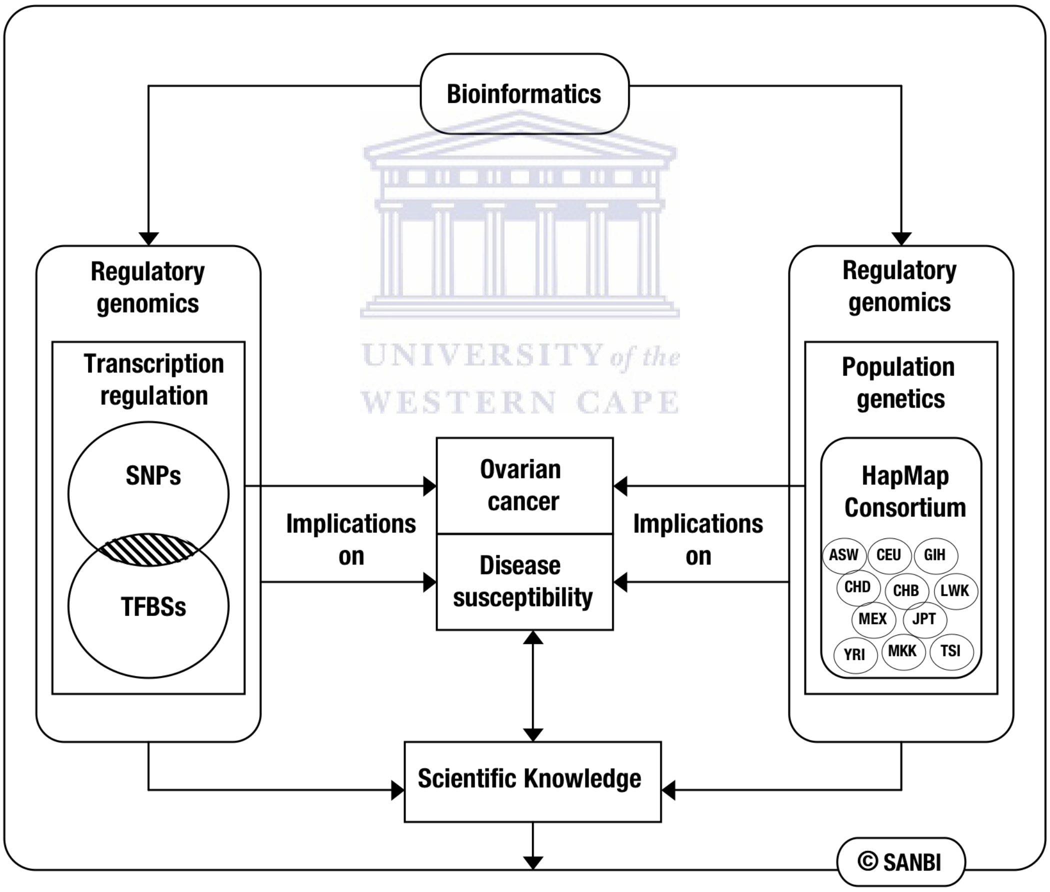
Incidence and Regulatory Implications of Single Nucleotide Polymorphisms among Established Ovarian Cancer Genes



KAVISHA RAMDAYAL

UNIVERSITY OF THE WESTERN CAPE

Incidence and Regulatory Implications of Single Nucleotide Polymorphisms among Established Ovarian Cancer Genes

Thesis presented in fulfillment of the requirements for the Degree
of *Magister Scientiae in Bioinformatics* at the South African National
Bioinformatics Institute (SANBI), Faculty of Natural Sciences,
University of the Western Cape (UWC)

by

Kavisha Ramdayal

SEPTEMBER 2009

Supervisor: Doctor Heikki Lehväslaiho
Co-Supervisor: Professor Vladimir B Bajic

Acknowledgements

Thanks to the National Bioinformatics Network (NBN) and National Research Foundation (NRF) for their provision of funding, without which this work would not have been possible.

Thanks to Dr. Heikki Lehväslaiho, Prof. Vladimir B Bajic and Dr. Adam Dawe for their learned advice, guidance and time with this project. I am grateful to the folks at SANBI and the NBN for granting me access to their reservoir of bioinformatics expertise and knowledge. My gratitude to the following people for their help in the editing and compilation of this thesis: Dr. Ashley Pretorius, Mark Wamalwa, Sebastian Schmeier, Ulf Schaefer, Mario Jonas, Ferial Mullins, Maryam Salie, Bonnie Opperman, Feziwe Mpondo, Sarah Wambui, Magbubah Essack, Monique Maqungo, Frederick Kamanu, Musa Nur Gabere, Dale Gibbs & Peter van Heusden.

Incalculable thanks to my family for their countless sacrifices, support, encouragement, advice and patience throughout my study career and to the great many friends and family I have made along the way, for their consolation, technical support and hospitality.

To my better half, Faisel, thank you for your uncompromising love, support and patience. You were everything I needed at every stage of this work.

Keywords

Ovarian cancer

Susceptibility

Single nucleotide polymorphism

SNP@Promoter

F-SNP

PupaSuite

Transcription factor binding site

Transcription factor

MATCH^{TM}

HapMap

Allele

Abstract

Incidence and Regulatory Implications of Single Nucleotide Polymorphisms among Established Ovarian Cancer Genes

K. Ramdayal

Magister Scientiae in Bioinformatics, Thesis,
Department of Biotechnology, University of the Western Cape

OVARIAN cancer research focuses on answering important questions related to the disease, determining whether new approaches are feasible to contribute towards improving current treatments or discovering new ones. This study focused on the transcriptional regulation of genes that have been implicated in ovarian cancer, based on the occurrences of single nucleotide polymorphisms (SNPs) within transcription factor binding sites (TFBSs). Through the application of several *in silico* tools, databases and custom programs, this research aimed to contribute toward the identification of potentially bio-medically important genes or SNPs for pre-diagnosis and subsequent treatment planning of ovarian cancer. A total of 379 candidate genes that have been experimentally associated with ovarian cancer were analyzed. This led to the identification of 121 SNPs that were found to coincide with putative TFBSs potentially influencing a total of 57 transcription factors that would normally bind to these TFBSs. These SNPs with potential phenotypic effect were then evaluated among several population groups, defined by the International HapMap consortium resulting in the identification of three SNPs present in five or more of the eleven population groups that have been sampled.

After analysis of the allele frequencies of each of the three SNPs and comparison to the ancestral alleles of each, SNPs rs12928665, rs20577 and rs4150842, present on genes *CIITA*, *TNFRSF10A* and *E2F5*, were observed as the only three SNPs that potentially influence the binding of the CDP, CP2/LBP-1c/LSF and/or C/EBP transcription factors at their putative transcription factor binding sites. These SNPs may be considered as potential diagnostic markers for the prognosis of ovarian cancer in high-risk women from specific population groups.

Furthermore, this study has highlighted a computational approach for the identification of SNPs coinciding with TFBSs that may play a role in the regulatory mechanisms encoded for by cancer-associated genes. This approach may be applied to the elucidation of SNPs influencing TFBSs in other complex diseases given the time and system requirements.

September 2009

Declaration

I declare that "**Incidence and regulatory implications of single nucleotide polymorphisms among established ovarian cancer genes**" is my own work that it has not been submitted for degree or examination at any other university, and that all the resources I have used or quoted, and all work which was the result of joint effort, have been indicated and acknowledged by complete references.

Kavisha Ramdayal
21 August 2009

Signed: Ramdayal

Table of Contents

List of Tables

List of Figures

Abbreviations

ASW	African ancestry in Southwest USA
bp	Base pair
CEU	Utah residents with Northern and Western European ancestry
CHB	Han Chinese in Beijing, China
CHD	Chinese in Metropolitan Denver, Colorado
CSS	Core Similarity Score
FISH	Fluorescent *In Situ* Hybridization
GIH	Gujarati Indians in Houston, Texas
JPT	Japanese in Tokyo, Japan
kb	Kilobase
LWK	Luhya in Webuye, Kenya
MEX	Mexican ancestry in Los Angeles, California
MKK	Maasai in Kinyawa, Kenya
MSS	Matrix Similarity Score
PFM	Position Frequency Matrix
PSSM	Position-Specific Scoring Matrix
PWM	Position Weight Matrix
RT-PCR	Reverse Transcriptase-Polymerase Chain Reaction
SNP	Single Nucleotide Polymorphism
TFBS	Transcription Factor Binding Site
TF	Transcription Factor
TSI	Toscans in Italy
TSS	Transcription Start Site
YRI	Yoruban in Ibadan, Nigeria

Definitions

Core Similarity Score

A measure of the quality of a match between the core sequence of a matrix (i.e. the five most conserved nucleotide positions within a matrix) and a part of the input sequence (Kel AE *et al.*, 2003).

Fluorescent *In Situ* Hybridization

A technique used to detect and localize the presence or absence of specific DNA sequences on chromosomes, in which a DNA probe is labeled with a fluorescent dye and then hybridized with target DNA usually on a microscopic slide, which can then be viewed under a florescent microscope (Nath *et al.*, 2000).

Haplotype

A combination of alleles at different markers along the same chromosome that are inherited as a unit (Crawford *et al.*, 2005).

Hardy-Weinberg

A classical mathematical principle in population genetics that describes the expected frequencies of genotypes for one locus after one generation of random mating if the allele frequencies in the parents are known (Hey J *et al.*, 2003).

Immuno-histochemistry

A technique for recognizing proteins according to the binding of specific antibodies to antigens in biological tissue (Rubinstein WS *et al.*, 2008).

Linkage analysis

Analysis of the segregation patterns of alleles or loci in families or experimental crosses commonly used to map genetic traits by testing whether a trait co-segregates with genetic markers whose chromosomal locations are known (Lusis *et al.*, 2008).

Matrix Similarity Score

This score describes the quality of a match between a matrix and an arbitrary part of the input sequences (Kel AE *et al.*, 2003).

Pharmacogenetics

The branch of science concerned with the influence of genetic variation on the effectiveness and side effects of drugs (Roche, 2008).

Reverse Transcriptase-Polymerization Chain Reaction Is a method of simultaneous DNA quantification and amplification that is highly sensitive technique for the detection and quantitation of messenger RNA (MedicineNet.com, 2009a & EverythingBio, 2007).

Single Nucleotide Polymorphisms Are heritable differences in individual base pairs of DNA that are distributed randomly over the genome (Roche, 2008).

Transcription factors A group of proteins that are essential for the process of transcription, involved in the formation of the pre-initiation complex and in the recruitment of RNA polymerase (Vaquerizas *et al.*, 2009).

Variation A measure of genetic variation obtained by a number of base differences between two genomes, divided by the total number of base pairs being compared (Chakravarti, 2001).

Western blotting A technique in molecular biology, used to separate and identify proteins based on their ability to bind to specific antibodies (MedicineNet.com, 2009b & Molecular Station, 2009).

Preface

The *Incidence and regulatory implications of single nucleotide polymorphisms among established ovarian cancer genes* was undertaken at the South African National Bioinformatics Institute (SANBI) situated at the University of the Western Cape (UWC) between November 2007 and May 2009 under the supervision of Prof. Heikki Lehväslaiho and Prof. Vladimir B. Bajic. The Microsoft Word (.doc) or Portable Document Format (.pdf) versions of this work can be requested from the author at the following address: kramdayal@gmail.com.

The focus of this study was to provide an insight into the use of SNPs as potential biomarkers for ovarian cancer and the possibility of using this information in the early detection and/or treatment planning of the disease. The thesis consists of two major parts.

The first part comprises of the identification of SNPs having potential phenotypic effect that coincide with TFBSs within genes that have been implicated (i.e. experimentally proven) in ovarian cancer. The second part consists of the exploration of these possible functional SNPs (identified in part one) that occur in several population groups defined by the International HapMap consortium, qualifying them as possible diagnostic markers that may be applied to the early detection of ovarian cancer.

September 2009 Kavisha Ramdayal

CHAPTER 1

Introduction

OVARIAN cancer affects more than 200,000 women around the world every year (Helm *et al.*, 2009). It has the highest mortality rate of all cancers of the female reproductive system and is the fifth leading cause of cancer-related death among women in the USA alone (U.S. National Institutes of Health, 2001). Due to the lack of early symptoms and proven ovarian cancer screening tests, ovarian cancer is often diagnosed at an advanced stage when the cancer has spread beyond the ovary (U.S. National Institutes of Health, 2001 & Helm *et al.*, 2009). Although recent scientific discoveries have led to new insights into cancer prevention, detection and treatment in the past, gynecological cancers are still claiming the lives of hundreds of thousands of women (U.S. National Institutes of Health, 2001).

1.1 Cancer

“Cancer is an extremely complex, heterogeneous disease, which displays a degree of complexity at the physiological, tissue and cellular levels”, (Wang *et al.*, 2007). Cancers arise from a loss of normal growth control and can originate almost anywhere in the body in the form of carcinomas, sarcomas, lymphomas or leukemias, with the most common being carcinomas that arise from cells covering the external or internal surfaces of the body, e.g. lung, breast and colon cancers (U.S. National Institutes of Health, 2001 & Bupa, 2007).

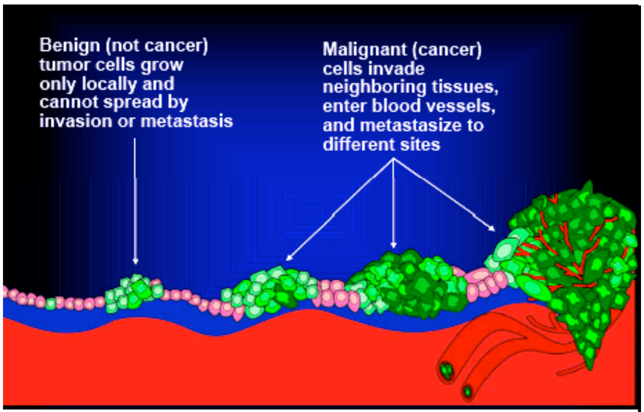

Benign (not cancer) tumor cells grow only locally and cannot spread by invasion or metastasis
Malignant (cancer) cells invade neighboring tissues, enter blood vessels, and metastasize to different sites

Figure 1.1 Malignant versus benign tumors (U.S. National Institutes of Health, 2001). Malignant tumors are a more serious health problem than the benign tumors, as these cancer cells can spread to distant parts of the body and can therefore be potentially life threatening (U.S. National Institutes of Health, 2001 & The New York Times, 2009).

Cancer occurs in the body in the form of tumors that are classified as being either benign or malignant, depending on whether or not abnormal cell growth can spread by invasion or metastasis (Bupa, 2007). Benign tumors are tumors that cannot spread by invasion or metastasis and can only grow locally whereas malignant tumors are capable of spreading through invasion and metastasis as described in Figure 1.1 (U.S. National Institutes of Health, 2001).

There are over 200 different kinds of cancer, named according to the part of the body where the cancer or tumor originates (e.g. brain, breast, cervix, colon, lung, ovary, prostate, skin, etc.) (Cancer Association of South Africa, 2008). Ovarian cancer falls within the category of gynaecological cancers, as it begins in the reproductive system of women with the most common types of gynaecologic malignancies being cervical, ovarian, and endometrial (uterus) cancer (Cancer Research UK, 2008 & CancerIndex, 2003).

1.2 Ovarian Cancer

Ovarian cancer develops within the cells of one or both of the ovaries present in the uterus of a women, with 90% of these beginning in the cells that cover the outer surface of the ovary (ecancermedia, 2008).

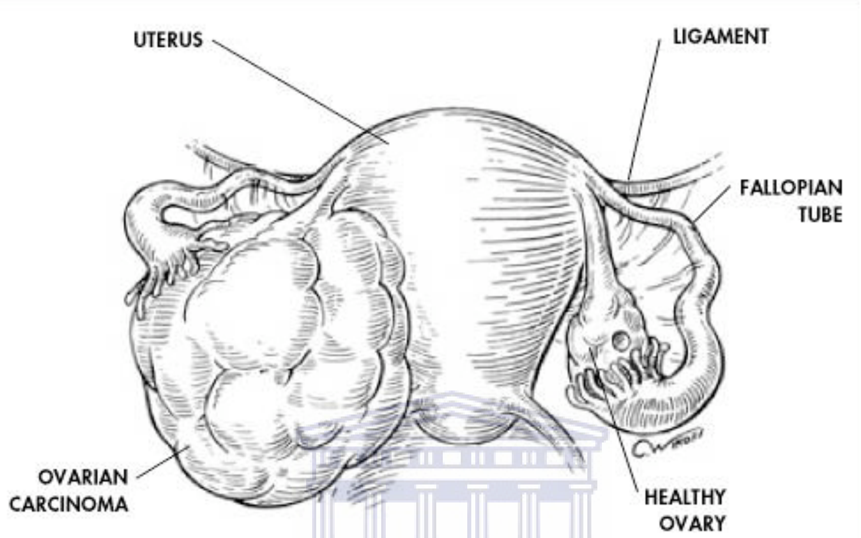


Figure 1.2 Healthy ovary compared to cancerous ovary (HealthSquare.com, 2009). Most often than not ovarian cancer develops without any troubling symptoms, leading to the tumor developing to a stage when it is eventually detected as a mass after physical examination of the pelvic area (HealthSquare.com, 2009).

Due to the lack of early detection strategies, most patients with ovarian cancer are diagnosed when the disease has progressed to an advanced stage, when there is only a 5-year overall survival rate of approximately 20% (Coukos *et al.*, 2008).

1.2.1 Stages of Ovarian Cancer

The staging of cancer is a crucial factor to consider in the development of a treatment regimen for each patient to improve the treatment outcome (National Ovarian Cancer Coalition, 2009). Ovarian cancer is divided into four major stages that are determined by the progression of the disease within the body as described by Figure 1.3 below.

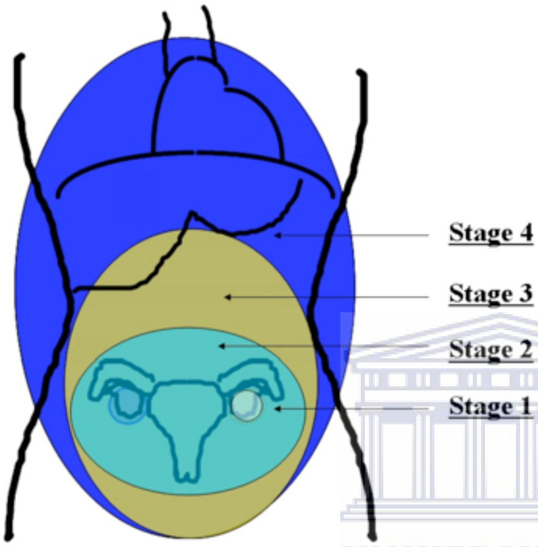


Figure 1.3 Ovarian cancer stages of disease progression (National Ovarian Cancer Coalition, 2009). As the disease progresses from Stage 1 to 4, survival rates have been shown to decline dramatically from an initial ±90% survival rate at Stage 1 to a less than 5% survival rate at Stage 4 (National Ovarian Cancer Coalition, 2009).

Stage 1: One or both ovaries is/are cancerous.

Stage 2: One or both ovaries is/are cancerous and the disease has spread to the uterus, fallopian tubes or other parts in the pelvic area of the body.

Stage 3: One or both ovaries is/are cancerous and the disease has spread to the lymph nodes or other parts of the body inside the abdomen.

Stage 4: One or both ovaries is/are cancerous and the disease has spread outside the abdomen and/or to the liver (National Ovarian Cancer Coalition, 2009).

1.2.2 Types of Ovarian Cancer

There are approximately 30 known types and subtypes of ovarian cancer malignancies, each with its own biological characteristics that may be grouped into three major categories (*Table 1.1*) according to the kind of cells from which they were formed (OncologyChannel, 2009).

Table 1.1 Three major types of ovarian cancer. Although some tumors that are found adjacent to ovarian tissues are viewed and treated as ovarian cancer (e.g. cancer of the membrane lining the walls of the pelvic cavity next to the ovaries) the three major categories include the following (OncologyChannel, 2009):

Type	Origin
Epithelial tumors	Cells that line or cover the ovaries
Germ cell tumors	Cells that develop into eggs within the ovaries
Sex cord-stromal cell tumors	Connective cells that hold the ovaries together and produce hormones

Epithelial tumors account for about 90% of all ovarian cancers, occurring in women between the ages of 30 and 80, and can be further subdivided into serous, endometrioid, mucinous and clear cell tumors (OncologyChannel, 2009). From all the diagnosed cases of these tumors, approximately 50% of serous tumors, 80% of endometrioid tumors, 5% of mucinous tumors and nearly all clear cell tumors are found to be malignant (OncologyChannel, 2009). Unlike epithelial tumors however, 60-70% of patients with germ cell tumors are diagnosed at stage 1 of the disease (*Figure 1.3*), whereas 75% of epithelial ovarian cancers are diagnosed when the disease has already progressed to stages 3 or 4 (OncologyChannel, 2009).

1.2.3 Risk Factors

"The overall lifetime risk of any woman developing ovarian cancer is low, however certain factors can increase that risk by 11% to 65%", (Anderson, 2009). The use of hormone-replacement therapy (HRT) may be one of the factors that contributes towards the development of ovarian cancer, as indicated by Blagden *et al*. (2008), who estimated this to be the cause of an additional 1300 cases of ovarian cancers since 1991. Another contributing factor identified by the Million Women Study, to be a cause of ovarian cancer, among a number of other tumors, was obesity (Blagden *et al.*, 2008).

Unlike other cancers that arise from a range of origins, mostly environmentally or lifestyle linked (Hertel *et al.*, 2008), the strongest risk factor for ovarian cancer comes from family history as a result of inherited cancer susceptibility genes such as *BRCA1* or *BRCA2* (Anderson, 2009; Blagden *et al.*, 2008; King *et al.*, 2003a). King *et al.* (2003a) found that risks (related to women that inherited mutations in the tumor suppressor genes *BRCA1* and *BRCA2*) appeared to be increasing with time, and found lifetime risks of ovarian cancer amounting to 54% for *BRCA1* and 23% for *BRCA2* mutation carriers (King *et al.*, 2003a).

Table 1.2 Factors responsible for an increased risk for ovarian cancer. Other factors that have been reported to cause an increased risk for ovarian cancer include the following (Anderson, 2009; Bupa, 2007 & ecancermedia, 2008):

Factor	Description
Age	Ovarian cancers occur most often after menopause, with 50% of these cases found in women above the age of 65
Children	There are slightly increased risks for women that do not have children or had their first child after the age of 30
Diet	Although obesity has been described as a contributing factor by Blagden *et al.*, (2008), some studies suggest that even a high fat diet may increase ovarian cancer risk
Menstrual Cycles	Women who started having periods early (before 12 years old) or those who go through menopause after the age of 50 have a slightly increased risk for ovarian cancer
Fertility Drugs	Prolonged use of fertility drugs may increase the risk for ovarian cancer, however infertility also increases the risk, even without the use of fertility drugs
Genetics	In addition to inherited gene mutations of the *BRCA1* and *BRCA2* genes, an inherited disease known as Hereditary Nonpolyposis Colon Cancer (HNPCC) increases the risk for ovarian cancer
Breast Cancer	Women that have had breast cancer have a higher risk for ovarian cancer
Estrogen-Replacement Therapy (ERT)	Long-term use (10 or more years) of ERT after menopause has been shown by most studies to slightly increase the risk for ovarian cancer
Carcinogen Exposure	These include asbestos, benzene and radioactive materials
Weak Immune System	This may be the result of medicines that suppress the immune system, e.g. high doses of radiation in the case of radiotherapy that may be employed to target another cancer

1.2.4 Symptoms

Symptoms are usually the result of the cancer growing and causing pressure or pain as a result and may include any of the following attributes (Bupa, 2007 & ecancermedia, 2008):

(1) Swelling of the stomach
(2) Abdominal pain
(3) Digestive problems (e.g. indigestion, constipation, appetite loss, etc.)
(4) Unexpected weight gain/loss
(5) Frequent need to urinate
(6) Unusual vaginal bleeding
(7) Back or leg pain
(8) Bowel or bladder changes

However, although there are a few identifiable symptoms of early-stage ovarian cancer, these are thought to be subtle or absent, making diagnosis difficult and patients reluctant to seek help (Anderson, 2009 & Bupa, 2007).

1.2.5 Diagnosis

Early stages of ovarian cancer are difficult to detect typically because the disease has very vague signs or symptoms, many of which could easily be mistaken or not considered as an indication of ovarian cancer (Anderson, 2009).

Table 1.3 Existing tests to confirm the presence of ovarian cancer. Traditionally ovarian cancer diagnosis techniques include the following tests or procedures (Bupa, 2007; ecancermedia, 2009; HealthSquare.com, 2009 & National Ovarian Cancer Coalition, 2009):

Technique	Description
Imaging studies	Will detect if there is a mass present in the pelvis, but cannot convey if it is cancer
Ultrasound	Uses sound waves to create an image on a video screen, distinguishing between tumors and normal tissue based on their varying reflection of sound waves
Computed Tomography (CT) scan	Uses a computer and an x-ray beam to take a series of pictures of the body from many angles and then combines them into a detailed image on the computer
Magnetic Resonance Imaging (MRI)	The MRI displays a cross-sectional picture of the body using radio waves and strong magnets instead of x-rays
CA125 Assay	CA125 is a protein that is a tumor marker and is measured via a blood sample. Elevated levels of CA125 are associated with the presence of ovarian cancer, but have been shown to produce numerous false positive results.
Laparoscopy	Entails the insertion of a thin light tube into the lower abdomen via a small incision and permits the doctor to examine the ovaries and other pelvic organs

The prospect of reducing ovarian cancer mortality rates through earlier diagnosis and treatment is a high priority, but available screening approaches such as those mentioned in Table 1.3 often fail to detect this disease at an early stage and/or can sometimes lead to unnecessary surgery (Coukos *et al.*, 2008).

1.2.6 Treatment

Ovarian cancer very often causes few symptoms until it has metastasized within the peritoneal cavity at which time the chance of cure is significantly reduced (Helm *et al.*, 2009). Even though ovarian cancer therapy has improved, the 5-year survival rates for stages I, II, III and IV are 74, 58, 30 and 19%, respectively (Steele *et al.*, 1994). Depending on the size, location, type and progression of the cancer, the most common types of treatment are surgery, chemotherapy and radiotherapy, or a combination of these treatments in varying degrees (Bupa, 2007 & ecancermedia, 2008).

1.3 Genetic Variation

DNA variations between individuals can be an indication of predisposition to disease or affect the degree of response to treatment (Stepanova *et al.*, 2006). Identification of functional genetic variation associated with increased susceptibility to complex diseases can elucidate genes and underlying biochemical mechanisms linked to disease onset or progression (Malin *et al.*, 2008).

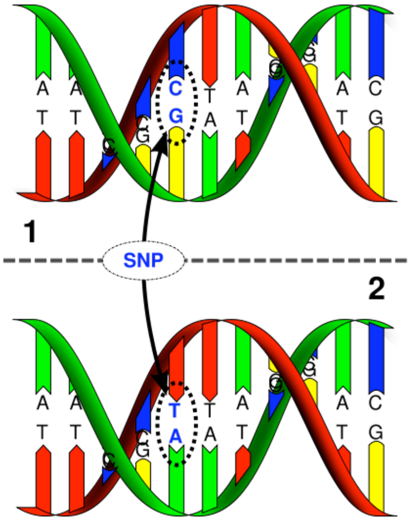


Figure 1.4 Illustration of a single nucleotide polymorphism between two DNA strands (GnpSNP, 2009).

Genetic differences occur in various ways, most with no medical consequences to our health, such as inheriting our mother's hair or father's eyes, whereas others have more significant effects, affecting protein activity or gene expression rendering some more genetically susceptible to diseases than others (Roche, 2008; U.S. National Institutes of Health, 2001; Tebbutt *et al.*, 2007). SNP detection is one of the most powerful tools in the search for disease susceptibility genes and drug response-determining genes (Conde *et al.*, 2006).

In contrast to more mutable markers such as microsatellites, SNPs have a low rate of recurrent mutation, making them stable indicators of human history (Sachidanandam *et al.*, 2001). In a study by Chowdhury *et al.* (2006), SNPs were identified within a region ±1200bp upstream and 1300bp downstream of the transcription start site (TSS) of the peptidase inhibitor 3 gene.

TFBSs were then detected from related and unrelated sites and SNPs genotyped from them (Chowdhury *et al.*, 2006). Results indicated a differential binding of transcription factors, providing evidence of functional promoter variants existing within genes (Chowdhury *et al.*, 2006). TFs acting upon these TFBSs containing genetic polymorphisms (i.e. SNPs) have also been shown to have diverse effects within the cell, with various indicators as by-products. An example of this is the USF-1 TF, which is associated with an increased adipocyte lipolysis (Hoffstedt *et al.*, 2005). Hoffstedt *et al.* (2005) describe the implications of an increased mRNA level of protein kinase A (a post-receptor enzyme) that highlighted the viability for monitoring 'SNP by-products' in the cell.

Another example of the effect of mutations within a TF binding domain, described by Sakazume *et al.* (2007) explains the effect of mutations affecting regulatory elements on the *PITX3* gene (responsible for normal eye development in vertebrates). The study focused on a subset of human patients wherein *PITX3* mutations demonstrated corneal anomalies with cataract/lens defects present in all cases (Sakazume *et al.*, 2007).

1.4 Regulatory Genomics

Biological diversity is governed by the biochemical processes that constitute gene regulation and is an important mechanism for the regulation of gene expression (Wasserman *et al.*, 2004 & Alkema *et al.*, 2004). Transcription constitutes this first step in the expression of genes and is central to the regulatory mechanisms within any biological cell (Wasserman *et al.*, 2004). Transcription regulation is shaped by the interactions between transcription factors (TFs) that bind to cis-regulatory elements in DNA and additional trans-acting proteins that aid/control the rate of transcription within each individual gene (Wasserman *et al.*, 2004). Transcriptional gene regulation is dependent on the "sequence-specific binding of TFs to regulatory regions of genes, thereby repressing or activating transcription", (Harbison *et al.*, 2004 & Alkema *et al.*, 2004).

A transcription factor; sometimes referred to as a "sequence-specific DNA binding factor", is a protein that binds to specific parts of DNA through its recognition of DNA binding domains and enables/disables the system that controls the transcription of DNA to RNA (Yang, 1998). The DNA binding domain (DBD) as described by Aranda *et al.* (2001) is composed of two zinc finger proteins made up of 60-70 amino acids, the first of which includes a region called the P box that is able to recognize the core DNA motifs (Aranda *et al.*, 2001). The second zinc finger protein has a D box region that is responsible for dimerisation and allows rotation of the DNA binding domain (Aranda *et al.*, 2001 & Robinson-Rechavi *et al.*, 2003).

Transcription factors play an important role in genetic regulation via the transcription process and have become a great research focus area as they make it possible to alter/impair physiological processes in a given disease (Genfit, 2009 & Janga, 2007). Transcription is sometimes performed solely by transcription factors or through the use of other proteins in complex by increasing or decreasing the presence of RNA polymerase (Berg *et al.*, 2004).

Most genes in the genome are controlled by a combination of "trans-acting factors", i.e. many TFs that bind cooperatively to their associated DNA sequences and subsequently recruit transcriptional cofactors (Hobert, 2008).

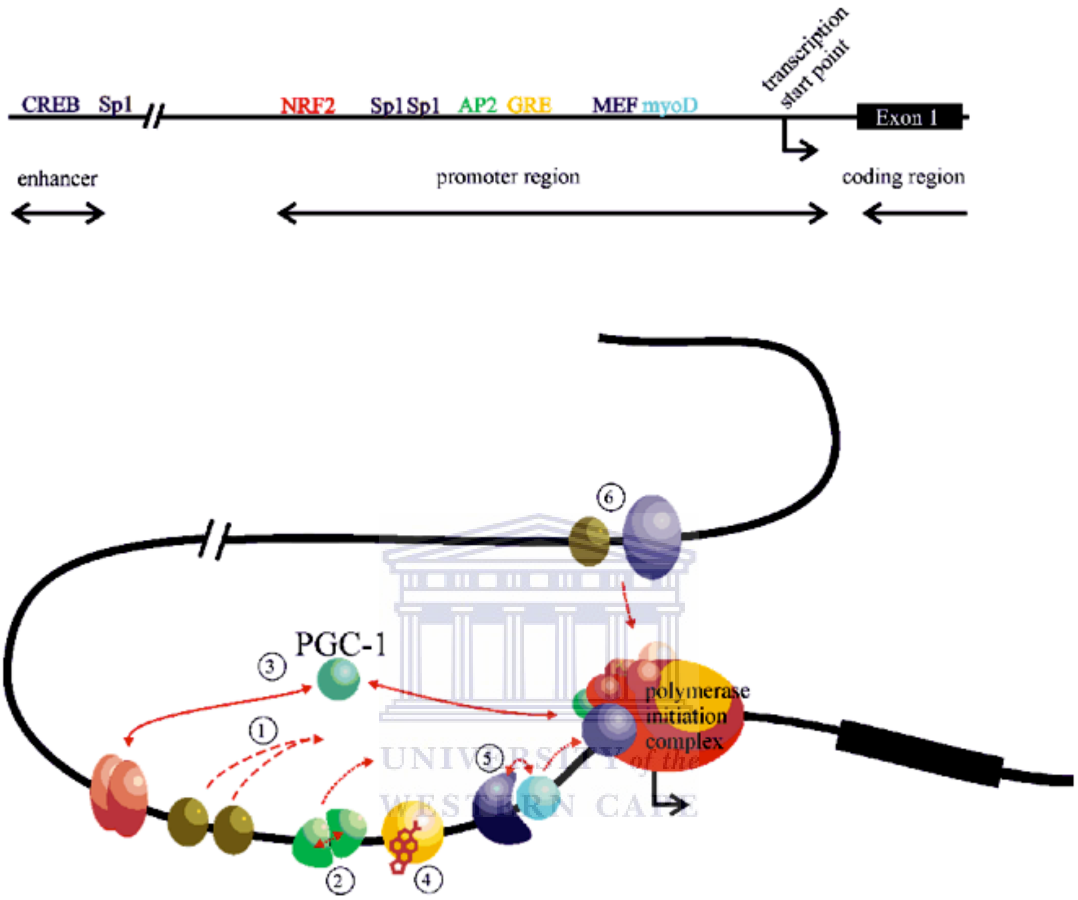


Figure 1.5 Cellular regulatory factors responsible for the transcription of RNA from DNA. In cells, transcription factors (TFs) are responsible for the control of tissue functionality and gene expression, depending on if they are activated, genes will be switched on (up-regulated) and others switched off (down-regulated) (Genfit, 2009 & Goffart *et al.*, 2003). TFs (1) bind to specific DNA consensus elements in the promoter region of DNA sequences and activate transcription by stabilizing the polymerase initiation complex (Goffart *et al.*, 2003). Other regulatory factors, such as protein complexes (2), co-activators (3), ligands (4) regulatory DNA sequences (5) and enhancers (6) may additionally work co-operatively in the mediation or enhancement of transcription regulation (Goffart *et al.*, 2003).

Numerous computational methods have been derived for the discovery of cis-regulatory elements including the use of "correlation with expression" techniques as described by Bussemaker *et al.* (2001). These researchers use a method of clustering genes based on their expression profiles, thus uncovering groups of genes that co-vary based on shared cis-regulatory regions. Bussemaker *et al.* (2001) were able to uncover patterns of combinatorial transcriptional control through analysis of mRNA levels and found that certain motifs were correlated to expression of transcription factor binding sites (TFBSs) in the 600bp upstream region of the transcription start site.

A few years ago, methods for the identification of TFBSs incorporated the use of weight matrices containing scores for all possible bases at each position in a binding site (Benos *et al.*, 2002). True binding sites score higher than sites that do not bind transcription factors. Benos *et al.* (2002) proposed a model where they measured the binding affinities of proteins to DNA. Despite the fact that their model did not fit the data perfectly, in most cases it provided a very good approximation for the discovery and prediction of genomic DNA binding sites.

Hannenhalli *et al.* (2002) introduced the concept of modules, present in a few hundred base pairs proximal to the gene. These various modules work together in the regulating of gene expression to constitute a promoter module (Hannenhalli *et al.*, 2002). The authors highlight the need for analyzing the transcription factors that bind to these regions and hypothesized that the transcription factors responsible for activity within the promoter sites are likely to be part of a transcriptional module on the human genome sequence (Hannenhalli *et al.*, 2002). McNutt *et al.* (2005), who have shown that TFBS composition is non-randomly distributed between gene promoters in a manner that defines gene class function, later confirmed the validity of this hypothesis.

Harbison *et al.* (2004) demonstrated eukaryotic transcription via their study of the yeast cell. Regulatory binding sites within the yeast cell were clustered between 100 – 500bp upstream of the coding region and not randomly distributed (Harbison *et al.*, 2004). They also observed four types of promoters, which were classified according to the architecture of the binding sites in terms of their organization (Harbison *et al.*, 2004). Pavesi *et al.* (2004) suggested that the problem associated with the data models was that many are based on the analysis of yeast genes, who inherently have a short regulatory region (<1000bp) whereas human genes have longer and more complex regulatory modules that include enhancers and silencers (Pavesi *et al.*, 2004).

Alternatively, in an investigation of prokaryotic gene expression, Elf *et al.* (2007) observed the kinetics of binding and dissociation of the repressor in response to metabolic signals. Moreover, they managed to characterize the nonspecific binding to DNA, observing the facilitated diffusion of the repressor along the DNA strand in their search for an 'operator' (Elf *et al.*, 2007). Although this study was performed in prokaryotes, the principles or dynamics of single molecule detection are essential in the investigation of how a transcription factor molecule identifies and binds to specific binding sites along the DNA strand. These results corroborate with those of Fessele *et al.* (2002) who showed that organizational features of sequence promoter regions contain information about the functional context of gene expression.

Berg *et al.* (2004) have also contributed to the area of TFBS identification by showing that the selection for transcription factor binding generally leads to specific correlations between nucleotide frequencies at different positions of a binding site. Different sites for the same transcription factor can differ by about 20-30% of the bases relevant for binding, making them difficult to identify (Berg *et al.*, 2004).

Berg *et al.* (2004) go on to describe binding co-operativity, wherein simultaneous binding at two nearby sites is energetically favored, as it requires lesser energy and this action may be related to various functions. The promoter region is the regulatory region of all DNA sequences, located upstream of a gene, it provides a control point for regulated gene transcription and is typically a few thousand base pairs long containing many different transcription factor binding sites (Berg *et al.*, 2004).

Other studies have demonstrated the need for TFBS profile libraries to be extensively used to identify regulatory elements in DNA sequences (Kielbasa *et al.*, 2005). Findings by Kielbasa *et al.* (2005) reveal two measuring yardsticks that compliment each other; i.e.: X^2 distances between *PFMs and correlation coefficients between position weight matrix (PWM) scores (Su *et al.*, 2006). This is in contrast to the representations of King *et al.* (2003b), who attempted to identify TFBSs using a collection of aligned known binding sites. King *et al.* (2003b) discovered that TFBSs are strongly conserved and tend to occur in clusters. Elnitski *et al.* (2006) corroborate this idea by reiterating the usefulness of representing information within regulatory sites in the form of position weight matrices (PWMs) or position-specific scoring matrices (PSSMs) which incorporate pattern variability by recording nucleotide frequencies at each site or by assigning penalties to nucleotides that should not be within a factor binding site.

Identifying TFBSs *in vitro* is in itself a problematic task, as these sites are miniscule in size and methods that scan sequences for matches to a consensus-binding site produce high false positive rates due to the low specificity (Alkema *et al.*, 2004). Hestand *et al.* (2008) have been one of the many groups to create a computational method able to reduce the identification of false positives in the identification/prediction of TFBSs.

Algorithmic approaches that have been developed for *de novo* pattern detection that search for recurring or overrepresented patterns in DNA include Hidden Markov Models, Gibbs sampling, greedy alignment algorithms (e.g. CONSENSUS), expectation-maximization (e.g. MEME), probabilistic mixture modeling (e.g. NestedMica) and exhaustive enumeration methods (Elnitski *et al.*, 2006).

In recent years, various computational models have been designed to understand gene regulatory networks in response to various stimuli. Hermsen *et al.* (2006) demonstrated one such model, wherein a computer model of transcriptional regulation that was allowed to evolve by mutation. The resulting cis-regulatory regions were thereafter observed to have tandem and often overlapping binding sites to which TFs could bind cooperatively and competitively, enabling the efficient integration of signals (Hermsen *et al.*, 2006). The emergence of phylogenetic footprinting to identify binding sites has also been another novel approach to the study of orthologous genes, for the analysis of common regulatory mechanisms even though regulatory sequences have diverged to render alignment non-existent as regulatory CRM (cis-regulatory module) models are another offset of functional binding site predictive algorithms, which are showing increased improvement in prediction as they become more optimized (Wasserman *et al.*, 2004).

Techniques to identify functional transcription factor binding sites in mammals, both experimentally and computationally (i.e. *in silico*, *in vitro* and *in vivo*), have also been described by Elnitski *et al.* (2006). Experimental techniques are methods that identify regulatory elements by indirectly measuring transcription factor/DNA interactions whereas computational analysis require data sets and are based on either pattern matching or pattern detection that make use of prior knowledge of all characterized DNA binding sites for a given transcription factor (Elnitski *et al.*, 2006).

In bacterial genetics, transcription factors (TFs) have been hypothesized as a major contributor to an organism's response to various external stimuli and a large amount of ongoing work has been focused on predicting the set of transcription factors responsible for gene regulation (Yang *et al.*, 2007). Most current methods attempt to identify possible binding sites from a genomic sequence but predicting transcription factors from these sequences often results in the inclusion of numerous false positives (Yang *et al.*, 2007). Little is known about their functional roles, expression dynamics and evolutionary scenarios (Janga, 2007). Relationships between homologous genes and structures imply correlations of evolutionary changes at different levels of biological organization and data from a variety of organisms have provided significant insight into the evolutionary relationship between genotype and phenotype (Wray, 2007).

Gene expression is extremely complex, but each discovery in the myriads of molecular interactions provides a building block for deciphering the regulatory mechanisms of each cell. According to Abnizova *et al.* (2007), not very much is known about the regulation of transcription in eukaryotes. More specifically, very little is known about the TFBSs and the interacting protein patterns. The authors presented a theory claiming that numerous transcription factors work together in a combinatorial manner to enable cells to respond to various signals/stimuli consisting of either a development or environmental nature. Abnizova *et al.* (2007) considered gene regulatory networks as being the key to understanding transcriptional regulatory mechanisms in eukaryotes. They motivate for the use of various TFBS motif-search algorithms to understand the enormous amounts of variant information encoded in genomic data. Furthermore they emphasize the use of algorithms to search for combinations of TFBS that are enriched in sets of co-regulated genes (Abnizova *et al.*, 2007). These tools will improve TFBS predictions and improve our understanding of gene regulatory network (GRN) construction (Stormo, 2000).

More recently, an important census study by Vaquerizas *et al.* (2009) has provided clues as to how transcription factors (TFs) may operate. These researchers express the lack of a reliable data set of TFs in the human genome and the problems associated with false predictions created through *in silico* studies. They were however, able to indicate where TFs are present with an analysis of chromosomal clusters of genes in relation to their evolutionary histories providing insights into how these regulators function.

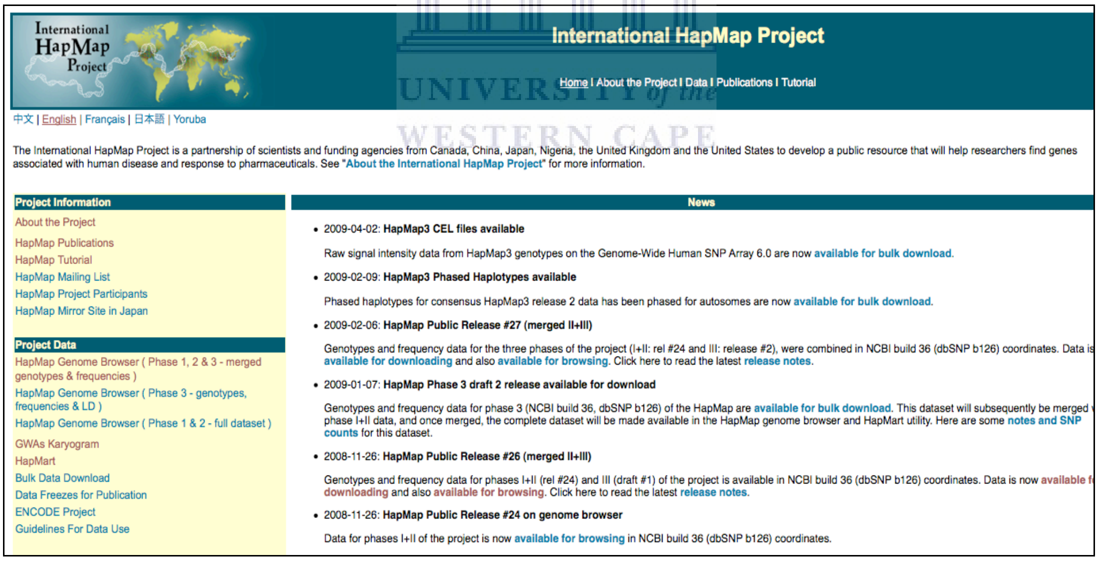
1.5 SNPs & Complex Diseases

Complex diseases are usually those attributed to a combination of environmental factors and genes that result in a phenotypic change (Lowe, 2001). Screening efforts to identify genes for complex diseases, such as ovarian cancer are complicated when considering the risk for developing this disease depends on a particular combination of susceptibility alleles in many linked or unlinked genes (Lowe, 2001). Despite these challenges there have been several reports describing the identification of genes that have been linked to complex diseases in some or other form. With the vast amounts of expression data generated in the past few years, researchers have been finely mapping the the expression levels of many genes, searching for the presence of single nucleotide variations that act as possible triggers in the development of complex diseases (Prokunina *et al.*, 2004). Linkage analysis provides researchers with this crucial information about where in the genome these genetic variations are located and have been “highly successful for many rare single-gene disorders”, (Gibbs *et al.*, 2003 & Prokunina *et al.*, 2004). The International HapMap Project has provided researchers with the foundation to do exactly this, through their provision of a freely available map of common patterns of DNA sequence variation within the human genome (Gibbs *et al.*, 2003).

1.6 The International HapMap Project

Designed to create a public genome-wide database of patterns of common sequence variation, the International HapMap Project emerged as a logical step in the characterization of human genomic variation (Manolio *et al.*, 2008). Aimed toward genetic studies of human health and disease, the HapMap Project has introduced a new paradigm of research in the form of genome-wide association studies (Manolio *et al.*, 2008).

The project collected and analyzed DNA samples from a multitude of population groups, initially of African, Asian and European descent, and identified SNPs within these samples in search of haplotypes with frequencies of 5% or higher within each population group (Nomikos, 2006). The goal of this project ultimately is to identify regions containing disease alleles or alleles that predispose individuals to a type of disease (i.e. from a particular environmental factor or medication) (Nomikos, 2006).



International HapMap Project

Home I About the Project I Data I Publications I Tutorial

中文 | English | Français | 日本語 | Yoruba

The International HapMap Project is a partnership of scientists and funding agencies from Canada, China, Japan, Nigeria, the United Kingdom and the United States to develop a public resource that will help researchers find genes associated with human disease and response to pharmaceuticals. See "About the International HapMap Project" for more information.

Project Information

About the Project
HapMap Publications
HapMap Tutorial
HapMap Mailing List
HapMap Project Participants
HapMap Mirror Site in Japan

Project Data

HapMap Genome Browser (Phase 1, 2 & 3 - merged genotypes & frequencies)
HapMap Genome Browser (Phase 3 - genotypes, frequencies & LD)
HapMap Genome Browser (Phase 1 & 2 - full dataset)
GWAs Karyogram
HapMart
Bulk Data Download
Data Freezes for Publication
ENCODE Project
Guidelines For Data Use

News

- 2009-04-02: **HapMap3 CEL files available**

  Raw signal intensity data from HapMap3 genotypes on the Genome-Wide Human SNP Array 6.0 are now available for bulk download.
- 2009-02-09: **HapMap3 Phased Haplotypes available**

  Phased haplotypes for consensus HapMap3 release 2 data has been phased for autosomes are now available for bulk download.
- 2009-02-06: **HapMap Public Release #27 (merged II+III)**

  Genotypes and frequency data for the three phases of the project (I+II: rel #24 and III: release #2), were combined in NCBI build 36 (dbSNP b126) coordinates. Data is available for downloading and also available for browsing. Click here to read the latest release notes.
- 2009-01-07: **HapMap Phase 3 draft 2 release available for download**

  Genotypes and frequency data for phase 3 (NCBI build 36, dbSNP b126) of the HapMap are available for bulk download. This dataset will subsequently be merged phase I+II data, and once merged, the complete dataset will be made available in the HapMap genome browser and HapMart utility. Here are some notes and SNP counts for this dataset.
- 2008-11-26: **HapMap Public Release #26 (merged II+III)**

  Genotypes and frequency data for phases I+II (rel #24) and III (draft #1) of the project is available in NCBI build 36 (dbSNP b126) coordinates. Data is now available f downloading and also available for browsing. Click here to read the latest release notes.
- 2008-11-26: **HapMap Public Release #24 on genome browser**

  Data for phases I+II of the project is now available for browsing in NCBI build 36 (dbSNP b126) coordinates.

Figure 1.6 The International *HapMap Project main page. Initially launched in 2002 as an international effort to identify and catalog genetic similarities and differences, the HapMap project was initiated to identify how these genetic variations are distributed among people within various population groups (Nomikos, 2006).

* http://hapmap.org/

The HapMap project has enabled the functional investigation and comparison of candidate disease genes across several population groups, providing researchers with new insights into the evolutionary pressures on the human genome (Manolio *et al.*, 2008). Moreover, it has led to the vast improvement of methodologies capable of reliably estimating genotypes of SNPs that have not been 'typed' on existing genotyping platforms, based on information from typed SNPs (Manolio *et al.*, 2008).

1.7 Using SNPs in Drug Development

The application of genetic variation data will enable scientists to discover sequence variants that affect common diseases, or facilitate the development of diagnostic tools that will enhance our ability to choose specific drug targets for therapeutic intervention (Gibbs *et al.,* 2003). Single base variations in the human genome may increase the risk of developing a disease or lower the likelihood of response to a specific medicine (Roche, 2008 & Ramaswamy *et al.*, 2003). The use of SNP markers in the evaluation of DNA samples and analysis of clinical data regarding drug safety and efficacy will make it possible to correlate patient response to medication with specific genetic profiles (GalaxoSmithKline, 2006).

Understanding how SNPs are involved in the susceptibility or resistance to a disease, or in the efficacy/toxicity of drugs is a major goal and the overall aim of pharmacogenomics (Xie *et al.*, 2005). Combining this knowledge with that of the molecular pathways involved in specific diseases and the role that SNPs have to play in these pathways, will provide researchers with an understanding of new potential drug targets, enabling improved intervention and more precise treatment strategies (Frazer *et al.*, 2009).

CHAPTER 2

Research Motivation

THIS year hundreds of thousands of women will be told that they have ovarian cancer. Last year 15520 women died from the disease of the 21650 affected in the United States alone (National Cancer Institute, 2009). As disturbing as these statistics may be, they fail to illustrate the extent of human suffering (Roberts, 1998). They fail to describe the several major surgeries, multiple courses of chemotherapy treatment (with their associated toxic effects), bouts of bowel dysfunction and psychological trauma of battling cancer that these women will undergo before they die of their disease (Roberts, 1998). It is an extremely aggressive and deadly disease, difficult to detect in its early development stages, allowing it to "progress silently until it has metastasized to other organ systems", (Anderson, 2009 & Helm *et al.*, 2009).

The focus of this study therefore concentrated on the genetic susceptibility of ovarian cancer patients through the observation of point mutations that occur within the transcription regulatory regions of genes that have been implicated in the disease. These tiny variations in the human genome known as single nucleotide polymorphisms (SNPs) are investigated here as a plentiful source of potential diagnostic markers to improve cancer diagnosis and treatment planning (Chakravarti, 2001 & Thomas *et al.*, 2004). Although typically SNPs have been used as markers to search for what may be considered the real determinant of a disease, the use of functional SNPs may be an important factor to be considered in the future of association studies and predictive medicine.

2.1 Predictive Medicine

Deciphering the regulatory control mechanisms that govern gene expression will enable us to understand the processes underlying gene regulation and can be of crucial importance to the unraveling of biological processes within cells of the thousands of existing and anticipated patients affected by ovarian cancer (Wasserman *et al.*, 2004 & Alkema *et al.*, 2004). Understanding the interplay between transcription factors and regulatory motifs in the upstream regions of genes will transform biological research and provide a means to interpret and model the responses of cells to diverse stimuli (Wasserman *et al.*, 2004).

2.2 Research Aims

The aims for this research were as follows:

(1) To assess the transcriptional regulation of ovarian cancer genes based on the inferred losses of putative transcription factor binding sites (TFBSs) caused by the occurrences of single nucleotide polymorphisms (SNPs).

(2) To contribute toward the identification of potentially bio-medically important genes or SNPs for pre-diagnosis and therapy of ovarian cancer by evaluating the possible regulatory effects of SNPs among several population groups defined by the International HapMap consortium.

2.3 Research Objectives

The objectives undertaken to achieve the above aims were as follows:

(1) Identification and mining of candidate gene dataset (i.e. genes implicated in ovarian cancer).

(2) Prediction of SNP occurrences within all candidate genes through the application of several online resources and custom programs.

(3) Prediction of TFBSs on promoter regions of all genes via a positional weight matrix (PWM) approach to TFBS motif identification.

(4) Determination of SNPs that overlap with TFBSs within candidate genes.

(5) Identification of SNPs (i.e. SNPs that overlap with TFBSs) allele patterns among population groups defined by the International HapMap consortium.

CHAPTER 3

Single Base Differences

MORE than a decade ago, the debate of whose genome was to be sequenced began with the start of the Human Genome Project, as the study of inherited genetic variation between individuals was envisaged (Chakravarti, 2001). Ultimately geneticists resolved to "not only a single history-making human genome sequence, composed of little bits from many humans, but also more than 1.4 million sites of variation mapped along that reference sequence", (Chakravarti, 2001 & Sachidanandam *et al.*, 2001). These variations (or polymorphisms) are the most common types of variation between humans and may account for as much as 90% of human genetic variation (Chakravarti, 2001 & Tian *et al.*, 2007). In 2001 the International SNP Map Working Group reported 93% of genes containing a SNP, when for the first time nearly every human gene and genomic region was marked by a sequence variation (Chakravarti, 2001). Today, more than 11 million SNPs have been described in databases such as dbSNP, and among them thousands that potentially impact on disease directly (Reumers *et al.*, 2007).

This chapter focuses on the identification of SNPs that potentially alter transcriptional activity and/or transcription factor binding among a collection of genes that have been differentially expressed in ovarian cancer. Through the application of several online resources and custom programs, the focal point of this chapter was on the elucidation of SNPs with potential phenotypic effects at a transcriptional level.

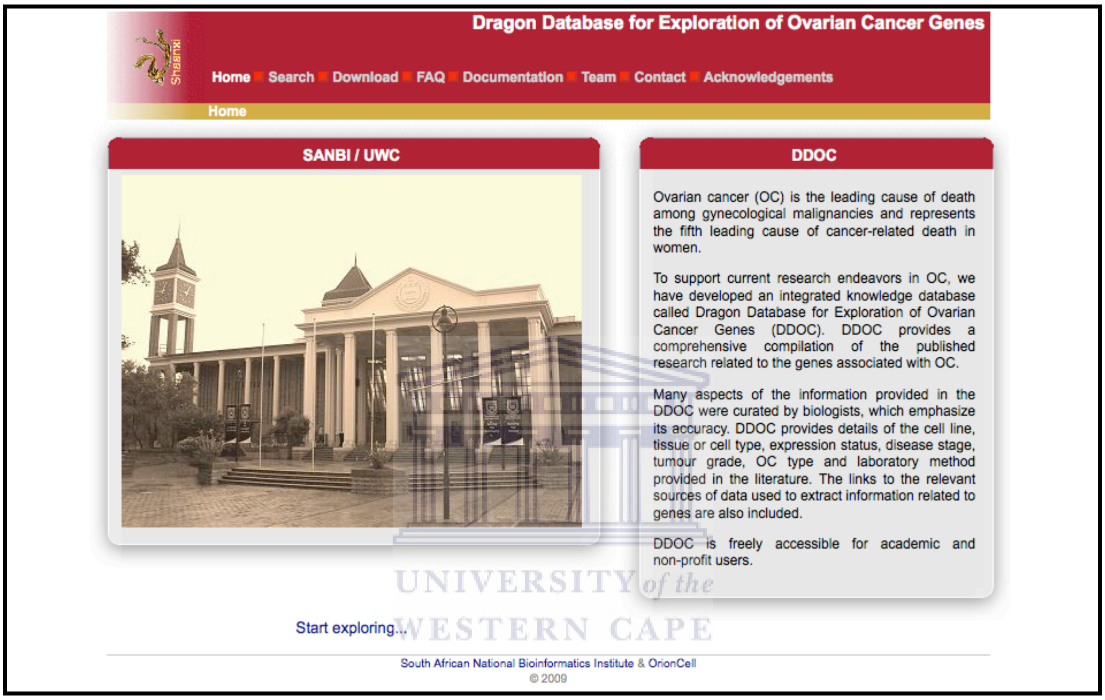
3.1 Introduction

Genetic variation is a factor that has been associated, by several studies; with the susceptibility of diseases through the modification of amino acid sequences in DNA encoded proteins (GuhaThakurta *et al.*, 2006). The phenotypes expressed by these variations include genetic susceptibility to diseases and resistance to therapeutic agents (Cardon *et al.*, 2001). Single nucleotide polymorphisms (SNPs) provide a means for the testing of associations between genetic variations and disease, and have been hypothesized to having causal roles in the susceptibility to genetic disorders as a result of interferences within the regulatory regions of genes (Flintoft, 2004).

With the onset of new expression techniques including data derived from microarray experiments, many *in silico* methods have been proposed and implemented for the identification of underlying biological information (Pavesi *et al.*, 2004). Online resources containing biological knowledge have become vital to the work of many scientists around the world, with individual groups having given rise to a diversity of biological databases, tools and applications in the field.

3.1.1 Dragon Database for Exploration of Ovarian Cancer Genes (DDOC)

The DDOC is a database dedicated to genes that have been implicated in ovarian cancer, developed to support exploration of functional characterization and analysis of biological processes related to the disease (Kaur *et al.*, 2008).


Dragon Database for Exploration of Ovarian Cancer Genes
Home Search Download FAQ Documentation Team Contact Acknowledgements
Home
SANBI / UWC
DDOC

Ovarian cancer (OC) is the leading cause of death among gynecological malignancies and represents the fifth leading cause of cancer-related death in women.

To support current research endeavors in OC, we have developed an integrated knowledge database called Dragon Database for Exploration of Ovarian Cancer Genes (DDOC). DDOC provides a comprehensive compilation of the published research related to the genes associated with OC.

Many aspects of the information provided in the DDOC were curated by biologists, which emphasize its accuracy. DDOC provides details of the cell line, tissue or cell type, expression status, disease stage, tumour grade, OC type and laboratory method provided in the literature. The links to the relevant sources of data used to extract information related to genes are also included.

DDOC is freely accessible for academic and non-profit users.

South African National Bioinformatics Institute & OrionCell
© 2009

Figure 3.1 The *DDOC main page (South African National Bioinformatics Institute & OrionCell, 2009). Integrating several other tools, DDOC provides users with detailed information relating to homologs, regulatory mechanisms, pathways and text-mining results associated to the genes contained in this database (Kaur *et al.*, 2008).

DDOC contains a total of 379 human genes that have been verified experimentally and through literature mining, resulting in the exclusion of 521 of the initially derived 900 genes that were unable to meet these criteria (Kaur *et al.*, 2008).

* http://apps.sanbi.ac.za/ddoc/

Initial data mining of genes were collated from the following repositories:

Repository	URL
Cancer Gene Census	http://www.sanger.ac.uk/genetics/CGP/Census/
GeneCards	http://www.genecards.org/index.shtml
SymAtlas	http://symatlas.gnf.org/SymAtlas/
OMIM	http://www.ncbi.nlm.nih.gov/
Ovarian Kaleidoscope Database	http://ovary.stanford.edu/
Entrez Gene	http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene
GenAtlas	http://www.genatlas.org/

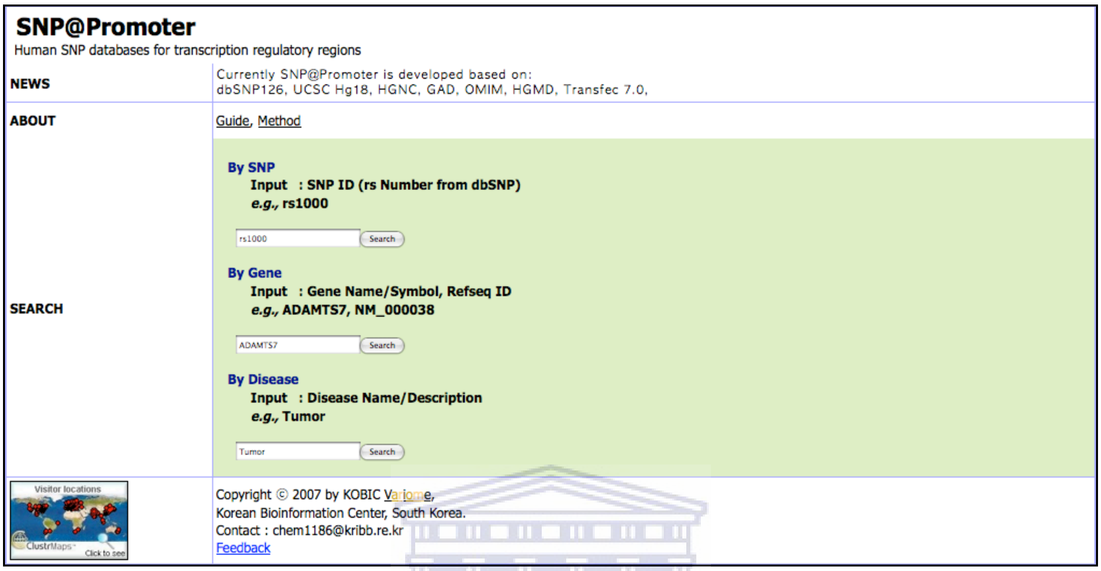
HUGO gene symbols, full gene names and Entrez IDs are available for all genes within the DDOC database, with HGNC IDs provided for 374 genes and Ensembl IDs for 370 of the total genes (Kaur *et al.*, 2008). The database additionally provides the user with *Gene Ontology (GO) annotations for 367 genes with 353 genes indexed in ©eVOC (Kaur *et al.*, 2008).

Among other features, DDOC offers information that is not freely available to all researchers via their delivery of TFBS predictions mapped to promoter regions contained in the 1000bp upstream and 200bp downstream of the transcription start sites (TSSs) on all genes (Kaur *et al.*, 2008). TFBSs were mapped to both the forward and reverse strands of these promoter regions using the MATCH^{TM} program from the Transfac® Professional (Version 11.4) database (Kaur *et al.*, 2008).

* Gene Ontology (http://www.ebi.ac.uk/GO/)
© eVOC (http://www.sanbi.ac.za/evoc/)

3.1.2 SNP@Promoter

The SNP@Promoter tool is an integrated computational system for the identification of SNPs in non-coding regulated regions of genes (SNP@Promoter, 2007).



SNP@Promoter
Human SNP databases for transcription regulatory regions

NEWS — Currently SNP@Promoter is developed based on:
dbSNP126, UCSC Hg18, HGNC, GAD, OMIM, HGMD, Transfec 7.0,

ABOUT — Guide, Method

SEARCH

By SNP
Input : SNP ID (rs Number from dbSNP)
***e.g.,* rs1000**

rs1000 [Search]

By Gene
Input : Gene Name/Symbol, Refseq ID
***e.g.,* ADAMTS7, NM_000038**

ADAMTS7 [Search]

By Disease
Input : Disease Name/Description
***e.g.,* Tumor**

Tumor [Search]

Copyright © 2007 by KOBIC Variome,
Korean Bioinformation Center, South Korea.
Contact : chem1186@kribb.re.kr
Feedback

Figure 3.2 The *SNP@Promoter main page. The main page may be queried via any of three term entries i.e. by entering a SNP identifier, gene name/symbol/RefSeq ID or by querying a disease term (Korean Bioinformation Center, 2007 & Kim *et al.*, 2008).

The underlying computational system of SNP@Promoter, defining the transcription regulatory region of a gene as "the sequence of *5kb upstream to 500bp downstream bases of a transcription start site"*, was developed specifically to determine the following key aspects of analysis (Korean Bioinformation Center, 2007 & Kim *et al.*, 2008):

(1) Prediction of TFBSs in putative promoter regions
(2) Identification of SNPs in putative promoter regions
(3) Select SNPs within predicted TFBSs
(4) Examine evolutionary conservation of predicted TFBSs
(5) Integration of a variety of gene annotation information

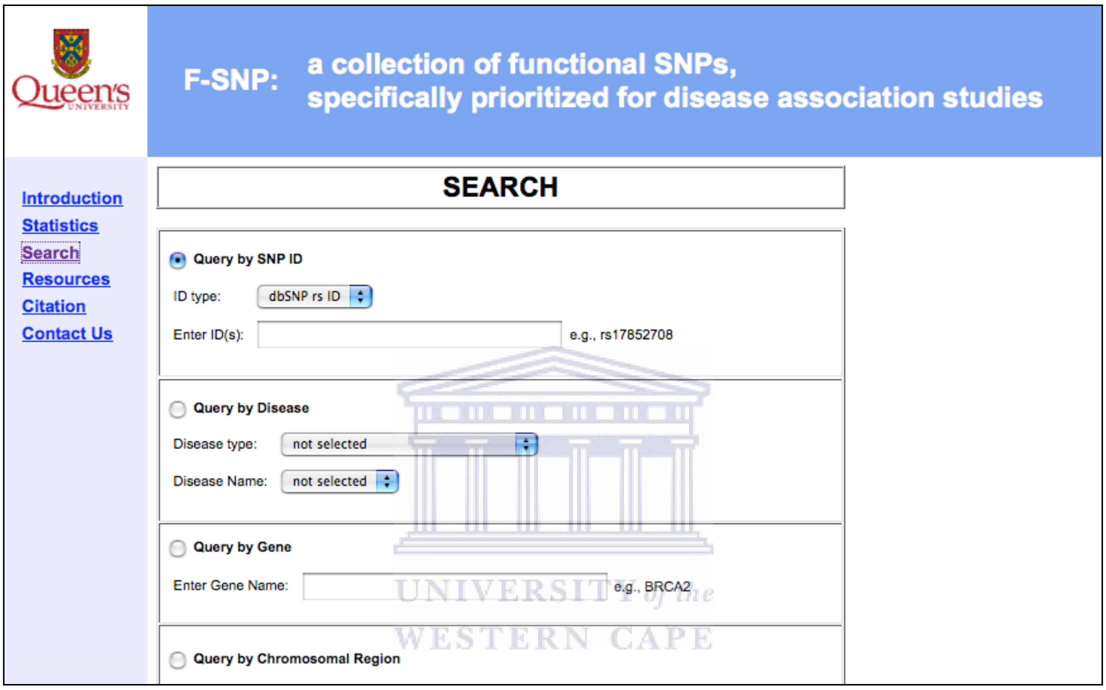
* http://variome.kobic.re.kr/SNPatPromoter/

SNPs present on putative promoter regions were derived from dbSNP (build 126) while TFBSs were predicted using the MATCH^{TM} (Matrix Search for Transcription Factor Binding Site) program from Version 8.4 of the Transfac$^{®}$ database (Kim *et al.*, 2008). As a result SNP@Promoter includes 1497317 TFBSs, from 28644 human genes mapped to 488452 SNPs, 47832 of which are located within the putative TFBSs (Kim *et al.*, 2008).

All annotation information that is mapped to the genes contained in this tool was obtained from the NCBI Gene database and can be viewed graphically for all queried genes and SNPs (Kim *et al.*, 2008).

3.1.3 Functional Single Nucleotide Polymorphism (F-SNP)

The F-SNP (Release 1.0) database integrates information obtained from 16 independent bioinformatics tools and databases (*Table 3.1*) relating to the functional effects of SNPs, based on their effects at a splicing, transcriptional, translational or post-translational level (Lee *et al.*, 2007).



Queen's UNIVERSITY

F-SNP: a collection of functional SNPs, specifically prioritized for disease association studies

SEARCH

● Query by SNP ID

ID type: dbSNP rs ID

Enter ID(s): e.g., rs17852708

○ Query by Disease

Disease type: not selected

Disease Name: not selected

○ Query by Gene

Enter Gene Name: e.g., BRCA2

○ Query by Chromosomal Region

Figure 3.3 The *F-SNP main page (Queen's University, 2007). Users may search the database by entering a SNP identifier, gene symbol, disease name/type or by selecting a chromosomal region.

F-SNP combines a total of 38550 human genes along with their related information (i.e. gene symbol, alias names, chromosomal location, etc.), obtained from the NCBI Entrez Gene database as well as SNP annotation data sourced from the dbSNP (build 126) and Ensembl (release 42) databases (Lee *et al.*, 2007). Consequently, a total of 4043147 SNPs located in the *5kb upstream and 5kb downstream* regions were mapped to 23630 of these human genes (Lee *et al.*, 2007).

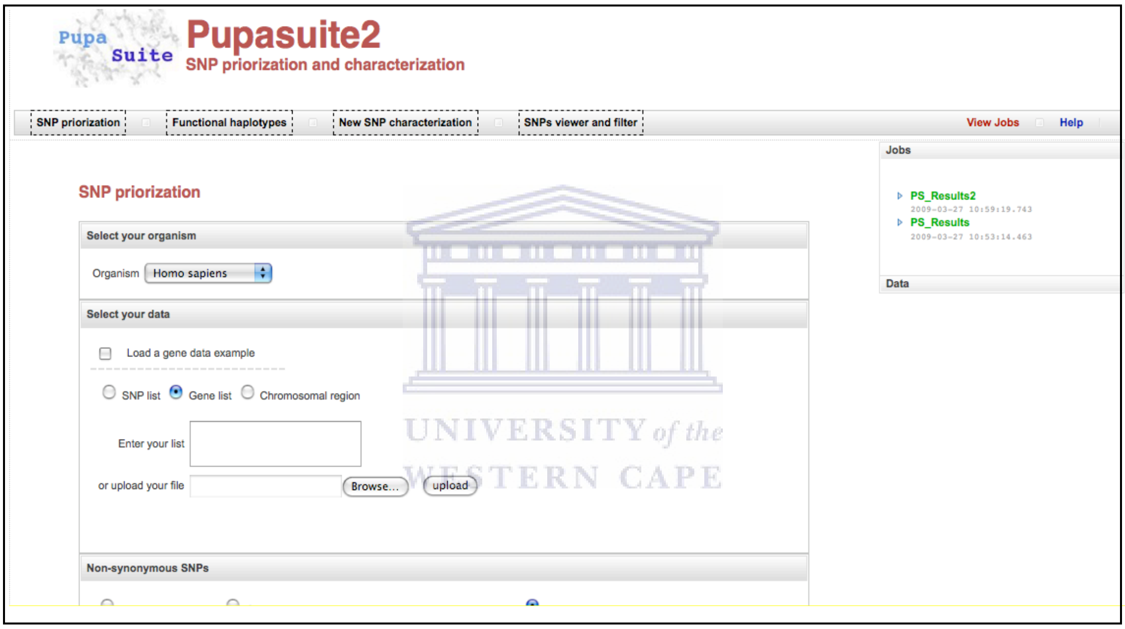
* http://compbio.cs.queensu.ca/F-SNP/

Table 3.1 F-SNP data incorporation from several sources (Lee *et al.*, 2007 & Karchin, 2008). To assess the functional effects of SNPs for each possible category of SNP type, F-SNP combines the functionalities of the following collection of tools:

Tool	Usage	URL
PolyPhen	Identification of non-synonymous deleterious SNPs	http://genetics.bwh.harvard.edu/pph/data/index.html
SIFT		http://blocks.fhcrc.org/sift/SIFT.html
SNPeffect		http://snpeffect.vib.be/index.php
SNPs3D		http://www.snps3d.org/modules.php?name=SNPtargets
LS-SNP		http://alto.compbio.ucsf.edu/LS-SNP/Queries.html
ESEfinder	Identification of SNPs in exonic splice regions	http://rulai.cshl.edu/cgi-bin/tools/ESE3/esefinder.cgi
RescueESE		http://genes.mit.edu/burgelab/rescue-ese/
ESRSearch		http://ast.bioinfo.tau.ac.il/
PESX		http://cubweb.biology.columbia.edu/pesx/
Ensembl	Identification of nonsense SNPs and SNPs in intronic splice sites	http://www.ensembl.org/index.html
TFSearch	Identification of transcriptional regulatory SNPs in promoter regions	http://www.cbrc.jp/research/db/TFSEARCH.html
Consite		http://asp.ii.uib.no:8090/cgi-bin/CONSITE/consite/
Ensembl	Identification of SNPs in other transcriptional regulatory regions (e.g. microRNA, cpgIslands)	http://www.ensembl.org/index.html
GoldenPath		http://genome.ucsc.edu/
KinasePhos	Examination of post-translational modification sites	http://kinasephos.mbc.nctu.edu.tw/
OGPET		http://ogpet.utep.edu/
Sulfinator		http://www.expasy.ch/tools/sulfinator/

3.1.4 PupaSuite

The PupaSuite (Version 2.0), developed by the Centro Nacional de Investigaciones Oncológicas (CNIO) is an interactive web-based SNP analysis tool that uses a collection of data on SNPs from several sources and combines the functionality of both PupaSNP and PupasView into a more integrated interface (Conde *et al.*, 2006). Moreover, it implements new facilities such as the analysis of user data to derive haplotypes with functional information, as well as predictions by the SNPeffect database (Conde *et al.*, 2006).



SNP priorization

Select your organism

Organism Homo sapiens

Select your data

□ Load a gene data example

○ SNP list ● Gene list ○ Chromosomal region

Enter your list

or upload your file Browse... upload

Non-synonymous SNPs

PS_Results2 2009-03-27 10:59:19.743

PS_Results 2009-03-27 10:53:14.463

Figure 3.4 The *PupaSuite main page (Centro de Investigacion Principe Felipe, 2008). Users may input their queries as *lists of SNPs/genes* or via *chromosomal regions*, which correspond to two common types of analyses, (1) genes that may be related to a disease because they are functionally related or (2) genes present in a chromosomal region linked to a disease (Conde *et al.*, 2006).

The PupaSuite tool selects SNPs for genotyping experiments that are often multigenic and reflective of disruptions in proteins that participate in protein complexes or pathways (Conde *et al.*, 2006). The program also includes predictions for SNPs present in TFBSs, splice sites, silencers and miRNAs including their targets (Reumers *et al.*, 2007).

* http://pupasuite.bioinfo.cipf.es or http://www.pupasnp.org

Table 3.2 Optimal SNP candidates as defined by the PupaSuite tool. In order for a SNP to be considered as an optimal candidate for genotyping purposes, the following three main features are taken into account (Conde *et al.*, 2006):

Criterion	Description
Minor allele frequency (MAF)	Sourced from the Ensembl database (maps dbSNP data onto corresponding chromosomal coordinates)
Linkage disequilibrium (LD)	Calculated as r^2 and D^0 with the Haploview program
Putative functional effect	Estimated in both exons and introns

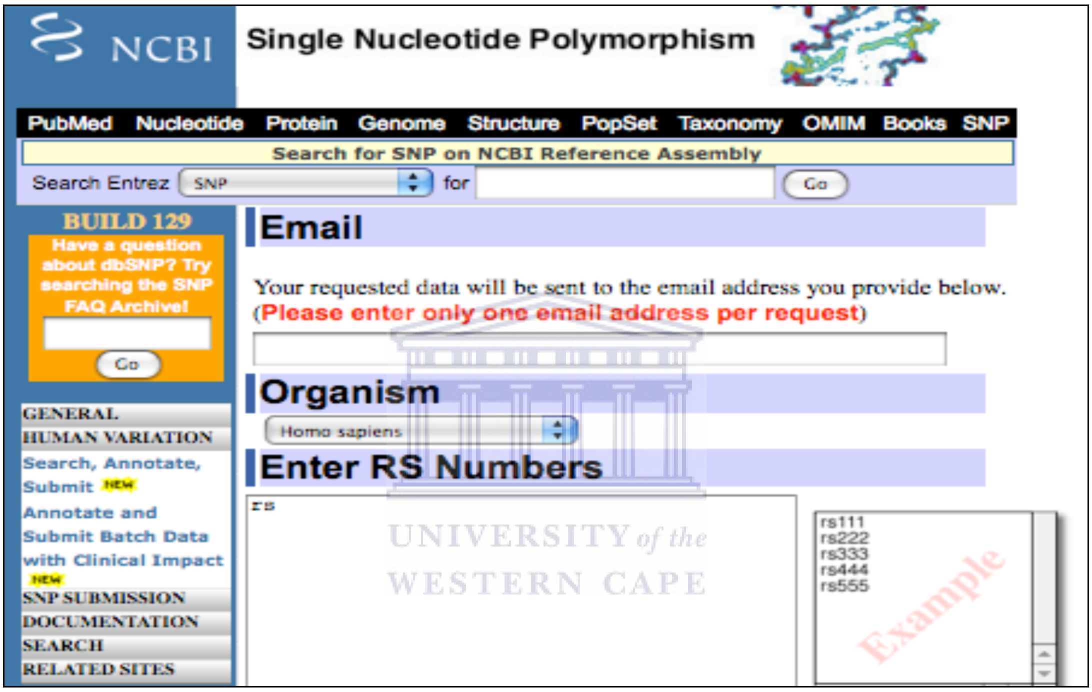
TFBS identification is done using position weight matrices (PWMs) stored in the ◉JASPAR and *Transfac® databases. JASPAR is an open-access database of annotated, high quality, matrix-based TFBS profiles for multi-cellular eukaryotes (Reumers *et al.*, 2007 & Sandelin *et al.*, 2004). Transfac® is another important database that uses PWMs to identify TFBSs (Fu *et al.*, 2005). Although the detailed algorithms to construct the PWMs in Transfac® have not been published, it has been applied to several notable studies (Fu *et al.*, 2005; Bozek *et al.*, 2007; Bozek *et al.*, 2008; Frericks *et al.*, 2008; Cai *et al.*, 2009; Ridder *et al.*, 2009 & Yang *et al.*, 2009).

PupaSuite uses matrices corresponding to vertebrates to search for TFBSs in the *5kb upstream region* of all human genes (Reumers *et al.*, 2007). The program additionally incorporates the use of the MatScan program (http://genome.imim.es) to search for binding sites in genomic sequences, however since it does not allow a cutoff to minimize false positives, the PupaSuite also uses the Meta program (http://genome.imim.es) to filter results by searching for coincidences of TFBSs in orthologous genes in the mouse genome (Reumers *et al.*, 2007).

◉ http://jaspar.cgb.ki.se/
* http://www.biobase-international.com/pages/index.php?id=transfac

3.1.5 dbSNP

The dbSNP database, created by the National Centre for Biotechnology Information (NCBI), is a public-domain archive of SNPs originally established to address the large-scale sampling designs required by association studies and to assist in gene mapping and evolutionary biology research (Sherry *et al.*, 2000 & Edvardsen *et al.*, 2006).



Your requested data will be sent to the email address you provide below. (Please enter only one email address per request)

Figure 3.5 The *dbSNP main page (National Center for Biotechnology Information, 2008). “Users may query dbSNP directly or start a search in any part of the NCBI discovery space to construct a set of dbSNP records that satisfy their search conditions”, (Sherry *et al.*, 2000).

In 2006, dbSNP contained over 10 million SNPs, collated from more than 97 registered groups with all records being cross-annotated within other NCBI-internal information resources (e.g. Pubmed, GenBank, LocusLink, etc.) (Sherry *et al.*, 2000 & Edvardsen *et al.*, 2006).

* http://www.ncbi.nlm.nih.gov/SNP

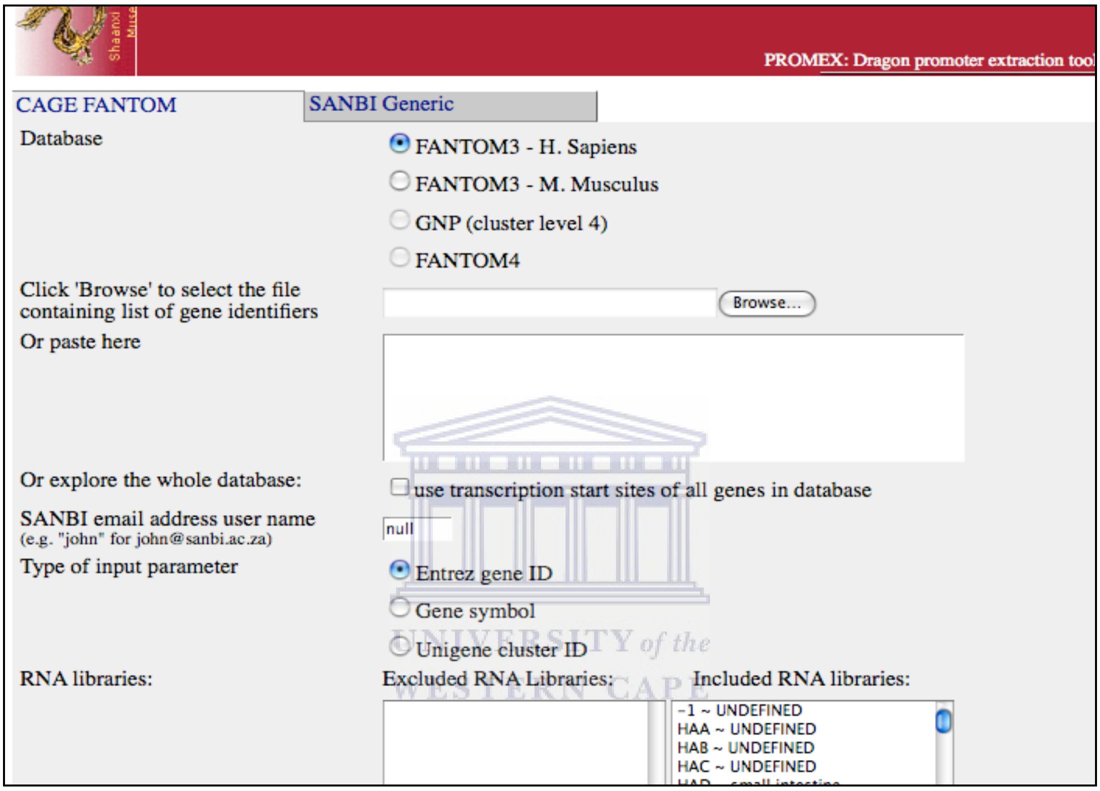
Today dbSNP contains more than 56 million SNPs mapped to 45 organisms (National Center for Biotechnology Information, 2006) designed to facilitate searches based on the following five key types of information (Sherry *et al.*, 2000):

(1) Sequence location
(2) Function
(3) Cross-species homology
(4) SNP quality or validation status
(5) Degree of population variation

It is suggested that SNPs in the dbSNP database that have been reported by at least two independent groups are most likely to be considered true variants and success rates of genotyping projects for a selected number of SNPs are improved if SNPs have been validated in dbSNP (Carlson *et al.*, 2003; Reich 2003 & Edvardsen *et al.*, 2006).

3.1.6 PROMEX: Dragon promoter extraction tool

PROMEX is a promoter retrieval tool designed for the identification of transcription start sites (TSSs) and extraction of user-specified promoter regions upstream and downstream of the identified TSSs.



PROMEX: Dragon promoter extraction tool

CAGE FANTOM | SANBI Generic

Database
● FANTOM3 - H. Sapiens
○ FANTOM3 - M. Musculus
○ GNP (cluster level 4)
○ FANTOM4

Click 'Browse' to select the file containing list of gene identifiers [Browse...]

Or paste here

Or explore the whole database: □ use transcription start sites of all genes in database

SANBI email address user name (e.g. "john" for john@sanbi.ac.za) null

Type of input parameter
● Entrez gene ID
○ Gene symbol
○ Unigene cluster ID

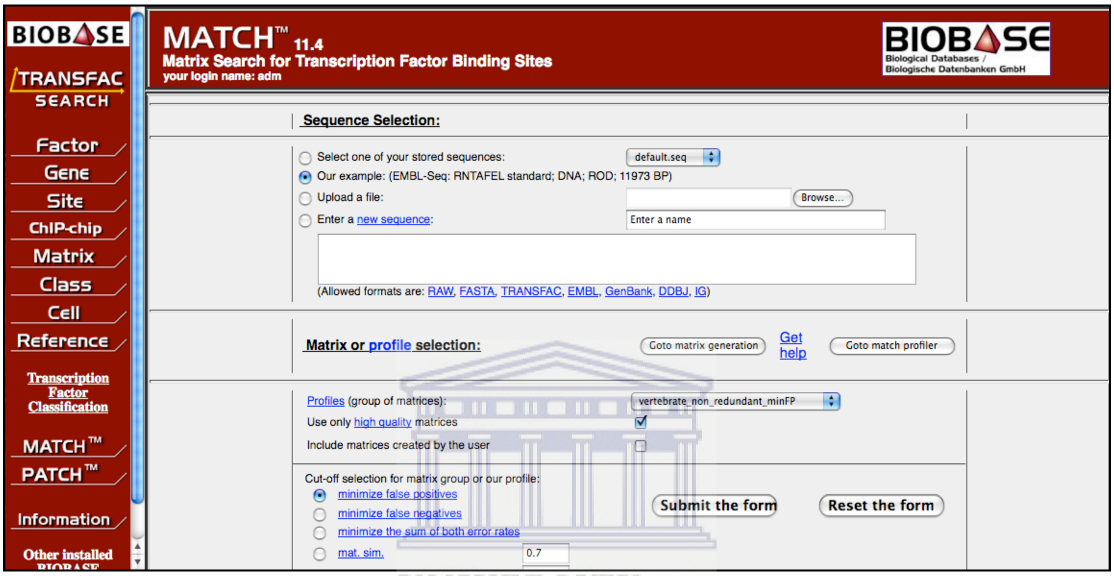
RNA libraries: Excluded RNA Libraries: Included RNA libraries:
-1 ~ UNDEFINED
HAA ~ UNDEFINED
HAB ~ UNDEFINED
HAC ~ UNDEFINED
HAD ~ small intestine

Figure 3.6 The *PROMEX: Dragon promoter extraction tool main page (Schaefer, 2009). Users may query multiple sequence files in a single query by clicking on the “Browse” button and selecting the sequence file or pasting it into the text query box. The inclusion or exclusion of “CAGE” RNA libraries is a customizable feature and genes may be queried in the form of *Entrez gene IDs*, *Gene symbols* or *Unigene cluster IDs*.

* http://tr.sanbi.ac.za/~ulf/promex/

3.1.7 Matrix Search for Transcription Factor Binding Site (MATCH^{TM})

MATCH^{TM} is a matrix-based search tool, designed to identify potential binding sites for transcription factors in DNA sequences (Kel *et al.*, 2003; Matys *et al.*, 2003 & Wasserman *et al.*, 2004).



Sequence Selection:

○ Select one of your stored sequences: default.seq
● Our example: (EMBL-Seq: RNTAFEL standard; DNA; ROD; 11973 BP)
○ Upload a file: Browse...
○ Enter a new sequence: Enter a name
(Allowed formats are: RAW, FASTA, TRANSFAC, EMBL, GenBank, DDBJ, IG)

Matrix or profile selection: Goto matrix generation | Get help | Goto match profiler

Profiles (group of matrices): vertebrate_non_redundant_minFP
Use only high quality matrices ☑
Include matrices created by the user □

Cut-off selection for matrix group or our profile:
● minimize false positives
○ minimize false negatives
○ minimize the sum of both error rates
○ mat. sim. 0.7

Submit the form | Reset the form

Figure 3.7 The *MATCH^{TM} query page. MATCH^{TM} accepts DNA sequences as input then searches for potential TFBSs using a library of PWMs. The program then outputs a list of these predictions in text or graphical format, illustrating their locations within the submitted sequence (Kel *et al.*, 2003).

The search algorithm uses the following two score measures to assess the quality of a match between the query sequence and the matrix (Kel *et al.*, 2003):

(1) Matrix Similarity Score (MSS)
(2) Core Similarity Score (CSS)

Ranging from 0.00 to 1.00, TFBS predictions with a score of 1.00 are classified as exact matches (Kel *et al.*, 2003).

* http://www.gene-regulation.com/pub/programs.html#match

Both MSS and CSS scores are calculated using the same formula (Kel *et al.*, 2003). While MSS is calculated using all matrix positions, CSS calculations are based only on the core positions within a matrix (Kel *et al.*, 2003). Two corresponding cut-off scores (customizable) are defined for every matrix and only matches for which both scores are higher than these cut-offs are reported (Kel *et al.*, 2003).

To improve on the efficiency of the algorithm, a hash table is constructed for every five nucleotides (pentanucleotide) in the query sequence (Kel *et al.*, 2003). The program then calculates and stores all CSS values for every five nucleotides into this hash table, before checking if the CSS value is higher than the initial cut-off value (Kel *et al.*, 2003). If the CSS value is higher than the cut-off score, this pentanucleotide is then searched for in the whole query sequence and is prolonged at both ends to ensure that it fits the matrix length (Kel *et al.*, 2003). The matrix similarity score is then calculated and only values higher than the cut-off value are reported (Kel *et al.*, 2003).

3.1.8 Python

Python is a general-purpose, freely available, object-oriented programming language that can be used for many kinds of software development, (Python, 2008). First released by Guido van Rossum in 1991, Python comes with extensive standard libraries and is a minimalist language both syntactically and semantically. It may be integrated into other languages and tools and is often used as a scripting language (Python, 2008).

Python was used extensively in the parsing and handling of data generated and analyzed in this study, in addition to data/results obtained from the various tools applied in Sections 3.2.2.1, 3.2.2.3 - 3.2.2.5, 3.2.3 - 3.2.5 and 4.2.1.

3.2 Methodology

The initial aspect of this project aimed at the prediction and validation of a panel of SNPs suitable for future disease association studies among women possessing one or more of the traits associated with an increased risk for ovarian cancer (*Table 1.2*).

3.2.1 Selecting Candidate Genes

The initial set of candidate genes were extracted from the Dragon Database for Exploration of Ovarian Cancer Genes (DDOC) (*Section 3. 1.1*). This data set included all entries stored within the database during February of 2008. All genes included in this data set were selected based on their gene expression via several experimentally proven techniques, excluding those tested by microarray technology. Since microarray technology only provides initial evidence of gene expression in certain cell types and is accompanied by associated limitations (i.e. high rate of false positives), determining any meaningful level of differential expression, statistical analysis or data interpretation were debatable and therefore excluded (Pritchard *et al.*, 2001 & Smyth *et al.*, 2003). Genes that were experimentally proven by wet-laboratory techniques (*Table 3.6*) such as immunohistochemistry, western blotting, FISH (Fluorescent *In Situ* Hybridization), RT-PCR (Reverse Transcriptase-Polymerization Chain Reaction), etc. formed the final candidate gene data set.

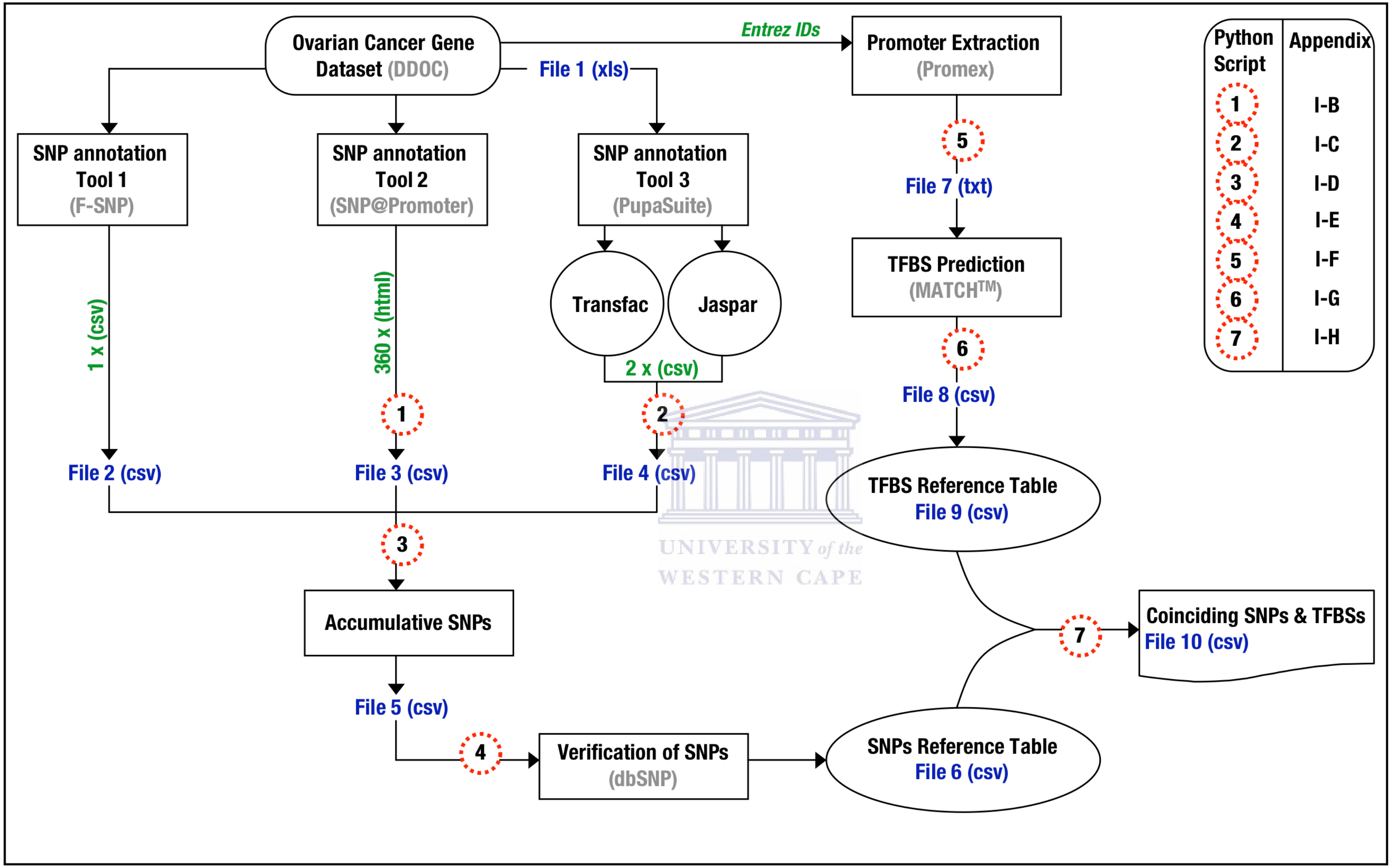

Ovarian Cancer Gene Dataset (DDOC)
Entrez IDs
File 1 (xls)
Promoter Extraction (Promex)
SNP annotation Tool 1 (F-SNP)
SNP annotation Tool 2 (SNP@Promoter)
SNP annotation Tool 3 (PupaSuite)
Transfac
Jaspar
1 x (csv)
360 x (html)
2 x (csv)
1
2
5
File 7 (txt)
TFBS Prediction (MATCHTM)
6
File 8 (csv)
File 2 (csv)
File 3 (csv)
File 4 (csv)
TFBS Reference Table File 9 (csv)
3
Accumulative SNPs
File 5 (csv)
4
Verification of SNPs (dbSNP)
SNPs Reference Table File 6 (csv)
7
Coinciding SNPs & TFBSs File 10 (csv)
Python Script
Appendix
1 I-B
2 I-C
3 I-D
4 I-E
5 I-F
6 I-G
7 I-H

Figure 3.8 Overview of methods applied to the candidate gene data set. Through the application of several tools and custom programs, the identification of SNPs coinciding with TFBSs was ultimately accomplished.

An overview of the methods applied in this chapter can be categorized into 3 key parts as follows:

PART 1

The initial set of candidate genes, obtained from the Dragon Database for the Exploration of Ovarian Cancer Genes (DDOC) (*File 1*) was queried through three independent SNP annotation tools 1, 2 & 3 (i.e. F-SNP, SNP@Promoter & PupaSuite respectively) resulting in the generation of three comma delimited files (*Files 2, 3 & 4*) containing SNP predictions by each tool respectively. Custom programs html_read.py (*Appendix I-B*) and pupasuite_read.py (*Appendix I-C*) were applied to the SNP results obtained from the SNP@Promoter and PupaSuite tools to enable/ensure the compilation of comparable data. Custom program all_snps.py (*Appendix I-D*) was then applied to the SNP results contained in Files 2, 3, & 4 (i.e. F-SNP, SNP@Promoter & PupaSuite SNP predictions) resulting in the generation of another comma delimited file (*File 5*) containing a combination of all SNP results per gene by each of the SNP annotation tools. All SNP results, then contained in File 5, were thereafter verified by querying all RefSNP IDs in a single batch query through the dbSNP database (*Section 3.1.5*). The resulting flat file was downloaded and read into a fourth custom program (*Appendix I-E*) before being written to a sixth comma delimited file (*File 6*) as the final SNP reference table to be compared with the corresponding TFBS reference table (*File 9*) acquired in Part 2 below.

PART 2

Following the generation of a SNP reference table (*File 6*), a corresponding TFBS reference table with which to compare the SNP results was then determined. To do this, the promoter regions of all candidate genes were extracted by querying the Entrez IDs of all genes through the Promex promoter extraction tool (*Section 3.1.6*). All promoter regions were then obtained in a single flat file illustrated in Figure 3.8 as File 7. This file was then read into custom program label_promoters.py (*Appendix I-F*) to refine and display selected information for each title of all promoter sequences contained in the flat file (*File 7*). The resulting re-labeled flat file was then queried through the MATCH^{TM} TFBS prediction tool (*Section 3.1.7*) before obtaining TFBS predictions for each promoter region of all candidate genes in Microsoft Excel (.xls) format.

This result file was then converted into comma delimited file format (*File 8*) for consistency and comparability (i.e. with the SNP reference table created above), before being read into the match_read.py custom program (*Appendix I-G*) and output to a final TFBS reference table (*File 9*).

PART 3

In the final part of this chapter's methodology, the SNP reference table (*File 6*) was compared to the TFBS reference table (*File 9*) through the application of another custom program (*Appendix I-H*). All SNPs that were found to coincide with TFBSs were then exported to a tenth comma delimited file (*File 10*).

3.2.2 Identification & Verification of SNPs

All candidate genes were screened for SNPs *in silico* through the implementation of three independent SNP annotation tools, as illustrated in Figure 3.8. These publicly available tools were utilized to avoid any bias as well as to ensure that any unforeseeable caveats within any one tool was compensated for by the application of a second and/or third tool.

3.2.2.1 SNP@Promoter

SNP predictions by SNP@Promoter (*Section 3.1.2*) were compiled by querying each candidate gene (*Appendix I-A*) one at a time through the tool via the “By Gene” text entry box, before clicking on the “Search” button. Entries matching the search query were then selected and all resulting webpages saved in Hyper Text Markup Language (.html) format to a local working directory. A custom program (*Appendix I-B*) designed to identify and extract all SNPs predicted per gene was then applied to the accumulated result files. Consequently, all SNP@Promoter results were compiled and exported to a comma delimited (.csv) file, reflecting the SNPs predicted in the following order:

HUGO gene symbol | RefSNP identifier | Chromosomal location | Strand orientation | Nucleotide base position

3.2.2.2 F-SNP

SNP predictions by F-SNP (*Section 3.1.3*) were manually collated by querying each candidate gene (*Appendix I-A*) one at a time through the tool. This was done first by selecting the “Query by Gene” option then entering the gene symbol into the text entry box entitled “Enter Gene Name” and clicking on the “Submit” button.

Only SNPs present in any of the following F-SNP-defined genomic regions were selected and manually entered into a Microsoft Excel (.xls) spreadsheet:

(1) “REGULATORY REGION, UPSTREAM”
(2) “REGULATORY_REGION, 3PRIME_UTR”
(3) “REGULATORY_REGION, DOWNSTREAM”
(4) “REGULATORY_REG”
(5) “REGULATORY_REGION, INTRONIC”
(6) “REGULATORY_REGION, 3P”
(7) “REGULATORY_REGION, 5”

Results were compiled in the following order:

HUGO gene symbol | RefSNP identifier | Chromosomal location | Strand orientation | Nucleotide base position

3.2.2.3 PupaSuite

All candidate genes (*Appendix I-A*) were queried through the "SNP Prioritization" page of the PupaSuite tool (*Section 3.1.4*) under the following parameter specifications:

Table 3.3 PupaSuite parameter specifications.

Option	Sub-Option	Status
Organism	-	Homo sapiens
Select your data	-	Gene list
Regulatory properties	TRANSFAC/Match predictions	Checked
	JASPAR/MatScan predictions	Checked

SNP predictions resulting from this query were received individually for Transfac and Jaspar results, in the form of two Microsoft Excel (.xls) files that were subsequently downloaded and converted into comma delimited file format to be used as an input file for the custom program shown in Appendix I-C. This script was designed to collate all SNPs predicted per gene by both Transfac and Jaspar, storing the sorted results into a new comma delimited file (.csv).

3.2.2.4 Accumulative SNP Predictions

Following the SNP predictions by the SNP@Promoter, F-SNP and PupaSuite tools, a custom program (*Appendix I-D*) designed to read through all result files from each of these tools was applied, creating a combined list of these SNP predictions. However, due to the absence of nucleotide base positions associated with each of the SNPs predicted by the PupaSuite tool, an initial file was created inclusive of all SNP predictions by the SNP@Promoter and F-SNP tools. This file was then scanned for RefSNP IDs contained within the PupaSuite result file, thereby confirming a percentage of the SNP predictions obtained by the PupaSuite tool.

3.2.2.5 SNP Verification

Verification of the accumulated SNP predictions began on the dbSNP main page, where under the "Batch" subheading, "References SNP ID (rs)" was selected. Subsequently, on the query page, a valid email address was provided (for receipt of results) in the "Email" text field; "Homo sapiens" was selected from the drop down menu below the "Organism" subheading and RefSNP (i.e. rs) IDs of all accumulative SNPs were then pasted into the query box below the "Enter RS Numbers" subheading. Finally, the "Flat file" option was selected as the preferred output format in the "Select Result Format" drop-down menu, before clicking on the "Submit" button.

Results were received in flat file format and downloaded to the working directory before being read into the verify_snps.py custom program (*Appendix I-E*) designed to search for SNPs reflected in both the accumulative SNP list derived in Section 3.2.2.4 and the dbSNP result file.

3.2.3 Promoter Sequence Retrieval

The first step in the identification of functional regulatory regions that control transcription rates was to locate and extract each of the genes' promoter regions. The promoter regions tend to be proximal to the initiation site(s) of transcription within any one gene (Qiu, 2003 & Wasserman *et al.*, 2004). Although there are no definite guidelines for the promoter-collection process, regulatory sequences are sought near transcription start sites (TSSs), as they are more likely to contain functionally important regulatory controls (Wasserman *et al.*, 2004).

In this study the promoter region was defined as the 2000bp upstream and 500bp downstream region, proximal to the gene's TSSs (i.e. -2000 - 500). Entrez IDs of all candidate genes were retrieved from the DDOC database (*Section 3.1.1*) and compiled into a flat file (one entry per line) before being uploaded into the *PROMEX: Dragon promoter extraction tool* (*Section 3.2.6*) under the following parameter specifications:

Table 3.4 PROMEX query specifications. Genes were queried via the CAGE FANTOM analysis page subject to the following parameter specifications:

Parameter	Setting
Database	FANTOM3 - H.Sapiens
Type of input parameter	Entrez gene ID
Distance	50000
Min. # of tags	5
Min. # of tags in representative tag	3
# of nucleotides upstream	2000
# of nucleotides downstream	500
Verify data	Checked

Results were received in the form of a flat file containing all promoter regions per gene and subsequently refined through the application of the label_promoters.py custom program (*Appendix I-F*).

This program (*Appendix I-F)* combined all HUGO gene symbols and corresponding Entrez IDs (sourced from the DDOC database) with the results from PROMEX, creating a new flat file with summarized titles for all promoter sequences predicted by PROMEX. All titles were concatenated into a string type label and arranged chronologically in the following order:

1. HUGO gene symbol
2. Strand orientation
3. Chromosomal location
4. Promoter sequence start position (bp)
5. Promoter sequence end position (bp)

The program (*Appendix I-F*) was run via the command line interface with the newly labeled promoter sequences piped to an output text file using the following command from the working directory:

```
“python label_promoters.py > promex_relabeled_promoters.txt”
```

3.2.4 Identification of TFBSs

In the search for SNPs with potential phenotypic effect, all promoter regions identified above (*Section 3.2.3*) consisting of 2000bp upstream and 500bp downstream region of all TSSs from each candidate gene was scanned for the presence of TFBSs.

To determine TFBSs present within these promoter regions, MATCH$^{\text{TM}}$ (Version 11.4) (*Section 3.1.7*) from the Transfac$^{®}$ Professional (Version 11.4) database was used under the following parameter specifications:

Table 3.5 MATCH$^{\text{TM}}$ parameter specifications.

Parameter	Setting
Upload a file	Promex_Relabeled_Promoters.txt
Profiles	vertebrate_non_redundant_minFP
Use only high quality matrices	Checked
minimize false positives	Checked

Results were received in the form of a flat file before being read into the match_read.py custom program (*Appendix I-G*) designed to filter through these results, isolating TFBS predictions per gene in the following order before storing results into a new comma delimited (.csv) file:

HUGO gene symbol | Strand orientation | Chromosomal location | TFBS (Nucleotide base position [Start-Stop]) | Transcription Factor | Matrix Identifier

3.2.5 Elucidating SNP-TFBS Overlap

For the elucidation of SNPs that occurred within TFBSs, the overlaps.py custom program (*Appendix I-H*) was applied to the data obtained from the MATCH^{TM} results in Section 3.2.4 and final list of verified SNPs resulting from Section 3.2.2.5. This program (*Appendix I-H*) was designed to identify SNP-TFBS overlap by cross matching two comma delimited (.csv) files (*Figure 3.8, Files 6 and 9*) for the elucidation of coinciding pairs.

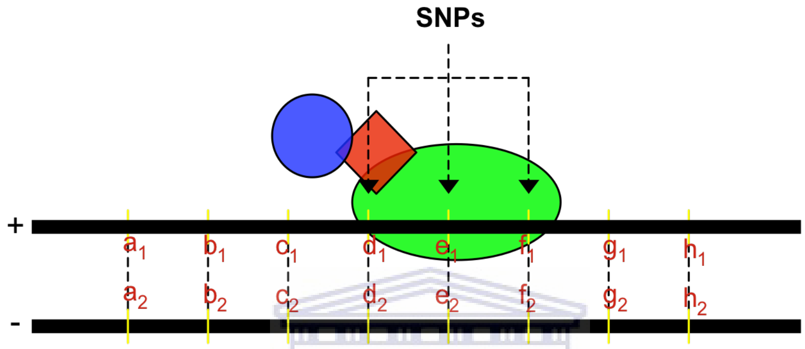


Figure 3.9 Determination of SNPs that coincide with TFBSs. Due to the binding of transcription factors (i.e. represented by the green molecule) across more than a single nucleotide base position (i.e. in this case positions d_1, e_1 & f_1), SNPs that were predicted at any one of these positions were classified as SNPs coinciding within TFBSs.

Coinciding SNPs and TFBSs were exported into a comma delimited output file (*Figure 3.8, File 10*) from the overlaps.py program (*Appendix I-H*) as illustrated by Figure 3.8.

3.3 Results

3.3.1 Ovarian Cancer Candidate Gene Dataset

A total of 379 candidate genes were extracted from the Dragon Database for the Exploration of Ovarian Cancer Genes (DDOC). These genes were observed among 22 of the 23 human chromosomes as illustrated below.

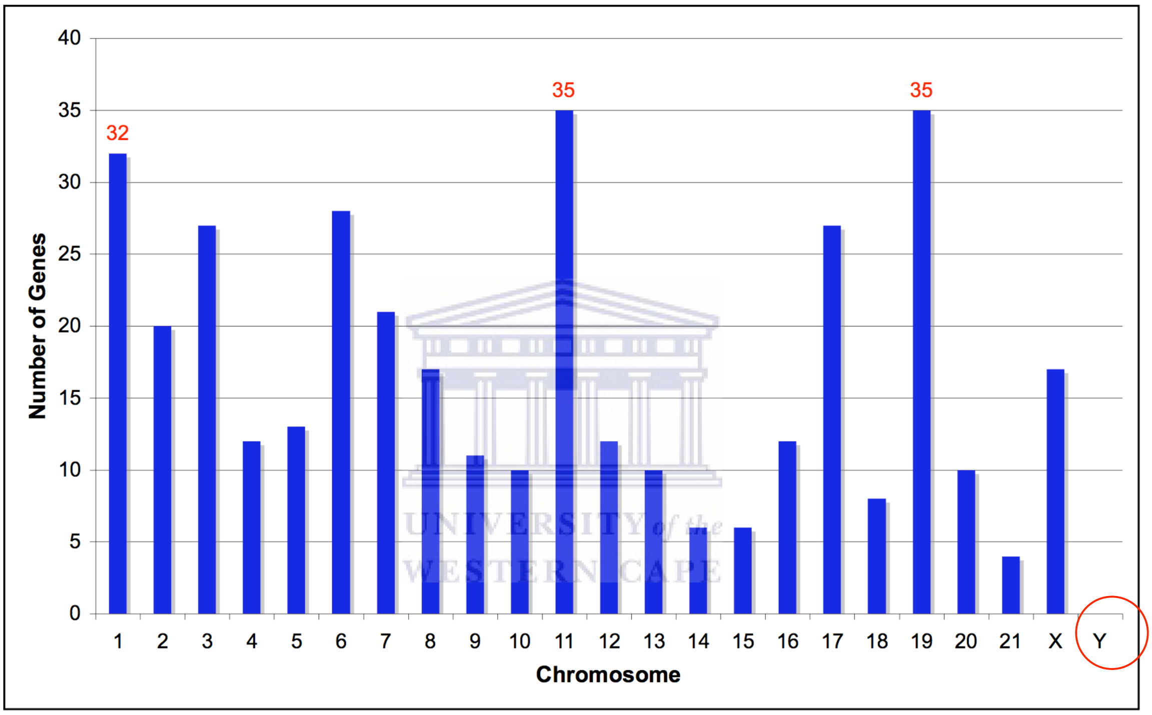


Figure 3.10 Chromosomal distributions of candidate genes. Chromosomes 1, 11 and 19 each comprise of more than 30 of the genes implicated in ovarian cancer with chromosome Y expectedly containing none of the candidate genes.

All candidate genes contained within the Dragon Database for the Exploration of Ovarian Cancer Genes (DDOC) were experimentally tested via any one of 28 techniques shown in Table 3.6.

Table 3.6 List of techniques employed in the experimental testing of the candidate gene data set. All techniques used to validate or experimentally prove the association between genes included in the DDOC and ovarian cancer, have been through laboratory-based analysis.

Number	Technique
1	Immunoassay
2	Cell lysis
3	Cell proliferation assays
4	Chemiluminescence Immunoassay
5	Combined Bisulfite Restriction Analysis (COBRA)
6	Denaturing High Performance Liquid Chromatography (DHPLC)
7	Electrophoretic mobility shift assays
8	Platinum/Paclitaxel-based Chemotherapy
9	Facs analysis
10	FISH (Fluorescent *In Situ* Hybridization)
11	Flow Cytometry
12	Immunoblotting
13	Immunocytochemistry
14	Immunofluorescence
15	Immunohistochemistry
16	*In situ* hybridization
17	Mass spectrometry
18	Methylation-specific PCR
19	Microscopy
20	Northern blotting
21	Polymerase Chain Reaction
22	ELISA
23	RT-PCR
24	Radioimmunoassay
25	SDS-PAGE Gelatin Zymography
26	SNP analysis
27	Southern blotting
28	Western blotting

Three techniques: immunohistochemistry, RT-PCR and western blotting were observed to have been used more extensively than the remaining 25 techniques listed in Table 3.6.

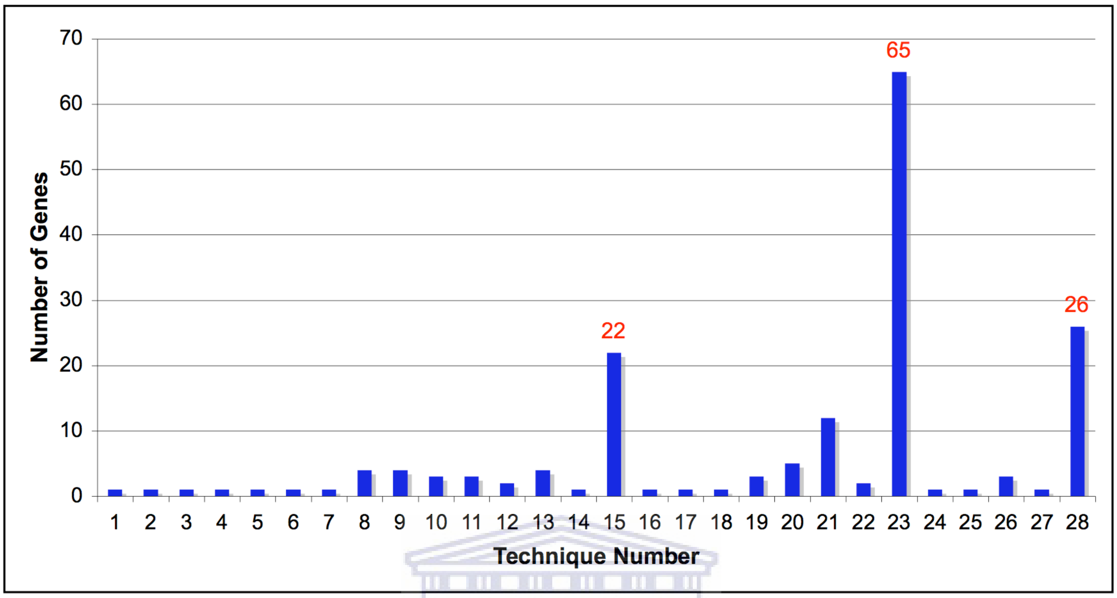


Figure 3.11 Foremost techniques employed in the experimental proof of the candidate gene data set. Technique numbers 15, 23 and 28 were observed to be the most widely used in the experimental testing of candidate genes. These correspond (*Table 3.6*) to the analysis of 22 candidate genes studied via immunohistochemistry; 65 studied via RT-PCR and 26 analyzed through western blotting.

3.3.2 Identification & Verification of SNPs

A total of 10863 SNPs were predicted by the SNP@Promoter, F-SNP and PupaSuite tools, denominated as shown in Figure 3.12 below. From the total of 10863 SNPs predicted, 97% of these SNPs were obtained from SNP@Promoter, with 3% arising from F-SNP. PupaSuite results (i.e. SNP predictions by Transfac and Jaspar) amounted to a total of 90 SNPs, 42 of which were identified in the accumulated SNP@Promoter and F-SNP result file (*Figure 3.8, File 6*).

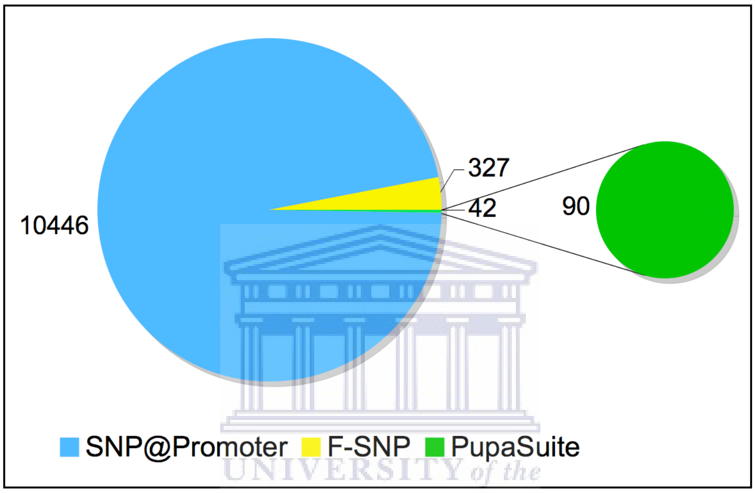


Figure 3.12 Total numbers of SNPs predicted by each of the SNP annotation tools used for ovarian cancer candidate gene analysis. SNP predictions by SNP@Promoter account for the largest percentage of SNP results as this tool was designed to analyze the regions 5kb upstream to 500bp downstream of the transcription start sites of genes and all SNP predictions obtained by this tool were included in this study. In comparison, only SNPs specifically occurring within putative TFBSs (*Table 3.3*) or regulatory regions (*Section 3.2.2.2*) were included from the PupaSuite and F-SNP tools.

The SNP@Promoter tool was able to analyze a total of 360 of the 379 candidate genes, with four genes observed to contain a noticeably high number of SNPs (i.e. more than 100 SNPs) when compared to the numbers of SNPs predicted on the remaining 356 candidate genes.

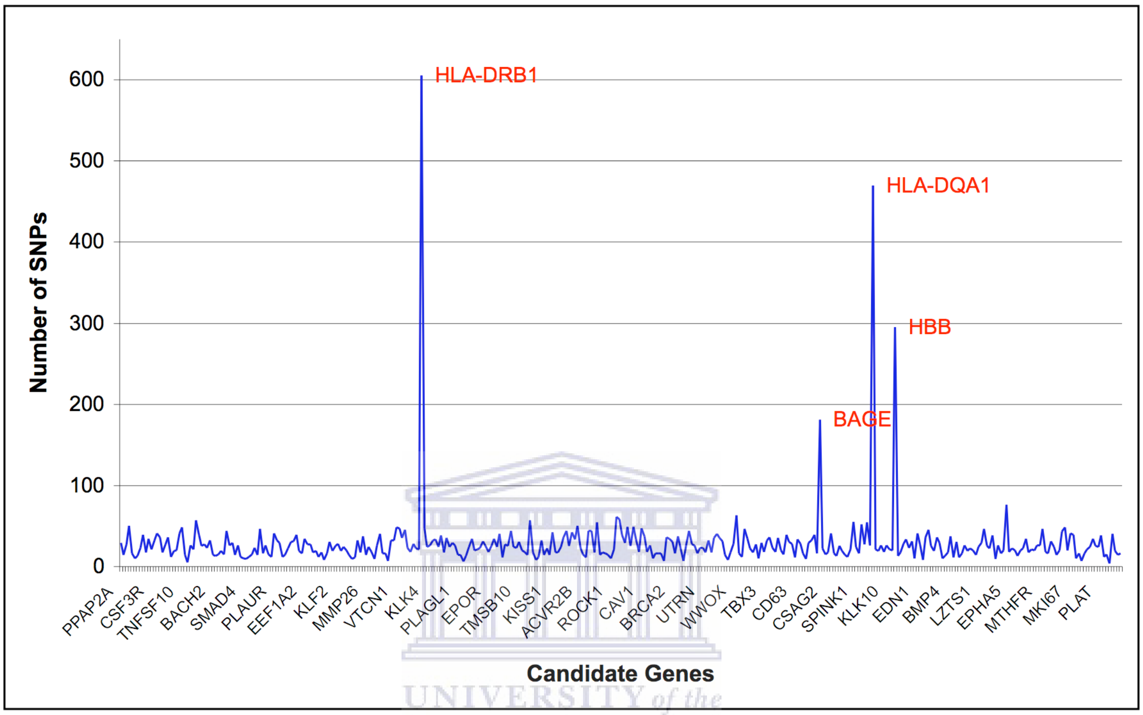


Figure 3.13 Frequency of SNP occurrences on candidate genes as predicted by the SNP@Promoter tool. HLA-type D genes *HLA-DRB1* and *HLA-DQA1* were observed to contain the highest numbers of SNPs when compared to the remaining 358 candidate genes that were analyzed, with 605 and 469 SNPs present on each respectively. *HBB* and *BAGE*, similarly, were found to contain 295 and 181 SNPs respectively. All other candidate genes contained an average of approximately 25 SNPs per gene.

The F-SNP tool was able to identify SNPs present within regulatory regions (Section 3.2.2.2) on a total of 128 of the 379 candidate genes (*Appendix I-A*), with four genes observed to contain a noticeably high number of SNPs (i.e. more than 15 SNPs) when compared to the numbers of SNPs predicted on the remaining 124 candidate genes.

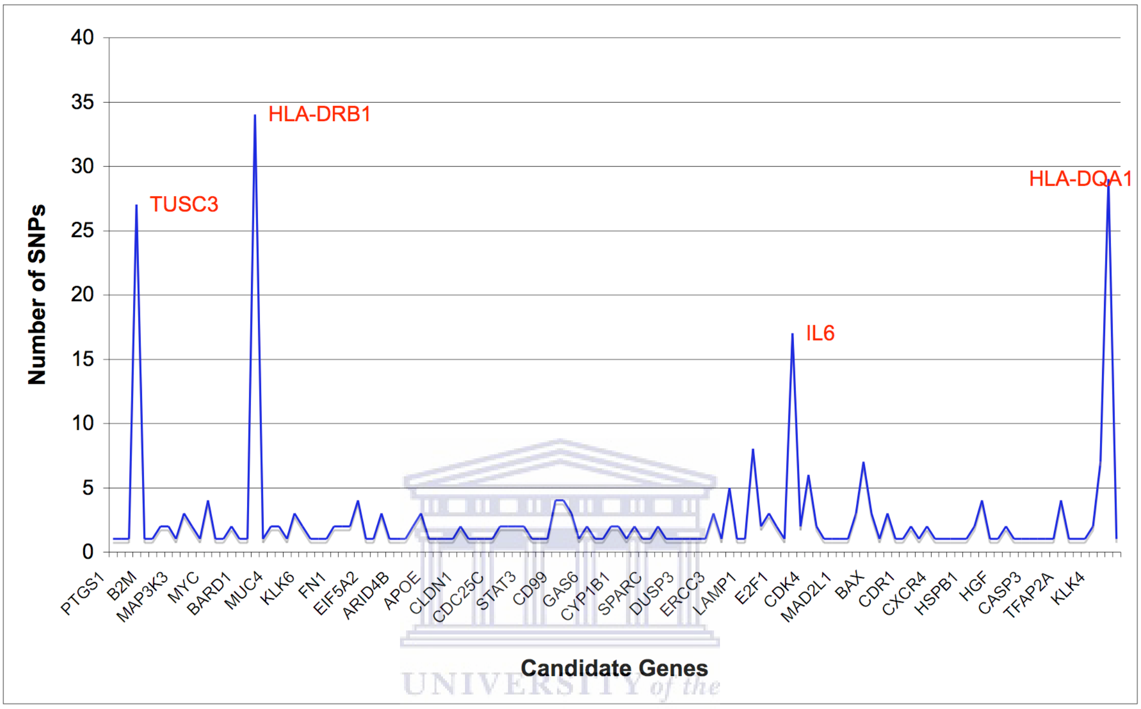


Figure 3.14 Frequency of SNP occurrences on candidate genes as predicted by the F-SNP tool. HLA-type D genes, *HLA-DRB1* and *HLA-DQA1* were observed to contain the highest numbers of SNPs once again when compared to the other 126 candidate genes that were analyzed. F-SNP identified 34 SNPs on *HLA-DRB1* with 29 SNPs found on *HLA-DQA1*. *TUSC3* and *IL6* were found to contain 27 and 17 SNPs respectively, while all other candidate genes (i.e. excluding *HLA-DRB1*, *HLA-DQA1*, *TUSC3* and *IL6*) contained an average of approximately 2 SNPs per gene identified within a regulatory region.

The PupaSuite tool was able to analyze 371 of the 379 candidate genes with one gene observed to contain a noticeably large number of SNPs predicted within putative TFBSs when compared to the numbers of SNPs predicted within TFBSs among the other 370 candidate genes that were analyzed.

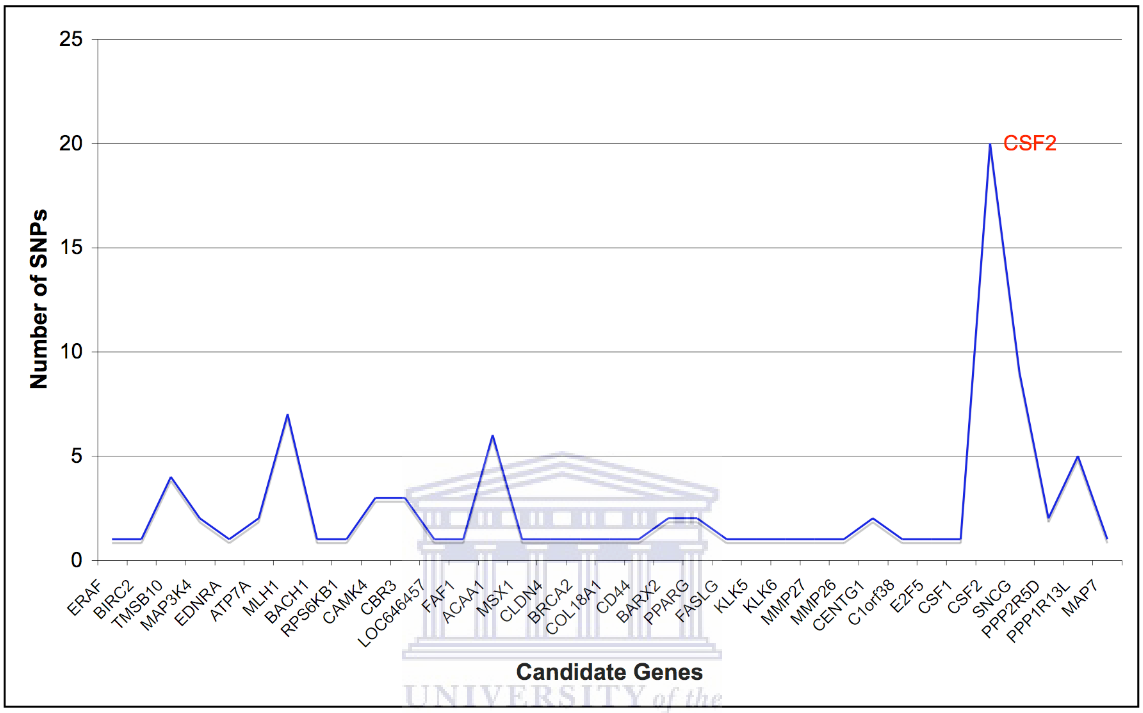


Figure 3.15 Frequency of SNP occurrences within putative transcription factor binding sites on candidate genes as predicted by the PupaSuite tool. A total of 20 SNPs were identified within putative TFBSs on the *CSF2* gene, noticeably higher than the SNP predictions obtained for any other candidate gene. From the total of 371 candidate genes that were analyzed, only 34 genes (excluding *CSF2*) were found to contain an average of approximately 2 SNPs within a putative TFBS per gene.

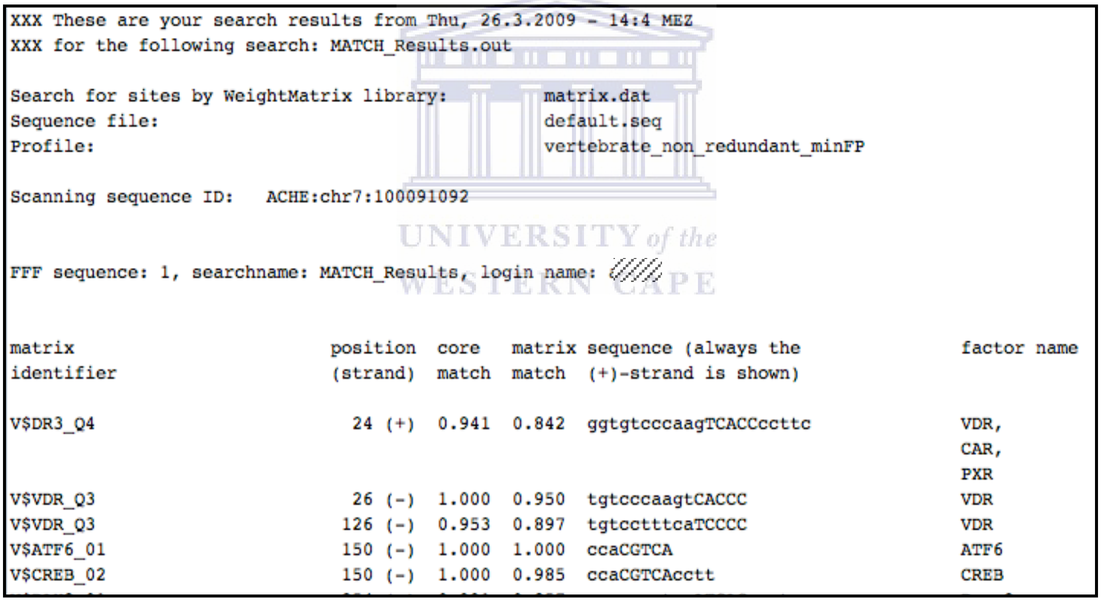
From the overall number of SNPs identified by the SNP@Promoter, F-SNP and PupaSuite tools, 10773 were verified via comparison with SNPs in the dbSNP (Build 129) database (i.e. 133 unverified) and included in the final comparison with TFBS predictions described in Section 3.2.5.

3.3.3 Promoter Sequence Retrieval

A total of 1155 promoter regions were obtained (*Section 3.2.3*) from the 379 candidate genes (*Appendix I-A*). This correlates to approximately 3.0395 transcription start sites (TSSs) predicted per gene within the 2000bp upstream and 500bp downstream genomic regions.

3.3.4 Identification of TFBSs

TFBS results from the MATCH^{TM} tool were based on the criteria specified in Table 3.5, which ensured the reduced number of random sites found by the tool and inclusion of putative sites only with a good similarity to the weight matrix selected.



```
XXX These are your search results from Thu, 26.3.2009 - 14:4 MEZ
XXX for the following search: MATCH_Results.out

Search for sites by WeightMatrix library:          matrix.dat
Sequence file:                                     default.seq
Profile:                                           vertebrate_non_redundant_minFP

Scanning sequence ID:   ACHE:chr7:100091092


FFF sequence: 1, searchname: MATCH_Results, login name: [redacted]


matrix                         position  core   matrix sequence (always the         factor name
identifier                     (strand)  match  match  (+)-strand is shown)

V$DR3_Q4                         24 (+)  0.941  0.842  ggtgtcccaagTCACCccttc         VDR,
                                                                                     CAR,
                                                                                     PXR
V$VDR_Q3                         26 (-)  1.000  0.950  tgtcccaagtCACCC               VDR
V$VDR_Q3                        126 (-)  0.953  0.897  tgtcctttcaTCCCC               VDR
V$ATF6_01                       150 (-)  1.000  1.000  ccaCGTCA                      ATF6
V$CREB_02                       150 (-)  1.000  0.985  ccaCGTCAcctt                  CREB
```

Figure 3.16 MATCH^{TM} tabulated result output. All matches that were higher than the cut-off scores and able to fulfill the parameter specifications described in Section 3.2.4 were reported in this result spreadsheet containing the matrix ID, position of the match, strand orientation (forward (+) or reverse (-)), core similarity score, matrix similarity score and corresponding nucleotide sequences and names of transcription factors associated with the TFBSs identified.

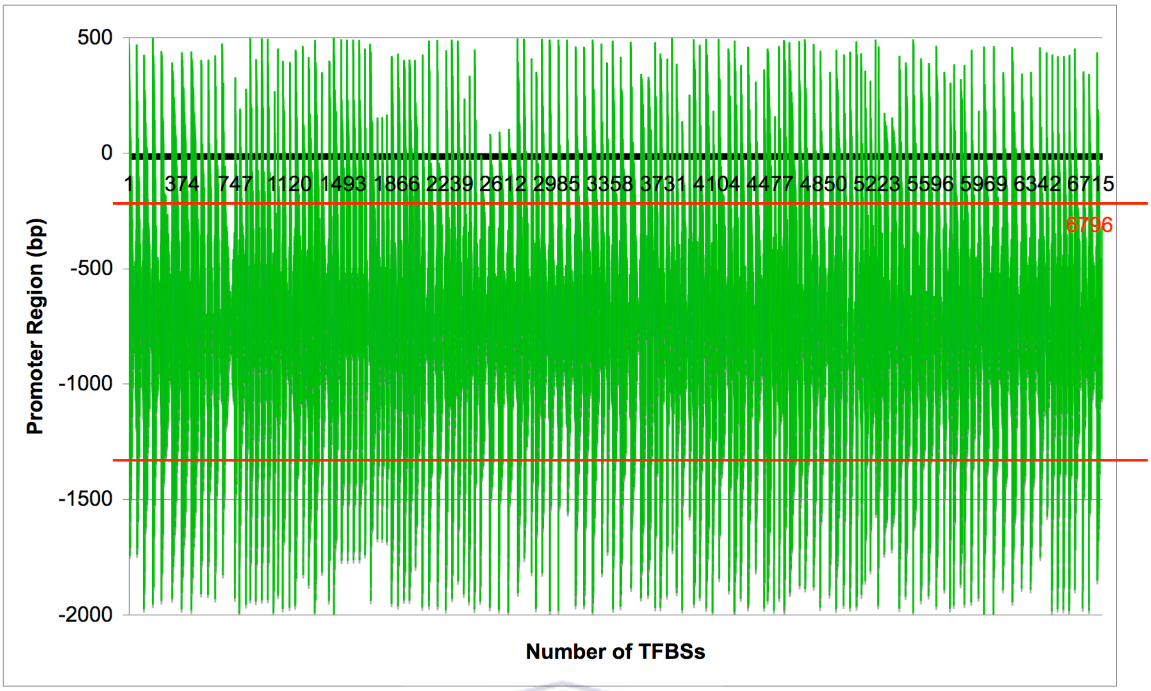


Figure 3.17 TFBS distribution between the 2000bp upstream to 500bp downstream promoter regions of all candidate genes. From the total of 6796 TFBSs predicted by the MATCH^{TM} tool, the highest concentration of TFBSs were found to be present within the upstream regions of approximately 200 to 1400bp of the TSS (indicated by the dense green frequency of TFBS predictions highlighted between the red margins (upper -200 to lower -1400)), with fewer TFBSs observed further away from the TSS (0bp).

3.3.5 Elucidating SNP-TFBS Overlap

Following the methodologies applied in Section 3.2.5 of this chapter, a list of 121 SNPs overlapping with TFBSs were found on a total of 121 of the ovarian cancer candidate gene data set (*Appendix I-I*). From these overlapping regions, 57 unique transcription factors that bind to these putative TFBSs were implicated and are shown in Table 3.7 below.

Table 3.7 Transcription factors found to occur at binding sites that coincide with the occurrence of SNPs. Shown here is a non-redundant list of the 121 transcription factors that were implicated by the occurrence of SNPs within the binding sites of these TFs.

Number	Transcription factor	Number	Transcription factor	Number	Transcription factor
1	AIRE	20	GATA-4	39	S8
2	Pax-4	21	c-Ets-1	40	Ets
3	HIF1	22	FAC1	41	LRF
4	TBX5	23	ETF	42	SREBP
5	Tal-1beta:E47	24	YY1	43	CP2/LBP-1c/LSF
6	myogenin	25	SREBP-1	44	GR
7	PLZF	26	ZF5	45	C/EBPdelta
8	Hand1:E47	27	CACD	46	v-Myb
9	Pax	28	POU3F2	47	AP-2alpha
10	CDP	29	Pax-8	48	PPARalpha:RXRalpha
11	Pax-5	30	Tax/CREB	49	CdxA
12	AP-2	31	Cart-1	50	c-Ets-1(p54)
13	Pax-3	32	Pax-2	51	POU6F1
14	RFX	33	C/EBP	52	E2F
15	GATA-X	34	ER	53	STAT
16	MyoD	35	Spz1	54	HIC1
17	Pax-6	36	HNF3alpha	55	VDR
18	Sp1	37	1-Oct	56	MRF-2
19	WT1	38	Egr-1	57	Muscle

3.4 Summary

This chapter focused on the elucidation of single nucleotide polymorphisms that overlapped with transcription factor binding sites within genes that have been differentially expressed in ovarian cancer. In doing so, a total of 379 candidate genes that have been experimentally associated with ovarian cancer were obtained for analysis. These genes were queried through three independent publicly available SNP annotation tools before being verified through a public SNP repository. All verified SNPs were collated into a SNP reference table including the following fields respectively: (1) HUGO gene symbol of gene within which SNP was identified, (2) RefSNP ID of SNP, (3) Chromosomal location of SNP, (4) Strand orientation of gene on which the SNP was located (i.e. forward or reverse) and (5) Nucleotide base position of SNP occurrence.

To determine if these SNPs coincided with putative TFBSs on any of the candidate genes, a TFBS reference table was also assembled with which to compare the SNP reference table to. This was done by extracting the promoter regions of all candidate genes and querying these regions through the MATCH$^{\text{TM}}$ tool. All results obtained from this step were collated into a TFBS reference table including the following fields respectively: (1) HUGO gene symbol of gene within which TFBS was predicted, (2) Strand orientation of promoter sequence on which the TFBS was predicted (i.e. forward or reverse), (3) Chromosomal location of gene on which TFBS was predicted, (4) TFBS start and stop nucleotide base positions, (5) Transcription factor that binds at the predicted TFBS, (6) the Matrix identifier of the TFBS predicted (7) Matrix Similarity Score (MSS) of the prediction and (8) the Core Similarity Score (CSS) of the prediction.

The use of a custom program (*Appendix I-H*) to compare the TFBS and SNP reference tables resulted in the elucidation of 121 SNPs that were found to coincide with TFBSs on 121 of the original 379 candidate genes (*Appendix I-I*). These SNPs are considered to have a putative phenotypic effect in the expression of the genes in which they were found by influencing the binding of transcription factors (*Table 3.7*) at these sites.

CHAPTER 4

Population Association

ASSOCIATING DNA variants with diseases has been used widely to identify regions of the genome and candidate genes that contribute to disease (Cardon *et al.*, 2001). SNPs are, as a result of this, generally used for association studies to identify genes partly responsible for complex diseases (Xu *et al.*, 2005). Theoretically, identifying common SNPs associated with a disease would involve the time-consuming and expensive task of genotyping millions of SNPs in individuals with and without a disease, before searching for sites that differ in frequency between the sampled groups (Manolio *et al.*, 2008).

SNPs have been proposed, by some, as the new frontier for population studies with several papers having presented evidence reporting the advantages and limitations of this type of diagnostic marker (Morin *et al.*, 2008). To benchmark the *in silico* identified SNPs established in chapter 3 of this study, each of the 11 population groups defined by the International HapMap Project were analyzed for the presence of these SNPs. The aim of this chapter therefore was to examine several population groups for the observation of SNPs or SNP patterns that have been found to coincide with transcription factor binding sites (TFBSs) in chapter 3 of this study (*Appendix I-I*). Through the combination of HapMap project data and the application of Perl (*Section 4.1.2*) and Python (*Section 3.1.8*) custom programs, these SNPs or SNP patterns are intended to constitute a suitable and more widely applicable basis for a SNP profile able to detect the presence of ovarian cancer before it is able to become invasive.

4.1 Introduction

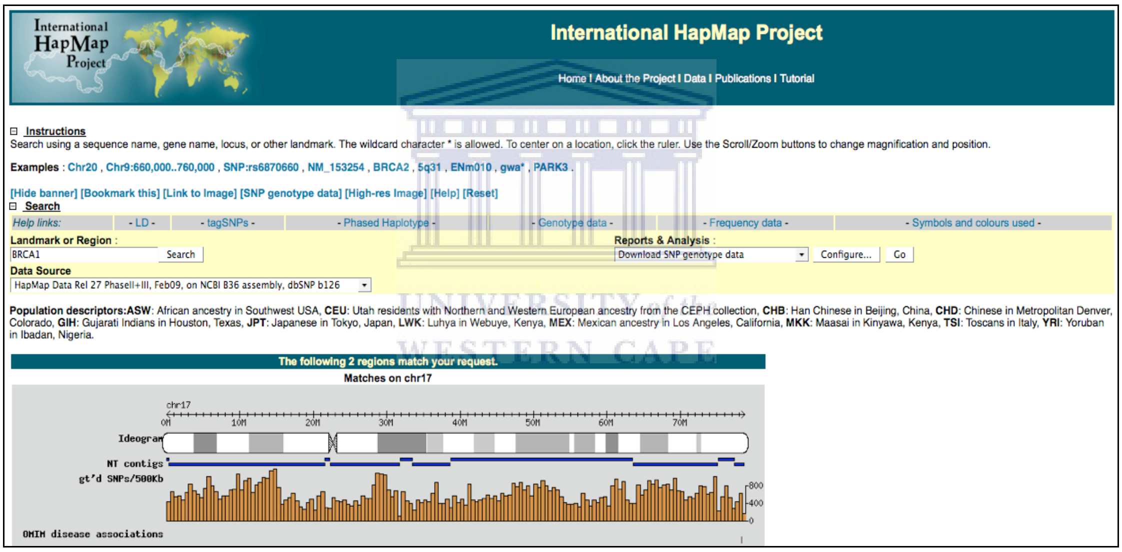
The use of population studies for insights into complex genes are giving researchers an advantage in identifying single gene defects despite their value in researching complex genes having been questioned (Peltonen *et al.*, 2000; Tabor *et al.*, 2002 & Luikart *et al.*, 2003). Association studies may be the best approach to the study of genetic features of population isolates and potentially unlock the genetics of complex diseases (Peltonen *et al.*, 2000 & Tabor *et al.*, 2002). Complex diseases differ in severity of symptoms and age of onset, and can show variance in the numerous biochemical pathways that they may influence, however small insights into the population dynamics of candidate genes have a greater statistical efficiency and a key advantage in the understanding of tissues, genes and proteins involved in disease (Tabor *et al.*, 2002 & Balding, 2006). With single gene defects, phenotypes can be diagnosed reliably and haplotype signatures used to map genes (Peltonen *et al.*, 2000). Careful dissection of disease phenotypes is required to minimize genetic heterogeneity, and haplotype mapping essential as follow-up to linkage analysis (Peltonen *et al.*, 2000).

Linkage disequilibrium (LD) analysis is required to map disease genes by association (Peltonen *et al.*, 2000 & Tabor *et al.*, 2002), with analysis of disease genes and variable markers or haplotypes depending strongly on the understanding of linkage disequilibrium (Tishkoff *et al.*, 2002). Haplotypes may have LD patterns that are distinct in various populations and may be subjected to SNP influence in varying levels of frequency (Tishkoff *et al.*, 2002). By investigating candidate SNPs in combination with population genetics scientists have discovered a formidable foundation within which to identify disease-related genes in humans through the application of computational and statistical methods (Luikart *et al.*, 2003).

Other statistical approaches in population studies have been suggested by Balding (2006), who found that use of the Hardy-Weinberg equilibrium test infers more accurately occurring SNPs or haplotypes from genotypes (Balding, 2006). Combined with association tests such as case-control phenotypes this has potentially provided future researchers with a greater insight into genetic associations (Tishkoff *et al.*, 2002 & Balding, 2006). These methods, when conducted in tandem with environmental epigenomics and disease susceptibility may provide the key to mapping the control of ovarian cancer. The epigenomics, which include genomic imprinting and monitors slight changes in gene expression, have already given rise to a greater understanding of developmental disorders associated with imprinted regions and genes (Jirtle *et al.*, 2007). Imprinted gene deregulation or mutation increases cancer risk, as a single mutation or epigenetic event is required to completely inactivate imprinted tumor suppressor genes, as imprinting functionally inactivates one allele (Jirtle *et al.*, 2007).

4.1.1 HapMap Genome Browser (GBrowse)

Despite there being a number of public online resources developed to provide high-volume genome-wide data sets such as the UCSC Genome Browser (http://genome.ucsc.edu) and EnsEMBL project (www.ensembl.org), the lack of flexibility for combining data from within and between each database does not allow for the calculation of key population variability statistics (Jorge *et al.*, 2008 & Smith, 2008a). The HapMap Genome Browser (GBrowse) was created with this distinct focus (Smith, 2008a). GBrowse aims to be a resource capable of retrieval, display and analysis of high-throughput and high quality genome-wide human genetic variation with an emphasis on disease association studies (Smith, 2008a).


International HapMap Project
International HapMap Project
Home I About the Project I Data I Publications I Tutorial
[Hide banner] [Bookmark this] [Link to Image] [SNP genotype data] [High-res Image] [Help] [Reset]
Help links:
- LD -
- tagSNPs -
- Phased Haplotype -
- Genotype data -
- Frequency data -
- Symbols and colours used -
Search
Configure...
Go
chr17
0M
10M
20M
30M
40M
50M
60M
70M
Ideogram
NT contigs
gt'd SNPs/500Kb
800
400
0
OMIM disease associations

⊟ Instructions

Search using a sequence name, gene name, locus, or other landmark. The wildcard character * is allowed. To center on a location, click the ruler. Use the Scroll/Zoom buttons to change magnification and position.

Examples : Chr20 , Chr9:660,000..760,000 , SNP:rs6870660 , NM_153254 , BRCA2 , 5q31 , ENm010 , gwa* , PARK3 .

⊟ Search

Landmark or Region : BRCA1

Reports & Analysis : Download SNP genotype data

Data Source: HapMap Data Rel 27 PhaseII+III, Feb09, on NCBI B36 assembly, dbSNP b126

Population descriptors:ASW: African ancestry in Southwest USA, **CEU**: Utah residents with Northern and Western European ancestry from the CEPH collection, **CHB**: Han Chinese in Beijing, China, **CHD**: Chinese in Metropolitan Denver, Colorado, **GIH**: Gujarati Indians in Houston, Texas, **JPT**: Japanese in Tokyo, Japan, **LWK**: Luhya in Webuye, Kenya, **MEX**: Mexican ancestry in Los Angeles, California, **MKK**: Maasai in Kinyawa, Kenya, **TSI**: Toscans in Italy, **YRI**: Yoruban in Ibadan, Nigeria.

The following 2 regions match your request.

Matches on chr17

Figure 4.1 View of the *HapMap Genome Browser (GBrowse) main page after submission of a query term. Depending on the computer language settings, this page can appear in one of several languages, displaying a range of results pertaining to the query term under user-specified subheadings (Smith, 2008b).

Users may query the GBrowse tool by entering any of the following search terms into the "Landmark or Region" search box:

(1) Chromosome name (e.g. "Chr10")

(2) Chromosomal start to stop position (e.g. "Chr10:25000..30000")

* http://hapmap.org/cgi-perl/gbrowse/hapmap27_B36/

(3) Reference SNP Identifier (e.g. "rs12345")

(4) NCBI RefSeq Accession Number (e.g. "NM 12345")

(5) HUGO Gene Symbol (e.g. "*BRCA1*")

(6) Chromosomal band (e.g. "5q31")

In this study the "*HapMap Genome Browser (Phase 1, 2 & 3 - merged genotypes & frequencies)*" project data was selected for the survey of SNPs occurring among a total of 11 population groups shown in Table 4.1.

Table 4.1 List of population group samples that are included in the HapMap Phase 3 project data set. The HapMap Phase 3 project data set includes the collection of 1301 samples (i.e. including the original 270 samples from Phase 1 and 2 of the project) from 11 population groups listed alphabetically here by their 3-letter labels (Broad Institute, 2008).

Label	Population Sample	Number of Samples
ASW	African ancestry in Southwest USA	90
CEU	Utah residents with Northern and Western European ancestry	180
CHB	Han Chinese in Beijing, China	90
CHD	Chinese in Metropolitan Denver, Colorado	100
GIH	Gujarati Indians in Houston, Texas	100
JPT	Japanese in Tokyo, Japan	91
LWK	Luhya in Webuye, Kenya	100
MEX	Mexican ancestry in Los Angeles, California	90
MKK	Maasai in Kinyawa, Kenya	180
TSI	Toscans in Italy	100
YRI	Yoruba in Ibadan, Nigeria	180

This release of HapMap project data includes SNP genotype data that has been generated from 1115 (558 males, 557 females) of the 1301 total samples collected, using either of two platforms (i.e. Illumina Human1M and Affymetrix SNP 6.0) (Broad Institute, 2008). It also includes PCR-based resequencing data across ten 100kb regions in 712 samples contributed by the Baylor College of Medicine Human Genome Sequencing Center (Broad Institute, 2008).

Data from the two platforms were merged using PLINK ("--*merge-mode 1*") and include only genotype calls with a consensus between non-missing genotype calls (i.e. merged genotype was set to missing if the two platforms gave different "non-missing calls") (Broad Institute, 2008). Quality control at the individual sample level was performed separately and includes only individuals with genotype data that were present on both platforms (Broad Institute, 2008). Only SNPs that satisfied the following criteria were included in the HapMap phase 3 data release (Broad Institute, 2008):

(1) Hardy-Weinberg p-value of more than 0.000001 (per population)
(2) Missingness value less than 0.05 (per population)
(3) Less than 3 Mendel errors (per population; only applied to YRI, CEU, ASW, MEX & MKK)
(4) SNP must have a RefSNP Identifier and map to a unique genomic location

4.1.2 Perl

Originally developed by Larry Wall in 1987 as a general purpose Unix scripting language, Perl is a "high-level, interpreted, dynamic" language that provides powerful text processing facilities (Wall *et al.*, 2000). Perl is used for graphics programming, system administration, network programming, CGI programming and applications that require access to databases (Wall *et al.*, 2000). It is a flexible and adaptable language that was designed to be practical while borrowing features from other programming languages such as “C, shell scripting, AWK and sed”, (Wall *et al.*, 2000).

4.2 Methodology

4.2.1 SNPs vs. Population Groups

The HapMap Genome Browser (GBrowse) was used to determine the prevalence of SNPs within the candidate gene dataset (*Appendix I-A*) among all population groups defined by the International HapMap Project. Each gene was queried one at a time through the GBrowse tool, via the application of the read_hapmap.pl custom program (*Appendix II-A*) created in Perl (*Section 4.1.2*). This program was run from the command line interface using the following commands, respectively:

1	"perl read_hapmap.pl candidate_genes.txt fwd fwd_results"
2	"perl read_hapmap.pl candidate_genes.txt rev rev_results"

Command 1 was used to obtain all SNPs on the forward strands of all population groups for each candidate gene before storing these results into an output folder entitled "fwd_results". Command 2 was used to obtain all SNPs on the reverse strands of all population groups for each candidate gene before storing these results into an output folder entitled "rev_results".

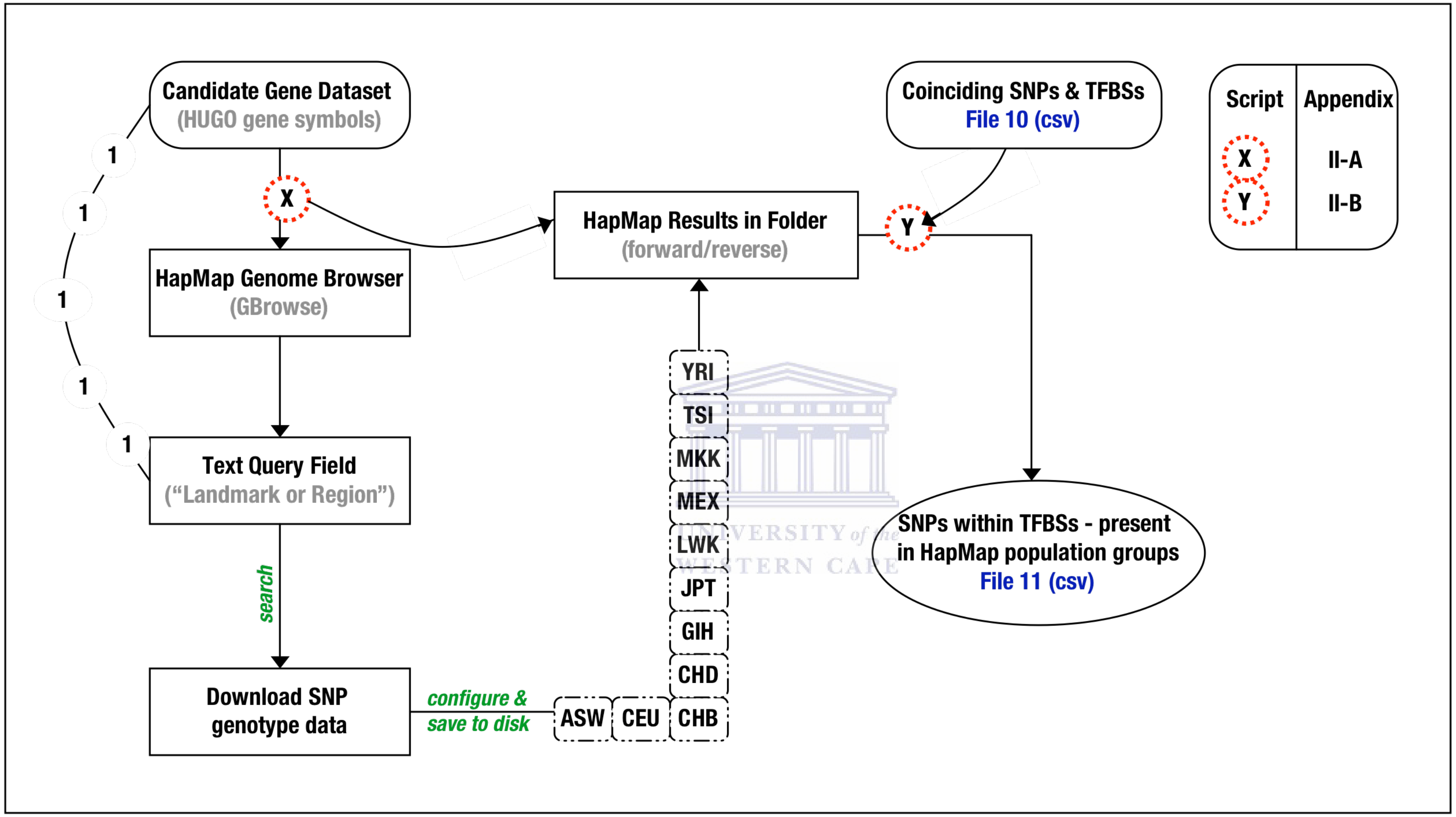


Figure 4.2 Overview of methods applied to the identification of commonly occurring SNPs within HapMap-defined population groups. Through the application of custom programs, the identification of commonly occurring SNPs within all 11 population groups in the HapMap project's Phase 3 data release was accomplished.

The overview of methods applied in this Chapter can be categorized into 2 key parts as follows:

PART 1

The HUGO gene symbols of all candidate genes (*Appendix I-A*) were queried one at a time through the HapMap Genome Browser (GBrowse) tool through the use of custom program read_hapmap.pl (*Appendix II-A*), resulting in the collection of genotyped SNP data for all 379 candidate genes per population group (i.e. included in HapMap Phase 3 data) on both the forward and reverse strands.

PART 2

All SNP genotype data obtained from the HapMap GBrowse tool were then filtered using the filter_hapmap.py custom program (*Appendix II-B*).

4.3 Results

4.3.1 SNPs vs. Population Groups

Following the application of the read_hapmap.pl program in Section 4.2.1, a total of 8338 flat files containing SNP predictions for each candidate gene (i.e. on the forward and reverse strands) among all 11 population groups were obtained and stored into a single working directory.

From the total of 5865604 SNPs predicted on the forward strands and 5878560 SNPs on the reverse strands of all population groups, 23852 common SNPs (i.e. present on all of the 11 population groups listed in Table 4.1) were identified on the forward strands with 24133 SNPs common SNPs identified on the reverse strands.

4.3.2 HapMap SNPs Coinciding with TFBSs

From the list of 121 SNPs identified in chapter 3 (*Table 3.7*) that were found to coincide with TFBSs, only 3 SNPs (*Table 4.2*) were identified among the 11 population groups in the HapMap phase 3 project data.

An overall average of the core similarity and matrix similarity scores for all three predictions by MATCH^{TM} revealed a 95.55% chance that the TFBSs for the C/EBP, CP2/LBP-1c/LSF and CDP transcription factors were present on genes *E2F5*, *TNFRSF10A* and *CIITA* respectively. SNPs rs4150842, rs20577 and rs12928665 presented new alleles to these genes in varying degrees among each of the population groups sampled by the HapMap consortium (*Table 4.1*) as indicated in Table 4.3.

Table 4.2 List of SNPs coinciding with TFBSs that were identified among population groups included in the HapMap Phase 3 project data. SNPs predicted on genes *E2F5*, *TNFRSF10A* and *CIITA* that potentially influence the binding of transcription factors C/EBP, CP2/LBP-1c/LSF and CDP respectively have been predicted with reasonably high core similarity and matrix similarity scores.

Gene	RefSNP ID	Nucleotide base position	Transcription factor affected	CSS	MSS
E2F5	rs4150842	86277150	C/EBP	0.997	0.997
TNFRSF10A	rs20577	23138422	CP2/LBP-1c/LSF	1	0.918
CIITA	rs12928665	10878975	CDP	1	0.839

CSS and MSS values of 0.997 for the prediction of C/EBP at nucleotide position 86277150 on the promoter region of the *E2F5* gene suggests a 99,7% chance that this transcription factor occurs at this position on the forward strand of this gene. Similarly, there was a 95% chance that the CP2/LBP-1c/LSF transcription factors occurred at position 23138422 of the reverse strand of the *TNFRSF10A* gene and a 91.95% chance that CDP occurred at position 10878975 on the forward strand of the *CIITA* gene.

Table 4.3 Allele frequencies of SNPs coinciding with TFBSs that were identified within HapMap population groups. SNP rs12928665 was observed within all 11 population groups, whereas SNPs rs4150842 and rs20577 were only identified in 5 and 9 population groups, respectively.

rs#	Allele	ASW	CEU	CHB	CHD	GIH	JPT	LWK	MEX	MKK	TSI	YRI
rs4150842	C	97%	100%	N/A	N/A	N/A	N/A	97%	99%	N/A	N/A	97%
	T	3%	0%	N/A	N/A	N/A	N/A	3%	0%	N/A	N/A	3%
rs20577	G	74%	0%	0%	98%	99%	0%	74%	N/A	89%	N/A	68%
	A	26%	100%	100%	2%	1%	100%	26%	N/A	11%	N/A	32%
rs12928665	A	78%	76%	44%	42%	64%	40%	84%	87%	81%	74%	88%
	G	22%	24%	56%	58%	36%	60%	16%	13%	19%	26%	12%

Subsequent to the observation of the ancestral nucleotide bases (within the TFBSs corresponding to transcription factors C/EBP, CP2/LBP-1c/LSF and CDP) on genes *E2F5*, *TNFRSF10A* and *CIITA*, the frequencies of each allele for all SNPs were examined. SNP rs4150842 was observed to be present in percentages of 97% or higher in the ASW, CEU, LWK, MEX and YRI population groups, indicating a 3% or lower chance that the minor allele (T) for these groups would be present. This would indicate that in 97% or above of the cases that the rs4150842 SNP occurred, the C/EBP transcription binding to its corresponding The TFBS constituting the ancestral allele of this SNP, would in most cases be present in its ancestral allele form. For percentages of 3% or lower, that showed the presence of the minor allele, SNP rs4150842 would present a hindrance or influence the binding of the C/EBP molecule to its corresponding TFBS in the ASW, CEU, LWK, MEX and YRI population groups.

SNP rs20577 represented as alleles G and A, with G being the ancestral allel, was present in high frequencies among population groups ASW, CHD, GIH, LWK, MKK, YRI implying only a 26%, 2%, 1%, 26%, 11% and 32% of the occurrence of the minor A allele respectively. However, the minor allele of the same SNP was observed to be present in 100% of the CEU, CHB and JPT population samples, implying the potential influence of this SNP in the binding of the CP2/LBP-1c/LSF transcription factors to their corresponding TFBS.

SNP rs12928665, present in all 11 population groups was observed on average in 68% of the population samples in the form of its ancestral allele (i.e. A) which corresponds to the TFBS configuration of the CDP transcription factor. The minor G allele was present in an average of 31% of the cases that this SNP occurred on the *CIITA* gene among all population samples.

4.4 Summary

Beginning with the collection of 121 SNPs found to overlap with TFBSs in chapter 3 (*Appendix I-I*), this chapter focused on the observation of any one of these SNPs among the 11 population groups defined by the International HapMap project. To do this, SNP genotype data was obtained from the HapMap Genome Browser tool with a custom program read_hapmap.pl (*Appendix II-A*). Subsequent to this, a custom program filter_hapmap.py filtered this data for the identification of SNPs (from the original list of 121) that were present within one or more of the 11 population groups. This resulted in the identification of 3 novel SNPs with potential phenotypic effect associated with 5, 9 and 11 population groups respectively.

SNP rs12928665, occurring within the *CIITA* gene, was found to be present in all population groups, with a higher percentage of the ancestral allele (A) being present in 7 of the 11 population groups, when compared to the presence of the minor allele (G). Population groups ASW, CEU, GIH, LWK, MEX, MKK, TSI and YRI indicated higher percentages of the ancestral allele (A), whereas CHB, CHD and JPT indicated higher percentages of the presence of the minor allele (G).

SNP rs20577, occurring within the *TNFRSF10A* gene, was found to be present in 9 population groups. Population groups ASW, CHD, GIH, LWK, MKK and YRI indicated higher percentages of the ancestral allele (G), whereas CEU, CHB and JPT indicated a 100% possibility of the minor allele (A) being present. This may imply that the binding of the CP2/LBP-1c/LSF transcription factors to their corresponding TFBSs may potentially be altered/influenced by the change in its TFBS nucleotide base configuration in the CEU, CHB and JPT population groups through the occurrence of the rs20577 SNP within the putative binding site of the CP2/LBP-1c/LSF transcription factors.

SNP rs4150842, occurring within the promoter region of the *E2F5* gene, was found to be present in 5 population samples. Population groups ASW, CEU, LWK, MEX and YRI were found to have higher percentages of the ancestral allele (C), with percentages of 3% or lower of the minor allele (T) present.

This may indicate that in 97% of the instances that the rs4150842 SNP occurs within the *E2F5* gene, the C/EBP transcription factor would bind to its corresponding (unaltered) TFBS whereas in only 3% or less of these occurrences it may potentially be influenced by the change in its TFBS nucleotide base configuration in these 5 population groups.

CHAPTER 5

Discussion

THE potential of a diagnostic marker that can be measured in blood is a high priority in view of the profuse amounts of patients that have been diagnosed with ovarian cancer all too late. Although promising markers have been reported in the last decade to contribute to this target (e.g. CA125, etc.) the use of these markers to detect ovarian cancer early enough to reduce mortality rates remains a challenge since these screening methods must be able to identify the cancer before it becomes invasive, i.e. early enough for the disease to be curable (Coukos *et al.*, 2008). Until now, only the CA125 assay has been able to detect ovarian cancer before symptoms arise (Coukos *et al.*, 2008) but with a high rate of false positive predictions (Vuillez *et al.*, 1997; McIntosh *et al.*, 2004 & Mahata, 2006).

The focus of this study was to identify SNPs that may have a potential to influence the expression of genes implicated in ovarian cancer. To do this, an original collection of 379 candidate genes were isolated from the Dragon Database for the Exploration of Ovarian Cancer Genes (DDOC), and their chromosomal distributions elucidated and mapped. The three highest concentrations of candidate genes were found to be located on chromosomes 1, 11 and 19, stimulating the assumption that these chromosomes are candidates to investigate for potential female-specific diseases. Chapter 3 (*Figure 3.11)* also presents the techniques whereby the candidate genes were experimentally proven, with the majority of the genes being proven via RT-PCR, western blotting and immunohistochemistry; wet-laboratory based techniques that have been implemented and practiced widely in the elucidation of molecular data.

From the 379 candidate genes studied, a total of 10863 SNPs were identified through the application of three SNP annotation tools. These SNPs constituted the total number of SNPs identified within the specific promoter regions analyzed by each SNP annotation tool, based on the criteria described in Sections 3.2.2.1, 3.2.2.2 and 3.2.2.3 of this study. A large difference in the number of SNPs obtained from each of the SNP annotation tools was observed, with 97% of the SNPs arising from the SNP@Promoter tool as illustrated in Figure 3.13. This was the result of the inclusion of all SNP results obtained from the SNP@Promoter tool within the 5kb upstream and downstream regions of all genes, whereas SNPs included from the F-SNP and PupaSuite tools only constituted those present within regulatory regions of the same promoter region or putative TFBSs (defined by Jaspar and Transfac position weight matrices) in the 5kb upstream promoter region of all genes, respectively (*Figures 3.14 and 3.15*).

Furthermore, the SNP results depicted above indicated higher SNP densities for the *HLA-DRB1*, *HLA-DQA1*, *HBB*, *BAGE*, *TUSC3*, *IL6* and *CSF2* genes (*Figures 3.13, 3.14 and 3.15*). The HLA genes are important in helping the immune system in distinguishing the body's own proteins from proteins made by foreign invaders such as viruses and bacteria (Genetics Home Reference, 2009). They are highly polymorphic (Pénzes *et al.*, 1999), and the HLA region has also been associated with genetic predisposition to diseases in Asian populations (Bouma *et al.*, 1997 & Keicho *et al.*, 1998). The *HBB* gene, located on chromosome 11, alongside *HBA* is responsible for normal adult haemoglobin structure, and mutations in this gene have been associated with diseases such as sickle-cell anemia and thallasemia (National Centre for Biotechnology Information, 2009).

BAGE is a gene located on chromosome 21 that codes for a the tumor antigen protein that are recognized by lymphocytes (National Centre for Biotechnology Information, 2009), which stimulates the body's immune system to find and eradicate cancer cells (National Centre for Biotechnology Information, 2009).

TUSC3 is a candidate tumor suppressor gene found on chromosome 8, while the *IL6* (chromosome 7) gene encodes a cytokine that functions in the inflammation and the maturation of B cells. *CSF2* found on chromosome 5 is responsible for the control of production, differentiation and function of granulocytes and macrophages (National Centre for Biotechnology Information, 2009), thereby playing a vital role in cellular health.

From the list of 10863 SNPs mapped to the original 379 candidate genes, 10773 were verified via comparison with SNPs catalogued in the dbSNP database (Build 129), to eradicate false positive SNP predictions that may have been reported by any of the three SNP annotation tools employed in sections 3.2.2.1-3.2.2.3 of this study. The dbSNP database has provided this study with an accurate grouping of SNPs to be compared with TFBS predictions obtained from the MATCH^{TM} tool (*Figure 3.16*).

To predict TFBSs present within the ovarian cancer related genes, promoter regions classified as the 2000bp upstream and 500bp downstream regions of all TSSs present on each of the 379 candidate genes, were extracted and queried through the MATCH^{TM} tool. This was done as most putative regulatory regions are identified within the promoter regions of genes (Hunninghake *et al.*, 1989; Ahlgren *et al.*, 1990; Horie *et al.*, 1996 & Savon *et al.,* 1997).

A total of 6796 high quality TFBSs (i.e. high core similarity and matrix match scores) were mapped to 1155 promoter regions on all 379 genes, with the highest concentration of TFBSs identified within the upstream regions of 200bp to 1400bp of the TSSs for the candidate gene data set as described by Figure 3.17. This suggested that the maximum numbers of TFBSs were included in the specified promoter region of 2000bp upstream and 500bp downstream regions (as classified by this study). The approach of the analyses utilized several custom programs for data handling and integration, facilitated by published, highly utilized databases and tools. This has led to the creation of a workflow on which SNPs with potential phenotypic effect may be elucidated in search of potential ovarian cancer biomarkers.

While, many biomarker studies have been unsuccessful as a result of variation within individuals' tissue localisations (Naylor, 2003 & Mayeux, 2007) as well as between individuals within a population (Nielsen *et al.*, 2005), utilizing SNP annotation technologies and prediction of SNPs within regulatory sequences of genes associated with ovarian cancer provides scientists with new potential therapeutic targets in response to the disease.

The SNP patterns that influence function may reflect common haplotypes in a population suggesting that there may exist functionally significant interaction between SNPs and regulatory regions according to the haplotype context (Chen *et al.,* 2001).

Chapter 4 aimed to identify SNP biomarkers present within several population groups defined by the International HapMap consortium. In doing so, a custom program designed to extract SNP genotype data was created and applied to all candidate genes. This resulted in the identification of 23852 common SNPs present on the forward strands of these genes in all 11 of the population groups, and 24133 common SNPs identified on the reverse strands. The reason for the great difference in the numbers of SNPs between the data collected by the HapMap Genome Browser and the SNP annotation tools applied in chapter 3, is that the SNP prediction tools analyzed specified promoter regions, whereas the HapMap project provides SNP predictions across the entire sequence length of the gene.

From the 121 SNPs found to overlap with TFBSs in chapter 3, only three were identified among these 11 population groups. Despite the absence of 198 SNPs in the HapMap data set, these SNPs are not considered as any less important to the phenotypic expression of ovarian cancer and may be investigated in future work.

SNPs rs4150842, rs20577 and rs12928665 identified on genes *E2F5*, *TNFRSF10A* and *CIITA* genes, were found to occur in 5, 9 and 11 population groups respectively. Because these are not limited to a specific population group, they may be broadly applicable as potential markers for ovarian cancer, when compared to the targeting of SNPs that are present in only one or two population groups. This observation however, excludes the predisposition of smaller population groups that may possess biomarkers or SNPs that may be implicated or linked to ovarian cancer but not represented in the HapMap project data set or represented in lower frequencies.

The tumor necrosis factor receptor (*TNFRSF10A*) gene, belonging to the superfamily A (member 10) encodes for the *death receptor 4* protein that mediates apoptosis. The regulatory region of the *TNFRSF10A* gene has been shown to reside in a conserved region of cysteine-rich domain, resulting in the development of prostate cancer when compromised via SNP rs20576 (Langsenlehner *et al.*, 2008). The putative TFBS undergoes a change from its original compilation including a guanine base to one of that including an adenine nucleotide base in this case.

In chapter 4 of this study it was shown that the rs20577 SNP within the same gene present as the ancestral allele (G) occurred in higher frequencies in the ASW, CHD, GIH, LWK, MKK and YRI population groups. This implies that the putative TFBS nucleotide base (coinciding with rs20577) constituting the G allele is less likely to be present as the minor allele (A) in these population groups. Alternatively, population groups CEU, CHB and JPT were shown to possess the minor allele (A) 100% of the instances that it occurred in these population samples, indicating the potential loss of the putative nucleotide base (G) on the *TNFRSF10A* gene. This change potentially influences the binding of the CP2 or LBP-1c or LSF transcription factors to their putative TFBSs on this gene. The presence of the minor allele (A) in these population groups corresponds to the Utah residents with Northern and Western European ancestry, Han Chinese in Beijing and Japanese in Tokyo population groups respectively. This suggests that the presence of this allele is specifically associated to these population groups. The CP2 or LBP-1c or LSF transcription factors also known as alpha-CP2, alpha CP2a, Late SV40 factor or SEF, has been shown to play a part in the activation of the SV40 late promoter transcription process (Lambert *et al.*, 2000).

Moreover, it has been reported to influence the risk of Alzheimer's disease (Lambert *et al.*, 2000). The presence and/or influence of the rs20577 SNP on this gene implies that the death receptor 4 protein, encoded by the *TNFRSF10A* gene may be rendered dysfunctional/altered in some way and potentially result in the onset or progression of ovarian cancer, affecting the role/s of the CP2/LBP-1c/LSF transcription factors.

The Class II transactivator (*CIITA*) gene plays a role in regulating cellular immune recognition (van der Stoep *et al.*, 2002). Inactivation or interferences with *CIITA* regulation has been closely associated with the absence of *HLA-DR* induction, implying that the body is unable to distinguish between its own proteins and those of foreign invaders (Satoh *et al.*, 2004). Chapter 4 of this study highlights the occurrences of the A/G alleles introduced by the rs12928665 SNP on the promoter region of this gene, indicating a higher prevalence of the ancestral allele (A) in population groups ASW, CEU, GIH, LWK, MEX, MKK, TSI and YRI. Alternatively, population groups CHB, CHD and JPT showed a higher presence of the minor allele (G), indicating the potential influence of the CDP transcription factor binding to its putative TFBS on this gene. The presence of the minor allele (G) in these population groups corresponds to the Han Chinese in Beijing, Chinese in Metropolitan Denver and Japanese in Tokyo population groups respectively. This suggests that the presence of this allele is specifically associated to the population class that may be categorized as those from Eastern/Southern Asian decent.

The CCAAT displacement protein (CDP) is a repressor protein that has been observed to interact with the special AT-rich sequence binding protein 1 (SATB1) (Liu *et al.*, 1999). These proteins have been reported to potentially be regulated by each other via protein-protein interaction (Liu *et al.*, 1999), suggesting the loss of regulatory control be either one in the absence of the other. The high incidences of the minor allele (G) if SNP rs12928665 located in the promoter region of the *CIITA* gene in population groups CHB, CHD and JPT may influence the binding between the CDP transcription factor and its putative TFBS, implicating the protein-protein interaction between the SATB1 and CDP molecules. The consequences of this hypothetical observation on the progression or pathogenesis of ovarian cancer may only be determined by the experimental testing of each of these scenarios.

The *E2F* genes (*E2F1*, *E2F3* and *E2F5*) have been documented in retinoblastoma, bladder, lung, ovarian and prostate cancers (Johnson *et al*., 2006). The protein encoded by the *E2F5* gene plays a crucial role in the control of cell cycle and action of tumor suppressor proteins (Reimer *et al*., 2006). This protein is differentially phosphorylated and is expressed in a wide variety of human tissues (GeneCards, 2009). Chapter 4 of this study illustrated the presence of the C and T alleles of SNP rs4150842 occurring within the regulatory regions of the *E2F5* gene. These alleles indicate a change from the ancestral allele (G), which suggests the potential influence of the binding of the C/EBP transcription factor to its putative TFBS on this gene.

The CCAAT enhancer-binding protein (C/EBP) is composed of 5 subfamilies of TFs (i.e. C/EBPdelta, C/EBPbeta, C/EBPalpha, C/EBPepsilon and C/EBPgamma), each with their own individual sets of interacting factors. More precisely the C/EBP molecule has been shown to interact with approximately 26 other factors, forming ± three protein complexes with a few of these factors (Yang *et al*., 2003). Regulated by glucose and insulin, this gene has been characterized as a putative insulin-responsive element in the rat genome (Maytin *et al*., 1999; Chen *et al*., 2001 & Sugiyama *et al*., 2001). Moreover, the occurrence of SNP rs4150842 on the *E2F5* gene that potentially affects the binding of the C/EBP molecule, may subsequently implicate the functionalities of these 26 other interacting factors in the expression profile of the *E2F5* gene.

CHAPTER 6

Conclusions

THE most promising approach to the reduction of mortality rates for ovarian cancer is through the early intervention of the disease. This will require the identification of a screening test that is able to detect the disease through the presence of a precursor within high-risk women before it is able to become invasive. This approach would ensure the reduction of disease incidence rates and increase of survival rates for those affected. Although SNP pattern recognition (i.e. SNP profiling) may not serve as a sole solution for the prognosis of high-risk women, it may be recognized that collaboration of this technique with existing ones may be required to identify and validate a more efficient diagnostic and early detection method (Coukos *et al.*, 2008).

This study has demonstrated the use of a simple protocol for the identification of SNPs that potentially affect transcription factor binding. These SNPs could be the causal factors for changes in the expression profile of genes. While this protocol illustrates a computational and scalable approach to the identification of SNPs related to diseases, it has in addition generated notable findings.

The combined application of approaches in this study has identified three SNPs (i.e. rs4150842, rs20577 & rs12928665) among genes *TNFRSF10A*, *CIITA* and *E2F5* that may be useful as diagnostic markers for the potential early detection of ovarian cancer. These SNPs and their associated implications on gene expression however, require experimental validation through the potential application of SNP microarrays. SNP microarrays are designed to genotype SNPs, capable of reporting hybridization of DNA fragments and therefore can be used for the purpose of detecting genomic fragments (McCann *et al.*, 2007).

The use of SNP arrays will be of great value to the ongoing search for ovarian cancer biomarkers if these SNPs are experimentally shown to be causal factors in the phenotypic expression of ovarian cancer, they could be considered as additive test measures (i.e. be coupled to existing ovarian cancer prognosis techniques, such as the CA125 assay). This may provide a plausible prognosis technique to many high-risk women.

The use of SNPs as early detection or susceptibility biomarkers for the prognosis of ovarian cancer has great potential in both diagnosing and strategizing the most optimal treatment plan for the disease. Because there are an estimated 3 billion SNPs along the human genome and are evolutionarily conserved, they are easy to follow in population studies and have attracted the attention of pharmaceutical companies with their potentially huge financial prospects (Nomikos, 2006).

Future Directions

SCREENING for ovarian cancer risk markers may be an important objective to explore in future work, because of the many challenges to early detection of curable invasive tumors. One of the aspects to be addressed in future work is the reduction of mortality rates from ovarian cancer and increase in cure rates by any number of feasible measures.

Following the results obtained from this study, the focus of future work should include the experimental validation of SNPs rs12928665, rs20577 and rs4150842, and their overlapping with the putative TFBSs to which the transcription factors CDP, CP2/LBP-1c/LSF and C/EBP bind.

Furthermore, the effects of these SNPs coinciding with TFBSs on genes *E2F5*, *TNFRSF10A* and *CIITA* should be investigated experimentally, to determine the potential influence on post-operative drug efficacy and/or drug resistance caused by the loss or gain of transcription factors that may bind to or form complexes with post-operative ovarian cancer drugs. These results may suggest how SNPs occurring within TFBSs (potentially influencing the binding of TFs) affect progression, regression or susceptibility of/for ovarian cancer.

Bibliography

Abnizova I, Subhankulova T & Gilks WR, *Recent Computational Approaches to Understand Gene Regulation: Mining Gene Regulation In Silico*. Current Genomics, 2007. **8**(79-91).

Ahlgren R, Simpson ER, Waterman MR & Lund J, *Characterization of the promoter/regulatory region of the bovine CYP11A (P-450scc) gene. Basal and cAMP-dependent expression*. Journal of Biological Chemistry, 1990. **265**(6).

Alkema WBL, Johansson Ö, Lagergren J & Wasserman WW, *MSCAN: identification of functional clusters of transcription factor binding sites*. Nucleic Acids Research. 2004. **32**(W195-W198).

Anderson L, *Ovarian cancer: The search for an accurate screening technique.* Journal of the American Academy of Physical Assistants, 2009. **22**(2).

Aranda A & Pascual A, *Nuclear Hormone Receptors and Gene Expression.* Physiological Reviews, 2001. **81**.

Balding DJ, *A tutorial on statistical methods for population association studies.* Nature Genetics, 2006. **7**(781-791).

Benos PV, Bulyk ML & Stormo GD, *Additivity in protein-DNA interactions: how good an approximation is it?*. Nucleic Acids Research, 2002. **30**(20).

Berg J, Willmann S & Lässig M, *Adaptive evolution of transcription factor binding sites*. BMC Evolutionary Biology, 2004. **4**(42).

Blagden S & Gabra H, *Future Directions in the Management of Epithelial Ovarian Cancer*. Future Oncology, 2008. **4**(3).

Bouma G, Pool MO, Crusius JBA, Schreuder GMT, Hellemans HPR, Meijer BUGA, Kostense PJ, Giphart MJ, Meuwissen SGM & Pena AS, *Evidence for genetic heterogeneity in IBD. HLA genes in the predisposition to suffer from ulcerative colitis and Crohn's disease*. Clinical and Experimental Immunology, 1997. **109**(1).

Bozek K, Kiełbasa SM, Kramer A, Herzel H, *Promoter analysis of mammalian clock controlled genes.* Genome Informatics, 2008. **18**(1).

Bozek K, Kiełbasa SM, Kramer A, Herzel H, *Promoter analysis of Mammalian clock controlled genes.* Genome Informatics. International Conference on Genome Informatics, 2007. **18**(65-74).

Broad Institute, *HapMap 3*. 2008 [cited May 08, 2009]; Available from: http://www.broad.mit.edu/mpg/hapmap3/.

Bupa. *Cancer – a general overview*. 2007 [cited September 11, 2008]; Available from: http://hcd2.bupa.co.uk/fact_sheets/html/cancer.html.

Bussemaker HJ, Li H & Siggia ED, *Regulatory element detection using correlation with expression*. Nature Genetics, 2001. **27**.

Cai Y, He J, Li X, Lu L, Yang X, Feng K, Lu W, Kong X, *A novel computational approach to predict transcription factor DNA binding preference.* Journal of Proteome Research, 2009. **8**(2).

Cancer Association of South Africa. *Types of cancer*. 2008 [cited May 05, 2009]; Available from: http://www.cansa.org.za/cgi-bin/giga.cgi?cmd=cause_dir_news&cat=751&cause_id=1056.

Cancer Research UK, *Women's cancer (gynaecological cancer)*. 2008 [cited May 05, 2009]; Available from: http://www.cancerhelp.org.uk/help/default.asp?page=17924.

CancerIndex, *Gynacological Cancers*. 2003 [cited May 05, 2009]; Available from: http://www.cancerindex.org/clinks3g.htm.

Cardon LR & Bell JI, *Association study designs for complex diseases*. Nature Reviews, 2001. **2**(91).

Carlson CS, Eberle MA, Reider MJ, Smith JD, Kruglyak L & Nickerson DA, *Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans*. Nature Genetics, 2003. **33**(518-521).

Centro de Investigacion Principe Felipe. *PupaSuite*. 2008 [cited March 25, 2009]; Available from: http://pupasuite.bioinfo.cipf.es.

Chakravarti A, *Single nucleotide polymorphisms… to a future of genetic medicine.* Nature, 2001. **409**.

Chen X, Zhang SL, Pang L, Filep JG, Tang SS, Ingelfinger JR & Chan JS, *Characterization of a putative insulin-responsive element and its binding protein(s) in rat angiotensinogen gene promoter: regulation by glucose and insulin.* Endocrinology, 2001. **142**(6).

Chowdhury MA, Kuivaniemi H, Romero R, Edwin S, Chaiworapongsa T & Tromp G, *Identification of novel functional sequence variants in the gene for peptidase inhibitor 3*. BMC Medical Genetics, 2006. **7**(49).

Conde L, Vaquerizas JM, Dopazo H, Arbiza L, Reumers J, Rousseau F, Schymkowitz J & Dopazo J, *PupaSuite: finding functional single nucleotide polymorphisms for large-scale genotyping purposes.* Nucleic Acids Research, 2006. **34**(W621-W625).

Coukos G, Berchuck A & Ozols R, *Ovarian Cancer: State of the Art and Future Directions in Translational Research.* Springer, 2008. ISBN 0387689664, 9780387689661.

Crawford DC & Nickerson DA, *Definition and clinical importance of haplotypes*. Annual Review of Medicine, 2005. **56**(303-320).

ecancermedia. *Ovarian cancer*. 2008 [cited September 02, 2008]; Available from: http://www.ecancermedia.com/Ovarian_Cancer_Factsheet.aspx.

Edvardsen H, Alnæs GIG, Tsalenko A, Mulcahy T, Yuryev A, Lindersson M, Lien S, Omholt S, Syva¨nen A-C, Børresen-Dale A-L & Kristensen VN, *Experimental validation of data mined single nucleotide polymorphisms from several databases and consecutive dbSNP builds.* Pharmacogenetics and Genomics, 2006. **16**(207-217).

Elf J, Li G-W & Xie S, *Probing Transcription Factor Dynamics at the Single-Molecule Level in a Living Cell*. Science, 2007. **316**(1191-1194).

Elnitski L, Jin VX, Farnham PJ & Jones SJM, *Locating mammalian transcription factor binding sites: A survey of computational and experimental techniques*. Genome Research, 2006. **16**(1455-1464).

EverythingBio. *Definition of real-time PCR*. 2007 [cited April 28, 2009]; Available from: http://www.everythingbio.com/glos/definition.php?word=real-time+PCR.

Fessele S, Maier H, Zischek C, Nelson PJ & Werner T, *Regulatory context is a crucial part of gene function*. TRENDS in Genetics, 2002.**18**(2).

Flintoft L, *No simple answer to complex disease*. Nature Reviews, 2004. **5**(886).

Frazer KA, Murray SS, Schork NJ, Topol EJ, *Human genetic variation and its contribution to complex traits*. Nature Reviews Genetics, 2009. **10**(4).

Frericks M, Burgoon LD, Zacharewski TR, Esser C, *Promoter analysis of TCDD-inducible genes in a thymic epithelial cell line indicates the potential for cell-specific transcription factor crosstalk in the AhR response.* Toxicology and Applied Pharmacology, 2008. **232**(2).

Fu Y & Weng Z, *Improvement of TRANSFAC Matrices Using Multiple Local Alignment of Transcription Factor Binding Site Sequences*. Genome Informatics, 2005. **16**(1).

GalaxoSmithKline. *Pharmacogenetics*. 2006 [cited September 11, 2008]; Available from: http://www.genetics.gsk.com/determin.htm.

GeneCards. *E2F transcription factor 5, p130-binding*. 2009 [cited May 15, 2009]; Available from: http://www.genecards.org/cgi-bin/carddisp.pl?gene=E2F5.

Genetics Home Reference. *HLA gene family*. 2009 [cited May 11, 2009]; Available from: http://ghr.nlm.nih.gov/geneFamily=hla.

Genfit. *How key biological processes are controlled by gene regulation.* 2009 [cited November 6, 2007]; Available from: http://www.genfit.com/en/science-discovery/technology-br-expertise/scientific-expertise/index.html.

Gibbs RA, Belmont JW, Hardenbol P, Willis TD, Yu F, Yang H, Ch'ang L-Y, Huang W, Liu B, Shen Y, Tam P K-H, Tsui L-C, Waye MMY, Wong J T-F, Zeng C, Zhang Q, *et al.*, *The International HapMap Project*. Nature, 2003. **426**(789-796).

GnpSNP. *SNP (Single Nucleotide Polymorphism)*. 2009. [cited August 6, 2009]; Available from: http://urgi.versailles.inra.fr/projects/GnpSNP/general_documentation.php.

Goffart S & Wiesner RJ, *Regulation and co-ordination of nuclear gene expression during mitochondrial biogenesis*. Experimental Physiology, 2003. **88**(33-40).

GuhaThakurta D, Xie T, Anand M, Edwards SW, Li G, Wang SS & Schadt EE, *Cis-regulatory variations: A study of SNPs around genes showing cis- linkage in segregating mouse populations*. BMC Genomics, 2006. **7**(235).

Hannenhalli S & Levy S, *Predicting transcription factor synergism*. Nucleic Acids Research, 2002. **30**(19).

Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne J-B, Reynolds DB, Yoo J, Jennings EG, Zeitlinger J, Pokholok D, Kellis M, Rolfe PA, Takusagawa KT, Lander ES, Gifford DK, Fraenkel E & Young RA, *Transcriptional regulatory code of a eukaryotic genome*. Nature, 2004. **431**.

HealthSquare.com, *Ovarian Cancer*. 2009 [cited May 07, 2009]; Available from: http://www.healthsquare.com/fgwh/wh1ch40.htm.

Helm CW & States JC, *Enhancing the efficacy of cisplatin in ovarian cancer treatment - could arsenic have a role*. Journal of Ovarian Research, 2009. **2**(2).

Hermsen R, Tans S & Wolde PR, *Transcriptional regulation by competing transcription factor modules*. PLOS Computational Biology, 2006. **2**(164).

Hertel JK, Johansson S, Raeder H, Midthjell K, Lyssenko V, Groop L, Molven A & Njolstad PR, *Genetic analysis of recently identified type 2 diabetes loci in 1,638 unselected patients with type 2 diabetes and 1,858 control participants from a Norwegian population-based cohort (the HUNT study)*. Diabetologia, 2008. **51**: 971-977.

Hestand MS, Galen MV, Villerius MP, Ommen G-JBV, Dunnen JTD & Hoen PAC’t, *CORE_TF: a user-friendly interface to identify evolutionary conserved transcription factor binding sites in sets of co-regulated gene*. BMC Bioinformatics, 2008. **9**(495).

Hey J & Machado CA, *The study of structured populations - new hope for a difficult and divided science*. Nature Reviews Genetics, 2003. **4**(535-543).

Hobert O, *Gene Regulation by Transcription Factors and MicroRNAs*. Science, 2008. **319(**1785-1786).

Hoffstedt J, Ryden M, Wahrenberg H, Harmelen VV & Arner P, *Upstream Transcription Factor-1 Gene Polymorphism Is Associated with Increased Adipocyte Lipolysis*. The Journal of Clinical Endocrinology & Metabolism, 2005. **90**(5356-5360).

Horie N & Takeishi K, *Identification of Functional Elements in the Promoter Region of the Human Gene for Thymidylate Synthase and Nuclear Factors That Regulate the Expression of the Gene.* The Journal of Biological Chemistry, 1997. **272**(29).

Hunninghake GW, Monick MM, Liu B & Stinski MF, *The promoter-regulatory region of the major immediate-early gene of human cytomegalovirus responds to T-lymphocyte stimulation and contains functional cyclic AMP-response elements.* Journal of Virology, 1989. **63**(7).

Janga SC, *Prediction and evolution of transcription factors and their evolutionary families in prokaryotes*. BMC Systems Biology, 2007. **1**(1).

Jirtle RL & Skinner MK, *Environmental epigenomics and disease susceptibility.* Nature Genetics, 2007. **8**(253-262).

Johnson DG & Degregori J. *Putting the Oncogenic and Tumor Suppressive Activities of E2F into Context.* Current Molecular Medicine, 2006. **6**(7).

Jorge A, Antonio S, Christopher P & Angel C, *SPSmart: adapting population based SNP genotype databases for fast and comprehensive web access.* BMC Bioinformatics, 2008. **9**(428).

Karchin R, *Next generation tools for the annotation of human SNPs.* Briefings in Bioinformatics, 2008. **10**(1).

Kaur M, Radovanovic A, Essack M, Schaefer U, Maqungo M, Kibler T, Schmeier S, Christoffels A, Narasimhan K, Choolani M & Bajic VB, *Database for exploration of functional context of genes implicated in ovarian cancer*. Nucleic Acids Research, 2008. **37**(D820-D823).

Keicho N, Tokunaga K, Nakata K, Taguchi Y, Azuma A, Bannai M, Emi M, Ohishi N, Yazaki Y & Kudoh S, *Contribution of HLA Genes to Genetic Predisposition in Diffuse Panbronchiolitis*. American Journal of Respiratory and Critical Care Medicine, 1998. **158**(3).

Kel AE, Gößling E, Reuter I, Cheremushkin E, Kel-Margoulis OV & Wingender E, *MATCH^{TM}: a tool for searching transcription factor binding sites in DNA sequences*. Nucleic Acids Research, 2003. **31**(13).

Kielbasa SM, Gonze D & Herzel H, *Measuring similarities between transcription factor binding sites*. BMC Bioinformatics, 2005. **6**(237).

Kim B-C, Kim W-Y, Park D, Chung W-H, Shin K-S & Bhak J, *SNP@Promoter: a database of human SNPs (Single Nucleotide Polymorphisms) within the putative promoter regions*. BMC Bioinformatics, 2008. **9**(1).

King M-C, Marks JH & Mandell JB, *Breast and Ovarian Cancer Risks Due to Inherited Mutations in BRCA1 and BRCA2*. Science, 2003a. **302**(643).

King OD & Roth FP, *A non-parametric model for transcription factor binding sites*. Nucleic Acids Research, 2003b. **31**(19).

Korean Bioinformation Center. *SNP@Promoter.* 2007 [cited August 14, 2008]; Available from: http://variome.kobic.re.kr/SNPatPromoter/.

Lambert JC, Goumidi L, Vrièze FW, Frigard B, Harris JM, Cummings A, Coates J, Pasquier F, Cottel D, Gaillac M, St Clair D, Mann DM, Hardy J, Lendon CL, Amouyel P & Chartier-Harlin MC, *The transcriptional factor LBP-1c/CP2/LSF gene on chromosome 12 is a genetic determinant of Alzheimer's disease.* Human Molecular Genetics, 2000. **9**(15).

Langsenlehner T, Langsenlehner U, Renner W, Kapp KS, Krippl P, Hofmann G, Clar H, Pummer K & Mayer R, *The Glu228Ala polymorhism in the ligand binding domain of death receptor 4 is associated with increased risk for prostate cancer metastases.* Prostate, **68**(3).

Lee PH & Shatkay H, F-SNP: computationally predicted functional SNPs for disease association studies. Nucleic Acids Research, 2007. **1**(5).

Liu J, Barnett A, Neufeld EJ & Dudley JP, *Homeoproteins CDP and SATB1 interact: potential for tissue-specific regulation.* Molecular and Cellular Biology, 1999. **19**(7).

Lowe WL, *Genetics of diabetes mellitus.* Springer, 2001. ISBN: 0792372522, 9780792372523.

Luikart G, England PR, Tallmon D, Jordan S & Taberlet P, *The power and promise of population genomics: from genotyping to genome typing.* Nature Genetics, 2003. **4**(981-994).

Lusis AJ, Attie AD & Reue K, *Metabolic syndrome: from epidemiology to systems biology*. Nature Reviews Genetics, 2008. **9**(819-830).

Malin C A, Engstrom P G, Lithwick S, Arenillas D, Eriksson P, Lenhard B, Wasserman W W & Odeberg J, *In Silico Detection of Sequence Variations Modifying Transcriptional Regulation*. Plos Computational Biology, 2008. **4**(1)

Manolio TA, Brooks LD & Collins FA, *A HapMap harvest of insights into the genetics of common disease*. The Journal of Clinical Investigation, 2008. **118**(5).

Matys V, Fricke E, Geffers R, Gößling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos D-U, Land S, Lewicki-Potapov B, Michael H, Münch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S & Wingender E, *TRANSFAC®: transcriptional regulation, from patterns to profiles*. Nucleic Acids Research, 2003. **31**(1).

Mayeux R, *Biomarkers: potential uses and limitations*. The Journal of the American Society for Experimental NeuroTherapeutics, 2004. **1**(2).

Maytin EV, Lin JC, Krishnamurthy R, Batchvarova N, Ron D, Mitchell PJ & Habener JF, *Keratin 10 gene expression during differentiation of mouse epidermis requires transcription factors C/EBP and AP-2*. Developmental Biology, 1999. **216**(1).

McCann JA, Muro EM, Palmer C, Palidwor G, Porter CJ, Andrade-Navarro MA & Rudnicki MA, *ChIP on SNP-chip for genome-wide analysis of human histone H4 hyperacetylation.* BMC Genomics, 2007. **8**(322).

McIntosh MW, Drescher C, Karlan B, Scholler N, Urban N, Hellstrom KE & Hellstrom I, *Combining CA 125 and SMR serum markers for diagnosis and early detection of ovarian carcinoma.* Gynecologic Oncology, 2004. **95**(1).

McNutt MC, Tongbai R, Cui W, Collins I, Freebern WJ, Montano I, Haggerty CM, Chandramouli GVR & Gardner K, *Human promoter genomic composition demonstrates non-random groupings that reflect general cellular function*. BMC Bioinformatics, 2005. **6**(259).

MedicineNet.com. *Definition of RT-PCR*. 2009a [cited April 28, 2009]; Available from: http://www.medterms.com/script/main/art.asp?articlekey=22766.

MedicineNet.com. *Definition of Western Blot*. 2009b [cited April 28, 2009]; http://www.medterms.com/script/main/art.asp?articlekey=6016.

Molecular Station. *Western Blot*. 2009 [cited April 28, 2009]; http://www.molecularstation.com/protein/western-blot/.

Morin PA, Martien KK & Taylor BL, Assessing statistical power of SNPs for population structure and conservation studies. Molecular Ecology Resources, 2008. **9**(1).

National Cancer Institute. *Ovarian Cancer*. 2009 [cited April 10, 2009]; Available from: http://www.cancer.gov/cancertopics/types/ovarian/.

National Center for Biotechnology Information. *Entrez Gene*. 2009 [cited May 11, 2009]; Available from: http://www.ncbi.nlm.nih.gov/sites/entrez?Db=gene&Cmd=ShowDetailView&TermToSearch=3043&ordinalpos=1&itool=EntrezSystem2.PEntrez.Gene.Gene_ResultsPanel.Gene_RVDocSum.

National Center for Biotechnology Information. *RELEASE: NCBI dbSNP Build 130*. 2006 [cited May 11, 2009]; Available from: http://www.ncbi.nlm.nih.gov/projects/SNP/snp_summary.cgi.

National Center for Biotechnology Information. *Single Nucleotide Polymorphism*. 2008 [cited April 28, 2009]; Available from: http://www.ncbi.nlm.nih.gov/projects/SNP/dbSNP.cgi?list=rslist.

National Ovarian Cancer Coalition, *About Ovarian Cancer*. 2009 [cited May 07, 2009]; Available from: http://www.dfwovarian.org/gillette.html.

Nath J & Johnson KL, *A review of fluorescence in situ hybridization (FISH): current status and future prospects*. Biotechnic and Histochemistry, 2000. **75**(2).

Naylor S, *Biomarkers: current perspectives and future prospects*. Expert Review of Molecular Diagnostics, 2003. **3**(5).

Nielsen, R. *Human genomics: disclosure of variation*. Nature, 2005. **434**(288–289).

Nomikos A. *Single Nucleotide Polymorphisms and Linkage Disequilibrium Mapping*. 2006 [cited May 12, 2009]; Available from: http://www.dartmouth.edu/~brenner/gene144-06/nomikos.html.

OncologyChannel, *Ovarian Cancer*. 2009 [cited May 06, 2009]; Available from: http://www.oncologychannel.com/ovariancancer/types.shtml.

Pavesi G, Mauri G & Pesole G, *In silico representation and discovery of transcription factor binding sites*. Briefings in Bioinformatics, 2004. **5**(3).

Peltonen L, Palotie A & Lange K, *Use of population isolates for mapping complex traits*. Nature Reviews, 2000. **1**(182-190).

Pénzes M, Rajczya K, Gyódia E, Rétib M, Fehéra É & Petrányia G, HLA-G gene polymorphism in the normal population and in recurrent spontaneous abortion in Hungary. Transplantation Proceedings, 1999. **31**(4).

Pritchard CC, Hsu L, Delrow J & Nelson PS, *Project normal: defining normal variance in mouse gene expression*. Proceedings of the National Academy of Sciences of the United States of America, 2001. **98**(23).

Prokunina L & Alarcón-Riquelme ME, *Regulatory SNPs in complex diseases: their identification and functional validation*. Expert Reviews in Molecular Medicine, 2004. **6**(1-15).

Python, *Python Programming Language – Official Website*. 2008 [cited October 16, 2008]; Available from: http://www.python.org/.

Qiu P, *Recent advances in computational promoter analysis in understanding the transcriptional regulatory network.* Biochemical and Biophysical Research Communications, 2003. **309**(495-501).

Queen's University. *F-SNP: a collection of functional SNPs specifically prioritized for disease association studies.* 2007 [cited February 27, 2009]; Available from: http://compbio.cs.queensu.ca/F-SNP/.

Ramaswamy SV, Reich R, Dou S-J, Jasperse L, Pan X, Wanger A, Quitugua T & Graviss EA, *Single Nucleotide Polymorphisms in genes Associated with Isoniazid Resistance in Mycobacterium tuberculosis.* Antimicrobial Agents and Chemotherapy, 2003. **47**(4).

Reich DE, Gabriel SB & Altshuler D, *Quality and completeness of SNP databases.* Nature Genetics, 2003. **33**(457-458).

Reimer D, Sadr S, Wiedemair A, Concin N, Hofstetter G, Marth C & Zeimet AG, *Heterogenous cross-talk of E2F familiy members is crucially involved in growth modulatory effects of interferon-gamma and EGF.* Cancer Biology & Therapy, 2006. **5**(7).

Reumers J, Conde L, Medina I, Maurer-Stroh S, Durme JV, Dopazo J, Rousseau F & Schymkowitz J, *Joint annotation of coding and non-coding single nucleotide polymorphisms and mutations in the SNPeffect and PupaSuite databases.* Nucleic Acids Research, 2007. **36**(D825-D829).

Ridder DA, Bulashevska S, Chaitanya GV, Babu PP, Brors B, Eils R, Schneider A & Schwaningera M, *Discovery of transcriptional programs in cerebral ischemia by in silico promoter analysis.* Brain Research, 2009. PubMed ID: 19344698.

Roberts J A, *Searching for a Biomarker for Ovarian Cancer.* Journal of the American Medical Association, 1998. **280**(8).

Robinson-Rechavi M, Escriva Garcia H & Laudet V, *The nuclear receptor superfamily*. Journal of Cell Science, 2003. **116**.

Roche, *SNPs: The great importance of small differences*. 2008 [cited September 12, 2008]; Available from: www.roche.com/pages/facets/22/snps_e.pdf.

Rubinstein WS & Weissman SM, *Managing hereditary gastrointestinal cancer syndromes: the partnership between genetic counselors and gastroenterologists.* Nature Clinical Practice Gastroenterology & Hepatology, 2008. **5**(569-582).

Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, Hunt SE, Cole CG, Coggill PC, Rice CM, Ning Z, Rogers J, *et al.*, *A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms.* Nature, 2001. **409**.

Sakazume S, Sorokina E, Iwamoto Y & Semina EV, *Functional analysis of human mutations in homeodomain transcription factor PITX3.* BMC Molecular Biology, 2007. **8**(84).

Sandelin A, Alkema W, Engström P, Wasserman WW & Lenhard B, *JASPAR: an open-access database for eukaryotic transcription factor binding profiles.* Nucleic Acids Research, 2004. **32**(D91-D94).

Satoh A, Toyota M, Ikeda H, Morimoto Y, Akino K, Mita K, Suzuki H, Sasaki Y, Kanaseki T, Takamura Y, Soejima H, Urano T, Yanagihara K, Endo T, Hinoda Y, Fujita M, Hosokawa M, Sato N, Tokino T & Imai K, *Epigenetic inactivation of class II transactivator (CIITA) is associated with the absence of interferon-gamma-induced HLA-DR expression in colorectal and gastric cancer cells.* Oncogene, 2004. **23**(55).

Savon SP, Hakimi P, Crawford DR & Klemm DJ, *The Promoter Regulatory Regions of the Genes for the Cytosolic Form of Phosphoenolpyruvate Carboxykinase (GTP) from the Chicken and the Rat Have Different Species-Specific Roles in Gluconeogenesis*. The Journal of Nutrition, 1997. **127**(2).

Schaefer U. *PROMEX: Dragon promoter extraction tool*. 2009 [cited April 28, 2009]; Available from: http://tr.sanbi.ac.za/~ulf/promex/.

Sherry ST, Ward M-H, Kholodov M, Baker J, Phan L, Smigielski EM & Sirotkin K, *dbSNP: the NCBI database of genetic variation*. Nucleic Acids Research, 2000. **29**(1).

Smith AV, *Generating HapMap Data Text Reports Using the Genome Browser*. Cold Spring Harbor Protocols, 2008a. doi:10.1101/pdb.prot5024.

Smith AV, *Browsing HapMap Data Using the Genome Browser*. Cold Spring Harbor Protocols, 2008b. doi:10.1101/pdb.prot5023. **98**.

Smyth GK, Yang YH & Speed T, *Statistical Issues in cDNA Microarray Data Analysis*. Methods in Molecular Biology, 2003. **224**(111-36).

SNP@Promoter. *SNP@Promoter Method*. 2007 [cited September 6, 2008]; Available from: http://variome.kobic.re.kr/SNPatPromoter/method.jsp.

South African National Bioinformatics Institute & OrionCell. *Dragon Database for Exploration of Ovarian Cancer Genes*. 2009 [cited March 25, 2009]; Available from: http://apps.sanbi.ac.za/ddoc/index.php.

Steele GD, Jr Osteen RT, Winchester DP, Murphy GP & Menck HR, *Clinical highlights from the National Cancer Data Base: 1994*. A Cancer Journal for Clinicians, 1994. **44**(71-80).

Stepanova M, Tiazhelova T, Skoblov M & Baranova A, *Potential regulatory SNPs in promoters of human genes: A systematic approach.* Molecular and Cellular Probes, 2006. **20** (348-358).

Stormo GD, *DNA binding sites: representation and discovery*. Bioinformatics, 2000. **16**(1).

Su G, Mao B & Wang J, *A web server for transcription factor binding site prediction.* Bioinformatician, 2006. **1**(5).

Sugiyama T, Uchida C, Oda T, Kitagawa M, Hayashi H & Ichiyama A, *Involvement of CCAAT/enhancer-binding protein in regulation of the rat serine:pyruvate/alanine:glyoxylate aminotransferase gene expression.* FEBS Letters, 2001. **508**(1).

Tabor HK, Risch NJ & Myers RM, *Candidate-gene approaches for studying complex genetic traits: practical considerations*. Nature Reviews, 2002. **3**(1).

Tebbutt SJ, James A & Paré PD, *Single-Nucleotide Polymorphisms and Lung Disease: Clinical Implications*. Chest, 2007. **131**(1216-1223).

The New York Times, *Health Guide*. 2009 [cited April 20, 2009]; Available from: http://health.nytimes.com/health/guides/disease/cancer/overview.html.

Thomas PD & Kejariwal A, *Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: evolutionary evidence for differences in molecular effects*. Proceedings of the National Academy of Sciences, 2004. **101**(43).

Tian J, Wu N, Guo X, Zhang J & Fan Y, *Predicting the phenotypic effects of non-synonymous single nucleotide polymorphisms based on support vector machines.* BMC Bioinformatics, 2007. **8**(450).

Tishkoff SA & Williams SM, *Genetic analysis of African populations: human evolution and complex disease. Nature Genetics*, 2002. **3**(611-621).

U.S. National Institutes of Health. *National Cancer Institute*. 2001 [cited June 20, 2008]; Available from: http://www.cancer.gov/.

van der Stoep N, Biesta P, Quinten E & van der Elsen PJ, *Lack of IFN-gamma-mediated induction of the class II transactivator (CIITA) through promoter methylation is predominantly found in developmental tumor cell lines*, International Journal of Cancer, 2002. **97**(4).

Vaquerizas JM, Kummerfeld SK, Teichmann SA & Luscombe NM, *A census of human transcription factors: function, expression and evolution*. Nature Genetics, 2009. **10**(252-263).

Vuillez JP, Levrot E, Mousseau M, Buffaz PD, Bolla M, Payan R, Comet M & Schaerer R, *Evaluation of the diagnostic usefulness of CA125 immunoscintigraphy for ovarian carcinoma follow-up after treatment: contribution of this technique in Grenoble University Medical Center*. Bulletin Du Cancer, 1997. **84**(11).

Wall L, Christiansen T & Orwant J, *Programming Perl (3rd Edition)*. O'Reilly Media, Incorporated, 2000. 3rd edition. ISBN-10: 0596000278.

Wang E, Lenferink A & O’Connor-McCourt M, *Cancer systems biology: exploring cancer-associated genes on cellular networks*. Cellular and Molecular Life Sciences, 2007, **64**(1752-1762).

Wasserman WW & Sandelin A, *Applied Bioinformatics for the Identification of Regulatory Elements*. Nature Reviews, 2004. **5**.

Wray GA, *Evolutionary Dissociations between Homologous Genes and Homologous Structures*. Novartis Foundation Symposia, 2007. ISBN: 9780471984931.

Xie Q, Ratnasinghe LD, Hong H, Perkins R, Tang Z-Z, Hu N, Taylor PR & Tong W, *Decision Forest Analysis of 61 Single Nucleotide Polymorphisms in a Case-Control Study of Esophageal Cancer; a novel method*. BMC Bioinformatics, 2005. **6**(2).

Xu H, Gregory SG, Hauser ER, Stenger JE, Pericak-Vance MA, Vance JM, Züchner S & Hauser MA, *SNPselector: a web tool for selecting SNPs for genetic association studies*. Bioinformatics, 2005. **21**(22).

Yang E, Simcha D, Almon RR, Dubois DC, Jusko WJ & Androulakis IP, *Context Specific Transcription Factor Prediction*. Annals of Biomedical engineering, 2007. **35**(6).

Yang H, Keane J, Bergman CM, Nenadic G, *Assigning roles to protein mentions: The case of transcription factors.* Journal of Biomedical Informatics, 2009. PubMed ID: 19364541.

Yang TT & Chow CW, *Transcription cooperation by NFAT.C/EBP composite enhancer complex.* The Journal of Biological Chemistry, 2003. **278**(18).

Yang VW, *Eukaryotic Transcription Factors: Identification, Characterization and Functions*. The Journal of Nutrition, 1998. **128**(11).

Appendix I

A. Ovarian Cancer Candidate Gene Data Set

HUGO Gene Symbols								
ABCB1	*BCL2L1*	*CDKN1C*	*DLEC1*	*FLT1*	*KLF2*	*MDM2*	*PCSK6*	*SPINK1*
ABCB11	*BCPR*	*CDKN2A*	*DLX4*	*FN1*	*KLF6*	*MEN1*	*PDGFRA*	*SPINT2*
ABCG2	*BECN1*	*CDKN2B*	*DNAJC15*	*FOLR1*	*KLK10*	*MKI67*	*PEBP4*	*SR-A1*
ACHE	*BIRC2*	*CDKN2C*	*DNMT1*	*FOXP3*	*KLK11*	*MLH1*	*PIK3CA*	*SST*
ACVR1B	*BIRC4*	*CDKN2D*	*DNMT3B*	*FRA9E*	*KLK13*	*MLLT4*	*PLA2G4C*	*ST8*
ACVR1C	*BMP2*	*CDR1*	*DPH1*	*FRAP1*	*KLK14*	*MMP1*	*PLAGL1*	*STAT3*
ACVR2A	*BMP4*	*CDR2*	*DPYD*	*FSHR*	*KLK15*	*MMP14*	*PLAT*	*STEAP1*
ACVR2B	*BRAF*	*CDX2*	*DUSP3*	*GADD45A*	*KLK3*	*MMP2*	*PLAU*	*STK11*
ADM	*BRCA1*	*CFLAR*	*E2F1*	*GALT*	*KLK4*	*MMP26*	*PLAUR*	*TACSTD1*
AES	*BRCA2*	*CGB5*	*E2F2*	*GAS6*	*KLK5*	*MMP3*	*PPAP2A*	*TBX3*
AGPAT2	*BRIP1*	*CHEK2*	*E2F3*	*GJA1*	*KLK6*	*MMP7*	*PPARG*	*TERC*
AGTR1	*BRMS1*	*CIITA*	*E2F4*	*GJB2*	*KLK7*	*MMP8*	*PPP2R4*	*TERT*
AKAP13	*C11orf30*	*CLDN1*	*E2F5*	*GNRH1*	*KLK8*	*MMP9*	*PPP2R5D*	*TFAP2A*
AKT1	*C1orf38*	*CLDN3*	*EBAG9*	*GPC1*	*KLK9*	*MPG*	*PRKACA*	*THBS3*
AKT2	*CALCA*	*CLDN4*	*ECGF1*	*GPLD1*	*KRAS*	*MPO*	*PRKCI*	*TMSB10*
AMH	*CAMK4*	*CLIP1*	*EDG4*	*GPR68*	*KRAS1P*	*MS4A1*	*PSD3*	*TNF*
AMPH	*CASP3*	*COL18A1*	*EDG7*	*GRN*	*LAMP1*	*MSH2*	*PTEN*	*TNFRSF10A*
APOA1	*CAV1*	*COMT*	*EDN1*	*GSK3B*	*LASP1*	*MSH3*	*PTGS1*	*TNFRSF10B*
APOE	*CBR3*	*CSAG2*	*EDNRA*	*GSTM1*	*LATS1*	*MSH6*	*PTK2*	*TNFSF10*

Continued…

HUGO Gene Symbols								
AR	*CCND1*	*CSF1*	*EEF1A2*	*GSTO2*	*LCFS2*	*MSLN*	*PTP4A3*	*TOC*
ARID4B	*CD24*	*CSF1R*	*EGF*	*GSTP1*	*LGALS1*	*MSN*	*PYGO2*	*TOP2A*
ARMCX1	*CD34*	*CSF2*	*EGFR*	*HBB*	*LOH11CR2A*	*MSR1*	*RAB25*	*TP53*
ARMCX2	*CD40*	*CSF3*	*EIF5A2*	*HGF*	*LTBP1*	*MSX1*	*RAF1*	*TP73L*
ARMCX3	*CD44*	*CSF3R*	*EPHA1*	*HIF1A*	*LUZP4*	*MTHFR*	*RASSF1*	*TTR*
ATF3	*CD46*	*CSK*	*EPHA2*	*HLA-DQA1*	*LZTS1*	*MUC1*	*RBP1*	*TUBB3*
ATM	*CD47*	*CTAG1B*	*EPHA5*	*HLA-DRB1*	*MAD2L1*	*MUC16*	*RNASE2*	*TUSC3*
ATP7A	*CD63*	*CTAG2*	*EPHB2*	*HMGA1*	*MAGEA1*	*MUC2*	*RNASET2*	*TYMS*
ATP7B	*CD80*	*CTNNA1*	*EPHB4*	*HOXA9*	*MAGEA4*	*MUC3A*	*ROCK1*	*UTRN*
ATR	*CD82*	*CTNNB1*	*EPHX1*	*HRAS*	*MAP2K1*	*MUC4*	*RPS6KB1*	*VEGFA*
AURKA	*CD86*	*CTSB*	*EPO*	*HSPB1*	*MAP2K3*	*MUC5AC*	*RSF1*	*VTCN1*
AXIN2	*CD9*	*CTSD*	*EPOR*	*HTRA1*	*MAP2K4*	*MVP*	*SDC1*	*WNT2B*
B2M	*CD99*	*CTSL*	*ERBB2*	*IGF2*	*MAP2K7*	*MYC*	*SELENBP1*	*WWOX*
BACH1	*CDC20*	*CXADR*	*ERBB3*	*IGFBP3*	*MAP3K1*	*MYCL1*	*SEMA3B*	*XPA*
BACH2	*CDC25A*	*CXCL1*	*ERBB4*	*IL13RA2*	*MAP3K2*	*MYO18B*	*SERPINB5*	*XRCC1*
BAD	*CDC25B*	*CXCL12*	*ERCC1*	*IL18*	*MAP3K3*	*NBN*	*SERPINE1*	*XRCC2*
BAG1	*CDC25C*	*CXCR4*	*ERCC2*	*IL6*	*MAP3K4*	*NBR1*	*SERPINF1*	
BAGE	*CDC42*	*CYP1B1*	*ERCC3*	*ILK*	*MAP3K5*	*NCOA3*	*SHMT1*	
BAK1	*CDH1*	*CYP3A4*	*ERCC5*	*INSR*	*MAP3K8*	*NEO1*	*SLC2A1*	
BARD1	*CDH13*	*DAB2*	*ETS1*	*ITGB3*	*MAPK1*	*NTRK2*	*SMAD4*	
BARX2	*CDK2*	*DCC*	*FASLG*	*KDR*	*MAPK3*	*OPCML*	*SNCG*	
BAX	*CDK4*	*DDR1*	*FASN*	*KISS1*	*MAPK8*	*OVCA2*	*SOD2*	
BCHE	*CDKN1A*	*DIRAS3*	*FBXW7*	*KIT*	*MCAM*	*PARK2*	*SPARC*	
BCL2	*CDKN1B*	*DLC1*	*FILIP1L*	*KITLG*	*MCC*	*PAX8*	*SPDEF*	

B. html_read.py

This program reads through multiple HTML files stored in a single working directory and extracts SNP results that are displayed in table format within each HTML webpage. Using the python table_parser module, this program was designed to extract the RefSNP ID, chromosomal location, strand orientation (i.e. forward or reverse) and nucleotide base position of each SNP result per gene and write these results into a single comma delimited output file (Figure 3.8, File 3).

```
#!/usr/bin/env python

import csv, os
from table_parser import *

#----------Accesses folder that stores all SNP@Promoter result files----------#
path = "/Users/kavisharamdayal/SNP@Promoter_Results"
snp_prom = os.listdir(path)
writer = csv.writer(file('SNP@Promoter.csv', 'w'))
#----------Selects gene symbol from SNP@Promoter result file name----------#
def get_genename(f):
    gene = str(f).split("\"")[1].split(".")[0]
    return gene

def get_info(line):
    strand = ""
    if line[1].startswith("SNP"):
        return ""
    elif line[1].startswith("Allele"):
        return ""
    elif line[1].startswith("NM"):
        return ""
    else:
#----------Return SNP info----------#
        snp = 'rs'+str(line[1])
        chr = line[2].split(":")[0].split("r")[1]
        pos = line[2].split("(")[1].split(",")[0]
        if len(line) >= 10:
            strand = line[4]
        else:
            strand = " "
    return snp, chr, strand, pos
#----------Reads through all files in directory----------#
for file in snp_prom:
    if file.endswith(".html"):
        f = open(file,"rb")
#----------Extract and read tables from HTML file----------#
        p = TableParser()
        p.feed(f.read())
#----------Reads through first table in HTML file----------#
        for line in p.doc[0]:
#----------Writes column containing RefSNPs if not a header----------#
            if get_info(line) != "":
                a,b,c,d = get_info(line)
                writer.writerow([get_genename(f),a,b,c,d])
f.close()
```

C. pupasuite_read.py

This program reads through Transfac and Jaspar result files (flat files) obtained from the completion of a batch query of candidate genes through the PupaSuite SNP annotation tool (Section 3.2.2.3). The program then reads through SNP results per gene, converting the Entrez gene IDs into HUGO gene symbols (for consistency & comparability). All SNP results (Transfac and Jaspar) per gene were then written to a single comma delimited output file (Figure 3.8, File 4).

```
#!/usr/bin/python

import csv

f = open("transfac.txt","rb")
g = open("jaspar.txt","rb")
h = open("IDconverterResults73486.csv","rb")
writer = csv.writer(file('PupaSuite.csv', 'w'))
ens = ""
dict = {}

#----------Convert Ensembl IDs to Gene Symbol----------#
def get_genename():
    for l in h.readlines():
        temp = l.rstrip().split(",")[1].rstrip()
        dict[temp] = l.rstrip().split(",")[0]
    return dict

#----------Compile Transfac Results----------#
for j in f.readlines():
    if j.startswith("rs"):
        snp = j.rstrip().split()[0]
        ens = j.rstrip().split()[1]
        gene = get_genename()[ens]
        writer.writerow([gene,snp])

#----------Compile Jaspar Results----------#
for k in g.readlines():
    if k.startswith("rs"):
        snp = k.rstrip().split()[0]
        ens = k.rstrip().split()[1]
        gene = get_genename()[ens]
        writer.writerow([gene,snp])

f.close()
g.close()
h.close()
```

D. all_snps.py

This program reads through the three result files obtained from each SNP annotation tool (Figure 3.8, Files 2, 3 & 4) and compiles all SNP results into a single comma delimited file, while collating SNP results in order of "HUGO gene symbol", "RefSNP ID", "Chromosomal location", "Strand orientation" and "Nucleotide base position".

```
#!/usr/bin/python

import csv

f = open("F-SNP_Results.csv","rb")
g = open("SNP@Promoter_Results.csv","rb")
h = open("PupaSuite_Results.csv","rb")
writer = csv.writer(file('All_SNPs.csv', 'w'))
count = 0

#----------Compile F-SNP Results----------#
for line in f.readlines():
    gene = line.rstrip().split(",")[0]
    snp = line.rstrip().split(",")[1]
    chr = line.rstrip().split(",")[2]
    strand = line.rstrip().split(",")[3]
    pos = line.rstrip().split(",")[4]
    writer.writerow([gene,snp,chr,strand,pos])

#----------Compile SNP@Promoter Results----------#
for line in g.readlines():
    gene = line.rstrip().split(",")[0]
    snp = line.rstrip().split(",")[1]
    chr = line.rstrip().split(",")[2]
    strand = line.rstrip().split(",")[3]
    pos = line.rstrip().split(",")[4]
    writer.writerow([gene,snp,chr,strand,pos])

#----------Crosscheck PupaSuite Results----------#
for line in h.readlines():
    gene = line.rstrip().split(",")[0]
    snp = line.rstrip().split(",")[1]
    i =open("SNP_Ref.csv","rb")
    for k in i.readlines():
        gn = k.rstrip().split(",")[0]
        sn = k.rstrip().split(",")[1]
        if gene==gn and snp==sn:
            print "Match",count,":",gn,sn
            count +=1
        # Total confirmed SNPs= 42 {Total PupaSuite SNPs = 90}

f.close()
g.close()
h.close()
i.close()
```

E. verify_snps.py

This program reads through the accumulated SNP results (Figure 3.8, File 5) obtained from the custom program all_snps.py (Appendix I-D) and the result flat file from querying all RefSNP IDs listed in File 5 (Figure 3.8) through the dbSNP database. The program then searches for SNPs that are listed in both files. All SNPs from the accumulated SNP results (Figure 3.8, File 5) that were present in the dbSNP result file (i.e. present in dbSNP Build 129) were classified as verified SNPs and output to a single comma delimited file (Figure 3.8, File 6).

```
#!/usr/bin/python

import csv
dbSNPs = []
count1 = 0
count2 = 0

f = open("All_SNPs.csv","rb")
g = open("dbSNP_BATCH_Results.txt","rb")
writer = csv.writer(file('SNP_Ref.csv', 'w'))

#----------Create dbSNP list of query SNPs----------#
for line in g.readlines():
    if line.startswith("rs"):
        dbSNPs.append(line.rstrip().split()[0])

#----------Verify all predicted SNPs & create SNP reference table----------#
for line in f.readlines():
    if line.rstrip().split(",")[1] in dbSNPs:
        writer.writerow([line.rstrip()])
        count1 += 1
    else:
        count2 += 1

# print "Verified:",count1 = 10640
# print "Unverified:",count2 = 133

f.close()
g.close()
```

F. label_promoters.py

This program was created for the relabeling of all promoter sequences that were retrieved by the Promex tool. Since Entrez IDs of all candidate genes were queried through the Promex tool, the resulting promoter regions were labeled according to these Entrez IDs instead of the HUGO gene symbols, as with all other results in this study. This program read through the Promex result file (Figure 3.8, File 7) and renamed the promoter sequences in the following order: “HUGO gene symbol”, “Strand orientation”, “Chromosomal location”, “Nucleotide base start position” and “Nucleotide base stop position”.

```
#!/usr/bin/python

import csv

ids = {}
f = open("Gene_Entrez_IDS.csv","rb")
g = open("Promex_Promoter_Regions[-2000,+500].txt","rb")

#----------Extract & store Entrez IDs in directory----------#
for line in f.readlines():
    ids[line.split(",")[1].rstrip()] = line.split(",")[0]

#-------Convert Entrez IDs to Gene Symbols & compile sequence details-------#
for line in g.readlines():
    if line.startswith(">"):
        entrez_id = line.split(",")[0].split()[2]
        chr = line.split(",")[3].split()[1]
        strand = line.split(",")[4].split()[1].rstrip()
        seq_start = line.split(",")[6].split()[1]
        seq_stop = line.split(",")[7].split()[1]
        new_label =
str(">"+ids[entrez_id])+str(":"+strand)+str(":"+chr)+str(":START,"+seq_start)+str(":STOP,"
+seq_stop)
        print new_label
    else:
        print line.rstrip()

f.close()
g.close()
```

G. match_read.py

This program reads through the results obtained from the MATCH^{TM} tool and creates a TFBS reference table containing the following fields of information: “HUGO gene symbol”, “Strand orientation”, “Chromosomal location”, “TFBS Start & Stop nucleotide base positions”, “Transcription factor”, “Core Similarity Score”, “Matrix Similarity Score”, “Transcription factor nucleotide sequence” and “Matrix identifier”. The script identifies TFBS nucleotide base positions per promoter sequence based on the length of the transcription factor identified and the start and end positions (bp) of the promoter sequence that was queried, e.g. if a transcription factor of length 8bp was identified on a specific promoter sequence, the TFBS range was calculated as follows:

For transcription factors (TF) predicted on the forward strand:

> *TFBS Start (bp) = Promoter start position (bp) (obtained from promoter sequence label (Figure 3.8, File 7)) + TF (MATCH^{TM}) nucleotide base position*
>
> *TFBS Stop (bp) = Promoter start position (bp) (obtained from promoter sequence label (Figure 3.8, File 7)) + length of TF (bp) - 1*

For transcription factors (TF) predicted on the reverse strand:

> *TFBS Start (bp) = Promoter end position (bp) (obtained from promoter sequence label (Figure 3.8, File 7)) - TF (MATCH^{TM}) nucleotide base position*
>
> *TFBS Stop (bp) = Promoter end position (bp) (obtained from promoter sequence label (Figure 3.8, File 7)) - length of TF (bp) - 1*

```
#!/usr/bin/python

import csv

f = open("MATCH_Promex(-2000+500)_Results.csv","rb")
writer = csv.writer(file('TFBS_Ref.csv', 'w'))
tf = ""
gene = ""
chr = ""

#----------Filter through MATCH™ results----------#
for l in f.readlines():
    line = l.split()
    temp = line
    for j in temp:

#----------Identify results for each promoter region----------#
        if "Scanning" in j:
            gene = temp[3].split(":")[0]
            strand = temp[3].split(":")[1]
            chr = temp[3].split(":")[2]
            seq_start = int(temp[3].split(",")[1].split(":")[0])
            seq_stop = int(temp[3].split(",")[2][:-1])
```

```
#------Compile results for each transcription factor predicted on the fwd strand------#
        elif (j.startswith("V$")) and temp[2] == "(+)":
            tf = temp[6]
            fam = temp[0]
            tfbs_start = seq_start+int(temp[1])
            tfbs_stop = tfbs_start+int(len(temp[5]))
            tfbs = str(tfbs_start)+"-"+str(tfbs_stop)
            css = temp[3]
            mss = temp[4]
            tf_seq = temp[5]
            writer.writerow([gene,strand,chr,tfbs,tf,css,mss, tf_seq,fam])

#-------Compile results for each transcription factor predicted on the rev strand------#
        elif (j.startswith("V$")) and temp[2] == "(-)":
            tf = temp[6]
            fam = temp[0]
            tfbs_start = seq_stop-int(temp[1])
            tfbs_stop = tfbs_start-int(len(temp[5]))
            tfbs = str(tfbs_stop)+"-"+str(tfbs_start)
            css = temp[3]
            mss = temp[4]
            tf_seq = temp[5]
            writer.writerow([gene,strand,chr,tfbs,tf,css,mss, tf_seq,fam])

f.close()
```

H. overlaps.py

This program compares the SNP results stored in the SNP reference table (Figure 3.8, File 6) to the TFBS results stored in the TFBS reference table (Figure 3.8, File 9). The program first isolates the HUGO gene symbol, RefSNP ID, Strand orientation, Chromosomal location and Nucleotide base position of all entries within the SNP reference table and stores this information in a dictionary with the nucleotide base position as the key and remaing information as values. The program then isolates the HUGO gene symbol, Strand orientation, Chromosomal location, TFBS start position, TFBS end position, Transcription factor, Matrix identifier, Transcription factor nucleotide sequence, Core similarity score and Matrix similarity score for each entry within the TFBS reference table. For each SNP result in the dictionary created above, the script then checked (for each result in the TFBS reference table) if (1) the SNP was located at or between the TFBS start and end nucleotide base positions, (2) the strand orientations were the same, and (3) chromosomal locations were the same. All SNPs that satisfied these criteria were classified as SNPs coinciding with TFBSs and written to an output comma delimited file (Figure 3.8, File 10).

```
#!/usr/bin/python

import csv

sref = open("SNP_Ref.csv","rb")
tref = open("TFBS_Ref.csv","rb")
writer = csv.writer(file('SNPs_within_TFBSs.csv', 'w'))
dict = {}
genes = []
count = 0

#----------Open SNP reference table & store info in directory----------#
for j in sref.readlines():
    gene1 = j.rstrip().split(",")[0]
    snp = j.rstrip().split(",")[1]
    strand1 = j.rstrip().split(",")[3]
    chr1 = j.rstrip().split(",")[2]
    pos1 = int(j.rstrip().split(",")[4].split('\"')[0])
    dict[pos1]=([gene1,snp,strand1,chr1,pos1])
writer.writerow(["Gene","SNP","Strand","Chromosome","Position","TF","Matrix
Identifier","TF Sequence","CSS","MSS"])

#----------Open TFBS reference table & store info in temporary variables----------#
for k in tref.readlines():
    gene2 = k.rstrip().split(",")[0]
    strand2 = k.rstrip().split(",")[1]
    chr2 = k.rstrip().split(",")[2].split("r")[1]
    s_pos = int(k.rstrip().split(",")[3].split("-")[0])
    e_pos = int(k.rstrip().split(",")[3].split("-")[1])-1
    tf = k.rstrip().split(",")[4]
    css = k.rstrip().split(",")[5]
    mss = k.rstrip().split(",")[6]
    fam = k.rstrip().split(",")[8]
```

```
#----------Compare SNP & TFBS strand orientation, chromosome & nucleotide base
positions for coinciding pairs----------#
    for i in dict:
        strand1 = dict[i][2]
        chr1 = dict[i][3]
        if (i >= s_pos) and (i <= e_pos) and (strand1 == strand2) and (chr1 == chr2):
            if dict[i][0].split('\"')[1] not in genes:
                genes.append(dict[i][0].split('\"')[1])
writer.writerow([dict[i][0].split('\"')[1],dict[i][1],dict[i][2],dict[i][3],str(dict[i][4]),tf,fam,tf_seq,cs
s,mss])

sref.close()
tref.close()
```

I. SNPs coinciding with TFBSs

Gene	SNP	Strand	Chr	Position	TF	TF Sequence	CSS	MSS
B2M	rs17235815	+	15	42789242	AIRE	gtTTTTAaattggttttccaagtga	0.673	0.712
MAP2K1	rs35534818	+	15	64465511	Pax-4	AAAAAatattgccaacatggtgaaaacccg	1	0.846
NEO1	rs8025535	+	15	71129558	HIF1	gtatACGTGcaggc	1	0.979
CSK	rs7496625	+	15	72861395	TBX5	caccACACCtat	1	0.973
AKAP13	rs12437885	+	15	83722873	Tal-1beta:E47	tagtaCAGATggcgtt	1	0.922
PCSK6	rs1472303	-	15	99849029	myogenin	gaccgcttggggACGGCgggcggccgggg	0.637	0.705
AR	rs34566600	+	X	66678721	PLZF	gtggaagcaacaTAAACtttggagtcttt	0.976	0.802
ARMCX1	rs6621104	+	X	100690370	PLZF	ttgtatttttaaTAAAGatggcgttttac	1	0.819
ARMCX3	rs2858167	+	X	100764000	Hand1:E47	attgCCAGAcacactg	1	0.981
BIRC4	rs7064224	+	X	122820033	Pax	CAGGCactcac	0.74	0.861
FOXP3	rs35851078	-	X	49008459	CDP	CTATAcacttttgtt	0.98	0.826
MMP14	rs1957371	+	14	22373891	Pax-5	aacctgggcgacaGGGCGagactccgtc	0.873	0.753
HIF1A	rs4902079	+	14	61230050	Pax-5	tccgagtgtggtgGTGCGtgcctgtaat	0.839	0.767
MUC2	rs35123704	+	11	1063794	Pax	CAGGAactcaa	0.857	0.847
ILK	rs2659860	+	11	6581236	AP-2	gaggccgCAGGCg	1	0.98
ADM	rs5001	+	11	10282837	Pax-5	ggggctaggactctcCTTTGccccttga	0.965	0.839
CD44	rs7944409	+	11	35115929	Pax-5	tgcctcgtgcCGCTGagcctggcgcaga	0.919	0.744
CD82	rs16914075	+	11	44543302	myogenin	ttcagggagaaaGCCAGctttgagggctt	0.9	0.724
GSTP1	rs4147581	+	11	67108161	Pax-3	agtttcgcCGTGAccttctgc	1	0.862
CCND1	rs954619	+	11	69163915	CDP	cacgaacaccTATCG	0.998	0.948
BIRC2	rs5794168	+	11	101723404	RFX	ggGCAACtg	1	0.988
ATM	rs3205808	+	11	107599080	GATA-X	tgtgcTTATCa	1	0.991
LOH11CR2A	rs1939849	+	11	123489876	AIRE	tcATCTTtatcggtttaattgtgta	0.679	0.707

Continued…

Gene	SNP	Strand	Chr	Position	TF	TF Sequence	CSS	MSS
IGF2	rs1050197	-	11	2116105	MyoD	ctggggCAGGTggcggct	1	0.984
MEN1	rs650277	-	11	64334871	Pax-5	agcctgagtgacaGAGCGagactctgtc	0.973	0.819
MMP1	rs17882592	-	11	102175154	Pax-5	cccagagtcaCGCTCagtctctttccag	0.973	0.773
IL18	rs5744223	-	11	111541778	Pax-6	tccatctgattCTTAAaatat	0.792	0.743
APOA1	rs2727786	-	11	116213733	Pax-5	gcggggcgggacgGAGCGgggcggcctc	0.973	0.744
MCAM	rs3923594	-	11	118693125	Sp1	cggGGCGGgg	1	0.993
MSX1	rs13104352	+	4	4908791	WT1	gggGGAGGg	1	1
PDGFRA	rs1800812	+	4	54789386	GATA-4	AGATAgaagcca	1	0.943
KIT	rs6554199	+	4	55217245	CDP	cacagagaccTTTGG	0.745	0.808
CXCL1	rs6825295	+	4	74952346	c-Ets-1	ccttccTTCCGgactc	1	0.952
EDNRA	rs3190169	+	4	148621911	myogenin	atttaggtaagtACCAAaaagtagaattg	0.929	0.731
MAD2L1	rs2934378	-	4	121207924	FAC1	actaACAACactca	1	0.94
EDN1	rs2854239	+	6	12396669	myogenin	cagcgctggcttCCGGCtcagtgccgcct	0.687	0.759
E2F3	rs9465729	+	6	20508517	ETF	CCGCCgc	1	1
DDR1	rs34119233	+	6	30956073	YY1	GCCATgttg	1	0.997
TNF	rs4645839	+	6	31651804	myogenin	acggggctgcgtTCCAGctcacccaggga	0.829	0.716
HLA-DQA1	rs9272454	+	6	32713503	SREBP-1	tgGGGTG	1	1
HMGA1	rs9380423	+	6	34312660	ZF5	gaGGGCGgcctcc	0.919	0.862
CDKN1A	rs34414143	+	6	36754408	myogenin	cctgggctcccaTCCCCacagcagaggag	0.597	0.71
PPP2R5D	rs6906393	+	6	43059321	CACD	ccaCACCC	1	1
VEGFA	rs1005230	+	6	43844474	CDP	CAATAgatctgtgtg	0.996	0.936
GJA1	rs35296339	+	6	121796608	POU3F2	TCATGgtaat	0.783	0.855
MAP3K4	rs9456608	+	6	161332310	Hand1:E47	ttaataaTCTGGaatt	1	0.944
MLLT4	rs9455902	+	6	167967870	Hand1:E47	ccgtCCAGAcccagaa	1	0.937
HLA-DRB1	rs17878475	-	6	32665458	Pax-8	tggtagggTGTGAat	0.953	0.91
MAP3K5	rs1474988	-	6	137155946	Tax/CREB	ctggaaatgCTTCAg	0.8	0.747

Continued…

Gene	SNP	Strand	Chr	Position	TF	TF Sequence	CSS	MSS
PLAGL1	rs28364590	-	6	144371507	Cart-1	aacTAATGgaaattgaga	0.951	0.901
LATS1	rs2297932	-	6	150081054	AP-2	aaCCCACgggcg	0.969	0.952
RNASET2	rs11557915	-	6	167289645	Pax-5	ccatgcgccctgcagCCCTGcgcggggc	0.988	0.888
SERPINF1	rs12948385	+	17	1611651	Pax-2	tgtgtaacccATGACccac	0.991	0.9
MAP2K4	rs9892151	+	17	11864845	C/EBP	gTTGCCtaatct	1	0.994
MAP2K3	rs28365971	+	17	21128276	Hand1:E47	ctgcgggTCTGGgggt	1	0.939
LASP1	rs3842366	+	17	34279059	Pax-2	tagttcgtacTTGACtatg	0.979	0.906
ERBB2	rs35771148	+	17	35098000	ER	gttGGTCAgggtggtcttg	1	0.952
CSF3	rs34616965	+	17	35420873	Spz1	tgcGGAGGgtgtact	1	0.96
GRN	rs4792937	+	17	39777388	HNF3alpha	taaaacAAACA	1	0.999
ITGB3	rs11871447	+	17	42686617	1-Oct	AATTTacataga	0.93	0.958
DLX4	rs4399574	+	17	45401010	Egr-1	ccgcGGGGGcgt	0.885	0.864
RPS6KB1	rs36013892	+	17	55324243	S8	tacattcAATTAacat	1	0.998
MAP3K3	rs11871767	+	17	59052455	Pax-8	ccTCACGccggagct	1	0.941
TP53	rs17882137	-	17	7533245	CDP	CAATAaacctgggtc	0.996	0.835
BRCA1	rs3092986	-	17	38531522	Pax-2	tccataactgTTGACaagt	0.979	0.854
MPO	rs34576380	-	17	53714238	Ets	gaGGAAGt	1	1
MAP3K8	rs8176941	+	10	30761201	myogenin	taaattttatttTTGGTtggccgcggtgg	0.929	0.702
PLAU	rs2227554	+	10	75340118	Pax-6	cccgtTAACActtcaatagga	0.659	0.765
SNCG	rs3793899	+	10	88708741	LRF	GGGGGcccc	1	1
PTEN	rs35361056	+	10	89611571	SREBP	gggttCACCCta	1	0.971
GSTO2	rs10509769	+	10	106018060	Pax-5	gctgtcggtcCGCTCcaattgtctggtt	0.973	0.874
CXCL12	rs2839682	-	10	44202341	ZF5	ctGCGCGcggctc	1	0.86
TUSC3	rs11545037	+	8	15442403	Hand1:E47	agagCCAGActgtcaa	1	0.94
E2F5	rs4150842	+	8	86277150	C/EBP	tTTGCAaaactt	0.997	0.997
EBAG9	rs1892762	+	8	110620840	Hand1:E47	agggacaTCTGGcaga	1	0.936

Continued…

Gene	SNP	Strand	Chr	Position	TF	TF Sequence	CSS	MSS
MYC	rs3895617	+	8	128816042	myogenin	catgtgtggggcTGGGCaactagctaagt	0.817	0.731
PTP4A3	rs28549814	+	8	142496729	ZF5	ctcgggCGCCCtc	0.919	0.873
TNFRSF10A	rs20577	-	8	23138422	CP2/LBP-1c/LSF	gcggccacacCCAGC	1	0.918
PLAT	rs8178669	-	8	42185902	Pax-4	ggtgtctttttgatgtaatgatttcTTTTT	1	0.831
PTK2	rs306954	-	8	142082141	GR	aAGAACacagtgttgg	1	0.872
COMT	rs9332307	+	22	18307629	Pax-4	AAAAAttagccaggcgtggtggcagatgcc	1	0.849
MYO18B	rs4369968	+	22	24467323	CACD	ccaCGCCC	0.983	0.988
LGALS1	rs4820293	+	22	36400867	GR	aAGGACagggtgcaca	0.978	0.874
ECGF1	rs28931613	-	22	49314874	Pax-3	gcatgaagCGAGAcggaggcc	0.818	0.849
PPARG	rs17029007	+	3	12304483	C/EBPdelta	agtggcGCAATc	1	0.973
MLH1	rs1800734	+	3	37009950	v-Myb	tCCGTTagt	1	0.993
ACVR2B	rs506993	+	3	38470967	c-Ets-1	aattacTTCCGttatc	1	0.968
CTNNB1	rs11564433	+	3	41215785	CDP	actttaaTCAATtgc	0.93	0.92
SEMA3B	rs36018346	+	3	50278462	Spz1	gcgGGGGGgtttctg	0.998	0.968
PRKCI	rs1082975	+	3	171422206	Pax-6	atattTTCTGgttgagtttct	0.633	0.756
PIK3CA	rs7615076	+	3	180346382	myogenin	caattatttaatTTTGAagtataccattt	0.746	0.708
CDC25A	rs3731483	-	3	48205120	AP-2alpha	GCCCGgggc	1	1
GSK3B	rs334557	-	3	121296168	v-Myb	tCCGTTcgg	1	0.939
ATR	rs35582603	-	3	143780806	PPARalpha:RXRalpha	cacctgaaggaAAAAGggca	0.875	0.751
BCHE	rs3806650	-	3	167038505	CdxA	aTTAATa	1	1
TNFSF10	rs3136581	-	3	173725855	c-Ets-1(p54)	gactTCCTGc	0.974	0.973
SST	rs35603672	-	3	188870742	ZF5	caGTGCGcgctgg	0.919	0.868
MSH3	rs1643646	+	5	79984397	FAC1	acccACAAGacaaa	0.919	0.94
CAMK4	rs34549881	+	5	110586803	PPARalpha:RXRalpha	aaaaaaaaggcCAAAAgtaa	0.751	0.776
CTNNA1	rs28365836	+	5	138116655	Sp1	gggGGCGGgg	1	1
DAB2	rs3812039	-	5	39462084	myogenin	acattgagctacCCCAAattacacagtgg	0.911	0.706

Continued…

Gene	SNP	Strand	Chr	Position	TF	TF Sequence	CSS	MSS
PPAP2A	rs2292279	-	5	54866630	POU6F1	gaataaTTTAT	1	0.941
CDC25C	rs11567954	-	5	137695646	Pax-6	agccaatgatgCGCCAggctc	0.737	0.777
NTRK2	rs3758317	+	9	86472435	E2F	cctTGGCGcgtc	1	0.948
CTSL	rs3118869	+	9	89530683	Pax-5	tccaggtcCACTGaggcaggcacgccca	1	0.836
PTGS1	rs10306109	+	9	124171452	Pax-3	gtttaaggTCACGctatggaa	1	0.946
PPP2R4	rs3124501	+	9	130913459	STAT	tccCAGAAgtaga	1	0.996
CDKN2A	rs3731190	-	9	21984282	Pax-6	gggctTGACGtctgatctgta	0.925	0.794
CXADR	rs211964	+	21	17807417	HIC1	gctcgcTGCCCgcgg	1	0.978
BACH1	rs7509867	+	21	29593108	ZF5	gccgggCGCTCtc	0.919	0.871
CBR3	rs8132243	+	21	36429035	RFX	caGTTGCca	1	0.994
COL18A1	rs12482579	+	21	45647797	Pax-5	gtgcctgtccCGCGCaggtgcccctggc	0.839	0.743
BAGE	rs2770494	-	21	10121586	VDR	tcacccttttCCCCC	0.956	0.971
BRCA2	rs9534160	+	13	31786021	AP-2	cgccgCCGGGag	0.952	0.958
DNAJC15	rs9594861	+	13	42493917	CDP	tATCGAtctg	1	0.93
ERCC5	rs4150250	+	13	102296426	YY1	aaacATGGC	1	0.994
LAMP1	rs9604056	+	13	112998011	MRF-2	acttaaAATACaaa	1	0.91
GAS6	rs8181793	+	13	113545319	CACD	GGGAGtgg	0.948	0.965
MPG	rs3176362	+	16	67489	ZF5	cggctgCGCACtg	0.919	0.859
MSLN	rs12597489	+	16	749162	Muscle	gtgcgccacCACACagggcct	0.995	0.926
CIITA	rs12928665	+	16	10878975	CDP	CCATAtccgtttgtt	1	0.839

Appendix II

A. read_hapmap.pl

This program implementing several Perl modules (e.g. Getopt::Long, WWW::Mechanize, HTML::Extract, URI, etc.), enabled the retrieval of SNP genotype data for each of the 379 candidate genes per population group in both the forward and reverse strands. The program reads through a flat file containing all HUGO gene symbols of the candidate genes (one per line), then inserts them one at a time into the text query field of the HapMap Genome Browser tool (i.e. "Landmark or Region"). It then performs the "Search" action, enabling the HUGO gene symbol to be queried through the tool. On the resulting webpage, the script then accesses the "Configure..." button, where it selects the specified strand orientation (i.e. provided as a commandline option when running the script) and each population group one at a time, saving the text results of each selection to an output folder created in the working directory.

```
#!/usr/bin/perl -w

use diagnostics;
use Getopt::Long;
use WWW::Mechanize;
use HTML::TableExtract;
use URI;
use strict;

#----------HapMap homepage----------#
my $url = "http://www.hapmap.org";
my $mech = WWW::Mechanize->new();

#----------Create command line options----------#
my ($gene,$strand,$output_dir);

GetOptions(
   "strand|s=s"  => \$strand,
   "output|o=s"  => \$output_dir,
);

#----------Open flat file containing query genes (one per line) & run through HapMap
Genome Browser----------#
open(MYINPUTFILE, "Candidate_Genes.txt");
while(<MYINPUTFILE>) {
   my($line) = $_;
   chomp($line);
   $gene = $_;
   my($processGene);
   for my $line ($gene){
      processGene();
   }
}
```

```
#----------Submit query gene & retrieve genotype data----------#
sub processGene {

#----------Create output directory----------#
unless(-d $output_dir){system("mkdir $output_dir");}
my $outfile="HapMap.html";
open(OUTFILE,">$output_dir/$outfile");
chomp($gene,$strand);

#----------Initialize population codes----------#
my @pop=("ASW","CEU","CHB","CHD","GIH","JPT","LWK","MEX","MKK","TSI","YRI");

#----------Download genotype data on fwd/rev strand----------#
foreach my $k(0..$#pop){
        $mech->get($url);

        #---------Access HapMap Genome Browser & edit relevant fields--------#
        $mech->follow_link(text => "HapMap Genome Browser ( Phase 1, 2 & 3 -
merged genotypes & frequencies )");

        #----------Submit query gene----------#
        $mech->set_fields(
                name => $gene,
                plugin => "Download SNP genotype data",
        );
        $mech->submit();
        print OUTFILE $mech->content();

        my $output_page= "$gene.$pop[$k].$strand.txt";
        open(OUT, ">$output_dir/$output_page");
        $mech->set_fields(
                "SNPGenotypeDataPhase3Dumper.pop_code" => $pop[$k],
                "SNPGenotypeDataPhase3Dumper.strand" => $ARGV[1],
                "SNPGenotypeDataPhase3Dumper.format" => "todisk",
                 plugin_action=> "Go",
        );
         $mech->click("plugin_action");

     #----------Progress status----------#
     # print "Downloading... $pop[$k] \n";
         print OUT $mech->content();
}
close(OUT);
close(OUTFILE);
@pop=();
}
```

B. filter_hapmap.py

This program opens and reads through all text files obtained from read_hapmap.pl program (i.e. genotyped SNPs occurring on all 379 candidate genes among 11 population groups on both the forward and reverse strands) and searches for SNPs present in any of the population groups corresponding to the SNPs that were observed to overlap with TFBSs in Chapter 3 (Section 3.2.5). All SNPs that were identified in any/all population groups and present in File 10 (Figure 3.8) were then piped to an output flat file from the commandline interface.

```
#!/usr/bin/env python

import csv, os
from table_parser import *
ref = open("SNPs_within_TFBSs.csv","rb")
writer = csv.writer(file('SNPs_in_all_Populations.csv', 'w'))

#----------Accesses folder that stores all HapMap result files----------#
path = "/Users/kavisharamdayal/Documents/FINAL_THESIS/8.HapMap_Results"
HapFolder = os.listdir(path)

a0 = b0 = c0 = f_0 = a1 = b1 = c1 = f_1 = a2 = b2 = c2 = f_2 = a3 = b3 = c3 = f_3 = a4 =
b4 = c4 = f_4 = a5 = b5 = c5 = f_5 = a6 = b6 = c6 = f_6 = a7 = b7 = c7 = f_7 = a8 = b8 =
c8 = f_8 = a9 = b9 = c9 = f_9 = a10 = b10 = c10 = f_10 = ""
f0 = f1 = f2 = f3 = f4 = f5 = f6 = f7 = f8 = f9 = f10 = snps_in_tfbs = []
total_fwd_count = total_rev_count = 0

for line in ref:
    ref_gene = line.strip().split(",")[0]
    ref_snp = line.strip().split(",")[1]
    ref_str = line.strip().split(",")[2]
    ref_chr = line.strip().split(",")[3]
    ref_pos = line.strip().split(",")[4]
    ref_tf = line.strip().split(",")[5]
    MI = line.strip().split(",")[6]
    tf_seq = line.strip().split(",")[7]
    CSS = line.strip().split(",")[8]
    MSS = line.strip().split(",")[9]
    tf_base = line.strip().split(",")[10]
    temp =
str(ref_gene+"|"+ref_snp+"|"+ref_str+"|"+ref_chr+"|"+ref_pos+"|"+ref_tf+"|"+MI+"|"+tf_seq
+"|"+CSS+"|"+MSS)
    snps_in_tfbs.append(temp)

def get_info(f):
    filename = str(f).split("'")[1]
    gene = filename.split(".")[0]
    pop = filename.split(".")[1]
    strand = filename.split(".")[2]
    return gene, pop, strand
```

```
def get_filecontent(f):
    temp = []
    for line in f:
        if (line.startswith("rs")) and ("#" not in line):
            snp = line.split()[0]
            temp.append(snp)
    return temp

#----------Analyze forward strand SNPs----------#
fwd_commons = []
for file in HapFolder:
    if file.endswith("ASW.fwd.txt"):
        f_0 = open(file,"rb")
        f0 = get_filecontent(f_0)
        a0,b0,c0 = get_info(f_0)
elif file.endswith("CEU.fwd.txt"):
        f_1 = open(file,"rb")
        f1 = get_filecontent(f_1)
        a1,b1,c1 = get_info(f_1)
elif file.endswith("CHB.fwd.txt"):
        f_2 = open(file,"rb")
        f2 = get_filecontent(f_2)
        a2,b2,c2 = get_info(f_2)
elif file.endswith("CHD.fwd.txt"):
        f_3 = open(file,"rb")
        f3 = get_filecontent(f_3)
        a3,b3,c3 = get_info(f_3)
elif file.endswith("GIH.fwd.txt"):
        f_4 = open(file,"rb")
        f4 = get_filecontent(f_4)
        a4,b4,c4 = get_info(f_4)
elif file.endswith("JPT.fwd.txt"):
        f_5 = open(file,"rb")
        f5 = get_filecontent(f_5)
        a5,b5,c5 = get_info(f_5)
elif file.endswith("LWK.fwd.txt"):
        f_6 = open(file,"rb")
        f6 = get_filecontent(f_6)
        a6,b6,c6 = get_info(f_6)
elif file.endswith("MEX.fwd.txt"):
        f_7 = open(file,"rb")
        f7 = get_filecontent(f_7)
        a7,b7,c7 = get_info(f_7)
elif file.endswith("MKK.fwd.txt"):
        f_8 = open(file,"rb")
        f8 = get_filecontent(f_8)
        a8,b8,c8 = get_info(f_8)
elif file.endswith("TSI.fwd.txt"):
        f_9 = open(file,"rb")
        f9 = get_filecontent(f_9)
        a9,b9,c9 = get_info(f_9)
elif file.endswith("YRI.fwd.txt"):
        f_10 = open(file,"rb")
        f10 = get_filecontent(f_10)
        a10,b10,c10 = get_info(f_10)
```

```
total_fwd_count +=
len(f1)+len(f2)+len(f3)+len(f4)+len(f5)+len(f6)+len(f7)+len(f8)+len(f9)+len(f10)

    if (a0==a1==a2==a3==a4==a5==a6==a7==a8==a9==a10) and
(c0==c1==c2==c3==c4==c5==c6==c7==c8==c9==c10):
        for snp in f0:
            if (snp in f1) or (snp in f2) or (snp in f3) or (snp in f4) or (snp in f5) or (snp in f6) or
(snp in f7) or(snp in f8) or (snp in f9) or (snp in f10):
                fwd_commons.append(snp)

#----------Analyze reverse strand SNPs----------#
rev_commons = []
for file in HapFolder:
    if file.endswith("ASW.rev.txt"):
        f_0 = open(file,"rb")
        f0 = get_filecontent(f_0)
        a0,b0,c0 = get_info(f_0)
elif file.endswith("CEU.rev.txt"):
        f_1 = open(file,"rb")
        f1 = get_filecontent(f_1)
        a1,b1,c1 = get_info(f_1)
elif file.endswith("CHB.rev.txt"):
        f_2 = open(file,"rb")
        f2 = get_filecontent(f_2)
        a2,b2,c2 = get_info(f_2)
elif file.endswith("CHD.rev.txt"):
        f_3 = open(file,"rb")
        f3 = get_filecontent(f_3)
        a3,b3,c3 = get_info(f_3)
elif file.endswith("GIH.rev.txt"):
        f_4 = open(file,"rb")
        f4 = get_filecontent(f_4)
        a4,b4,c4 = get_info(f_4)
elif file.endswith("JPT.rev.txt"):
        f_5 = open(file,"rb")
        f5 = get_filecontent(f_5)
        a5,b5,c5 = get_info(f_5)
elif file.endswith("LWK.rev.txt"):
        f_6 = open(file,"rb")
        f6 = get_filecontent(f_6)
        a6,b6,c6 = get_info(f_6)
elif file.endswith("MEX.rev.txt"):
        f_7 = open(file,"rb")
        f7 = get_filecontent(f_7)
        a7,b7,c7 = get_info(f_7)
elif file.endswith("MKK.rev.txt"):
        f_8 = open(file,"rb")
        f8 = get_filecontent(f_8)
        a8,b8,c8 = get_info(f_8)
elif file.endswith("TSI.rev.txt"):
        f_9 = open(file,"rb")
        f9 = get_filecontent(f_9)
        a9,b9,c9 = get_info(f_9)
elif file.endswith("YRI.rev.txt"):
        f_10 = open(file,"rb")
        f10 = get_filecontent(f_10)
        a10,b10,c10 = get_info(f_10)
    total_rev_count +=
len(f1)+len(f2)+len(f3)+len(f4)+len(f5)+len(f6)+len(f7)+len(f8)+len(f9)+len(f10)

    if (a0==a1==a2==a3==a4==a5==a6==a7==a8==a9==a10) and
(c0==c1==c2==c3==c4==c5==c6==c7==c8==c9==c10):
        for snp in f0:
            if (snp in f1) or (snp in f2) or (snp in f3) or (snp in f4) or (snp in f5) or (snp in f6) or
(snp in f7) or (snp in f8) or (snp in f9) or (snp in f10):
                rev_commons.append(snp)

for i in snps_in_tfbs:
    gene = i.split("|")[0]
    snp = i.split("|")[1]
    strand = i.split("|")[2]
    chr = i.split("|")[3]
    pos = i.split("|")[4]
    tf = i.split("|")[5]
    MI = i.split("|")[6]
    tf_seq = i.split("|")[7]
    CSS = i.split("|")[8]
    MSS = i.split("|")[9]
    tf_base = i.split("|")[10]

if (strand == "+") and (snp in fwd_commons):
        print gene,snp,strand,chr,pos,tf,tf_seq,CSS,MSS,tf_base
    elif (strand == "-") and (snp in rev_commons):
        print gene,snp,strand,chr,pos,tf,tf_seq,CSS,MSS,tf_base

f_0.close()
f_1.close()
f_2.close()
f_3.close()
f_4.close()
f_5.close()
f_6.close()
f_7.close()
f_8.close()
f_9.close()
f_10.close()
ref.close()

#print "snps_in_tfbs",len(snps_in_tfbs) = 988
#print "fwd_commons",len(fwd_commons) = 23852
#print "rev_commons",len(rev_commons) = 24133
#print "total fwd snps",total_fwd_count = 5865604
#print "total rev snps",total_rev_count = 5878560
```