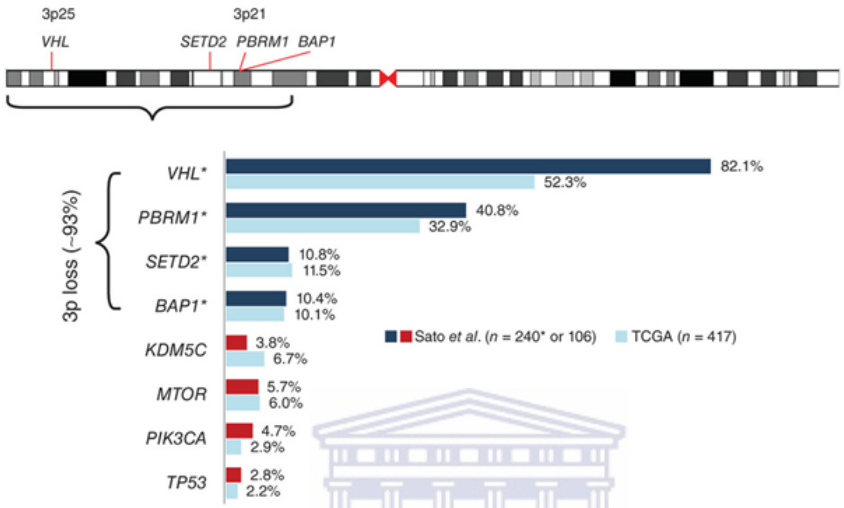


Figure 5: The frequency at which each gene of interest in ccRCC samples are mutated as a result of loss of heterozygosity (LOH) in the chromosome 3p arm (Hakimi et al., 2013). VHL PBRM1, SETD2 and BAP1 are four of the genes most commonly implicated in ccRCC.

1.8. Lack of understanding of disease

Understanding the pathogenicity of renal cancer on a molecular level (both genetically and epigenetically) is therefore essential for its early diagnosis, prognosis and for drug development (Girgis *et al.*, 2014), which could ultimately lessen the burden of this disease. Our current knowledge on RCC is limited, due to it being based solely on studies that involved interpreting, analysing and drawing inferences from the 2% of the human genome that encodes proteins (Skubitz and Skubitz, 2002). This has left us with not much success in elucidating the disease pathogenesis. For years, differential gene expression (GE) studies have shown that changes in the expression levels of genes and proteins are critical for malignant transformation (Skubitz and Skubitz, 2002). Yet, not many studies have delved into why and how these genes are dysregulated in the tissue or organ of interest (Skubitz and Skubitz,

2002). Albeit, several studies within the last decade have highlighted that it may be worthwhile to explore how the non-coding regions of the genome may contribute to the diseased phenotype (Linehan, 2012). Determining the locations of the regulatory regions and how they influence transcription or translation could reveal the links between gene dysregulation and disease. The next section will give a brief overview of the synthesis and functions of the non-coding regions surrounding genes in order to explore whether these regions could possibly control or impact gene dysregulation.

1.9. The Human Genome and Cancer

1.9.1. Intragenic and Intergenic DNA

The human genome is made up of genes that are protein-coding, genes that encode functional ribonucleic acid (RNA) products, as well as large regions of non-coding DNA (Gaffney and Keightley, 2006) (Wong et al., 2000). In higher eukaryotes, approximately 97% of the genome does not code for proteins. Furthermore, the chromatin (DNA wrapped around histone proteins) in the nucleus is divided into the euchromatic portion and the heterochromatic portion (e.g. centromeres and telomeres) (Wong et al., 2000). The latter portion consists of highly repetitive DNA and is largely devoid of genes (Wong et al., 2000). The euchromatic portion will be the focus of this study. Euchromatin can be further subcategorized into the intergenic region (stretches of non-coding DNA *between* adjacent genes) and the intragenic regions (stretches of DNA *within* the same gene) (Wong et al., 2000) (Figure 6 illustrates the difference between intergenic and intragenic DNA). The intergenic region contains the regulatory elements, while the intragenic region can be further subdivided into the introns and exons (Wong et al., 2000). The introns, which may also contain regulatory components, are excised and the exons are spliced together to form mRNA, after which they are translated into proteins (Wong et al., 2000) (Mignone et al., 2002). Because they are translated into proteins, they were commonly believed to be the most important components of the genome.

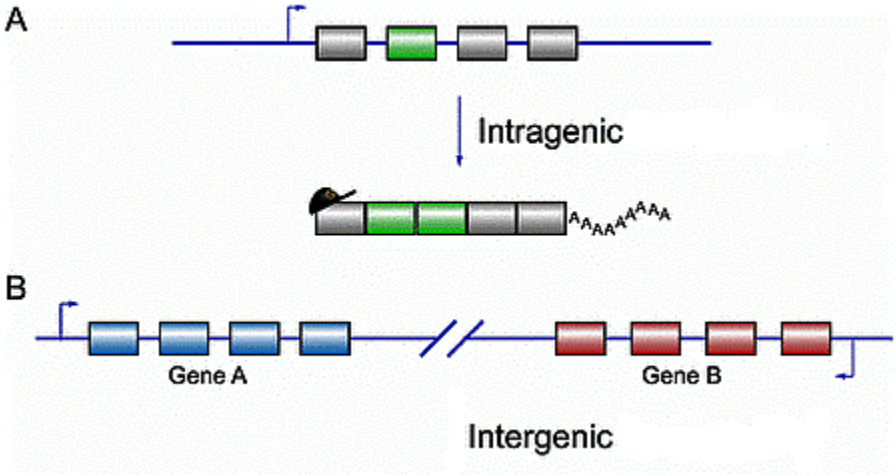


Figure 6: An illustration of the intragenic region of a gene (A) which contitutes the exons (represented as blocks) and the introns (represented as lines) and how the introns may be excised to form mRNA. (B) shows the intergenic region between Gene A and Gene B (adapted from Horiuchi and Aigaki, 2006).

1.9.2. Regulation in the Human Genome

The central dogma of molecular biology states that DNA is transcribed into RNA, which is translated into proteins and these proteins ultimately determine the organism's phenotype (Shapiro, 2009). However, as stated before, not all DNA encodes for proteins and most of the transcribed DNA remains in its final RNA state - as ribosomal RNA, transfer RNA and other small non-coding RNA (Shapiro, 2009). These RNA molecules may have catalytic roles (ribozymes), function in ribosomal RNA (rRNA) processing, in translation and in gene regulation (Shapiro, 2009). In fact, most genomic DNA functions in the regulation of gene expression and is not transcribed (Mignone et al., 2002). Gene regulation may be exerted on a transcriptional level; controlling whether a gene should be transcribed and if it is, to what extent it is transcribed. Post-transcriptionally, the fate of the transcribed RNA may also be regulated. This is achieved by controlling the stability, translation efficiency and subcellular localization of RNA molecules (Mignone et al., 2002). Generally, **transcriptional** control is mediated by RNA polymerase, transcription factors and *cis*-regulatory elements (e.g. promoters, enhancers, silencers, introns

etc.), that regulate the expression of the gene on the same DNA strand on which the regulatory element is located), which play a role in the production of pre-messenger RNAs (pre-mRNA) from DNA (Mignone et al., 2002).

1.9.2.1. Processing of pre-mRNA to mature messenger RNA (mRNA)

After production of the pre-mRNA (the initial transcript), the molecule is further processed by the removal of introns (Figure 7 (B)), and the addition of a 7-methyl-guanylate cap at the five prime end of the first exon and a poly adenine (poly A) tail at the three prime end of the last exon (Mignone et al., 2002). The functionally mature mRNA is a tripartite that consists of a five prime untranslated region (5'UTR) and a three prime untranslated region (3' UTR) on either end of a coding sequence (CDS) (Mignone et al., 2002), as shown in Figure 7 (C). Mutations in any of these regions could result in disease. However, as emphasised before, countless studies have already focused on mutations within the CDS of genes. Therefore, together with the two *cis*-regulatory elements, the promoters and the introns, this study will primarily focus on the non-coding regions of the tripartite, namely: the 5'UTR and the 3'UTR - all of which are predicted to be the major regions involved in regulating gene expression for a given gene (L.W. Barrett et al., 2013).

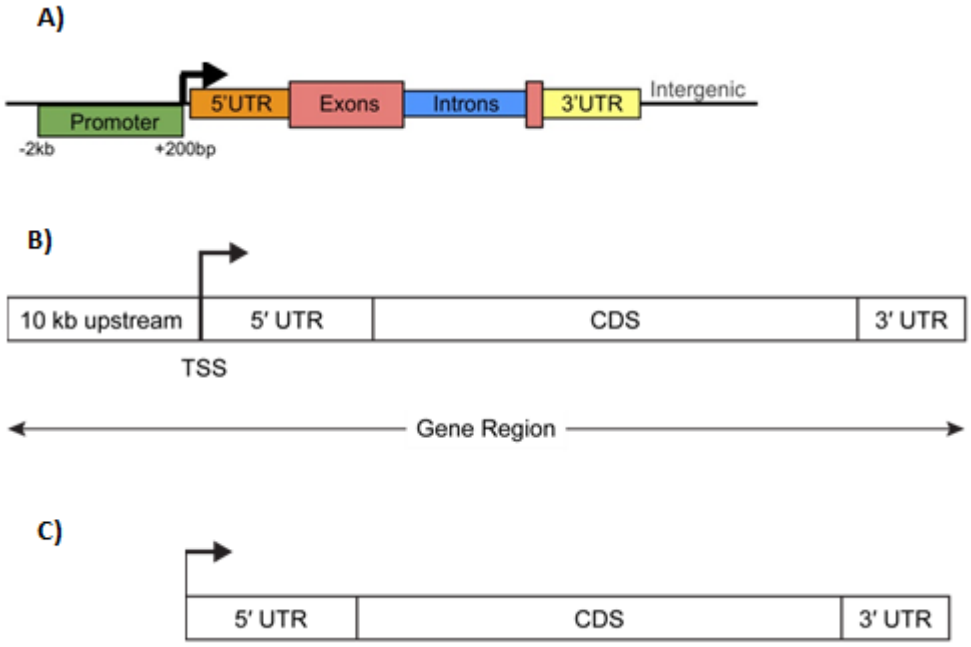


Figure 7: A) A complete gene with the promoter and introns before transcription. **B)** The gene region containing the transcription start site (TSS) and the 10kb upstream region (promoter) after the introns are removed. **C)** The mature mRNA consisting of just the 5'UTR, the 3' UTR and the CDS after splicing (Dere et al., 2011).

The subsequent subsections will discuss the non-coding regions of the gene in a little more detail and what the effects are of mutations within the region of interest, starting with the largest region, the introns.

1.9.2.2. Introns

Introns are spliced out of precursor RNA during splicing and it is well-known that this splicing is also necessary to produce a variety of proteins from a single gene copy in a process termed alternative splicing (Zhang and Edwards, 2012). Alternative splicing is demonstrated in Figure 8. Introns are also the sources of many small, non-coding RNA molecules that regulate the expression of other genes and are therefore carriers of an array of transcriptional regulation elements (Zhang and Edwards, 2012). Furthermore, they are enhancers of meiotic crossing over within the coding regions and therefore drivers of evolution. Finally, they are signals for mRNA export

from the nucleus and function in nonsense-mediated decay – a surveillance pathway that functions to reduce gene expression errors, by deleting mRNA transcripts with premature stop codons (Lucy W. Barrett et al., 2013). Most mutations that affect splicing are single nucleotide substitutions within introns or exons. These mutations may either result in the incorrect protein being produced or the introduction of a premature termination codon, which ultimately results in loss of function of the mutated allele (Faustino and Cooper, 2003).

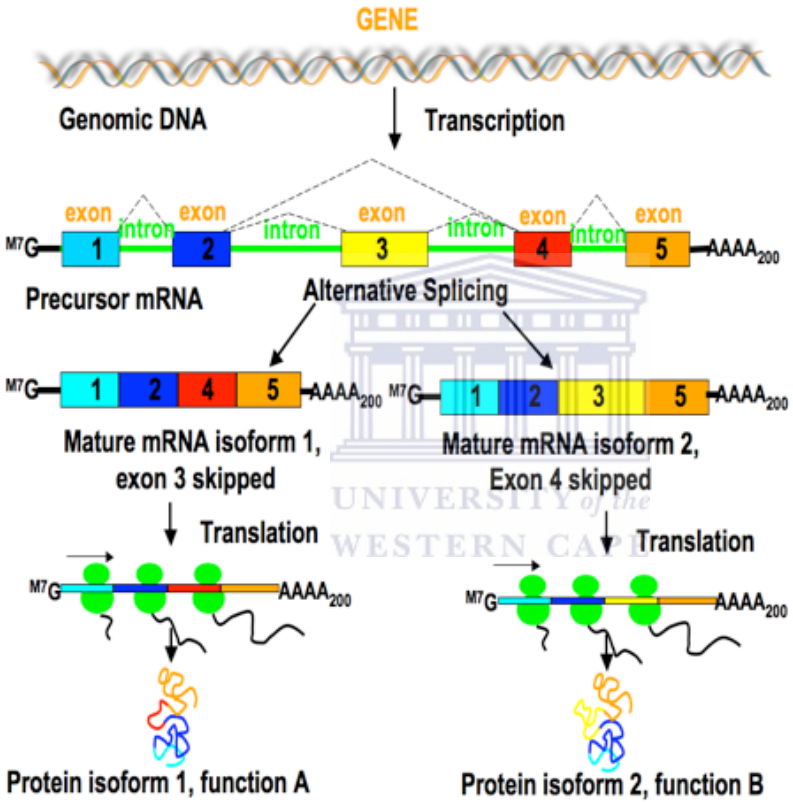


Figure 8: An illustration of how one gene can give rise to two different isoforms (splice variants), by carefully controlling which introns are spliced out. When the introns before and after exon 3 are spliced out, isoform 1 is formed. Alternatively, a different isoform (isoform 2) is created by the splicing out of exon 4. These two isoforms will give rise to two different proteins (Grigoryev, 2013).

1.9.2.3. The 3'- and 5' untranslated regions

Similarly, the untranslated regions flanking the CDS of the gene do not directly produce proteins, but they play pivotal roles in transcriptional control and in translation (Reamon-Buettner et al., 2007). The 3'UTR contains regulatory elements that are involved in the control of nuclear transport, subcellular targeting, polyadenylation signalling and mRNA degradation and translation (Reamon-Buettner et al., 2007). Mutations in this region can result in the production of non-functional proteins or downregulation of functional proteins and have been associated with breast cancer, papillary thyroid carcinoma and other diseases (Pal et al., 2001). Likewise, the 5'UTR plays a role in regulating translation by influencing RNA stability and translation efficiency (Chatterjee et al., 2001). Functional elements in this region serve to modify protein expression in response to cellular requirements. Genetic mutations, such as SNPs, that disrupt the motif in the 5'UTR have been linked to certain cancers such as multiple myeloma, oesophageal cancer (Chatterjee et al., 2001) and breast cancer (Wang et al., 2007). Nevertheless, they are also known to harbour promoters, another reason why this non-coding region is impossible to ignore.

1.9.2.4. Locations of Promoters in non-coding regions and the significance of Transcription factor binding site disruptions

As displayed in Figure 7 (A), promoters are regulatory sequences which are generally located near the 5' end of genes (Holloway et al., 2008) and act as a binding sites for transcription factors (TFs) (Cartharius et al., 2005). A fundamental step of regulatory control is the association of TFs with their DNA binding sites, also called transcription factor binding sites (TFBSs), at the onset of transcription (Holloway et al., 2008). TFBSs are short stretches of DNA (~6-20bp) located within promoters. They are recognized by the DNA binding domains of TFs, where they bind and activate or repress gene expression (Zhang et al., 2006). Since promoters harbour TFBSs, they therefore play a crucial role in transcription initiation. In some organisms, TFs may exert their control from several kilobases away from the

transcription start site (TSS), which is located upstream of the 5'UTR (See Figure 7 (B)) for the position of the TSS). However, in complex genomes, such as humans, these binding sites may be found in the 5'UTRs, the introns or the 3'UTRs (Holloway et al., 2008). Mutations occurring in these regulatory sites often result in an altered rate or control of transcription. For example, a mutation in a promoter may result in the RNA polymerase that is no longer able to bind at the corresponding promoter region (Ringo, 2004). When RNA polymerase cannot bind, transcription cannot initiate and the necessary proteins cannot be produced. Similarly, if a mutation produces a new transcription binding site, a gene may be upregulated, expressing proteins at higher levels than required (Laurila and Lähdesmäki, 2009) or they may be inappropriately suppressed, accomplishing the opposite. It is therefore not surprising that mutations in promoter regions have previously been linked to renal cell carcinoma (Hirata et al., 2003) (Havranek et al., 2005). Finally, the promoter region is particularly interesting, because they are one of the most common sites of single nucleotide polymorphisms (SNPs) compared to the rest of the genome (Li et al., 2014).

1.9.3. Single Nucleotide Polymorphisms (SNPs)

Various studies have found that functional SNPs occur in 30-60% of human promoters and tend to cluster in close proximity around the TSS (Linehan, 2012). SNPs are single-nucleotide substitutions of one base for another. They are the most common type of sequence variants, accounting for roughly one mutation per 1000 base pairs (bps). SNPs are found in both the coding and non-coding regions. They are said to hold the key to understanding disease susceptibility and progression in complex diseases involving multiple gene interplay, such as cancers (Linehan, 2012). SNPs which occur in regulatory regions (e.g. promoters, enhancers, silencers), called regulatory SNPs (rSNPs), may dysregulate allele-specific gene expression and eliminate or create a new TFBS (Buroker, 2014). Andersson et al. (2014) reported that disease-associated SNPs were overrepresented to a greater extent in regulatory regions than in exons. Linehan (2012) therefore suggested that locations of transcriptional regulatory elements may represent a major site where mutations may

contribute to disease. However, except for an activating mutation in the *TERT* promoter, no other regulatory variants have been functionally characterized as cancer drivers and this is largely due to poor exploration of non-coding sequences (Fredriksson et al., 2014) (Fu et al., 2014). Since research has shown that many regulatory elements are located within the UTRs and introns, these non-coding regions may be promising targets for the discovery of variants for further functional analysis.

1.9.4. Previous Research on Mutations in the Intragenic and Non-genic regions

In a report by Ley et al. (2008), which describes somatic mutations in acute myeloid leukaemia (AML) patients using whole genome sequencing (WGS), of the 11000 mutations detected more than 97% were in the intronic regions, while a further 1% were in the UTRs. The mutations were, however, not further analysed for their functional significance, due to difficulty in the interpretation of the generated data (Hindorff et al., 2009). When a similar study was carried out by Mardis et al. (2009), another mutation was identified in an evolutionarily conserved non-genic region. Genome-wide association studies conducted by De Gobbi et al. (2006), Easton and Eeles (2008) and Steidl et al. (2007) also reported inherited non-genic alterations in cancer genomes. Undeniably, several studies within the last decade have demonstrated the immense impact of mutations in these non-coding regions, in the context of disease. Nonetheless, cancer is a heterogeneous disease that often involves crosstalk between multiple mutated genes (Loeb et al., 2003). More importantly, since cancer cells are renegade normal cells, a systematic approach to such studies is required to distinguish which genes or gene variants perform normal metabolic functions and which truly underlie the disease phenotype.

1.9.5. Somatic Mutations and Cancer

Linehan (2012) suggested that genomic studies such as whole-genome sequencing need to be exploited in order to gain a complete understanding of the genetic basis of

kidney cancer and its pathways. Previous studies have shown that more than 1% of human genes are implicated in cancer. Of these genes, more than 90% are dysregulated by somatic mutations; 20% show germline mutations and 10% are causally implicated by both (Futreal et al., 2004). Since most cancers can therefore be traced back to somatic events, by comparing the somatic mutations in the normal and cancer genomes of the same individual, one can confidently detect the specific mutations that may be implicated in the disease as well as discern how they are related to the disease stage, metastasis and drug resistance (Strausberg and Simpson, 2010). Furthermore, unlike whole exome sequencing (WES) and targeted sequencing, WGS allows one to identify mutations in the non-coding and regulatory regions that may contribute to carcinogenesis (Huang et al., 2013).

However, beyond somatic and germline mutations, a considerable amount of interest has been shown in the contribution of epigenetic factors, especially the role of DNA methylation in disease.

1.10. Aberrant DNA Methylation in cancer

Epigenetic changes are reversible modifications that are heritable. They affect gene expression without altering the DNA sequence and have also been identified as hallmarks of cancer (Martin-Subero et al., 2009). DNA methylation is one of the most extensively studied epigenetic changes in mammals, because of its importance in cell, tissue and organismal phenotypes. Furthermore, neither genetic mutations (nucleotide changes, deletions, recombinations), nor cytogenetic abnormalities, are as common in human tumours as DNA methylation alterations (Baylin et al., 2001). DNA methylation refers to a chemical conversion of a cytosine to a 5-methyl-cytosine (m5C) residue by DNA methyltransferases (DNMTs), in regions where a cytosine is immediately followed by a guanine (CpG). These CpGs are generally globally depleted in mammals except for at short CpG-rich DNA stretches called CpG islands (CGIs), which are preferentially located at the TSS of promoters. CGIs may harbour hypermethylated promoter regions (discussed later in this section), which results in gene repression, an important method of inactivation of tumour suppressor genes, DNA repair genes and apoptotic genes in neoplastic cells (Kim and

Kim, 2014). This aberrant methylation has been shown to be ubiquitous in human cancers (Du et al., 2010), and has been implicated in both oncogenesis and cancer progression (Ibragimova et al., 2013).

In ccRCC, it has been shown that the VHL tumour suppressor gene (TSG) is inactivated due to promoter hypermethylation in ~15% of cases (Ibragimova et al., 2013). In fact, some common cancer-associated genes, such as RASSF1, are frequently hypermethylated and rarely mutated. Recently, intragenic methylation (within introns and exons) has also been linked to transcriptional and splicing events, suggesting its associated regulatory potential (Heyn et al., 2013). Hence, it has become apparent that in human cancers, heritable loss of gene function can be mediated as often by epigenetic, as by genetic abnormalities (Baylin et al., 2001).

High throughput profiling of the methylation status at CpG sites is therefore crucial for our insight into the impact of the epigenome in disease (Du et al., 2010) Microarray-based Illumina Infinium methylation assays have been introduced to further epigenomic studies due to their high throughput, good accuracy, small sample requirement and relatively low cost (http://www.illumina.com/products/-methylation_450_beadchip_kits.html). To estimate the methylation status, the assay uses a pair of methylated and unmethylated probes to measure the intensity of the methylated and unmethylated alleles at the interrogated CpG site (Du et al., 2010). The Beta value (β-value) is a widely used method to measure the level of methylation. It is a quantitative measure of DNA methylation at CpG islands or the **ratio** of the intensity of the **methylated bead** type/probe **to** the **overall locus intensity** (which is the sum of methylated and unmethylated probe intensities). The β-value is a continuous number ranging between 0 for completely unmethylated, to 1, which is completely methylated (Du et al., 2010). A β-value of 0.5 indicates a similar intensity between methylated and unmethylated probes, which means the CpG site is about half-methylated (Du et al., 2010). Higher β-values represent hypermethylation, while lower β-values represent a lower level of DNA methylation, classified as hypomethylation. Both hypermethylation and hypomethylation events have been associated with human cancers (Ehrlich, 2002).

After considering somatic mutations, regulatory information as well as aberrant

methylation events in their contribution to cancers, the allele frequency (AF) of the significant variants becomes significant.

1.11. Allele Frequency of Variants in the African Population

Genetic variants associated with diseases identified via high-throughput technologies such as whole genome sequencing are determined using case control studies (Cross et al., 2010). The individuals participating in the study are classified as affected or unaffected, but one of the challenges of translating the (putative) associated biomarkers, such as the SNP, as a causal agent or risk factor in disease is the lack of allele frequency data (Cross et al., 2010). The aim of this section was to compare the frequency of allele variants in the genomes of individuals with ccRCC to healthy individuals in the general population in order to observe if they are common or rare allelic mutations. Common gene variants are usually defined as those present at a minor allele frequency (MAF) of >5% within the general population, whereas a low-frequency is defined as being between 0.5%–5% and very rare alleles at a MAF of <0.5%. MAF is defined as the frequency at which the less abundant allele of a SNP is present in a given population. In order to identify pathological candidates, the rare alleles are generally more interesting, because variants causing disease risk are likely to be segregating at a very low frequency (The 1000 Genomes Project Consortium, 2012). For this analysis the 1000 Genomes Project dataset was used, because in their study over 2000 samples were sequenced within five super-populations, namely: East Asian, South Asian, African, European and American ancestries. These genomes are available to researchers to assist with establishing reference allele frequencies within healthy population groups.

1.12. Why Africans?

By taking this study one step further and analysing these mutations in the African population, one can further increase the depth and scope of one's findings. Because Africa is the ancestral homeland of modern humans, African populations have an appreciable amount of genetic and phenotypic diversity that is much greater than in

the rest-of-world populations. Stated differently, modern non-Africans carry only a subset of the genetic variation present in Africans (Campbell and Tishkoff, 2008). African populations possess a much larger pool of population-specific alleles and are known to have less linkage disequilibrium (LD – the non-random association between alleles at different loci) among loci, relative to non-Africans. Genomic studies involving data from these diverse ethnic groups are thus essential for understanding how genetic variations influence complex disease susceptibility (Campbell and Tishkoff, 2008). The mortality rate in Africa is high, but this may change as access to anti-retrovirals and better health-care systems are becoming more widely available. Since kidney cancer is a late-onset disease and the life-expectancy of Africans is predicted to increase, this may lead to an increase in the incidence of kidney cancers in Africans (Bor et al., 2013).

Therefore, by **functionally annotating** the non-coding data generated by whole genome sequencing and by comparing them to AF data from African ethnic groups, we may unequivocally add great insight into our current understanding of how non-coding genetic variants relate to disease susceptibility. For this reason, the data released by the ENCODE Project will be invaluable in this research project. ENCODE recently publicly released functional annotation data on both protein-coding and non-coding genes of the human genome. Their ultimate aim was to enhance the understanding of the human biology and disease within the scientific and medical communities, by aiding with their interpretation of the functions of the human genome (The ENCODE Project Consortium, 2011).

1.13. The Encyclopedia of DNA elements (ENCODE) Project

Recent analysis has indicated that at least 80% of the entire genome is biologically active and is either transcribed, binds to regulatory proteins or participates in another important biochemical activity (The ENCODE Project Consortium, 2011). Exploring the function and evolutionary origins of non-coding DNA is an important goal of contemporary genome research, considering it encompasses 98% of the human genome.

The ENCODE Project (http://genome.ucsc.edu/ENCODE/), initiated by the National

Human Genome Research Institute (NHGRI) in 2003, aimed to use multiple scientific technologies and approaches in order to functionally annotate the dynamic aspects of the human genome (See Figure 9 for an overview of some of the technologies used). During their pilot project phase, which spanned from 2003 until 2007, a variety of computational and experimental methods were exploited and compared to functionally analyse a defined 1% of the genome. By 2007 they capitalized on the technologies developed during the pilot phase to study the entire human genome. The goal was to annotate all genes (coding and non-coding), all transcripts, transcriptional regulatory regions as well as chromatin states and DNA methylation patterns (The ENCODE Project Consortium, 2011).

To achieve this, seven ENCODE Data Production Centres, encompassing 27 institutions, were established to generate multiple complementary types of genome-wide data. This data included the identification and quantitation of RNA species in whole cells and subcellular compartments, the mapping of protein-coding regions and the delineation of chromatin and DNA accessibility and structure using nucleases and chemical probes. Histone modifications and transcription factor binding sites were mapped by chromatin immunoprecipitation (ChIP) and the genomic DNA methylation was measured (these data types are discussed in the section that follows). In parallel to the large-scale projects, several smaller-scale production efforts analysed long-range chromatin interactions, localizing binding proteins on RNA, identifying transcriptional silencing elements and examining detailed promoter sequence architecture in a portion of the genome. To ensure data quality, they have an on-going emphasis on the development and application of standards to allow for the reproducibility of the data and to record the metadata of experiments. Massively parallel DNA sequence technologies have been implemented to facilitate standardized data processing, comparison and integration. Primary- and processed data as well as experimental data are collected for curation, quality review, visualization and dissemination by a Central Coordination Centre, after which the data is rapidly released to the public via a web-accessible database. RegulomDB is an example of an easily accessible web-interface that has made extensive use of the ENCODE data.

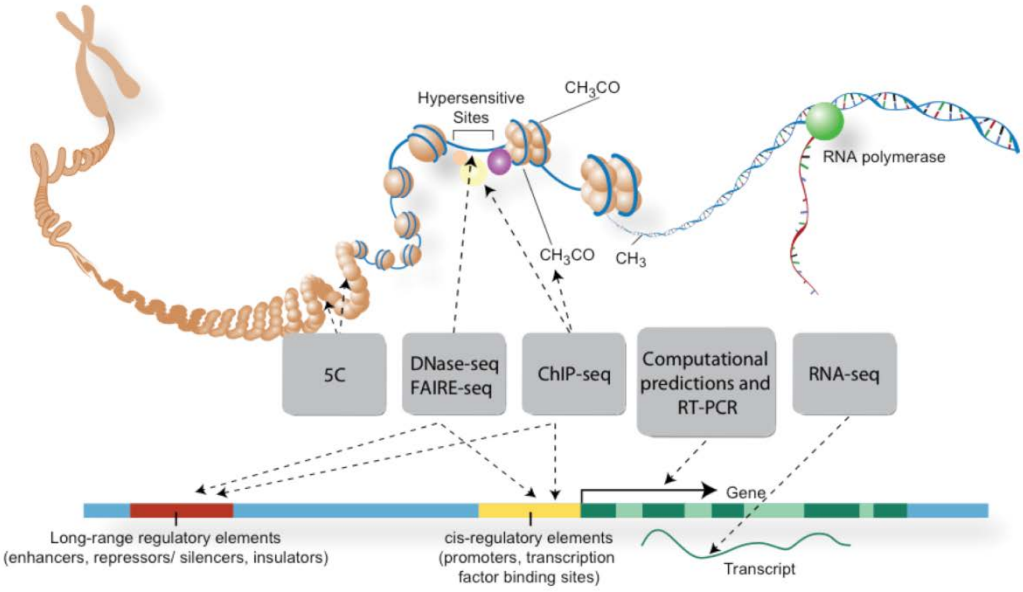


Figure 9: Various assays and methods were employed to identify functional elements in the ENCODE Project (Darryl Leja and Ian Dunham, 2011).

1.14. RegulomDB

RegulomDB is a database which functionally annotates non-coding and intergenic regions of the human genome in order to identify regions/elements with putative regulatory potential and variants that are truly functional within the human genome, by making use of the data generated by ENCODE. All ENCODE (2012 Freeze) TF ChIP sequencing (ChIP-seq), histone ChIP-seq, Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE), DNase1 hypersensitive data, DNA Methylation data and Chromatin states from the Roadmap Epigenome Consortium (unpublished) have been integrated into the database. Furthermore, TF ChIP-seq data from the National Centre for Biotechnology Information (NCBI) Sequence Read Archive, a large collection of Expression Quantitative Trait Loci (eQTLs), manually curated enhancer regions from VISTA Enhancer Browser and various computationally predicted data which are supplemented with manually curated annotations have been incorporated (Boyle et al., 2012). A brief description of the various data types and technologies used in epigenetic studies follows below (DNA methylation was already discussed in detail in section 1.10, so it is excluded here).

a) *Expression Quantitative Trait Loci (eQTLs)*

Genome-wide association studies (GWAS) have shown that most single nucleotide variants (SNVs) associated with complex multifactorial diseases lie within the non-coding portion of the genome. Due to their locations, they do not modify amino acid sequences, but evidence strongly suggests that these variants have an impact on gene expression and they have therefore been termed expression Quantitative Trait Loci (eQTLs) (Costa et al., 2013).

b) *DNAse I hypersensitive data*

DNA is usually tightly wrapped around histone proteins and packed into nucleosomes. This tight packaging effectively shields the DNA from the cleavage by DNAse I enzymes. Thus the ability of DNA to be digested by DNAse I, is indicative of nucleosome-depleted DNA and therefore suggests that the DNA must be active and presumably occupied by transcription factors. Mapping DNase hypersensitive sites (DHSs) is therefore a valuable tool for identifying active regulatory elements such as promoters and enhancers (Song and Crawford, 2010).

c) *Chromatin immunoprecipitation sequencing (ChIP-seq) and histone ChIP-seq*

The epigenome is defined as the methylated DNA and the modified histone proteins around which both methylated and unmethylated DNA is wrapped. During transitions through the developmental stages and within diseases such as cancers, the DNA methylation states and the histone modifications (such as histone acetylation and histone methylation) undergo global changes, thereby drastically altering the chromatin states. The chromatin state is defined as being either open; meaning the DNA is accessible for transcription; or closed, meaning the DNA is inaccessibly wound around their histone proteins and packaged tightly into nucleosomes. ChIP assays allow investigators to identify protein-DNA interactions *in vivo,* by covalently

crosslinking proteins such as histones and/or TFs to their genomic DNA substrates. After the isolation and fragmentation of the chromatin, the protein-DNA complexes are captured using antibodies specific to the protein bound to the DNA in the complex. ChIP followed by high-throughput sequencing (ChIP-seq) allows for the high-resolution characterization of genome-wide profiles of TFs, histone modifications, DNA methylation and nucleosome positioning (O'Geen et al., 2011).

d) Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE)

Similar to ChIP assays, FAIRE is used to isolate nucleosome-depleted DNA from human chromatin. The FAIRE procedure involves cross-linking chromatin with formaldehyde which is then sheared by sonication, before it is phenol-chloroform extracted. The DNA in the aqueous phase is recovered, fluorescently labelled and hybridized to a DNA microarray. FAIRE can therefore show the position of DNase hypersensitive sites, transcriptional start sites and active promoters in combination with techniques such as quantitative Polymerase Chain Reaction (qPCR) or FAIRE-sequencing (FAIR-seq) (Tsompana and Buck, 2014).

1.14.1. Scoring system of RegulomDB

Using the observed modifications in these data types, RegulomDB developed a heuristic scoring system based on the functional consequence of the variant. The functional consequence then gets assigned to a class ranging from Category 1 to Category 6, as illustrated in Table 1. Category 1 represents the highest level of confidence that the functional location of the variant likely results in a functional consequence, such as altered TFB or altered gene regulation. This is based on known eQTLs for genes, which as previously discussed in section 1.14a, have been shown to be associated with expression. Categories 4-6 lack evidence of the variant actually disrupting the site of binding (Boyle et al., 2012).

Table 1: A breakdown of the RegulomDB scoring system and the corresponding annotations. Category 1 represents the highest level of confidence that the variant has functional consequences. Category 3 is border line and category 4-6 means that there is insufficient evidence that the variant has functional consequences.

Score	Supporting data
1a	eQTL + TF binding + matched TF motif + matched DNase Footprint + DNase peak
1b	eQTL + TF binding + any motif + DNase Footprint + DNase peak
1c	eQTL + TF binding + matched TF motif + DNase peak
1d	eQTL + TF binding + any motif + DNase peak
1e	eQTL + TF binding + matched TF motif
1f	eQTL + TF binding / DNase peak
2a	TF binding + matched TF motif + matched DNase Footprint + DNase peak
2b	TF binding + any motif + DNase Footprint + DNase peak
2c	TF binding + matched TF motif + DNase peak
3a	TF binding + any motif + DNase peak
3b	TF binding + matched TF motif
4	TF binding + DNase peak
5	TF binding or DNase peak
6	Other

1.15. Aims and objectives

The aim of this project was to implement a Bioinformatics approach to extract the coordinates of the promoter, intronic and untranslated regions of genes known to be implicated in ccRCC, to compare these regions in paired normal-matched-tumour ccRCC genomes generated by whole genome sequencing, in order to understand how these non-coding somatic mutations as well as their regulatory mechanisms and epigenetic modifications may contribute to the ccRCC. Analysis of allele frequencies of these variants will be used to assess whether these mutations may also affect Africans with respect to ccRCC when compared to rest-of-world populations.

The objectives of this study were, to:

a) Identify a set of previously identified RCC-associated disease genes, and an equal number of non-disease genes as a control set. Comparison of mutation frequency in RCC genes and control genes will indicate whether RCC genes are mutated more frequently than non-RCC genes or whether genomic instability is genome-wide (non-specific) in ccRCC tumours.
b) Extract the promoter, CDS, 5'UTR, 3'UTR and intronic regions of these genes. The CDS region serves as a second control to assess the frequency of mutations observed in the non-coding regions compared to those in the coding region.
c) Retrieve a publicly available whole genome sequenced ccRCC somatic mutation dataset (tumour and matching normal samples).
d) Extract the somatic mutations that fall within the different genomic regions of ccRCC-associated genes and the control set (e.g. 5'UTR, introns etc.)
e) Find the functional annotation of the ccRCC tumour-specific variants using ENCODE released data.
f) Identify the transcription factor binding sites (TFBSs) in RCC genes that may have been disrupted by the somatic variants.
g) Identify aberrant methylation patterns specific to ccRCC tumours around the promoter regions of the RCC genes.
h) Relate the TFBS, methylation and somatic mutation data specific to ccRCC tumours to the gene expression levels of the genes of interest, for the ccRCC tumours compared to unaffected tissue.
i) Extract the allele frequencies (AFs) of the somatic variants identified in this study, in the African population, to assess whether their frequencies in African populations differ from in rest-of-world populations and may predispose Africans to ccRCC.

CHAPTER 2

2. METHODS

In this section, the methodology and the reasoning behind each step will be described. The first step involved the selection of RCC and normal genes and the identification of an appropriate ccRCC somatic mutation dataset on which the entire successive analyses would rest. After the extraction of the somatic mutations within genomic regions of interest, the functional annotations of the non-coding variants were retrieved from RegulomDB. The variants were also checked for their locations within TFBS and for differential methylation within the genomic regions. Furthermore, gene expression data was utilized to ascertain if these variants may have resulted in gene dysregulation. The interplay between protein-protein interactions and the connections between the genes within a network were also analysed for the genes often targeted. Finally, the allele frequencies of all ccRCC non-coding somatic variants within the normal population were investigated. The flowchart in Figure 10 shows a brief overview of the most important steps, which are discussed in more detail within the various subsections. The shortened version of the Read-Me containing the **names** of the Python scripts can be viewed in Appendix VII. The slightly longer version of the Read-Me and all the Python scripts were uploaded to https://github.com/ and can be viewed at https://github.com/tralynca/Thesis-scripts/tree/master.

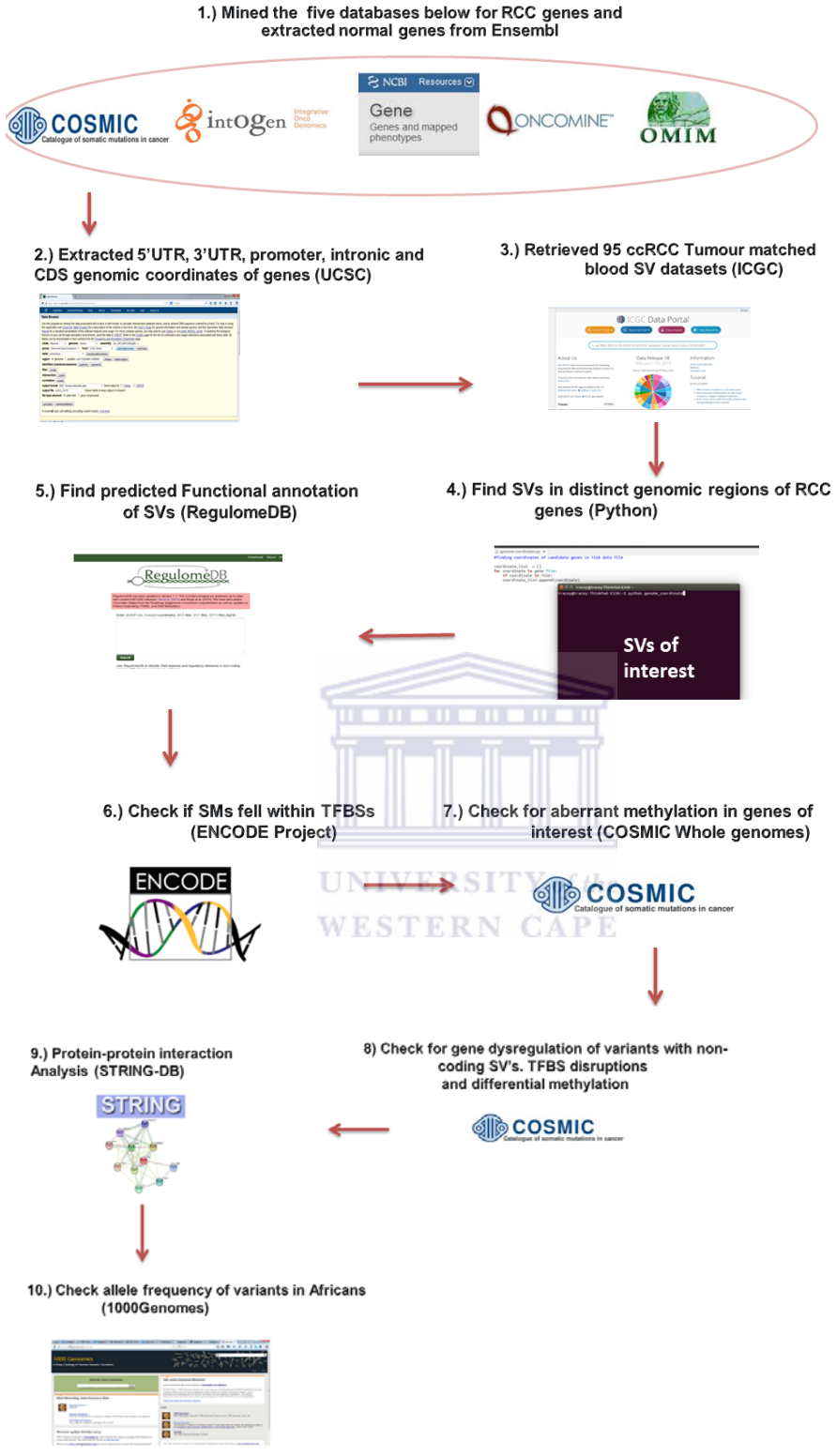


Figure 10: An overview/flowchart of the methodology used in this analysis. Intricate steps such as filtering and the selection of controls are not included, but they are discussed under the different subsections of the methods.

2.1. Mining of non-disease genes and RCC disease genes

Many RCC genes have already been identified and captured in multiple databases, ready for mining. By initiating this study with confirmed disease genes, it was easier to deduce whether the somatic variants, regulatory- and epigenetic data extracted within the genomic regions of these genes, could be disease-related. Furthermore, linking regulatory variants (which are generally located within non-coding regions) with coding genes that are well known cancer drivers may aid in shedding light on the regulatory mechanisms that govern oncogenesis (Fu et al., 2014). This was given further credibility by comparing the results of the RCC-disease genes to the non-disease genes used as the control. The list of ccRCC genes was compiled by interrogating various publicly available databases, detailed below.

2.1.1. Selection of RCC disease genes

Five databases were queried and genes that were present in two or more of the five databases were included in the list of RCC genes. The databases queried were:

2.1.1.1 The Online Mendelian Inheritance in Man (OMIM)

OMIM was queried via NCBI (http://www.ncbi.nlm.nih.gov/omim, Accessed 10/06/2015) for genes implicated in RCC. OMIM is a compendium of manually curated human genes and their genetic phenotypes, as well as extensive text summaries of all known Mendelian disorders (Hamosh et al., 2005). It also offers links to literature, sequences, maps and many other resources (Hamosh et al., 2005). Using *OMIM*, ***all*** *RCC genes were extracted.*

2.1.1.2. The Integrative Onco Genomics (IntOGen)

IntOGen 2.4.0 (http://www.intogen.org/, Accessed 13/06/2014) is a web platform that integrates the results of tumour genomes analysed with various different mutation-

calling workflows to summarize pathways, genes and somatic mutations involved in tumourigenesis (Gonzalez-Perez et al., 2013). Over 4600 somatic mutations from 31 different projects and 13 distinct tumour types were analysed (Gonzalez-Perez et al., 2013). *IntOGen didn't allow genes of cancer subgroups to be mined, so all high confidence driver kidney cancer genes were mined.*

2.1.1.3. The Catalogue of Somatic Mutations in Cancer (COSMIC)

COSMIC V70 (http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/, Accessed 13/06/2014) combines data from the Wellcome Trust Sanger Institute's Cancer Genome Project with ***manually curated*** cancer mutation data from scientific literature (Forbes et al., 2011). It is maintained by the Institute's Cancer Genome Project and it contains data for over two million point mutations and over six million non-coding mutations in over 1 million tumour samples and 12 000 cancer genomes (Forbes et al., 2015). COSMIC includes a substantial amount of data sets from The Cancer Genome Atlas (TCGA; http://cancergenome.nih.gov/) and the International Cancer Genome Consortium (ICGC; https://dcc.icgc.org/) projects. Approximately half of COSMIC's cancer genomes are curated from these consortium data portals, while the other half results from curations of published literature. COSMIC has also commited to a data release every two months to ensure updated data implicated in human cancers (Forbes et al., 2011). *From Cosmic the top 300 ccRCC-implicated genes were extracted.*

The mutation data was obtained from the Sanger Institute Catalogue Of Somatic Mutations In Cancer web site, http://www.sanger.ac.uk/cosmic. Bamford et al (2004). The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. Br J Cancer, 91,355-358.

2.1.1.4. Oncomine

Oncomine 4.4.4.4 (https://www.oncomine.org/resource/main.html, Accessed 10/06/2014) systematically compiles microarray data for analyses and *curation* before making it available via a web-based data mining platform (Rhodes et al., 2004). This includes gene annotation data from various genome resources to facilitate the interpretation of a genes possible role in cancer pathogenesis (Rhodes et al., 2004). *From Oncomine the top 10% of over- and underexpressed genes in ccRCC were retrieved.*

2.1.1.5. Entrez-Gene

The content of Entrez-Gene (http://www.ncbi.nlm.nih.gov/gene, Accessed 10/06/2014) represents the results of *curated and automated* data integration from NCBI's Reference Sequence project (RefSeq), collaborating data and other databases in NCBI. This database allowed for more refined filtering. Hence, alternatively spliced and annotated ccRCC genes were selected.

Because IntOGen and OMIM didn't allow for a refined filtering to retrieve ccRCC-specific genes, the gene list was essentially a RCC gene list and not a ccRCC gene list. All disease genes were in the Human Genome Organization Gene Nomenclature Committee (HGNC) format.

2.1.1.6. Final Selection of RCC disease genes

The gene lists were prioritized using a gene ranking procedure. Duplicate genes were eliminated from individual lists first, whereafter the genes were ranked in order of the frequency of occurrence using a command line in Ubuntu 13.10 (sort input_filenames | uniq –c| sort –nr > output_filenames). The –c counts the number of unique entries and the sort –nr records the number of times the same entry is observed across the five gene lists. The aim was to winnow down the set of genes for genes commonly

associated with RCC, in order to increase the confidence of the genes' implication in RCC. This was followed by the selection of the genes most commonly identified by the five databases. After the selection of RCC disease genes, a control set of non-disease genes was required for comparative purposes.

2.1.2. Random Selection of non-disease genes

A list of non-disease genes were randomly selected using Ensembl (http://www.ensembl.org/index.html). Ensembl is a joint project between the European Molecular Biology Library-European Bioinformatics Institute (EMBL-EBI) and the Wellcome Trust Sanger Institute to provide a publicly available web interface which stores automatically annotated data on genomes of multiple species. A Python V2.7.6 script was used in order to ***randomly*** select the equivalent number of non-disease genes as the RCC disease genes.

The genes were not matched by length, because the selection of non-disease genes was meant to be truly random. Also, matching by length could favour the selection of potentially unknown disease genes, since the protein coding- and intronic regions of disease genes have been shown to be longer in disease genes than in non-disease genes (Polymenidou et al., 2011) (Smith and Eyre-Walker, 2003). However, gene length was taken into consideration when calculating the density of the hits in the various genomic regions of both the disease and non-disease genes, as discussed in section 2.5.

2.1.2.1. Conversion of Ensembl gene ID's to transcript ID's in BioMart Ensembl

These genes (disease and non-disease) were then submitted to BioMart Ensembl Genes 79 (http://www.ensembl.org/BioMart/martview/) to retrieve the Ensembl gene- and transcript ID's, the HGNC symbols and strand orientation for all genes, since they were all required for submission to University of California at Santa Cruz's (UCSC's) Table Browser Tool and for the analysis that followed.

2.2. Extraction of genomic coordinates of genes of interest from UCSC

A common mode of impact of disease variants is through disruption of regulatory elements that modulate the target gene (Macintyre et al., 2014). As previously discussed, many of these regulatory elements are located within the non-coding regions of genes. This section aimed to identify the non-coding genomic coordinates for the RCC disease genes and the random non-disease genes. The hg19 genomic coordinates of the introns, 5'UTR, 3'UTR and 1000 bases upstream of each gene were extracted from UCSC using the Table Browser feature (https://genome.ucsc.edu/cgi-bin/hgTables). Thus, for this section and for all subsequent data sets the NCBI human reference genome build 37 (GRCh37)/hg19 was used. The 1000 bases upstream were considered since this includes the TATA box, which acts as a basal promoter element for transcription by RNA polymerase 2 (RNAP II) and RNA polymerase 3 (RNAP III) (Wang et al., 1996). This region will therefore henceforth be called the promoter region. The coordinates of the CDS were extracted, as a control to compare the results of the non-coding regions with that of the coding region. The locations of a random selection of genomic coordinates from each genomic section were manually confirmed using UCSC and Ensembl, in order to validate the scripts used by these parties in selecting these regions. Files were stored as Browser Extensible data (.bed) files for which the format is:

Chromosome start coordinate end coordinate strand

After the genomic coordinates of the genes were extracted the whole genome sequenced somatic variant data was obtained.

2.3. Extraction of somatic mutations

2.3.1. Why ICGC somatic mutation data and not COSMIC or TCGA?

Initially the TCGA data was considered for the whole genome analysis, but it was later determined to be whole exome sequenced data and therefore not appropriate for this study. Since the COSMIC database hosts updated TCGA and ICGC data for ccRCC somatic variant calls, its datasets were then retrieved. However after

obtaining more somatic mutations within the CDS region compared to the non-coding regions, a query was made with the COSMIC information team and this data was also found to be whole exome sequenced and not whole genome sequenced as anticipated. Finally the ICGC data proved to be the WGS dataset that truly contained whole genome sequence data, and was thus suitable for this study.

2.3.1.1. Extraction of Whole Genome Sequencing Data from ICGC

In order to identify the non-coding somatic variants within ccRCC patients, data from ccRCC whole genomes with their matched normal genomes were required. The ICGC simple somatic mutation dataset, current Release 18 for ccRCC tumour with matching blood control samples were obtained from the ICGC Data Repository (https://dcc.icgc.org/repository/-current/Projects/RECA-EU, Accessed 05/05/2015). The EU/FR project was selected due to the availability of whole genome sequenced data as opposed to the other ccRCC ICGC projects that performed only whole exome sequencing. The Illumina HiSeq sequencing platform was used in these studies to carry out full genome sequencing on the tumour and matching controls (blood). CASAVA version 1.7 was used as their general sequence analysis workflow, which includes multiple processing steps such as base-calling, demultiplexing, alignment and genotyping. Here Burrows-Wheeler Aligner (BWA) was used for the alignment and the variant calling was performed with Samtools mpileup. The other analysis algorithms used were Genome Analysis Toolkit, Picard, SNpEff, VCF Tools and BVA Tools. The output file available for public data and downloaded for this study was a tab separated file (simple_somatic_mutations.open.RECA-EU.tsv).

2.3.2. Checking the validity of the ccRCC genomes data

The data was checked using Linux (sort –kn –u input file > output file) to confirm that there was an equal number of donor IDs and specimen IDs, since these numbers would later be required in downstream calculations. The –k command can be used multiple times, but here it was used twice to specify the columns for the donor ID and

specimen ID, where –k is the command to signal that columns will be looked at and n is the number of the column. Finally, the –u command prints only the unique entries. Similarly grep –c GRCh37 showed that the number of times the human reference code occurred correlated with amount of entries in the file; that is, all entries contained genomic coordinates from the same human reference genome. The grep command allows one to search for a pattern specified by the user. In this case the pattern was GRCh37. Again, the –c counts the number of times the pattern occurred. The number of entries were determined using wc –l, where wc –l executes the command to count the number of lines. The same procedure was followed to check if all entries were generated via WGS (grep –c WGS).

2.3.3. Detecting somatic variants (SVs) in the ccRCC disease and non-disease genes

2.3.3.1. Locating the somatic variants (SVs) in disease genes

Non-coding somatic variants were detected by using a Python script that scanned both the bed files (containing the genomic coordinates of the genes) and the somatic variant call file. If the somatic variant fell **within** the bed range genomic coordinates, the desired data was directed to and stored in a new file.

The pseudo code for extracting the somatic mutations was:

```
for (every) line in the ccRCC somatic variant (SV) file:
        for (every) line in the .bed file (genomic range file):
                if the chromosome in SV file = = the chromosome in .bed file AND the start coordinate in SV file >= start coordinate in .bed file and  end coordinate in SV file <= last coordinate in .bed file:
                        Write the matching line to a new file
```

2.3.5.2. SVs in non-disease genes

By considering the non-disease genes the overarching question was: “If the same number of random *non-disease* genes were used instead of RCC disease genes, was an equivalent amount of SVs observed as in the RCC genes?” The same Python

script was modified to use the genomic regions of these non-disease genes and the data was stored separately as the control gene variants set.

All hits for this section and every section that followed were always first printed to Stdout (screen) for validation before the data was written to an output file.

2.4. Processing and filtering of somatic variant list

Many duplicate entries cluttered the file due to one gene having many Ensembl transcript IDs. The unique entries were sorted and retained by using a Linux command sort –t Ctrl+v+Tab -kn -u SV_results_*.txt > SV_results_*.txt_SORTED.txt

*The ***** was the genomic region, since these files were initially kept separately.*

The –t allows one to specify the character on which the data is split in the file, e.g. –t 'Ctrl+v+Tab' means it is split on a tab. Sometimes, just using the Tab button doesn't work effectively, but 'Ctrl+v+Tab' always executes the command accurately. The –k was used multiple times for the selection of all the columns except for the column containing the Ensembl transcript IDs. That is, sort and select unique entries by observing all columns, but ignore the Ensembl transcript IDs. The accuracy of the Python script to select somatic mutations was also checked by selecting random hits and checking them manually in UCSC's genome browser to verify if they were located within the appropriate targeted genomic region (e.g. if the position of the somatic variant was within the 5'UTR region of that specific gene, as indicated by the script). In so doing some discrepancies were observed. Firstly, it was already known that the somatic variant file retrieved from ICGC reported only the Ensembl IDs and **no** HGNC symbols. Therefore, genes that were not within the original disease gene list were not easily spotted based solely on their Ensembl IDs. Due to many overlapping genes in the human genome and because the strand of the genes couldn't be specified in the Python script, many of the hits were not linked to the gene IDs within the selected RCC gene list. The Ensembl and HGNC IDs obtained from BioMart Ensembl Grch37 were therefore matched with the SVs in the output

file to retrieve only those entries linked to the genes of interest. The correct gene with the correct HGNC symbol would again become imperative when doing the gene expression analysis since that dataset reports only HGNC symbols. Having the incorrect HGNC symbol would mean that the gene expression levels for a specific variant would not be extracted.

The pseudo code for the extraction of the correct somatic variants was:

```
if Ensembl gene ID in candidate gene list == Ensembl gene ID in SV hits list:
        Print the data to a new file
```

The Python script was modified to accept the data files for the non-disease genes and the same processing was applied. All the results were tabulated for direct comparison between somatic variants for the RCC disease and non-disease genes, as shown in the Results section.

2.4.1. Manual confirmation of one of the genes of interest

A second check was performed in order to confirm appropriate variant selection. One of the genes was selected at random and **all** its hits were manually checked by extracting the genomic region coordinates, which were exclusive to that specific gene, from UCSC (.bed file). All the somatic variants linked to this gene within the WGS somatic mutations file were extracted using:

`grep gene_name simple_somatic_mutation.open.RECA-EU.tsv > output_file` and stored in a separate file.

The hits for this gene generated by the Python script were also extracted using:

`grep gene_name file_with_python_generated_somatic_mutations.txt > gene_name_mutations.txt`

The genomic position of each of the gene's somatic variants generated by whole genome sequencing were checked manually to see how many of them fell within the specific ranges captured in the .bed file and the number of hits per genomic region was recorded. For example, all of gene X's WGS variants were validated against the bed ranges in the introns bed file, to ascertain whether the mutation fell within one of the introns in that file. These manually extracted hits were compared to those generated by the script to ascertain if the same number of hits was obtained per genomic region.

After the variants were extracted and validated, the density of these hits was investigated.

2.5. Density of hits

Cancer mutations are not always distributed uniformly across the genome, but instead the local density may vary by up to five-fold in distinct genomic regions (Polak et al., 2015). Furthermore, it is expected that non-coding regions will have more variation than coding regions as explained in section 1.9.4. In order to compare the density of these hits per genomic region a Python script was used. Here the number of unique hits was used, considering multiple patients reported variants at the same position causing the same somatic variant to appear more than once, and therefore skewing the results. The Linux sort –t –k –u function allowed for duplicate somatic variants to be removed if the genomic position was exactly the same despite the patient ID being different. The –k here specified the genomic region (i.e. observe the chromosome number and gene start- and end coordinate **only**) in order to select unique entries.

Pseudo code:

Add up all the bases that were scanned within the distinct genomic region

Add up the number of unique hits within the distinct genomic region

Density =

number of unique hits within the distinct genomic region /

the bases that were scanned within the distinct genomic region

2.5.1. Manual confirmation of density with the PTEN gene

Similarly, in order to confirm that the density of the variants was correctly calculated by the script, a gene was selected for manual confirmation. Its genomic regions were extracted from UCSC's Table browser and manually added. The number of hits for that gene was manually divided by the number of bases allocated to the distinct genomic region of that gene. The same python script was then adapted to perform the calculation on the files for this gene to validate that the script was doing the density calculation correctly for the actual data.

2.6. Annotation of Somatic Mutations using RegulomDB

The ultimate aim of this study was to retrieve and understand the functional consequences of the non-coding somatic variants, in order to determine whether they have a predicted detrimental effect on transcription or translation. In order to therefore assign the potential functional annotations to the non-coding variants, the epigenetic data repository of RegulomDB V1.1 was mined. To submit the somatic variants to RegulomDB, the data had to be converted to 0-based coordinates (Format = chromosome_number : minimum_coordinate maximum_coordinate), which was the format required by the web front-end of the database.

2.6.1. Conversion of the genomic coordinates for RegulomDB input

All the ccRCC somatic mutations identified in this study were single nucleotide substitutions, hence the start and end coordinates were identical. To facilitate the conversion, a mini-script was written (Python) to add one to the maximum coordinate.

This resulted in some duplicates since many SVs had the same genomic coordinates and only differed by the type of nucleotide substitution (the allele that was mutated) in different patients. The duplicates were removed with sort –u, but this meant that fewer SVs were retained than in the original SV hits files. The SVs with a Category

1-2 rating were highlighted due to the highest level of confidence displayed in the functional consequence of these variants. A python script was used to connect the RegulomDB data to the somatic mutation data which was previously removed for input to the Regulome database. A pivot table was created to observe the genes that contained the most somatic mutations (which could be due to numerous variants in the same patient) and the genes commonly targeted in multiple patients.

Pseudo code for adding one to the end coordinate:

for (every) line in file:

    Split the file into columns at every position where there is a tab

    Extract the chromosome, start and end coordinate

    if the end coordinate == the start coordinate:

        += 1 (add one)

    else: (i.e. if the mutation was an insertion or deletion and the end coordinate was already greater than the start coordinate)

    end coordinate +=0

Print the chromosome, start and end coordinate to a new file

2.7. Analysis of somatic variants

After processing, the files were analysed to check for trends, overlapping genes/variants, correlations and other relationships. Pivot tables were used to observe the regional comparisons were also made (5'UTR, 3'UTR, intronic, CDS and promoter) to observe the genomic regions where mutations or certain observations were most frequent. The Gene Ontology (GO) term names and definitions of the genes with somatic mutations were also retrieved from Ensembl BioMart (hg19) in order to observe if the molecular functions these genes participate in may be related to the hallmarks of cancer.

2.8. ICGC patient clinical data

The clinical data for the 95 patients was also made available by ICGC (https://dcc.icgc.org/api/v1/download?fn=/release_19/Projects/RECA-EU/donor.-RECA-EU.tsv.gz), which included amongst other: the ICGC donor id, the donor sex, vital status, disease status at last follow up, relapse type, age at diagnosis, relapse interval, tumour stage at diagnosis and survival time. This was retrieved in the event that further inferences could be made after the initial analyses were complete.

2.8.1. Connecting clinical information to somatic mutations

The total number of somatic mutations as well as the number of predicted deleterious mutations was connected to the patient's clinical information using a Python script so as to see if this information could be translated to the condition of the patient. After connecting the clinical and somatic mutation data, the next step was to relate the somatic variants/SNPs to transcription factor binding sites that might have been disrupted by these SVs. One potential approach for prioritizing non-coding variants for further analysis is to identify the variants located in regulatory elements/regions, because they may enhance or disrupt transcription factor binding at enhancers or promoters (Soumya, 2013).

2.9. Somatic variants in Transcription Factor Binding Sites (TFBS)

rSNPs in the **non-coding** regions have been shown to alter the DNA landscape where TFs bind, effectively altering the TFBSs (Buroker, 2014). Although RegulomDB also displays TF ChIP-seq data, the ENCODE TFBS data allowed for a more efficient data-specific extraction and observation of the TFs affected, which later became useful in the ensuing analysis. The Transcription Factor Binding Site clustered V3 data was retrieved from UCSC/ENCODE (http://genome.ucsc.edu/-ENCODE/downloads.html). Data was based on ChIP-seq experiments spanning 7 human cell types and mapped to GRCh37 (hg19). The whole genome SVs detected in

the non-coding and CDS regions of RCC disease genes were compared to the TFBSs identified by ENCODE in ccRCC genomes, to discover the total number TFBSs that may have been disrupted by these variants as well as the genes and TF's commonly involved. The TFBS ranges were larger than the SV range, hence, the script was written to find the SV positions within the TFBS ranges.

The pseudo code was as follows:

```
For each somatic mutation:
        For each TFBS coordinate:
                if first coordinate of non-coding ccRCC SV >= first coordinate of TFBS
AND last coordinate of the somatic mutation <= last coordinate of TFBS:
                        Write matching line to the output file
```

Doing this part of the analysis was essential, since SVs in TFBS could alter TF binding by either destroying/altering the binding capability of the TFBS or by creating a new binding capability, as previously discussed. No new TFBSs were, however, expected to be discovered with the data exploited in this study, since the TFBS-dataset was comprised of predetermined potential TFBSs.

After somatic mutations as well as regulatory information were examined, taking an epigenetic approach was still necessary for a more holistic analysis. Crucial modifications that contribute to cancer onset and progression may not be detected by common DNA analysis as they may affect gene expression and/or DNA methylation patterns rather than the protein function (Costa et al., 2013). Aberrant DNA methylation, particularly promoter hypermethylation, has been hypothesized to play a pivotal role in the development of ccRCC with various reports showing over 60 candidate tumour suppressor genes demonstrating evidence of tumour-specific hypermethylation. As previously stated, the VHL tumour suppressor gene is inactivated by promoter hypermethylation in approximately 15% of ccRCC cases (Ricketts et al., 2014). Hence, the genes of interest (GOIs) were also analysed for potential differential methylation patterns and their relationship with tissue-specific gene expression.

2.10. Aberrant Methylation in GOI

Methylation data (CosmicCompleteDifferentialMethylation.tsv) was extracted from the Cosmic Whole genomes database (http://cancer.sanger.ac.uk/wgs-/files?data=/files/grch38/cosmic/v73/CosmicCompleteDifferentialMethylation.tsv.gz, Accessed 06/2015) via their sftp server. COSMIC, however, obtained the data from the ICGC portal, after which it was processed by the COSMIC team. Methylation data was generated using the Infinium HumanMethylation450 bead chip. The HM450 array, targets 482 421 CpG sites throughout the genome; that is, 96% of CpG islands, with additional coverage in island shores and flanking regions. The TCGA Level 3 data was used by COSMIC, since normal samples were included which could be used to calculate differential methylation. The Beta values (M/M+U), where *M* is methylated and *U* is unmethylated, were already calculated for each interrogated locus. Probes with a SNP coordinate within 10bp of the interrogated CpG site or that had a "within 15bp from the CpG site" overlap with a REPEAT element, were masked as NA across all samples. Probes with a non-detection probability (p-value) greater than 0.05, were also masked as NA. Lastly, probes that mapped to multiple sites on hg19 were similarly annotated as NA. The differential methylation analysis was then carried out by COSMIC. The beta-values from tumour and normal populations for each locus (probe/CpG) were compared using the Mann-Whitney test. The Mann-Whitney U test is used to compare two independent groups, when the dependent variable is either ordinal or continuous and the data is not normally distributed (Laerd Statistics, 2013). The correction for multiple testing was carried out using the Bonferroni correction as follows:

the p-value of each locus (CpG) is multiplied by the total number of CpGs in the list. If the corrected p-value is still below the error rate, the locus was considered significant:

Corrected p-value = p-value * n (number of CpGs in the test) <0.05. In practice this means that a p-value < 0.0000001655 is significant.

2.10.1. Representation of the data

The methylation level was classified as High, Medium, or Low (Beta-value > 0.8; 0.2-08; < 0.2 respectively) and the methylation state (altered=Y or N). For each locus, the state was defined as 'altered' when the absolute difference between the average beta value in the normal population and tumour sample was > 0.5. The CosmicCompleteDifferentialMethylation file only displayed results for loci where the p-value < 0.0000001655 and where the methylation level was High or Low and the state was 'altered' (http://cancer.sanger.ac.u-k/wgs/analyses).

2.10.2. Filtering and analysis of methylation of data

This data, however, consisted of the combined methylation profiles of all analysed cancers. Hence, the ccRCC data was extracted from all the other cancer data using

grep clear_cell_renal_cell_carcinoma CosmicCompleteDifferentialMethylation.tsv > ccRCC_methylation.tsv.

The sort –t –kn ccRCC_methylation.tsv -u > uniq_patients.txt Linux command, (where n was the column number containing patient ID) revealed that the differential methylation data of 307 ccRCC patients was reported in this study. COSMIC also annotated the probes and reported those that were within the promoter region as "Promoter_Associated. These were then extracted using grep Promoter_Associated ccRCC_methylation.tsv > promoter_meth_ccRCC.txt. Lastly Python was used to extract all the aberrantly methylated promoter positions within the genomic regions of the 173 disease and non-disease genes (using the .bed files).

Pseudo code:

For genomic coordinate of differentially methylated promoter in ccRCC:

For range of genomic region of RCC genes in bed file:

If the methylated position >= start coordinate of the range of the genomic region of RCC genes **AND** the methylated position <= end coordinate of range of the genomic region range of RCC genes:

Print the match to output file

The genes, to which the methylated positions belonged, had to be manually retrieved from UCSC hg19, since both input files didn't report the HGNC symbols of promoter associated methylation points. A pivot table was once again drawn up to determine the genes for which the promoter regions were most often aberrantly methylated and the number of patients affected. These results were compared to the genes that were somatically mutated and specifically to the genes that potentially disrupted TFBSs in order to see if there was a relationship.

Nevertheless, another way to add credibility to the contribution of SVs, disruptions in TFBSs or aberrant methylation to the disease, is to relate these mutations to gene dysregulation in the genes of interest. Therefore, gene expression levels for the genes mutated/modified data were compared.

2.11. Gene Expression

2.11.1. Gene Expression changes of somatic variants

Changes in gene expression levels are known to be associated with cancerigenesis (Kasowski et al., 2010). However, as stated previously, whereas changes in the CDS of genes may alter the amino acid sequence, *cis*-regulatory mutations alter gene expression (Stern and Orgogozo, 2008). Therefore, linking non-coding SVs, TFBS modifications or any other forms of epigenetic changes with gene dysregulation could strongly implicate these mutations in the disease. Gene Expression data for ccRCC was initially retrieved from ICGC, to be used with the somatic variants. This would have allowed access to the same tumour specimens sequenced for somatic

variant calling. However, ICGC only sequenced the tumour samples, hence, no differential analysis (Fold change) could be calculated. Therefore, COMSIC's Whole Genomes genes expression data was downloaded (http://cancer.sanger.ac.uk/wgs-/files?data=/files/grch37/cosmic/v73/CosmicCompleteGeneExpression.tsv.gz, Accessed June 2015). COSMIC made use of the TCGA Level 3 Gene expression data generated via IlluminaHiSeq RNASeqV2, IlluminaGA RNASeqV2, IlluminaHiSeq RNASeq, and IlluminaGA RNASeq. For the RNASeq platforms, the Reads Per Kilobase of transcript per Million (RPKM) was used as a method of quantifying gene expression from RNA sequencing data by normalizing for total read length and the number of sequencing reads. The RNASeqV2 platforms contain data produced using MapSplice to do the alignment and RSEM to perform the quantitation.

The mean and sample standard deviation (STDEV) of the gene expression values were calculated from the tumour samples that are diploid for each corresponding gene, platform and study. Based on these mean and STDEV values, the standard scores for gene expression for each corresponding gene, platform, and study were calculated. In order to display if a gene is over or under expressed, a threshold of 2 STDEV, plus or minus was selected. In the cases where a sample was analysed with more than one platform for the specific study and gene, if the scores from **all** platforms were reported as above or below the threshold, then only was over or under displayed, respectively. If they didn't agree then they were not displayed. The z-core displayed across the website (an indicative score of the expression level) was taken from one platform in order of preference: IlluminaHiSeq_RNASeqV2, IlluminaGA RNASeqV2, IlluminaHiSeq RNASeq, IlluminaGA RNASeq (http://cancer.sanger-.ac.uk/wgs/analyses).

This file contained the patient ID, sample name, the HUGO gene symbol, whether the gene expression level was normal, up- or down-regulated and the z-score. However, COSMIC combined all gene expression data for the numerous tissue types sequenced by TCGA. Because TCGA followed a barcode system by which the tissue type could be identified, a python script was used to extract all data belonging to the set of unique ccRCC-specific tumour IDs recovered from https://tcga-

data.nci.nih.gov/datareports/codeTablesReport.htm?codeTable=Tissue%20Source%20Site.

Pseudo code:

```
for barcode in TCGA barcode file:
    for the barcode in the gene expression (GE) file:
        if the barcode of TCGA file in barcode of the GE file:
            print matching line to output file
```

A total of 95 unique ccRCC patient datasets were then randomly selected using Python (pseudo code not shown) in order to make sure that the patient sample size matched that of the somatic mutations dataset (n = 95) and that the same patients were considered when a comparison was made between gene expression levels in disease and non-disease genes.

Thereafter, the data for the gene expression levels of all 173 ccRCC disease and non-disease **genes** were extracted and saved in a sub file (using Python).

Pseudo code:

```
for HGNC symbol in RCC gene list:
    for HGNC symbol in GE file:
        if HGNC symbol in gene list == HGNC symbol in GE file:
            print matching line to output file
```

A comparison was made based on all somatic variants for RCC disease and non-disease genes that were initially reported before they were functionally annotated by RegulomDB. Reporting data after functional annotation would skew the results considering the non-disease genes reported fewer deleterious hits and therefore would have fewer genes for which the gene expression levels would be reported. Up-

or downregulation was deemed significant if the Fold change was above two (FC >2).

A Pivot table was again constructed in Excel for the gene expression data, in order to see which genes were often dysregulated.

2.11.2. Gene Expression for SV in TFBS

Additionally, the genes for which there were no deleterious annotations, but many possible TFBS disruptions, were also compared to see if there was a relationship with the TFBS disruptions and their GE levels. A pivot table became useful again to correlate SMs with TFBS disruptions and gene dysregulations per gene and per patient.

However, a drawback of many gene-specific studies is that physical interactions between genes or proteins are often not taken into account despite the knowledge that many diseases, such as cancers, are known to be multifactorial (Glaab et al., 2012). The top genes that frequently accumulated deleterious, non-coding variants, the genes with the most TFBS disruptions at the locations of non-coding variants and all the genes with aberrantly methylated promoter regions were therefore submitted to STRING-DB in order to see if these genes and their proteins somehow interplayed with each other.

2.12. STRING-DB protein-protein interactions

The analysis in STRING 9.0 (http://string-db.org/) allowed for common pathways via Kyoto Encyclopedia of Genes and Genomes (KEGG) to be observed in a gene-specific manner, because it was easier to observe the actual number of the GOIs enriched for a specific pathway. Here the GOIs refer to the subset of genes selected at the end of the previous section 2.11.2 for the STRING-DB analysis. STRING-DB grouped and functionally annotated the genes according to their biological process and molecular function within their protein-protein interaction network. Most

importantly, the direct interactions, co-regulations and co-expression of certain genes as well as the potential hub proteins, became more apparent. This information makes it easier to translate the genes into their relevance to the disease. All of STRING-DB's Active prediction methods were initially exploited, which include: the Neighbourhood method, gene fusion data, co-occurrence-, co-expression- and experimental data and finally, data gathered from data-mining and text mining. The highest confidence together with the Bonferroni correction method ($P < 0.05$) were used as the thresholds for prediction. The Markov Cluster Algorithm (MCL) was applied to cluster molecules that are directly associated with each other. MCL is a fast and scalable, unsupervised cluster algorithm for networks, based on stochastic flow in graphs ("MCL - a cluster algorithm for graphs," n.d.). Conversely, only the *experimental data* was used when the network was expanded to observe the surrounding molecules that interplay with the GOIs and to identify hub proteins that may elucidate the role of the genes/proteins.

After all the data was compared and variants of interest were highlighted, it was important to check the frequency of these variants. Variants that are more common in the diseased genomes compared to controls could indicate increased risk of developing the disease while the less common variants could be associated with having a more protective function. Therefore the allele frequency of the variants was analysed using 1000Genomes data (The 1000 Genomes Project Consortium, 2012).

2.13. Allele Frequency of Somatic Variants in the African Population

Whole genome sequencing files were retrieved from the ENCODE 1000genomesdatabase (ALL.wgs.phase3_shapeit2_mvncall_integrated.v5.20130502.sites.vcf.gz). This release contains an integrated set of SNPs, indels, multi-nucleotide polymorphisms (MNPs), long insertions and deletions, copy number variations and other structural variants discovered and genotyped in 2504 unrelated individuals. All variants in the 1000 Genomes vcf files are reported on the forward strand. Python was once again used to extract the allele frequency of the variants of interest. This was done for all

CDS and non-coding **ccRCC** variants as well as CDS and non-coding **non**-disease genes.

Pseudo code:

For line in somatic mutation file:
    For line in allele frequency file:
        If chromosome and genomic position as well as reference and alternate allele in the SM file == chromosome and genomic position as well as reference and alternate allele in the allele frequency file:
            Print the data from both files to a new file.

This section was responsible for answering the questions:

- *What is the allele frequency of the variants in the African population compared to other population groups?*
- *What could it mean if the allele frequency is more or less frequent?*
- *If the variant is completely absent from the 1000genomes data set, does this mean that it is a truly rare/novel and functional variant generated by genomic instability in the cancer tumour or that 1000genomes does not report sufficient data (i.e. the sequencing depth must be increased) to complete this analysis?*
- *If the AF of the variant is higher in Africans compared to other population groups, does this mean that we should be seeing more ccRCC cases in Africa, that they do not contribute to disease or that these variants have a protective effect in Africans?*

CHAPTER 3

3. RESULTS AND DISCUSSION

In this chapter the most important findings generated from the methodology are shown and discussed. Detailed findings and tables can however be found in the Appendices. In some cases, the results that were generated motivated the implementation of new methods subsections, not found in the original methods chapter. For this chapter the most important questions that were answered were:

a) Were there more somatic variants in the noncoding genomic regions compared to the coding regions; and were these results ccRCC-specific or were the variants generated as a result of general genomic instability across the cancer genomes of the 95 patients (determined by comparing the results of the disease genes to that of the non-disease genes)?

b) Could these non-coding variants be causally implicated in ccRCC based on their functional annotation by RegulomDB?

c) Could these non-coding variants in any way potentially disrupt TFBSs, and if they do, how many TFs would be affected?

d) Were there any aberrantly methylated positions within the non-coding genomic regions of these genes, since literature has shown a link between somatic mutations, TFBS disruptions and differential methylation patterns?

e) After the genes with frequent non-coding mutations and those often targeted by these epigenetic and regulatory changes are isolated, can one see interplay between these genes or their proteins within a network?

f) Could the allele frequency data help to deduce which variants increase the susceptibility and which variants have a protective effect within the African population?

3.1. Selection of RCC disease genes and random non-disease genes

3.1.1. RCC disease genes

Table 2 shows the number of genes obtained from each database, how the data was generated and whether the databases are curated. Apart from IntOGen, all of the databases contained either manually curated or a combination of manually curated and computer-automated validated data. When the filtering procedure of selecting genes present in three or more gene lists was implemented, only 27 genes remained. Hence, genes found in two or more of the five genes lists were retained for pragmatic purposes. Of the 3567 genes, 175 genes remained for the combined sense and antisense strands. After eliminating genes that didn't have a matching Ensembl ID and HGNC symbol, 173 RCC disease genes were carried forward for further downstream analysis, as the RCC-associated gene set. See Appendix I for the complete list of ccRCC genes used in this study.

3.1.2. Selection of non-disease genes

Because 173 RCC- implicated disease genes were used, 173 random non-disease genes were also chosen from the Ensembl list of non-disease genes for a fair comparison. See Appendix II for the complete list of non-disease genes used in this study.

3.2. Bed files from UCSC

The genomic positions of the non-coding and CDS regions of all genes were captured in .bed files. They contain the reference chromosome, the start coordinate, end coordinate and the strand orientation for the GOIs. The start and end coordinates of the bed files were always captured in a **range** of a few hundred to a few 1000 bases (with the start coordinate reported first, followed by the end coordinate). Therefore the coordinates of WGS SVs (which, being SNPs, were reported as a single

coordinate), were located within the bigger bed ranges when the Python script was written.

Table 2: The number of genes extracted from the various databases and how these databases sourced their data. Apart from IntOGen, all of the databases contained either manually curated or a combination of manually curated and computer-automated validated data.

Database	Sources	Data type	No of genes	Reviewed
IntOGen	Publicly available cancer genomic studies: Gene Expression Omnibus (**GEO**), **Array Express**, **COSMIC**, **Progenetix**, **Sanger Cancer Genome Project**, Cancer Genome Atlas (**TCGA**)	Genomic, transcriptomic, pathways	263	Unknown
COSMIC	Cancer Genome Project (CGP), TCGA,	Somatic mutations	224	Curated
Oncomine	DNA Microarray experiment (t-statistics), Therapeutic Target Database, GO Ontology Consortium	Gene expression	2844	Curated
Entrez-Gene (NCBI)	NCBI Reference Sequence Project (RefSeq), collaborating model organism databases, other databases in NCBI	Genes	134	curated and automated
OMIM (NCBI)	Biomedical literature	Genes and genetic phenotypes	81	Curated

3.3. Extraction of somatic mutations

3.3.1. WGS data from ICGC

A total of 53 males and 42 females (total number of participants; n = 95) between the ages 30-79 were enrolled for the ICGC WGS somatic mutation study. 95 primary site tumour biospecimens with paired blood samples were obtained from the 95 patients. The biospecimens obtained were distributed across the various stages of the disease with the majority of the samples collected from stage 1 (n = 53). Upon collection of the samples, none of the patients underwent any chemo- or radiation therapy. ICGC did not record the HGNC symbols, but used the Ensembl gene and transcript IDs in their data analyses.

3.3.2. Validity of the ccRCC genomes data

After all entries in the ICGC dataset was confirmed to be WGS and no extra tumour specimens per patient ID were found, the data was ready to be processed to extract the somatic variants (mutations within the tumour sample and not in the non-tumour sample) within the 173 disease and non-disease genes.

3.3.3. Detecting somatic variants (SVs) in the RCC disease and non-disease genes

After the somatic mutations were retrieved for the both gene sets, the output files were cluttered with many duplicates due to one gene having many Ensembl transcript IDs. Furthermore, when some hits were manually checked to confirm their genomic position, many of the hits were linked to Ensembl gene IDs that were not in the RCC gene list. As previously stated, this was due to some overlapping genes in the human genome and due to genes at the same location but on the opposite DNA strand (the ICGC somatic mutation file didn't record the strand of the gene for which their somatic mutations were captured).

3.4. Processing and filtering of somatic variant list

The *Total (before filtering)* column in Table 3, displays the total number of SVs **with** their duplicates. The unique hits were then extracted based on the Ensembl ID, genomic position, donor ID and alleles, while the transcript ID was ignored. The genes with an Ensembl gene ID that didn't occur in the gene list were also rejected and unique results were captures in See Table 3, *'After filtering'*. Table 3 also shows that when the disease and non-disease genes were compared; the promoter and 5'UTR regions contained at least twice as many variants in the RCC disease genes than in the non-disease genes (after filtering). However, the 3'UTR harboured 9x more variants, while the CDS regions had 15x more variants in the disease genes compared to the non-disease genes. Finally, the intronic region accumulated almost 22x more variants in the RCC genes than the non-disease genes, indicating that the increase in variants in the RCC genes were true disease-associated variants targeting specific genes and not just as a consequence of global cancer-induced instability across the genome.

3.4.1. Manual confirmation of the PTEN gene

The PTEN gene was selected for the manual confirmation to ascertain that the script was selecting appropriate data. This gene was chosen, because the Python script detected hits in multiple genomic regions (e.g. 3'UTR, CDS) for this gene, so the accuracy of variant detection could be reviewed using ALL the distinct genomic regions of interest. The manual tally matched that of the automated Python script for each genomic region, so the script was determined to be accurately selecting the somatic variants.

3.5. Density of hits

Before the density study could be carried out, some of the hits had to be filtered out, since, especially in the intronic regions, many patients had somatic variants at the exact same position (which are effectively duplicate positions in this regard). Only one entry of each unique position was therefore used in the calculation of the density (See Table 3, *'Unique positions'*). Thus a total of 4295 unique non-coding SVs (total not shown) and 153 CDS variants were identified. Initially when the density of the hits was compared the results were unexpected, because previous WGS studies have shown that most spontaneous mutations and even disease-associated SNPs were outside of the coding region. When this density analysis was carried out, the distribution of the hits seemed to not vary too greatly across the genomic regions. Upon further investigation it became apparent that other studies have pointed out similar results. Weinhold et al. (2014) demonstrated that there was no distinction in the regional mutation density of variants in transcribed regions: which includes the CDS, introns, 3'-UTRs, 5'-UTRs and even promoter and enhancer regions. Much of the previously reported mutation enrichment in the non-coding regions of the genome was attributed to the complete intergenic regions, which, except for the promoter regions (1000 bases), were not really considered in this study. The vast intergenic region is understood to be largely uninvolved in gene regulation and is subsequently also under weaker selective constraint, explaining the higher number of mutations in other studies (Weinhold et al., 2014).

A one-tailed, paired t-test was carried out using an R-script (V2.3.2) to test if the difference in the mutation rate between the disease and non-disease genes was statistically significant. The H_0 stated that there was no significant difference in the mutation rate of the disease and non-disease genes in the group of 95 patients. Alternatively, H_A stated that the mutation rate in the disease genes was significantly higher than the mutation rate in the non-disease genes. The results of the t-test showed that the mutation rate was 5.9x higher in disease genes than in the non-disease genes ($p = 0.040$).

A surprisingly high number of hits were generated in the CDS region, as shown in Figure 11. This could be attributed to the fact that the genes selected were disease genes and since it was a cancer dataset, the genomic instability was still expected to be higher within the tumours compared to the non-disease genomes, even within coding regions (which are generally under greater negative selection). However, as anticipated, it was still clear that in terms of sheer numbers and comparatively, the intronic region carried the highest mutational burden, with more than 90% of the total somatic variants (See Figure 11).

The intronic region was expected to accrue the most mutations due to them being much longer stretches of DNA than the other genomic regions. Furthermore, one gene usually has multiple introns, but only one of each of the other genomic regions (with the exception of the CDS which usually constitutes more than one exon) (Ivashchenko et al., 2009).

Table 3: The results for the somatic variants in the RCC disease genes (**blue**) and in the random non-disease genes (green). The **first** column of each sub- table represents the genomic regions of interest (e.g. 5'UTR). The **second** column represents the total number of variants per genomic region before any filtering was applied (duplicates variants due to splice variants creating multiple Ensemble transcript IDs and genes linked to incorrect Ensembl IDs due to overlapping genes are therefore included) The t**hird** column is a count of the total number of variants after the transcript IDs and overlapping genes not in the original gene list were removed. Column **four** displays the unique positions where the variants were located by eliminating duplicates based on donors that have somatic mutations at the exact same position .The **fifth** column displays the number bases that were scanned within each genomic region in order to find the somatic variants and column **six** shows the density of the hits (number of unique hits at unique position/number of bases scanned).

ccRCC Somatic Variants in RCC disease genes non-disease genes

	RCC disease genes						Random Non-disease Genes				
	Total (before filtering)	After filtering	Unique positions	No bases scanned	Density of hits		Total (before filtering)	After filtering	Unique positions	No bases scanned	Density of hits
Promoter	256	62	62	639 000	9.70×10^{-5}	Promoter	545	34	33	497 000	6.64×10^{-5}
5'UTR	155	43	43	242 001	1.78×10^{-4}	5'UTR	262	18	17	172 158	9.87×10^{-5}
CDS	560	154	153	1065 807	1.4×10^{-4}	CDS	306	10	9	98 452	9.14×10^{-5}
Introns	21988	4205	4115	49 257 081	8.35×10^{-4}	Introns	3699	192	192	4 666 766	4.11×10^{-5}
3'UTR	165	75	75	610 175	1.23×10^{-4}	3'UTR	123	8	8	271 819	2.94×10^{-5}

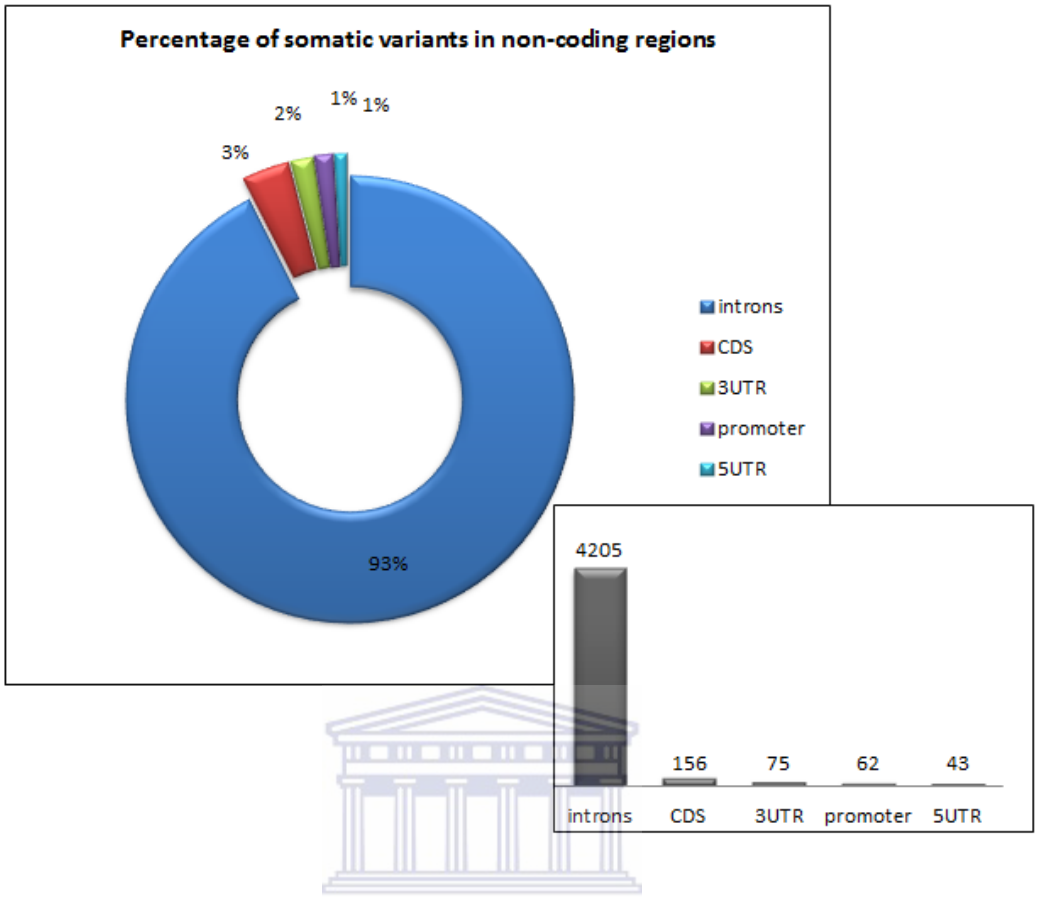


Figure 11: Most of the non-coding somatic variants were located within the intronic regions of the RCC disease genes, although a surprisingly high number of hits were also in the CDS region. This high number of mutations in the CDS region may be due to general genomic instability in the tumour genomes compared to normal, non-disease genomes.

3.5.1. Manual confirmation of density with the PTEN gene

The PTEN gene was once again used for the manual confirmation of the density of hits. The manual calculation showed that the automated calculation of the density of the hits was accurately calculated with the Python script.

After the somatic variants were acquired and filtered the primary question of interest could be answered: How many of these non-coding variants are actually predicted to have a functional effect on the disease phenotype? The variants were therefore submitted to RegulomDB for their functional annotation.

3.6. Annotation of Somatic Mutations using RegulomDB

Although RegulomDB is a database for non-coding variant annotations, the CDS variants were also submitted in order to observe how they fared compared to the non-coding regions. Similarly, the functional annotations of the variants for the non-disease genes were also queried for comparison. The hypothesis was that there would be far *fewer deleterious* variants in the *non-disease* genes than in the RCC genes since they are not implicated in disease. First all the genomic positions had to be converted (using a Python script) to comply with the submission format of RegulomDB.

3.6.1. Conversion of the genomic coordinates for RegulomDB input and RegulomDB output

As stated in section 2.6.1, after the conversion, there were once again many duplicates due to the many patients having the same somatic variant (exact same position). Furthermore, for many patients the position of the somatic mutation was the same, but the alternate allele was different. The RegulomDB web front does not take the respective alleles into account. Hence, the variant count once again reduced, but they were later recoupled to the original patient ID and allelic data. The variants were then separated based on the RegulomDB score. Variants with a score of 1-2 were deemed deleterious. A variant with a score of three was border line and was excluded because the RegulomeDB database still classifies it as having too little evidence of functional consequences. Naturally, the rest of the variants with a score higher than three were also not considered.

As shown in Table 4 below, there were no deleterious SVs within the 3'UTR, 5'UTR or CDS regions of the **non**-disease genes. The Fisher's exact test was carried out to determine whether there was a significantly higher ratio of deleterious hits for the disease genes compared to non-disease genes, per genomic region. However, the p-values for all categories were 0.8 on average and so the H_0 (there is no significant difference between the number of disease and non-disease deleterious hits) could not

be rejected. There was, therefore, no statistically significant evidence that the number of deleterious hits was higher for the disease genes than for the non-disease genes. However, the power of the Fisher test depends directly on the magnitude of the counts and the number of variants analysed were consistently below 10 for the non-disease genes, which therefore doesn't provide enough power to be confident in the results of the test.

In terms of percentages, despite the **percentage** of deleterious hits in the CDS and intronic regions of both sets of genes appearing to be very similar; this was based on a smaller number of input data (SVs) for non-disease genes and so there were only six total deleterious mutations in the non-disease genes compared to the 128 reported in the disease genes.

As hypothesized, for all genomic classifications, the **total** number of somatic variants far outweighed the predicted **deleterious** variants (between ~8x more for promoters to ~38x more for introns). Figure 12 below shows a graphical interpretation of this distinction for the ccRCC disease variants (only).

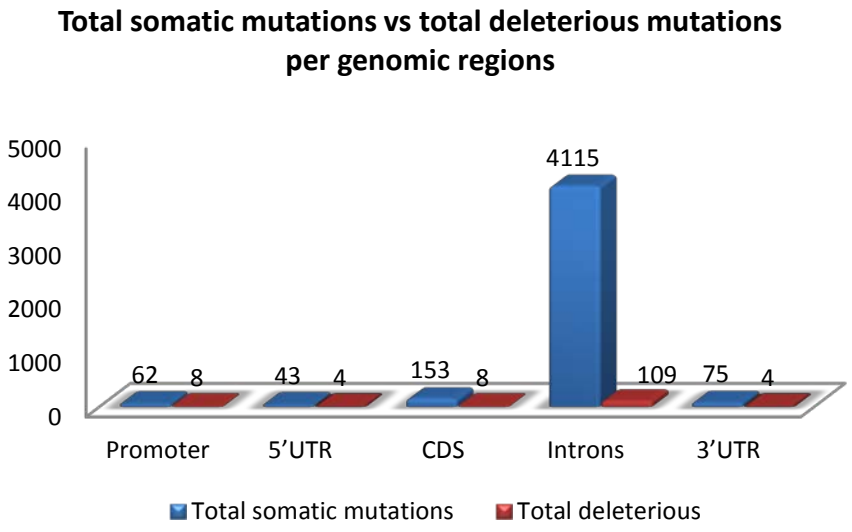


Figure 12: As expected the total number of deleterious variants (**red**) were always far fewer than the general mutations (**blue**) accumulated in the genome.

A Python script was used to link all the variants with their newly acquired RegulomDB scores back to their allelic and patient ID information. The non-coding variants were combined and a pivot table was used to further analyse the variants. However, before an in-depth analysis was carried out the “multi-hit” hypothesis was tested.

Table 4: The results for the somatic variants in the RCC disease genes (**blue**) and the random-non-disease genes (green). The first column of each sub- table represents the genomic regions of interest. The number of somatic variants for the RCC disease genes that were submitted to RegulomDB is shown in Column 2 of each sub-table and the number of variants with deleterious and borderline scores in columns 3 and 4, respectively.

RegulomDB annotation score for SVs in RCC disease genes and non-disease genes							
	RCC disease genes				Random Non-disease genes		
	Total submitted to RegulomDB	Category 2 (Likely to affect binding)	Category 3 (Less likely to bind)		Total submitted to RegulomDB	Category 2 (Likely to affect binding)	Category 3 (Less likely to bind)
Promoter	62	8 (12.9%)	2 (3.2%)	Promoter	33	1 (3.03%)	0
5'UTR	43	4 (9.3%)	5 (11.6%)	5'UTR	17	0	0
CDS	153	3 (1.9%)	3 (1.9%)	CDS	9	0	0
Introns	4115	109 (2.65%)	124 (3.01%)	Introns	192	5 (2.56%)	6 (3.13%)
3'UTR	75	4 (5.33%)	2 (2.67%)	3'UTR	8	0	0

The conversion of a normal cell to a neoplastic (malignant) cell is known to require multiple mutations. Each increasing mutation in the progeny cells confers a greater growth advantage on its subsequent offspring (Lodish et al., 2000b). To see if, generally, multiple hits may also be driving ccRCC tumourigenesis, especially in the context of non-coding mutations, the total number of hits per patient as well as the total number of deleterious hits per patient were investigated. This was of course an estimate since some patients had a far greater number of mutations than others.

a) Total variants

1.) Non-coding regions

The total number of **all** non-coding variants (i.e. *excluding* the CDS region) was divided by the total number of **patients** for which there were non-coding somatic variants (**4385/95**), as shown in Table 5 (**blue**), and an estimated **46.16** total non-coding variants per patient was calculated. As expected, the *non-disease genes* accumulated a 14 fold lower ratio, with 252 mutations in 79 patients; roughly equating to 3.19 *total* non-coding variants per patient (SV: patient ratio is illustrated in *Table 5* in **green**).

2.) CDS region

For all the variants in the CDS region, this total was 2.2 (154/70) hits per patient (Table 5 in **blue**). This was understandable since mutations in the coding regions are generally perceived to be more detrimental and so are under a strong *negative* selection, keeping the number of mutations low. The non-disease genes reported half this value with 1.1 (9/10) total CDS variants (**green**); thus, still half that of the RCC genes.

b) Deleterious variants

When the same calculation was carried out for the total *deleterious variants* in the non-coding and CDS region, the numbers were again considerably reduced.

1.) Non-coding regions

As Table 5 shows (**blue**), there was projected to be just 2.19 (125/57) deleterious hits per patient for the non-coding regions (down from 46.16) of the disease genes. Hence, there were 21 times more total non-coding variants (46.16 per patient) compared to deleterious non-coding variants (2.19 per patient) in the disease genes. Interestingly, the genomic region with the variants that were generally more functionally significant (deleterious) were located within the promoter region. Table 4 (**blue**, under heading Category 2) shows that 12.9% of these variants had a predicted functional effect, which was of course expected, since the promoter region is critical for transcription initiation. The non-disease genes had just one (1.0) deleterious hit per patient in the non-coding regions (6/6), as illustrated in Table 5 (**green**).

2.) CDS region

Within the coding regions there was also exactly one (3/3) deleterious hit per patient (Table 5 in **blue**). This means that there were at least twice as many *deleterious non-coding* mutations (2.19 per patient) compared to *deleterious CDS* mutations (1.0 per patient). As shown in Table 5 (**green**), the non-disease genes contained **no** deleterious CDS hits. As stated before, mutations in the coding sequences are generally more likely to cause disease, so serious base modifications in the CDS region are usually repaired by DNA-repair systems (Paz-Elizur et al., 2005). It was therefore not surprising that the variants in the CDS regions of the non-disease genes were not deemed deleterious. Stated differently, if the CDS mutations commonly occurred in

these genes, then they would have been highlighted as RCC disease genes in prior studies. The Fisher's exact test was once again performed, where the H_0 stated that the reduction from the number of total hits to deleterious hits for the combined non-coding regions was NOT significantly different from that of the CDS region. However, with a p-value of 0.35, the null hypothesis could not be rejected and thus, there was no statistically significant distinction between the CDS and the combined non-coding region in terms of their reduction from total mutations to deleterious mutations. Again, the observations were in some cases very low, which didn't give sufficient power to be confident in the results of the test.

Table 5: The total number of hits in the non-coding and CDS regions for the RCC genes (**blue**) and the non-disease genes (**green**) contrasted with the RegulomDB predicted deleterious variants per category are shown under the respective colours. The number of patients affected is shown in brackets. In general the total number of deleterious variants was just a fraction of the total number of somatic variants per category. The non-disease genes also generally reported far fewer variants in all categories compared to the disease genes.

RCC genes				Random non-disease genes			
Total SV		Deleterious SV		Total SV		Deleterious SV	
CDS	Non-coding	CDS	Non-coding	CDS	Non-coding	CDS	Non-coding
154 (70)	4385 (95)	3 (3)	125 (57)	10 (9)	252 (79)	0	6 (6)

In summary, there was an estimated 48 total mutations (non-coding and CDS) per patient, but just **three potential deleterious** mutations per patient, of which two predicted deleterious mutations were non-coding. Although this was just a small subset of cancer genes, these findings were substantiated with that of Tomasetti et al. (2015), who documented that only three successive mutations are required for cancer development, despite previous predictions that this number has to be much higher.

After the potentially deleterious variants were identified and the number of hits per patient were calculated, the genes associated with the most somatic variants were

isolated. This was done in order to see the types of genes that were commonly targeted, bearing in mind that the results would be biased because the subset of genes that were selected could have been biased towards a certain category.

3.7. Analysis of somatic variants

a) All non-coding variants

When all non-coding somatic mutations were taken into account, a total of 165 of the 173 RCC disease genes contained non-coding somatic mutations. Every one of the 95 patients had *multiple* somatic mutations within their non-coding regions. In contrast, only 40 of the 173 ***non***-disease genes harboured random mutations in 79 of the patients. The genes such as VHL, MET, PBRM1 and BAP1 that are observed in literature as those frequently accumulating variants implicated in ccRCC were notably NOT highlighted in the top genes with the most total non-coding somatic variants category, as seen in Figure 14 (**green** bar graph). In fact the PTPRD gene on chromosome 9 was the most targeted gene, with 473 non-coding variants in 92% of patients (87/95). Chromosome 9 was also the most frequently targeted chromosome in the same number of patients (87/95), while chromosome 3 accumulated the second most mutations, accruing 419 variants in 89% of patients (85/95). However, only five genes harboured mutations on chromosome 9, compared to 15 genes on chromosome 3. Interestingly, although chromosome 3 often comes up in ccRCC somatic mutations studies, a different set of genes, such as FHIT and MECOM carried the mutational burden as opposed to VHL and BAP1 which were minimally impacted (results not shown.)

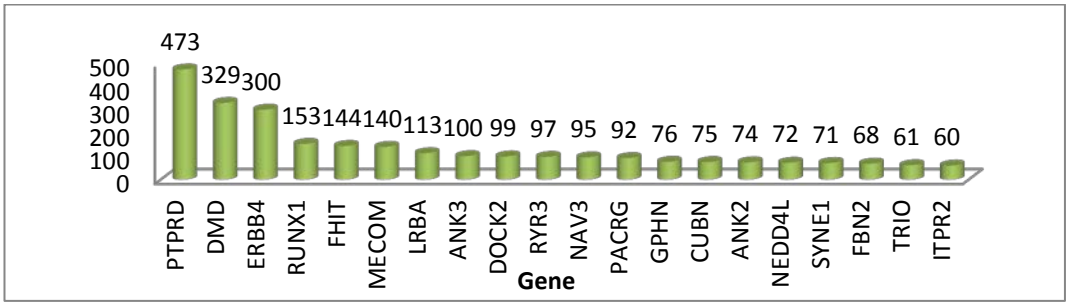
b) Deleterious non-coding variants

When this was contrasted with the deleterious non-coding mutations, 60 genes harboured deleterious non-coding mutations within patients. The total number of non-coding mutations versus the total number of deleterious mutations was compared

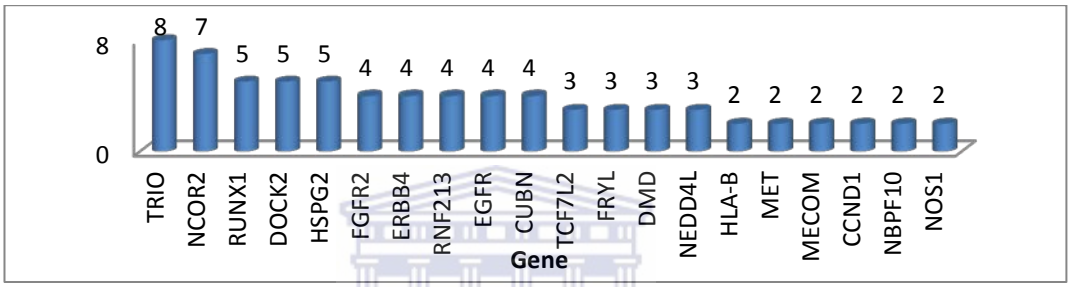
per gene and their statistical significance was tested using the chi-square method. The p-values per gene can be viewed in Appendix III.

In terms of the genes implicated, one of the genes (MET) commonly observed in ccRCC studies, did rank in the top 20 as seen in Figure 14 (**blue** bar graph).

A) The total number of non-coding mutations (ccRCC-specific)



B) The number of deleterious, non-coding mutations (ccRCC-specific)



C) Total number of coding sequence (CDS) mutations (ccRCC-specific)

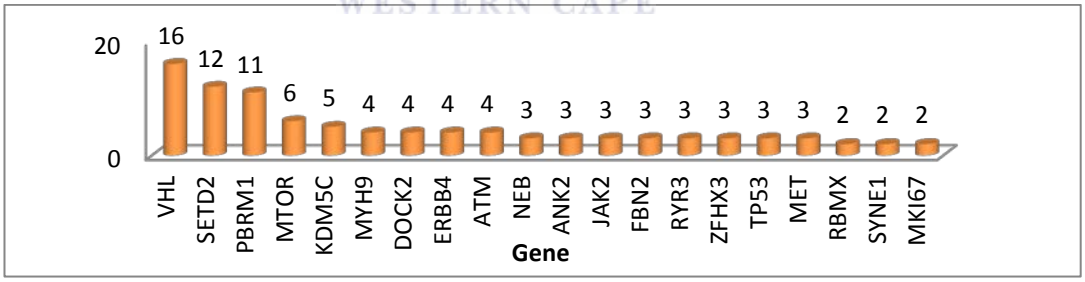


Figure 13: A) The top 20 genes with regards to the most **total** non-coding somatic variants across the 95 patient tumours. The genes with the most variants were not those genes commonly implicated as being the most frequently mutated in exome-related ccRCC studies. **B)** The top 20 genes with regards to the most **deleterious**, non-coding somatic variants across the 95 patient tumours. Except for the MET gene, these genes are still not the most commonly mutated genes in ccRCC exome-related studies. **C)** The top 20 genes, (especially the top five genes VHL, SETD2, PBRM1, MTOR and KDM5C) with the most variants in the **CDS** region were the genes commonly highlighted as the most frequently mutated in ccRCC exome-related studies.

a) CDS variants

Similarly, the coding region accumulated mutations in only 68 of its 173 genes within 70 patients. Hence, 25 of the 95 patients enrolled in this study, had **no** mutations in the coding sequences of any of these selected RCC genes in their ccRCC tumours. Moreover, although eight out of the 173 GOIs didn't acquire any non-coding somatic mutations, there were also no new CDS mutations in these eight genes that could implicate these genes in ccRCC. When the commonly mutated genes within the CDS regions were observed, an almost completely different gene set was showcased as displayed in Figure 14 (**orange** bar graph). Now the top five genes: VHL, SETD2, PBRM1, MTOR and KDM5C were all the genes commonly mutated in patients in previous exome-related ccRCC studies. In fact, this was almost a replica of the genes reported by (Hakimi et al., 2013), shown in Figure 13. It should be noted that this was an exploratory analysis in which gene length was not yet taken into account. This factor was however considered in subsection 3.7.1.

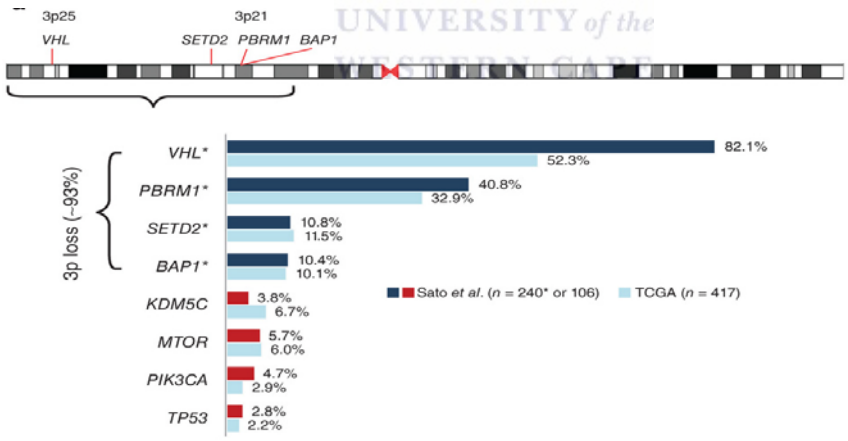


Figure 14: The top genes most commonly affected by the loss of the a piece of chromosome 3p in ccRCC (Hakimi et al., 2013). VHL, SETD2, PBRM1, MTOR and KDM5C also accrued the most somatic variants within the coding sequences of the patient tumours used in this study.

These initial results do, however, add some credibility to the accuracy of this study in terms of its agreement with previous studies. However, this also demonstrates that by

continuing to do exome sequencing studies one may limit the potential discovery of ccRCC gene variants that may play a crucial role in the development and/or progression of the disease, due to their locations outside the coding sequence. By focusing on the CDS region, the functional significance of other genes associated in ccRCC may have been overlooked. More emphatically, this suggests that perhaps it is not mutations in the non-coding regions of *already identified* ccRCC-implicated genes that eventually prime the gene for additional mutations in the CDS region of that same gene, in order to establish the gene's selective growth advantage; but rather, that there may be an entirely new set of undiscovered ccRCC genes that may be implicated in ccRCC through their frequency of harbouring non-coding DNA mutations. This was demonstrated by the 25 patients with **no CDS** variants and the 165 genes which accrued non-coding mutations compared to the mere 68 genes with CDS mutations.

Consequently, in this study the non-coding variants were of greater interest, but as stated previously, it was also because the coding portion of the genome has been exhaustively queried and analysed in various studies. Special attention was given to the variants that were predicted to be deleterious by RegulomDB, although frequent comparisons were made to other mutation categories where necessary. A total of 125 deleterious non-coding mutations were identified, which affected 60 genes in 57 patients. It was obvious that some genes were frequently targeted in many patients, such as with the RUNX1 gene, demonstrated below in Table 6 (**blue**).

In some cases the same patient had multiple mutations in a distinct gene such as with TRIO, FRYL and AKT1 (Table 6 in **green**). However, while **both** mutations occurred within the *intronic* region of the TRIO gene in patients DO47150 and DO46905, the two mutations within the FRYL and AKT1 genes were located within the promoter and intronic regions of the same genes. This suggested that unlike the TRIO gene, where mutations were in close proximity to each other, there was a wider distribution of genomic instability around the FRYL and AKT1 genes.

A hotspot analysis was therefore carried out on all CDS, total non-coding and non-coding, deleterious variants. The aim of the analysis was to evaluate if mutations were confined to shorter stretches of the gene, as opposed to being evenly distributed

throughout the gene. It was determined that a few genes which incurred non-coding mutations (i.e. total non-coding), had recurrent mutations in several patients at the exact same position. For example, five patients had a mutation at the same position in the genes: ADCY1, ANK3, CUBN and VWF, while two patients acquired variants at the same location in the VHL gene.

For the genes with deleterious non-coding and CDS mutations, no mutations were identified at identical positions. For all three categories (total non-coding, deleterious non-coding and CDS) several variants were within close proximity to each other (within 1kb), however, most of the variants were broadly distributed across numerous genomic positions within that specific gene, even in the genes that were highly enriched for mutations. This demonstrates again the general genomic instability hypothesis within certain genes. Still, an additional analysis was carried out for a direct comparison between gene length and total mutations.

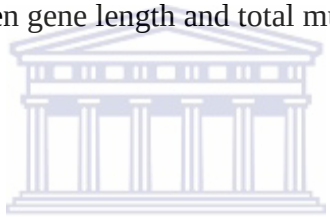


Table 6: The donor ID is bolded and always starts with '**DO**' followed by a five-character number. Some patients had multiple mutations in the same genes, such as in the RUNX1 gene (**blue**). In some cases, such as in the TRIO gene (**green**), some patients had more than one mutation in the same gene and within the same genomic region. In other patients, despite having more than one mutation in the same gene, the mutations actually occurred within different genomic regions such as with the FRYL and AKT1 genes (**green**).

Donor ID	3UTR	5UTR	Introns	Promoter
DO47150			**5**	
TRIO			2	
SYNE1			1	
HSPG2			1	
NOS1			1	
DO47100			**5**	
DO47092			**4**	**1**
FRYL			1	1
RUNX1			1	
CUBN			1	
DOCK2			1	
DO46905			**4**	**1**
AKT1			1	1
NCOR2			1	
RUNX1			1	
ERBB4			1	

3.7.1. Gene length versus number of mutations

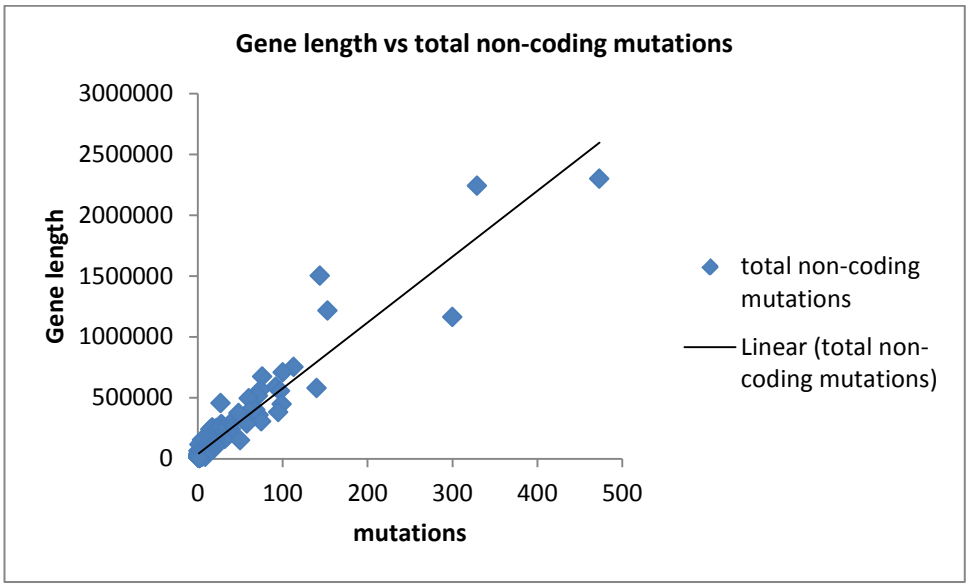
In order to see if the genes such as RUNX1, AKT1, TRIO and FRYL just naturally incurred more mutations due to being longer that the other genes, a scatter plot was drawn up. The genes' start and end coordinates were extracted from Ensembl BioMart (hg19) and the gene lengths were calculated using the built-in Excel formula: =*sum (end – start*). Excel's Chart feature was used create a scatter plot of the gene lengths versus the total non-coding mutations and the gene length versus the deleterious non-coding mutations.

As shown in Figure 15 (A), the hypothesis was mostly true in that the genes with shorter lengths generally had fewer mutations than the longer genes. However, some

exceptions were apparent. Some of the median length genes incurred a higher number of total non-coding mutations. Likewise, several longer genes had the median number of mutations. The genes highlighted in the previous section as having mutations at the same location in multiple patients and the genes with multiple mutations occurring in the same gene belonging to just one patient, were then individually scrutinized. While the RUNX1, ADCY1, ANK3, CUBN, VWF, TRIO and FRYL genes were distributed among the top half of the genes with longest lengths, the AKT1 and VHL genes were two of the smallest genes in the gene lists, substantiating again that these mutations are not just arbitrary.

Table 7 illustrates that when the GO annotations of these genes were retrieved from Ensembl BioMart (hg19), all of the genes participated in cancer hallmark pathways and molecular activities. However, besides the long RUNX1 gene which functions in the most cancer trademark events, the two smallest genes, AKT1 and VHL are involved in more cancer-related metabolic activities than the other longer genes, which could explain why they would be targeted in cancers. When the same number of genes (n = 9) with just one mutation was randomly selected (Hence, a mutation in just one genomic region and in just one patient), three of the genes were functionally annotated to play a role in three cancer-hallmark events, but the majority of genes (67%) participate in just one or no cancer-related activities (not shown).

A)



B)

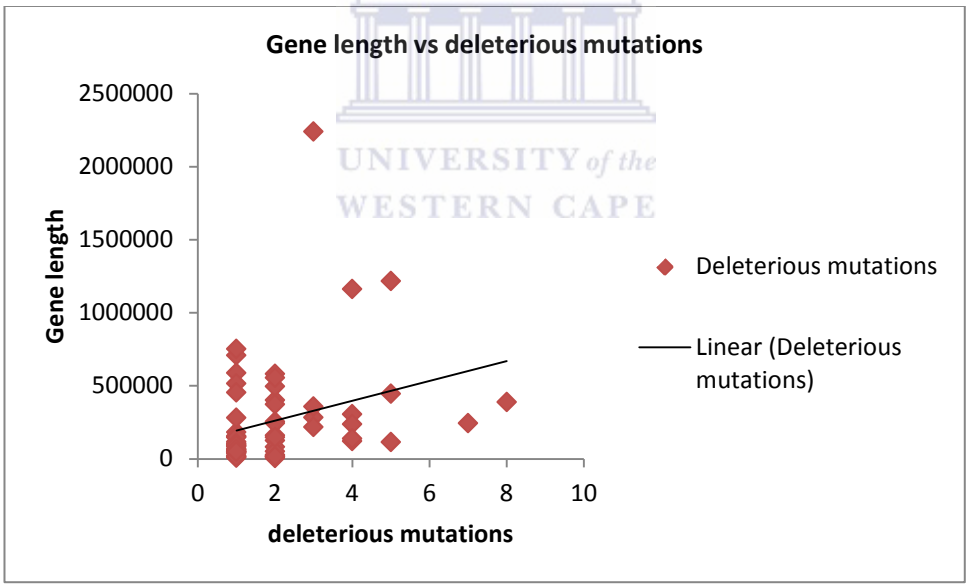


Figure 15: A) The gene lengths and the total non-coding mutations per gene. The longest genes generally incurred the most non-coding mutations. **B)** The 60 genes with their deleterious variants. Contrary to total non-coding variant s, with the deleterious variants, the smaller genes generally accumulated the most mutations.

Table 7: The cancer-related activities in which the genes mutated at the exact *same position* in several patients (**blue**) and the genes mutated at *multiple genomic regions* in the same patient (**green**) function in. The two smallest genes, VHL and AKT1 function in more cancer-hallmark events than most of the much longer genes.

Gene	Signalling	Adhesion	Proliferation	Transcription regulation	Differentiation	Apoptosis	Angiogenesis
ADCY1	√			√			
ANK3	√	√			√		
CUBN		√					
VWF		√					
VHL			√	√	√	√	
FRYL				√			
RUNX1	√		√	√	√		√
TRIO	√					√	
AKT1	√		√		√	√	

Similarly, when this analysis was repeated with the eight genes from the original list of 173 RCC genes (shown in Appendix IV), which had no somatic mutations in any of the patients, the lengths of these genes also ranged from long to short and except for the RAF1 gene (which is linked to three of the selected cancer-related events), most of the genes (75%) partake in just one or no cancer-hallmark activities (Shown in Appendix V). This suggests that it is very likely that it is not specific lengths or types of genes that are targeted in cancer but any pathway that may give the cell a growth advantage when perturbed.

Interestingly, when the gene lengths versus non-coding, **deleterious** mutations were considered, there was **no** definite relationship between the two, as shown in Figure 15 (B). Therefore, although the **mostly** linear relationship between gene length and **total** non-coding mutations suggests that the mutations were most probably randomly acquired across the length of the gene, the longer genes with a moderate number of total mutations as well as the relationship between gene length and deleterious non-coding mutations yet again suggests that the higher instability *around certain genes* cannot just be attributed to chance. In fact Martincorena et al.(2012) carried out a phylogenetics and population genetics study to show that the mutation rate has been evolutionary optimized; such that the rate of mutation is generally lower in more highly expressed genes and in those under stronger negative selection. Albeit, the Martincorena *et al.* observations were carried out on *Escherichia coli* (*E. coli*).

Hence, even the general instability seems to be controlled or targeted to some degree. In the next subsection the categories of mutations (categories = total non-coding, deleterious non-coding and CDS mutations) were observed within individual patients to ascertain if there were any apparent trends, i.e., if any category of mutation was absolutely necessary for the disease phenotype to be expressed.

3.7.2. Total non-coding, deleterious and CDS mutations per patient

The total number of non-coding variants, CDS variants and non-coding deleterious variants were compared using Excel's Vlookup feature, to observe the number of

mutations per mutation category (categories: all non-coding, all CDS, deleterious non-coding). When the data was analysed on a patient-by-patient basis, as stated previously, all 95 patients had *multiple total non-coding* somatic mutations.

There wasn't much of a generalized trend with regard to the number of deleterious mutations required; in the sense that one patient would have multiple deleterious mutations in multiple genes, while others reported just one or two deleterious mutations in the just one gene. In fact, nearly half of the patients (27/57) with deleterious non-coding mutations had *only one deleterious* mutation (not shown). Albeit, for **most** (but not all) patients, there was either a deleterious non-coding variant and/or a CDS variant as displayed in Table 8. However, 14 patients, accrued no CDS or deleterious non-coding variants (only **tolerated** non-coding mutations) within these RCC-implicated genes and yet these patients expressed the disease phenotype (One example is highlighted in Table 8 in **green**).

With regards to the genes, as discussed in section 3.7a, 105 of the 173 RCC genes, (i.e. 60%) accrued **no** somatic mutations within their CDS region. Again, had this been a whole exome sequenced data set, where the total coding sequence mutations are considered, these genes would not be discovered nor classified as significant. Obviously this could imply that additional criteria may be required over and above that used by RegulomDB to classify variants as deleterious, since these patients did develop ccRCC. However, since this is just a subset of RCC genes, it could also suggest that the genes that are disease-associated in these patients were not selected for this particular study. This also demonstrates the complexity of polygenic and multifactorial diseases such as cancers, since no two patients may have the same distinct profile or markers signifying the disease.

To see if the clinical data could add richness to the analysis, the total number of hits as well as the total number of deleterious hits were associated with each donor's clinical information using a Python script.

Table 8: (Subset of the original table shown in Appendix IV) If there were no mutations in that mutation category, it is shown #N/A. Although all patients had multiple non-coding somatic mutations (column 2), some patients had either deleterious non-coding mutations **or** CDS mutations, but not both. For fourteen patients, there were no deleterious non-coding or CDS mutations (example shown in **green**).

Patient ID	ALL non-coding	ALL CDS mutations	Deleterious mutations ONLY
DO46877	221	7	10
DO46933	66	2	3
DO47159	66	6	#N/A
DO46827	65	#N/A	2
DO47174	59	5	3
DO46873	57	2	#N/A
DO47136	54	1	1
DO47012	52	5	#N/A
DO46957	50	#N/A	#N/A

3.8. ICGC patient clinical data

Only 29 of the 95 donors contained complete clinical data, but the data was nevertheless coupled, to see if a story could unfold from this small sample size.

3.8.1. Connecting clinical information to somatic mutations

No apparent trend could be observed in the donors who were reported to be deceased, nor was there any indication as to why some patients have been able to survive the disease, despite an exorbitantly high number of mutations. In fact, some **deceased** patients had neither deleterious non-coding mutations nor CDS mutations. One

patient had no CDS mutations and only one deleterious non-coding mutation (although he had 83 tolerated, non-coding mutations); he was only 46 years old and was only in stage I of ccRCC and yet succumbed to the disease. None of the deceased patients had at least one gene mutated in all patients and except for in the introns, none of them had variants within the same genomic position that could signal if the distinct genomic region of the variant reflects on the severity of the phenotype. Hence, both the somatic mutation data and the clinical information provided no sufficient basis for why one patient would develop cancer or have a worse prognosis, despite having only one deleterious mutation, while another required multiple mutations. Of course it has been shown that there is generally genomic instability within cancer genomes, so the mutations could have accumulated **after** the patient's sample was taken at the initial visitation, or there may have been other environmental factors that came into play to accelerate the disease. Furthermore, as stated before, the additional genes that contributed to this particular donor's disease may not have been considered in this study. At this stage the conventional approach would be to look at the functional annotations of all the genes involved, but since regulatory data was made available by the ENCODE Project, it seemed more fitting to first observe how these mutations related to transcription factor binding site (TFBS) disruptions; it could be that this provided the clues for the functional significance of the mutations, especially for those patients with no deleterious or CDS mutations. Also, unlike CDS SNPs or variants that may cause disease by altering the amino acid sequence, rSNPs located in non-coding regions, are more likely to have an effect on the **transcription** of their neighbouring genes; usually by affecting the binding affinity of TFs or by altering promoter methylation (Li et al., 2014).

3.9. Somatic variants in Transcription Factor Binding Sites (TFBS)

In a GWAS study in 2010, Kasowski et al. showed that TF binding variations were frequently due to SNPs and that these mutations often resulted in gene dysregulation, suggesting a functional effect for these variants. More recently, disruptions in these TFB motifs due to genetic variants, have also been shown to be an underlying

mechanism of *chromatin variation* within humans (Kasowski et al., 2013). A discussed in section 1.14c) , the state of the chromatin **impacts** the accessibility of DNA and therefore the **transcription** of the gene. The rs2125230 rSNP in the AKT3 gene, for example, has been implicated in aggressive prostate cancer by **modifying** the IRF1 **TFBS**, which possibly inhibits the antiviral agents used to combat the tumour (Buroker, 2014).

For this section of the study the location of SVs in TFBSs were therefore investigated. The data was reported separately for the coding (CDS) and non-coding mutations within TFBS of the ***RCC disease genes***. The script was then modified to generate the TFBS data for the variants reported within **the *non-disease genes***. After the distinction was made between disease and non-disease genes, an evaluation was made based on the difference between TFBS disruptions for the **deleterious** and **tolerated** non-coding mutations (as categorized by RegulomDB) within the RCC disease genes. The motivation behind the search was that the *function* of the gene within which the somatic mutation was reported may not be of primary significance, but that the exact genomic position within the gene of interest played an even greater role due to its effect on multiple TFBSs.

For the *non*-disease genes, out of the total 40/173 genes that accumulated somatic variants, only 14 genes had possible TFBSs modifications. On the contrary, **most** non-coding somatic variants (132/165), located within the RCC disease genes, were located within TFBSs; meaning they could possibly alter TF binding affinity. There were over 3000 potential TFBS disruptions according to the positions of these non-coding variants and more than half of the patients had variants at positions in the genome where ten or more TFs bind. The 3000 disruptions do not necessarily mean that they reflect 3000 distinct TFBSs or 3000 distinct TFs. Many genes are regulated by the same TFs and therefore provide binding sites for these TFs at a specific motif (DNA sequence) in that gene. These motifs do not need to be an exact complementary match for the TF to bind, but its sequence could vary slightly and the TF would still recognize it as a binding site (i.e. they are degenerate). Numerous patients therefore acquired variants at these TFBSs, which affected the same TFs. Because of these degenerate motifs, one motif/binding site within the same gene was

at times the binding site of multiple distinct TFs. Therefore, it was not surprising that the number of total somatic mutations within a patient mostly fell far below the number of potentially disrupted TFBSs. However, two general observations could be made at this point in the study. Contrary to the non-disease genes, the RCC disease genes frequently accumulated somatic mutations even if these mutations were not considered deleterious by the criteria used by RegulomDB and even if the somatic mutations did not occur within the CDS region. In addition, the ccRCC variants were often located within TFBSs and more specifically, often at TFBS locations that could affect the binding of multiple TFs.

a) ***Number of total non-coding somatic mutations vs number of TFBS disruptions***

When **all** non-coding disease variants (deleterious and tolerated) and their associated TFBS disruptions were considered, there were no direct relationships between the genes that accumulated the most somatic variants and the genes with the most TFBS disruptions as shown in Figure 16. As stated previously, this was not really surprising since the position of the variant within the gene determined the number of TFBSs that could be altered. One gene could have multiple mutations, but none of them could be located within TFBS positions, whereas another gene could have just one variant, but that position may fall within the binding site of a myriad of TFs. Here, in the case of **all** non-coding variants, the *former* scenario was evident since there were generally more non-coding somatic variants than potential TFBS disruptions.

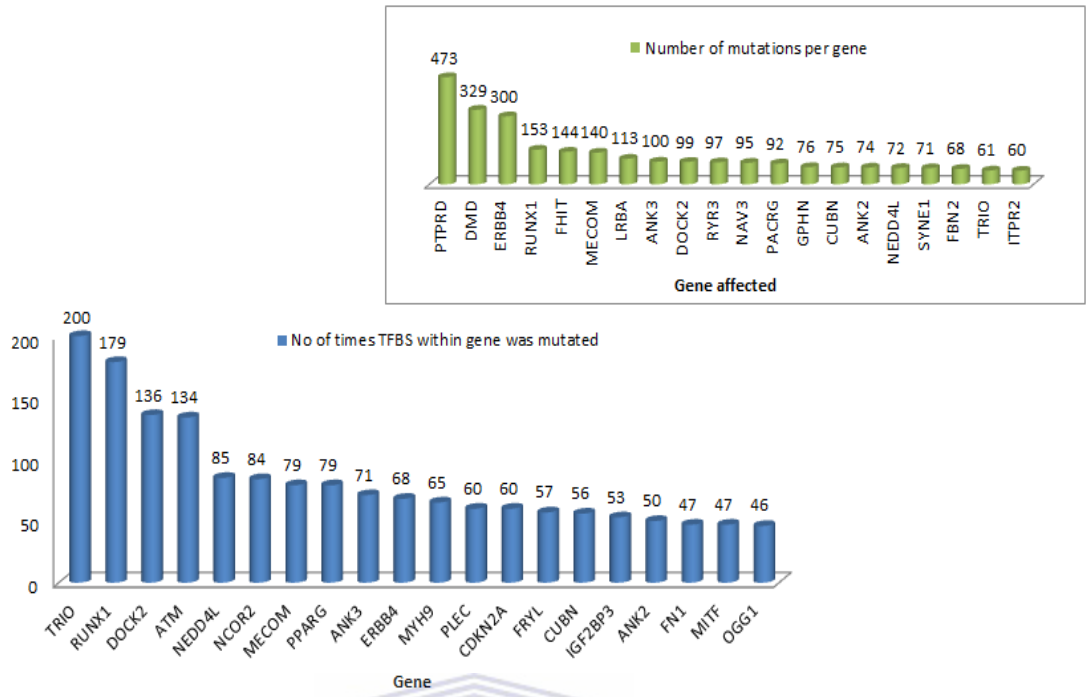


Figure 16: The bar graph in **green** on top displays the genes which had the most **non-coding** somatic mutations (**deleterious + tolerated**) while the bar graph at the bottom (in **blue**), shows the genes which had the most TFBS disruptions based on the position of the variant. When genes with the most non-coding somatic mutations were contrasted with the genes that harboured the most TFBS disruptions based on the position of the variants, these genes didn't overlap. Many of the *total* non-coding mutations (deleterious + tolerated) therefore did not fall within multiple TFBSs.

b) ***Number of deleterious non-coding somatic mutations vs number of TFBS disruptions***

This was however very different when the RegulomDB scored deleterious variants were considered. As displayed in Figure 17 , there was a much better correlation in terms of the top 20 genes with the most deleterious somatic mutations (as scored by RegulomDB) and the top 20 genes with the most potential TFBS disruptions (using ENCODE TFBS data). There was still a big distinction between the number of times the gene was mutated (as depicted by the number of accumulated variants within that gene) and the number of TFBSs the mutation may have disrupted. However, as opposed to when all non-coding variants were considered, with the *deleterious* non-

coding variants, there were **more TFBS** disruptions for a particular gene than somatic variants. Hence, by using the RegulomDB scoring system one is better able possibly isolate the somatic variants that have a functional regulatory effect, while the ENCODE TFBS data allows for the observation of the actual number of distinct TFs **and** TFBSs affected. For TRIO, PPARG and RUNX1 for example, there were an estimated 14-15 fold more possible disruptions in the TFBS of these genes compared to the number of non- coding somatic mutations as depicted by Figure 17. This could demonstrate a greater functional consequence for the mutations at those positions.

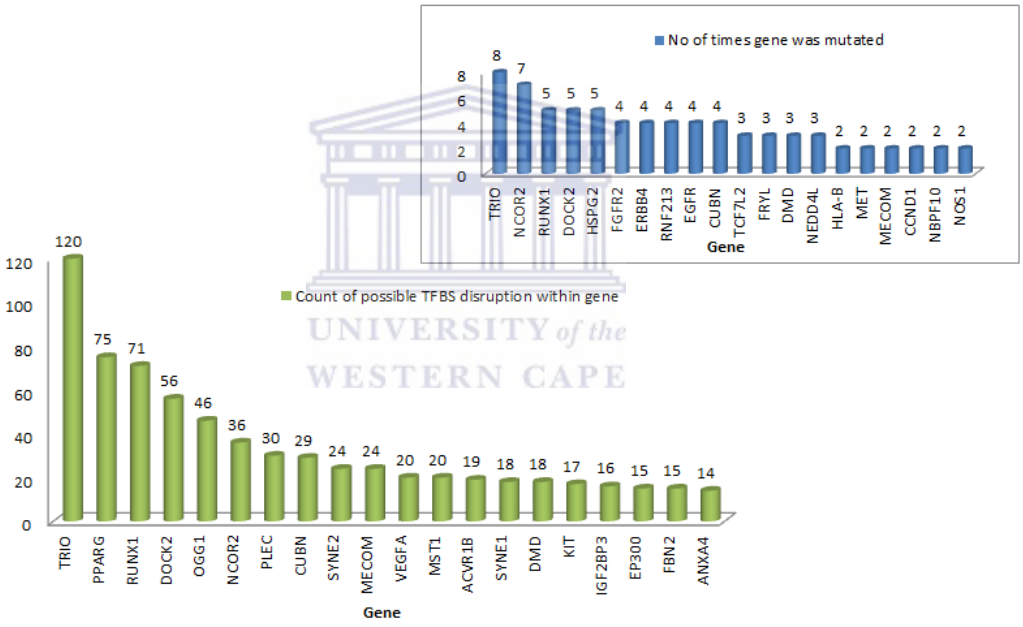


Figure 17: The top 20 genes with the most deleterious non-coding variants (blue) and the top 20 deleterious genes with the most TFBS disruptions (green). There was a much better overlap between these genes as seen by TRIO, RUNX1, DOCK2 and NCOR2, compared to when all non-coding (deleterious and tolerated) somatic variants were considered. This indicates that by using the RegulomDB scoring system one is better able to observe the non-coding somatic variants that may have an adverse effect on transcriptional regulation.

a) ***Number of CDS somatic mutations versus number of TFBS disruptions***

As when the total non-coding variants were taken into account, when the total number of *CDS* variants in the RCC genes was compared to the number of possible TFBS disruptions, for the most part, there was no strong association between the two. Although it may seem unexpected at first to detect TFBSs in CDS regions, this phenomenon was previously observed in a study carried out by Stergachis et al. (2013). Genomes contain both regulatory code that specifies TFB recognition sequences and genetic code for codons that specify amino acids. These codes have always been assumed to be segregated physically into the coding and non-coding compartments of the genome and to operate independently of one another. However, the potential for some coding sequences to accommodate splicing signals and transcriptional enhancers has long been recognized (Stergachis et al., 2013). In fact, by using DNaseI footprinting in 81 different cell types, Stergachis et al. (2013), found that more than 15% of human codons are dual used codons, termed duons. These duons simultaneously specify amino acids and TFBSs. TFBSs have already been detected in the exons of keratin18 in humans (Stern and Orgogozo, 2008). Approximately 17% of single nucleotide variants (SNVs) that fall within duons alter TF binding (Stergachis et al., 2013). The 153 CDS somatic variants were determined to possibly impact the binding of TFs to their 304 TFBS (not shown). This amounted to an estimated **two** TFBS disruptions per CDS mutations. In contrast, there were 891 possible TFBS disruptions for the 125 non-coding deleterious variants or approximately **seven** possible TFBS disruptions per non-coding deleterious variant. In addition, out of the 70/95 *patients* with CDS somatic variants (previously shown in Table 5), less than 40% of them (only 27 patients) had mutations that fell within TFBSs and those variants that **did** fall within TFBSs affected only 23 genes.

Moreover, upon further investigation, these TFBS disruptions caused by CDS variants were mostly attributed to one gene in one donor that had a somatic mutation in a location where 68 TFs potentially bind. In terms of genes,

although the VHL and SETD2 genes, which were previously two of the top 20 genes (with regard to the most CDS variants), remained in the top 20 genes with the most potential TFBS alterations, as displayed in Figure 18 (**blue**), most of the mutations in the CDS (**orange**), did not seem to be specifically targeted at a genes TFBSs.

Meanwhile, the peroxisome proliferator activated receptor gamma (PPARG) gene was previously not highlighted in any of the categories with the top 20 frequently mutated genes, yet the position of its CDS variant potentially alters numerous TFBSs (**orange**). PPARG is a nuclear receptor that has been shown to **inhibit** cell growth in ccRCC, but only if it is expressed and remains unaltered (Collet et al., 2011). A disruption at a critical location in its TFBS that eliminates this binding site could cause the gene not to be expressed and would thereby **promote** tumourigenesis. This shows again that evaluating the *position of a variant* with respect to their TFBSs could shed better light on the functional consequences of that specific variant. This holds true even for synonymous/silent CDS mutations which are now widely accepted as disease-causing due to their impact on the efficiency of protein folding, aberrant splicing, the folding energy and structure of pre-mRNA, as well as the creation and modification of binding sites (Buske et al., 2013) (Faa′ et al., 2010) (Sauna and Kimchi-Sarfaty, 2001).

Despite finding that CDS variants do not generally fall within TFBSs, it is still worthwhile to investigate their position in relation to their binding motifs. Since 13.5% of common disease- and trait-specific SNVs identified in GWAS studies fall within duons, they may have a profound effect on pathogenesis (Stergachis et al., 2013). Correct classification of mutations, such as identifying the previously unknown molecular defects caused by a mutation, especially the largely unexplored non-coding variants or the previously dismissed synonymous (neutral) coding mutations, could allow us to define the pathogenic role of many sequence variants, which could ultimately lead to better therapies. The unravelling of its role could very well be as a result of defining them with regard to their TFBS disruptions. In the

next subsection a few significant TFs were investigated for their role in the severity of the ccRCC.

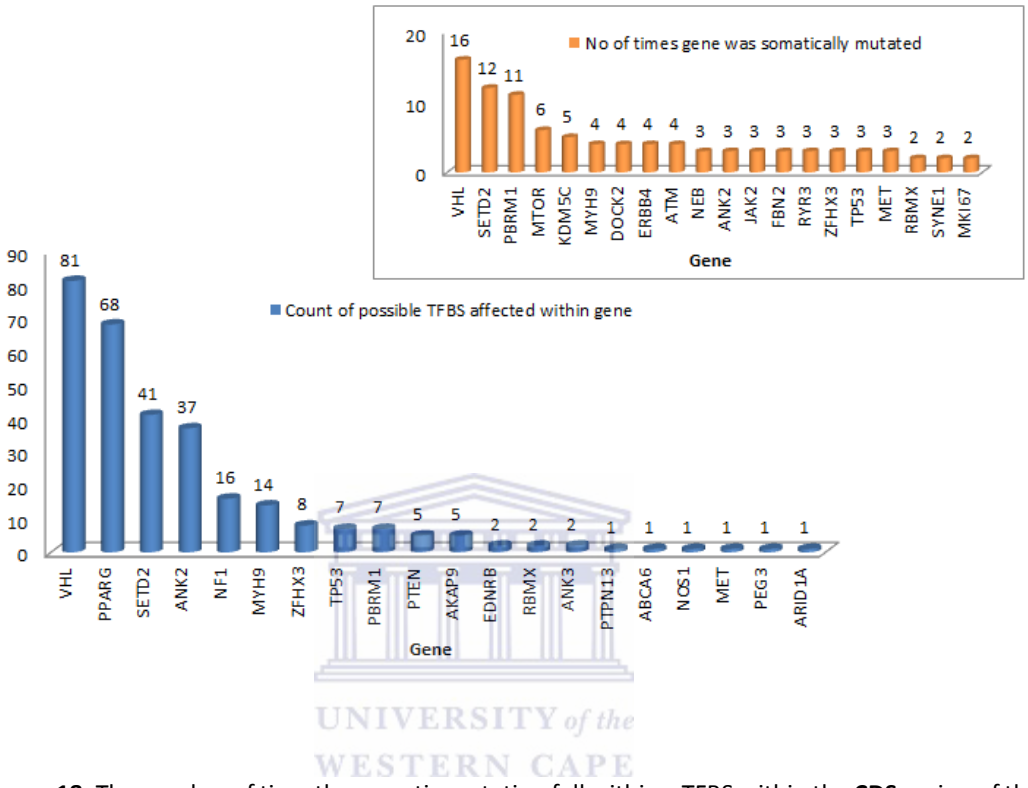


Figure 18: The number of time the somatic mutation fell within a TFBS within the **CDS** region of the gene (**blue**) compared to the number of times the gene was somatically mutated (orange). Although VHL and SETD2 remained in the top five, many of the genes frequently mutated did not fall within TFBSs, while the location of others such as PPARG which previously didn't even come up in the top 20, affected many TFBSs.

3.9.1. Transcription factors (TFs) in deceased patients

The TFs that bind at these possibly disrupted TFBSs were also investigated within all **five** *deceased* donors, to ascertain if specific TFs were common to all of the deceased patients, but absent in the surviving patients, thereby serving as biomarkers for a poor prognosis. Mutations in the binding sites of three of the TFs: CEBPB, EBF1 and CTCF were commonly found within the five deceased patients, but only two of the

patients had mutations that may simultaneously affect the binding of all three TFs, as Table 9 indicates.

CTCF and EBF1 binding site disruptions were present in 3/5 patients, while disrupted CEBPB binding was common to 4/5 of the deceased patients. Early B-cell factor 1 (EBF1) is a transcriptional activator, just as CCAAT/enhancer binding protein beta (CEBPB), while CCCTC-binding factor (CTCF) is a chromatin binding protein that acts as a transcriptional activator or repressor depending on the gene. It became apparent that a mutation in the binding site of CTCF may be critical, since one patient (DO47004) who lost their battle with ccRCC, had only one deleterious variant located in the binding site of this TF and *no other* deleterious mutations that potentially impacted TFBSs. CTCF has been shown to modulate cell differentiation, cellular senescence, as well as the control and progression of the cell cycle, and may also potentially act as a tumour suppressor (Fiorentino and Giordano, 2012). In 2014 Kemp et al. showed that CTCF hemizygous (having only one copy of a gene instead of the normal two copies) knockout mice were more susceptible to cancers and these cancers were particularly characterized by their increased aggressiveness.

One deceased patient (DO47249) who did **not** have a mutation in the CTCF binding site, did however, have mutations in the TFBSs of the other two TFs, namely: CEBPB and EBF1. When this analysis was carried out in the surviving patients, none of them had deleterious variants that fell within TFBS of the CTCF transcription factor and none of them, except for one patient had the combination of CEBPB/EBF1 TFBS disruptions. This exception was a patient for whom no vital status or other clinical data was recorded and so their results could be included in this study.

It would thus be worthwhile to explore these TFs as potential targets for cancer therapeutics or to investigate its function in tumourigenesis or cancer progression. Although this analysis was beyond the scope of this study, in the next sub-section, a STRING-DB analysis was carried out to view the connections these three TFs make with other molecules in order to understand their possible influence on the disease.

Table 9: The binding of the three transcription factors (CEBPB, EBF1 and CTCF) within the five deceased donors may be altered by non-coding variants located within their TFBSs. Every one of the three TFs affected was present in at least three of the patients at a time, but no single affected TF was ever common to all five patients. All the deceased patients had a combination of the disruptions in the binding sites of the TFs CEBPB/EBF1 or a disruption in the binding site of CTCF. None of the surviving patients had the former combination or a disruption in the binding site of CTCF. To see if any of these TFs interplay with each other or if they are attached to other known genes that are causally implicated in cancers, a protein-protein interaction analysis was carried out.

DO47249	DO46885	DO46828	DO46827	DO47004
CEBPB	BATF	ATF2	CCNT2	ARID3A
EBF1	CEBPB	ATF3	CEBPB	ATF3
EP300	CTCF	BCLAF1	CHD1	BRF2
FOXM	EBF1	BRCA1	CTCF	CREB1
IKZF1	FOS	CBX3	EBF1	CTCF
MTA3	JUN	CCNT2	EGR1	EGR1
NFATC	MAFF	CEBPB	FOXA1	EP300
NFIC	MAFK	CHD2	FOXP2	ETS1
NR2C2	REST	CTCF	HDAC	FOS
PAX5		E2F1	KDM5B	FOSL2

3.9.2. Network analysis of TFs

The Bonferroni statistical method ($p < 0.05$) was chosen at the highest confidence levels (0.900) and only experimental evidence was considered to build the network. Figure 19 demonstrates that CEBPB and CTCF interplay via two intermediates. If a disruption in the binding site of CTCF is as detrimental as hypothesized, then CTCF's relationship with CEBPB could explain how the cancer in the patients without the potential CTCF disruptions, could accelerate to its final state. EBF1, on

the contrary, remains independent of the other proteins in the network and even when EBF1was considered individually and the network around this protein was expanded to observe other directly-interacting molecules, this protein remained unassociated. Hence, not much is known about its interactions with other proteins in metabolic processes. However, EBF1 has been shown to cooperate with STAT5 (Heltemes-Harris et al., 2011), which is a signal transducer and transcription activator that has been causally implicated in several cancers, including renal cancer (Cavalcanti et al., 2010)(Yamashita and Iwase, 2002). Therefore, this protein could also be further explored as a potential genetic marker or to better understand the regulatory architecture of ccRCC.

When the proteins were grouped according to their Biological process, all three TFs participate in positive regulation of *gene expression*, metabolic- and biosynthetic processes. Dysregulation of gene expression and cellular processes are known cancer mechanisms (Srihari et al., 2014)(Skubitz and Skubitz, 2002) and thus, it is not surprising that disruptions in the binding sites of these TFs could result in cancerigenesis or metastasis. This was however, an extremely small sample size so no firm conclusions could be drawn.

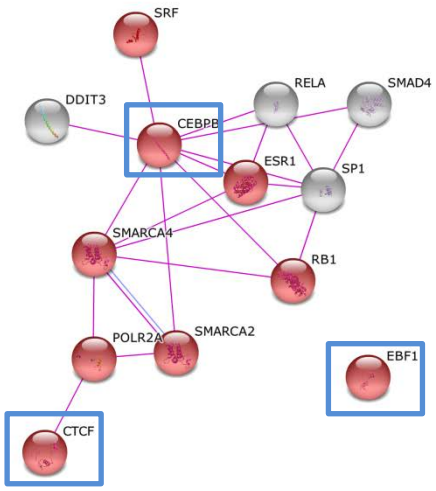


Figure 19: All three proteins participate in the positive regulation of gene expression, metabolic and biosynthetic processes according to the Gene ontology category 'Biological process' in STRING-DB.

Finally, the TFBS data was linked to the clinical data of the donors to see if trends could be observed that related the number or location of potential TFBS disruptions to the severity of the disease.

3.9.3. TFBS data combined with clinical data

When the TFBS data was added to the clinical data of the patients, one patient's data deserved special attention, in that no mutations in the CDS region was observed, and despite harbouring six non-coding somatic mutations, these mutations were not deemed deleterious by RegulomDB. Moreover, none of the mutations fell within TFBS locations. More explicitly, there were no (CDS) somatic variants that affected the amino acid sequence or (non-coding) variants that could influence regulation by directly altering TFBSs.

The fundamental question was therefore, is it only somatic mutations that potentially disrupt TFBSs; disruptions which ultimately influence transcription or gene dysregulation? In recent years there have been reports of DNA methylation's possible role in directly impacting the binding of TFs to TFBSs. In 2011, Chen et al. showed that TF binding affinities are directly influenced by the methylation of cis-regulatory elements (such as **promoters**, introns, enhancers etc.). It has been shown that some disease genes go under the radar with regard to DNA mutations and TFBS disruptions, yet fall victim to aberrant methylation events that eventually result in the expression of the disease phenotype. So the next section of this study focused on assessing the signatures of methylation with regards to TFBSs and gene regulation.

3.10. Aberrant Methylation in GOI

Gene regulation is driven by TFs that bind to TFBSs at promoters, and this binding affinity is controlled by **promoter methylation** (Luu et al., 2013). The COSMIC methylation data was therefore expedient for this study, since the promoter-associated positions were already defined by using the position of the probes on the

array chip. Initially **all** the differentially methylated promoter-associated positions within the 173 RCC disease genes were extracted, within their genomic regions of interest, in order to see the overall extent of aberrant methylation within their promoters. That is, the relationship between differential methylation and its effect on gene dysregulation was not yet considered. The motivation behind scanning all of the genomic regions is predicated on the awareness that promoters may be located in the 5'UTRs, the introns or even downstream of genes, in the 3'UTRs of human genomes, as explained earlier in section 1.9.2.4 (Holloway et al., 2008). Later this data was contrasted with gene dysregulation to characterize the aberrant methylation as potentially being functionally significant.

As shown in Table 10, no promoter-associated aberrant methylation was reported in the CDS region, which was expected, since promoters are not located within protein coding sequences. The values reported in the table are more a reflection of the number of patients with differential methylation at a specific genomic location than the number of unique differentially methylated events (e.g. patient **A** at chr**3:500548**, patient **B** at chr**3:500548**, patient **C** at chr**3:500548,** patient **A** **again**, at chr**17: 2124587**). Initially there were 1024 methylation events, however, many of the hits in the introns and promoter regions were at the same genomic positions. After removing duplicates there were 616 methylation events that affected 17 genes, in 191 of the original 307 patients enrolled in the methylation study. Thus, nearly two thirds of patients (62%) had aberrant DNA methylation, even though not many distinct genes were affected. Out of the 616 events there were just 38 unique differentially methylated positions.

Table 10: The number of differentially methylated positions in the various genomic regions and the associated RCC genes. The CDS region showed no promoter-associated aberrant methylation. There was aberrant promoter methylation in the 3'UTR, which was initially of concern, considering promoters are usually upstream of the transcription start site of genes. However, as stated before, in complex organisms such as humans, promoters may even be located in the 3'UTRs of genes.

	Promoter	5'UTR	CDS	Introns	3'UTR
Aberrantly methylated position (promoter associated)	303 (OGG1, PDGFRA, VEGFA, IGF2BP3, RUNX1, LCP1)	124 (RUNX1, PACRG, BAP1)	0	586 (VHL, PACRG, HLA-B, NF1, RUNX1, LCP1, MAX, SCD, NCOR2, PLEC, NBPF10)	11 (KMTD2)

In contrast, the non-disease genes had **no** promoter-associated differential methylation in the 3'-, 5'UTR or the CDS regions, as displayed in Table 11. When the duplicate positions between the promoter and intronic aberrant DNA methylation events were removed, there were only 31 unique methylation events that affected just four of the 173 non-disease genes (2%). Only 25 out of the 307 patients were affected; that is 8% of patients, compared to 62% of patients when the RCC disease genes were considered. Again, the 31 events were based more on the number of individual patients **with their** distinct aberrant methylation events (e.g. patient **A** at chr**3:500548**, patient **B** at chr**3:500548**, patient **C** at chr**3:500548,** patient **A again**, at chr**17: 2124587**), but the unique genomic positions for which there was differential methylation totalled just six. In terms of the sizes of the genes affected, the 17 differentially methylated genes were scattered across all gene lengths. In this part of the analysis gene length is still relevant, because the methylation patterns were observed within the individual genomic regions (5'UTR, 3'UTR, introns, CDS and promoter) and not just in the 1000 bases chosen as the promoter region for all genes at the start of the study.

Since size wasn't a factor, the differential methylation could not have been coincidental. This again showed that methylation events are directed towards certain genes and because they were less common in non-disease genes, differential methylation events are also more enriched in the context of disease. Hence, working backwards from differentially methylated positions to the variants/genes affected, may help in the discovery of novel variants/genes and enhance our understanding of the role epigenetics plays in complex diseases. In the next section aberrant methylation and its association with TFBS is investigated for their possible link to TFBS affinities.

Table 11: The number of differentially methylated positions in the various genomic regions for the non-disease genes. The CDS, 3UTR and 5UTR regions showed no promoter-associated aberrant methylation.

	Promoter	5'UTR	CDS	Introns	3'UTR
Aberrantly methylated position (promoter associated)	29 (EIF3L, ZNF502, TMCO6)	0	0	2 (ZNF502, PRKRIP1)	0

3.10.1. Aberrant methylation in relation to TFBSs

As stated previously, there have been multiple reports that DNA methylation mediates cell differentiation and gene regulation by altering the interactions between TFs and their TFBSs (Chen et al., 2011). When the **TFBS** is **methylated**, the **TF cannot bind** at its binding site and the gene cannot be expressed (Chen et al., 2011). As Figure 20 illustrates, there was a good overlap between the genes that had differentially methylated promoters (**C**), the genes with the most potential TFBS disruptions located in the genes that had the most *total* noncoding somatic variants (deleterious and tolerated) (**A**) as well as the genes with the most potential TFBS disruptions in the genes with the most *deleterious* somatic variants **(B)**. Many

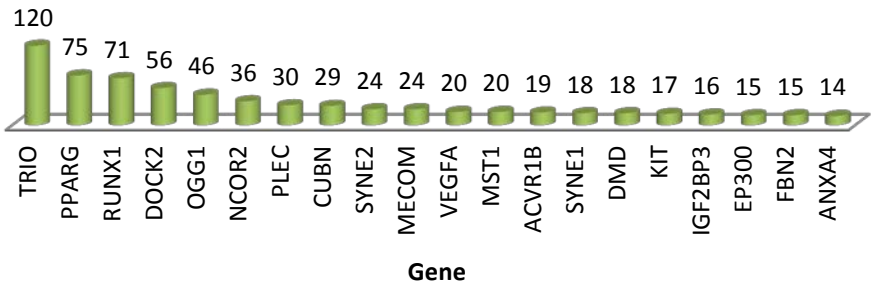
differentially methylated genes such as: RUNX1, NCOR2, PLEC, OGG1 etc., were also amongst the top 20 genes of both of the latter two categories. Other aberrantly methylated genes such as VEGA and IGF2BP3 were reported in at least one of the two categories of genes with frequent TFBS disruptions (categories: genes with the most *total* non-coding SVs with corresponding levels of potential TFBS disruptions **or** genes with the most *deleterious* SVs with corresponding levels of potential TFBS disruptions).

Interestingly, the BAP1 gene **(C)**, which didn't really come to the forefront within the other parts of this study, despite being frequently implicated in ccRCC, was highlighted in this section of the study. Albeit, the VHL gene which has been found to be differentially methylated in 15% of ccRCC cases (Ricketts et al., 2014), *was* one of the frequently methylated genes in this analysis, corroborating the findings of this study.

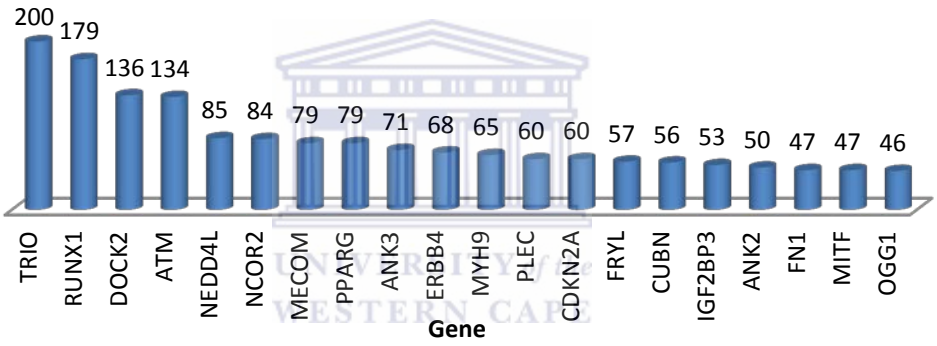
There were, however, far fewer genes with differential promoter methylation than genes with somatic variants. As discussed previously, only 17 of the 173 disease genes reported aberrant methylation, despite 165 of them containing non-coding somatic variants. When the GO annotations for these 17 differentially methylated genes were investigated, no apparent trend could be observed in terms of identical shared molecular functions. However, 81% of the genes participate in a combination of signalling, proliferation or the activation of certain pathways (not shown). These are all well-known proto-oncogenic activities, although this observation will only be relevant if the differential methylation also results in the upregulation of these genes.

Interestingly, no additional genes were shown to be differentially methylated that were not previously shown to be somatically mutated (this will be elaborated on later). Another interesting observation was the type of differential promoter methylation frequently observed. Most methylation studies report promoter *hyper*methylation, but in this study almost all (97%) of the non-coding promoter-associated positions were hypomethylated. Only 16 of the 616 methylation events were as a result of *hyper*methylation. In 2002, Ehrlich showed that although genomic hypermethylation is often seen in the CpG islands of cancers, frequent hypomethylation often underlies transcriptional control sequences.

A) The number of TFBS disruptions in the genes with the most total non-coding mutations



B) The number of TFBS disruptions in the genes with the most deleterious non-coding mutations



C) The genes with the most differentially methylated positions in their promoters

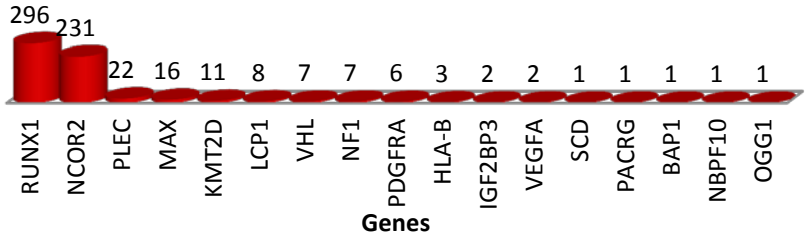


Figure 20: A) The top 20 genes with the most **TFBS** disruptions that overlapped with the genes with the most **total** (deleterious + tolerated) non-coding variants. **B)** The top 20 genes with the most **TFBS** disruptions that overlapped with the genes with the most **deleterious** non-coding variants. **C)** All the genes with **aberrantly methylated** positions in their promoters**.** Many of the differentially methylated genes such as RUNX, NCOR, PLEC, OGG1, IGF2BP3, etc. in **(C)** overlapped with the genes in panels A and B, showing a possible relationship between methylation and TFBS disruptions.

Thus, although hypermethylation is frequently studied, inappropriate hypomethylation of oncogenes has been shown to be key feature in cancerigenesis, in recent years (Chen et al., 2011).

In as early as 1983, Gama-Sosa et al. reported that the malignant tissues of 103 human tumours had significantly lower genomic m5C content than benign and normal tissues. In fact they demonstrated that the level of hypomethylation correlated with the extent of the cancer, with metastases showing the highest percentage of hypomethylation. This may also reflect a role for the extent of hypomethylation in relation to tumour progression (Gama-Sosa et al., 1983). Later it was shown that ovarian carcinomas and Wilms tumours respectively showed up to 25% and 60% less m5C in their DNA than their normal counterparts (Ehrlich, 2002). It has therefore been suggested that assays for DNA hypomethylation may become a clinically useful addition to hypermethylation of CpG islands (Ehrlich, 2002).

At the end of section 3.9.3, the question was posed as to the relevance of an association between of methylation events and TFB affinities. Since there seems to be a relationship between the genes reporting the highest differential promoter methylation events and the genes with the highest number of TFBS disruptions and because this relationship **only** appears to exist for the non-coding variants (**not** the **CDS** variants, hence it seems to be targeted at transcriptional regulation) the possibility exists that the two events may somehow be tied and that aberrant methylation may be used as a mechanism by cancer cells to pervert normal cellular processes to their advantage. Furthermore, earlier it was highlighted that no new genes were differentially methylated that were not also somatically mutated, suggesting that there is a possible relationship between somatic variants and differential methylation, and since the non-coding variants are often located within multiple TFBSs, the association may even be more specifically between methylation and TFBS disruptions.

Of course knowing whether the somatic variant (or SNP) or the aberrant methylation came first, is a bit harder to ascertain. Nevertheless, rSNPs have been reported to alter CpG methylation, representing one of the mechanisms that link genetic variations, such as somatic variants, to epigenetic changes (Li et al., 2014).

Albeit, aberrant methylation only becomes relevant to disease when there is a correlation with gene dysregulation of the gene expression of the target gene for the associated promoter. Hence, a Python script was again used to extract only the instances where the differential status was indirectly proportional with the gene expression levels (since promoter **methylation represses** gene expression).

3.11. Gene Expression

3.11.1. Gene Expression and aberrant methylation

When the relationship between the aberrant DNA methylation and the associated gene expression levels (Fold change >2) were taken into account, six genes harboured 102 differentially methylated events with corresponding differential gene dysregulation in 30 of the patients. These six genes were also the top genes which had the most differential methylation. In contrast the non-disease genes reported **no** correlation between the two events for any of the genes. All of the aberrantly methylated positions for the RCC genes were the **hypo**methylated promoter regions with the **over**expression of the target gene as shown in Figure 21. That is, **none** of the **hype**rmethylated promoter regions showed inversely proportional dysregulated genes.

In 2002 Ehrlich stated that there is evidence that cancer-associated hypomethylation of proto-oncogene promoters is correlated with **activation** of the gene expression counterpart. This was similarly shown to be true this study. Since no '**hyper**methylated promoter-**downregulated** gene' relationships existed, the purpose of hypomethylation in ccRCC may be exclusively targeted at the upregulation of proto-oncogenes. Now the purpose behind the targeted promoter **hypo**methylation of the 17 genes becomes apparent, since the upregulation of proto-oncogenic activities is one of the hallmarks of cancer.

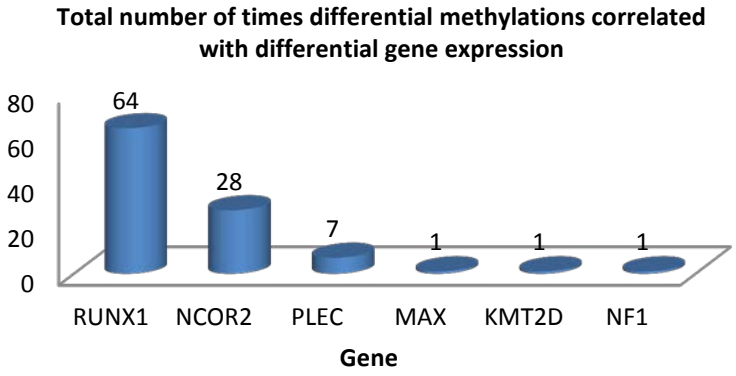


Figure 21: The total number of times a gene displayed aberrant DNA methylation and an indirectly proportional gene expression levels. All of the genes were hypomethylated with concordant upregulation of the same gene.

In 2013, Luu et al. used an algorithm to show that *hypo*methylated-associated motifs often target conserved regions near TSS and *tend to occur in TFBS*s. Through a series on intermittent mechanisms this results in the conversion of enhancers from a repressed state (H3K27me) to an active state (H3K27a), which naturally have implications in the overexpression of the target gene (Luu et al., 2013).

After contrasting differential methylation and gene dysregulation, the relationship between non-coding SVs and gene dysregulation was investigated, since it was expected that there would be a better correlation between the two events if the SV was also a rSNP; that is, a SNP that alters TFBSs (Buroker, 2014).

3.11.2. Gene expression and somatic mutations

As stated before in section 3.9, rSNPs often result in gene expression changes due to their impact on TF binding. A total of 158 of the 165 somatically mutated genes were dysregulated in ccRCC tumours compared to their normal tissue. The seven genes which were not dysregulated in ccRCC patients, despite accumulating multiple non-coding somatic mutations in the WGS study, were investigated to observe the number of potential TFBSs they may disrupt. Although two of the genes had variants that

could alter multiple TFBSs, the other four gene variants would result in minimal TFBS modifications, with two of them just potentially disrupting one binding site each. Furthermore, none of these seven genes acquired aberrant promoter methylation. Therefore, non-coding mutations are possibly more likely to have an impact on gene expression if variant lies at a position that potentially alters multiple TFBSs or if the promoter of the gene is aberrantly methylated.

Once again, when the specific genes **most frequently** dysregulated (Fold change > 2) were analysed, many of those genes (MST1, RAN, PBRM1, RUNX1 etc.) were also the genes that had deleterious non-coding mutations, disrupted TFBSs and some, such as the RUNX1 gene, were also hypomethylated with concordant overexpression of their genes (See Figure 22). Also, most of these genes, as with previous sections of this study, are not the genes most commonly highlighted in ccRCC somatic mutation studies (such as VHL) where WES is the empirical platform and the coding sequence is the focus of the analysis. Interestingly, the KDM5C and NSD1 genes were dysregulated, despite having no variants at *TFBSs* and no *differential promoter methylation*, but only a few **tolerated**, non-coding somatic variants in the patients enrolled in the WGS somatic mutation study.

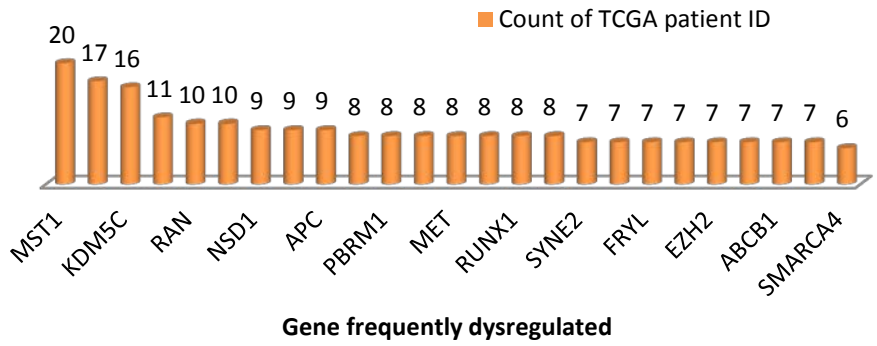


Figure 22: The genes often dysregulated were the same genes that were previously shown to have deleterious, non-coding somatic variants. For these genes the genomic location of the mutation affected multiple TFBSs and some such as the RUNX1 gene contained deleterious, non-coding SVs, possible TFBS disruptions, differentially methylated promoters and concordant gene dysregulation.

KDM5C did harbour some CDS mutations, but coding mutations generally affect the amino acid sequence, rather than gene expression levels. Therefore, for genes such as KDM5C and NSD1, there may be an alternative mechanism at work that elucidates its observed gene dysregulation. Then again, it may also be that many of the non-coding variants annotated as tolerated by RegulomDB, may in fact be more functionally significant than anticipated.

Nonetheless, the relationship between the TFBS disruptions, non-coding somatic mutations and differential methylation and their gene expression levels, demonstrates that there is definitely a niche for considering variants in the non-coding regions of the genome as well as the epigenetics and regulatory information in their contribution to ccRCC.

This therefore also adds weight to the theory that one should look beyond just the mutations in the genome to **a)** what causes them and **b)** the consequences of the same. However, the sample IDs that were genotyped for somatic mutations (processed by ICGC), differed from the sample IDs of the methylation and gene expression data (processed by TCGA). A direct comparison across data types could therefore not be made on a patient-by-patient basis, despite the use of the same samples and hence, no strong associations could be made.

Nevertheless, with complex multifactorial diseases such as cancers, the pathogenesis often depends on interplay between various dysregulated genes and/or their regulatory and epigenetic events. Taking a systems biology approach could therefore shed more light on which genes are more relevant in the expression of the disease phenotype. To this end the significant genes were submitted to STRING-DB to analyse the interactions of these genes and proteins in a network.

3.12. STRING-DB protein-protein interactions

The genes with the most deleterious, non-coding somatic mutations (n = 31, genes), the top 30 genes with the most TFBS disruptions in the non-coding regions, all the genes which incurred differential methylation in their promoters (n = 17) and the

three TFs for which the TFBSs are commonly disrupted in the deceased patients, were extracted and duplicates were removed in Excel. A total of 57 unique genes/proteins were submitted to STRING-DB (shown in Appendix VI) and the analysis was carried out using the highest confidence score in STRING-DB together with the Bonferroni correction method ($P < 0.05$).

When an enrichment analysis was carried out to group the proteins according to Biological function, 25 of the 57 genes were involved in negative regulation of biological processes, as illustrated by the red molecules in Figure 23 (A). According to its Gene Ontology definition (https://www.ebi.ac.uk/QuickGO/GTerm?id=GO:0048519), this is any process that **inhibits** or **downregulates** the frequency or the extent of a **biological process**. This influence may be exerted by means of dysregulation of **gene expression** or protein modifications and interactions.

When the Markov Cluster Algorithm (MCL) was applied, a surprisingly high number of genes were observed to be **directly** interacting with each other, as shown in yellow in Figure 23 (B). Thus, mutations in any of these genes or their TFBSs that alter the levels of proteins present could have a snowball effect on downstream molecules in the pathway and possibly influence numerous neighbouring metabolic processes. If the targeted genes being downregulated are tumour suppressor proteins, then decreasing their levels could result in the cell cycle not being arrested if DNA repair is required. Similarly, if proteins responsible for the degradation of proto-oncogenes upon completion of their function are downregulated, this would increase proliferation, thereby contributing to tumourigenesis.

Some of frequently mutated genes such as TRIO and DOCK2 do not form any links within this network. Many of them such as KMT2D (shown in STRING-DB by its synonym *MLL2*), NOTCH1, HSPG2, RUNX1, ERBB4 and ATM do, however, interact with each other **and** interplay with multiple other proteins that form the backbone of the network (as shown in Figure 23(A) in the **green** blocks). These backbone genes/proteins which include EGFR and HRAS and NFkB1 have already been implicated in many cancers (Minner et al., 2012) (Fujita et al., 1988) (Hoesel and Schmid, 2013). Two of the noteworthy genes in ccRCC, VHL and MET (shown in Figure 23(A) in the **blue** block), form connections on the outskirts of this network,

but also interplay with many cancer genes such as PDGFRA and VEGFA, demonstrating how modifications in many of these genes could result in the cancer phenotype.

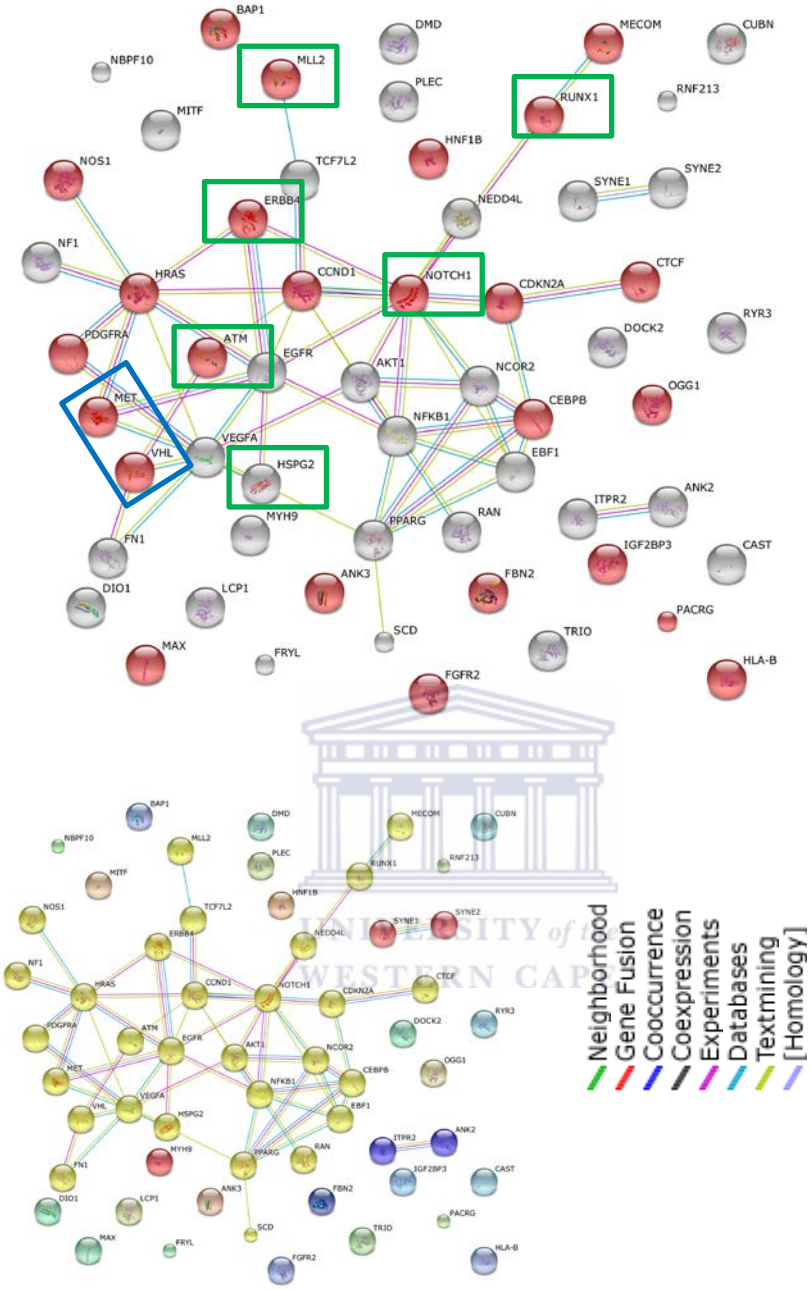


Figure 23: (Top) Most of the proteins (25/57) were grouped under 'Negative regulation of biological process' according to the Gene ontology criterion: Biological Process (shown in red). Many genes such as KMT2D (shown as MLL2), NOTCH1, HSPG2, RUNX1, ERBB4 and ATM interact and interplay with multiple proteins that form the backbone of the network (**green** blocks). Two of the noteworthy genes in ccRCC, VHL and MET, form connections on the outskirts of this network (shown in the **blue** block), but also interplay with many cancer genes such as PDGFRA and VEGFA. **(Bottom)** When MCL clustering was applied, one can clearly see that many of these proteins directly interact with one another (shown in yellow) (Adapted from String-DB, 2015).

Additionally, in this small network (n=57), ten of these proteins participate in the phosphoinositide 3-kinase/Akt (PI3K/Akt) pathway and nine out of the ten proteins from a **tight network** (shown by the red molecules in Figure 24). Since these are the genes frequently modified in ccRCC patients across various studies and because this pathway has already been associated with ccRCC (Guo et al., 2015; Porta and Figlin, 2009), this also demonstrates how catastrophic a mutation or the inactivation of just one gene could be to a pathway. Alternatively, it could imply that more than one of these genes needs to be altered or inactivated in order for the disease phenotype to be expressed.

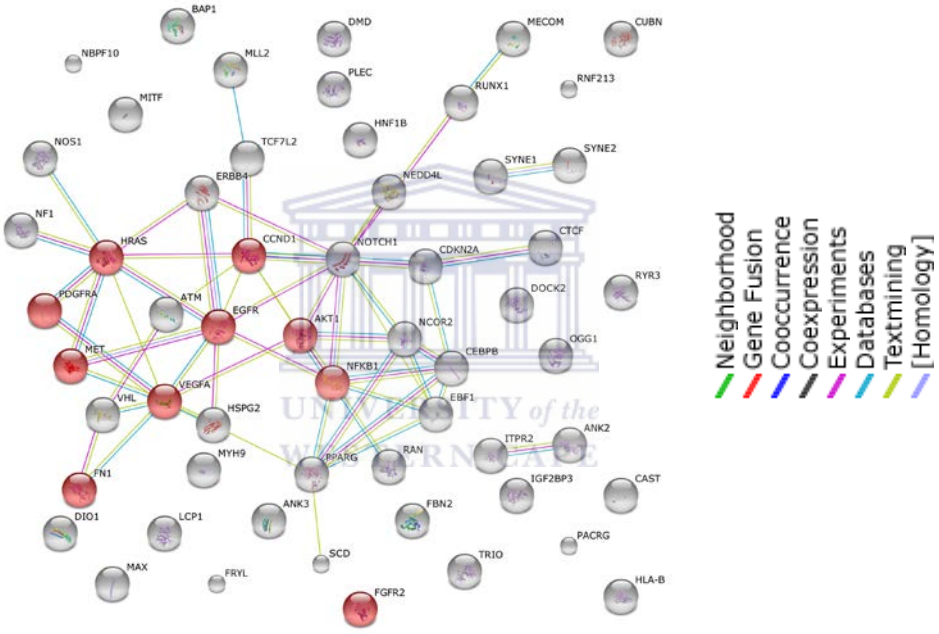


Figure 24: The ten genes/proteins that participate in the PI3K/Akt pathway. Most of them were interlinked with the exception of FGFR2.

Finally, the network was expanded within the context of STRING-DB's Active prediction methods to view the surrounding molecules. The intention was to observe other proteins (not included in the original submitted list) that may interplay with the 57 proteins of interest and to identify hub proteins that may further elucidate the role of the genes/proteins.

A total of 26 of the 57 proteins can be seen interplaying with ubiquitin C (UBC), as illustrated in Figure 25. Table 12 shows that four of these proteins are linked to UBC via an intermediate protein, while 22 interact directly with UBC. Several studies have reported and discussed frequent mutations in the ubiquitin-proteasome pathway (UPS) and more specifically, the link between the VHL gene and ubiquitin (Guo et al., 2012) (Roos et al., 2011) (Corn, 2007)(Ishizawa et al., 2004). However, not many studies have focused on the roles of these **other** genes and their association with ubiquitin C in ccRCC. Ubiquitin has been shown to be involved in DNA repair, cell cycle progression, the modification of polypeptide receptors, the biogenesis of ribosomes (ribosomes are essential for cell proliferation) and transcription regulation (Shi and Grossman, 2010) (Kanayama et al., 1991).

Cyclins, which play a fundamental role in modulating the cell cycle during cell proliferation are, for example, degraded by the ubiquitin proteins (Kanayama et al., 1991). The level of cell surface signal receptors, which play a role in cell migration, proliferation, survival and differentiation, are also tightly regulated by the ubiquitin pathway (Diehl et al., 2010)(Fischer et al., 2009). Due to ubiquitin's dynamic role in protein regulation by triggering the degradation of target proteins, they can act as either oncogenes or tumour suppressor genes, depending on which of these genes are targeted for degradation at any given time (Diehl et al., 2010). Interestingly, in as far back as1991, Kanayama et al., already showed that the levels of poly-ubiquitin C were higher in malignant renal tumours than in their normal counterparts.

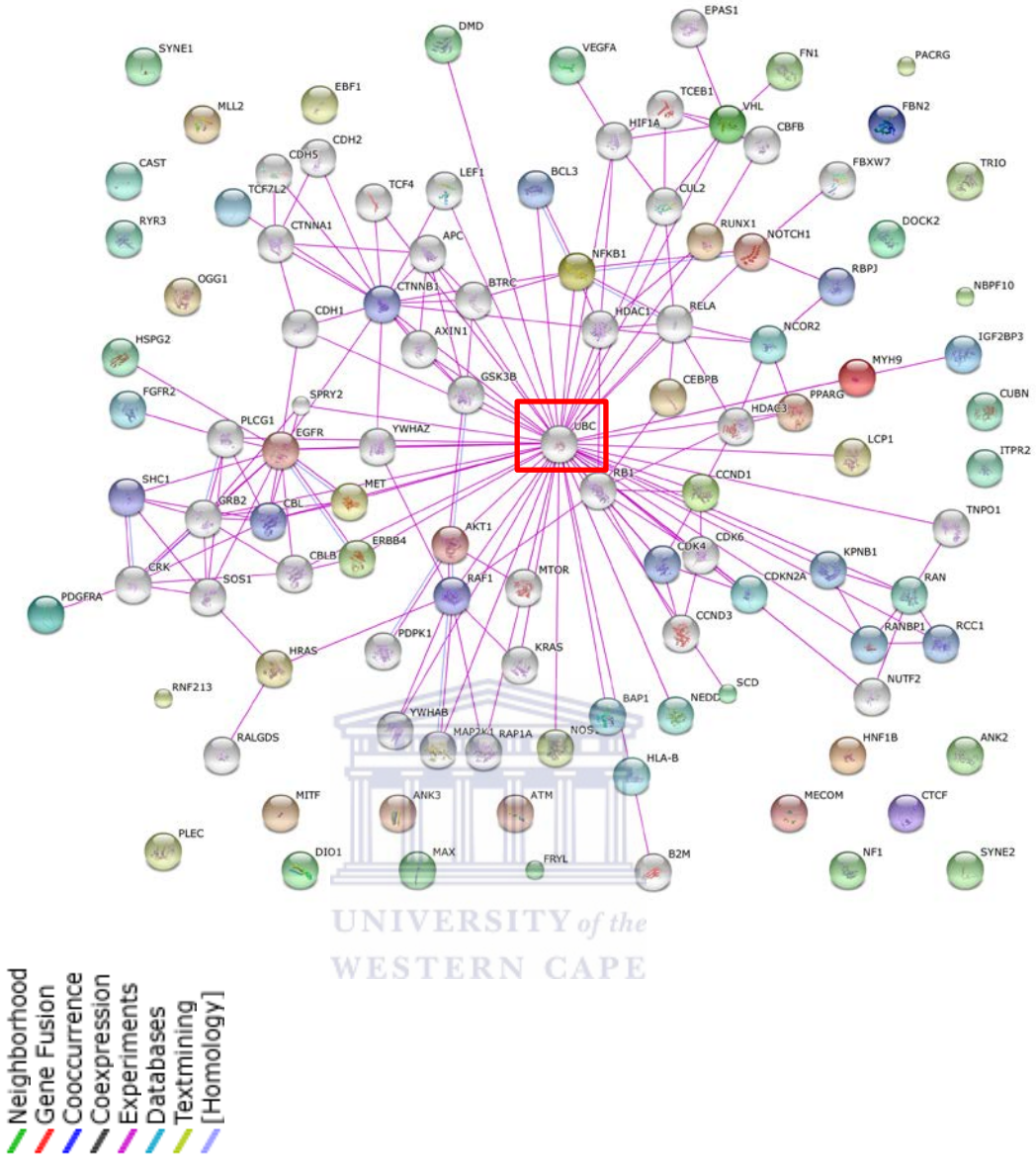


Figure 25: Most of the genes of interest were centred on Ubiquitin C (centre in **red square**). Some were also linked to EGFR or RB1 (Adapted from String-DB, Accessed 29/09/2015).

Table 12: The **g**enes/proteins interacting directly with **ubiquitin C (UBC)** or via an intermediate gene/protein.

Protein/gene	Direct link	Via another gene/protein
CDKN2A	√	
VHL	√	
NOTCH1	√	
RUNX1	√	
NOS1	√	
ERBB4	√	
MYH9	√	
IGF2BP3	√	
HLA-B	√	
NEDD4L	√	
LCP1	√	
SCD	√	
BAP1	√	
CEBPB	√	
PPARG	√	
NCOR2		via HDAC1
DMD	√	
VEGFA	√	
EGFR	√	
TCF7L2		via CTNNB1
NFKB1	√	
HRAS		via RAF1
RAN	√	
AKT1	√	
CCND1	√	
FN1		via VHL

Hence, because of the target specificity of ubiquitin and by reason of ubiquitin's general upregulation in renal tumours, the cancer-hallmark events would therefore be enhanced in malignant cells compared to normal cells, contributing to tumour progression. Frezza et al. (2011), therefore, discusses the ubiquitin-proteasome pathway's pivotal role in the upregulation of cell growth and in the downregulation of apoptosis (which is not further elaborated here).

In terms of therapy, since many of the RCC genes are directly associated with ubiquitin, as demonstrated in Figure 25, combinatorial drug therapy could be advantageous in combatting renal tumours. This approach has been shown to overcome the complexity of treating multifactorial diseases such as cancers (Frezza et al., 2011). One such example is in diabetes mellitus, another complex disease, where a combination of three **non-coding** SNPs was associated with increased risk in Mexican Americans, as discussed by Pritchard and Cox (2002) and were therefore suggested as a triad biomarker for the disease.

Moreover, the E3 ubiquitin ligase, which is the enzyme responsible for ubiquitin's keen substrate-specificity for its target molecule (Frezza et al., 2011), could also be used as a vector to target tumours by pursuing them in the opposite direction; that is, to target the specific protein/gene attached to ubiquitin without interfering with other ubiquitin-interacting molecules. For this reason, the various components of the ubiquitin-proteasome pathway, especially the ubiquitin-related targets, have emerged as crucial targets for novel anti-cancer therapy (Ande et al., 2009). Hence, analysing the protein-protein interactions surrounding ubiquitin could assist with the targeted design of more effective anti-cancer therapeutics.

Regarding the independent proteins, many of the proteins which appeared completely unassociated with other proteins in the network, were individually submitted to STRING-DB to observe their connections, and upon further investigation were also observed to be linked to UBC via an intermediate, as demonstrated in Figure 26. The two frequently mutated genes, DOCK2 and TRIO for example, interact with UBC via the RAC1 protein. Therefore, further expansion of this network could eventually interconnect most of these disease molecules to UBC via at least one intermediate.

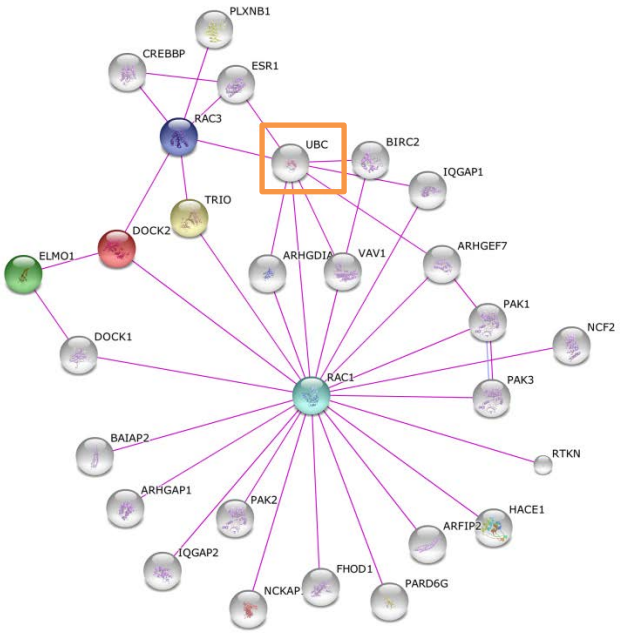


Figure 26: Despite not being associated with UBC in the larger network where all 57 genes were considered, DOCK2 (red) and TRIO (yellow) can be seen here as interplaying with UBC (**orange** square) via the intermediate protein RAC3 (**dark blue**) or RAC1 in the centre (**light blue**).

On the other hand, Table 13 shows that four of the genes that did not interplay with UBC, did interact with the epidermal growth factor receptor (EGFR). EGFR is involved in the progression of many cancer types, including RCC, representing an important therapeutic target (Minner et al., 2012). However, although many inhibitors have been designed to target EGFR, they have failed to produce objective responses (Dancey, 2004). The molecular heterogeneity of ccRCC has been proposed as one of the reasons for the low activity and thus, the relative resistance of ccRCC's **single-agent** EGFR inhibitors (Dancey, 2004). This again suggests that combinatorial treatment may be a more promising strategy for eradicating ccRCC tumours (Miles et al., 2014) (Cooper et al., 2012). This network analysis therefore validates that by analysing just the non-coding variants with the most frequent mutations, one could successfully select for the most appropriate targets by either directly targeting a specific gene or a combination of genes, or by targeting a hub gene/protein they all form connections with.

Table 13: The genes/proteins interacting directly with the epidermal growth factor receptor (EGFR) or via an intermediate gene/protein.

Protein/gene	Direct link	Via another gene/protein
HSPG2	√	
MET	√	
FGFR2		via PLCG1
PDGFRA		via CRK

However, even with non-coding variants, the allele frequency of a rSNP can vary between different ethnic or racial groups due to population bottlenecks (Buroker, 2014). This would influence the occurrence of TFBSs and the TFs regulating specific genes; thereby also impacting the susceptibility of certain populations to a disease. The final step in the analysis was to therefore, to observe and compare the allele frequencies of all non-coding somatic variants within the different super populations.

3.13. Allele Frequency of Variants in the African Population

For both the ccRCC **disease** variants and the **non-disease** variants, most of the alleles were not reported in the 1000Genomes data file, as shown in Table 14. However, finding such a small subset of AFs was unexpected. Thus, a random set of about 30 variants were manually checked using *grep 'variant position' 1000genomes_filename*, but they were still not found. To confirm their absence, the variants were submitted to BioMart using the Ensembl Variants dataset (hg19). When they were also not found, the scope of the range was extended on either end of the start- and end coordinate in order to ascertain whether other variants could be picked up around this region. For example, if the variant was at position 14: 10**5**254288, then the region that was verified was 14:10**4**254288 - 14:10**6**254288. Many SNPs were then flagged within this range, but the ccRCC variants were once again not found. The allele frequencies of these variants could therefore not be determined, but

it is likely that they do not occur frequently in populations, i.e. that they are very rare variants (present in <0.5% of the population).

Table 14: The total number of alleles for which the allele frequencies were obtained in the 1000Genomes dataset. The **blue** columns show the total non-coding and the CDS **ccRCC variants**, while the **green** columns display the total non-coding and the CDS **non-disease variants**. For all variants within their distinct categories, very few alleles were found.

RCC disease genes		Non-disease genes	
All **non-coding** SVs	**Non-coding** variants found in 1000genomes	All **non-coding** SVs	**Non-coding** variants found in 1000genomes
4385	73	244	5
All **CDS** SVs	**CDS** variants found in 1000genomes	All **CDS** SVs	**CDS** variants found in 1000genomes
154	3	10	0

It is also likely that these variants contribute to ccRCC, since they were located *within ccRCC disease genes*. Rare alleles are often more recent, due to not having been subjected for a prolonged period of time to the influences of purifying negative selection (Soumya, 2013). Purifying selection aims to remove modifications with functional consequences from a population (Elyashiv et al., 2010). Hence, rare variants are expected to be relatively deleterious mutations (Soumya, 2013) and thus interesting for this study. Alternatively, these may also be novel individual-specific variants generated due to the *natural mutation rate* in *Homo sapiens* or they may be spontaneous variants as a result of increased *genomic instability* within the tumours of these donors. Nevertheless, discovering very rare/novel variants are certainly very

promising, because they may provide pioneering insights into the underlying mechanisms of a disease and hence, possible new targets for the diagnosis, prognosis and treatment of a disease. However, the allelic variants that were present in the 1000Genomes dataset and therefore, more common, were also specifically investigated, since **common** variants that **do** contribute to disease are more likely to act on the **non-coding genome** (Soumya, 2013).

3.13.1. No distinction in AF between Africans and other super populations

When the alleles were compared across the macro-populations, of the 73 non-coding variants that WERE present in the 1000genomes dataset, 11 alleles did not show any distinctly different frequencies in Africans compared to the other population groups. That is, they were either the same as another super population or the AF was somewhere in the middle range of the extremes of the other population groups. It is therefore unlikely that these AF differences may have a large effect on the differences in ccRCC incidence in Africans compared to other populations.

3.13.2. Higher AF in Africans compared to other super populations

For 28 alleles, the AFs were higher in Africans and are therefore likely to be implicated in ccRCC when they are observed in Africans. It has been proposed that in genetically complex diseases like cancers, weakly penetrant, high-risk alleles may be present at higher frequency (>1%) in the population (Hirschhorn et al., 2002). Higher frequency or common alleles tend to be ancient, because they have survived the effects of negative purifying selection and are likely to have a small to modest functional effect of the disease phenotype. They often act very subtly to cause disease without sabotaging their evolutionary fitness. That is, while they may confer a high risk for a one disease, they may also protect the individual from another disease (Soumya, 2013) and are thereby able to ‘survive’ the selection pressure.

Most of the SNPs with comparatively higher AFs in Africans were just slightly more frequent in Africans and so are not individually discussed. However, a number of these AFs were much more common in this super population compared to other population groups and they are the ones being discussed in more detail below.

The most common of these SNPs, rs261597 (G->A), in the Dedicator of cytokinesis 2 gene (**DOCK2)**, for example, is found in 76% (AF = 0.76) of Africans, although this particular SNP is rather common across populations, being present in between 26% and 46% of other super populations. This DOCK2 gene is interesting, since it was in the top 20 genes with the most total and deleterious SVs, the top 20 genes with the most CDS mutations and the top 20 genes with the most possible TFBS disruptions. DOCK2 functions in mitosis and has been shown to be **induced in RCC** (Lenburg et al., 2003). Although this variant may contribute to ccRCC, it is unlikely that it is highly penetrant; otherwise we would see a lot more ccRCC cases across super populations. However, it could be one of a subset of variants required for the disease onset, since DOCK2 modifications **are** ccRCC-implicated.

In contrast, the rs116331317 SNP (A->C) in the cubilin gene (**CUBN)** occurred in 53% of Africans, but it was only detected in between 2% and 10% of other population groups. CUBN is found mainly in the proximal tubules of the kidneys (the primary site for ccRCC development) and is responsible for vitamin B1 uptake, which is essential for DNA and protein formation (Maria Aminoff, 1999). This gene was also one of the top 20 genes with the most SVs, deleterious SV, total TFBS disruptions and the most potentially deleterious TFBS disruptions.

Similarly, the rs61415991 SNP (T->G) in the pre-mRNA processing factor 8 gene (**PRPF8)** occurs in 41% of Africans, but only in between 2% and 8% of other populations. This gene, however, never featured in any of the top 20 categories, which is interesting considering its high allele frequency in the general population. In addition, not much literature was found linking this gene to ccRCC. Therefore, except for its selection from databases, this gene doesn't appear to be somatically mutated very often in ccRCC (based on literature), so there must be another mechanism at work to link this gene to ccRCC in a subset of patients. However, this

also suggests that mutations in this gene might be under a stronger purifying selection than DOCK2 and CUBN and hence, more detrimental/penetrant.

Lastly, the rs79266366 SNP (G->T) in the microphthalmia-associated transcription factor gene (**MITF)** occurs in 14% of Africans, but only occurs in 3%-8% of other population groups. This gene was also in the top 20 genes with the highest number of possible TFBS disruptions and has additionally been causally implicated in paediatric RCC (Fall et al., 2011).

Identifying common variants in complex disease studies is interesting for a second reason. It has been suggested that these common variant associations may be as a consequence of **undiscovered rare** variants with dramatic functional consequences, which are present at the same locus (Dickson et al., 2010). For example, the allele of the **MECOM** SNP, rs1918961 (C->T), is a *common* allele variant (MAF >5%) in the Asian, American, African and European super populations. However, this common allele also has 138 rare/novel variants (i.e., they were not found in 1000Genomes) in close proximity to it, and these rarer variants may be the variants actually contributing to tumourigenesis, or rather, they may have the greater functional effect on ccRCC.

3.13.3. Lower AF in Africans compared to other super populations

Eleven of the non-coding variants had a lower AF in Africans compared to the other super populations - three of each in the protein tyrosine phosphatase receptor type delta (PTPRD) gene. This gene was previously shown to be the most frequently mutated of all genes. Focal deletions are often seen in the 9p23 arm of PTPRD in many ccRCC tumours (The Cancer Genome Atlas Research Network, 2013). These 9p23 SNPs have been linked to several other cancers, such as cutaneous squamous cell sarcomas and non-small cell lung cancer (Hendriks and Pulido, 2013) (Purdie et al., 2007). However, Du et al. (2013) also described a SNP in PTPRD as a genetic risk factor for developing ccRCC, but since these PTPRD SNPs are infrequent in Africans, they are more likely to result in ccRCC when seen in other populations.

Alternatively, mutations in this gene could be kept to a minimum in Africans, because it may be more deleterious in this populations group. Concerning other populations, because the somatic mutation patients in this study were of European/French ancestry, this may be a promising gene to investigate as a risk factor/biomarker for ccRCC in this population group.

3.13.4. Alleles not found in Africans

Lastly, a total of 15 alleles was present in other population groups and NOT in Africans. These SNPs could represent a locus that has a protective function in Africans, due to being under a strong purifying selection, although the theory can only be tenable with a proper study design that evaluates these loci within Africans with and without ccRCC.

3.13.5. Alleles only in Africans

Lastly, nine alleles (mostly in the NAV3 gene) were distinctly unique to the African population. The rs183605535 SNP was found in 1% of Africans, the rs183605535 SNP in 7% of the population, while the particularly **rare SNP**, rs112941962, was present in 0.2% of Africans (i.e. <1%). Similarly, the ANK3 SNP, rs150147334 (C->T), was also present in <1% of this super population, while the RYR3, RUNX1, PTPRD, FGFR2 and MECOM SNPs, were all low frequency SNPS common to 1% of Africans. SNP rs150147334 in ANK3 has been shown to potentially disrupt multiple TFBSs and this gene incurred 100 non-coding somatic mutations in 60/95 patients. So although this **allele** is very rare, this gene is often targeted in ccRCC tumours. Therefore, DNA repair enzymes might be activated to keep the total number of mutations low, by correcting mismatched bases at distinctly deleterious locations. Nevertheless, all of these genes were within one or more categories of the top 20 most mutated genes or top 20 genes with the most TFBS disruptions. The runt-related transcription factor gene (RUNX1) is of particular interest, as it was present in all

categories of the top 20 most mutated genes and TFBS disruptions. Additionally, its promoters were frequently differentially methylated, with subsequent gene dysregulation in many patients. Since its variant's AF is quite low across all super populations, despite being quite a common target for non-coding somatic mutations and regulatory and epigenetic modifications, it could serve as an important molecular marker for ccRCC diagnosis or prognosis and possibly also for its treatment. Its direct link to UBC could also make targeting this molecule very achievable (targeting proteins/genes linked to UBC was explained earlier in section STRING-DB protein-protein interactions 3.12).

However, the cancer patients used in this study were not Africans, creating ascertainment bias due to the greater genetic variation in Africans compared to the rest-of-world populations. Since no conclusive supporting data was available, no strong associations could therefore be made about the relationship between allele frequencies and their potential risk or contribution to disease in Africans.

3.13.6. Biomarker implications

The fact that these variants were **somatic** mutations (present in the tumour and not in the normal tissue), means that they are most likely contributing to the ccRCC phenotype. Again, it is difficult to determine whether most of them accumulated as a result of the malfunctioning or inactivation of pathways/genes after the cancer developed (i.e. whether these variants were passengers) or whether they were causally implicated in ccRCC (i.e. drivers of the disease).

Nevertheless, given that all of the patients had **multiple** non-coding variants, and assuming that the variants did not accrue as a general result of the disease, then it is plausible that a large number of SNPs **cumulatively** contribute to ccRCC in each individual. Unlike Mendelian disorders that have high-penetrance mutations (the disease-specific allele directly expresses the disease phenotype) and are under strong purifying selection, disease-susceptibility variants in complex diseases are known to

have low to medium penetrance and are under weaker selection (Pritchard and Cox, 2002).

Therefore, it is possible that the three SNPs that are very common in Africans, compared to other super populations, the SNPs that were distinctly found **only** in Africans and even those found less frequently in Africans, represent an opportunity for the design of a genetic signature for disease susceptibility/risk in Africans.

In summary non-coding variants and their relationships with TFBS disruptions, DNA methylation patterns and gene dysregulation, show promise for a better understanding of the mechanistic architecture of ccRCC and potentially other non-Mendelian diseases. This also provides a gateway for the selection of the most relevant genes/variants to simplify a topological network analysis for the possible design of cancer therapies. Lastly, the AFs of the variants may help us to better identify which individuals are truly at risk for developing ccRCC in different population groups for a pharmacogenomics approach to treating and possibly eradicating the disease.

CHAPTER 4

4. CONCLUSION AND FUTURE WORK

4.11. Summary of the findings

As hypothesized, most of the somatic variants specific to tumours were found in the intronic regions of the genes of interest. Only a small subset of these variants were predicted to be detrimental by RegulomDB, but the variants that fell within the promoter regions were generally more likely to have a functional effect- understandably so, since the promoter region is crucial for transcription initiation. When the positions of these variants in relation to TFBSs were observed, only a few of the variants within the CDS regions, termed duons, were located within TFBSs. However, all of non-coding variants were preferentially located within the TFBSs. More than 60% of patients enrolled in the DNA methylation study also had epigenetic modifications in the form of differential DNA methylation. More than 97% of these aberrant methylation events were hypomethylation of the gene promoters with a subsequent upregulation of the affected gene, indicating that the activation of proto-oncogenes could potentially be a mechanism of action for DNA methylation changes in ccRCC. There was substantial overlap between the genes with the most non-coding variants, the genes linked to variants often located within TFBS and those often hypomethylated at promoter regions. Many of the genes often targeted were also directly linked to each other in a protein-protein interaction network.

4.12. Novel Findings of this study

An interesting observation was that the genes most often dysregulated and those frequently associated with the non-coding mutations, were not the usual RCC genes that are often highlighted in whole exome sequencing and other studies that focus only on the protein-coding variants. Hence, identifying and examining the variants in the non-coding regions of disease genes may offer

substantial insight into the underlying mechanisms of ccRCC oncogenesis. When the protein-protein interactions of a selection of these genes were observed, many more genes/proteins were directly associated with ubiquitin C, which represents many new potential targets for anti-cancer therapeutics. Also, for most of the variants the AFs could not be determined due to their absence from 1000genomes data, which suggests that many novel/very rare variants are generated within non-coding regions of ccRCC tumours, possibly as a result of genomic instability. Albeit, this is promising, since detecting *de novo* or very rare variants may hold the key to identify genetic regions with unknown disease mechanisms, which may ultimately lead to genetic screening to identify individuals at risk or to the development of more efficient therapeutics to target the disease.

4.13. Limitations of this study

However, it should be noted that even in this study many of the potentially significant and novel variants may have been missed since the entire intergenic region were not taken into account. Identifying variants in the intergenic regions may also lead to the discovery of variants affecting the enhancer regions, since these regions are usually located far from the genes they regulate.

In terms of other epigenetic modifications, it has been shown that there is a better correlation between chromatin states and gene expression levels than with differential methylation or non-coding somatic mutations and their associated gene expression levels. Including this data in this study may give rise to more clues about the underlying mechanisms that influence all stages of the disease.

In addition, this was also a small sample size and therefore only constitutes a pilot study. On the issue of bias; in future it would be best to extract the complete set of all non-coding somatic variants in a WGS ccRCC study, in order to avoid omitting variants of interests that may be present in genes not included in a predetermined disease-associated gene list. Using a list of

'known' disease genes will inevitably bias the study towards existing knowledge of the disease mechanisms.

It is also better to use the same tissue type as the normal/control tissue counterpart, as opposed to using blood.

Also, most whole genome sequencing is performed at 30-60x depth of coverage, which may not be sufficient to identify rare alleles with a high level of confidence. Tumours can constitute several distinct genetic subclones for which the AF of somatic mutations in small subclones may be very low. A sequencing depth of 1000X may therefore be required to accurately identify variants present in only 1% of the sample (Stead et al., 2013). Therefore identifying a dataset sequenced at greater depth may increase the discovery of very rare/novel variants with a functional impact on the disease.

Finally, it would be worthwhile to extract the somatic mutations for the individuals for which complete clinical data is available and to cross compare the results with the TFBS disruptions and methylations statuses of all of those deceased and alive. This could give a truer picture of the genomic profile of those who are more severely affected than others.

4.14. Future directions

This is a pilot study that indicates some interesting avenues for further research, particularly:

a) That with ENCODE genome annotation now available, analysis of the non-coding regions can give profound insights into the contribution of non-CDS to disease aetiology
b) That relationships appear to exist between non-coding somatic variants, differential methylation of promoters, TFBSs disruptions and gene dysregulation in tumours – which suggest that these should also be explored in future studies of disease mechanisms
c) That genes identified as disease-associated through CDS analysis are not always the same as those identified through non-coding region analysis,

and extending research to analyse whole genome sequence may identify new disease associated genes/variants

d) That observing the protein-protein interactions of flagged genes may allow us to identify new or a combination of targets for the production of less resistant tumour inhibitors

Despite the limitations associated with this study, the outcomes have strongly suggested that the non-coding portions of the genome may have a dynamic impact on disease development and progression, as a result of the numerous processes they affect when even a single position is mutated; and that these regions therefore deserve consideration within the context of genetic studies of variants that may underlie complex diseases.

REFERENCES

American Cancer Society, 2014. URL http://www.cancer.org/cancer/kidneycancer/detailedguide/kidney-cancer-adult-survival-rates (accessed 10.28.15).

Ande, S.R., Chen, J., Maddika, S., 2009. The ubiquitin pathway: an emerging drug target in cancer therapy. Eur. J. Pharmacol. 625, 199–205. doi:10.1016/j.ejphar.2009.08.042

Barrett, L.W., Fletcher, S., Wilton, S.D., 2013. Untranslated Gene Regions and Other Non-coding Elements.

Barrett, L.W., Fletcher, S., Wilton, S.D., 2013. Untranslated Gene Regions and Other Non-coding Elements: Regulation of Eukaryotic Gene Expression. Springer Science & Business Media.

Baylin, S.B., Esteller, M., Rountree, M.R., Bachman, K.E., Schuebel, K., Herman, J.G., 2001. Aberrant patterns of DNA methylation, chromatin formation and gene expression in cancer. Hum. Mol. Genet. 10, 687–692. doi:10.1093/hmg/10.7.687

Bor, J., Herbst, A.J., Newell, M.-L., Barnighausen, T., 2013. Increases in adult life expectancy in rural South Africa: valuing the scale-up of HIV treatment. Science 339. doi:10.1126/science.1230413

Boyle, A.P., Hong, E.L., Hariharan, M., Cheng, Y., Schaub, M.A., Kasowski, M., Karczewski, K.J., Park, J., Hitz, B.C., Weng, S., Cherry, J.M., Snyder, M., 2012. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. 22, 1790–1797. doi:10.1101/gr.137323.112

Buroker, N.E., 2014. Regulatory SNPs and transcriptional factor binding sites in ADRBK1, AKT3, ATF3, DIO2, TBXA2R and VEGFA. Transcription 5, e964559. doi:10.4161/21541264.2014.964559

Buske, O.J., Manickaraj, A., Mital, S., Ray, P.N., Brudno, M., 2013. Identification of deleterious synonymous variants in human genomes. Bioinformatics 29, 1843–1850. doi:10.1093/bioinformatics/btt308

Campbell, M.C., Tishkoff, S.A., 2008. AFRICAN GENETIC DIVERSITY: Implications for Human Demographic History, Modern Human Origins, and Complex Disease Mapping. Annu Rev Genomics Hum Genet 9, 403–433. doi:10.1146/annurev.genom.9.081307.164258

Cavalcanti, E., Gigante, M., Mancini, V., Battaglia, M., Ditonno, P., Capobianco, C., Cincione, R.I., Selvaggi, F.P., Herr, W., Storkus, W.J., Gesualdo, L., Ranieri, E., Cavalcanti, E., Gigante, M., Mancini, V., Battaglia, M., Ditonno, P., Capobianco, C., Cincione, R.I., Selvaggi, F.P., Herr, W., Storkus, W.J., Gesualdo, L., Ranieri, E., 2010. JAK3/STAT5/6 Pathway Alterations Are Associated with Immune Deviation in T Cells in Renal Cell Carcinoma Patients, JAK3/STAT5/6 Pathway Alterations Are Associated with Immune Deviation in T Cells in Renal Cell Carcinoma Patients. BioMed Research International, BioMed Research International 2010, 2010, e935764. doi:10.1155/2010/935764, 10.1155/2010/935764

Chatterjee, S., Berwal, S.K., Pal, J.K., 2001. Pathological Mutations in 5' Untranslated Regions of Human Genes, in: eLS. John Wiley & Sons, Ltd.

Chau, B.N., Wang, J.Y.J., 2003. Coordinated regulation of life and death by RB. Nature Reviews Cancer 3, 130–138. doi:10.1038/nrc993

Chen, P.-Y., Feng, S., Joo, J.W.J., Jacobsen, S.E., Pellegrini, M., 2011. A comparative analysis of DNA methylation across human embryonic stem cell lines. Genome Biol 12, R62. doi:10.1186/gb-2011-12-7-r62

Chow, W.-H., Dong, L.M., Devesa, S.S., 2010. Epidemiology and risk factors for kidney cancer. Nat Rev Urol 7, 245–257. doi:10.1038/nrurol.2010.46

Collet, Théoleyre, Rageul, Mottier, Jouan, Rioux-Leclercq, Fergelot, Patard, Masson, Denis, 2011. PPARγ is functionally expressed in clear cell renal cell carcinoma. International Journal of Oncology 38, 851–857. doi:10.3892/ijo.2010.891

Corn, P.G., 2007. Role of the ubiquitin proteasome system in renal cell carcinoma. BMC Biochemistry 8, S4. doi:10.1186/1471-2091-8-S1-S4

Costa, V., Aprile, M., Esposito, R., Ciccodicola, A., 2013. RNA-Seq and human complex diseases: recent accomplishments and future perspectives. Eur J Hum Genet 21, 134–142. doi:10.1038/ejhg.2012.129

Cross, D.S., Ivacic, L.C., Stefanski, E.L., McCarty, C.A., 2010. Population based allele frequencies of disease associated polymorphisms in the Personalized Medicine Research Project. BMC Genetics 11, 51. doi:10.1186/1471-2156-11-51

Dall'Oglio, M.F., Coelho, R., Lopes, R., Antunes, A.A., Crippa, A., Camara, C., Leite, K.R.M., Srougi, M., 2011. Significant heterogeneity in terms of diagnosis and treatment of renal cell carcinoma at a private and public hospital in Brazil. International braz j urol 37, 584–590. doi:10.1590/S1677-55382011000500003

Dancey, J.E., 2004. Epidermal Growth Factor Receptor and Epidermal Growth Factor Receptor Therapies in Renal Cell Carcinoma: Do We Need a Better Mouse Trap? JCO 22, 2975–2977. doi:10.1200/JCO.2004.04.934

De Gobbi, M., Viprakasit, V., Hughes, J.R., Fisher, C., Buckle, V.J., Ayyub, H., Gibbons, R.J., Vernimmen, D., Yoshinaga, Y., de Jong, P., Cheng, J.-F., Rubin, E.M., Wood, W.G., Bowden, D., Higgs, D.R., 2006. A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. Science 312, 1215–1217. doi:10.1126/science.1126431

Dickson, S.P., Wang, K., Krantz, I., Hakonarson, H., Goldstein, D.B., 2010. Rare variants create synthetic genome-wide associations. PLoS Biol. 8, e1000294. doi:10.1371/journal.pbio.1000294

Diehl, J.A., Fuchs, S.Y., Haines, D.S., 2010. Ubiquitin and Cancer New Discussions for a New Journal. Genes & Cancer 1, 679–680. doi:10.1177/1947601910383565

Du, P., Zhang, X., Huang, C.-C., Jafari, N., Kibbe, W.A., Hou, L., Lin, S.M., 2010. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics 11, 587. doi:10.1186/1471-2105-11-587

Du, Y., Su, T., Tan, X., Li, X., Xie, J., Wang, G., Shen, J., Hou, J., Cao, G., 2013. Polymorphism in protein tyrosine phosphatase receptor delta is associated with the risk of clear cell renal cell carcinoma. Gene 512, 64–69. doi:10.1016/j.gene.2012.09.094

Easton, D.F., Eeles, R.A., 2008. Genome-wide association studies in cancer. Human Molecular Genetics 17, R109–R115. doi:10.1093/hmg/ddn287

Ehrlich, M., 2002. DNA methylation in cancer: too much, but also too little. Oncogene 21, 5400–5413. doi:10.1038/sj.onc.1205651

Elyashiv, E., Bullaughey, K., Sattath, S., Rinott, Y., Przeworski, M., Sella, G., 2010. Shifts in the intensity of purifying selection: An analysis of genome-wide polymorphism data from two closely related yeast species. Genome Res 20, 1558–1573. doi:10.1101/gr.108993.110

Faa', V., Coiana, A., Incani, F., Costantino, L., Cao, A., Rosatelli, M.C., 2010. A Synonymous Mutation in the CFTR Gene Causes Aberrant Splicing in an Italian Patient Affected by a Mild Form of Cystic Fibrosis. J Mol Diagn 12, 380–383. doi:10.2353/jmoldx.2010.090126

Fall, B., Sarr, A., Sow, Y., Diao, B., 2011. Renal cell carcinoma with MiTF/TFE3 translocation in children: report of a case at the stage of lymph node involvement. Afr J Paediatr Surg 8, 317–319. doi:10.4103/0189-6725.91669

Faustino, N.A., Cooper, T.A., 2003. Pre-mRNA splicing and human disease. Genes Dev. 17, 419–437. doi:10.1101/gad.1048803

Fiorentino, F.P., Giordano, A., 2012. The tumor suppressor role of CTCF. J. Cell. Physiol. 227, 479–492. doi:10.1002/jcp.22780

Fischer, O.M., Gschwind, A., Ullrich, A.A., 2009. Cell Surface Growth Factor Receptor Molecules as Targets for Cancer Therapy. Discovery Medicine 4, 166–171.

Forbes, S.A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., Bamford, S., Cole, C., Ward, S., Kok, C.Y., Jia, M., De, T., Teague, J.W., Stratton, M.R., McDermott, U., Campbell, P.J., 2015. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucl. Acids Res. 43, D805–D811. doi:10.1093/nar/gku1075

Forbes, S.A., Bindal, N., Bamford, S., Cole, C., Kok, C.Y., Beare, D., Jia, M., Shepherd, R., Leung, K., Menzies, A., Teague, J.W., Campbell, P.J., Stratton, M.R., Futreal, P.A., 2011. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucl. Acids Res. 39, D945–D950. doi:10.1093/nar/gkq929

Fredriksson, N.J., Ny, L., Nilsson, J.A., Larsson, E., 2014. Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. Nat Genet 46, 1258–1263. doi:10.1038/ng.3141

Frezza, M., Schmitt, S., Dou, Q.P., 2011. Targeting the ubiquitin-proteasome pathway: an emerging concept in cancer therapy. Curr Top Med Chem 11, 2888–2905.

Fujita, J., Kraus, M.H., Onoue, H., Srivastava, S.K., Ebi, Y., Kitamura, Y., Rhim, J.S., 1988. Activated H-ras oncogenes in human kidney tumors. Cancer Res. 48, 5251–5255.

Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., Stratton, M.R., 2004. A census of human cancer genes. Nat Rev Cancer 4, 177–183. doi:10.1038/nrc1299

Fu, Y., Liu, Z., Lou, S., Bedford, J., Mu, X.J., Yip, K.Y., Khurana, E., Gerstein, M., 2014. FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. Genome Biology 15, 480. doi:10.1186/s13059-014-0480-5

Gaffney, D.J., Keightley, P.D., 2006. Genomic Selective Constraints in Murid Noncoding DNA. PLoS Genet 2, e204. doi:10.1371/journal.pgen.0020204

Gama-Sosa, M.A., Slagel, V.A., Trewyn, R.W., Oxenhandler, R., Kuo, K.C., Gehrke, C.W., Ehrlich, M., 1983. The 5-methylcytosine content of DNA from human tumors. Nucleic Acids Res. 11, 6883–6894.

GLOBOCAN 2012 [WWW Document], n.d. URL http://globocan.iarc.fr/old/bar_sex_site.asp?selection=10210&title=Kidney&statistic=2&populations=6&window=1&grid=1&info=1&color1=5&color1e=&color2=4&color2e=&submit=%C2%A0Execute%C2%A0 (accessed 8.26.15).

Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Tamborero, D., Schroeder, M.P., Jene-Sanz, A., Santos, A., Lopez-Bigas, N., 2013. IntOGen-mutations identifies cancer drivers across tumor types. Nat Meth 10, 1081–1082. doi:10.1038/nmeth.2642

Grigoryev, Y., 2013. What is Alternative Splicing, and Why is it Important? Bitesize Bio.

Guo, G., Gui, Y., Gao, S., Tang, A., Hu, X., Huang, Y., Jia, W., Li, Z., He, M., Sun, L., Song, P., Sun, X., Zhao, X., Yang, S., Liang, C., Wan, S., Zhou, F., Chen, C., Zhu, J., Li, X., Jian,

M., Zhou, L., Ye, R., Huang, P., Chen, J., Jiang, T., Liu, X., Wang, Y., Zou, J., Jiang, Z., Wu, R., Wu, S., Fan, F., Zhang, Z., Liu, L., Yang, R., Liu, X., Wu, H., Yin, W., Zhao, X., Liu, Y., Peng, H., Jiang, B., Feng, Q., Li, C., Xie, J., Lu, J., Kristiansen, K., Li, Y., Zhang, X., Li, S., Wang, J., Yang, H., Cai, Z., Wang, J., 2012. Frequent mutations of genes encoding ubiquitin-mediated proteolysis pathway components in clear cell renal cell carcinoma. Nat Genet 44, 17–19. doi:10.1038/ng.1014

Guo, H., German, P., Bai, S., Barnes, S., Guo, W., Qi, X., Lou, H., Liang, J., Jonasch, E., Mills, G.B., Ding, Z., 2015. The PI3K/AKT Pathway and Renal Cell Carcinoma. Journal of Genetics and Genomics 42, 343–353. doi:10.1016/j.jgg.2015.03.003

Hakimi, A.A., Pham, C.G., Hsieh, J.J., 2013. A clear picture of renal cell carcinoma. Nat Genet 45, 849–850. doi:10.1038/ng.2708

Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A., McKusick, V.A., 2005. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 33, D514–D517. doi:10.1093/nar/gki033

Havranek, E., Howell, W.M., Fussell, H.M., Whelan, J.A., Whelan, M.A., Pandha, H.S., 2005. An interleukin-10 promoter polymorphism may influence tumor development in renal cell carcinoma. J. Urol. 173, 709–712. doi:10.1097/01.ju.0000152493.86001.91

Heltemes-Harris, L.M., Willette, M.J.L., Ramsey, L.B., Qiu, Y.H., Neeley, E.S., Zhang, N., Thomas, D.A., Koeuth, T., Baechler, E.C., Kornblau, S.M., Farrar, M.A., 2011. Ebf1 or Pax5 haploinsufficiency synergizes with STAT5 activation to initiate acute lymphoblastic leukemia. J Exp Med 208, 1135–1149. doi:10.1084/jem.20101947

Hendriks, W.J.A.J., Pulido, R., 2013. Protein tyrosine phosphatase variants in human hereditary disorders and disease susceptibilities. Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease 1832, 1673–1696. doi:10.1016/j.bbadis.2013.05.022

Heyn, H., Moran, S., Hernando-Herraez, I., Sayols, S., Gomez, A., Sandoval, J., Monk, D., Hata, K., Marques-Bonet, T., Wang, L., Esteller, M., 2013. DNA methylation contributes to natural human variation. Genome Res 23, 1363–1372. doi:10.1101/gr.154187.112

Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., Manolio, T.A., 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. PNAS 106, 9362–9367. doi:10.1073/pnas.0903103106

Hirata, H., Naito, K., Yoshihiro, S., Matsuyama, H., Suehiro, Y., Hinoda, Y., 2003. A single nucleotide polymorphism in the matrix metalloproteinase-1 promoter is associated with conventional renal cell carcinoma. Int. J. Cancer 106, 372–374. doi:10.1002/ijc.11229

Hirschhorn, J.N., Lohmueller, K., Byrne, E., Hirschhorn, K., 2002. A comprehensive review of genetic association studies. Genet Med 4, 45–61. doi:10.1097/00125817-200203000-00002

Hoesel, B., Schmid, J.A., 2013. The complexity of NF-κB signaling in inflammation and cancer. Molecular Cancer 12, 86. doi:10.1186/1476-4598-12-86

Holloway, D.T., Kon, M., DeLisi, C., 2008. In silico regulatory analysis for exploring human disease progression. Biology Direct 3, 24. doi:10.1186/1745-6150-3-24

Horiuchi, T., Aigaki, T., 2006. Alternative trans-splicing: a novel mode of pre-mRNA processing. Biology of the Cell 98, 135–140. doi:10.1042/BC20050002

Huang, F.W., Hodis, E., Xu, M.J., Kryukov, G.V., Chin, L., Garraway, L.A., 2013. Highly Recurrent TERT Promoter Mutations in Human Melanoma. Science 339, 957–959. doi:10.1126/science.1229259

Ibragimova, I., Slifker, M.J., Maradeo, M.E., Banumathy, G., Dulaimi, E., Uzzo, R.G., Cairns, P., 2013. Genome-Wide Promoter Methylome of Small Renal Masses. PLoS ONE 8, e77309. doi:10.1371/journal.pone.0077309

Ishizawa, J., Yoshida, S., Oya, M., Mizuno, R., Shinojima, T., Marumo, K., Murai, M., 2004. Inhibition of the ubiquitin-proteasome pathway activates stress kinases and induces apoptosis in renal cancer cells. Int. J. Oncol. 25, 697–702.

Ivashchenko, A.T., Khailenko, V.A., Atambaeva, S.A., 2009. Variations of the length of exons and introns in human genome genes. Russ J Genet 45, 16–22. doi:10.1134/S1022795409010025

Kanayama, H., Tanaka, K., Aki, M., Kagawa, S., Miyaji, H., Satoh, M., Okada, F., Sato, S., Shimbara, N., Ichihara, A., 1991. Changes in expressions of proteasome and ubiquitin genes in human renal cancer cells. Cancer Res. 51, 6677–6685.

Kasowski, M., Grubert, F., Heffelfinger, C., Hariharan, M., Asabere, A., Waszak, S.M., Habegger, L., Rozowsky, J., Shi, M., Urban, A.E., Hong, M.-Y., Karczewski, K.J., Huber, W., Weissman, S.M., Gerstein, M.B., Korbel, J.O., Snyder, M., 2010. Variation in Transcription Factor Binding Among Humans. Science 328, 232–235. doi:10.1126/science.1183621

Kasowski, M., Kyriazopoulou-Panagiotopoulou, S., Grubert, F., Zaugg, J.B., Kundaje, A., Liu, Y., Boyle, A.P., Zhang, Q.C., Zakharia, F., Spacek, D.V., Li, J., Xie, D., Olarerin-George, A., Steinmetz, L.M., Hogenesch, J.B., Kellis, M., Batzoglou, S., Snyder, M., 2013. Extensive Variation in Chromatin States Across Humans. Science 342, 750–752. doi:10.1126/science.1242510

Kemp, C.J., Moore, J.M., Moser, R., Bernard, B., Teater, M., Smith, L.E., Rabaia, N., Gurley, K.E., Guinney, J., Busch, S.E., Shaknovich, R., Lobanenkov, V.V., Liggitt, D., Shmulevich, I., Melnick, A., Filippova, G.N., 2014. CTCF haploinsufficiency destabilizes DNA methylation and predisposes to cancer. Cell Rep 7, 1020–1029. doi:10.1016/j.celrep.2014.04.004

Kim, J., Kim, S., 2014. *In silico* Identification of *SFRP1* as a Hypermethylated Gene in Colorectal Cancers. Genomics & Informatics 12, 171. doi:10.5808/GI.2014.12.4.171

Laerd Statistics, 2013. Mann-Whitney U Test in SPSS Statistics | Setup, Procedure & Interpretation | Laerd Statistics [WWW Document]. Mann-Whitney U Test using SPSS Statistics. URL https://statistics.laerd.com/spss-tutorials/mann-whitney-u-test-using-spss-statistics.php (accessed 11.18.15).

Laurila, K., Lähdesmäki, H., 2009. Systematic analysis of disease-related regulatory mutation classes reveals distinct effects on transcription factor binding. In Silico Biol. (Gedrukt) 9, 209–224.

Lenburg, M.E., Liou, L.S., Gerry, N.P., Frampton, G.M., Cohen, H.T., Christman, M.F., 2003. Previously unidentified changes in renal cell carcinoma gene expression identified by parametric analysis of microarray data. BMC Cancer 3, 31. doi:10.1186/1471-2407-3-31

Ley, T.J., Mardis, E.R., Ding, L., Fulton, B., McLellan, M.D., Chen, K., Dooling, D., Dunford-Shore, B.H., McGrath, S., Hickenbotham, M., Cook, L., Abbott, R., Larson, D.E., Koboldt, D.C., Pohl, C., Smith, S., Hawkins, A., Abbott, S., Locke, D., Hillier, L.W., Miner, T., Fulton, L., Magrini, V., Wylie, T., Glasscock, J., Conyers, J., Sander, N., Shi, X., Osborne, J.R., Minx, P., Gordon, D., Chinwalla, A., Zhao, Y., Ries, R.E., Payton, J.E., Westervelt, P., Tomasson, M.H., Watson, M., Baty, J., Ivanovich, J., Heath, S., Shannon, W.D., Nagarajan, R., Walter, M.J., Link, D.C., Graubert, T.A., DiPersio, J.F., Wilson, R.K., 2008. DNA sequencing of a cytogenetically normal acute myeloid leukemia genome. Nature 456, 66–72. doi:10.1038/nature07485

Li, G., Pan, T., Guo, D., Li, L.-C., 2014. Regulatory Variants and Disease: The E-Cadherin -160C/A SNP as an Example, Molecular Biology International 2014, e967565. doi:10.1155/2014/967565

Linehan, W.M., 2012. Genetic basis of kidney cancer: Role of genomics for the development of disease-based therapeutics. Genome Res 22, 2089–2100. doi:10.1101/gr.131110.111

Linehan, W.M., Vasselli, J., Srinivasan, R., Walther, M.M., Merino, M., Choyke, P., Vocke, C., Schmidt, L., Isaacs, J.S., Glenn, G., Toro, J., Zbar, B., Bottaro, D., Neckers, L., 2004. Genetic Basis of Cancer of the Kidney Disease-Specific Approaches to Therapy. Clin Cancer Res 10, 6282S–6289S. doi:10.1158/1078-0432.CCR-050013

Ljungberg, B., Campbell, S.C., Cho, H.Y., Jacqmin, D., Lee, J.E., Weikert, S., Kiemeney, L.A., 2011. The Epidemiology of Renal Cell Carcinoma. European Urology 60, 615–621. doi:10.1016/j.eururo.2011.06.049

Lodish, H., Berk, A., Zipursky, S.L., Matsudaira, P., Baltimore, D., Darnell, J., 2000a. Proto-Oncogenes and Tumor-Suppressor Genes.

Lodish, H., Berk, A., Zipursky, S.L., Matsudaira, P., Baltimore, D., Darnell, J., 2000b. Tumor Cells and the Onset of Cancer.

Loeb, L.A., Loeb, K.R., Anderson, J.P., 2003. Multiple mutations and cancer. Proc Natl Acad Sci U S A 100, 776–781. doi:10.1073/pnas.0334858100

Luu, P.-L., Schöler, H.R., Araúzo-Bravo, M.J., 2013. Disclosing the crosstalk among DNA methylation, transcription factors, and histone marks in human pluripotent cells through discovery of DNA methylation motifs. Genome Res. doi:10.1101/gr.155960.113

Macintyre, G., Jimeno Yepes, A., Ong, C.S., Verspoor, K., 2014. Associating disease-related genetic variants in intergenic regions to the genes they impact. PeerJ 2. doi:10.7717/peerj.639

Mardis, E.R., Ding, L., Dooling, D.J., Larson, D.E., McLellan, M.D., Chen, K., Koboldt, D.C., Fulton, R.S., Delehaunty, K.D., McGrath, S.D., Fulton, L.A., Locke, D.P., Magrini, V.J., Abbott, R.M., Vickery, T.L., Reed, J.S., Robinson, J.S., Wylie, T., Smith, S.M., Carmichael, L., Eldred, J.M., Harris, C.C., Walker, J., Peck, J.B., Du, F., Dukes, A.F., Sanderson, G.E., Brummett, A.M., Clark, E., McMichael, J.F., Meyer, R.J., Schindler, J.K., Pohl, C.S., Wallis, J.W., Shi, X., Lin, L., Schmidt, H., Tang, Y., Haipek, C., Wiechert, M.E., Ivy, J.V., Kalicki, J., Elliott, G., Ries, R.E., Payton, J.E., Westervelt, P., Tomasson, M.H., Watson, M.A., Baty, J., Heath, S., Shannon, W.D., Nagarajan, R., Link, D.C., Walter, M.J., Graubert, T.A., DiPersio, J.F., Wilson, R.K., Ley, T.J., 2009. Recurring Mutations Found by Sequencing an Acute Myeloid Leukemia Genome. New England Journal of Medicine 361, 1058–1066. doi:10.1056/NEJMoa0903840

Maria Aminoff, J.E.C., 1999. Aminoff, M. et al. Mutations in the CUBN gene encoding the intrinsic factor vitamin B12 receptor, cubilin cause hereditary megaloblastic anemia 1 (MGA-1) Nature Genet. 21, 309−313. Nature genetics 21, 309–13. doi:10.1038/6831

Martincorena, I., Seshasayee, A.S.N., Luscombe, N.M., 2012. Evidence of non-random mutation rates suggests an evolutionary risk management strategy. Nature 485, 95–98. doi:10.1038/nature10995

Martin-Subero, J.I., Ammerpohl, O., Bibikova, M., Wickham-Garcia, E., Agirre, X., Alvarez, S., Brüggemann, M., Bug, S., Calasanz, M.J., Deckert, M., Dreyling, M., Du, M.Q., Dürig, J., Dyer, M.J.S., Fan, J.-B., Gesk, S., Hansmann, M.-L., Harder, L., Hartmann, S., Klapper, W., Küppers, R., Montesinos-Rongen, M., Nagel, I., Pott, C., Richter, J., Román-Gómez, J., Seifert, M., Stein, H., Suela, J., Trümper, L.,

Vater, I., Prosper, F., Haferlach, C., Cigudosa, J.C., Siebert, R., 2009. A Comprehensive Microarray-Based DNA Methylation Study of 367 Hematological Neoplasms. PLoS ONE 4, e6986. doi:10.1371/journal.pone.0006986

MCL - a cluster algorithm for graphs [WWW Document], n.d. URL http://micans.org/mcl/ (accessed 10.2.15).

Mignone, F., Gissi, C., Liuni, S., Pesole, G., 2002. Untranslated regions of mRNAs. Genome Biol 3, reviews0004.1–reviews0004.10.

Miles, K.M., Seshadri, M., Ciamporcero, E., Adelaiye, R., Gillard, B., Sotomayor, P., Attwood, K., Shen, L., Conroy, D., Kuhnert, F., Lalani, A.S., Thurston, G., Pili, R., 2014. Dll4 Blockade Potentiates the Anti-Tumor Effects of VEGF Inhibition in Renal Cell Carcinoma Patient-Derived Xenografts. PLoS ONE 9, e112371. doi:10.1371/journal.pone.0112371

Minner, S., Rump, D., Tennstedt, P., Simon, R., Burandt, E., Terracciano, L., Moch, H., Wilczak, W., Bokemeyer, C., Fisch, M., Sauter, G., Eichelberg, C., 2012. Epidermal growth factor receptor protein expression and genomic alterations in renal cell carcinoma. Cancer 118, 1268–1275. doi:10.1002/cncr.26436

O'Geen, H., Echipare, L., Farnham, P.J., 2011. Using ChIP-Seq Technology to Generate High-Resolution Profiles of Histone Modifications. Methods Mol Biol 791, 265–286. doi:10.1007/978-1-61779-316-5_20

Pal, J.K., Chatterjee, S., Berwal, S.K., 2001. Pathological Variations in 3'-untranslated Regions of Human Genes, in: eLS. John Wiley & Sons, Ltd.

Paz-Elizur, T., Brenner, D.E., Livneh, Z., 2005. Interrogating DNA Repair in Cancer Risk Assessment. Cancer Epidemiol Biomarkers Prev 14, 1585–1587. doi:10.1158/1055-9965.EPI-14-7-ED

Polak, P., Karlić, R., Koren, A., Thurman, R., Sandstrom, R., Lawrence, M.S., Reynolds, A., Rynes, E., Vlahoviček, K., Stamatoyannopoulos, J.A., Sunyaev, S.R., 2015. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. Nature 518, 360–364. doi:10.1038/nature14221

Polymenidou, M., Lagier-Tourenne, C., Hutt, K.R., Huelga, S.C., Moran, J., Liang, T.Y., Ling, S.-C., Sun, E., Wancewicz, E., Mazur, C., Kordasiewicz, H., Sedaghat, Y., Donohue, J.P., Shiue, L., Bennett, C.F., Yeo, G.W., Cleveland, D.W., 2011. Long pre-mRNA depletion and RNA missplicing contribute to neuronal vulnerability from loss of TDP-43. Nat Neurosci 14, 459–468. doi:10.1038/nn.2779

Porta, C., Figlin, R.A., 2009. Phosphatidylinositol-3-kinase/Akt signaling pathway and kidney cancer, and the therapeutic potential of phosphatidylinositol-3-kinase/Akt inhibitors. J. Urol. 182, 2569–2577. doi:10.1016/j.juro.2009.08.085

Pritchard, J.K., Cox, N.J., 2002. The allelic architecture of human disease genes: common disease–common variant… or not? Hum. Mol. Genet. 11, 2417–2423. doi:10.1093/hmg/11.20.2417

Purdie, K.J., Lambert, S.R., Teh, M.-T., Chaplin, T., Molloy, G., Raghavan, M., Kelsell, D.P., Leigh, I.M., Harwood, C.A., Proby, C.M., Young, B.D., 2007. Allelic Imbalances and Microdeletions Affecting the PTPRD Gene in Cutaneous Squamous Cell Carcinomas Detected Using Single Nucleotide Polymorphism Microarray Analysis. Genes Chromosomes Cancer 46, 661–669. doi:10.1002/gcc.20447

Reamon-Buettner, S.M., Cho, S.-H., Borlak, J., 2007. Mutations in the 3'-untranslated region of GATA4 as molecular hotspots for congenital heart disease (CHD). BMC Medical Genetics 8, 38. doi:10.1186/1471-2350-8-38

Rhodes, D.R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., Barrette, T., Pandey, A., Chinnaiyan, A.M., 2004. ONCOMINE: A Cancer Microarray Database and Integrated Data-Mining Platform. Neoplasia 6, 1–6.

Ricketts, C.J., Hill, V.K., Linehan, W.M., 2014. Tumor-Specific Hypermethylation of Epigenetic Biomarkers, Including SFRP1, Predicts for Poorer Survival in Patients from the TCGA Kidney Renal Clear Cell Carcinoma (KIRC) Project. PLoS ONE 9, e85621. doi:10.1371/journal.pone.0085621

Roos, F.C., Evans, A.J., Brenner, W., Wondergem, B., Klomp, J., Heir, P., Roche, O., Thomas, C., Schimmel, H., Furge, K.A., Teh, B.T., Thüroff, J.W., Hampel, C., Ohh, M., 2011. Deregulation of E2-EPF ubiquitin carrier protein in papillary renal cell carcinoma. Am. J. Pathol. 178, 853–860. doi:10.1016/j.ajpath.2010.10.033

Salehipoor, M., A, K., A, B.-B., B, G., M, R., M, A., Ma, A., 2012. Role of viruses in renal cell carcinoma. Saudi Journal of Kidney Diseases and Transplantation 23, 53.

Sauna, Z.E., Kimchi-Sarfaty, C., 2001. Synonymous Mutations as a Cause of Human Genetic Disease, in: eLS. John Wiley & Sons, Ltd.

Shapiro, J.A., 2009. Revisiting the Central Dogma in the 21st Century. Annals of the New York Academy of Sciences 1178, 6–28. doi:10.1111/j.1749-6632.2009.04990.x

Shi, D., Grossman, S.R., 2010. Ubiquitin becomes ubiquitous in cancer. Cancer Biol Ther 10, 737–747. doi:10.4161/cbt.10.8.13417

Skubitz, K.M., Skubitz, A.P.N., 2002. Differential gene expression in renal-cell cancer. Journal of Laboratory and Clinical Medicine 140, 52–64. doi:10.1067/mlc.2002.125213

Smith, N.G.C., Eyre-Walker, A., 2003. Human disease genes: patterns and predictions. Gene 318, 169–175. doi:10.1016/S0378-1119(03)00772-8

Song, L., Crawford, G.E., 2010. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. Cold Spring Harb Protoc 2010, pdb.prot5384. doi:10.1101/pdb.prot5384

Soumya, R., 2013. Mapping rare and common causal alleles for complex human diseases [WWW Document]. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3198013/ (accessed 11.12.15).

Srihari, S., Madhamshettiwar, P.B., Song, S., Liu, C., Simpson, P.T., Khanna, K.K., Ragan, M.A., 2014. Complex-based analysis of dysregulated cellular processes in cancer. BMC Systems Biology 8, S1. doi:10.1186/1752-0509-8-S4-S1

Stead, L.F., Sutton, K.M., Taylor, G.R., Quirke, P., Rabbitts, P., 2013. Accurately Identifying Low-Allelic Fraction Variants in Single Samples with Next-Generation Sequencing: Applications in Tumor Subclone Resolution. Human Mutation 34, 1432–1438. doi:10.1002/humu.22365

Steidl, U., Steidl, C., Ebralidze, A., Chapuy, B., Han, H.-J., Will, B., Rosenbauer, F., Becker, A., Wagner, K., Koschmieder, S., Kobayashi, S., Costa, D.B., Schulz, T., O'Brien, K.B., Verhaak, R.G.W., Delwel, R., Haase, D., Trümper, L., Krauter, J., Kohwi-Shigematsu, T., Griesinger, F., Tenen, D.G., 2007. A distal single nucleotide polymorphism alters long-range regulation of the PU.1 gene in acute myeloid leukemia. J. Clin. Invest. 117, 2611–2620. doi:10.1172/JCI30525

Stergachis, A.B., Haugen, E., Shafer, A., Fu, W., Vernot, B., Reynolds, A., Raubitschek, A., Ziegler, S., LeProust, E.M., Akey, J.M., Stamatoyannopoulos, J.A., 2013. Exonic transcription factor binding directs codon choice and impacts protein evolution. Science 342, 1367–1372. doi:10.1126/science.1243490

Stern, D.L., Orgogozo, V., 2008. The Loci of Evolution: How Predictable is Genetic Evolution? Evolution 62, 2155–2177. doi:10.1111/j.1558-5646.2008.00450.x

Strausberg, R.L., Simpson, A.J.G., 2010. Whole-genome cancer analysis as an approach to deeper understanding of tumour biology. Br J Cancer 102, 243–248. doi:10.1038/sj.bjc.6605497

The 1000 Genomes Project Consortium, 2012. An integrated map of genetic variation from 1,092 human genomes [WWW Document]. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3498066/ (accessed 9.1.15).

The Cancer Genome Atlas Research Network, 2013. Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature 499, 43–49. doi:10.1038/nature12222

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

Tijani, K.H., Anunobi, C.C., Ezenwa, E.V., Lawal, A., Habeebu, M.Y.M., Jeje, E.A., Ogunjimi, M.A., Afolayan, M.O., 2012. Adult renal cell carcinoma in Lagos: Experience and challenges at the Lagos University Teaching Hospital. African Journal of Urology 18, 20–23. doi:10.1016/j.afju.2012.04.005

Tomasetti, C., Marchionni, L., Nowak, M.A., Parmigiani, G., Vogelstein, B., 2015. Only three driver gene mutations are required for the development of lung and colorectal cancers. PNAS 112, 118–123. doi:10.1073/pnas.1421839112

Tsompana, M., Buck, M.J., 2014. Chromatin accessibility: a window into the genome. Epigenetics & Chromatin 7, 33. doi:10.1186/1756-8935-7-33

Wang, J., Lu, C., Min, D., Wang, Z., Ma, X., 2007. A Mutation in the 5' Untranslated Region of the BRCA1 Gene in Sporadic Breast Cancer Causes Downregulation of Translation Efficiency. Journal of International Medical Research 35, 564–573. doi:10.1177/147323000703500417

Wang, Y., Jensen, R.C., Stumph, W.E., 1996. Role of TATA Box Sequence and Orientation in Determining RNA Polymerase II/III Transcription Specificity. Nucl. Acids Res. 24, 3100–3106. doi:10.1093/nar/24.15.3100

Weinhold, N., Jacobsen, A., Schultz, N., Sander, C., Lee, W., 2014. Genome-wide analysis of non-coding regulatory mutations in cancer. Nat Genet 46, 1160–1165. doi:10.1038/ng.3101

White, N.M.A., Masui, O., DeSouza, L.V., Krakovska-Yutz, O., Metias, S., Romaschin, A.D., Honey, R.J., Stewart, R., Pace, K., Lee, J., Jewett, M.A., Bjarnason, G.A., Siu, K.W.M., Yousef, G.M., 2014. Quantitative proteomic analysis reveals potential diagnostic markers and pathways involved in pathogenesis of renal cell carcinoma. Oncotarget 5, 506–518.

Wong, G.K.-S., Passey, D.A., Huang, Y., Yang, Z., Yu, J., 2000. Is "Junk" DNA Mostly Intron DNA? Genome Res 10, 1672–1678.

You, J.S., Jones, P.A., 2012. Cancer Genetics and Epigenetics: Two Sides of the Same Coin? Cancer Cell 22, 9–20. doi:10.1016/j.ccr.2012.06.008

Zeng, Z., QUE, T., ZHANG, J., HU, Y., 2014. A study exploring critical pathways in clear cell renal cell carcinoma. Exp Ther Med 7, 121–130. doi:10.3892/etm.2013.1392

Zhang, C., Xuan, Z., Otto, S., Hover, J.R., McCorkle, S.R., Mandel, G., Zhang, M.Q., 2006. A clustering property of highly-degenerate transcription factor binding sites in the mammalian genome. Nucleic Acids Res 34, 2238–2246. doi:10.1093/nar/gkl248

Zhang, Q., Edwards, S.V., 2012. The Evolution of Intron Size in Amniotes: A Role for Powered Flight? Genome Biol Evol 4, 1033–1043. doi:10.1093/gbe/evs070

Appendices

Appendix I

The 173 ccRCC disease genes with their Ensembl IDs, HGNC symbols, chromosome number and strand, where 1 is the forward strand and -1 is the reverse strand.

ENSG00000154262	ABCA6	17	-1
ENSG00000085563	ABCB1	7	-1
ENSG00000135503	ACVR1B	12	1
ENSG00000164742	ADCY1	7	1
ENSG00000185567	AHNAK2	14	-1
ENSG00000127914	AKAP9	7	1
ENSG00000142208	AKT1	14	-1
ENSG00000145362	ANK2	4	1
ENSG00000151150	ANK3	10	-1
ENSG00000196975	ANXA4	2	1
ENSG00000134982	APC	5	1
ENSG00000117713	ARID1A	1	1
ENSG00000149311	ATM	11	1
ENSG00000085224	ATRX	X	-1
ENSG00000166710	B2M	15	1
ENSG00000163930	BAP1	3	-1
ENSG00000138376	BARD1	2	-1
ENSG00000176171	BNIP3	10	-1
ENSG00000074410	CA12	15	-1
ENSG00000107159	CA9	9	1
ENSG00000064989	CALCRL	2	-1
ENSG00000153113	CAST	5	1
ENSG00000105974	CAV1	7	1
ENSG00000067955	CBFB	16	1
ENSG00000110148	CCKBR	11	1
ENSG00000110092	CCND1	11	1
ENSG00000004897	CDC27	17	-1
ENSG00000039068	CDH1	16	1
ENSG00000179776	CDH5	16	1
ENSG00000147889	CDKN2A	9	-1
ENSG00000181885	CLDN7	17	-1
ENSG00000141367	CLTC	17	1
ENSG00000182871	COL18A1	21	1
ENSG00000169031	COL4A3	2	1
ENSG00000118260	CREB1	2	1

ENSG00000105662	CRTC1		19	1
ENSG00000144677	CTDSPL		3	1
ENSG00000107611	CUBN		10	-1
ENSG00000108094	CUL2		10	-1
ENSG00000121966	CXCR4		2	-1
ENSG00000083799	CYLD		16	1
ENSG00000211452	DIO1		1	1
ENSG00000198947	DMD	X		-1
ENSG00000134516	DOCK2		5	1
ENSG00000136160	EDNRB		13	-1
ENSG00000146648	EGFR		7	1
ENSG00000100393	EP300		22	1
ENSG00000116016	EPAS1		2	1
ENSG00000141736	ERBB2		17	1
ENSG00000178568	ERBB4		2	-1
ENSG00000106462	EZH2		7	-1
ENSG00000168309	FAM107A		3	-1
ENSG00000138829	FBN2		5	-1
ENSG00000066468	FGFR2		10	-1
ENSG00000068078	FGFR3		4	1
ENSG00000091483	FH		1	-1
ENSG00000189283	FHIT		3	-1
ENSG00000115414	FN1		2	-1
ENSG00000075539	FRYL		4	-1
ENSG00000128242	GAL3ST1		22	-1
ENSG00000107485	GATA3		10	1
ENSG00000125166	GOT2		16	-1
ENSG00000171723	GPHN		14	1
ENSG00000113249	HAVCR1		5	-1
ENSG00000112406	HECA		6	1
ENSG00000095951	HIVEP1		6	1
ENSG00000206503	HLA-A		6	1
ENSG00000234745	HLA-B		6	-1
ENSG00000108753	HNF1B		17	-1
ENSG00000174775	HRAS		11	-1
ENSG00000109971	HSPA8		11	-1
ENSG00000142798	HSPG2		1	-1
ENSG00000086758	HUWE1	X		-1
ENSG00000136231	IGF2BP3		7	-1
ENSG00000123104	ITPR2		12	-1
ENSG00000096968	JAK2		9	1
ENSG00000136636	KCTD3		1	1
ENSG00000126012	KDM5C	X		-1
ENSG00000147050	KDM6A	X		1

ENSG00000157404	KIT		4	1
ENSG00000167548	KMT2D		12	-1
ENSG00000135480	KRT7		12	1
ENSG00000198910	L1CAM	X		-1
ENSG00000136167	LCP1		13	-1
ENSG00000131981	LGALS3		14	1
ENSG00000198589	LRBA		4	-1
ENSG00000188906	LRRK2		12	1
ENSG00000139329	LUM		12	-1
ENSG00000107968	MAP3K8		10	1
ENSG00000125952	MAX		14	-1
ENSG00000131844	MCCC2		5	1
ENSG00000137337	MDC1		6	-1
ENSG00000085276	MECOM		3	-1
ENSG00000108510	MED13		17	-1
ENSG00000042429	MED17		11	1
ENSG00000068305	MEF2A		15	1
ENSG00000105976	MET		7	1
ENSG00000187098	MITF		3	1
ENSG00000148773	MKI67		10	-1
ENSG00000196549	MME		3	1
ENSG00000087245	MMP2		16	1
ENSG00000147065	MSN	X		1
ENSG00000173531	MST1		3	-1
ENSG00000198793	MTOR		1	-1
ENSG00000100345	MYH9		22	-1
ENSG00000067798	NAV3		12	1
ENSG00000163386	NBPF10		1	1
ENSG00000196498	NCOR2		12	-1
ENSG00000165795	NDRG2		14	-1
ENSG00000183091	NEB		2	-1
ENSG00000049759	NEDD4L		18	1
ENSG00000196712	NF1		17	1
ENSG00000109320	NFKB1		4	1
ENSG00000166741	NNMT		11	1
ENSG00000147140	NONO	X		1
ENSG00000089250	NOS1		12	-1
ENSG00000148400	NOTCH1		9	-1
ENSG00000165671	NSD1		5	1
ENSG00000114026	OGG1		3	1
ENSG00000070756	PABPC1		8	-1
ENSG00000112530	PACRG		6	1
ENSG00000163939	PBRM1		3	-1
ENSG00000134853	PDGFRA		4	1

ENSG00000261371	PECAM1		17	-1
ENSG00000198300	PEG3		19	-1
ENSG00000121879	PIK3CA		3	1
ENSG00000178209	PLEC		8	-1
ENSG00000132170	PPARG		3	1
ENSG00000163932	PRKCD		3	1
ENSG00000174231	PRPF8		17	-1
ENSG00000171862	PTEN		10	1
ENSG00000163629	PTPN13		4	1
ENSG00000153707	PTPRD		9	-1
ENSG00000164611	PTTG1		5	1
ENSG00000132155	RAF1		3	-1
ENSG00000132341	RAN		12	1
ENSG00000153201	RANBP2		2	1
ENSG00000139687	RB1		13	1
ENSG00000147274	RBMX	X		-1
ENSG00000173821	RNF213		17	1
ENSG00000108375	RNF43		17	-1
ENSG00000122406	RPL5		1	1
ENSG00000159216	RUNX1		21	-1
ENSG00000198838	RYR3		15	1
ENSG00000151835	SACS		13	-1
ENSG00000099194	SCD		10	1
ENSG00000117118	SDHB		1	-1
ENSG00000181555	SETD2		3	-1
ENSG00000104332	SFRP1		8	-1
ENSG00000141646	SMAD4		18	1
ENSG00000127616	SMARCA4		19	1
ENSG00000065526	SPEN		1	1
ENSG00000118046	STK11		19	1
ENSG00000131018	SYNE1		6	-1
ENSG00000054654	SYNE2		14	1
ENSG00000135111	TBX3		12	-1
ENSG00000148737	TCF7L2		10	1
ENSG00000168769	TET2		4	1
ENSG00000105329	TGFB1		19	-1
ENSG00000171914	TLN2		15	1
ENSG00000164342	TLR3		4	1
ENSG00000141510	TP53		17	-1
ENSG00000038382	TRIO		5	1
ENSG00000103197	TSC2		16	1
ENSG00000156298	TSPAN7	X		1
ENSG00000038427	VCAN		5	1
ENSG00000112715	VEGFA		6	1

ENSG00000134086	VHL	3	1
ENSG00000134258	VTCN1	1	-1
ENSG00000110799	VWF	12	-1
ENSG00000109685	WHSC1	4	1
ENSG00000184937	WT1	11	-1
ENSG00000140836	ZFHX3	16	-1

Appendix II

The 173 non-disease genes with their Ensembl IDs, HGNC symbols, chromosome number and strand, where 1 is the forward strand and -1 is the reverse strand.

ENSG00000243532	RN7SL19P	8	-1
ENSG00000258668	COX6CP11	14	-1
ENSG00000122432	SPATA1	1	1
ENSG00000257853	MED15P1	14	-1
ENSG00000252535	RNA5SP254	8	-1
ENSG00000229455	RPS10P18	10	-1
ENSG00000235294	RPL7AP25	3	-1
ENSG00000206228	HNRNPA1P4	8	-1
ENSG00000250585	LINC00604	5	1
ENSG00000233543	CHTF8P1	X	1
ENSG00000184208	C22orf46	22	1
ENSG00000199876	RN7SKP131	22	-1
ENSG00000247240	UBL7-AS1	15	1
ENSG00000252174	RNU7-18P	3	1
ENSG00000249031	SUMO2P6	5	1
ENSG00000180440	SERTM1	13	1
ENSG00000226617	RPL21P110	13	1
ENSG00000182700	IGIP	5	1
ENSG00000200028	RNA5SP98	2	1
ENSG00000104177	MYEF2	15	-1
ENSG00000113119	TMCO6	5	1
ENSG00000227203	SUB1P1	8	1
ENSG00000207962	MIR30C1	1	1
ENSG00000223656	HMGB3P10	1	1
ENSG00000163958	ZDHHC19	3	-1
ENSG00000243510	RN7SL111P	2	1
ENSG00000128563	PRKRIP1	7	1
ENSG00000227766	HCG4P5	6	-1
ENSG00000207138	RNU6-869P	8	1
ENSG00000184586	KRTAP7-1	21	-1
ENSG00000176953	NFATC2IP	16	1
ENSG00000227242	NBPF13P	1	-1
ENSG00000215325	ASS1P10	5	1
ENSG00000235703	LINC00894	X	1
ENSG00000204652	RPS26P8	17	1
ENSG00000219891	ZSCAN12P1	6	1
ENSG00000169900	PYDC1	16	-1
ENSG00000230833	RPEP3	1	1

ENSG00000253936	IGHV3-63		14	-1
ENSG00000178947	LINC00086	X		1
ENSG00000230873	STMND1		6	1
ENSG00000234617	SNRK-AS1		3	-1
ENSG00000178852	EFCAB13		17	1
ENSG00000014914	MTMR11		1	-1
ENSG00000181689	OR8K3		11	1
ENSG00000201170	RNU1-132P		1	-1
ENSG00000237864	LINC00322		21	-1
ENSG00000086619	ERO1LB		1	-1
ENSG00000251260	WDFY3-AS1		4	1
ENSG00000241350	PMS2P11		7	1
ENSG00000248366	TRAJ51		14	1
ENSG00000180658	OR2A4		6	-1
ENSG00000200480	SNORD114-8		14	1
ENSG00000240545	RN7SL492P		4	-1
ENSG00000141699	FAM134C		17	-1
ENSG00000242360	RN7SL272P		13	-1
ENSG00000242707	RN7SL362P		18	-1
ENSG00000231920	NEBL-AS1		10	1
ENSG00000223921	MTND1P27		2	-1
ENSG00000199529	RNU6-462P		4	1
ENSG00000236753	MKLN1-AS		7	-1
ENSG00000151470	C4orf33		4	1
ENSG00000200086	RNU6-433P		2	-1
ENSG00000254825	OR9G2P		11	-1
ENSG00000132424	PNISR		6	-1
ENSG00000243366	RN7SL60P		13	1
ENSG00000239490	RPS4XP18		18	1
ENSG00000256193	LINC00507		12	1
ENSG00000207453	RNU6-535P		10	-1
ENSG00000181778	TMEM252		9	-1
ENSG00000250337	LINC01021		5	1
ENSG00000202261	SNORD115-4		15	1
ENSG00000225713	RPL30P1		1	-1
ENSG00000147041	SYTL5	X		1
ENSG00000188626	GOLGA8M		15	-1
ENSG00000121766	ZCCHC17		1	1
ENSG00000185182	GOLGA8DP		15	-1
ENSG00000218902	PTMAP3		20	1
ENSG00000229546	LINC00428		13	-1
ENSG00000228050	TOP3BP1		22	-1
ENSG00000198833	UBE2J1		6	-1
ENSG00000234901	MTND6P13	X		-1

ENSG00000226942	IL9RP3		16	-1
ENSG00000214313	AZGP1P1		7	1
ENSG00000143162	CREG1		1	-1
ENSG00000226245	ZNF32-AS1		10	1
ENSG00000204670	IGKV1OR2-3		2	1
ENSG00000231166	TUBB4BP6		9	1
ENSG00000254925	OR4C9P		11	-1
ENSG00000201210	RNA5SP139		3	1
ENSG00000162592	CCDC27		1	1
ENSG00000249464	LINC01091		4	1
ENSG00000241225	TRNAS30P		17	-1
ENSG00000130640	TUBGCP2		10	-1
ENSG00000224344	KNOP1P3		2	1
ENSG00000167595	C19orf55		19	1
ENSG00000251859	RNU6-1288P		2	-1
ENSG00000252782	RNU6-341P		14	-1
ENSG00000212160	RNU6-205P		4	1
ENSG00000237200	ZBTB40-IT1		1	1
ENSG00000207622	MIR619		12	-1
ENSG00000196653	ZNF502		3	1
ENSG00000100129	EIF3L		22	1
ENSG00000235379	RPL7P31		7	1
ENSG00000187536	TPM3P7		2	-1
ENSG00000237443	OR13D2P		9	1
ENSG00000251813	RNU6-983P		1	-1
ENSG00000207406	SNORA41		2	1
ENSG00000226653	OR13Z1P		1	1
ENSG00000249421	ADAMTS19-1		5	-1
ENSG00000128694	OSGEPL1		2	-1
ENSG00000180846	CSNK1G2-S1		19	-1
ENSG00000207721	MIR186		1	-1
ENSG00000252311	RNU1-103P		16	-1
ENSG00000235892	PKMP2	X		1
ENSG00000068654	POLR1A		2	-1
ENSG00000212014	MIR509-3	X		-1
ENSG00000164970	FAM219A		9	-1
ENSG00000164385	C6orf195		6	-1
ENSG00000235169	SMIM1		1	1
ENSG00000230069	LRRC37A15P		4	-1
ENSG00000249459	ZNF286B		17	-1
ENSG00000102055	PPP1R2P9	X		-1
ENSG00000171987	C11orf40		11	-1
ENSG00000236156	CHCHD4P3		9	1
ENSG00000215943	MIR892A	X		-1

ENSG00000148468	FAM171A1		10	-1
ENSG00000255238	RFPL4AP1		19	1
ENSG00000199179	MIRLET7I		12	1
ENSG00000049319	SRD5A2		2	-1
ENSG00000196240	OR2T2		1	1
ENSG00000022840	RNF10		12	1
ENSG00000176312	OR4H12P		14	1
ENSG00000198414	TATDN2P1	X		1
ENSG00000205871	RPS3AP47		15	-1
ENSG00000139239	RPL14P1		12	1
ENSG00000257482	ZNF727		7	1
ENSG00000251729	RNA5SP401		15	-1
ENSG00000142609	C1orf222		1	-1
ENSG00000237665	GRM7-AS2		3	-1
ENSG00000258710	CT60		15	1
ENSG00000239151	RNU7-195P		15	1
ENSG00000201499	RNU6-312P		2	-1
ENSG00000238842	RNU7-106P		12	-1
ENSG00000256037	MRPL40P1		12	1
ENSG00000187156	LINC00221		14	1
ENSG00000123933	MXD4		4	-1
ENSG00000064763	FAR2		12	1
ENSG00000187867	PALM3		19	-1
ENSG00000257704	PRR24		19	1
ENSG00000164818	HEATR2		7	1
ENSG00000254161	IGLVIV-65		22	1
ENSG00000130684	ZNF337		20	-1
ENSG00000221598	MIR1249		22	-1
ENSG00000216090	MIR937		8	-1
ENSG00000153975	ZUFSP		6	-1
ENSG00000217330	SSXP10		6	1
ENSG00000200889	RNU4-13P		17	-1
ENSG00000253229	HIGD1AP6		8	1
ENSG00000177693	OR4F4		15	-1
ENSG00000147036	LANCL3	X		1
ENSG00000128891	C15orf57		15	-1
ENSG00000104979	C19orf53		19	1
ENSG00000207757	MIR93		7	-1
ENSG00000230418	ARL2BPP7		9	-1
ENSG00000185028	LRRC14B		5	1
ENSG00000151806	GUF1		4	1
ENSG00000145107	TM4SF19		3	-1
ENSG00000214530	STARD10		11	-1
ENSG00000204588	LINC01123		2	1

ENSG00000236253	SLC25A3P1	1	-1
ENSG00000168890	TMEM150A	2	-1
ENSG00000201302	SNORA65	9	-1

Appendix III

The total number of non-coding mutations compared to the total number of deleterious mutations per gene and their corresponding p-value based on the chi-square statistical test.

HGNC symbol	Total non-coding mutations	Deleterious mutations	p-value
PTPRD	473	0	7.1277E-105
DMD	329	3	6.17363E-71
ERBB4	300	4	9.06696E-64
FHIT	144	0	3.55296E-33
RUNX1	153	5	6.50503E-31
MECOM	140	2	1.41249E-30
LRBA	113	1	1.59413E-25
RYR3	97	0	6.93273E-23
ANK3	100	1	1.12586E-22
NAV3	95	0	1.90385E-22
PACRG	92	1	6.40314E-21
DOCK2	99	5	3.72501E-19
GPHN	76	0	2.83665E-18
ANK2	74	0	7.8117E-18
SYNE1	71	1	2.63858E-16
NEDD4L	72	3	7.35785E-15
FBN2	68	2	8.41693E-15
CUBN	75	4	1.02196E-14
NF1	58	0	2.62118E-14
ITPR2	60	2	4.8453E-13
SYNE2	48	0	4.26219E-12
ADCY1	50	1	1.13521E-11
VWF	34	0	5.51121E-09
TRIO	61	8	8.32909E-09
MME	32	0	1.54173E-08
ABCB1	29	0	7.23783E-08
ZFHX3	27	0	2.03455E-07
FRYL	38	3	2.09072E-07
NSD1	25	0	5.73303E-07
TCF7L2	36	3	5.73303E-07
ATRX	28	1	8.94473E-07
TLN2	27	1	1.49988E-06
PBRM1	23	0	1.62001E-06
MITF	22	0	2.7265E-06
NEB	22	0	2.7265E-06
APC	21	0	4.59283E-06

SACS	21	0	4.59283E-06
MEF2A	20	0	7.74422E-06
PTPN13	20	0	7.74422E-06
WHSC1	20	0	7.74422E-06
NOS1	27	2	9.58401E-06
AKAP9	19	0	1.30718E-05
ATM	19	0	1.30718E-05
MED13	19	0	1.30718E-05
SMAD4	19	0	1.30718E-05
IGF2BP3	25	2	2.66915E-05
HUWE1	17	0	3.73798E-05
PTEN	17	0	3.73798E-05
RB1	17	0	3.73798E-05
TET2	17	0	3.73798E-05
HIVEP1	20	1	5.69941E-05
NCOR2	39	7	6.24904E-05
JAK2	16	0	6.33425E-05
KIT	19	1	9.61659E-05
KDM6A	15	0	0.000107511
LRRK2	15	0	0.000107511
EGFR	28	4	0.000157052
COL18A1	14	0	0.000182811
COL4A3	14	0	0.000182811
FN1	14	0	0.000182811
SMARCA4	17	1	0.000274727
CALCRL	13	0	0.000311491
MTOR	13	0	0.000311491
VCAN	13	0	0.000311491
CDC27	12	0	0.000532006
CDH1	12	0	0.000532006
ANXA4	15	1	0.000789113
EP300	15	1	0.000789113
SETD2	15	1	0.000789113
MYH9	11	0	0.000911119
PDGFRA	11	0	0.000911119
RANBP2	11	0	0.000911119
RNF43	11	0	0.000911119
CRTC1	14	1	0.001340641
VTCN1	14	1	0.001340641
ABCA6	10	0	0.001565402
EPAS1	10	0	0.001565402
EZH2	10	0	0.001565402
LCP1	10	0	0.001565402
CAST	17	2	0.001616222

MET	17	2	0.001616222
NBPF10	17	2	0.001616222
NFKB1	13	1	0.002281937
ARID1A	9	0	0.002699796
CAV1	9	0	0.002699796
CDH5	9	0	0.002699796
CUL2	9	0	0.002699796
MCCC2	9	0	0.002699796
VHL	9	0	0.002699796
CBFB	8	0	0.004677735
CLTC	8	0	0.004677735
HECA	8	0	0.004677735
TP53	8	0	0.004677735
HNF1B	11	1	0.006655605
SPEN	11	1	0.006655605
RNF213	20	4	0.007290358
BARD1	7	0	0.008150972
CA12	7	0	0.008150972
CREB1	7	0	0.008150972
CTDSPL	7	0	0.008150972
KCTD3	7	0	0.008150972
MED17	7	0	0.008150972
PIK3CA	7	0	0.008150972
CDKN2A	6	0	0.014305878
EDNRB	6	0	0.014305878
KMT2D	6	0	0.014305878
MAX	6	0	0.014305878
CYLD	5	0	0.025347319
MSN	5	0	0.025347319
NNMT	5	0	0.025347319
PABPC1	5	0	0.025347319
PRPF8	5	0	0.025347319
SFRP1	5	0	0.025347319
AHNAK2	4	0	0.045500264
FGFR2	16	4	0.045500264
MKI67	4	0	0.045500264
PEG3	4	0	0.045500264
TSPAN7	4	0	0.045500264
PPARG	10	2	0.057779571
GATA3	7	1	0.058781721
L1CAM	7	1	0.058781721
B2M	3	0	0.083264517
BAP1	3	0	0.083264517
FH	3	0	0.083264517

GOT2	3	0	0.083264517
NDRG2	3	0	0.083264517
RBMX	3	0	0.083264517
RPL5	3	0	0.083264517
CXCR4	2	0	0.157299207
DIO1	8	2	0.157299207
GAL3ST1	2	0	0.157299207
HAVCR1	2	0	0.157299207
HLA-B	2	2	0.157299207
KDM5C	2	0	0.157299207
LUM	2	0	0.157299207
MMP2	2	0	0.157299207
SDHB	2	0	0.157299207
TSC2	2	0	0.157299207
BNIP3	5	1	0.179712495
NOTCH1	7	2	0.256839258
CA9	1	0	0.317310508
FAM107A	1	0	0.317310508
FGFR3	1	0	0.317310508
HRAS	1	0	0.317310508
KRT7	1	0	0.317310508
MAP3K8	1	0	0.317310508
MDC1	1	0	0.317310508
PLEC	4	1	0.317310508
PRKCD	1	0	0.317310508
PTTG1	1	0	0.317310508
RAN	1	0	0.317310508
SCD	1	0	0.317310508
STK11	1	0	0.317310508
TLR3	1	0	0.317310508
WT1	4	1	0.317310508
CCND1	3	2	0.563702862
ERBB2	3	1	0.563702862
MST1	3	1	0.563702862
VEGFA	3	2	0.563702862
ACVR1B	2	1	1
AKT1	4	2	1
HSPG2	10	5	1
OGG1	2	1	1
TBX3	2	1	1
TGFB1	2	1	1

Appendix IV

The eight RCC with had no somatic mutations

CCKBR

CLDN7

HLA-A

HSPA8

LGALS3

NONO

PECAM1

RAF1

Appendix V

Table 8: (full) If there were no mutations in that mutation category, it is shown #N/A. Although all patients had multiple non-coding somatic mutations (column 2), some patients had either deleterious non-coding mutations **or** CDS mutations, but not both. For fourteen patients, there wasn't any deleterious non-coding or CDS mutations (example shown in **green**).

Patient ID	ALL non-coding	ALL CDS mutations	Deleterious mutations ONLY	No deleterious or CDS mutation
DO46877	221	7	10	
DO46897	200	3	4	
DO46905	131	4	5	
DO47100	129	5	5	
DO47240	103	4	2	
DO46836	87	2	3	
DO47088	85	3	2	
DO47004	83	1	1	
DO46847	83	5	3	
DO46929	83	2	1	
DO46992	80	1	#N/A	
DO47140	77	4	#N/A	
DO47112	76	1	#N/A	
DO47150	75	2	5	
DO46844	73	1	2	
DO46980	72	2	3	
DO47060	67	3	1	
DO46933	66	2	3	
DO47159	66	6	#N/A	

DO46827	65	#N/A		2	
DO47174	59	5		3	
DO46873	57	2	#N/A		
DO47162	57	2		3	
DO46841	56	2		2	
DO47104	55	3		2	
DO46881	55	3		2	
DO47136	54	1		1	
DO47012	52	5	#N/A		
DO46957	50	#N/A	#N/A		1
DO47168	50	#N/A	#N/A		2
DO47237	50	3	#N/A		
DO47171	49	#N/A		4	
DO46984	49	#N/A		1	
DO46828	49	4		1	
DO47120	48	2		1	
DO46949	46	#N/A		1	
DO47128	46	1		1	
DO47072	45	#N/A	#N/A		3
DO46937	45	2		1	
DO46838	45	2		1	
DO47092	45	3		5	
DO46988	45	2	#N/A		
DO47124	44	2		3	
DO46973	43	2		3	
DO46869	42	1		1	
DO46856	41	2		2	

DO46889	41		1		2	
DO46961	41		2		1	
DO46859	40		1	#N/A		
DO46853	40		1		1	
DO47144	39		4	#N/A		
DO46925	37		2		1	
DO46953	36		1	#N/A		
DO46885	35	#N/A			1	
DO46977	35		1		1	
DO46917	35		2	#N/A		
DO46909	34		1		1	
DO46850	34		1	#N/A		
DO47153	34		1		1	
DO47165	30	#N/A		#N/A		4
DO46913	30		1	#N/A		
DO46996	29		1		1	
DO47056	29		1	#N/A		
DO46830	29		3	#N/A		
DO46945	28	#N/A		#N/A		5
DO47064	27		2	#N/A		
DO46941	27	#N/A		#N/A		6
DO46893	27		4		2	
DO47147	27	#N/A			1	
DO47096	26		2		4	
DO46832	26		1		2	
DO47016	26		1		3	
DO47052	26	#N/A			1	

DO46865	24	#N/A	#N/A		7
DO46969	23	1	#N/A		
DO47116	23	2		1	
DO46862	22	1	#N/A		
DO47080	21	1		1	
DO47246	21	2		4	
DO47084	18	1	#N/A		
DO47156	16	#N/A		1	
DO47234	15	3		1	
DO47243	15	1	#N/A		
DO46965	14	#N/A	#N/A		8
DO47068	14	#N/A		3	
DO47000	12	#N/A	#N/A		9
DO47108	12	#N/A		1	
DO47076	12	#N/A	#N/A		
DO47132	10	#N/A	#N/A		10
DO46901	10	1	#N/A		
DO46921	8	2	#N/A		
DO47249	8	#N/A	#N/A		11
DO46834	8	#N/A	#N/A		12
DO46826	6	#N/A	#N/A		13
DO47048	6	#N/A	#N/A		14

Appendix VI

The 57 genes submitted to STRING-DB

TRIO
RUNX1
DOCK2
ATM
NEDD4L
NCOR2
MECOM
PPARG
ANK3
ERBB4
MYH9
PLEC
CDKN2A
FRYL
CUBN
IGF2BP3
ANK2
FN1
MITF
OGG1
DMD
RAN
ITPR2
NF1
SYNE1
EGFR
SYNE2
HRAS
HNF1B
NFKB1
HSPG2

FGFR2
RNF213
TCF7L2
HLA-B
MET
CCND1
NBPF10
NOS1
AKT1
NOTCH1
RYR3
FBN2
DIO1
CAST
VEGFA
MAX
KMT2D
LCP1
VHL
PDGFRA
HLA-C
SCD
PACRG
BAP1
CEBPB
CTCF
EBF1

Appendix VII

Read-Me

1. Selection of ccRCC WGS somatic mutation dataset from ICGC

Initially TCGA and then COSMIC ccRCC datasets were used, but when the results were evaluated, more coding SVs were obtained. The contact centres of the respective databases were contacted and it was confirmed that these were actually whole exome sequencing datasets, despite being under the heading 'whole genome.'

The ICGC dataset was then retrieved and confirmed to be whole genome sequenced somatic variants.

1.1. Downloaded simple somatic mutation in ccRCC from ICGC
https://dcc.icgc.org/repository/current/Projects/RECA-EU

File: simple_somatic_mutation.open.RECA-EU.tsv (Release18 January 21, 2015)

1.2. There were 95 unique donors, specimens and samples as shown by the Linux command: **sort –t 'Cntl + V + tab' –k2,2 (donor)/-k4,4 (specimen)/ -k5,5 (sample) simple_somatic_mutaion.open.RECA-EU.tsv –u > uniq_donor/specimen/sample_ID.txt**

1.3. I checked if they were all GRCh37 coordinates using **grep –c GRCh37 simple_somatic_mutaion.open.RECA-EU.tsv.** Had 1522025 hits. The same as the number of lines in the file using wc –l.

1.4. I Did the same to check if they were Whole genome sequenced **grep –c WGS simple_somatic_mutaion.open.RECA-EU.tsv** and this was found to

be the same count as the same as the number of lines in the file using **wc –l**.

1.5. I also downloaded the ICGC clinical data file from https://dcc.icgc.org/api/v1/download?fn=/release_18/Projects/RECA-EU/clinical.RECA-EU.tsv.gz

File: **donor.all_projects.tsv**

The file contains: ICGC donor ID, project code, study donor involved in (eg. EU or US), submitted donor ID, donor sex, vital status, disease status at last follow up, relapse type (progression), donor age at diagnosis, age at enrolment, age at last follow up, relapse interval, donor diagnosis_icd10, donor tumour staging system at diagnosis, tumour stage at diagnosis, tumour stage at diagnosis supplemental, survival time, interval of last follow up

2. Selection of genes of interest

2.1. ccRCC genes

The following databases were queried and the below-mentioned criteria were used to select the genes of interest.

NCBI (OMIM) - 81 genes

http://www.ncbi.nlm.nih.gov/omim

Too few genes (ONLY 10) if I select “clear cell renal cell carcinoma”

Advanced search: “renal cell carcinoma”

Selected only those preceded by asterix (*)

NCBI (Gene) - 134 genes

http://www.ncbi.nlm.nih.gov/gene

Criteria: “clear cell renal cell carcinoma” AND “Homo sapiens”

Homo sapiens (side panel)

Gene source: Genomic

Categories: Alternatively spliced, Annotated genes

Sequence content: Ensembl

Oncomine - (2844 genes)

https://www.oncomine.org/resource/login.html **(requires Login details)**

Analysis Type: clear cell renal cell carcinoma vs. Normal Analysis

Analysis Type: Kidney cancer vs. Normal Analysis

Analysis Type: clear cell renal cell carcinoma

Molecular subtype: Mutation

Sample Type: clinical specimen

Selected top 10% over and under expressed genes.

Intogen - (263 genes)

Filtered by: Cancer site: Kidney cancer

Driver category: HCD (High confident drivers)

COSMIC– 300 genes

https://www.intogen.org/search

Criteria: Clear cell renal cell carcinoma

Genes with mutations TAB

Selected the top 300 genes based on amount of mutated samples out of the amount of samples tested

2.2. Non-disease genes

I wrote a script to select 500 random non-disease genes from a file containing 38256 non-disease genes compiled by a previous Post-doc student Dr Wendy Kroger. I took

500 at first because when I selected the correct amount of genes, many of them didn't have HGNC symbols and I needed genes with both Ensembl and HGNC symbols.
Python script: **random_nondisease_genes.py**

I converted the 500 random Ensembl IDs to Ensembl Transcript IDs with their strand orientation and HGNC symbols using BioMart Ensembl.

I used the sort function in excel to extract only the genes that had HGNC symbols, because it would later be needed to confirm that the hits found where representative of my GOIs. The chromosome number and strand were also extracted for interest's sake.

The **random_nondisease_genes.py script** was modified to extract JUST 173 genes from the script with all the non-disease genes (that had both an Ensembl ID and HGNC symbol)
sort -k2,2 173_non_disease_genes.txt -u > 173_non_disease_genes_UNIQ.txt was used to confirm that the HGNC Ids were unique and **wc -l** was used to confirm that the number of ccRCC genes I was working with was 173.
The Ensembl ID's of the 173 non-disease genes were again submitted to BioMart in order to retrieve the transcript ID's for only those 173 genes, because UCSC didn't recognize many of the HGCN symbols and doesn't accept Ensembl Ids

3. Extraction of bed files form UCSC

3.1. ccRCC disease genes

It was found that the + strand in the ICGC dataset was actually representative of the reference genome strand on which the genotype alleles are located and it has nothing to do with the strandedness of the gene that contains the somatic mutation.

3.1.1. Therefore the HGNC symbols of the 175 GOI where again submitted to Ensembl in order to extract the Ensembl Ids for the genes (the ICGC doc

only reports Ensembl Ids).

Those that had Ensembl ID's were retained because it would later be needed to confirm that the hits found where representative of my GOIs. Also the chromosome and strand was extracted.

3.1.2. The **sort -k2,2 173_GOI_with_unique_Ensembl_ID.txt -u > 175_ccRCC_disease_genes_with_or_without_ensembl_ID_UNIQ.txt** was used to confirm that the HGNC Ids were unique and **wc -l** was used to confirm that the number of ccRCC genes I was working with was 173.

3.1.3. The regional genomic coordinates of the 5'-UTR, 3'-UTR, introns, CDS and promoter regions (1kB upstream of the TSS) were retrieved from UCSC using the Table function (http://genome.ucsc.edu/cgi-bin/hgTables) and the GRCh37/hg19 human genome reference.

The UCSC track was chosen over Ensembl and Refseq, because the latter two could not pick up some of the 175 GOI. The format is:

Chromosome genomic_ start genomic_end ucsc_description score strand

3.1.4. A proportion of coordinates for each genomic region were manually checked to see if they were located within the regions specified by UCSC. For each file there were more entries than the original number of genes, but this was accounted for by the presence of many splice variants per gene.

3.1.5. **sort -k1,1 -k2,2 -k3,3 -k6,6 promoter_175_ccRCC_genes.txt -u > promoter_175_ccRCC_genes_UNIQ.txt** was used to extract the unique lines based on chromosome, genomic range and strand.

3.2. Non-disease genes

3.2.1. The transcript IDs were submitted to UCSC's table browser to extract the 5'UTR, 3'UTR, promoter, CDS and intronic regions of the 173 non-disease genes.

The unique bed range coordinates were again extracted using:

sort -k1,1 -k2,2 -k3,3 -k6,6 promoter_173_non_disease_genes.txt -u > promoter_173_non_disease_genes_UNIQ.txt

4. SV Discovery

4.1. Disease genes

4.1.1. Wrote a script to find all the somatic variants for distinct genomic regions

Python: icgc_variants.py

The hits were first printed to Stdout with the bed range in order to check if the variant position did fall within the bed range. This was however not retained in the files of interest.

4.1.2. Many duplicate genes existed due to one gene having many different Ensemble transcript ID's. The unique entries were maintained and sorted by using a Linux command:

sort -t'Ctrl+V+Tab' -k1,1 -k2,2 -k5,5 -k8,8 -k9,9 -k10,10 -k16,16 -u SV_results_173ccRCC_3UTR.txt > SV_results_173ccRCC_3UTR_UNIQ.txt

4.1.3. However, because there are many overlapping genes in the human genome, many of the hits based on genomic position were not within the genes within the GOI list. The Ensembl and HGNC Ids were therefore retrieved using BioMart Ensembl Grch37 (12 June 2015) and matched with the SV file to retrieve only those linked to the genes of interest

Python script:

checking_if_ensembl_IDs_of_173_GOI_are_in_SV_hits.py

4.1.4. Duplicates were eliminated based on a unique genomic position, gene, donor ID, structural annotations, alleles and Ensembl ID.

sort -k1,1 -k2,2 -k3,3 -k4,4 -k5,5 -k7,7 -k8,8 -k9,9 -k10,10 -k11,11 actual_hits_175_ccRCC_introns.txt -u > actual_hits_175_ccRCC_introns_UNIQ.txt

There were no duplicates

4.1.5. The number of unique somatic mutations based on just the genomic position, Ensembl ID and alleles, were also subtracted because this was required for the density study

sort -k1,1 -k2,2 -k3,3 -k4,4 -k7,7 -k8,8 -k9,9 -k11,11 actual_hits_175_ccRCC_promoter.txt -u > actual_hits_175_ccRCC_promoter_UNIQ_positions.txt

4.2. Non-disease genes

4.2.1. I wrote a script to find all the somatic variants for distinct genomic regions

Python: icgc_non_disease_genes_variants.py

The hits were first printed to Stdout with the bed range in order to check if the variant position did fall within the bed range. This was however not retained in the files of interest.

4.2.2. Many duplicate genes existed due to one gene having many different Ensemble transcript ID's. The unique entries were maintained and sorted by using a Linux command:

sort -t'Ctrl+V+Tab' -k1,1 -k2,2 -k3,3 -k5,5 -k8,8 -k9,9 -k10,10 -

k16,16 -u SV_results_173_non_disease_introns.txt > SV_results_173_non_disease_introns_UNIQ.txt

4.2.3. However because there are many overlapping genes in the human genome, many of the hits based on genomic position were not within the genes within the GOI list. The Ensembl and HGNC Ids were therefore retrieved using BioMart Ensembl Grch37 (12 June 2015) and matched with the SV file to retrieve only those linked to the genes of interest

Python script:
checking_if_ensembl_IDs_of_173_GOI_are_in_SV_hits.py

4.2.4. Duplicates were eliminated based on a unique genomic position, gene, donor ID, structural annotations, alleles and Ensembl ID.
sort -k1,1 -k2,2 -k3,3 -k4,4 -k5,5 -k7,7 -k8,8 -k9,9 -k10,10 -k11,11 actual_hits_175_non_disease_3UTR.txt -u > actual_hits_175_non_disease_3UTR_UNIQ.txt
There were no duplicates

4.2.5. The number of unique somatic mutations based on just the genomic position, Ensembl ID and alleles, were also subtracted because this was required for the density study using:
sort -k1,1 -k2,2 -k3,3 -k4,4 -k7,7 -k8,8 -k9,9 -k11,11 actual_hits_175_non_disease_3UTR.txt **-u >** actual_hits_175_non_disease_3UTR**_UNIQ_positions.txt**

5. Density of hits

5.1.The density of the hits were check based on the unique position of the somatic mutation, its genes ID and alleles involved using
Python script: density_hits_ccRCC.py

5.2. The **density_hits_ccRCC.py** script was checked by extracting the genomic bases for **JUST** the PTEN gene (ENSG00000171862) from UCSC. Duplicates were also removed by sort –t 'cntrl+v+tab' –k1,1 –k2,2 –k3,3 -u PTEN_gene > PTEN_gene_*_UNIQ. The bases were manually added per distinct genomic region to calculate the total number of bases. The density was also manually checked by dividing the number of hits over the total number of bases.

5.3. The **density_hits_ccRCC.py** script was modified to accommodate the non-disease genes.

6. RegulomeDb

6.1. In order to enter the data to RegulomeDB the end coordinates of the somatic mutations had to be converted to zero based format using a **Python script: add_one_to_coordinate2_RCC.py**

6.2. Thereafter the coordinates were submitted to RegulomeDB
However the introns file was too big and had to be split into manageable sizes for input
split -l 200 regulome_coord_coverted_ccRCC_introns.txt was used, which created 20 files named xaa – xau

6.3. The intronic file was stored separately in
/ICGC/Regulome_annotations/Results/Annotations_from_web_interface/173_ccRCC_genes/introns/regulomedb_results_introns_xaa.bed up until **regulomedb_results_introns_xau.bed.** The same was done for the .bed files.
sort -k4,4 -k1,1n -k2,2n -k3,3n regulomedb_results_introns_xa*.bed -u > regulomedb_results_ccRCC_introns.bed was used to combine the

files and **wc -l** was used to check that the original amount that was submitted was in the output file

6.4. The input files of the same script were modified to accommodate the non-disease variants using the **add_one_to_coordinate2.py** script.

7. Matching RegulomeDB results to ICGC SV

7.1. I wrote a script to check link up the RegulomeDB annotation with their original ccRCC variants in order to eventually extract the total number of variants and the number of deleterious variants that are linked to a specific tumour; as well as other trends in terms of the mutational landscape. I also used this script in order to see if any of mutations were indels.

Python script: variants_linked_to_regulome_score_and_annotation.py

7.2. The output was checked to see if there were the same amount as in the original actual hits file with a unique Ensembl ID, patients ID, genomic positions and alleles.

7.3. The **variants_linked_to_regulome_score_and_annotation.py** script was modified for the non-disease genes

8. Analysis of somatic mutations

8.1. The non-coding ccRCC variant were combined in order to see which genes came up frequently in different donors , which chromosomes are often affected etc.

sort -k1,1 -k2,2 -k3,3 -k4,4 -k5,5 -k7,7 -k8,8 -k9,9 -k10,10 -k11,11 -

k12,12 173_ccRCC_regulome_annotation_introns_sv.txt 173_ccRCC_regulome_annotation_3UTR_sv.txt 173_ccRCC_regulome_annotation_5UTR_sv.txt 173_ccRCC_regulome_annotation_promoter_sv.txt -u > combined_non_coding_regulome_annotation_of_sv.txt

8.2. This data was placed into a pivot table in order to do the analysis.

8.3. The ccRCC CDS data was also copied to the pivot containing all non-coding to do a comparison on which genomic regions contained the most mutations.

8.4. The deleterious variants were then extracted and placed into a separate excel sheet.

8.5. A Vlookup was carried out in Excel to compare the number of all non-coding somatic mutations, CDS somatic mutations and deleterious non-coding somatic mutations.

8.6. To see if the reason why some genes accumulated more mutations was simply due to the gene being longer and hence having more targets for variation, a scatter plot was drawn up using Excel. The gene lengths of the 60 genes with contained RegulomeDB deleterious mutations were submitted to BioMart Ensembl (hg19) and the gene lengths were extracted from the Structures attribute. The difference in gene lengths were calculated in excel and the deleterious mutations were placed in a column next to it in excel. Excels built-in scatter plot was used to draw up the scatter plot.

8.7. The same procedure was followed for all non-coding mutations (=VLOOKUP(A2,$E$2:$F$166,2,FALSE)).

8.8. The genes within which one patient had multiple mutations (TRIO, FRYL and AKT1) and the genes where several patients had mutations at the exact

same location (ADCY1, ANK3, CUBN, VWF and VHL) were also scrutinized to see if they generally occurred in the longer genes

9. BioMart Gene Ontology annotations

9.1. For the 164 genes with non-coding somatic mutations the GO was extracted for Ensembl BioMart

From the Attributes: (Features) External option, the GO term and GO definition were chosen.

9.2. The GO annotation of the TRIO, FRYL and AKT1, ADCY1, ANK3, CUBN, VWF and VHL genes were copied from the **original** file and pasted onto its own sheet.

The **Conditional Formatting -> Highlight cell rules -> Text that contains** was used in order to highlight the cells containing the text I specified e.g. 'proliferation'

9.3. Nine genes with one mutations as opposed to multiple mutation in many genomic regions were randomly selected to see if they also participate in as many cancer-related activities

MDC1 (adhesion)
CA9 (transcription)
SCD (none)
STK11 (apoptosis, signalling, receptor activity
FAM107A (none)
PRKCD (apoptosis, signalling, receptor activity
KRT7 (none)
TLR3 (apoptosis, signalling, receptor activity
PTTG1 (transcription)

9.4. The eight RCC genes with no SM were also selected and the gene lengths and GOs were retrieved.

The genes were:

CCKBR

CLDN7

HLA-A

HSPA8

LGALS3

NONO

PECAM1

RAF1

10. Clinical Information

10.1. The donor id with the number of mutations per mutation category (ALL non-coding, non-coding deleterious and all CDS mutations) were copied to a separate file.

This was then used with the python script **match_mutations_to_donor_ID.py** to link the number of mutations per category to the donor ID.

10.2. The donor IDs were also checked for common genes or common genomic position in order to ascertain if variants in certain genomic regions resulted in a more severe phenotype.

11. TFBS analysis

11.1. The file containing the combined ccRCC non-coding variants was used in the TFBS analysis in order to see whether these SM fell within

TFBSs

A python script was used to extract all TF and TFBS information for the non-coding somatic mutations.
Python script: **som_mut_in_TFBS.py**

11.2. The results were placed into a pivot table
The deleterious mutations were separated to place into the previous Vlookup table.

11.3. The script was modified to accommodate the ccRCC CDS mutations

11.4. A pivot table was also created for the CDS TFBS data in order to extract donor to TF and gene to TF relationship data

11.5. Similarly the **som_mut_in_TFBS.py** script was modified for the combined non-coding non-disease gene variants and for the non-disease genes CDS mutations

11.6. These results for the both were saved separately into pivot tables.

11.7. The TFBS data was added to the Vlookup sheet with the results for the somatic variants and charts were made to see the comparison in terms of how the somatic mutations affected the TFBSs.

11.8. The genomic position were also analysed to see if any individuals had variants at the exact same genomic positions.

12. Methylation

12.1. I ran the **methylation_in_non_coding_regions.py** to find the methylation statuses of the promoter regions for the 173 disease and non-disease genes by modifying the input and output files of the specific variants.

12.2. The HCNC symbols were manually added to all files using UCSC and Ensembl

12.3. The genomic regions were placed into the methylation file (e.g. **introns** if the variant originally came from the intron bed file) and the bed regions were deleted in excel. The methylation files were combined using

sort –t'Ctrl + V + Tab' –k1,1 –k2,2 –k3,3 –k,4 –k8,8 –k10,10 –k13,13 –k14,14 methylation……. -u > combined_methylation_in_non_coding_regions.txt

(the genomic region from the methylation data file as well as beta values was ignored for now)

12.4. The results were placed into a pivot table.

12.5. 16 RCC genes had aberrant methylation. The GO annotations for these genes were also observed an pasted in into a separate Excel sheet.

13. Gene Expression and methylation data

13.1. The ccRCC methylation data was coupled to the gene expression data using

Python: **python gene_expr_at_diff_methyl_regions.py**

13.2. The significant methylation versus gene expression data was recorded (inverse proportionality between gene expression and diff methylation) Python: **significant_correlat_gene_exp_and_methyl.py**

13.3. The same was done for the non-disease genes by modifying the scripts with the input files of non-disease data

14. Gene Expression and non-coding somatic mutations

A lot of processing had to be done before the GE data could be extracted. First the entries with the 173 disease and non-disease genes had to be extracted. The gene expression file from ICGC had no matching normal tissue that was sequenced so differential analysis (Fold change) could not be determined. Hence the gene expression data from COSMIC/TCGA was used just to check simply on gene level, if the gene was found to be frequently differentially methylated in many ccRCC tumours.

a) The Differential gene expression for all cancers were retrieved from **COSMIC Whole genomes** http://cancer.sanger.ac.uk/wgs/files?data=/files/grch38/cosmic/v73/CosmicCompleteGeneExpression.tsv.gz, Accessed June 2015.

b) The **sort -t 'Ctrl + v + Tab ' -k2,2 CosmicCompleteGeneExpression.tsv -u > uniq_patients_CosmicCompleteGeneExpression.tsv** was used to ascertain the number of patients for which the gene expression data was recorded.

c) However since TCGA combined all the gene expression data from all their cancer patients there were a total of 8348 unique patient ID's.

d) Hence, the tissue bar codes for ccRCC tissue were retrieved from TCGA and a Python script was used to extract the GE data for ccRCC tissues/genes only.

Python: extract_ccRCC_gene_exp_data.py

e) To make sure that the use of study subjects was unbiased, the GE data for 95 patients were extracted to be used for both disease and non-disease genes.

Python script: extract_GE_data_for_95_patients_from_521_patients.py

f) The same Python script was then used to extract the GE data for the ccRCC non-coding and CDS variants and the non-disease non-coding and CDS variants and the input and output file names were changed as necessary.

Python script: SV_and_COSMIC_gene_exp.py

g) I created a pivot table for the RCC genes for further analysis.

15. STRING-DB

15.1. The genes with the most deleterious, non-coding somatic mutations (2 or more, n = 31, genes), the top 30 genes with the most TFBS disruptions in the non-coding regions (33 or more disruptions), the genes which incurred differential methylation in their promoters (n = 17) and the three TFs of which the TFBSs are commonly disrupted in the deceased patients were extracted and duplicates were removed in Excel. A total of 57 unique genes/proteins were therefore submitted to String-DB and the analysis was carried out using the highest confidence score in String-DB.

15.2. The Biological Process was considered as the Gene ontology: Negative regulation of biological process contained the most molecules.

15.3. Then the network was zoomed out in order to establish hub proteins. Experimental was used as the sole parameter. UBC was identified.

15.4. Some of the genes that were unattached in the network were checked individually on the highest confidence and experimental evidence only

PLEC linked to ITGB4 (integrin receptor protein for laminin)

TRIO linked to UBC via RAC1 (GTPase which in its active state regulates apoptotic cells)

MLL2 linked to RBBP5 retinoblastoma binding protein 5 that plays a crucial role in cell differentiation and regulates H3K4 methylation at important developmental loci

DOCK2 also linked to UBC via RAC1 (ras-related C3 botulinum toxin substrate 1)

OGG1 was also linked to UBC via POLH DNA polymerase specifically involved in DNA repair.

15.5. When the pathways that could be perturbed by these mutations were considered, ten of the molecules functioned in the phosphoinositide 3-kinase (PI3K)/ Akt pathway.

16. Allele Frequency

a) The AF of the combined non-coding ccRCC and non-disease genes variants were checked as well as the variants in the CDS regions for both categories.

b) Since the combined non-coding file was big (over 4000 variants) and the 1000 genomes dataset doesn't look at the donor ID, the file was sorted to filter out duplicate entries by ignoring the donor ID.

sort -k1,1 -k2,2 -k3,3 -k4,4 -k7,7 -k8,8 -k9,9 -k11,11 combined_non_coding_regulome_annotation_of_sv.txt -u >

combined_non_coding_regulome_annotation_of_sv_UNIQ_positions.txt

This brought the number down from 4385 to 4226

c) A python script was used to extract the variants in the files: ccRCC non-coding somatic mutations, ccRCC CDS mutations, non-disease non-coding somatic mutations and non-disease CDS mutations, if the allele for that variant was found in the 1000Genome dataset.

Python script: **1000genomes_variant.py**

16.1. Most of the variants were not found so several random variants (~30) were checked using **grep variant position 1000genomes_filename**

16.2. The variants were also checked using Ensembl BioMart (hg19) using the Ensembl Variants database and the Homo sapiens Somatic Short Variants (SNPs and indels) (GRCh37.p13) dataset.

The filters section was used to choose the chromosome and to enter the variant start and end region. To ascertain that the search was being carried out correctly and that the results indicate that my variant was definitely not found, the range was increased to include about 100000 bases before and after the start and end coordinate, respectively. The generated many results. The View all feature was used together with Control+Find in order to check if my variant was not in this list of variants.

16.3. Finally, the variants were then run through Variant Effect Predictor (VEP) on 06/11/2015 to confirm that these variants were novel with the following criteria for output

http://grch37.ensembl.org/Homo_sapiens/Tools/VEP