POTENTIALLY FUNCTIONAL SINGLE NUCLEOTIDE POLYMORPHISMS (PFSNPS) IN THE HUMAN GENOME AND THEIR POTENTIAL APPLICATIONS

JINGBO WANG

(B.Sc.) NUS

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHYLOSOPHY

DEPARTMENT OF BIOCHEMISTRY

NATIONAL UNIVERSITY OF SINGAPORE

2012

ACKNOWLEDGEMENTS

I would like to take the opportunity to thank the following people for making this thesis possible.

 Associate Professor Caroline Lee, Department of Biochemistry, National University of Singapore, for giving me the opportunity to pursuit my post-graduate degree and providing priceless guidance and support throughout my candidature.

 Associate Professor Wing-Kin Sung, Department of Computer Science, School of Computing, National University of Singapore and Assistant Professor Liang Kee Goh, DUKE-NUS Graduate Medical School, for being referees of my Ph.D qualifying exam and thesis advisory committee member.

 Associate Professor Poh San Lai, Department of Paediatrics, Yong Loo. Lin School of Medicine, National University of Singapore for being the chair person of my Ph.D qualifying exam.

 My senior Dr. Tang Kun, Dr. Ren Jianwei, Dr. Wang Zihua, Dr. Gwee Pai Chong, Mr. Zhang Dongwei and Mr. Wang Baoshuang for teaching me indispensable lab techniques and providing much needed advice on troubleshooting.

 My fellow post-graduate friends Dr. Grace Pang, Department of Biochemistry, National University of Singapore for many discussions on the subject of this thesis and kind help of proof-reading part of the thesis.

 My colleagues Mr. Teo Wei Bing, Miss Wong Yin Yee, Ms. Choy Mingzi Juliana and Ms. Xiao Peiyun for their invaluable support in carrying out certain experiments.

 Dr. Rebecca Jackson, Department of Biochemistry, National University of Singapore for her kind help of proof-reading the thesis.

 My fellow post-graduate friends and colleagues Dr. Lu Yiwei, Dr. Wang Yu, Dr. Cao Yi, Dr. Gao Yun, Dr. Cheryl Chan, Miss Toh Soo Ting, Mr Mah Wei Champ, Mr. Maulana Bachtiar and Miss Jin Yu for the friendship.

 The people in the Department of Biochemistry and the Dean's office of the Faculty of Medicine for their assistance in administrative matters.

 My wife Liu Ying for her love and understanding during difficult times.

 My family for their love, support and encouragement.

TABLE OF CONTENTS

SUMMARY

As of 2009, about 56,000,000 single nucleotide polymorphisms (SNPs) have been deposited in the NCBI dbSNP database, Build 129 (dbSNP 129). Of these, >14,000,000 represent non-redundant SNPs and ~6,600,000 SNPs have been validated. Some of these SNPs are likely to affect phenotype and be causally associated with disease risk or drug response. However, identifying these phenotype-affecting SNPs from a pool of >14 million poses significant challenges. Several strategies have been proposed for the selection of a subset of SNPs that may be useful for disease- or drug-response gene-based or genome-wide association studies (GBAS/GWAS) (Jorgenson and Witte 2006; Rebbeck, Spitz and Wu 2004).

This dissertation attempts to annotate the potential functionality of SNPs in dbSNP129 by integrating >40 different algorithms/resources to interrogate >14,000,000 SNPs from this dbSNP database for SNPs of potential functional significance based on previously published reports, inferred potential functionality from genetic approaches as well as predicted potential functionality from sequence motifs. A bioinformatics pipeline was first established and a data warehouse for potential functional SNPs (pfSNPs) identified from >40 resources was built. A comprehensive, well-annotated, integrated pfSNP (potentially functional SNPs) Web Resource (http://pfs.nus.edu.sg/) which is aimed to facilitate better hypothesis generation through knowledge syntheses mediated by better data integration and a user-friendly web interface was then developed. Its query-interface has the user-friendly 'auto-complete, prompt-as-you-type' feature and is highly customizable facilitating different combination of queries using Boolean-logic. Additionally, to facilitate better understanding of the results and aid in hypotheses generation, gene/pathway-level information with text-clouds highlighting enriched tissues/pathways as well as detailed related information are also

provided on the results page. Hence, this pfSNP resource will be of great interest to scientists focusing on association studies as well as those interested to experimentally address the functionality of SNPs.

To demonstrate the applicability of the pfSNP resource, I employed this resource and examined pfSNPs to identify SNPs that are significantly associated with response to anti-cancer drugs (5FU/capecitabine and oxaliplatin) in metastatic colorectal cancer patients. Notably, the minor allele of 2 non-synonymous SNPs potentially affecting exon-splice enhancers (ESE) and nonsense mediated decay (NMD) as well as a 3'UTR SNP affecting miRNA binding at the UMPS gene was observed only in non-responders but not in responders when >100 patients are examined. Due to the limited number of patients examined, these SNP did not pass multiple test correction. Future studies could examine more patients to evaluate if these SNPs with potential function are in fact the causative SNPs associated with response to these anti-cancer drugs.

In summary, I have developed a useful resource that annotates the potential functionality of SNPs in the human genome. Employing these pfSNPs in pathways associated with 5FU, oxaliplatin and colorectal cancer, I identified 3 pfSNPs affecting ESE, NMD and miRNA binding that may play a role in determining differences in drug response of different individuals to these drugs.

LIST OF PUBLICATIONS DURING PHD TENURE

Publications resulting from research carried out during the period of postgraduate program are listed below.

International Peer Reviewed Primary Publications:

1. **Jingbo Wang**, Mostafa Ronaghi, Samuel S Chong, Caroline GL Lee*. pfSNP: An Integrated Potentially Functional SNP Resource that facilitates Hypotheses Generation through Knowledge Syntheses. ***Human Mutation*** 32(1):19-24 (2011)

2. Grace SY Pang, **Jingbo Wang**, Zihua Wang, Caroline GL Lee*****. The G-allele of SNP E1/A118G at the Mu Opioid Receptor Gene locus shows Genomic Evidence of Recent Positive Selection. ***Pharmacogenomics*** 10(7):1101-1109 (2009)

3. Wang Y, Lee AT, Ma JZ, **Jingbo Wang**, Ren J, Yang Y, Tantoso E, Li KB, Ooi LL, Tan P et al: Profiling microRNA expression in hepatocellular carcinoma reveals microRNA-224 up-regulation and apoptosis inhibitor-5 as a microRNA-224-specific target. ***J Biol Chem***, 283(19):13205-13215 (2008).

4. Zihua Wang, **Jingbo Wang**, Erwin Tantoso, Baoshuang Wang, Amy YP Tai, London L.P.J Ooi, Samuel S Chong, and Caroline GL Lee*. Signatures of Recent Positive Selection at the ATP-Binding Cassette (ABC) Drug Transporter Superfamily Gene Loci. ***Human Molecular Genetics*** 16(11):1367-1380 (2007)

International Peer Reviewed Reviews Articles:

1. **Jingbo Wang**, Grace SY Pang, Samuel S. Chong and Caroline GL Lee*. SNP web resources and their potential applications in personalized medicine. ***Current Drug Metabolism***. 13(7):978-990 (2012)

2. Wolf SJ, Bachtiar M, **Jingbo Wang**, Sim TS, Chong SS, Lee CG*. An update on ABCB1 pharmacogenetics: insights from a 3D model into the location and evolutionary conservation of residues corresponding to SNPs associated with drug pharmacokinetics. ***The Pharmacogenomics Journal***, 11(5):315-325 (2011).

3. Grace SY Pang, **Jingbo Wang**, Zihua Wang, Caroline GL Lee*. Predicting Potentially Functional Single Nucleotide Polymorphisms in Drug Response Genes. ***Pharmacogenomics*** 10(4): 639-653 (2009)

4. Zihua Wang, **Jingbo Wang**, Samuel S. Chong and Caroline G.L. Lee*****. Mining Potential Functional Polymorphisms at the ATP-Binding Cassette Transporter Family. ***Current Pharmacogenomics and Personalized Medicine*** 7(1):40-58 (2009).

International Conference Poster Presentations:

1. **Jingbo Wang**, Mostafa Ronaghi, Samuel S Chong, Caroline GL Lee*. pfSNP: An Integrated Portal To Identify Potential SNP Biomarkers For Cancer And Complex Diseases. ***International Symposium on Integrative Bioinformatics 2011 (7th annual meeting),*** March 21-23, 2011, Wageningen University, The NETHERLANDS.

2. **Jingbo Wang**, Mostafa Ronaghi, Samuel S Chong, Caroline GL Lee*. pfSNP: An Integrated Potentially Functional SNP Resource that facilitates Hypotheses Generation through Knowledge Syntheses. ***Third Asian Young Researchers Conference on Computational and Omics Biology,*** March 10-12, 2010, National Cheng Kung University, Taiwan.

LIST OF PENDING PATENTS DURING PHD TENURE

Patent applications resulting from research carried out during the period of postgraduate program are listed below.

1. **US Provisional Patent Application No. 61/663,867**
   Title: Potentially Functional SNP (pfSNP) Panel Identification Methodology and pfSNP Marker Panel for 5-Fluorouracil (5-FU) and/or FOLFOX Response Prediction in Colorectal Patients

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

5-FU	5-fluorouracil
ACNC	Accelerated conserved non-coding
AJAX	Asynchronous Javascript and XML
BDIS	Background Data Integration Server
CAST	Cohort Allelic Sums Test
CDCV	Common Disease, Common Variant
CDRV	Common Disease, Rare Variant
CEPH	Centre d'Etude du Polymorphisme Humain
CEU	U.S. residents with northern and western European ancestry
CFTR	Cystic fibrosis trans-membrane conductance receptor
CHB	Han Chinese in Beijing, China
CMC	Combined Multivariate and Collapsing
CR	Complete Response
CRC	Colorectal cancer
dbSNP	NCBI SNP database
DPYD	Dihydropyrimidine Dehydrogenase
EHH	Extended Haplotype Homozygosity
ESE	Exonic splicing enhancers
ESS	Exonic splicing silencers
G6PD	Glucose-6-phosphate dehydrogenase
GAD	Genetic Association Database
GBAS	Gene-based association studies
GO	Gene Ontology
GWAS	Genome-wide association studies
HGMD	Human Gene Mutation Database
HH	Haplotype Homozygosity
IAS	Internet Application Server
IDS	Internet Data Sources
ISRE	Intronic splicing regulatory elements
JPT	Japanese in Tokyo, Japan
KEGG	Kyoto Encyclopedia of Genes and Genomes
LAS	Linux Application Server
LD	Linkage disequilibrium
LOVD	Leiden Open source sequence Variation Database
LSDB	Locus specific database
MAF	Minor allele frequency
mCRC	Metastatic colorectal cancer
miRNA	micro RNA
mSigDB	Molecular Signatures Database
MTHFR	Methylenetetrahydrofolate reductase
NCBI	National Center for Biotechnology Information
NIEHS	National Institute of Environmental Health Sciences
NMD	Nonsense mediated mRNA decay

ODBC	Open Database Connectivity
OMIM	Online Mendelian Inheritance in Man
OR	Odds Ratio
OS	Overall Survival
PD	Progressive Disease
pfSNP	SNP predicted to be potentially functional
PK	pharmacokinetics
PoDA	Pathways of Distinction Analysis
PR	Partial Response
PWM	Positional Weight Matrices
QR-SNPs	Quasi-randomly selected SNPs in Affymetrix SNP Chip
RECIST	Response Evaluation Criteria In Solid Tumours
rEHH	Relative Extended Haplotype Homozygosity
RFS	Relapse-Free Survival
rsNo	NCBI dbSNP reference cluster numbers
SD	Stable Disease
SNP	Single nucleotide polymorphism
TF	Transcription factor
TFBS	Transcription factor binding sites
t-SNPs	Tagging SNPs in Illumina SNP chip
TYMP	Thymidine Phosphorylase
TYMS	Thymidylate Synthase
UMD	Universal Mutation Database
UMPS	Uridine Monophosphate Synthetase
VB	Visual Basic
VB.NET	Visual Basic .NET
VBA	Visual Basic for Application
VNTR	Variable number tandem repeat
WAS	Windows Application Server
YRI	Yoruban population in Ibadan, Nigeria

CHAPTER I: GENERAL INTRODUCTION

1.1 Overview of Single Nucleotide Polymorphisms (SNPs)

1.1.1 SNPs and their role in human health

Single nucleotide polymorphisms (SNPs) refer to single base differences at particular chromosome loci. More simply put, SNPs are copying errors that create variations between people, influencing a variety of differences in traits, disease susceptibility and drug response. SNPs arise from founder germ-line mutations, and such mutations are retained in the population through random genetic drift (when the mutation is evolutionarily neutral) (Kimura 1979) or positive selection if the mutation carries a selective advantage. In recent years, the distinction between rare SNP and mutation are blurred therefore in this thesis the term SNP may also refer to mutation as well.

SNPs can act as useful genetic markers to identify genes that influence human health-related traits. Two methods, namely the family-based "linkage" mapping (Lander and Schork 1994) and population-based genetic association studies (Freimer and Sabatti 2004), can be used for gene discovery. The "linkage" mapping approach uses various genetic markers, such as microsatellites (otherwise known as short tandem repeats) and SNPs, to study segregated genetic markers amongst related individuals to ascertain the genetic basis of a certain phenotype, such as hair and eye colour. By comparison, genetic association studies attempt to find markers that have different allele distributions amongst affected and un-affected groups of unrelated individuals in the population. For these studies, SNPs are the marker of choice, as

recombination occurs extensively at the population scale and genetic fine-mapping using more markers will increase the power of such studies (Kruglyak 1997; Evans and Cardon 2004). As compared with microsatellites, SNPs are more abundant, with an initial estimated frequency of one per 100 - 300 base pairs (Cargill, Altshuler et al. 1999). With the continuous effort of SNP discovery and the advent of new sequencing technology, SNPs are found widely spread across the human genome and the list of known SNPs is growing at an accelerated rate. More than 52 million SNPs have been catalogued in the latest release of the National Center for Biotechnology Information (NCBI) SNP database (dbSNP) (db135, as of June 2012), of which 12 million are newly reports.

Notably, many SNPs reside in functionally important regions of genes and affect the molecular functions of the gene. Some of them may represent causal SNPs that can directly impact upon human health rather than merely acting as markers. One classical example is the SNPs in the glucose-6-phosphate dehydrogenase (G6PD) gene. G6PD gene deficiency is known to be the cause of sickle cell anaemia, but also confers protection against malaria (Ruwende, Khoo et al. 1995). It has been demonstrated that two SNPs (rs1050828, Val68Met and rs1050829, Asn126Asp) in the G6PD gene, which are both non-synonymous SNPs in the coding region, together cause a severe drop in the enzyme yield (Town, Bautista et al. 1992), and thus directly causes the G6PD gene deficiency. Currently, there are more than 160 SNPs or mutations found to cause G6PD deficiency in different world populations (Mason, Bautista et al. 2007). Besides G6PD deficiency, SNPs are also found to be directly causal in several other diseases, including cystic fibrosis (e.g. rs1800085, Val322Met in CFTR gene) (Bobadilla, Macek et al. 2002) and alpha thalassemia (e.g. 5'UR/T149709C in HBA1 gene) (De Gobbi, Viprakasit et al. 2006).

1.1.2 SNPs with varied molecular functions can be causal to human traits

1.1.2.1 SNPs change protein functions

Many of the identified causal SNPs are non-synonymous SNPs (Mason, Bautista et al. 2007); that is, these SNPs cause changes in the amino acids in the protein. Such SNPs may abolish or alter protein function, especially if the amino acid change occurs at a key residue in a motif that is crucial for protein function. In addition to GD6P gene deficiency, another example of a non-synonymous, causal SNP is the recently reported SNP (rs1804645, P1272S) in the SRC-1 gene (Hartmaier, Richter et al. 2012). This SNP causes a disruption to the phosphorylation site in the protein, and has been associated with decreased bone mineral density in breast cancer patients treated with an antagonist against the oestrogen receptor, called Tamoxifen. A non-synonymous SNP may also reside outside the functional domain and affect protein conformation, leading to a change in binding affinity to its substrate or to another protein partner.

Besides non-synonymous SNP, other types of SNPs can also alter protein conformation and substrate specificity. Synonymous SNPs, which result in the coding of the same amino acid, had long been thought of as “silent”. However, studies now suggest the role of synonymous SNPs in the direct alteration of protein conformation/substrate specificity (Soranzo, Cavalleri et al. 2004; Kimchi-Sarfaty, Oh et al. 2007), with accumulating evidences such as the recent mutations identified in the cystic fibrosis trans-membrane conductance receptor (CFTR) gene (Bartoszewski, Jablonsky et al. 2010). Studies also show that synonymous SNPs implicated in human disease extensively (Sauna and Kimchi-Sarfaty 2011) with effect size similar to that

of the non-synonymous counterparts (Chen, Davydov et al. 2010).

1.1.2.2 SNPs change protein expression level

The G6PD deficiency example also shows that gene expression level is important in determining human traits and SNPs in different gene regions may well affect gene expression via different mechanisms.

In the coding region, a SNP forming a stop codon instead of an amino acid residue change has long been implicated in causing human disease. Such SNPs reduce gene expression levels through a well-known mechanism named "Nonsense mediated mRNA decay" (NMD), which leads to the degradation of the mRNA containing the premature stop codon. In the CFTR gene, one such mutation reduces the mRNA level by more than 90% as compared with the wild-type allele (Hamosh, Rosenstein et al. 1992). It is now known that a number of SNPs/mutations causing premature stop codon are implicated in cystic fibrosis, accounting for 5-10% of the total mutant alleles in cystic fibrosis patients (Kerem 2004).

SNPs at the promoter region are known to affect gene expression at the transcriptional level by creating or disrupting transcription factor binding sites and thus leading to altered mRNA production. An example is the SNP in the α-globin cluster which results in α-thalassemia (De Gobbi, Viprakasit et al. 2006). This SNP (5'UR/T149709C, rsNo unknown) creates a binding site for the transcription factor GATA-1 and thus induces a promoter-like element. The newly created promoter then competes with the native promoter of the α-globin cluster, leading to a reduction in the expression α-globin in the α-thalassemia phenotype. In the MDM2 gene, a similar promoter-forming SNP (rs2279744, I/1/T309G) leads to an enhanced SP1 binding site.

This SNP was later found to be associated with earlier tumour onset (Bond, Hu et al. 2005).

At the 3' UTR region, SNPs may affect gene expression by altering micro RNA (miRNA) binding. miRNA binds to its target mRNA at the 3'UTR region and blocks translation. This is a mechanism recently demonstrated to be universal in gene expression regulation, with 30% of all the genes estimated to be regulated (Lewis, Burge et al. 2005). Although it is a relatively new gene regulation mechanism, as compared with transcription factor binding, some SNPs have already been shown to affect human health via this mechanism. In the papillary thyroid carcinoma, a SNP (rs17084733, E21/3UTR /G217A) in the miR221/222 binding site in the 3'UTR region of the KIT gene deregulates KIT expression via miR-221/222, which are up-regulated and postulated to affect cancer progression (He, Jazdzewski et al. 2005). Another 3' UTR SNP in dihydrofolate reductase leads to gene overexpression and therefore methotrexate resistance, as the T allele disrupts miR-24 binding to the mRNA and abolishes miRNA-mediated suppression (Mishra, Humeniuk et al. 2007).

1.1.2.3 SNPs cause differential splicing

Protein function may also be affected by differential mRNA splicing. Differential mRNA splicing is seemingly universal, with more than 95% of all human genes estimated to produce more than one protein product (Pan, Shai et al. 2008; Wang, Sandberg et al. 2008). Traditionally, SNPs at the splicing donor and acceptor sites were considered important in this process (Zhang 1998). However, with the discovery of splicing enhancers and silencers in the exon and intron (Fairbrother, Yeh et al. 2002; Yeo, Nostrand et al. 2007), a higher incidence of SNP involvement may

be implicated. Furthermore, 15-50% of human disease mutations are estimated to affect splice site selection (Wang and Cooper 2007). The following **Table 1.1** summarizes the aforementioned major gene regulatory mechanisms affected by SNPs, and their links to human health conditions.

Table 1.1: Gene regulatory mechanisms affected by SNPs in human disorders.

	Function of SNP	Examples in Human Health
mRNA	Reduce mRNA copy number by "Nonsense mediated mRNA decay"	Cystic fibrosis (Hamosh, Rosenstein et al. 1992)
	Change transcript level by changing TF binding site	1. α-thalassemia (De Gobbi, Viprakasit et al. 2006) 2. Earlier tumour onset (Bond, Hu et al. 2005)
	Change translational efficiency of mRNA by changing miRNA binding site	1. Progression of papillary thyroid carcinoma (He, Jazdzewski et al. 2005) 2. Methotrexate resistance (Mishra, Humeniuk et al. 2007)
	Change gene product by causing differential splicing	15-50% of human disease mutations estimated to affect splice site (Wang and Cooper 2007)
Protein	Reduce protein level by causing conformational change and destabilize the protein	Glucose-6-phosphate dehydrogenase deficiency (sickle cell anaemia) (Town, Bautista et al. 1992)
	Affect protein function by changing functional motif directly	Decreased bone mineral density in tamoxifen-treated women (Hartmaier, Richter et al. 2012)
	Affect protein function by changing protein conformation	Cystic fibrosis (Bartoszewski, Jablonsky et al. 2010)

1.1.3 Strategies and challenges to the identification of functional SNPs

Functional SNPs underlying human health conditions can be identified in two ways: (1) identification of relevant functional SNPs by demonstrating their biological functions and associating these SNPs to the possible conditions related to such

biological plausibility; or (2) lineage mapping/genetic association followed by functional testing of the SNPs to potentially explain the biological effect.

The first approach is usually employed when there is extensive prior knowledge to support the function of the SNP as well as the gene. This approach was employed for the MDM2 study discussed previously, where the promoter SNP in the MDM2 gene was demonstrated to enhance SP1 binding to the MDM2 promoter and later found to be associated with earlier tumour onset (Bond, Hu et al. 2005). By comparison, when prior knowledge is weak or new biological plausibility is required, the second approach is used. "Linkage mapping" or genetic association studies are first carried out to find the regions or SNPs likely to be responsible for the condition. This is later followed by functional testing of the SNPs for possible biological plausibility to explain the observed association. This approach led to the discovery of the SNP in the α-globin cluster associated with α-thalassemia (De Gobbi, Viprakasit et al. 2006). The region containing the SNP was first found to be associated with the disease in the Melanesian population. The SNP was later found to create a promoter and suppresses the expression of α-globin by competing with the native promoter. Between the two approaches, the latter reveals new biological mechanisms (Li 2006) because it is less restricted by prior knowledge. The drop in sequencing and large-scale genotyping costs also makes this approach more popular in recent years. This is evidenced by the wide adoption of the "hypothesis-free" genome-wide association studies (GWAS) for various human health conditions.

The successful application of both approaches requires a complete list of all possible functional SNPs to be known, such that no important SNPs are omitted. Further, limiting this list to only the most important SNPs would be more time- and cost-effective, and introduce less noise into the association analysis. Unfortunately, no

such list is readily available. The next section briefly discusses some of the computational resources that I employed to help derive such a list.

1.1.4 Computational methods and resources to facilitate the identification of functional SNPs

1.1.4.1 SNP cataloguing projects to reveal SNP architecture in gene and genome

To facilitate the identification of functional SNPs, a catalogue that contains a complete list of SNPs is required. This is particularly important in cases where more than one SNP function together to cause the phenotype and an incomplete list that does not contain one of these SNPs will not be helpful. A complete list of SNPs is also useful in GWAS studies to get a better view of the linkage disequilibrium profile of the population and subsequently to facilitate the selection of more appropriate tagging SNPs. **Table 1.2** tabulates some of these recent efforts to catalogue SNPs.

Table 1.2: Some of the SNP cataloguing efforts.

Database Name	Scope	Information Provided
LSDB (e.g. IARC TP53 database)	Single gene	1. Literature reported function 2. Allele frequency
Environmental Genome Project	648 genes important in gene-environment interactions	Allele frequency
Seattle SNPs Project	319 genes implicated in inflammation	Allele frequency
HapMap	Whole genome	1. Allele frequency in different world population for common SNP (MAF >5) 2. LD and haplotype information
1000 Genome Project	Whole genome	1. Allele frequency in different world population for all SNP 2. LD and haplotype information
NCBI dbSNP	Whole genome	1. Allele frequency by voluntary reporting from individual researchers, as well as large scale studies 2. Gene and genome context

Small scale catalogues in the format of a locus specific database (LSDB) were the first to be introduced (Claustres, Horaitis et al. 2002). These databases catalogue SNPs in one or few gene loci and serve as encyclopaedias for SNPs in the gene of interest because detailed phenotype relation/molecular function of the SNPs are included as a result of expert manual curating. These databases are generally biased towards more well-studied genes (Claustres, Horaitis et al. 2002), such as p53 (Olivier, Eeles et al. 2002; Hamroun, Kato et al. 2006). Due to the high popularity of LSDB, specific toolkits, such as the Universal Mutation Database (UMD) (Beroud, Hamroun et al. 2005) and Leiden Open source sequence Variation Database (LOVD) (Fokkema, Taschner et al. 2011) has been developed to facilitate database construction. LSDBs continue to be under heavy development and extensive efforts have been made to establish standards (Celli, Dalgleish et al. 2012) and provide guidelines for building such databases (Cotton, Auerbach et al. 2008; Vihinen, den Dunnen et al. 2012).

Medium-scale SNP catalogues are useful for genetic association studies that focus on more than one gene locus. These catalogues focus on SNPs in genes from a particular pathway or related to certain phenotype. One example of such resource is the National Institute of Environmental Health Sciences (NIEHS) Environmental Genome Project (Rieder, Livingston et al. 2008) (http://egp.gs.washington.edu/welcome.html), which houses information about 648 genes that have been previously reported in the literature to be important in gene-environment interactions. SNPs in these 648 genes are then postulated as important in gene-environment interactions. Similarly, SNPs catalogued in the Seattle SNPs Project (http://pga.gs.washington.edu/), a database of 319 genes previously reported to be involved in inflammation, could also be hypothesized to affect pathways involved in inflammation. Beside obtaining genotype information directly from

experiments, integrating genotype information into a central repository from scattered sources in different LSDBs may be useful (den Dunnen, Sijmons et al. 2009).

Large-scale projects cataloguing SNPs and targeting genome-wide SNP allele frequency information in different human populations are also available. Resources, such as ALFRED (Osier, Cheung et al. 2001) and SPSmart (Amigo, Salas et al. 2008), constructed SNP catalogues by collecting SNP population specific allele frequency information from the literature. A different approach, which is to generate this information by genotyping a group of individuals from each population via high-throughput methods, was undertaken by collaborative efforts under The International HAPMAP Project. In the pilot phase of the project, samples representing the four major world populations were genotyped. These included thirty sets of trios from a USA Utah population, with Northern and Western European ancestry collected by the Centre d'Etude du Polymorphisme Humain (CEPH); thirty sets of trios from a Yoruban population in Ibadan, Nigeria (YRI); 45 unrelated Japanese in Tokyo, Japan (JPT); and 45 unrelated Han Chinese in Beijing, China (CHB). The pilot phase targeted to genotype 600,000 SNPs spaced at approximately one per 5-kilobase intervals. As large-scale genotyping technology was relatively new at the time, and the accuracy was of concern, a locus had to have minor allele frequency (MAF) of more than 5% in at least one population in order to pass the quality control. Today, ~4M SNPs have been genotyped in the four populations. The MAF requirement of at least 5% has also been removed as the genotyping platforms improved over the years. Seven additional populations have been included, with 1.6 million SNPs genotyped in each population through Affymetrix or Illumina genotyping Chip assays (Altshuler, Gibbs et al. 2010). The number of SNPs covered by HAPMAP is currently restricted by the technology used, and the 1000 Genomes Project using whole genome

sequencing platforms sought to carry on and identify more than 95% of the variants in the human genome from different ancestral populations. These variants in the 1000 Genomes Project included low frequency and rare variants, whose minor allele frequencies were 0.5%-5% and less than 0.5%, respectively, that were previously not well-covered (Durbin, Abecasis et al. 2010). Information pertaining to allele frequencies, genotype and haplotype of known and unknown variants discovered was catalogued, and is available to the public (http://www.1000genomes.org/home).

By far, the most important and thorough large-scale SNP catalogue is the NCBI dbSNP (Sayers, Barrett et al. 2011). The NCBI dbSNP provides SNP allele and genotype frequencies reported by HAPMAP and the 1000 Genomes Project, as well as those submitted by individual researchers in their population of interest. Information submitted by individual researchers may help to cross-validate the information provided by HAPMAP and the 1000 Genomes Project, as well as provide allele information in populations not covered. It also provides the genomic location and gene context of the SNP, which offers some basic functional effect of the SNP (e.g., the SNP is a non-synonymous SNP). Therefore, the NCBI dbSNP can act as a useful starting point for researchers to mine for SNP-related information. One important observation, however, is that the NCBI dbSNP database is very dynamic, due to the ever increasing number of entries, as well as constant updating from other data sources on which it depends, such as RefSeq and Genome Build. Consequently, changes to information provided by the NCBI dbSNP can create confusion, even for users that are familiar with it. For instance, the assignment of different reference cluster numbers (rsNo, the unique identifier of SNP in dbSNP) to the same genomic location adds confusion. Another example would be mapping of a SNP to a different gene location due to different genome coordinates assigned in a new Genome Build entry or due to some updates in

the gene reference sequence (e.g., change of exon length). Although these types of changes are infrequent, they can be problematic if researchers use different builds of the human genome at different points in time for their work.

1.1.4.2 Computational methods and resources to facilitate the identification of functional SNPs

As the number of SNPs documented in the databases continues to grow, finding functional SNPs in the human genome is like trying to "find a needle in a haystack". Fortunately, a number of computational methods and resources have been developed recently to help pinpoint the potentially functional SNPs (pfSNP) that are likely to be causal due to their known biological plausibility.

1.1.4.2.1 Literature-reported "potentially functional SNPs"

Literature-reported functional SNPs in related or even apparently unrelated phenotypes may affect the phenotype to be studied due to a phenomenon called "pleiotropy". This is because the SNP may affect a gene that has multiple molecular functions, or alternatively, the single molecular function of the gene may be shared by different phenotypes (Becker 2004). Some classical examples of "pleiotropy" can be found in immune-related diseases (Zhernakova, van Diemen et al. 2009; Lees, Barrett et al. 2011) and a recent report (Sivakumaran, Agakov et al. 2011) provided evidence that 233 (16.9%) genes and 77 (4.6%) SNPs causing human disease show pleiotrophic effect.

Therefore, mining literature for previously reported SNPs implicated in related

traits or traits that share a common pathway may aid in pinpointing the causal SNP. However, this is not a trivial task. Although natural language processing has been proposed as a promising way to help the mining process, there are a few obstacles to be overcome before this will be successfully implemented. The first is the non-standardized nomenclature used for SNP reporting. Since non-synonymous SNPs were the primary focus in the past, a SNP in the literature is usually referred to by the amino acid change it caused; and citing the reference SNP number (rsNo) assigned by NCBI dbSNP was not a common practice until recently. SNPs were also referred to as an "allele" of the gene, and extensive domain knowledge is required to translate the allele number mentioned in the literature into a corresponding SNP, if possible. Furthermore, the nomenclature used in SNP reporting is not standardized, and it is very difficult to create an accurate rule to sift through the literature (Xuan, Wang et al. 2007). Therefore, LSDB and other general purpose databases relying on extensive manual compilation, such as MutaGeneSys (Stoyanovich and Pe'er 2008) and the Human Gene Mutation Database (HGMD, http://www.hgmd.cf.ac.uk/ac/index.php), are still the preferred resources for finding literature reported functional SNPs. Manual curation is also used for summarizing GWAS results currently available. The Genetic Association Database (GAD) (http://geneticassociationdb.nih.gov/) and the NHGRI 'A Catalog of Published Genome-Wide Association Studies' (http://www.genome.gov/gwastudies) are maintained by curators rather than automated data integration pipelines.

1.1.4.2.2 Predicted "potentially functional SNPs"

As our knowledge about SNP function accumulates, various methods have

been proposed to predict SNP functions. I broadly classified these methods into two categories, namely "SNP character-based method" and "Motif-based method", according to the underlying mechanisms used to derive the predictions in each method. There are also methods that use both approaches, forming a third "Combined method" category. **Table 1.3** lists some of the methods commonly found in the literature.

Table 1.3: List of the methods used for SNP function prediction and some of the tools available.

Method Class	SNP Category	Function Category	Method Short Description	Reference
SNP Character-Based	Coding	Cause NMD	Nonsense Mediated Decay	(Nagy and Maquat 1998)
	Coding and Intronic	Disrupts Splice Site	Cause aberrant 5' site (SD-Score)	(Sahashi, Masuda et al. 2007)
Motif- Based	Promoter	Change TF Binding sites	TransFac etc.	
	Coding	Change ESE/ESS Site	RESCUE-ESE etc.	(Fairbrother, Yeh et al. 2002; Cartegni, Wang et al. 2003; Zhang and Chasin 2004; Zheng 2004)
	Intronic	Change ISRE sites	Conserved intronic sequence within 400bps of splicing site	(Yeo, Nostrand et al. 2007)
	3UTR	Change Mir Binding	MirRanda etc.	(Krek, Grun et al. 2005; Grimson, Farh et al. 2007; Betel, Wilson et al. 2008)
		In 3UTR Conserved Region	Functional region predicted by multiple species sequence alignment	(Xie, Lu et al. 2005)
	Coding	Protein Domain / Functional Sites	Interpro Scan	(Quevillon, Silventoinen et al. 2005)
Combined	Non-synonymous	Change functional motif	NetPhos etc.	(Hansen, Lund et al. 1998; Blom, Gammeltoft et al. 1999)
		Non-Synonymous SNP function based on multiple criteria	Polyphen etc.	(Ramensky, Bork et al. 2002; Ng and Henikoff 2003; Thomas and Kejariwal 2004; Yue, Melamud et al. 2006)

The "SNP character-based method" is based on the character of the SNP, such as the amino acid change caused. One example of such methods is the empirical rule to identify SNPs leading to the degradation of mRNA (nonsense-mediated decay, NMD) by generating a premature stop codon in certain coding region (Nagy and Maquat 1998). Another example is the SD score (Sahashi, Masuda et al. 2007), which is the precise formula for predicting disruption effects of the SNPs found at the 5' Intron-Exon boundary. It's noteworthy that methods strictly belonging to this category are scarce, because SNPs usually function in the context of a "motif", and, consequently, the "Combined method" is usually more useful. This can be illustrated by the SNPs causing phosphorylation site changes, where a single SNP changing the amino acid would be functional, only if such residue is in a phosphorylation motif.

Compared to "SNP character-based" methods, "Motif-based" methods are more abundant and can be used to find pfSNP in all gene regions. However, most of them were not developed to find pfSNP directly and extra efforts are needed to adapt them for pfSNP screening. Therefore, this section will cover some of the extension efforts and review their strengths and weaknesses.

Promoter SNPs may change transcription factor binding sites (TF binding sites), and their effects on gene expression are well-established (Chorley, Wang et al. 2008). The methods currently available – TRANSFAC Gene Transcription Factor database http://www.biobase-international.com/pages/index.php?id=transfac and MatInspector http://www.genomatix.de/online_help/help_matinspector/matinspector_help.html – largely focus on predicting transcription factor binding sites in a given DNA sequence. These tools scan the input DNA sequence against a library of Positional Weight Matrices (PWM) for different transcription factors curated from the literature.

A hit will be reported if the similar sequence gets a score above a general or a tissue-specific threshold. Since these methods do not predict the effect of the SNP on the TF binding site, Kim et al. (Kim, Kim et al. 2008) scanned the entire human genome for possible TF binding sites, and regarded all the SNPs residing within such sites as important. This method suffers from a few drawbacks. Firstly, it ignores the fact that TF binding sites are highly degenerate, and a single nucleotide change may have a limited impact on the binding affinity (Chorley, Wang et al. 2008). Second, this approach ignores SNPs that create novel TF binding sites. Therefore, there is a need to develop a tool to compare and contrast the TF binding site scanning results for different SNP alleles.

The 3' UTR SNPs were recently in the limelight because of the newly discovered microRNA regulation mechanism implicated in the mRNA translation process. Similar to TF binding site prediction, a series of programs, such as MirRanda (Betel, Wilson et al. 2008), TargetScanS (Grimson, Farh et al. 2007) and PicTar (Krek, Grun et al. 2005) can predict miRNA binding sites in a given nucleotide sequence. The Patrocles database (Georges, Clop et al. 2006) used MirRanda to scan for miRNA binding sites unique to the original and alternative allele of the SNP and nicely summarized the SNPs causing differences to miRNA binding sites. Regions of conservation in the 3'UTR can also be used to identify functional motifs, a number of which have been identified by Xie et al. (Xie, Lu et al. 2005).

Other than the SNPs on the miRNA target sites that would affect the efficacy of miRNA, SNPs in the coding region, especially the seeding sequence of miRNA, would lead to reduced or even abolished miRNA binding to its designated target. The SNP may also lead the miRNA to target sequences which are not supposed to be regulated and create an "off target" effect. At the time of the project started, no

database focuses on this group of SNPs, and a SNP mapper to identify these SNPs would be highly valuable for this task.

In addition to these well-established mechanisms and their associated tools, differential splicing of mRNA has been recently recognized as the next layer of complexity in gene regulation, which may, in part, explain the much smaller human gene pool as it was once thought to be. Intron-Exon boundaries were thought to be the key motif facilitating exon recognition and the conserved sequences were identified (Zhang 1998). It was also proposed that exon sequences hosted other motifs, which may enhance or suppress the splicing of the particular exon by exonic splicing enhancers (ESE) and exonic splicing silencers (ESS), respectively (Fairbrother, Yeh et al. 2002; Cartegni, Wang et al. 2003; Zhang and Chasin 2004; Zheng 2004). Yeo et al. (Yeo, Nostrand et al. 2007) later proposed that the 400 base pairs of intron flanking the exon-intron junction may also contain motifs that either enhance or silence the splicing process, and these motifs were termed intronic splicing regulatory elements (ISRE). Similar to the TF binding site prediction tools, a number of tools are available for identifying these motifs in a given sequence (Fairbrother, Yeh et al. 2002; Cartegni, Wang et al. 2003; Zhang and Chasin 2004; Zheng 2004). However, they are not designed to compare and contrast the motif scanning results for different SNP alleles.

The “Combined method” is largely exclusive to non-synonymous SNPs. It can be further divided into two sub-categories. The first contains methods, such as NetPhos (Blom, Gammeltoft et al. 1999) and NetOGlyc (Hansen, Lund et al. 1998), which use relative simpler “SNP character-based” rules to determine the SNP functional impact. The second sub-category contains methods using more complex rules and can be further divided into two sub-groups. The first tries to predict the

deleterious potential of a SNP by looking at a group of criteria, such as (1) whether it falls in an annotated active motif or binding site; (2) whether it affects the interaction with ligands present in the crystallographic structure; (3) whether it leads to hydrophobicity or electrostatic charge changes in a buried site; (4) whether it destroys a disulphide bond; (5) whether it affects the protein's solubility; (6) whether it inserts a proline in an α-helix; or (6) whether it is incompatible with the profile of amino acid substitutions observed at this site in the set of homologous proteins. PolyPhen (Ramensky, Bork et al. 2002) uses this method, and regards a SNP "deleterious" if it satisfies some empirical rules summarized from current knowledge. Later, LS-SNP (Karchin, Diekhans et al. 2005) and SNP3D (Yue, Melamud et al. 2006) employed machine learning methods to identify the proper threshold from a training data set rather than relying on a human-defined threshold. The second sub-group looks for SNPs in evolutionarily conserved regions amongst the same protein family members. It is used by SIFT (Ng and Henikoff 2003) and Panther SNP (Thomas and Kejariwal 2004). The difference between these two tools is that SIFT uses sequence similarity searches to identify potential family members, while Panther SNP relies on their expert-curated database for protein sub-family members.

1.1.4.2.3 Inferred "potentially functional SNPs"

As with SIFT and Panther SNP, which suggest that evolutionary history may help to identify non-synonymous SNPs with deleterious effects, other methods using the same principle (which is to find a SNP with evolutionary evidence of functionality) have been developed for other genic and/or inter-genic regions. Unlike non-synonymous SNP, these SNPs usually lack any known biological function, and

are therefore termed "inferred functional" SNPs.

A widely used method is to look for SNPs in the ultra-conserved regions under the assumption that these regions are conserved for a reason (Lee and Shatkay 2008). The pre-calculated 28-way vertebrate multiple sequence alignment (Miller, Rosenbloom et al. 2007) by UCSC genome browser (Rhead, Karolchik et al. 2010) is usually used for this purpose. There is also a related concept, accelerated conserved non-coding ("ACNC") region, proposed by several authors (Pollard, Salama et al. 2006; Prabhakar, Noonan et al. 2006; Kim and Pritchard 2007; Bush and Lahn 2008). These regions are conserved among multiple species with the exception of humans, and are thought to host elements crucial for human speciation. These papers do not list SNPs within such regions; as such, a SNP mapper is also needed here.

Similarly, evolutionary forces may have left some mark on the human genome. Sabati et al. suggests that regions with a signature of recent positive selection should be functional. Our lab has extended this method further to identify the SNP with a signature of recent positive selection (Tang, Wong et al. 2004; Wang, Wang et al. 2005). The signature refers to the higher than usual haplotype conservation across a relatively long distance. Before we look at how such a signature can be found, we would first need to understand a group of related statistics to which such signatures are heavily related. They are Haplotype Homozygosity (HH), a derived term called Extended Haplotype Homozygosity (EHH) as well as a further derived term called Relative Extended Haplotype Homozygosity (rEHH). HH is a statistic to measure the likelihood of getting a pair of identical haplotypes randomly drawn from a pool (Sabatti and Risch 2002). EHH is the HH for a pool of haplotypes all containing one allele of a SNP at a specific distance away from the SNP locus. The rEHH marks a comparison of the EHH values between 2 alleles of the SNP, and is obtained simply

by dividing the EHH of one allele by the EHH of the alternative allele.

A signature of recent positive selection in SNPs can be illustrated by this example. When a SNP first appears as a new mutation in the genome, a new haplotype comprising this mutational allele together with alleles of other existing SNPs on the same chromosome is created. Assuming that the new allele is strongly selected for, and a large number of copies of this haplotype were passed down intact to the next generation because homologous recombination has not happen due to the short time frame. If we measure EHH for this mutational allele, we would observe that it remains as one throughout the whole chromosome because the haplotype pool from which we draw comprises solely the intact copies. On the contrary, the ancestral allele has existed on the genome for a long time, and extensive crossing over should have occurred. As the likelihood of crossing over increases with length, the EHH for the ancestral allele would gradually decrease from 1 to 0 along the chromosome length. Therefore, if we observe rEHH values for the mutational allele along the chromosome length, we should see that it continue to rise from 1 to infinity. For the simplicity of calculation, we usually choose to look at the rEHH value at one specific distance, and a high rEHH value for one allele is the signature of recent positive selection. Random genetic drift could not produce such signatures because a mutation under random genetic drift would take a long time to become a SNP, and crossing over should have happened extensively to the mutational allele as well. The rEHH value would therefore remain low for both alleles. Recombination hot/cold spot would not affect the signature, since such a disturbance would equally affect both of the alleles and get cancelled out when rEHH values were calculated. The only factor that may produce a similar signature besides the selective force is population demographic history. This factor can be accounted for by showing that the SNP of interest

demonstrates a significantly higher rEHH value than its simulated counterpart under such a scenario.

Besides conservation-based methods, genotype heterozygosity-based methods have also been proposed. Akey et al. (Akey, Zhang et al. 2002) inspired by the Wright's fixation index F_{st} (a measure of population subdivision) proposed that SNPs showing a high F_{st} value should be driven by the selective force exerted during the formation of different human races. More specifically, since modern human races all come from a common ancestor, the genotype heterozygosity for a SNP should be the same when each subpopulation (race) and the entire human population as a whole are examined. The authors suggest that any discrepancy from this would signify a selection on such a SNP during human race formation. Although the same phenomenon may be caused by random genetic drift, the authors argue that such events would affect all the SNPs on a genomic scale, and can be ruled out by setting an empirical threshold based on the distribution of F_{st} values for all of the SNPs. The authors proceeded to calculate the F_{st} values for more than 26,000 SNPs in populations of African American, East Asian and European American descents, and SNPs with F_{st} values within the top 5^{th} percentile were declared to be functional (Akey, Zhang et al. 2002).

Since Akey et al. published their results, the HapMap project released a more densely populated SNP genotype dataset and enabled us to have a closer look at F_{st} distributions genome-wide. However, a fast F_{st} calculator would be needed for handling such data.

1.2 Overview of genetic association studies

Since genetic association studies are becoming the preferred starting point for functional SNP discovery, this section is designated to review the basics of these studies and to highlight the challenges they currently face.

1.2.1 Candidate gene-based and genome-wide association studies (GBAS and GWAS)

Candidate gene-based association studies (GBAS) focus on SNP markers in a single or a few candidate genes that have been identified from prior knowledge. These candidate genes can be those reported to play certain roles in pathways related to a certain phenotype, previously implicated in a certain phenotype having similar characteristics, or directly linked to the phenotype of interest (e.g., Inflammatory Bowel Disease and Colon Cancer). This approach is ideal if the phenotype were a Mendelian trait, with extensive prior knowledge.

In candidate gene-based association studies, a literature review is normally carried out to identify the gene of interest together with those previously reported associated SNP markers. Currently, there are a number of online knowledge bases that can facilitate accomplishing this task, such as PharmGKB (Thorn, Klein et al. 2010). PharmGKB is a useful web-based resource for researchers interested in the genetic basis of drug response or pharmacogenomics. It is a repository containing previously published associations of drug response genes with various drug response phenotypes. As the drug response phenotype is the consequence of the pharmacokinetic and pharmacodynamic effects of a drug, information about drug response genes in

PharmGKB is placed within the context of the pharmacokinetic and pharmacodynamic pathways of various drugs. Structuring information according to pathways that the drug response gene affects enables researchers to visualize the effect of a SNP as well as possible SNP-SNP interactions within a pharmacologically and biologically meaningful context. Besides application-specific PharmGKB, the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Aoki and Kanehisa 2005), Gene Ontology (GO) (Ashburner, Ball et al. 2000) and Molecular Signatures Database (mSigDB) C2 gene sets (Liberzon, Subramanian et al. 2011) are the general purpose repositories frequently consulted for inferring which genes are related to the phenotype of interest. KEGG attempts to classify genes into different functional pathways. GO, on the other hand, uses controlled vocabulary – a set of preselected, defined, authorized terms – to describe gene function and location. mSigDB C2 gene sets use extensive curation to collect gene information from various sources, such as online pathway databases, publications in PubMed, and knowledge garnered from domain experts.

If the phenotype is complex and multiple genes in different pathways can contribute or there is no prior knowledge existing for the particular phenotype, researchers can use genome-wide association studies to discover causal variants across the whole genome. To perform a genome wide association study (GWAS), there is no need to do a literature search. If whole genome sequencing were used, all the SNPs in the genome would be genotyped and analysed. If SNP genotyping Chips were used, researcher would have little, if any, freedom to choose markers to genotype, due to the limited availability of genotyping Chip suppliers (McCarthy, Abecasis et al. 2008).

1.2.2 Linkage disequilibrium and its implications in genetic association studies

One of their ultimate goals for HapMap and 1000 Genomes Project is to facilitate genetic association study by identifying the human genome architecture in terms of SNP haplotype structure and linkage disequilibrium pattern. This section will discuss the impact of linkage disequilibrium on genetic association studies.

1.2.2.1 Linkage disequilibrium

Although "linkage disequilibrium" (LD) seems to be related to "linkage", the terms describe different phenomenon. "Linkage" is used to describe the tendency of genes to be inherited together during meiosis due to their proximity. "Linkage disequilibrium" is defined as the non-random association of alleles at two or more loci. Formally, LD can be measured by two statistics, namely $|D'|$ and r^2. Both are based on the Lewontin's D, which is the deviation of the observed frequency of a haplotype from the expected. $|D'|$ is obtained by dividing D over D_{max}, the maximum D possible for a given set of allelic frequencies at any two loci. $|D'|$ ranges from 0 to 1, where 0 signifies complete linkage equilibrium and 1 represents complete linkage disequilibrium (Zondervan and Cardon 2004). When there is no recombination between two markers, $|D'|$ will be 1.0. $|D'|$ is sensitive to allele frequencies and can be inflated when one SNP in the pair has low minor allele frequency while the other has high minor allele frequency (Zondervan and Cardon 2004). Therefore, $|D'|$ is only a good measure of recombination rate between two SNPs. However, knowing the genotype of one SNP may not enable us to predict the genotype of another SNP with a perfect $|D'|$ value. In contrast, r^2 is a measure of the correlation of alleles at two loci.

Like |*D*'|, it can take a value from 0 to 1. r^2 = 0 means perfect linkage equilibrium and r^2 = 1 stands for perfect linkage disequilibrium. r^2 will only be 1 when every occurrence of an allele at each of the markers perfectly predicts the allele at the other locus (Zondervan and Cardon 2004).

1.2.2.2 Linkage disequilibrium in genetic association studies

For a group of SNPs with r^2 = 1, only one SNP in the group needs to be genotyped for association analysis because knowing its genotype is equivalent to knowing all the other genotypes across the loci. Even for SNPs with lower r^2 values, it is possible to detect an association between the marker SNP and phenotype by increasing the sample size by a factor of $1/r^2$ over the number required when the effect of the causative SNP is measured directly (Kruglyak 1999; Pickrell, Clerget-Darpoux et al. 2007). Therefore, r^2 is widely used in association studies to reduce the number of markers to be examined. It is worthwhile to note that this strategy would only work for common SNPs (MAF >5%) that tend to have a higher r^2 with other common SNPs. Surrogate markers do not adequately represent rare SNPs (MAF<5%) because they generally have low r^2 values with both common and other rare SNPs (Gorlov, Gorlova et al. 2008; Li and Leal 2008). Hence, tagging SNP strategy described in the next section is applicable only to common SNPs.

When examining the LD structure among SNPs in the genome, Gabriel et al. (Gabriel, Schaffner et al. 2002; Tishkoff and Verrelli 2003) discovered that a large portion of the human genome comprised blocks of LD separated by recombination hotspots. In each LD block, haplotype diversity was found to be low and a few SNPs

could capture most of the diversity. This kind of SNP was thus called haplotype tagging SNPs.

Haplotype-based association studies were first used to define disease-causing regions in the human genome. Within the regions that were significantly associated with the disease phenotype, sequencing would then be performed to identify the disease variant (Kerem, Rommens et al. 1989; Hastbacka, de la Chapelle et al. 1992; Edwards, Ritter et al. 2005; Haines, Hauser et al. 2005; Klein, Zeiss et al. 2005). The LD block structure of the human genome found by Gabriel allowed haplotype tagging SNPs to replace haplotypes for correlating with phenotypes in association studies. However, identification of a haplotype tagging SNP would require knowing the haplotype composition in the region of interest. This is not directly available by using popular genotyping methods and would require a phasing process to generate the most probable haplotype pool in the region of interest by looking at the genotype data. Currently, phasing is normally done with computational intensive programs which implement the Expectation-Maximization algorithm (Excoffier and Slatkin 1995) by themselves (Zhang, Qin et al. 2005) or use phased output (Anderson and Novembre 2003; Avi-Itzhak, Su et al. 2003) from other programs, such as “Phase” (Stephens, Smith et al. 2001) or “SNPHap” https://www-gene.cimr.cam.ac.uk/staff/clayton/software/snphap.txt. In regions with dense markers and relatively low LD, a long time is required to perform such procedures, and it is very likely to have a long list of low frequency haplotypes with a frequency less than 5% (Crawford, Akey et al. 2005). Moreover, SNPs residing in different haplotype blocks may still exhibit high LD (Lai, Bowman et al. 2002). To avoid such shortcomings, methods for generating tagging SNPs based primarily on pair-wise r^2 values without resolving the haplotype were also developed (Ke and

Cardon 2003; Stram, Haiman et al. 2003; Wang and Xu 2003; Weale, Depondt et al. 2003; Carlson, Eberle et al. 2004; Halldorsson, Bafna et al. 2004). There are also tagging SNP selection strategies that rely on principle component analysis, which is a method used to reduce the number of variables (SNPs in this case) by finding correlations between them, which is able to reduce the SNP set to a smaller and more manageable one (Meng, Zaykin et al. 2003; Lin and Altman 2004).

Associating tagging SNPs with a phenotype has practical advantages. Testing every SNP individually for association with the phenotype-of-interest has its limitations, as the causal SNP is often unknown to researchers or not easily genotyped. It is also impractical and expensive to genotype all the SNPs in a particular genomic region in order to identify the causal SNP. By using tagging SNPs, genotyping costs can be reduced, as only the tagging SNP for the group of SNPs in the high LD need to be genotyped. Genotyping fewer SNPs also relieves, to a certain extent, the burden of having to control for multiple testing (Balding 2006), which is discussed later. In association studies using customizable SNP marker panel where the total number of SNPs is fixed, researchers could have the benefit to cover more genetic regions of interest. For researchers studying populations of highly similar ancestry to those in HAPMAP or the 1000 Genomes Project, using the pre-scanned tagging SNPs from the reference populations in HAPMAP or the 1000 Genomes Project would save much effort. However, using tagging SNPs in areas of low LD as a result of many recombination hotspots negates the benefits of using tagging SNPs, as the number of SNPs genotyped is unlikely to be substantially reduced.

Genotyping tagging SNPs only would not necessarily lead to loss of information. Genotyping data for the tagging SNPs could be used to “impute” the genotype of un-genotyped SNPs to enable tests for association (Marchini, Howie et al. 2007; Scott,

Mohlke et al. 2007). HapMap samples can be used to deduce the LD structure between genotyped tagging SNPs and the un-genotyped ones, and such information is used to impute the genotype at the un-genotyped loci. Although this may seem as "sleight-of-hand with no information gained" (Clark and Li 2007), it has been shown that better association results can be obtained (Marchini, Howie et al. 2007; Scott, Mohlke et al. 2007).

1.2.3 SNP selection strategies

In this section, we will discuss why marker selection is needed, the options that are available, how different factors may affect the choice of SNP selection strategy, and how SNP selection strategy may, in turn, affect a follow-up study.

1.2.3.1 Rationale for SNP selection: Multiple-testing problem and power

Marker selection is a crucial process not only because fewer markers would cost less to genotype but also due to the problem of multiple-testing. Each statistical test has a false-positive rate equivalent to the alpha value set. If the alpha-value were set to 0.05 for a series of 20 tests to establish an association between different SNP markers and one phenotype, and only one of them have a P-value marginally less than 0.05, we would have difficulty in justifying the validity or significance of the result simply because we would expect one false positive out of the 20 simultaneous tests. The above-mentioned scenario illustrates the multiple-testing problem.

One way to solve this problem is Bonferroni correction, the simplest but most conservative method. In the Bonferroni correction, the P-value cut-off is set to be the

alpha value divided by number of tests carried out. For example, in a small scale gene-based association study where 5 SNPs were tested for association with the phenotype, a threshold of 0.01 would be needed (assuming alpha to be 0.05). As gene-based associations now focus on genes in the whole pathway and a dozen of the SNPs in these genes can be tested, it is increasingly difficult to detect markers with statistical significance. In the case of GWAS, the gold standard is to set the threshold to 5 E-8 (Khoury and Yang 1998; Hoggart, Clark et al. 2008), and this creates a lot of debate on whether Bonferroni correction should be used (Zhang, Liu et al. 2007; McCarthy, Abecasis et al. 2008). There are other correction methods proposed to work by controlling the false discovery rate (FDR); for example, the Benjamini-Hochberg Procedure (Benjamini and Hochberg 1995) and Q-value (Hoggart, Clark et al. 2008). Benjamini-Hochberg procedure ranks the P-values in ascending order and subjects each P-value to a cut-off as the product of the rank multiplied by the Bonferroni corrected cut-off. In other words, only the smallest P value need to pass the most stringent Bonferroni correction while others would be subjected to a less stringent (higher) cut-off proportional to its rank. The Q-value method is based on the theory that all P-values should distribute equally from 0 to 1 if no association exists, and it declares markers deviating from the theoretical distribution as a true association. However, such methods only moderate the dropping of the threshold, and it would still be challenging to establish an association when tens of thousands of markers are to be tested.

As can be seen, multiple-testing corrections would inevitably lead to a loss of power for an association study. This loss can be compensated for by increasing the sample size (McCarthy, Abecasis et al. 2008) and it is now a common practice to recruit controls in the thousands for GWAS in order to accommodate the stringent

threshold (5 E-8). This, however, adds extra costs in patient recruitment as well as in genotyping, and has become a limiting factor in the wider adoption of association studies.

1.2.3.2 SNP selection strategies: Direct association and indirect association

In SNP selection, we have the choice of including the SNPs potentially causing such differences in phenotypes or targeting the surrogate SNPs which are in LD with the causal ones (Cordell and Clayton 2005). We call the first strategy "direct association" and the second "indirect association".

1.2.3.3 Factors affecting the choice of SNP selection strategy

A number of factors may affect the choice of SNP selection strategy:

1. Choice of the region covered
2. The genetic basis of the phenotype
3. The LD structure in the human genome

The choice of region covered is an important factor. For gene-based association studies, "direct association" by targeting non-synonymous SNPs has been the preferred choice. This is because non-synonymous SNPs were thought to be the most important ones in affecting gene function and other SNPs were largely ignored. In recent years, SNPs affecting gene expression are also getting attention, since gene expression level is another key factor determining the overall gene function (Chorley, Wang et al. 2008).

For GWAS, the SNP selection strategy used is further dependent on the other

two factors: the genetic basis of the phenotype, and the LD structure in the human genome. There are different models regarding the genetic basis of common disease and human phenotype in general. The "Common Disease, Common Variant" (CDCV) model (Reich and Lander 2001; Wang, Barratt et al. 2005) hypothesizes that common disease is caused by the same variants that are widely distributed in the population, with limited penetrance. Since the causal SNPs are those common ones which tend to have a higher LD with other common and non-functional SNPs (Montgomery 2011), targeting any SNP in the LD block other than the causal one would result in limited power loss (Pickrell, Clerget-Darpoux et al. 2007). Under the proportionality assumption, it is possible to detect an association between the marker SNP and phenotype by increasing the sample size by a factor of $1/r^2$ over the number required when the effect of the functionally significant SNP is measured directly (Kruglyak 1999; Pickrell, Clerget-Darpoux et al. 2007). It was proposed that ~500K well-spaced markers would sufficiently cover the whole human genome (Collins, Guyer et al. 1997; Reich and Lander 2001), and this forms the basis of the "indirect association" approach, which tries to examine surrogate SNPs that are linked to the causal ones (Cordell and Clayton 2005). In current GWAS using DNA Chips, the most popular approach is the "indirect association" approach (Jorgenson and Witte 2006). Affymetrix and Illumina are the two most popular genotyping technologies for GWAS and their latest chips can interrogate ~1,000,000 SNPs concurrently. SNPs on the Illumina platform are primarily selected based on tagging SNPs (t-SNPs) that exceed a predetermined linkage disequilibrium (LD) threshold specified by the International HapMap Consortium (Perkel 2008). On the other hand, SNPs in Affymetrix chips are quasi-randomly selected (QR-SNPs) to cover the entire genome, although their selection of SNPs are also limited to SNPs that can accommodate

sequence constraints imposed by the Affymetrix technology (Perkel 2008).

However, the proportionality assumption (the possibility to detect an association between the marker SNP and phenotype by increasing the sample size by a factor of $1/r^2$ over the number required when the effect of the functionally significant SNP is measured directly) has been criticized as an over-simplification of the many factors influencing the strength of the marker (Terwilliger and Hiekkalinna 2006). Additionally, the usefulness of tagging SNPs to act as good surrogate markers for functional or causal SNPs has been challenged, since the presence of two or more susceptibility loci, which is assumed in complex disease genetics, may significantly impact the power of tagging SNP-based strategies (Terwilliger and Hiekkalinna 2006; Pickrell, Clerget-Darpoux et al. 2007). It has been suggested that, in some situations, although a causal SNP can unequivocally be detected when it is directly genotyped, marker SNPs in high LD with the causal SNP may never show sufficient evidence of disease-association even with infinite sample size (Terwilliger and Hiekkalinna 2006). As a result, a trimmed-down version of the "direct association" approach, termed "gene-centric" approach, has also been used in the GWAS setting (Keating, Tischfield et al. 2008; Webb, Broderick et al. 2009). These "gene-centric" studies either focus on SNPs that are capable of representing variants in genes or focus on the putative causal variants directly (Jorgenson and Witte 2006). The advantages of the "gene-centric" approach include decreasing the genotyping and Type I error burden, since SNPs within genes have a higher likelihood of being functionally important. Additionally, as it was reported that SNPs within many genes seem to have a lower LD, with SNPs residing outside genes (Smith, Thomas et al. 2005), the "indirect association" approach may not provide adequate coverage for genic regions. Conversely, the "gene-centric approach" suffers the disadvantage of having less

power to detect non-genic causal SNPs as compared with the "indirect approach".

The "Common Disease, Rare Variant" (CDRV) model attributes the common disease to different variants that are each unique to one of a few patients with high penetrance (Gorlov, Gorlova et al. 2008; Li and Leal 2008). In the case of the CDRV model, the "direct association" approach fits better because rare SNPs (MAF<5%) are generally low in LD with all other SNPs and "indirect association" would be insufficient to readily capture them (Pritchard and Cox 2002; Zeggini, Rayner et al. 2005; Montgomery 2011). The CDRV model is now gaining popularity, as previous GWAS studies relying on the CDCV model had difficulty in finding causal markers for a number of phenotypes (Montgomery 2011).

1.2.4 Methods of association analysis

Single marker-based association analysis is still the most widely used data analysis method in candidate gene-based association studies and GWAS. P-values for every marker are calculated as the measurement for strength of association. If no assumption about the underlying genetic model is made, genotype-based tests can be used with the 2-df Pearson test or Fisher's exact test focusing on a 2X3 matrices tabulating the count of different genotypes in case and control. If an "additive" model is assumed, the allele-based test or the Cochran-Armitage test can be used. The allele-based test can take the format of 1-df fisher's exact / permutation test, with assumption of HWE in the case and control population as a whole. Cochran-Armitage test has the advantage of independence on the assumption of HWE. The 1-df permutation test would be more advantageous by being less conservative as compared with the Cochran-Armitage test. Currently, single marker-based analysis methods

suffer from the stringent cut-off set to tackle the multiple testing problems caused by the ever-growing number of markers tested. The stringent cut-off makes it difficult to find markers with moderate-effect size. Moreover, even if such markers would survive after the weeding process, the custom to focus on a few top scoring ones would still leave them out for further analysis.

One way to enhance the power of detecting a moderate-effect size marker is to look at the combined effects of multiple SNPs. Haplotype-based association analyses can help to capture the combined effects of tightly linked *cis*-acting SNPs. It is only applicable to "direct association" studies, because "indirect association" studies normally use tagging SNPs with low pair-wise LD between each other (Balding 2006). "Indirect association" studies use regression-based methods to measure the combined effects, and these methods have the advantage of looking at multiple SNPs, even if they are not on the same chromosome. Furthermore, they have the added advantage of being able to take other clinical parameters into consideration together with SNPs. Such regression-based methods can also be used by "direct association" with one extra step of finding tagging SNPs, since too many correlated parameters would render the regression less effective (Balding 2006).

However, even combined effects may not have enough power, since the addictive effect may not be strong enough to produce a small P-value to endure the multiple testing corrections. It has been suggested a gene set-/pathway-based analysis may better suit the need to recover more markers with moderate-effect size. The principle of these methods is that the group of truly associated genes must show consistent yet moderate deviation from chance and such consistent changes would not be visible when looking at the SNPs individually. These methods also tackle multiple testing problems by looking at gene set/pathway and reduce the number of tests

carried out from millions to just a few.

The deviation from the "no enrichment of associated SNP in a particular gene set/pathway" null hypothesis can be detected either by comparing against expected values under the null hypothesis, which is called the "self-contained" method, or measured against other gene sets or pathways, which form the basis of "competitive" methods. For candidate gene-based association studies, the "self-contained" method is the better choice, simply because genes outside the pathway of interest are inadequately covered (Wang, Li et al. 2010). For GWAS, both methods can be used. However, there is no guideline on whether the "competitive" or "self-contained" method should be used in terms of performance, and even the "competitive" method has several implementations using different statistical methods (Wang, Li et al. 2010). The choice of a better fit is still on a trial-and-error basis.

1.3 Aims of the current study

While pfSNPs can be broadly classified as "literature-reported functional" (p12) , "predicted functional" (p13) and "inferred functional" (p18) based on the computational methods and resources used to identify them, the methods available to identify them were quite limited to the "predicted functional" category and new methods to identify pfSNPs from the other two categories were needed. Integrating results from these diverse methods via a semi-automated pipeline is required so that a pfSNP database at the genome scale can be made possible. Once such a database is built, a web portal will need to be established to facilitate the scientific community to tap into the pfSNP dataset for various applications. In the meantime, the usefulness of the pfSNP dataset and the web portal in finding real causal SNPs will be validated. Formally,

the three major aims for the current study can be summarized as below:

Aim 1: Identification and characterization of pfSNPs in the human genome.

Aim 2: Development of a pfSNP web-resource.

Aim 3: Validating the usefulness of pfSNPs in association study

This thesis aims to meet each of the above mentioned aims sequentially, and the thesis is structured accordingly. This section will briefly describe each of the aims together with the specific aims/deliverables to give an overview of the content of the thesis, and the following chapters will present the deliverables achieved for each aim together with detailed materials and methods.

1.3.1 Aim 1: Identification and characterization of pfSNPs in the human genome

Chapter Two will be devoted to the process of identification and characterization of pfSNPs in the human genome. I will first review the gaps in the existing tools and the proposed solutions to these gaps including some new methods proposed by me. I will then define the general workflow of the pfSNP identification process, the architecture of the pfSNP data warehouse, as well as the design of three specific tools needed to build a semi-automated pipeline to facilitate the extraction, transformation and loading of the pfSNP dataset into this data warehouse. Storing the genome-scale pfSNP dataset into a database can be beneficial in several ways. First, it would enable the integration of all our knowledge for such SNPs and facilitate analysis of the distribution and characteristics of such SNPs. Second, it would enable the easy retrieval of a subset of such pfSNPs for various needs by using customizable

criteria. Finally, it would facilitate the easy extension of the collection when a new resource or tool becomes available.

Once the database has been built, I will characterize the pfSNP dataset in terms of genomic coverage and MAF spectrum etc. and compare these attributes against the SNP marker sets from popular genotyping platforms like Illumina Bead Chip 1M Duo and Affymetrix SNP Chip 6.0. As our knowledge about the human genome is still far from complete, it is acknowledged that the pfSNP set may not thoroughly cover all functional SNPs, and the possibility of using pfSNP as tagging SNPs to cover other common functional variants will thus be explored.

1.3.2 Aim 2: Development of a pfSNP web-resource

Once the pfSNP data set has been collected and characterized, a sophisticated web portal will be required for the scientific community to tap into the information-rich content and various potential applications the pfSNP dataset can offer. Therefore, Chapter Three will be devoted to the design and implementation issues related to the web portal. Specifically, the pfSNP web portal need features to serve the following purposes:

1. *Designing experiments to address the predicted functionality of SNPs*. This web-resource needs to be able to inform the scientist about the predicted functionality of the SNP-of-interest so that appropriate experiments may be designed. It must also provide information of other previous reports that have examined the SNP-of-interest.

2. *Candidate gene-based association studies*. This web-resource will facilitate the selection of potentially functional SNPs in candidate genes for genetic association studies. Significantly, this resource will enable the selection of a subset of pfSNPs in genes within a certain chromosome region, expressed in specific tissues, and/or has been associated with a certain disease/phenotype/drug response/pathway. It should also facilitate the selection of pfSNPs that occur at/above a specific threshold frequency in specific populations or pfSNPs residing only in exons or promoter regions or at 5'/3' UTRs, etc.

3. *Whole-genome association studies (GWAS)*. As discussed earlier, this resource should be made useful to current GWAS scientists, as it will provide information with regard to the distance as well as the LD measured by r^2 of a close by HapMap genotyped pfSNP (SNP predicted to be potentially functional) that is at the highest LD to the genotyped SNP-of interest. Importantly, the integration of gene/pathway level information into the result interface with text clouds highlighting enriched terms in each category will be useful to these researchers who have a list of SNPs that are associated with a phenotype and wish to formulate hypotheses about their findings.

With these applications in mind, I will explain, in detail, how the web portal is architected to suit these needs and the features built to facilitate such applications.

1.3.3 Aim 3: Validating the usefulness of pfSNPs in association study

In Chapter Four, I aim to demonstrate the usefulness of the pfSNP dataset and the web-resource by showing their value in helping to pinpoint the real functional SNPs causing differential response in colorectal cancer patients. I will carry out a gene based association study using pfSNPs. The web resource will be used to choose the pfSNP markers suitable for the phenotype of interest. pfSNPs will be genotyped and tested for any association with the difference in patients' drug response.

In the meantime, it will be interesting to test if tagging SNPs will be able to capture the pfSNPs associated with the differential drug response. Although the "indirect approach" using tagging SNPs are claimed to be able to capture other SNPs adequately via LD, it is still largely based on simulation results focusing on a few genetic regions, and evidence that it would hold true in all regions is still wanting. The lack of power for the "indirect approach" may be more severe in candidate gene-based association studies in which sample sizes are normally small. Previous studies using this strategy tend to fail, and it is still unclear if such failure is caused by the "indirect approach" used. This study will thus provide a first-hand measure of the loss of power by looking at the efficiency of tagging SNPs in capturing the association signals shown by pfSNPs.

CHAPTER 2: IDENTIFICATION AND CHARACTERIZATION OF PFSNPS IN THE HUMAN GENOME

2.1 Introduction

2.1.1 The gaps present in the methods to identify pfSNP at genome scale and the solutions proposed

As mentioned in Chapter One, pfSNPs can be broadly classified as "literature-reported functional", "predicted functional" and "inferred functional" based on the methods used to identify them. However, identification of pfSNPs at genome scale was still limited by one or both of the following gaps:

1. The SNP coverage was not genome scale.
2. Some of the methods used to identify pfSNPs were not accurate and need further enhancement.

Table 2.1 lists the gaps in the methods for each of the pfSNP categories and the solutions proposed by me.

For the "literature-reported functional" category, resources like OMIM and HGMD are largely restricted to gene coding region SNPs. As genome wide association study (GWAS) is getting popular and a large portion of the current GWAS found their association signals outside the gene region (Hindorff, Sethupathy et al. 2009), I proposed including SNPs identified from ever-growing GWAS as pfSNP which may help to extend the SNP coverage of the "literature-reported functional" category outside of the coding region. NCBI dbGaP (Mailman, Feolo et al. 2007), which was introduced in 2007 as a central depot for large scale GWAS studies carried out by NIH institutes, can be a source of well-powered GWAS studies. Since focusing only at the strongest signal in current GWAS may miss markers with moderate effect

Table 2.1: Gaps present in the tools and methods to identify pfSNP at genome scale and the solutions proposed.

pfSNP Category	Gaps in Tools	Gap Detail	Proposed Solution	Publications Using this Idea
Reported	SNP coverage is not genome scale	OMIM and HGMD are limited to gene region SNPs	Collect pfSNPs with P<0.01 from GWAS studies in NCBI dbGaP project	(Li, Wang et al. 2012)
Inferred	SNP coverage is not genome scale	The genome-wide scan for SNPs with high Fst was limited to 26K SNPs	Scan 4 Million SNPs genotyped by HapMap for high Fst SNPs	(Duan, Zhang et al. 2008)
		It is unknown if signature of recent positive selection can be used to identify pfSNPs at genome scale	Show SNPs with such signatures are wide-spread and can be found outside of the coding region	
	Inaccuracy in existing methods	The validity of the claim that "SNPs in the ultra conserved regions are important" is questionable	Proposed that SNPs in the accelerated conserved non-coding ("ACNC") region are important	
Predicted	SNP coverage is not genome scale	Existing tools are limited to SNPs that are non-synonymous	Develop tool to identify SNPs affecting newly discovered biological functions like intronic splicing regulatory element	
	Inaccuracy in existing methods	Declare all SNPs reside in motifs important without proper evaluation of motif disruption and creation caused by SNP	Develop tool to identify SNPs causing such perturbation	
		A single tool may not cover all the pfSNPs due to limited training dataset	Combine results from different algorithms for predicting SNPs affecting same function	(Liu, Jian et al. 2011)

size, I further proposed to use a less stringent P value cut off (1E-2) which may help to salvage SNPs with weaker association signals. The cut-off of 1E-2 is arbitrary and I propose to use it as a clue to guide further experimental validation. This idea was later supported by Li et al (Li, Wang et al. 2012) and their GWASdb used a similar P value cut-off of 1E-3 to identify important SNPs from dbGaP.

For "inferred functional" pfSNPs, existing methods also suffer from limited genome coverage. The genome scan for high Fst SNPs carried out by Akey et al (Akey, Zhang et al. 2002) is limited to 26K SNPs available at that time. I proposed to search for high Fst SNPs amongst the 4 Million SNPs genotyped by HapMap which is 150 times denser than Akey's study. This scan may reveal a more thorough picture of high Fst SNPs in the human genome. Duan et al (Duan, Zhang et al. 2008) carried out such scan and compiled the results into a database. The possibility of using signature of recent positive selection as an indicator of pfSNP at genome scale was still unclear because prior work had only shown that it could be found in a few genes with limited evidence supporting SNP with such signature is functional (Tang, Wong et al. 2004; Wang, Wang et al. 2005). We used HapMap genotype data to show that such signatures are widespread in the ABC transporter family and can be readily found outside of the gene coding region (Wang, Wang et al. 2007). We also provided more evidence that the SNPs with such signature are functional.

One method proposed to identify "inferred functional" pfSNPs may not be valid. SNPs in the ultra-conserved regions were considered important in causing disease (Lee and Shatkay 2008). I argue that the SNPs found in these regions should be functionally neutral rather than deleterious. In theory, these ultra-conserved regions should be under strong purification selection and any deleterious mutation in such regions should have been eliminated. Only functionally neutral mutation in these regions could escape from the purifying selection and become SNP. I proposed another method to identify pfSNPs using conservation related information which is to look for SNPs in the accelerated conserved non-coding ("ACNC") region. It has been shown that there are regions conserved amongst multiple species with the exception of humans, and are thought to host elements crucial for human speciation (Pollard,

Salama et al. 2006; Prabhakar, Noonan et al. 2006; Kim and Pritchard 2007; Bush and Lahn 2008). Unlike ultra-conserved regions, these ACNC regions should be subjected to strong positive selection and any SNP found in these regions might be a result of such selection force. To date, there is no other paper implementing this idea yet and the pfSNPs belonging to this category can only be found in our pfSNP resource.

The coverage of "predicted functional" SNPs is also limited to non-synonymous SNPs in coding region. The tools available (e.g. Polyphen and SIFT) were applicable to non-synonymous SNPs only. To extend the coverage, I proposed to identify "predicted functional" pfSNPs that may affect new gene regulatory mechanisms, such as non-sense mediated mRNA decay (Nagy and Maquat 1998), intronic SNPs regulating gene differential splicing (Yeo, Nostrand et al. 2007) as well as synonymous SNPs causing high codon usage difference (Kimchi-Sarfaty, Oh et al. 2007). To date, pfSNP is the only resource providing "predicted functional" SNPs with such functions.

It is convenient to claim the SNPs mapped into certain DNA motifs (e.g. Transcription factor binding site) are important (Kim, Kim et al. 2008). However, such claim is inaccurate because the SNP may not be changing the motif due to the high degeneracy of the motif. I proposed to compare the motif scan results for each of the alleles and only those SNPs whose alleles leading to a change in motif will be included as pfSNP.

There are also different methods available to identify pfSNPs belonging to the same function category. Because the methods use different information and are based on different training data (e.g: Both PolyPhen (Ramensky, Bork et al. 2002) and SIFT (Ng and Henikoff 2003) are tools trying to identify non-synonymous SNPs with deleterious effect. PolyPhen is mainly based on whether the SNP will change the

biochemical property of the protein motif while SIFT is based on whether the SNP is conserved.), combining the results of these methods may ensure a more thorough coverage. I proposed to use five tools, namely Polyphen (Ramensky, Bork et al. 2002), SIFT (Ng and Henikoff 2003), SNP3D (Yue, Melamud et al. 2006), LS-SNP (Karchin, Diekhans et al. 2005) and PantherSNP (Thomas and Kejariwal 2004), for predicting deleterious non-synonymous SNPs. A recent resource dbNSNP (Liu, Jian et al. 2011) which tried to identify all the deleterious non-synonymous SNPs used the same idea despite different tools were used.

2.1.2 Challenges in building genome scale pfSNP database

With my proposed solutions to existing gaps, a genome-scale pfSNP database would be made possible. The pfSNP database would be beneficial in three ways. First, it would enable integration of all pfSNP knowledge and facilitate the analysis of their distribution and characteristics. Second, through the use of customizable criteria, it would enable the easy retrieval of a certain subset of pfSNPs. Third, it would be expandable when new resources or tools are available.

Table 2.2 lists the tools that can be used for pfSNP identification. However, there were still technical challenges that stop us from tapping into them. The technical challenges can be broadly classified into three categories, namely:

1. Lack of suitable tools for information integration.
2. Heterogeneous tools with complex input data requirement.
3. No cross referencing and cross checking of information available and possible erroneous information provided.

Table 2.2: List of tools and methods that can be used to identify pfSNPs genome wide.

Molecule Level	SNP Cat.	Function Cat.	Tools Available	Input Requirement	URL for Tool	Remarks	References
SNP	All SNP	Under Recent Positive Selection	1. Haplotter		http://haplotter.uchicago.edu/selection/	Provides iHS, Tajima's D as well as Fay and Wu's H.	BF Voight et al, Plos Biology, 2006, 4:446
			2. WGLRH		--	No standalone web resource. However, pfSNP has results from this algorism.	C Zhang et al, Bioinformatics, 2006, 22:2122
			3. LDD		--	No standalone web resource. However, pfSNP has results from this algorism.	ET Wang et al, PNAS, 2006, 103:135
			4. SNP@Evolution		http://bighapmap.big.ac.cn/	Provides Fst, iHS and Heterozygosity Score	F Cheng et al, BMC Evol Biol. 2009 Sep 5;9:221
		Accelerated Human CNC Region	5. --		--	No standalone web resource. However, pfSNP has results from this algorism.	SY Kim et al, Plos Genetics, 2007, 3:1572 KS Pollard et al, Plos Genetics, 2006, 2:1599 S Prabhakar et al, Science, 2006, 314:786 EC Bush et al, BMC Evol Biol, 2008, 8:17
		Evolutionarily Conserved	6. 28 way conserved region in UCSC Genome Browser	#	http://genome.ucsc.edu/cgi-bin/hgGateway		W. Miller et al, Genome Research, 2007. 17: 1797
mRNA	Promoter Region	Change TF Binding sites	7. TransFac	*	http://www.biobase-international.com/pages/index.php?id=transfac		E Casamar et al, Nucleic Acids Res. 2010 Jan;38(Database
			8. JASPAR	*	http://jaspar.cgb.ki.se		
			9. MATInspector	*	http://www.genomatics.de		
			10. SNP@Promoter		http://variome.kobic.re.kr/SNPatPromoter		BC KIM et al, BMC Bioinformatics. 2008;9 Suppl 1:S2
	Coding Region SNP	Affect NMD	11. SNP2NMD		http://variome.kobic.re.kr/SNP2NMD/		Han et al, Bioinformatics 2007 23(3):397
		Change ESE/ESS Site	12. ESE Finder	*	http://rulai.cshl.edu/cgi-bin/tools/ESE3/esefinder.cgi?process=home		Fairbrother WG et al, Science, 2002, 297:1007
			13. RESCUE-ESE	*	http://genes.mit.edu/burgelab/rescue-ese/		Zhang XH et al, Gene and Development, 2004, 18:1241
			14. PESX	*	http://cubweb.biology.columbia.edu/pesx/		ZM Zhang, J Biomed Sci, 2004, 11:278
	Intronic SNP	Change Intronic Splicing Regulation Elements	15. --		--	No standalone web resource. However, pfSNP has results from this algorism.	Yeo GW et al, Plos Genetics, 2007, 3:814
		Cause Abberant Splice Site	16. --		--	No standalone web resource. However, pfSNP has results from this algorism.	Sahashi K et al, Nucleic Acid Res, 2007, 35:5995
	3UTR SNPs	Change Mir Binding	17. MicroSNiPer		http://cbdb.nimh.nih.gov/microsniper		M Barenboim et al, Hum Mutat. 2010 Nov;31(11):1223
			18. PolymiRTS Database		http://compbio.utmem.edu/miRSNP/		L Bao et al, Nucleic Acids Res, 2007, 35:D51
			19. Patrocles Database		http://www.patrocles.org/		M Georges et al, Cold Spring Harb Symp Quant Biol. 2006, 71:343
			20. TarBase	#	http://www.diana.pcbi.upenn.edu/tarbase		P Sethupathy et al, RNA. 2006 Feb;12(2):192-7
			21. miRanda	#	http://www.microrna.org/microrna/getGeneForm.do	Target site pre-scanned and stored in microrna.org	B John et al, PLoS Biol, 2004 2:E363
			22. TargetScan	#	http://www.targetscan.org/		BP Lewis, Cell, 2003 115:787
			23. PicTar	#	http://pictar.mdc-berlin.de/		A Krek et al, Nat Genetics, 2005 37:495
		In 3UTR Conserved Region	24. --		--	No standalone web resource. However, pfSNP has results from this algorism.	XH Xie et al, Nature, 2005, 434:338
	miR Coding Region SNP	Change Mir Sequence	25. Patrocles Database		http://www.patrocles.org/		M Georges et al, Cold Spring Harb Symp Quant Biol. 2006, 71:343
Protein	Non-Synonymous SNP	Resides / Affects Functional Domain or Structure	26. Interpro Scan	* #	http://www.ebi.ac.uk/InterProScan/	A collection of 14 protein domain database scanner	Zdobnov EM et al, Bioinformatics. 2001, 17:847
			27. Tm HMM	* #	http://www.cbs.dtu.dk/services/TMHMM/		
			28. NetOGlyc	* #	http://www.cbs.dtu.dk/services/NetOGlyc/		JE Hansen et al, Glycoconj J. 1998 Feb;15(2):115
			29. NetNGlyc	* #	http://www.cbs.dtu.dk/services/NetNGlyc/		--
			30. NetPhos	* #	http://www.cbs.dtu.dk/services/NetPhos/		--
			31. KinasePhos	* #	http://kinasephos.mbc.nctu.edu.tw/		YH Wong et al, Nucleic Acids Res. 2007 Jul;35
			32. OGPET	* #	http://ogpet.utep.edu/OGPET/		--
			33. Sulfinator	* #	http://ca.expasy.org/tools/sulfinator/		F Monigatti, Bioinformatics. 2002 May;18(5):769
			34. SIFT	* #	http://blocks.fhcrc.org/sift/SIFT.html		
			35. Polyphen		http://genetics.bwh.harvard.edu/pph/	Straightforward empirical rules applied to the sequence, phylogenetic and structural information	V Ramensky et al, Nucleic Acids Res, 2002, 30:3894
			36. SNP3D		http://www.snps3d.org/	SVM approximate the stability effect on a protein structure	P Yue et al, BMC Bioinformatics, 2006, 7:66
			37. LS-SNP		http://alto.compbio.ucsf.edu/LS-SNP/	SVM to integrate sequence, evolution, and structure information with a combination of knowledge-based rules	R Karchin et al, Bioinformatics, 2005, 21:2814
			38. Panther-SNP	* #	http://www.pantherdb.org/tools/csnpScoreForm.jsp	Conserved seq by HMM Domain analysis and Ka/Ks	PD Thomas et al, PNAS, 2004, 101:15398

2.1.2.1 Lack of suitable tools for information integration

In **Table 2.2**, a number of methods are listed without a URL or tool name. These are the methods with no proper implementation, and they require special effort to tap into them. Furthermore, some tools (e.g. ESE finder (Cartegni, Wang et al. 2003)) only provide a web-based interface, which is not suitable for scanning large-scale data, and for which scalable local implementation is still required. Further, a number of tools are not designed for SNPs (e.g. ESE finder), and post-processing is required to identify SNPs with alleles causing a difference in the results (e.g., only SNPs that cause a change in the ESE motif will be interesting).

2.1.2.2 Heterogeneous tools with complex input data requirement

Although some tools only require the rs number (rsNo) as an input field, most tools require additional information, such as a SNP flanking sequence or even a corresponding gene or amino acid (AA) sequence (**Table 2.2** with '*' and '#' in the "Input Requirement" column). For example, PantherSNP (Thomas and Kejariwal 2004) requires input of the AA sequence change caused by the SNP together with the AA sequence of the corresponding splicing variants of the protein. Preparing such information requires the use of extra tools and data sources.

2.1.2.3 No cross referencing and cross checking of information available and possible erroneous information provided

There are often inconsistencies in the information provided, such as the SNP

allele reported and the actual allele found on the mRNA, because SNP allele reporting is not required to be on the same strand of mRNA. Furthermore, the SNP allele may not correspond to the amino acid change reported for non-synonymous SNPs. Such inconsistencies may create unnecessary confusion in interpreting the biological impact of the SNP.

2.1.3 Tools needed for the semi automated pipeline

In light of the aforementioned gaps in the available tools, three tools were deemed necessary to semi-automate a pfSNP collection pipeline: 1) a multi-purpose SNP mapper; 2) a general-purpose motif scanner; and 3) a SNP statistics generator for large-scale data sources.

1. A multi-purpose SNP mapper:

Why it's needed: The primary goal of the mapper is to obtain the gene context of the SNP (e.g., which exon it is in and which codon it is in). The mapper will use this information further in two ways. First, it will apply location based rules to tell if a SNP is a pfSNP. For example, it can pinpoint non-sense SNPs causing mRNA nonsense-mediated decay (NMD), because only non-sense SNPs located more than 50 bps from the last intron-exon boundary will lead to NMD (Nagy and Maquat 1998). It will also help determine whether a SNP is likely to affect miRNA targeting, because SNPs in the seed sequence of miRNA are more important than SNPs in other regions. Second, it will deduce new information from related information and check for

inconsistency automatically. For example, the difference in synonymous SNP codon usage can be deduced by retrieving the codon from the corresponding mRNA sequence using the gene context of the SNP. A check can be carried out to see if the reported allele of the SNP is consistent with the nucleotide found on the mRNA at the same location. Furthermore, the SNP allele change reported and the amino acid change caused by the non-synonymous SNPs may also not tally. For example, if a SNP were reported to have an allele change of A/G and an amino acid change of Arg/Ala, it may appear as if the A allele corresponds to an Arg residue; there is no guarantee for such correspondence. Such inconsistencies may create unnecessary confusion in interpreting the biological impact of the SNP. By using the mapper, the amino acid change can be deduced from the codon sequence retrieved from the mRNA sequence, helping to guarantee consistency between allele change and amino acid change.

Special considerations: Although the input data of the mapper will be primarily from NCBI, it should work with other data sources as well. Therefore, minimal information should be needed, and any information that can be derived from known ones should not be required. (e.g. AA change caused by a non-synonymous SNP is redundant once the alleles and mRNA sequence are known)

2. A general-purpose motif scanner.

Why it's needed: A general-purpose motif scanner is beneficial in two ways.

First, because a number of papers simply report the important motifs found, without describing the tool to screen for such a motif. Second, there is no facility to compare and contrast the output for different alleles of a SNP, which is the key to pinpointing those SNPs whose alleles would cause a change in motif.

Special considerations: Although the motifs that currently need to be screened are only DNA motifs, the need to scan for the protein motif in the future is likely, and therefore the scanner should be able to handle protein sequences as well.

3. A SNP statistics generator for large-scale data sources (such as HapMap).

Why it's needed: The main purpose for developing a SNP statistics generator for large-scale data sources is to be able to calculate pair-wise comparisons and overall Fst amongst different HapMap populations. In the meantime, other descriptive statistics, such as minor allele frequency and HWE P values, are also calculated.

Special considerations: Since the statistics generator may be handy for other applications, such as analysing our own genotyping results, building the generator as a Microsoft Excel-based tool is highly desirable.

2.2 Materials and methods

Table 2.3 shows the input, output and algorithms/methods to be included into the three tools.

Table 2.3: The input, output and algorithms/methods to be included into the three tools.

Tool	Input	Output	Algorithms/methods
Multi-purpose SNP Mapper	1. SNP genome location and allele information. 2. Gene exon and UTR genome coordinates for each splicing variants 3. mRNA sequence containing the sequence for the splicing variants of each gene	1.SNP mRNA location 2.SNP protein residue change	NMD algorithm
General purpose motif scanner	SNP flanking sequences with each allele embedded	The motif created and disrupted by the allele change	1.ISRE motif 2.TF binding sites 3.ESE/ESS motif 4.3'UTR conserved site
SNP Stats Calculator	SNP genotype information	1.HWE P value 2.Genotype frequency 3.Allele frequency 4.Fisher's exact test for population difference 5.Fst	1. HWE 2. Fst 3. Fisher's exact test

For the three tools mentioned, I chose to implement them using the Microsoft Visual Basic (VB) family of languages. This is mainly because the VB family of language is widely used in Microsoft Excel and ASP.NET web framework, and the tools can thus be easily built as Microsoft Excel-based desktop applications as well as web-based applications with shared core program components and libraries.

Specifically, the general-purpose SNP mapper was developed as a Microsoft Excel macro-based application using Microsoft Visual Basic for Application (VBA). Although this mapper is Microsoft Excel-bound, it only uses Microsoft Excel as a

running host, and the input and output are all channelled from and to Open Database Connectivity (ODBC) compatible databases via Microsoft Excel ActiveX Data Objects for the genome scale mapping purpose. To satisfy the "minimal information required" requirement, only three pieces of information is needed for the mapper as input. The first is a "SNP" table containing genome location and allele information for all SNPs. The second is a "Gene" table containing the exon and UTR genome coordinates for each splicing variants of every gene. The third is an "mRNA sequence" table containing the sequence for the splicing variants of each gene. This architecture also makes it easy to adapt the mapper to be used with low-throughput data, if needed. Source data in the right format can be read directly from Microsoft Excel worksheets and output can be saved as a Microsoft Excel workbook. All processes are facilitated by the familiar user interface provided by Microsoft Excel.

The general-purpose motif scanner was developed in Visual Basic .NET (VB.NET). It has two components. The first component is a wrapper to feed the two flanking sequences containing each allele of a SNP into respective motif-scanning packages. The "PS Scan tool" (ftp://ftp.expasy.org/databases/prosite/tools/ps_scan) was selected as a tool to scan for ESE/ESS motifs (Fairbrother, Yeh et al. 2002; Wang, Rolish et al. 2004; Zhang and Chasin 2004; Zheng 2004), ISRE motifs (Yeo, Nostrand et al. 2007) and 3' UTR conserved sites (Xie, Lu et al. 2005) in a DNA/RNA sequence. The DNA motif signatures from the abovementioned publications were compiled into a Prosite-compatible format by our Microsoft Excel macro. Although the "PS Scan tool" was originally designed to screen motifs in AA sequences, it can be used to screen for motifs in DNA sequences because the four nucleotide code for DNA is within the AA code collection. The Match program from TransFac (Matys, Fricke et al. 2003) is used for TF site screening. The second

component is a post processor to screen through the motif scan results and determine if any of the altered motifs can be found by comparing the outputs of the two SNP alleles. This was implemented by VB.NET. Using VB.NET also enabled the addition of a web front for this general-purpose scanner with ease, which has been made available at http://pfs.nus.edu.sg/seqmotifscan/. A sequence uploading facility was built with the help of a Perl script utilizing the BioPerl sequence manipulation package to provide functionalities of taking different input formats and screening for reverse complementary sequences.

The SNP statistic generator was also developed as a Microsoft Excel-bound macro-based tool for the same reasons. It can take in genotype in HapMap format either from the database connection or from selected Microsoft Excel worksheets, and output to a database or a Microsoft Excel workbook, accordingly.

The three tools form the backbone of a semi-automated pipeline to facilitate the pfSNP data warehousing (**Figure 2.1**).

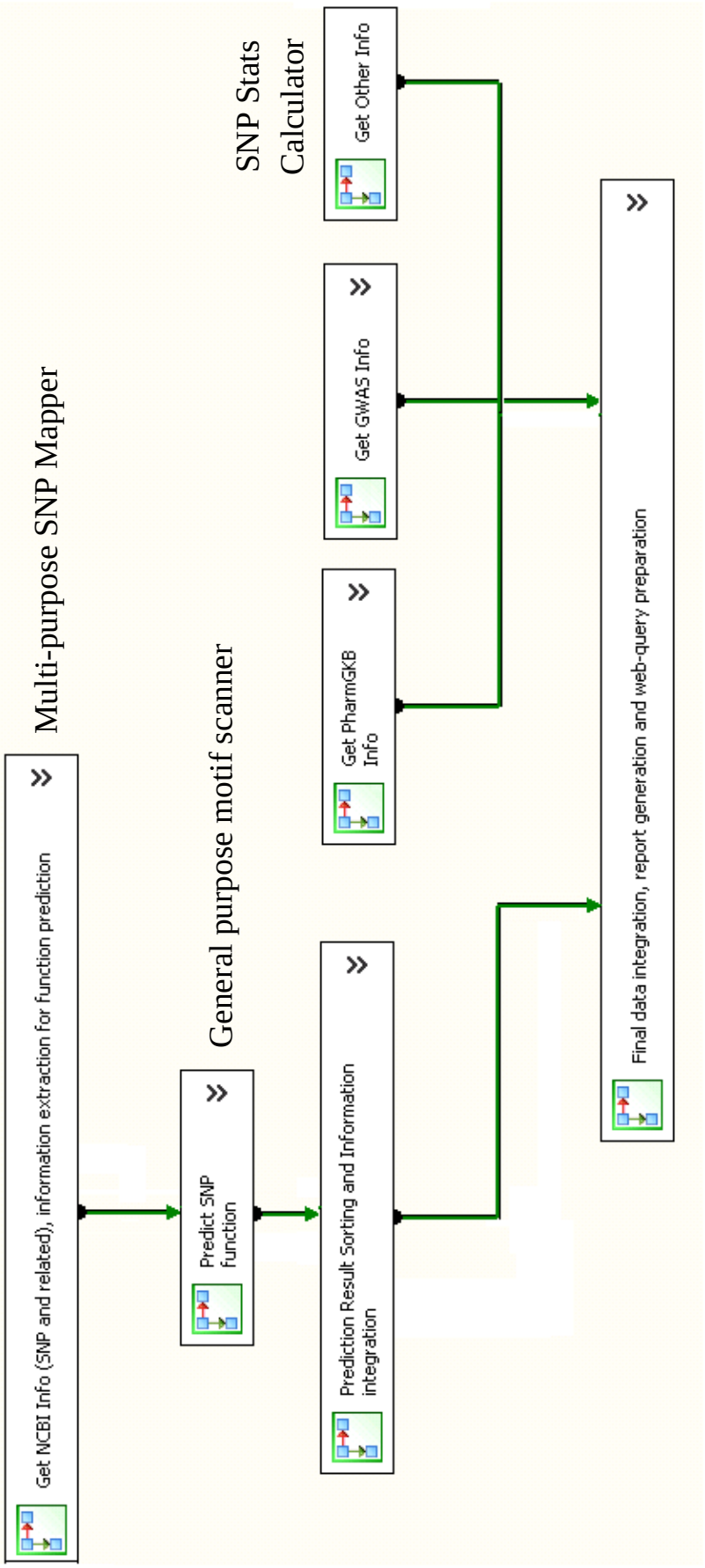


Figure 2.1: The data flow of pfSNP data warehousing process. A. General data flow. B. Detailed data flow for NCBI SNP-related information retrieval.

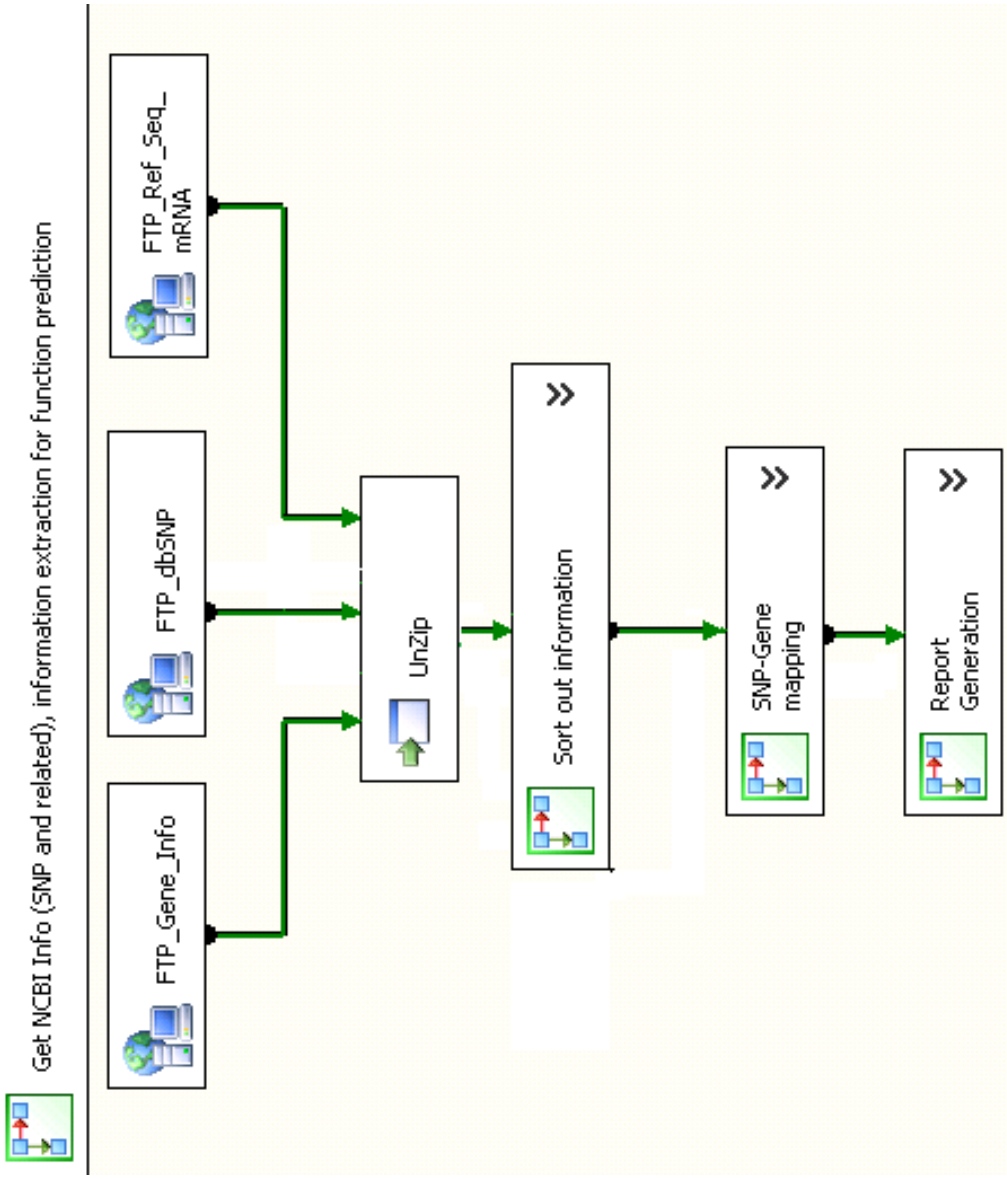


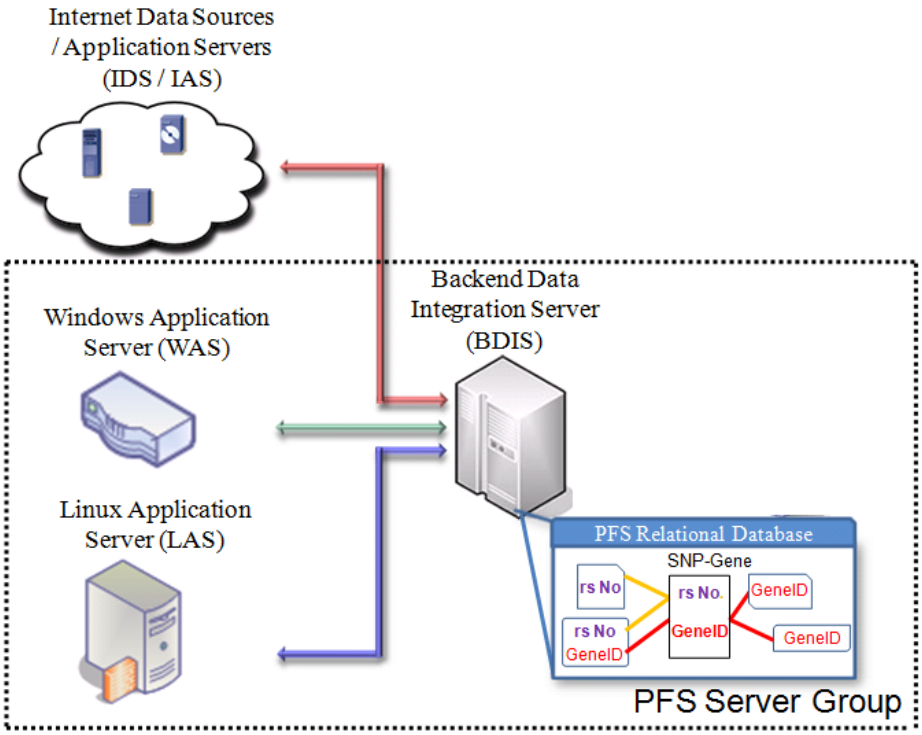
Figure 2.1 (Cont.) B

The workflow begins with retrieving relevant NCBI information (dbSNP and Entrez RefSeq (ftp://ftp.ncbi.nih.gov/refseq/) (**Figure 2.1 A**; more detail in **Figure 2.1 B**), and the SNP mapper will be utilized to generate SNP-gene mapping and summary reports (e.g., which gene and splicing variant the SNP maps into, which region the SNP maps into, what function the SNP has, etc.). Input information for the motif scanner will be prepared and fed into the "Predict SNP function" step. SNPs that change various motifs will be identified in this step and a sorting and integration step

will follow to generate SNP function summaries by combining the dynamic prediction results with static results, such as that garnered from previous literature function reports and other static functional predictions and inferences. Further down the pipeline, a gene level summary will be generated by integrating the SNP level summary together with gene level information; for example, PharmGKB pathway information.

Besides the three tools developed, which would run primarily in Microsoft Windows environment, the database would need to be able to incorporate data from Linux-specific programs or internet-based application servers. **Figure 2.2** outlines the overall architecture of the hardware and software components of the semi-automated pfSNP collection pipeline. The pfSNP database will be held in the Background Data Integration Server (BDIS) in the format of a relational database. Since flexible database architecture would be needed to enable querying from one or all layers of the SNP-Gene-Pathway hierarchy, a star schema centred on a SNP-Gene mapping table generated by the multi-purpose mapper was used. Dimension tables link to the central table via SNP ID, gene ID or a combination of the two.

Figure 2.2: The hardware and software design of the pfSNP database.



To characterize the pfSNP dataset, the numbers of pfSNPs satisfying various criteria were obtained by SQL queries carried out in the pfSNP database. The chromosome coverage diagrams were generated by a Microsoft Excel-based tool developed in-house. The list of markers in the tagging SNP (t-SNP) and quasi-random SNP (QR-SNP) sets were obtained from Illumina 1M Duo Bead Chip and Affymetrix SNP Array 6.0, respectively. The list of SNPs in HapMap Release 23, their genotypes and minor allele frequencies in the four HapMap populations were obtained from HapMap FTP server (ftp://ftp.ncbi.nlm.nih.gov/hapmap/). The population names are: Beijing Chinese (CHB), Tokyo Japanese (JPT), U.S. residents with northern and western European ancestry (CEU) and the Yoruba people of Ibadan, Nigeria (YRI). HapMap Release 23 data was used because this is the release Illumina used to benchmark the genome coverage of the t-SNP set. The Tagger program in Haploview (Barrett, Fry et al. 2005) was used to check the tagging efficiency of pfSNP and also

the list of tagging SNPs needed to cover all the SNPs in HapMap Release 23.

2.3 Results and discussion

2.3.1 The multi-purpose SNP mapper

The output of the mapper is a single multi-column table with one SNP per row. **Figure 2.3** shows a few sample outputs. For the ease of reading, columns are grouped as "RNA-related" (**Figure 2.3A**) and "Protein-related" (**Figure 2.3B**). The "rsNo" and "GeneID" are duplicated in each for easy reference.

Figure 2.3: The structure of SNP-Gene mapping table and sample output. A "RNA-related" columns. B "Protein-related" columns

A

	rsNo	GeneID	RNAAccession	Feature	FeatureDetail	FeatureLocat	Allele	CorrectedAllele	InPutativePromoter	EUpBps	EDnBps	InLastExon
1	rs17884410	GeneID:55135	NM_018081.1	5UR		-1017	T/C		YES	0	0	
2	rs17884410	GeneID:7157	NM_000546.3	E	1	-110	T/C	A/G	YES	140	-82	
3	rs1800372	GeneID:7157	NM_000546.3	E	6	639	T/C	A/G	NO	79	-33	NO
4	rs2291078	GeneID:7372	NM_000373.1	E	4	1050	T/A	T/A	NO	67	-108	NO
5	rs34686922	GeneID:55135	NM_018081.1	5UR		-1209	A/G		YES	0	0	
6	rs34686922	GeneID:7157	NM_000546.3	I	1	110	A/G		YES	0	0	

B

	rsNo	GeneID	ProteinAccession	ProteinLocat	AAChange	AAChangeType	CodonChange	CodonUseDiff
1	rs17884410	GeneID:55135	NP_060551.1					0
2	rs17884410	GeneID:7157	NP_000537.3					0
3	rs1800372	GeneID:7157	NP_000537.3	213	Arg/Arg	Syn	CGA/CGG	5.2
4	rs2291078	GeneID:7372	NP_000364.1	350	Cys/End	NonSyn	TGT/TGA	0
5	rs34686922	GeneID:55135	NP_060551.1					0
6	rs34686922	GeneID:7157	NP_000537.3					0

In the "RNA-related" columns, the "RNAAccession" field records the specific splicing variant the mapping refers to. The three following "Feature" columns record the gene feature (5' upstream region, exon, intron or 3' UTR), the serial number (which exon or intron) and the exact location of the mapping. For example, rs1800372 in **Figure 2.3A**, row 3, has "E", "6" and "369" in the three columns, respectively. It

means the SNP is in exon 6 of the gene and 369 bps after the translational start site.

It is noteworthy that the naming convention is different for different regions and the logic for handling such naming conventions has been built into the mapper. For example, rs1784410 in **Figure 2.3A**, row 1, is in the 5' upstream region and "-1017" refers to 1017 bps upstream of the transcription start site instead of translational start site, which is for 5'UTR and coding region SNPs, as with previously mentioned rs1800372. The mapper correctly handles SNPs in different regions and chooses naming conventions accordingly.

The "Allele" column refers to the allele reported by the "SNP" table, while the "CorrectedAllele" records the mRNA corrected allele by checking the nucleotide found on the mRNA at the position of the SNP. This is to see if the reported allele is of the same strand of the particular mRNA. For example, rs1800372 has a reported allele change of "T/C", but the mRNA sequence at this particular position has an A; so, the "CorrectedAllele" reports an "A/G" change, which reflects the strand difference. Altogether, 54% (115336/214390) UTR or coding SNPs were observed to contain strand inconsistency. This was expected, since SNP allele assignment can be on either strand of DNA and are done without taking into consideration the mRNA in which it may reside. Intronic and 5' upstream SNPs cannot be checked, due to the lack of reference mRNA sequences.

I defined the putative promoter as 5000 bps and 500 bps up- and down-stream of the transcription start site, respectively. The "InPutativePromoter" column informs if a 5' UTR and exonic SNP are in the promoter region. For example, rs17884410 in **Figure 2.3A**, row 1, is a promoter SNP in the 5'UTR region.

The "EUpBps" and "EDnBps" columns show the distance of the coding SNP from the upstream and downstream intron-exon boundary, respectively, which would

be helpful to determine if a SNP falls into the conserved motif within the boundary. Combining information from "EDnBps" and "InLastExon" would further indicate if a stop codon created by the SNP would lead to NMD of the mRNA by predefined rules (Nagy and Maquat 1998).

Figure 2.3B contains columns recording protein level information for the SNP. It's noteworthy that since we required minimal input information, most of the protein level information is derived on the fly, based on the previous mRNA mapping information. "CodonChange" information is obtained directly from the actual mRNA sequence with the help of previous mRNA mapping information, together with the "CorrectedAllele". The "CodonChange" information further helps to derive the "AAChange", and the status of synonymous or non-synonymous is also generated on the fly. "CodonChange" is further used to deduce the "CodonUseDiff", which is obtained by comparing the codon frequency as reported in the codon usage database (http://www.kazusa.or.jp/codon/). Synonymous SNPs with top 5% codon usage difference were declared as pfSNP.

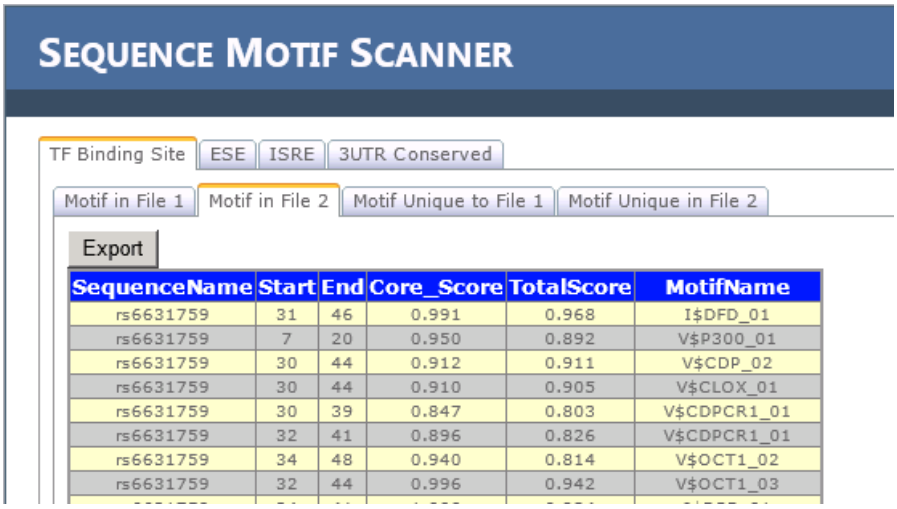
2.3.2 The general-purpose motif scanner

The general-purpose motif scanner, with built-in compare and contrast facility, was developed and has been made accessible at http://pfs.nus.edu.sg/seqmotifscan/. **Figure 2.4A** shows a screen shot of the typical output screen. It is currently capable of screening for transcription factor binding sites, exonic splicing enhancers/exonic splicing silencers (ESE/ESS), intronic splicing regulatory elements (ISRE), as well as 3'UTR conserved sequences. Results for each category of motif are presented in different tabs, each with four sub-tabs. The sub-tabs contain the raw scan outputs

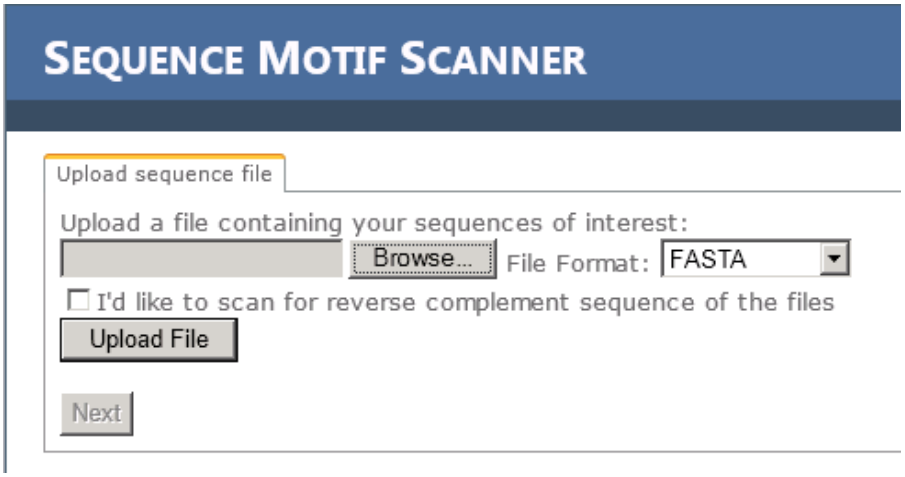
from the two input files, as well as the unique motif found in each input. The sequence upload facility is shown in **Figure 2.4B**.

Figure 2.4: The general purpose motif scanner. A. Sample output of the motif scanner. B. The input files uploading facility for the motif scanner

A.



B.



2.3.3 The SNP stats calculator

The sample output of the SNP stats calculator is shown in **Figure 2.5**. The SNP stats calculator helps to calculate the overall Fst for all input populations as well as pairwise Fst for each population pair; there are four populations in this case, but more populations can be handled. It also calculates pairwise Fisher's exact P values

for allele frequency as another measure of population difference. Genotype and allele frequencies, together with HWE P values, are also given for easy checking of possible genotyping error, if any. The calculator can take input data in HapMap format directly from Microsoft Excel and output easy to read summary tables, as shown in **Figure 2.5**. It can also handle large scale data from database and store output into databases directly (not shown).

Figure 2.5: Sample output of the SNP stats calculator in Microsoft Excel.

SNP ID	Gene Location	Population	n	HWE P-Value	Genotype frequency (%)			Allele frequency (%)		Pairwise differences Fisher's exact P-value				Fst	Pairwise Fst Value			
					AA	AG	GG	A	G	CHB	JPT	CEU	YRI		CHB	JPT	CEU	YRI
rs1202184	I4/A928G	CHB	45	0.46	53.33	42.22	4.44	74.44	25.56		0.06	1.18E-05	1.18E-14	0.33		0.04	0.16	0.67
		JPT	45	0.17	31.11	57.78	11.11	60.00	40.00			0.03	1.14E-14				0.04	0.52
		CEU	60	0.05	13.33	61.67	25.00	44.17	55.83				2.27E-12					0.32
		YRI	60	0.63	0.00	11.67	88.33	5.83	94.17									

2.3.4 The semi-automated pfSNP collection pipeline

In addition to the three major tools developed, a collection of auxiliary scripts were written with Bio Perl library to facilitate sequence manipulation and looping through a collection of inputs for Linux-based programs. The number of pfSNPs identified by each method is listed in **Table 2.4**.

Efforts are currently underway to facilitate the continuous update of the pfSNP resource in an automated fashion so that this resource can remain useful for a long time. As shown in **Figure 2.2** previously, pfSNP resource will comprise three servers, namely, the Backend Data Integration Server (BDIS), the Windows Application Server (WAS) and the Linux Application Server (LAS). The BDIS will monitor various Internet Data Sources (IDS) by SQL Server Integration Service for any new updates, as well as obtain SNP function prediction results from various application servers, including the Internet Application Server (IAS) , the WAS and the LAS. The

Table 2.4: No. of pfSNPs identified by the semi-automated pipeline for each tool/method

Molecule Level	SNP Cat.	Function Cat.	Method Short Description	No. pfSNP Identified
mRNA	Promoter	Change TFBS	TransFac	176550
	Coding Region SNP	Affect NMD	Nonsense Mediated Decay	917
		Change ESE/ESS Site	RESCUE-ESE	58881
			Detect enrichment compared to 5UTR and Intronless exons	
			Hex mer liberary screening	
			Experimentally validated ESE/ESS	
	Intronic SNP	Change ISRE sites	Conserved intronic sequence within 400bps of splicing site	134674
		Reside on Splice Site	Cause abberant 5' site (SD-Score)	379
			Change 3' Consensus sequence	603
	3'UTR SNPs	Change Mir Binding	MiRanda to predict binding change and also functional validation	344
			PolymiRTS Database	10682
			Patrocles Database	11732
		In 3UTR Conserved Region	Functional region predicted multiple species sequence alignment	5035
		Change TFBS	TransFac	39920
	5Kb Downstream	Change TFBS	TransFac	208470
	miR Promoter / Coding Region SNP	Change TFBS	TransFac	7407
		Reside on miRNA seed, etc region	NCBI and miRBase	167
			Patrocles Database	117
Protein	Coding Region SNP	Protein Domain / Functional Sites	Interpro Scan	34461
			Tm HMM (Transmembrane domain)	1207
			NetOGlyc (O Glycosylation site)	439
			NetNGlyc (N Glycosylation site)	93
			NetPhos (Phosphorylation site)	4969
		Non-Synonymous SNP	Polyphen	10383
			SNP3D	2040
			LS-SNP	812
			Panther-SNP	121
			All Non-Synonymous SNP	49733
		Synonymous SNP	Codon usage differences falling in top and last 5%	2933
SNP	All SNP	Disease/Function Related	OMIM reported SNPs	67
			PubMed Reported SNPs	531
			NCBI dbGaP Reported SNPs (Cut off 0.01)	22308
			HGMD for Reported SNPs	2043
		Positively Selected	mLRH (modified LRH)	82
			WGLRH (Whole Genome LRH based on LRH)	4110
			LDD (LD Decay)	24596
			"iHs" (Integrated Haplotype Score)	264040
			Fst	183
			Fst_HM	15368
			abnormal CNC Regions	788
			"Human Accelerated Region"	115
			Accelerated Human CNC Region	680
			Accelerated Human CNC Region	565

WAS and LAS will execute SNP function prediction software locally and the standardized web-interface wrapper will be implemented to facilitate a unified client-server mode of communication between the BDIS and the Application Servers. New algorithms/applications can then be easily included manually through a simple customization of the standardized web-interface wrapper.

Minor updates affecting only a single satellite table (e.g., table for dbGAP, which is linked to the main table through rsNo) will be downloaded automatically, and pre-defined data processing will be carried out by BDIS, which will then update the corresponding tables in the BDIS. For major changes, in which the main SNP-Gene mapping table has to be altered (e.g., new SNPs are identified through ultra-high throughput sequencing or new release of dbSNP database), the database will be updated in an incremental manner, where new SNP entries will be identified by the BDIS, and various Application Servers will evaluate the potential functionality of the new SNP entries. BDIS will then compile a new release of PFS database.

2.3.5 The pfSNP database

The pfSNP database was established in Microsoft SQL Server 2005 as a relational database with star schema centred on the SNP-Gene mapping table. **Table 2.5** lists the major tables in the pfSNP database providing SNP- and gene-related information. Specifically, **Table 2.5** lists how the major tables are linked to the central table and the information content provided. The number of tables in **Table 2.5** is less than the methods and tools used in **Table 2.4** because results from the methods of a similar purpose are aggregated into one table to facilitate downstream analysis and filtering. For example, there are four papers identifying different ESE/ESS motifs in

Table 2.5: The list of major tables providing SNP and gene level information

Table No.	Purpose	Information Content	Link to SNP-Gene mapping table
1	The SNP-Gene Mapping Table	rs No., GeneID, mRNA Location summary in all the splicing variants, mRNA Location Count, AAChange, SNP location summary (5' UR,5' UTR, Promoter, Coding, NonSyn, Syn, Intron, 3UTR, 3'DR)	N.A
2	SNP chromosome location and selection summary	SNP chromosome location, whether it shows RPS signature or in ACNC regions	rs No.
3	SNP showing high Fst value	Fst value	rs No.
4	SNP minor allele frequency	SNP minor allele frequency in all HapMap populations	rs No.
5	SNP changing Transcription Factor binding site in gene promoter	Summary of TF binding site altered by the SNP	rs No.
6	Non-synonymous SNP deleterious effect prediction	Polyphen, SIFT, SNP 3D, LS-SNP, Panther SNP	rs No, GeneID
7	Protein domain in which the non-synonymous SNP reside	InterPro Scan results, Transmembrane domain by TMHMM, O- and N-glycosilation site, Phosphorylation site.	rs No., GeneID
8	SNP changing ESE/ESS site	ESE/ESS site altered by the SNP	rs No.
9	Synonymous SNP with high codon usage difference	Codon usage difference	rs No., GeneID
10	SNP causing NMD	SNP causing NMD in respective gene	rs No., GeneID
11	SNP causing abberant 5' splicing site	SNP causing abberant 5' splicing site	rs No.
12	SNP causing abberant 3' splicing site	SNP causing abberant 3' splicing site	rs No.
13	SNP changing ISRE site	ISRE site altered by the SNP	rs No.
14	SNP changing 3' UTR conserved motif	3' UTR conserved motif altered by SNP	rs No.
15	SNP changing 3' UTR miRNA binding site	miRNA whose binding site got altered by SNP	rs No.
16	SNP reported to be associated with phenotype by NCBI dbGaP	Phenotype associated, P value	rs No.
17	SNP reported by HGMD	Phenotype associated and related information if any	rs No.
18	SNP reported to be associated with phenotype by MedLine Search	Phenotype associated and related information if any	rs No.
19	SNP reported in OMIM	OMIM disease relation information	rs No.
20	Gene Information Collection	GO term, KEGG path, mSigDB path, tissue expression, Pharm GKB knowledge and Genetic Assocaiton Database information.	GeneID
21	Gene reported by OMIM	OMIM disease in which the gene implicated	GeneID
22	SNP changing Transcription factor binding site in miRNA promoter	Summary of TF binding sites altered by the SNP	rs No.
23	SNP mapped into miRNA coding and related regions	The miRNA and region the SNP maps into	rs No.

Table 2.4, but their motifs are combined in the scanning process and there is only one final table (Table No.8 in **Table 2.5**) to summarize pfSNPs changing ESE/ESS motifs.

Other than these major tables, the pfSNP database also holds auxiliary tables which contain supplementary information to the tables in the core star schema (e.g., a table for the ESE/ESS motif reporting their origin and the validation status) as well as those tables to facilitate query of the database (e.g., a SNP may have multiple rs No. given to it on different occasions, and a table to record such aliases is used in the query process so that querying any of the rs No will return the same result). Since they are not the focus of the pfSNP database, I will not go into any further detail.

There are multiple SNP annotation projects similar to pfSNP in recent years (**Table 2.6**). In terms of SNP function coverage, a number of them only cover the "predicted" pfSNP or "reported" pfSNP category. Only SNPNexus (Chelala, Khan et al. 2009) collects SNPs with reported and predicted functions but it does not compare and contrast the motifs that are created or disrupted by different alleles. It also does not contain SNPs with inferred functionalities. In contrast, pfSNP also provides auxiliary information, such as Linkage Disequilibrium (LD) data to help identify causal SNPs when a non-functional SNP is searched.

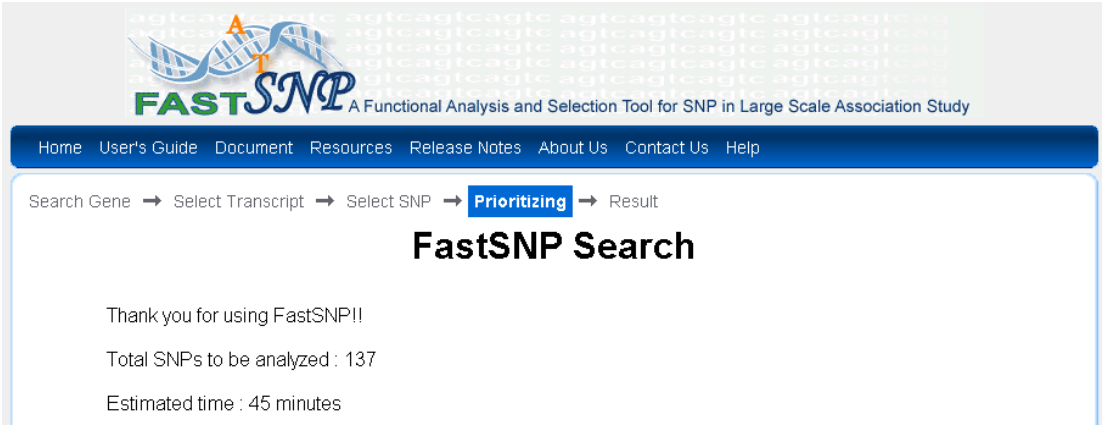
Full automation of the whole pipeline is desirable, but daunting. Recently, Goodswen et al. proposed an R-based fully automated data integration pipeline (Goodswen, Gondro et al. 2010). However, the number of tools tapped into is small (namely dbSNP, GO, KEGG, UniProt, QTLdb, OMIA and HomoloGene) and restricted to ready-to-use tools with little or no downstream processing required. To date, there has yet to be a fully automated pipeline capable of handling complex data processing comparable to pfSNP.

Table 2.6: The data content provided by recently published SNP function resources.

	Tools with predicted and reported function		Tools with reported function			Tools with predicted function		
	pfSNP	SNPNexus	Varietas	Gwas Analyser	Scan	CandiSNPer	SNPInfo	SNPLogic
	2010	2009	2010	2010	2010	2010	2009	2009
1. SNP Covered	dbSNP 129	?	dbSNP 130	?	dbSNP 129	HapMap R27	?	?
2. Gene Context	√	√	√	√	√	√	√	√
3. Reported SNP Function	√	√	√	√	√	--	--	--
4. Predicted SNP Function	√	√	--	--	--	√	√	√
5. Inferred SNP Function	√	--	--	--	--	--	--	--
6. Allele Frequency	√	√	--	√	√	--	√	√
7. LD Information	√	--	--	√	√	√	√	√

Other than the data warehousing approach, which stores data locally, federated data integration is offered by selected resources, such as FastSNP (Yuan, Chiou et al. 2006) and SNPit (Shen, Carlson et al. 2009) as lightweight alternatives in term of storage space. However, federated data integration sacrifices performance for the benefit of storage efficiency and the need for information to be regularly updated. As **Figure 2.6** shows, querying 137 SNPs will take 45 minutes as compared to less than a minute on data warehousing based solutions. The need to process information on the fly may also burden the server and the scalability may quickly become a concern.

Figure 2.6: The FastSNP takes 45 minutes to process a simple query.



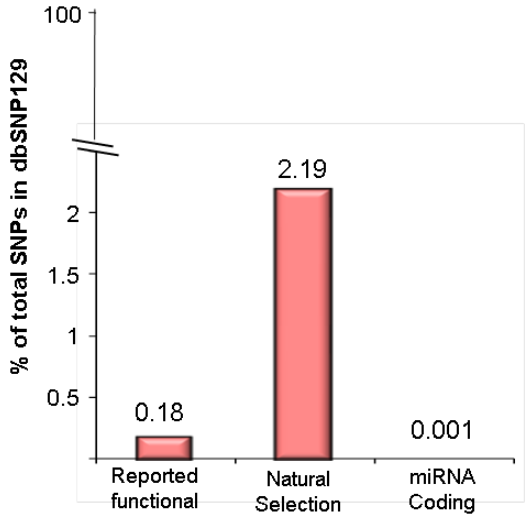
2.3.6 Characterization of the pfSNP data set

Out of >14,000,000 SNPs in the dbSNP129 database, a total of 972,673 SNPs were found to be of potential functional significance (list of SNPs can be downloaded from the pfSNP web-resource http://pfs.nus.edu.sg/PickGwas_V2_3.aspx). As shown in **Figure 2.7A**, only <0.2% of the unique and non-redundant dbSNP129 SNPs were previously reported to be functional or associated with a disease/phenotype, based on information of SNPs culled from the NCBI dbGAP and HGMD databases and from two publications that performed literature mining from PubMed (Xuan, Wang et al. 2007) or OMIM (Stoyanovich and Pe'er 2008). Another ~2% of the SNPs in the dbSNP129 database were "inferred" to be functional because they are under selective pressure, while only 0.001% were found to reside within miRNA coding regions.

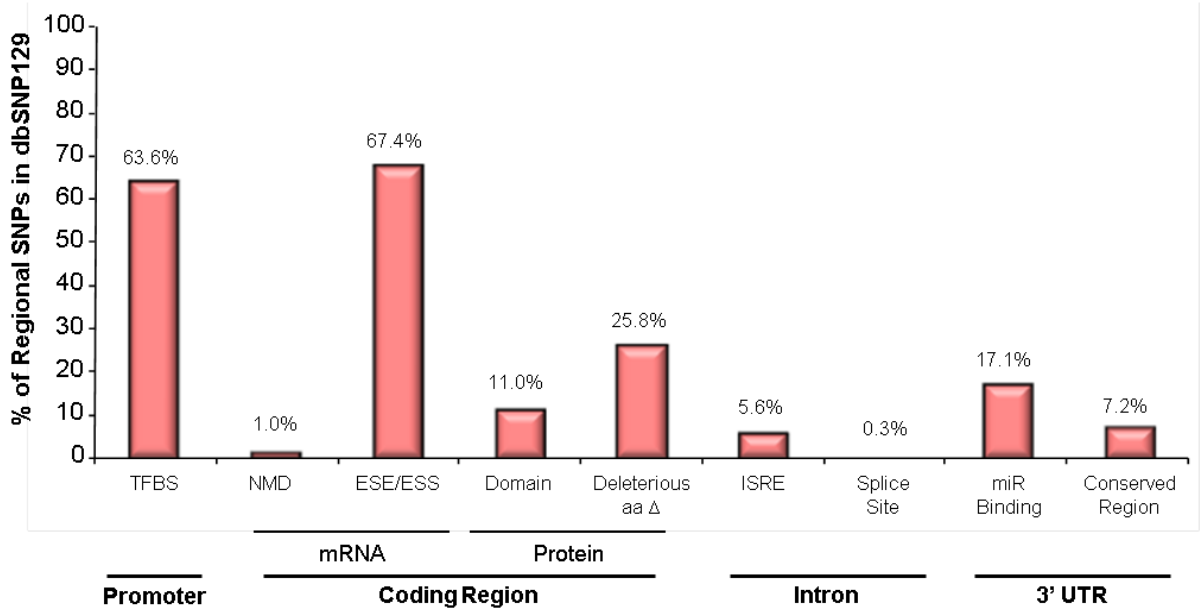
The percentage of dbSNP129 SNPs residing in the various genomic regions predicted to be potentially functional (pfSNPs) was investigated and the results are shown in **Figure 2.7B**. Greater than 63% of promoter SNPs in dbSNP129 was predicted to be potentially important, affecting transcription factor binding sites (TFBS). Of the coding dbSNP129 SNPs, 1% were predicted to result in nonsense-mediated decay (NMD) while 67.4% were predicted to alter exon-splice-enhancer (ESE)/exon-splice-silencer (ESS) sites at the transcript level. At the protein level, 11% of coding dbSNP129 SNPs were predicted to reside in functionally important domains or alter phosphorylation/glycosylation sites, while 25.8% were predicted to result in deleterious amino acid changes. Within the introns, 5.6% of SNPs were predicted to be functionally significant, primarily due to their effects on intron splice regulatory element (ISRE) sites, and only 0.3% of dbSNP129 intronic SNPs were predicted to alter splice sites. At the 3'UTR regions, >32% of

Figure 2.7: Characterization of potentially functional SNPs (pfSNPs) in dbSNP129 database. A. Graph showing the percentage of total SNPs in the dbSNP129 database that is reported to be associated with function/disease (left bar), show evidence of natural selection (middle bar), or reside within miRNA coding regions (right bar). B. Graph showing the percentage of regional (i.e. promoter, coding region, intron or 3'UTR) SNPs in the dbSNP129 database that is predicted to be potentially functional, including affecting transcription factor binding sites (TFBS) in the promoter; affecting non-sense mediated decay (NMD) or exon-splice enhancer (ESE)/exon-splice silencer (ESS) in coding region at the mRNA level; residing in important domains (e.g. p53 tetramerization domain) (Domain) or predicted to result in a deleterious amino acid changes (deleterious aa) within the coding region at the protein level; affecting intron splicing regulatory elements (ISRE) or splice sites (Splice Site) within introns, as well as affecting miRNA binding sites (miR binding) or residing within conserved regions at the 3'UTR.

A



B



dbSNP129 SNPs were predicted to have potential functional importance, including 17.1% that affect miR binding sites and 7.2% that reside within highly conserved regions. Due to our lack of understanding of the functional significance of inter-genic regions, there is a paucity of algorithms to predict the potential functionality of SNPs in this region. Hence, only 2.1% of inter-genic SNPs in dbSNP129 were predicted to have potential functional significance. These inter-genic pfSNPs are primarily those that are reported to be associated with disease/phenotype through genome-wide association studies or those under selective pressure.

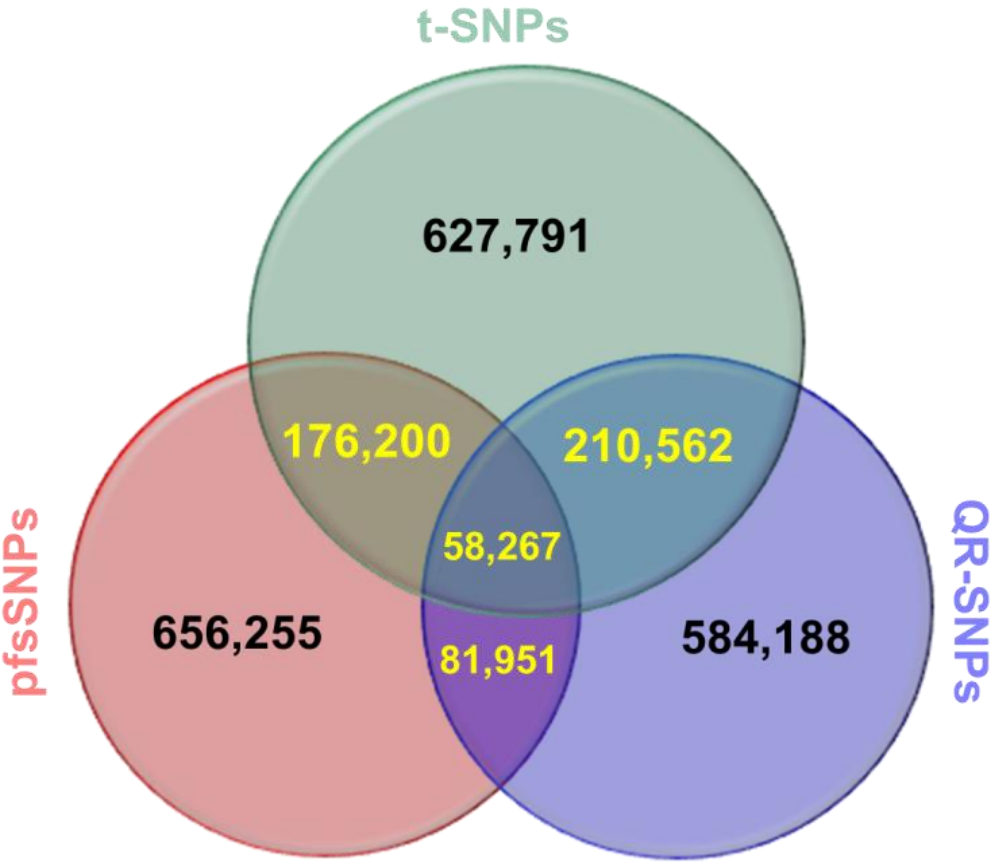
As the total number of pfSNPs (972,673) was similar to the number of QR-SNPs used in the Affymetrix SNP Array 6.0 (934,968) and the t-SNPs in the Illumina Human 1M-Duo DNA-Analysis Beadchip (1,072,820), the characteristics of these three SNP sets were compared.

The extent of similarity amongst the SNPs selected from the three different SNP-selection platforms is depicted as a Venn diagram in **Figure 2.8**. Only 58,267 SNPs (<7%) in each SNP set are shared amongst all three. Greater than 21% of the SNPs are common between QR-SNPs/pfSNPs. The least sharing of SNPs (<15%) was between the pfSNP and QR-SNP sets. An additional ~11% and 13% of Illumina and Affymetrix SNPs are in high linkage disequilibrium (LD) with a pfSNP (r^2>0.8, data not shown).

The gross chromosome coverage of the three different SNP-selection datasets was then examined. As evident in **Figure 2.9**, the SNP coverage for pfSNPs (red), t-SNPs (green) and QR-SNPs (blue) are similar and even across all chromosomes, except for the sex chromosomes. The coverage and distribution of pfSNPs in the X and Y chromosomes are slightly reduced as compared with the t-SNPs/QR-SNP sets,

which is probably due to the reduced density of validated genes in these chromosomes.

Figure 2.8: Venn diagram showing the number of common SNPs amongst the three different SNP-selection platforms. The three SNP selection platforms are the currently reported putatively functional SNPs (pfSNPs), the Illumina tag-SNPs (t-SNPs) and the Affymetrix quasi-random SNPs (QR-SNPs).



Since SNPs within the pfSNP set were selected from the dbSNP129 database without any consideration for the allele frequency of the SNPs, it is not surprising that nearly 50% of the pfSNPs have not been genotyped in the HapMap (**Table 2.7**). Of the non-monomorphic genotyped SNPs, the distribution of MAF amongst the three SNP sets was quite similar, except that more pfSNPs have a slightly lower MAF (**Figure 2.10**).

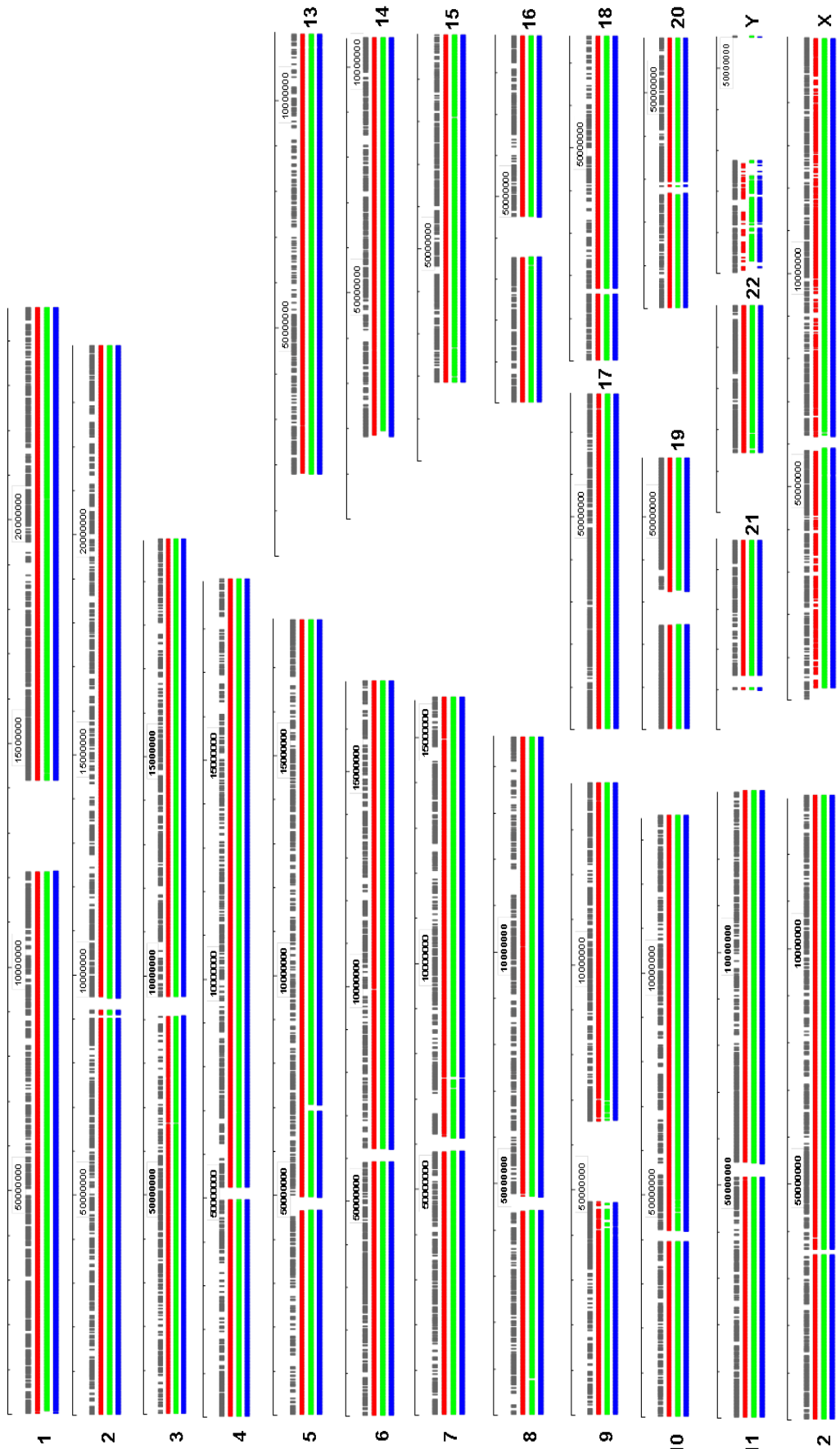
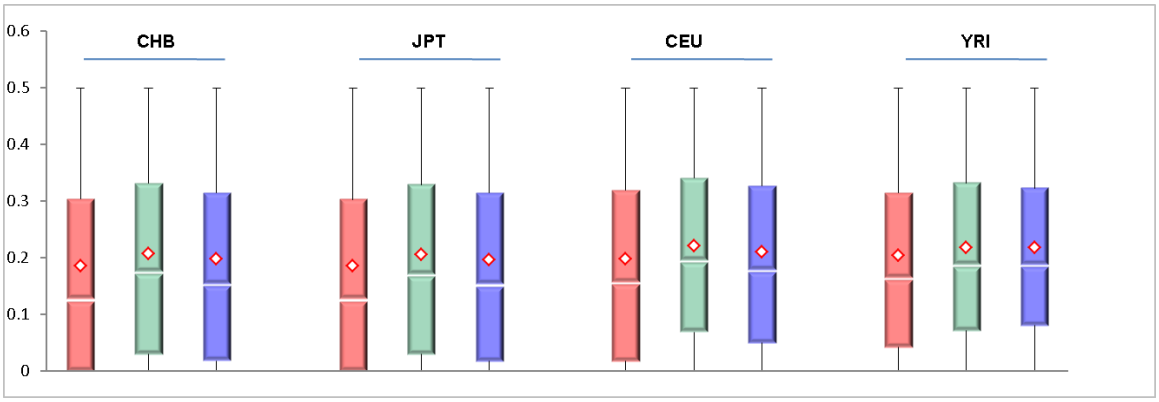


Figure 2.9: Chromosome coverage of the three SNP-selection platforms. One pixel represents 250 kilobases. GRAY track shows NCBI Ref Seq mRNA spanned regions. RED track depicts region that contains pfSNP. GREEN and BLUE tracks represent regions containing the t-SNPs and the QR-SNPs, respectively.

Table 2.7: The percentage of SNPs from the three different SNP-selection platforms that were not genotyped in the four HapMap populations.

	pfSNP				t-SNP				QR-SNP			
	CHB	JPT	CEU	YRI	CHB	JPT	CEU	YRI	CHB	JPT	CEU	YRI
% Not Genotyped	47.89	47.88	47.98	48.22	4.57	4.53	4.11	3.76	0.74	0.75	0.80	1.16

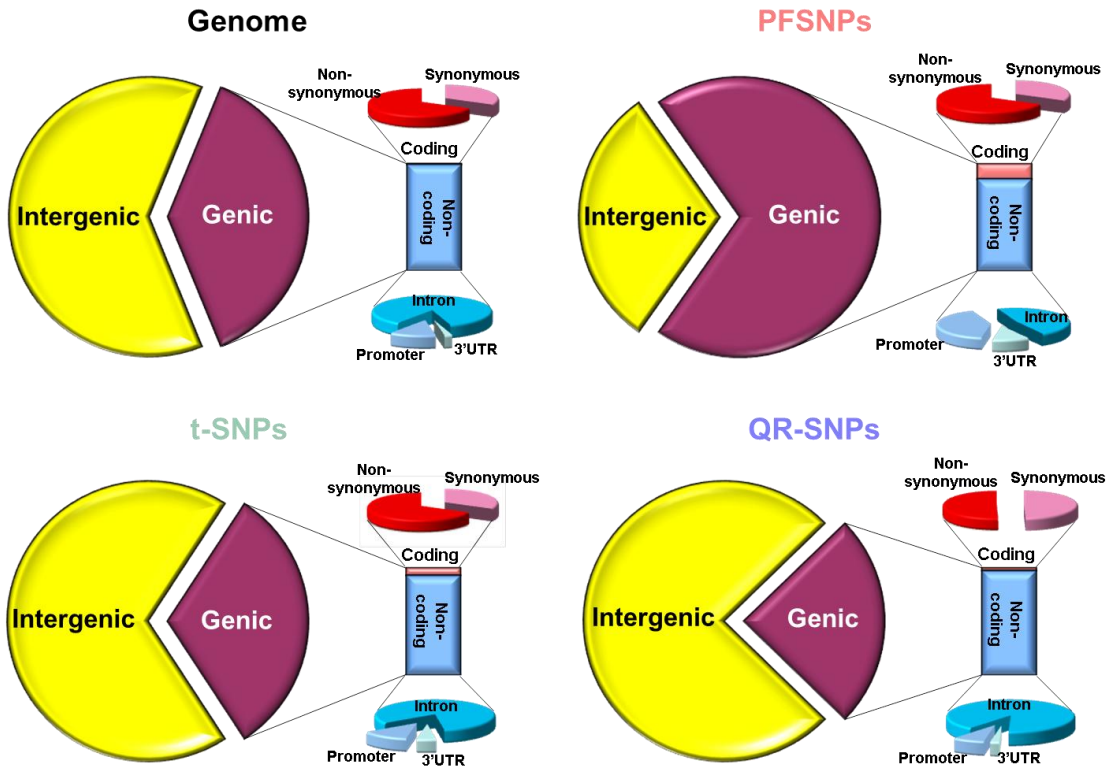
Figure 2.10: Box-and-whisker plot showing the lowest, lower quartile, median, mean (diamond), upper quartile, and highest MAF in the 4 different populations from the three different SNP-selection platforms. pfSNP is in red, t-SNP is in green and QR-SNP is in blue.



The genomic distribution of the SNPs in the 3 SNP sets was then compared (**Figure 2.11**). Unlike the t-SNP and QR-SNP sets, the majority of the pfSNPs (>75%) were found in the genic region. Within the genic regions, pfSNPs (10.36%) were more enriched in coding regions as compared with t-SNPs (6.27%) or QR-SNPs (2.43%). Moreover, a greater percentage of coding pfSNPs (67.26%) was non-synonymous as compared with coding t-SNPs (57.59%) or coding QR-SNPs (50%). In the genic non-coding regions, a greater percentage of pfSNPs was localized in the promoter (27.81%) and 3'UTR (7.32%) regions as compared with t-SNPs (11.07% and 4.57%) and QR-SNPs (7.22% and 2%). Hence, both commercial SNP sets, especially the t-SNP set, appear to reflect the relative distribution of all SNPs in the human genome, while pfSNPs are more enriched in regions of potential functional significance like promoter and 3'UTR. Of the two commercial SNP sets, the t-SNP

set appears more enriched in regions of potential functional significance compared with the QR-SNP selection strategy.

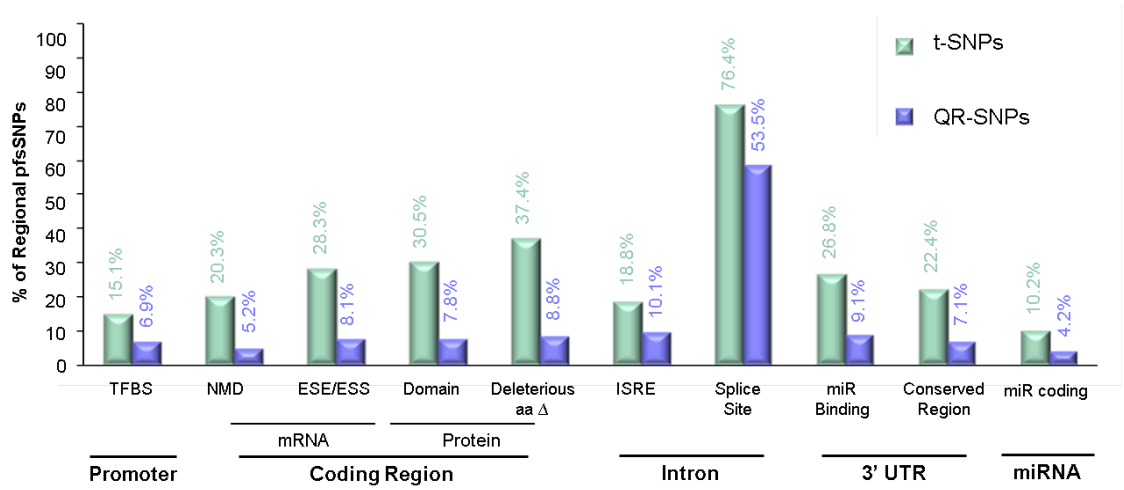
Figure 2.11: Distribution of SNPs in the various genomic regions from the pfSNP, t-SNP and QR-SNP datasets. Pie-chart shows the proportion of SNPs in the inter-genic (yellow) versus genic (maroon) regions. Within the genic region, the proportion of SNPs in the coding (peach) and non-coding (blue) regions is indicated as colored bars. The proportion of genic synonymous (pink) and non-synonymous (red) coding SNPs are shown as pie-charts above the coding SNP bar. The proportion of intronic (dark blue), promoter (light blue) and 3'UTR (green) SNPs are represented as a pie-chart below the non-coding SNP bar.



The proportion of potentially functional SNPs (pfSNPs) that are represented in the t-SNP and QR-SNP sets was then investigated (**Figure 2.12**). Except for SNPs predicted to potentially alter splice sites which are represented at >50% in the two platforms, all the other potentially functional SNPs are represented at <40% in both the t-SNP and QR-SNP platforms. SNPs that are predicted to have potential

functional significance seem to be better represented in the t-SNP set (10-38%) compared with the QR-SNP set (<10%).

Figure 2.12: Proportion of potentially functional SNPs (pfSNPs) represented in the t-SNP and QR-SNP selection platforms. The proportion of putatively functional SNPs (pfSNPs) within the various regions (promoter, coding, intron, 3'UTR and miRNA) predicted to affect the specified function (e.g. TFBS, NMD, etc.) that are represented in the t-SNP (green) and QR-SNP (blue) selection platforms are shown.



Due to the limited knowledge of SNP function and the availability of algorithms used to identify pfSNPs, there might be more pfSNPs yet to be discovered. Therefore, it would be interesting to test if the SNPs in the current pfSNP collection would be good tagging SNPs so that they would cover other pfSNPs yet to be identified. **Table 2.8** shows the tagging efficiency of the polymorphic pfSNP markers in different HapMap populations. Tagging efficiency is measured as the average number of SNPs covered per marker. **Table 2.8** shows that the tagging efficiency of pfSNP is generally comparable to t-SNP set except for the YRI population.

Table 2.8: The tagging efficiency of pfSNP in different HapMap populations compared to t-SNP using HapMap Release 23 data.

	CEU			ASN			YRI		
	# Polymorphic Marker	# SNP covered r2>0.8	# SNP covered Per Marker	# Polymorphic Marker	# SNP covered r2>0.8	# SNP covered Per Marker	# Polymorphic Marker	# SNP covered r2>0.8	# SNP covered Per Marker
t-SNP	0.83 M	2.49 M	3.00	0.79 M	2.36 M	2.99	0.85 M	2.19 M	2.59
pfSNP	0.44 M	1.28 M	2.87	0.43 M	1.26 M	2.94	0.47 M	0.75 M	1.75

Since the tagging efficiency of pfSNP is good and there were less than 0.5M polymorphic pfSNPs in each HapMap population (**Table 2.8**), it may be possible to add additional tagging SNPs outside the 0.5M pfSNP set to cover the whole genome. In later sections, this tagging method is called "pfSNP-centric tagging". **Figure 2.13** shows the number of "pfSNP-centric tagging" markers needed to cover all the SNPs in the HapMap Release 23, with different minimal r^2 values. Around 1M SNPs would be able to cover all HapMap Release 23 SNPs with an r^2 more than 0.6 in the YRI population. Although r^2 of 0.6 would be considered low in practice (Pe'er, de Bakker et al. 2006), this is still better than the t-SNP set, which would only cover all the SNPs in the YRI population at a much lower r^2 value of 0.1 (**Figure 2.14**). For the CEU and ASN populations, 1M "pfSNP-centric tagging" SNPs would cover all the HapMap Release 23 SNPs at high r^2 value of 0.8 and 0.9 (**Figure 2.13**), respectively. This is much better than the t-SNP set, which covers all the SNPs only at an r^2 value of 0.2 (**Figure 2.14**).

Figure 2.13: The number of "pfSNP-centric tagging" SNPs needed to cover all of the HapMap Release 23 SNPs at different minimal r^2 values in different populations.

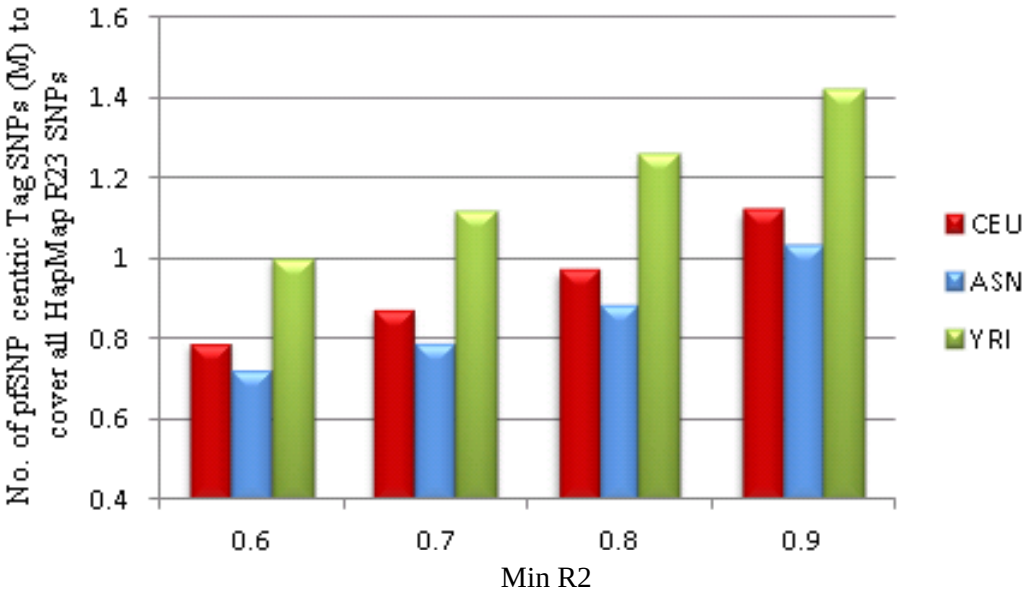
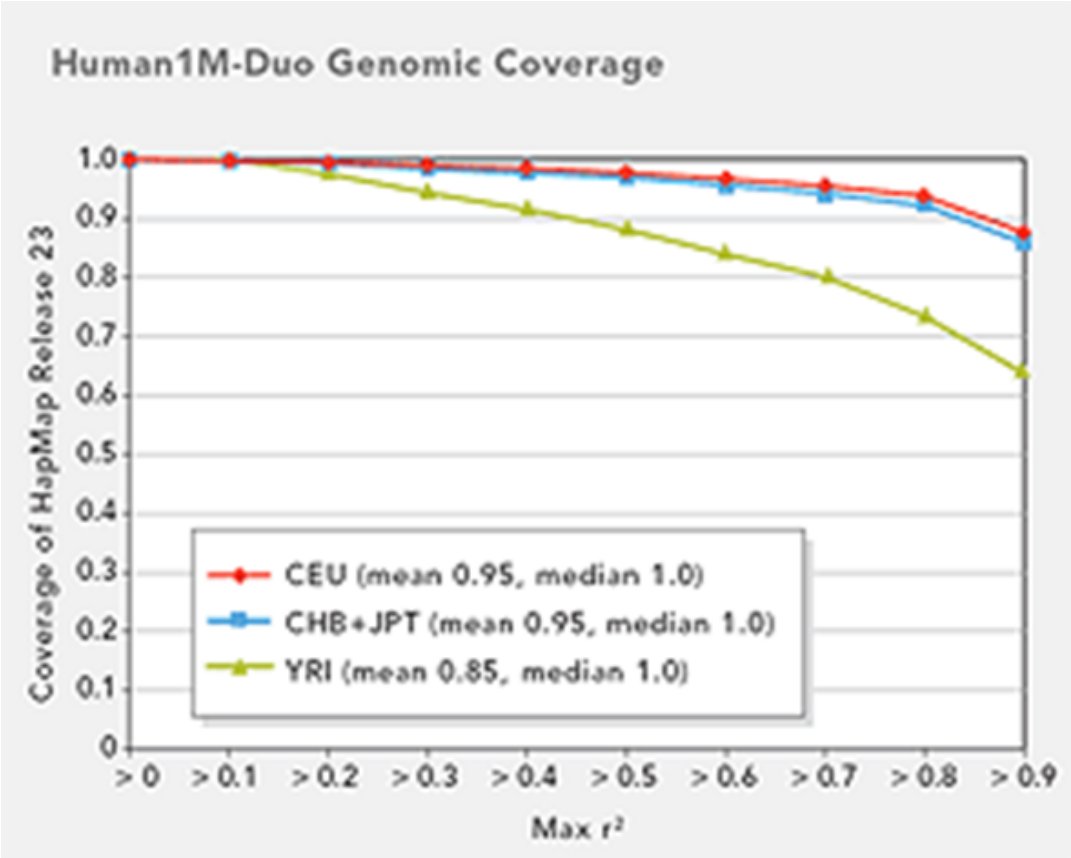


Figure 2.14: The genomic coverage of Illumina t-SNP set.



2.4 Summary

To enable accurately identifying pfSNP at genome scale, I proposed a few new methods to extend the pfSNP coverage outside of the gene coding region or to enhance the accuracy of existing methods. A semi-automated pipeline for collecting pfSNP from various resources was established with three major tools developed in the Visual Basic family of languages and a collection of Perl scripts to facilitate data processing. The tools are designed with an emphasis on flexibility and extensibility in terms of incorporating new functionalities easily.

A database for storing and querying the pfSNP collection was established as a relational database with star schema. This database currently holds more than 30 GB of data and uses Microsoft SQL Server 2005 as the relational database management system. It can be easily queried by using criteria from one or all layers of the SNP-gene-pathway hierarchy. Compared with other SNP functional data integration projects and pipelines, this semi-automated pipeline targets more data sources and has greater data processing capability. The data warehousing solution is also seen as a better alternative to the federated option for data integration in term of query performance.

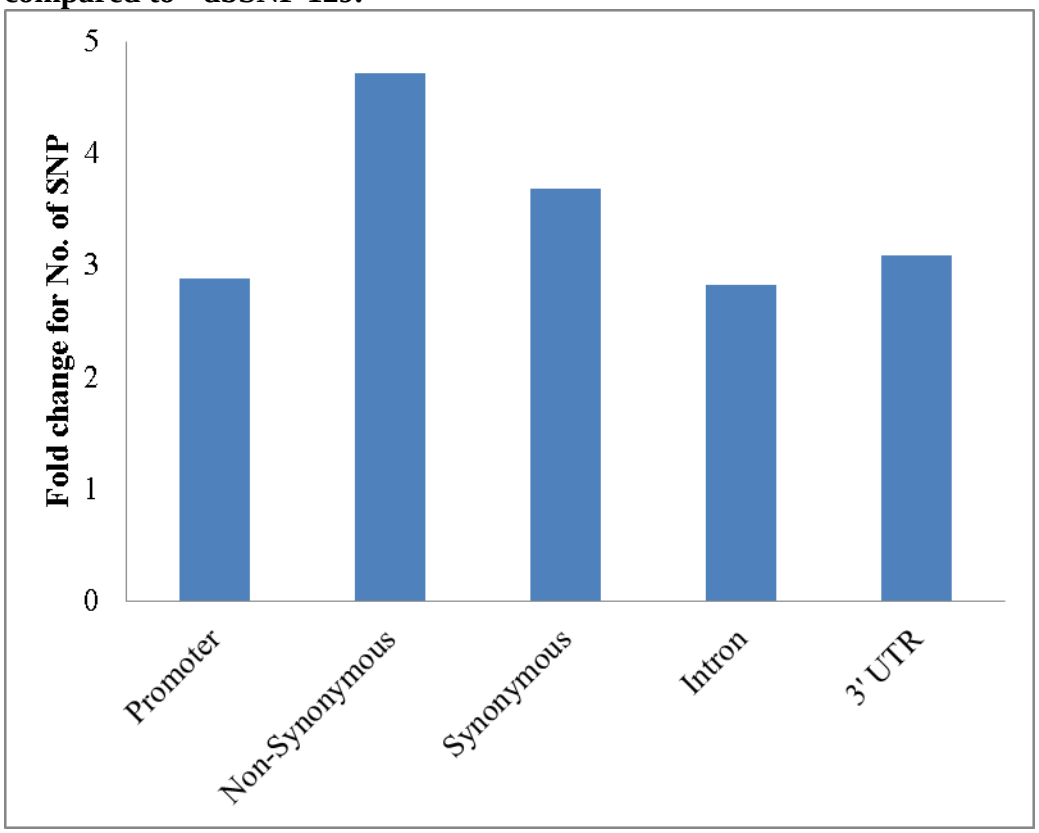
Through the integration of various databases and resources, >900,000 SNPs out of >14 million unique non-redundant SNPs in the dbSNP129 database were of potential functional significance (pfSNPs). Due to the current paucity of information and algorithms to predict functionality in the non-genic regions, >75% of pfSNPs reside in genic regions with significant enrichment pfSNPs in the promoter, coding, and 3'UTR regions of genes, but not in introns or inter-genic regions. Interestingly, a very high percentage of dbSNP129 promoter and coding SNPs (>60%) were predicted to affect

TFBS and ESE/ESS, while a very low percentage of coding and intronic SNPs (≤1%) were predicted to affect NMD or splice sites.

A comparison of the pfSNP set with the popular commercially available t-SNP (Illumina) and QR-SNP (Affymetrix) sets revealed that the three SNP sets were unique with less than 7% of SNPs common amongst the 3 platforms. Nonetheless, the gross chromosome coverage of the three different SNP sets was similar. Due to the strategy of selection, the pfSNP set has the greatest proportion of SNPs (~50%) that were not genotyped in HapMap Release 23 compared with the other sets (<5%). Of the genotyped SNPs, the pfSNP set also contained the greatest percentage of monomorphic SNPs (15-27%) compared with the t-SNP (7-18%) and the QR-SNP (5-21%) sets. Nevertheless, the current pfSNP set may act as good tagging SNPs with comparable tagging efficiency compared to Illumina t-SNP set. With additional tagging SNPs picked outside of the pfSNP set, the "pfSNP-centric tagging" SNP set would cover all the SNPs in HapMap Release 23 with higher r^2 value compared to Illumina t-SNP set.

As the adoption of next generation sequencing technique has grown drastically along the years, more and more SNPs have been identified. Currently, the number of SNP in each gene region has increased to be more than 3 times compared to dbSNP 129 from which I generated the list of pfSNPs (**Figure 2.15**). Noticeably, more SNPs in the coding region which include non-synonymous SNP and synonymous SNP have been reported compared to other regions. In the current release of pfSNP collection, pfSNP identified by any method would be called a pfSNP. To minimize the false positive rate for pfSNP identification, it is possible to choose those predicted by more than one algorithm.

Figure 2.15: The fold change of SNP numbers in each gene region in dbSNP 137 compared to dbSNP 129.



CHAPTER 3: DEVELOPMENT OF A PFSNP WEB-RESOURCE

(Adapted from "pfSNP: An integrated potentially functional SNP resource that facilitates hypotheses generation through knowledge syntheses" Hum Mutat. 2011 Jan;32(1):19-24)

3.1 Introduction

As of 2010, a few web-resources had been developed that attempted to integrate SNP function information and facilitate a better annotation of the functionality of SNPs in the human genome. For example, one of the earliest resources is the SNPselector (Xu, Gregory et al. 2005), which integrates only a few resources to prioritize a list of SNPs based on their linkage disequilibrium (LD) tagging potential, allele frequencies, source, function, regulatory potential and repeat status. However, the resource has been retired from public service as of 2010. One of the more comprehensive tools is the F-SNP (Lee and Shatkay 2008); this tool integrates 16 bioinformatic resources to annotate the functionality of ~500,000 SNPs in ~18,000 genes associated with 85 major human diseases from OMIM. SNPLogic (Pico, Smirnov et al. 2009) is another web-resource that combines <15 bioinformatic tools to annotate potentially functional SNPs. The user can construct lists of SNPs based either on genes, chromosome region or pathways (including OMIM, Biocarta, GO, Biocyc, Wikipathway, etc). Another SNP integration tool, SNPit (Shen, Carlson et al. 2009) employs less than five bioinformatic tools to annotate the functionality of SNPs, based on functional impact of non-synonymous coding SNPs, evolutionary conservation of non-coding regions as well as positive selection. Yet another tool, the GWAS Analyzer (Fong, Ko et al. 2010), links statistical results from association studies, HapMap data, known genes (UCSC Genome Bioinformatics), microRNA

(Sanger miRBase), splice sites, and expression data from microarray experiments (GENEVAR project).

Nonetheless, there is still a need for a single comprehensive, integrated SNP web-resource that will facilitate hypotheses generation through knowledge synthesis mediated by better data integration and a biologist-friendly web interface. Ideally, this resource should provide all the pertinent information about the SNPs in the human genome, including allele frequency and LD information of SNPs. Furthermore, this resource should comprehensively integrate all the available resources that predict the potential functionality of SNPs based not only on predicted functionality due to sequence motifs but also inferred functionality from a genetic approach and reported functionality and disease association from published reports or association studies or expression data. To facilitate hypotheses generation through knowledge syntheses, the query interface should be biologist-friendly and highly customizable to facilitate several combinations of queries. Additionally, the query results interface should be similarly user-friendly and also integrate tissue distribution as well as pathway information where enriched tissues/pathways are highlighted, and detailed related information of the query results are provided.

I have thus developed the pfSNP web-resource to address the above-mentioned need. The pfSNP resource builds on the pfSNP database mentioned in the previous chapter. Significantly, to facilitate hypotheses generation through knowledge syntheses, the pfSNP resource has a biologist-friendly, highly customizable query interface as well as a user-friendly results interface that integrates gene, pathway and tissue distribution information with text clouds highlighting the enriched tissues/pathways. Hyperlinks providing detailed related information about specific results in the result interface are also given. It is anticipated that this pfSNP

resource will be useful to most researchers interested in SNPs, including those interested in association studies and those focusing on designing experiments to address SNP functionality in association with a phenotype.

3.2 Materials and methods

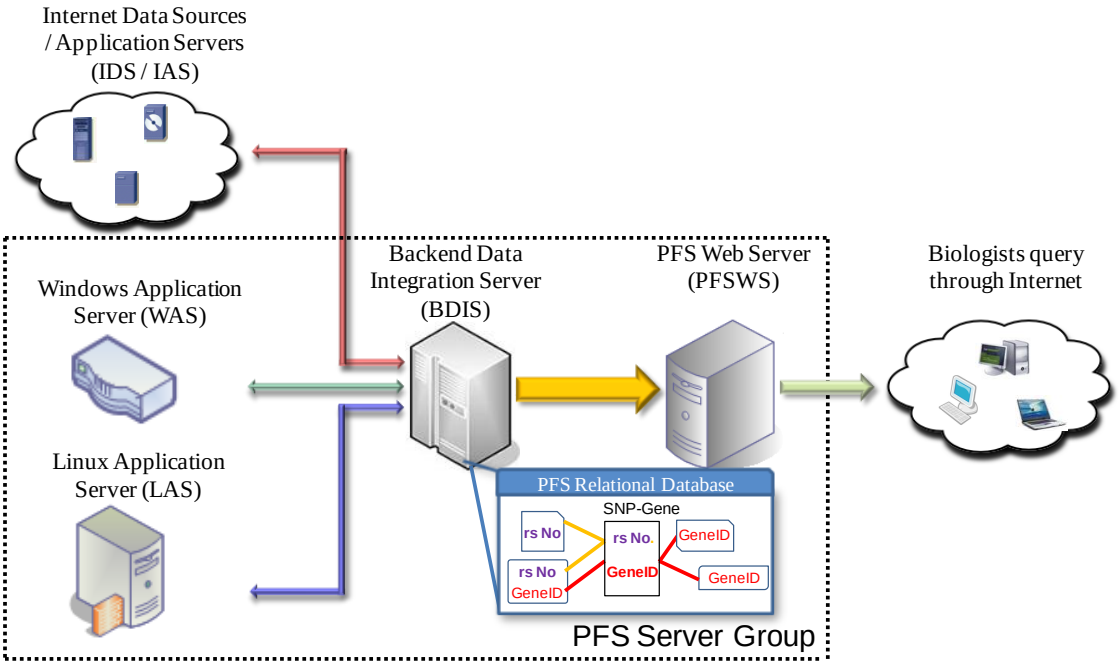
pfSNP web-resource (http://pfs.nus.edu.sg/) integrates >40 different algorithms/resource to evaluate the potential functionality of SNPs from the dbSNP database (http://www.ncbi.nlm.nih.gov/projects/SNP/), based on previous published reports, inferred potential functionality from genetics approaches, and predicted potential functionality based on sequence motifs (**Table 3.1**).

The design of pfSNP web-resource is based on the snowflake relational database schema centered on a SNP-Gene (rs number and Gene ID) mapping table to which various other data tables are linked either via the rs number or the Gene ID or both (**Figure 3.1**). This schema has the advantage of facilitating queries or combinations of queries based on different SNPs (including genomic, mRNA or protein location) and/or Gene (tissue of expression, biological pathways, etc) information. It also ensures appropriate SNP annotation, even when the same SNP is found to reside in two different location/genes. For example, the pfSNP resource will inform the user that SNP rs1048913 actually resides in two different genes and the effect of that SNP in each gene will be given. Additionally, new algorithms for potential functionality can readily be included as new data tables linked to the same SNP-Gene mapping table as they are discovered.

Table 3.1: The data sources used in pfSNP web-resource

Molecule Level	SNP Cat.	Function Cat.	Method Short Description	References	Resource Link
SNP	All SNP	Population Frequency	HapMap		http://www.hapmap.org
		Disease/Function Related	OMIM reported SNPs	Julia Stoyanovich et al, Bioinformatics, 2008, 24:440	
			PubMed Reported SNPs	Weijian Xuan et al, Bioinformatics, 2007, 23:2477	
			NCBI dbGaP Reported SNPs		http://www.ncbi.nlm.nih.gov/entrez/query/Gap/gap_tmpl/about.html
			HGMD for Reported SNPs		www.hgmd.cf.ac.uk
			A Catalog of Published Genome-Wide Association Studies	Hindorff LA et al, PNAS, 2009 Jun 9;106(23):9362	http://www.genome.gov/gwastudies
		Positively Selected	mLRH (modified LRH)	K Tang et al, Hum Mol Genet, 2004, 13:783	
				Wang Z et al, Hum Mol Genet. 2005, 14:2075	
				Wang Z et al, Hum Mol Genet. 2007, 16:1367	
				Unpublished	
			WGLRH	Chun Zhang et al, Bioinformatics, 2006, 22:2122	http://portal.acm.org/citation.cfm?id=1182321
			LDD (LD Decay)	ET Wang et al, PNAS, 2006, 103:135	
			"iHs"	BF Voight et al, Plos Biology, 2006, 4:446	http://haplotter.uchicago.edu/selection/
			Fst	JM Akey et al, Genome Research, 2002, 12:1805	
			HapMap SNP with Fst value in the top 1% among all HapMap SNPs		
			abnormal CNC Regions	Su Yeon Kim et al, Plos Genetics, 2007, 3:1572	
			"Human Accelerated Region"	KS Pollard et al, Plos Genetics, 2006, 2:1599	
			Accelerated Human CNC Region	S Prabhakar et al, Science, 2006, 314:786	
			Accelerated Human CNC Region	EC Bush et al, BMC Evol Biol, 2008, 8:17	
		Evolutionarily Conserved	Ultra Conserved Region	G Bejerano et al, Science, 2004, 28;304(5675):1321	
		Expression Level Assoc	Affy Exon Tiling Array	Tony Kwan et al, Nat Genetics, 2008, 40:225	
			Illumina Sentrix Chip	BE Stranger et al, Science, 2007, 315:848	
mRNA	Promoter Region	Change TF Binding sites	TransFac		http://www.biobase-international.com/pages/index.php?id=transfac
	Coding Region SNP	Affect NMD	Nonsense Mediated Decay	Nagy et al, Trends Biochem. Sci, 1998, 23:198	
		Change ESE/ESS Site	RESCUE-ESE	Fairbrother WG et al, Science, 2002, 297:1007	http://rulai.cshl.edu/cgi-bin/tools/ESE3/esefinder.cgi?process=home
			Detect enrichment compared to 5UTR and Intronless exons	Zhang XH et al, Gene and Development, 2004, 18:1241	http://genes.mit.edu/burgelab/rescue-ese/
			Hexmer liberary screening	Wang Z et al, Cell, 2004, 19:831	http://ast.bioinfo.tau.ac.il/
			Experimentally validated ESE/ESS	Zhi-Ming Zhang, J Biomed Sci, 2004, 11:278	http://cubweb.biology.columbia.edu/pesx/
	Intronic SNP	Change ISRE sites	Conserved intronic sequence within 400bps of splicing site	Yeo GW et al, Plos Genetics, 2007, 3:814	
		Reside on Splice Site	Cause abberant 5' site (SD-Score)	Sahashi K et al, Nucleic Acid Res, 2007, 35:5995	
			Change 3' Consensus sequence		
	3UTR SNPs	Change Mir Binding	MiRanda to predict binding change and also functional validation	Saunders MA et al, PNAS, 2007, 104:3300	
			PolymiRTS Database	L Bao et al, Nucleic Acids Res, 2007, 35:D51	http://compbio.utmem.edu/miRSNP/
			Patrocles Database	M Georges et al, Cold Spring Harb Symp Quant Biol. 2006, 71:343	http://www.patrocles.org/
		In 3UTR Conserved Region	Functional region predicted multiple species sequence alignment	Xiaohui Xie et al, Nature, 2005, 434:338	
		Change TF Binding sites	TransFac		http://www.biobase-international.com/pages/index.php?id=transfac
	5Kb Downstream region of gene	Change TF Binding sites	TransFac		http://www.biobase-international.com/pages/index.php?id=transfac
	miR Promoter / Coding Region SNP	Change TF Binding sites	TransFac		http://www.biobase-international.com/pages/index.php?id=transfac
		Reside on miRNA seed, etc region	NCBI and miRBase		
			Patrocles Database	M Georges et al, Cold Spring Harb Symp Quant Biol. 2006, 71:343	http://www.patrocles.org/
Protein	Coding Region SNP	Protein Domain / Functional Sites	Interpro Scan	Zdobnov EM et al, Bioinformatics. 2001, 17:847	http://www.ebi.ac.uk/InterProScan/
			Tm HMM (Transmembrane domain)		http://www.cbs.dtu.dk/services/TMHMM/
			NetOGlyc (O Glycosylation site)		http://www.cbs.dtu.dk/services/NetOGlyc/
			NetNGlyc (N Glycosylation site)		http://www.cbs.dtu.dk/services/NetNGlyc/
			NetPhos (Phosphorylation site)		http://www.cbs.dtu.dk/services/NetPhos/
		Non-Synonymous SNP	Polyphen	V Ramensky et al, Nucleic Acids Res, 2002, 30:3894	http://genetics.bwh.harvard.edu/pph/
			SNP3D	P Yue et al, BMC Bioinformatics, 2006, 7:66	http://www.snps3d.org/
			LS-SNP	R Karchin et al, Bioinformatics, 2005, 21:2814	http://alto.compbio.ucsf.edu/LS-SNP/
			Panther-SNP	PD Thomas et al, PNAS, 2004, 101:15398	http://www.pantherdb.org/tools/csnpScoreForm.jsp
			All Non-Synonymous SNP		
		Synonymous SNP	Codon usage differences falling in top and last 5%		http://www.kazusa.or.jp/codon/
Gene	N.A	Gene alias	NCBI GeneHistory table		ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/
		GO Term	Gene Ontology		ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2go.gz
		KEGG Pathway	Kyoto Encyclopedia of Genes and Genomes		ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/
		Gene Tissue Expression	Gene Expression Atlas		http://symatlas.gnf.org/SymAtlas/
		mSigDB C2	Molecular Signature Database (C2 collection)		http://www.broad.mit.edu/gsea/msigdb/collections.jsp#C2
		Genetic Association Database	Genetic Association Database		http://geneticassociationdb.nih.gov
		PharmGKB	Pharmacogenetics and Pharmacogenomics Knowledge	Tina HB et al, Nuclear Acids Research, 2008, 36:D913	http://www.pharmgkb.org/

Figure 3.1: The hardware and software architecture of the pfSNP web-resource



The pfSNP web portal is implemented as a 3-tier web application using ASP.NET 3.5 with extensive AJAX (Asynchronous Javascript and XML) functionality. The 3-tier architecture ensures clear separation of query logic and web site presentation, which not only eases the maintenance of the site but is also crucial to the functionality of the web portal. In this web portal, various new queries focusing on a subgroup of original query results (e.g., the queried SNPs may belong to multiple genes and user wants to focus on SNPs in a specific gene) or providing extensive supporting information (e.g. view pfSNPs in high LD with the SNPs queried for) may be formed naturally as the user navigates through the query results. The separation of query logic will facilitate the initiation of a new query without losing track of the original query results. AJAX not only enables users to re-arrange their query criteria visually (which is the key functionality of the web portal) but also ensures a rich user experience, which will be discussed in detail later.

3.3 Results

3.3.1 Biologist-Friendly Features of the pfSNP Web-Resource

The web interface was designed to provide features that are biologist-friendly using ASP.NET 3.5 with AJAX (Asynchronous Javascript and XML) for an enriched user experience.

1. *Auto-Complete Prompt-As-You-Type feature in the Query Interface*. The pfSNP database integrates information from a variety of resources that have different terminologies for the same phrase (e.g. Parkinson's Disease versus Parkinson Disease). As the list of the terms can be very long, it is cumbersome to provide a drop-down list for the user to select. This Auto-Complete Prompt-As-You-Type feature in the query interface relies on ASP.NET 3.5 web-service to provide the user with a list of query terms contained within the various data sources as the user begins to type so that the user can rapidly select the appropriate term from the list that is most meaningful to him/her (**Figure 3.2**, an online demonstration of this feature as a flash movie can be found at http://pfs.nus.edu.sg/demo_src/supp_info_s1.html). This feature not only helps the user to quickly type their term of interest, but prevents the user from typing a specific term that is not the exact term found in the query database; this may lead to no results for his/her query.
2. *Retrieved all related terms to query*. It is also possible to retrieve and query a number of related terms containing or beginning with a keyword (**Figure 3.3A**, an online demonstration of this feature as a flash movie can be found at http://pfs.nus.edu.sg/demo_src/supp_info_s1.html). A list of

related terms will be displayed so that the user can select several or all of the relevant related terms for the query (**Figure 3.3B**).

Figure 3.2: Auto-Complete Prompt-As-You-Type feature in the Query Interface. Related term will be retrieved as user key in the query term.

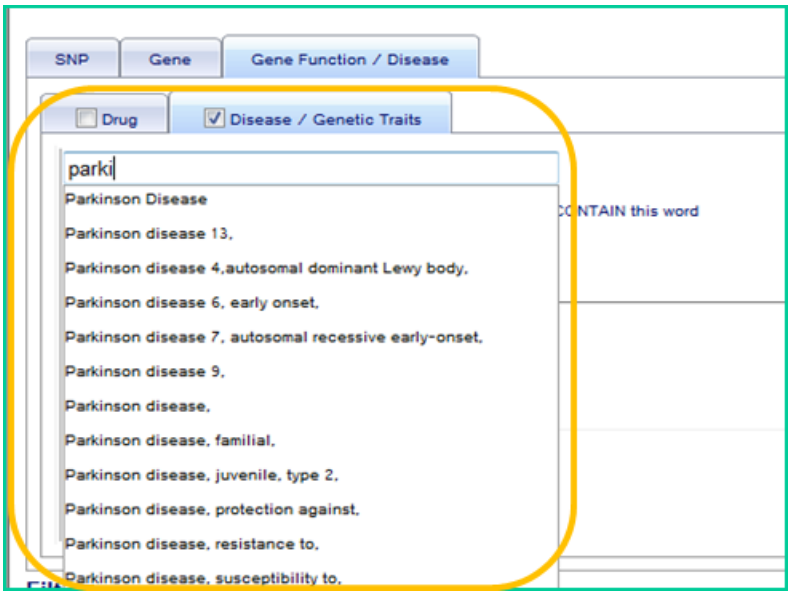
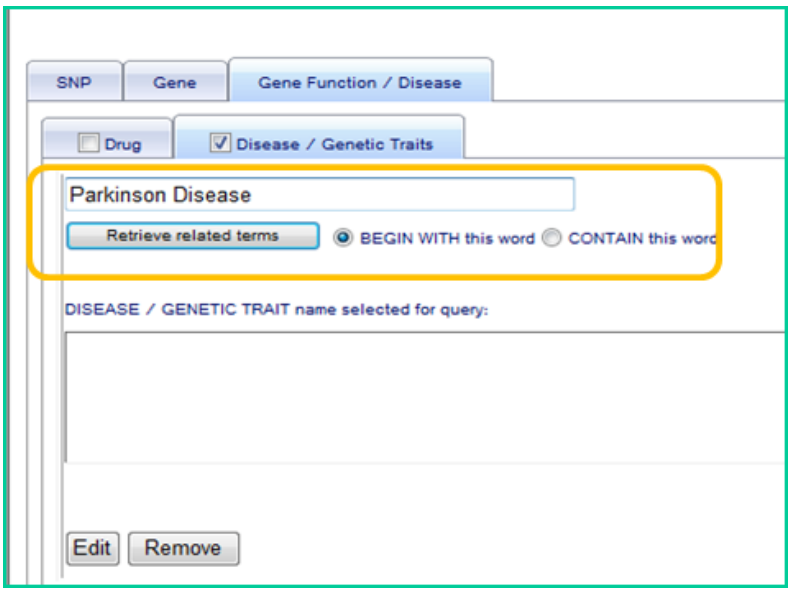
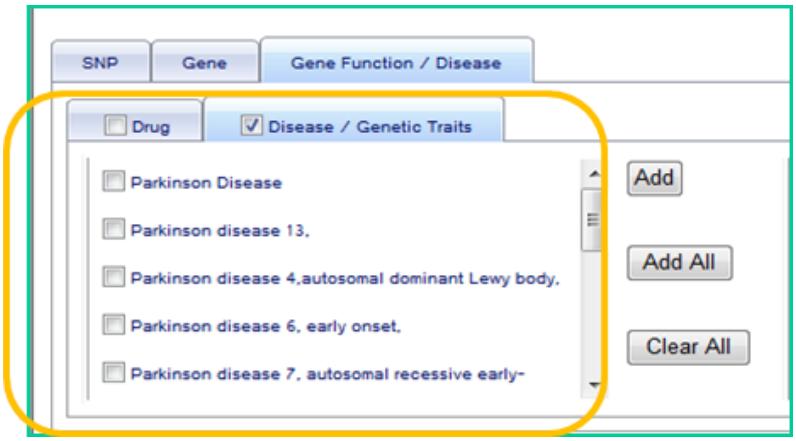


Figure 3.3: Retrieve all related terms to query. A. User can retrieve related terms either beginning with or containing the key word. B. The list of related terms retrieved.

A



B

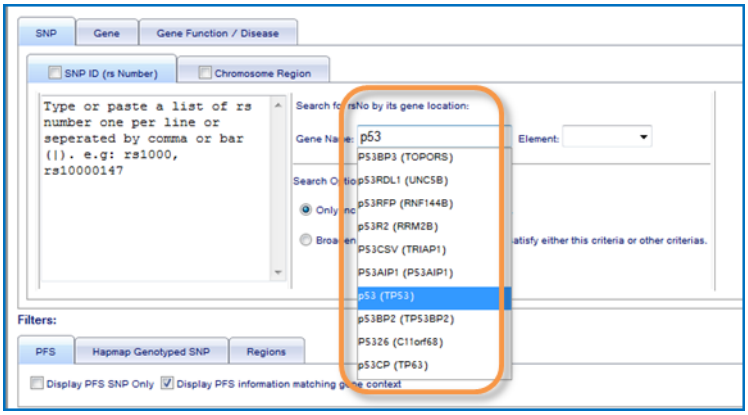


3. *Auto Retrieval of SNP ID*. Another useful feature is that the exact SNP ID (i.e. rs number) need not be known as this web-resource is capable of providing the rs number for any SNP within any gene region so long as pertinent information with regards to the SNP is known. This feature would be very useful, as most publications (except for more recent ones) reporting functionally important SNPs, especially those in the coding region, often provide only the amino acid residue changes (e.g. R273H polymorphism in p53). For example, to convert amino acid residue change to rs number, the SNP ID (rs number) tab can be selected and on the right panel, there will be an option to "Search for rs No. by its gene location". As one begins to type the name of the gene, the Auto-Complete Prompt-As-You-Type feature will provide a pull-down list of gene names and the user can then select the gene name that is the most relevant (**Figure 3.4A**, a step-by-step demonstration of this feature can be found at http://pfs.nus.edu.sg/demo_src/supp_info_s2.html). Thereafter, the user can select the appropriate gene region that the SNP resides in (**Figure.3.4B**). If the SNP resides in the coding region and the amino acid

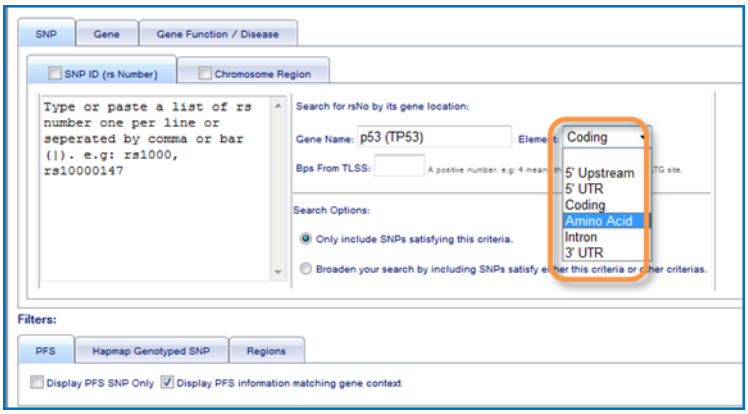
position is known, the user then provides the amino acid position (**Figure 3.4C**) and the rs number will be automatically retrieved and added to the list of rs numbers on the SNP ID panel.

Figure 3.4: The feature to retrieve SNP ID based on its gene context.

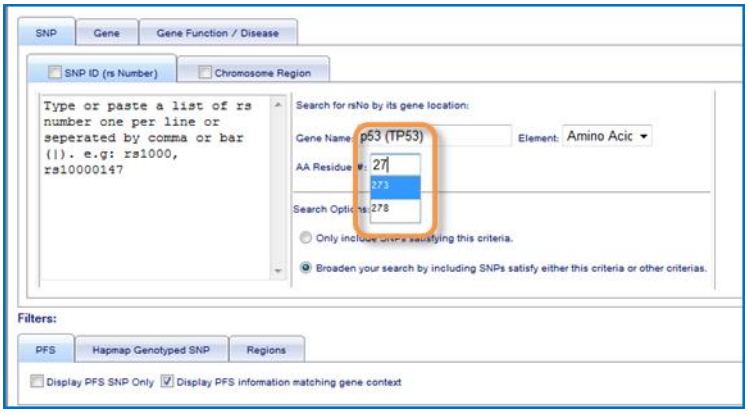
A



B



C



4. *Excel-Like Result Presentation with In-Depth Resource Reference*. The results of queries from this web-resource are presented in a familiar excel-like format. This excel-like presentation has customizable freezing of headers and columns to facilitate reading and referencing. Most of the predicted functions are hyperlinked to either to the original resource itself or a summary of important features of the prediction, including information of the original resource. The query results are also downloadable in a fully hyperlinked well-formatted Excel file.

3.3.2 Features in pfSNP resource that facilitate Hypotheses Generation

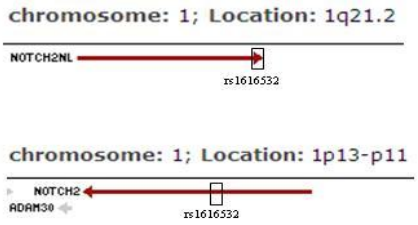
The pfSNP web-resource contains the following features that will facilitate hypotheses generation:

1. *Ambiguity in SNP information appropriately addressed*. Unlike other SNPs tools, all ambiguity in SNP information is appropriately addressed in the pfSNP resource. A single SNP (with a unique rs number) may map to different splice variants or different genes at the same chromosome location (overlapping genes) or to different genes on different chromosome locations. A SNP with a unique rs number, but mapping to different chromosome locations, could merely represent artifactual single nucleotide differences (SNDs) that are caused by the erroneous comparison of two or more duplicated (paralogous) sequences (Day, 2010; Musumeci, et al., 2010). This web-resource will inform the user if a particular SNP is likely to be a SND (**Figure 3.5A**) or if it represents a genuine SNP, mapping to different splice variants/genes at the same

chromosomal location. Detailed information about that SNP in the different splicing variant/gene context will be provided (**Figure 3.5B**).

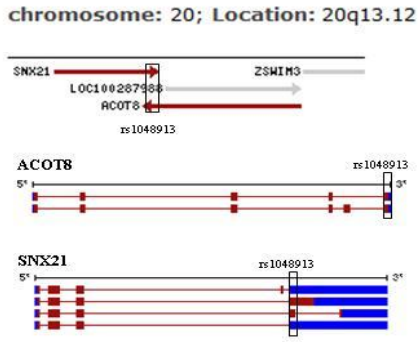
Figure 3.5: Ambiguity in SNP information appropriately addressed.
A. User will be warned if the SNP is SND. B. Detailed information matching gene context will be shown if the SNP is genuine and mapped into different genes and/or splice variants.

A



Serial	rs No.	pfSNP in Highest LD	Gene Name	mRNA Location	#mRNA Location	AA Change
1	rs1616532 Warning: This SNP is probably an SND!	This SNP is pfSNP.	NOTCH2NL N2N	E/4/T417C	1	K139K
2	rs1616532 Warning: This SNP is	This SNP is pfSNP.	NOTCH2 AGS2 hN2	E/4/T534C	1	K178K

B



Serial	rs No.	pfSNP in Highest LD	Gene Name	mRNA Location	#mRNA Location	AA Change
1	rs1048913	This SNP is pfSNP.	ACOT8 HNAACTE PTE-2 PTE1 PTE2	E/5/G747A E/6/G942A	1 1	V249V V314V
2	rs1048913	This SNP is pfSNP.	SNX21 C20orf161 MGC29895 PP3993	E4/3UTR/G1199A E4/3UTR/G543A E5/3UTR/G1199A I/4/G-177A	1 1 1 1	-- -- -- --

2. *Linkage Disequilibrium information of genotyped SNPs provided*. Many current researchers interested in Genome Wide Association Studies (GWAS) have utilized either the Affymetrix or the Illumina SNP chips, which select SNPs quasi-randomly or based on 'tag' SNPs, respectively. Many of the SNPs on these chips shown to be associated with a phenotype during GWAS may not be predicted to be functional. This is because the SNPs examined in the chips may only serve as surrogate SNPs in strong

LD with a causative/functional SNP. One useful feature of this resource is that for any SNP in the human genome that has been genotyped by HapMap, it will inform the user of the distance as well as the LD measured by r^2 between the SNP-of-interest and the closest pfSNP. Details of the potential functionality of that particular pfSNP will also be given. LD information of the nearby genotyped pfSNPs is available for greater than 75% of Illumina and Affymetrix SNPs. This will thus guide current GWAS scientists to make an informed decision about the potential functionality of SNPs in LD with their SNP-of-interest and design appropriate experiments to address them.

3. *Highly Customizable Query Interface*. Biologists often formulate hypotheses about the unknown based on currently known information. For example, to design a gene-based study to associate SNPs with Parkinson's Disease (PD) in Asians, one may wish to identify SNPs in genes that were previously reported to be associated with the disease. On the query interface page, the user can choose the 'Gene Function/Disease' tab and type 'Parkinson Disease' (**Figure 3.6A**). This web-resource utilizes information from previous GWAS studies, HGMD as well as limited information from OMIM and PubMed to inform the user of the genes (and their associated pfSNPs) that may be associated with PD. It may also be possible to limit the focus to SNPs in genes associated with PD that are expressed in only a specific organ, such as the brain, or those that are found in a specific chromosome region or belonging to a specific pathway. This can be done by selecting the 'Gene' and 'Expressed in' Tab. A list of

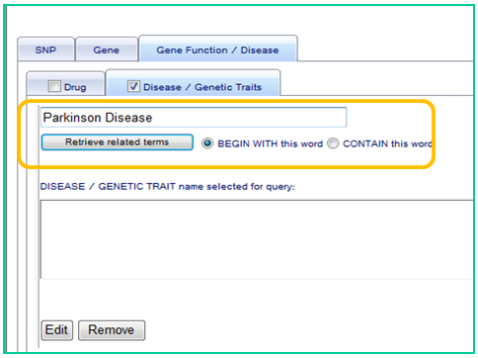
tissue options will be available and the user can then select the tissue of interest; e.g., Brain (**Figure 3.6A**, a step-by-step demonstration of this feature can be found at http://pfs.nus.edu.sg/demo_src/supp_info_s4.html). This web-resource utilizes information from Gene Expression Atlas to enable the user to limit the search to only genes that are expressed in a specific tissue. The user may further limit the search to the genes in a particular pathway. pfSNP web-resource provides a choice of either GO terms, KEGG pathway or mSigDB C2 pathway or a combination of any of these terms.

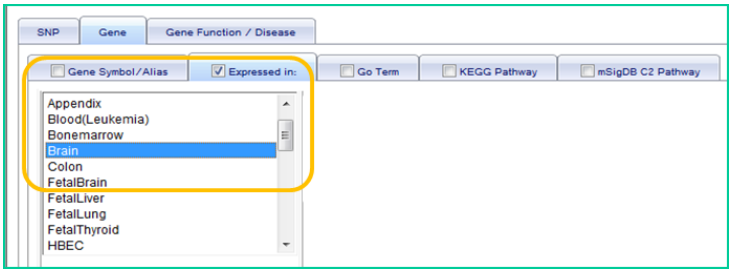
Another unique feature of this web-resource is that users would be able to arrange their query criteria as they wish by "dragging and dropping" the different tabs containing their query criteria and different logic operators to form a highly customized query (**Figure 3.6B**, a step-by-step demonstration of this feature can be found at http://pfs.nus.edu.sg/Demo_Src/ArrangeCriteria.html).

Using this web-resource, it is also possible to further limit the search to SNPs that occur in Asians above a specific threshold minor allele frequency (**Figure 3.6C**). One can also just look at pfSNPs that are located in coding regions, promoters, 5'UTR, 3'UTR, introns or a combination of different regions or all regions. Hence, the query interface of this pfSNP resource is not only user-friendly but highly customizable such that different combinations of queries are possible using Boolean logic.

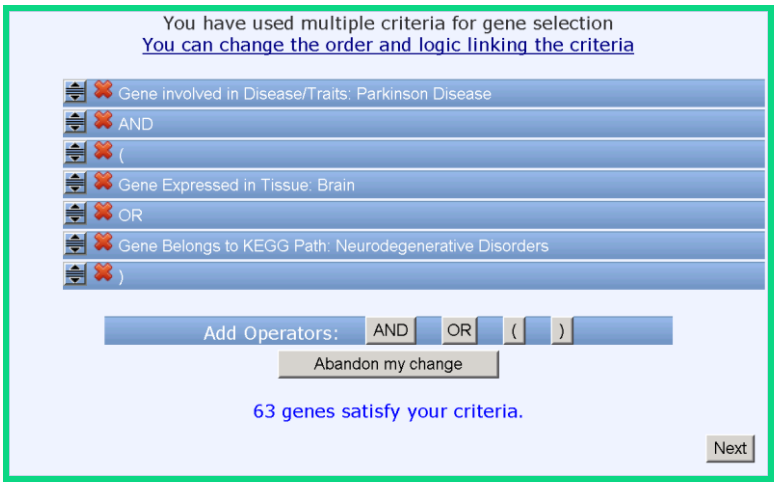
Figure 3.6: Highly Customizable Query Interface. A. Multiple query criteria can be provided in the query interface. B. User can re-arrange the criteria provided and adding appropriate logic operators. C. Multiple filter criteria can be applied.

A

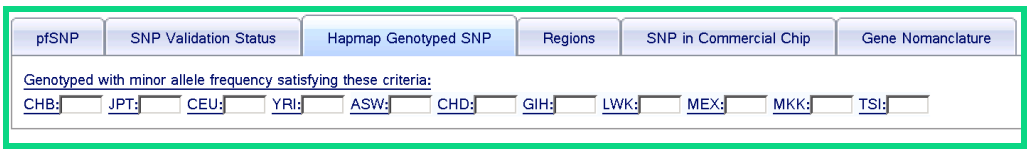


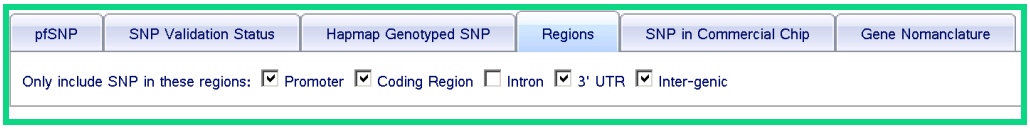


B



C



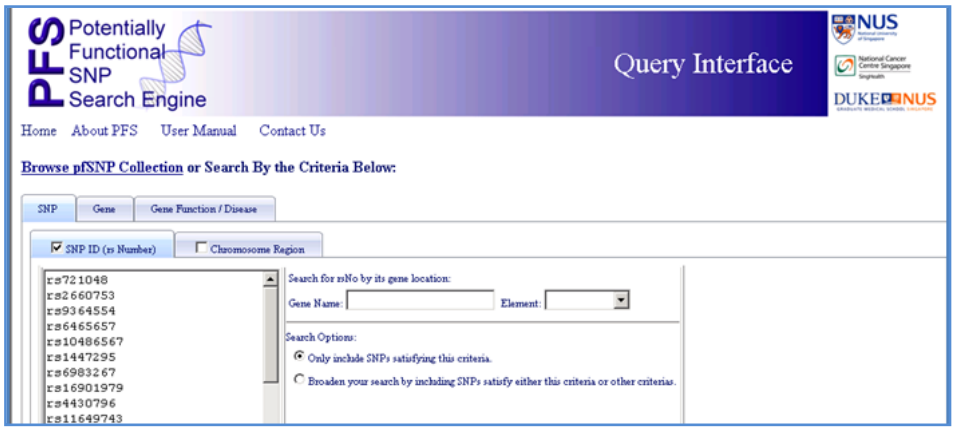


4. *Integration of Gene/Pathway Level Information into the Results Interface*. To facilitate better hypotheses generation, the pfSNP resource integrates gene/pathway level information into the Result interface. This feature would be particularly useful for researchers who have identified a list of SNPs that may be associated with a particular disease/phenotype from, for instance, GWAS and want to know if this list of associated SNPs is enriched in any particular pathway so as to generate some testable hypotheses. A step-by-step demonstration of this feature can be found at http://pfs.nus.edu.sg/Demo_Src/ supp_info_s5.html. When this list of SNPs are queried in the pfSNP database, not only is the potential functionality of the SNP given, the genes associated with these SNPs as well as the tissue distribution of the genes and the pathway/GO term to which these genes belong are also given. Importantly, text clouds are generated to highlight the enriched terms in each category. As an illustration, a list of 16 SNPs that were previously associated with prostate cancer were queried in the pfSNP resource (**Figure 3.7A**). The web-resource highlighted that six of these SNPs are inter-genic while the other 10 SNPs reside in seven genes (**Figure 3.7B**). The Query Result showed information regarding the seven genes and a summary of the properties of the seven genes can be obtained by moving the cursor over the various headers. For example, the text clouds under the header, 'Tissue Expression' and 'mSigDB Pathway' inform the user that majority of these seven genes are expressed in the testis as well as prostate and are in the pathway of hematopoietic stem cell progenitors (HSC_Early Progenitors), since these terms are written in larger fonts compared to the other terms

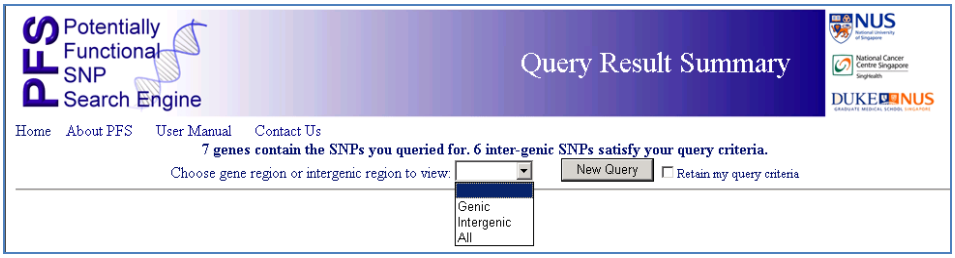
(**Figure 3.7C**). Such integration of information will help scientists to better formulate hypotheses of their findings and focus their efforts on designing appropriate experiments.

Figure 3.7: Integration of gene/pathway level information into the results interface.

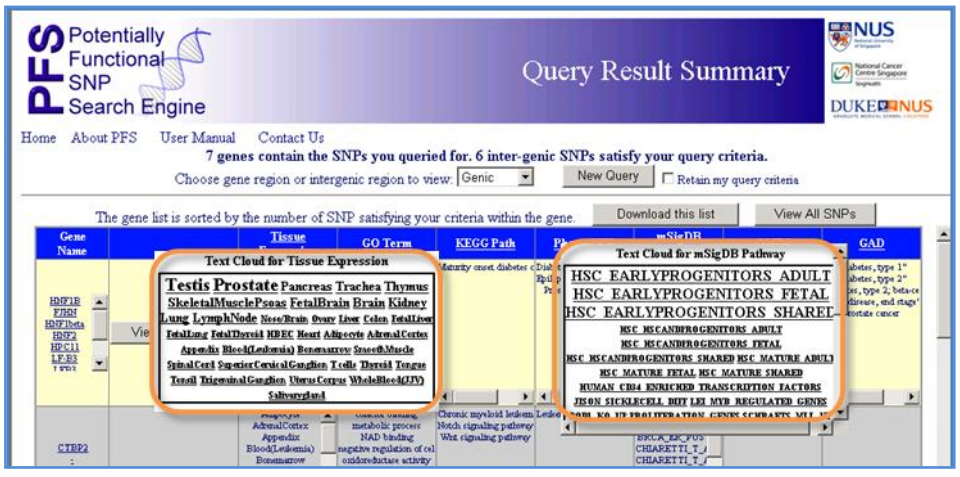
A



B



C

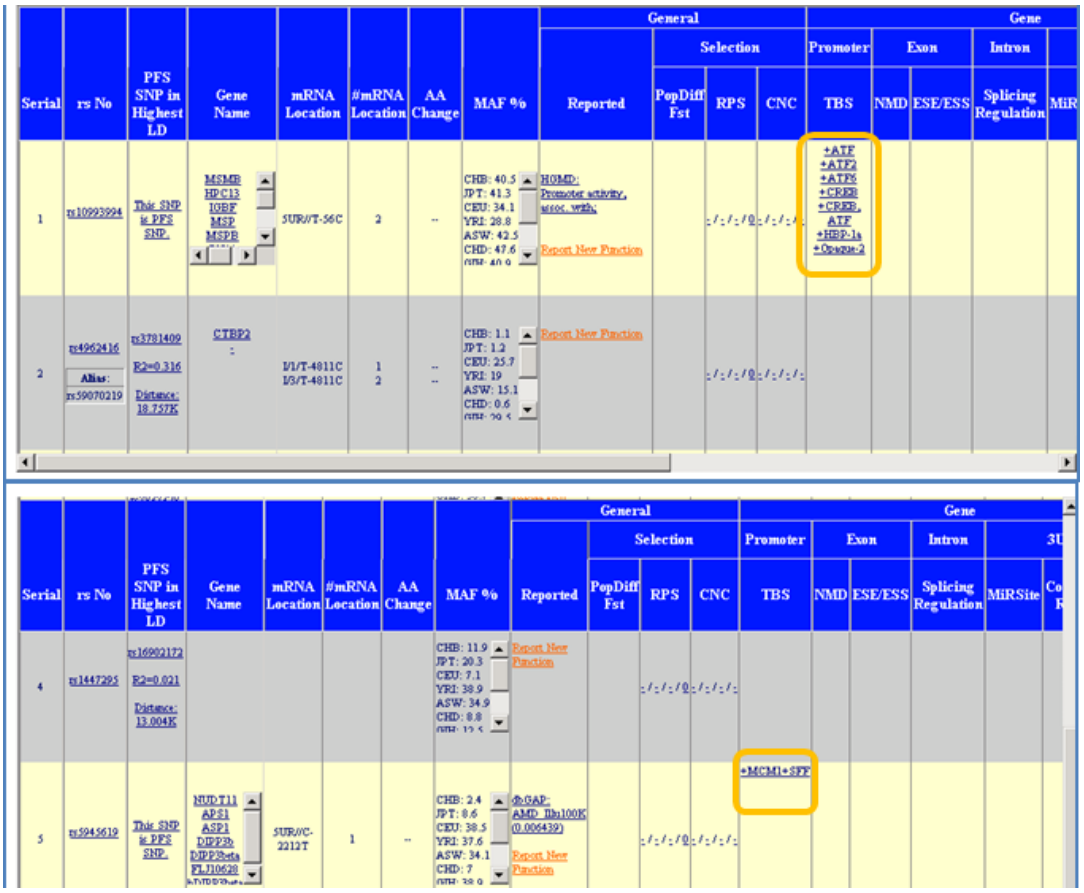


5. *Detailed Related Information of the Query Result Provided*. To facilitate a better understanding of the results, as well as to aid in the generation of hypotheses, detailed related information is also given on the query result. A step-by-step demonstration of this feature can be found at http://pfs.nus.edu.sg/Demo_Src/ supp_info_s6.html. For example, from the list of 16 SNPs associated with prostate cancer, two were predicted to affect particular transcription factor (TF) binding sites (**Figure 3.8A**). The user has the option to click on the hyperlink to that TF to better understand the role of each one. A new page providing information regarding the tissue expression, GO, KEGG pathway, PharmGKB drug association, mSigDB Pathway, OMIM or GAD reports of that TF will be shown. Alternatively, the list of TFs can also be queried using the pfSNP web-resource and the general properties of the TFs affected by the two SNPs can be inferred by placing the mouse cursor over the various headers. Text clouds under the header, 'Tissue Expression' and 'mSigDB Pathway' reveal that the TFs affected by these two SNPs are expressed reasonably significantly in the prostate and are primarily human CD34-enriched transcription factors (**Figure 3.8B**). This is consistent with the tissue distribution and pathway of the list of SNPs associated with prostate cancer. The researcher could then perhaps make the hypotheses that the polymorphisms in the promoter of these genes created binding sites for stem cell-promoting transcription factors that are expressed in the prostate, which, in turn, will activate early progenitor genes in the prostate. The scientist can then design appropriate experiments to test this hypothesis. Such information may be useful for the researcher to prioritize particular SNP(s) from a list of SNPs for functional validation. Additionally, in that TF page, there is another

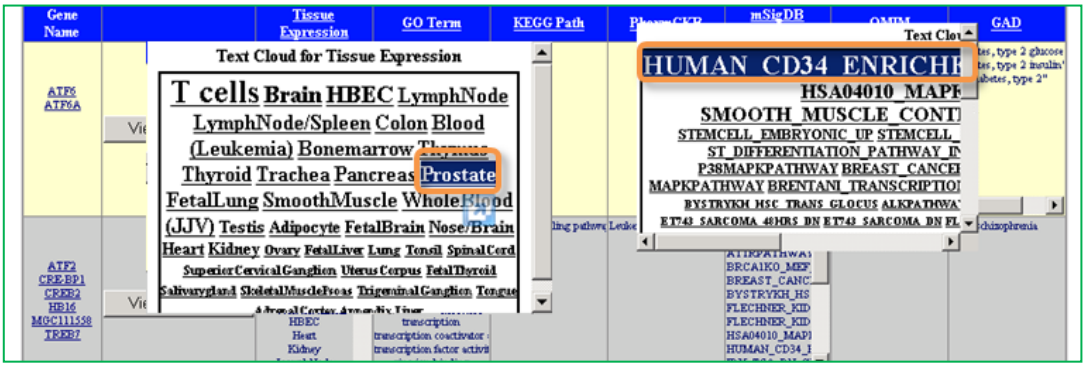
option for the user to click and view all the pfSNPs for that TF. This functionality of the pfSNP web-resource will facilitate a better and more complete understanding of the potential functionality of the SNP.

Figure 3.8: Detailed related information of the query result provided

A



B



3.4 Discussion

The pfSNP web resource aims to provide researchers interested in SNPs with a comprehensive and integrated resource that not only annotates the potential functionality of all SNPs in the dbSNP database, but also facilitates hypotheses generation through knowledge syntheses mediated by better data integration and a biologist-friendly web-interface. The comprehensiveness of this resource is achieved through the integration of numerous and diverse algorithms/resources to identify potentially functional SNPs not only based on predicted sequence motifs, but also based on previously reported functionally significant SNPs and inferred potential functionality from genetics approaches.

With the abovementioned features, I believe that the pfSNP resource will be useful for various groups of researchers focusing on SNPs, including those interested in:

1. *Designing experiments to address the functionality of specific SNPs*. This web-resource will inform the scientist about the predicted potential functionality of the SNP-of-interest so that appropriate experiments may be designed, and provide information from previous reports that have examined the SNP-of-interest.

2. *Gene-based association studies (GBAS).* This web-resource will facilitate the selection of potentially functional SNPs in any gene-of-interest for studies associating a particular gene with a phenotype. Significantly, this resource has the potential to select subset of pfSNPs in

genes within a certain chromosome region; expressed in specific tissues; and/or has been associated with a certain disease/phenotype/drug response/pathway. It is also possible to select pfSNPs that occur at/above specific threshold frequencies in specific populations or pfSNPs residing only in exons or promoter or 5'/3' UTRs, etc.

3. *Genome-wide association studies (GWAS).* As discussed earlier, this resource will be useful to current GWAS scientists, as it provides information with regards to the distance as well as the LD measured by r^2 of a nearby HapMap genotyped pfSNP that is at the highest LD to the genotyped SNP-of interest. Importantly, the integration of gene/pathway level information into the Results interface with text clouds highlighting enriched terms in each category will be useful to those researchers who have a list of SNPs associated with a phenotype and would like to formulate testable hypotheses about their findings.

When compared against other similar resources (**Table 3.2**), few other resources would be suitable to handle the GBAS- and/or GWAS-related applications either due to limited data content or web portal functionality.

Table 3.2: Comparing pfSNP against other similar web-resources.

		Tools with predicted and reported function		Tools with reported function					Tools with predicted function					
		PFS Database (pfSNP)	SNPNexus	Varietas	Gwas Analyser	Scan	Diseasome	MarkerInfoFinder	CandiSNPer	SNPInfo	SNPLogic	F-SNP	SNP Function Portal	FastSNP
Year of Publication		2010	2009	2010	2010	2010	2008	2007	2010	2009	2009	2008	2006	2006
Data Content														
	1. SNP Covered	dbSNP 129	?	dbSNP 130	?	dbSNP 129	?	?	HapMap R27	?	?	dbSNP 126	dbSNP 126	dbSNP 130
	2. Gene Context	√	√	√	√	√	√	--	√	√	√	√	√	√
	3. Allele Frequency	√	√	--	√	√	√	--	--	√	√	--	--	√
	4. LD Information	√	--	--	√	√	--	√	√	√	√	--	√	--
	5. Predicted SNP Function	√	√	--	--	--	--	--	√	√	√	√	√	√
	6. Reported SNP Function	√	√	√	√	√	√	√	--	--	--	--	--	--
Web Portal Functionality														
Search Criteria	1. SNP ID	√	√	√	√	√	√	√	√	√	√	√	√	√
	2. SNP Sequence	--	--	--	--	--	--	--	--	--	--	--	--	√
	3. Gene	√	--	√	--	√	√	--	--	√	√	√	--	√
	4. Pathway/Disease/Drug	√	--	--	--	--	√	--	--	--	√	√	--	--
	5. Allow Combination of the Criteria	√	--	--	--	--	--	--	--	--	--	--	--	--
Search Filter	1. Minor Allele Frequency	√	--	--	√	--	--	--	--	√	--	--	--	√
	2. Gene Region	√	--	--	--	--	--	--	--	--	--	--	--	√
	3. SNP Function Category	√	√	--	--	--	--	--	--	√	--	--	--	√
Search Result	1. Presentation Style	Excel-Like Web Table	Web Table	Web Table	Web Pages	Web Table	Web Pages	Web Table	Web Page	Web Table	Web Pages	Web Table	Web Table	Web Table
	2. Detailed Information about the Effect of a SNP (e.g: Creates / Disrupts a Fucntion Site)	√	--	--	--	--	--	--	--	√	√	√	--	--
	3. Comprehensive Supporting Information	√	--	√	--	--	--	--	--	--	--	√	√	√
	4. Further Search on the Result (e.g: If a SNP changes a TF site, what's known about the TF?)	√	--	--	--	--	--	--	--	--	--	--	--	--
	5. Access Supporting Information from Downloaded Results	√	*	--	--	*	--	--	--	--	--	--	--	--

Table 3.2: Continued

		Tools with predicted and reported function		Tools with reported function					Tools with predicted function					
		PFS Database (pfSNP)	SNPNexus	Varietas	Gwas Analyser	Scan	Diseasome	MarkerInfoFinder	CandiSNPer	SNPInfo	SNPLogic	F-SNP	SNP Function Portal	FastSNP
Anticipated Application														
	1. GBAS Marker Selection	√	*	*	--	--	*	*	--	√	*	*	*	√
	2. GWAS Marker Selection	√	--	--	--	--	--	--	--	√	--	--	--	--
	3. GWAS Result Annotation	√	√	√	√	√	√	√	√	√	√	√	√	√
	4. GWAS Result Analysis (Explore other potentially associated markers through LD)	√	--	--	√	--	--	--	--	√	--	--	--	--

Legend:

√	Feature present
--	Feature absent
*	Limits to the features present
?	Information not known

Abbreviations:

SNP: Single nucleotide polymorphism
LD: Linkage disequilibrium

GBAS: Gene based association study

GWAS: Genomewide association study

The inclusion of all publicly available information about the functionality of SNPs proved challenging and not all published information about SNPs associated with function/disease/drug response has been captured in this database. I have attempted to be as comprehensive as possible by culling data from the NCBI dbGAP, GAD and HGMD databases, as well as that from two previous publications that performed literature mining from PubMed (Xuan, Wang et al. 2007) or OMIM (Stoyanovich and Pe'er 2008). Data from dbGAP, GAD and HGMD databases can be updated periodically. However, the two publications which mined either PubMed (Xuan, Wang et al. 2007) or OMIM (Stoyanovich and Pe'er 2008) database does not seem complete nor updatable as some of the well-known previously reported, functionally significant SNPs were not captured in these two publications due to algorithm constraints in the PubMed literature mining and imperfection in manual curation of the OMIM database. Since many different nomenclatures have been used to describe the same SNP in previous published literature, there is no single comprehensive algorithm that can completely mine the PubMed literature databases for SNPs associated with function or disease/drug response.

Therefore, the pfSNP web-resource also provides ways for the user to deposit and share published information not found on the website. There are two ways to deposit published information with regards to a particular SNP. The first is within the Results page of the query interface. In the column titled 'Reported', there will be a phrase "Report New Function" which is hyperlinked to a page for the user to input the association of that SNP with disease/function. This feature will be useful for users who have previously published on that SNP being associated with disease/function, but realized that the previously published association was not reflected in this web-resource. An alternative way to input published information for a

SNP-of-interest is by clicking the "Submit Published SNP Info" button found on the home page after the login page.

The webpage to submit published SNP information is simple and contains several user-friendly features. The user will first have to submit the reference of the publication that reports the association of the SNP with function/disease/drug response. Although the PubMed ID is required, for users who do not know the PubMed ID, the button 'Search for PubMed ID' (**Figure. 3.9A**) can be used and requires only one or preferably more information with regards to the first author, journal, volume, and page number of the publication (**Figure. 3.9B**). From this, a list of titles from PubMed that fits this description will appear for selection. The user then inputs the SNP-of-interest as the rs number. However, if the rs number is not known, the user can merely provide other information with regards to the SNP-of-interest by clicking the 'Search for rs number by its gene location' tab (**Figure 3.9C and D**) and the web-resource will then help match the rs number with the information provided. The user then provides information about the reported functional significance of the SNP – whether the SNP affects function (e.g. disrupt protein function or cause mRNA destabilization, etc) or is associated with drug response or disease risk (**Figure 3.9E**). Here, again, the Auto-Complete Prompt-As-You-Type feature for the 'drug' and 'disease' will facilitate the user to select the most appropriate drug or disease, and if none of the drugs/diseases in the database matches what the user has in mind, a new drug/disease can also be input. The user is then given an opportunity to elaborate (up to a maximum of 200 characters) about the associated functional significance of the SNP (**Figure 3.9F**). When all the information has been provided, the user clicks on the 'Submit the Literature Reference', and the information will then be associated with that particular SNP and presented in the Results page when that SNP is queried,

with acknowledgement to the individual who submitted the information. With support from the scientific community, this pfSNP resource can then be kept comprehensive and current.

Figure 3.9: User-friendly web-interface to submit published functionally significant SNPs

Nonetheless, because our current understanding of the functional consequences of the sequences in the human genome remains incomplete, there remains many yet-to-be discovered pfSNPs. As our understanding about the functional importance of various regions/sequences in the human genome improves, more new pfSNPs will be identified and these new pfSNPs can then be updated in

later versions of our resource. Our aim is for this pfSNP web-resource to remain relevant even as many more new SNPs are identified through whole-genome sequencing strategies. Through automation, we hope that these novel SNPs can be input or automatically captured into the pfSNP resource to be evaluated for potential functionality. Structural variation, including copy-number variation, represents another important source of genetic variation associated with complex disease/phenotypes. As our understanding of the role of these structural variants in diseases advances, and as algorithms are developed to better predict the potential functionality of these variants, such potentially functional variants could also be included in later versions of the pfSNP resource.

3.5 Summary

A comprehensive, well-annotated, integrated pfSNP Web Resource (http://pfs.nus.edu.sg/) which is aimed to facilitate better hypothesis generation through knowledge syntheses mediated by better data integration and a user-friendly web interface was developed. Its query-interface has the user-friendly 'auto-complete, prompt-as-you-type' feature and is highly customizable facilitating different combination of queries using Boolean-logic. Additionally, to facilitate better understanding of the results and aid hypotheses generation, gene/pathway-level information with text-clouds highlighting enriched tissues/pathways as well as detailed related information are also provided on the results page.Hence, this pfSNP resource will be of great interest to scientists focusing on association studies as well as those interested to experimentally address the functionality of SNPs.

CHAPTER 4: VALIDATING THE USEFULNESS OF PFSNPS IN ASSOCIATION STUDY

4.1 Introduction

Every year, more than one million individuals worldwide will develop colorectal cancer (Cunningham, Atkin et al. 2010), accounting for 10% of the global cancer burden. Colorectal cancer (CRC) is the most frequently diagnosed cancer in Singapore (7,909 new cases between 2005-2009) (SC 2011). More than half of CRC patients develop metastatic disease (stage 4) either at diagnosis or at relapse following initial curative intent therapy. This means that a substantial proportion of patients may need treatment for the metastasis or relapse of colorectal cancer.

The efficacy of anti-cancer drug treatment for CRC can be measured by either tumor response or patient survival. Tumor response is usually used in the neo-adjuvant setting where shrinkage of tumor is the primary focus. Response Evaluation Criteria In Solid Tumours (RECIST)(Sohaib 2012) is usually used to determine tumor response. According to the RECIST criteria, patient responses are classified into four categories, namely Complete Response (CR, disappearance of all target lesions), Partial Response (PR, at least a 30% decrease in the target lesions), Progressive Disease (PD, at least a 20% increase in the target lesions or new lesions) and Stable Disease (SD, neither sufficient shrinkage to qualify for PR nor sufficient increase to qualify for PD). Patient survival is used in the adjuvant setting and multiple clinical endpoints, such as Overall Survival (OS) and Relapse-Free Survival (RFS) can be used to measure patient survival. OS is an endpoint to capture all causes of mortality and therefore is a composite measurement of both drug efficacy and adverse effects. RFS is defined as the time to the first relapse or death from any cause

excluding second primaries or other cancers. It is perhaps the endpoint most sensitive to a treatment effect while still accounting for any imbalance in adverse event-related deaths (Chua, Sargent et al. 2005). Toxicity of anti-cancer drug treatment is usually measured by the severity of adverse effects, specific to the drug used.

5-fluorouracil (5-FU) and its pro-drug, Capecitabine, are widely used to treat CRC. It has been proposed that there are two distinct modes of action for 5-FU. First, it acts as anti-metabolite whereby its active form, FdUMP, produced by Thymidine Phosphorylase (TYMP), inhibits Thymidylate Synthase (TYMS). Second, it can induce cell death, whereby incorporation of its active products FUTP and FdUTP into RNA and DNA, respectively, leads to subsequent cell apoptosis (Sobrero, Aschele et al. 1997). Uridine Monophosphate Synthetase (UMPS, also known as OPRT) is responsible for converting 5-FU to FUMP, which is the first step of producing FUTP and FdUTP. It was previously thought that the "anti-metabolite" pathway was more important than the "cell toxicity" route. However, the two pathways may overlap, because the intermediate product in the "cell toxicity" pathway, FUDP, may also be converted to FdUDP and subsequently FdUMP and participate in the "anti-metabolite" pathway. 5-FU is catabolized into the inactive form of DHFU by Dihydropyrimidine Dehydrogenase (DPYD), and DPYD is the rate-limiting enzyme in degrading 5-FU.

The expression level of the abovementioned genes, namely TYMP, UMPS, DPYD and TYMS, have long been suggested to affect 5-FU pharmacokinetics and pharmacodynamics and thus determine the efficacy and toxicity of 5-FU. Various studies have demonstrated that the expression levels of TYMP, UMPS, DPYD and TYMS genes would be markers for predicting CRC patient survival (Koopman, Venderbosch et al. 2009). However, such reports remain controversial and not readily

replicable possibly due to the inaccuracy of the Immunohistochemistry method used for measuring the protein abundance in tissue samples.

SNPs and other genetic markers in these genes have thus been tested as possibly representing better markers for predicting efficacy or toxicity of 5-FU in CRC patients (Gosens, Moerland et al. 2008). **Table 4.1** shows the variants in either "5-FU PD"- or "5-FU PK"-related genes found to be associated with 5-FU efficacy or toxicity in CRC patients. The most studied variant is the 28 bp-tandem repeat (**Table 4.1 A**, rs34743033, either 2 or 3 repeats and denoted 2R and 3R allele) in the 5'UTR region of the TYMS gene, together with an embedded SNP (rs2853542, E/1/G-58C). Patients with genotype 2R/2R, 2R/3C (3C denotes a haplotype comprising 3R allele for the tandem repeat and the C allele of the SNP), and 3C/3C were shown to have lower expressions of the gene and were associated with a better response to 5-FU, as evidenced by longer patient survival (Kawakami and Watanabe 2003). This VNTR and SNP have also been shown to be associated with tumor response (Graziano, Ruzzo et al. 2008). In the 3' UTR region of the TYMS gene, there is also one 6 bp indel (rs16430) associated with tumor response (Salgado, Zabalegui et al. 2007). Two SNPs in the MTHFR genes (**Table 4.1 A**), rs1801133 (E/5/G677A, A222V) and rs1801131 (E/8/T1298G, E429A), have been associated with patient survival in various cancers individually (De Mattia and Toffoli 2009) and a model combining these SNPs with TYMS polymorphism/status was proposed to predict the response better (Etienne, Formento et al. 2004; Jakobsen, Nielsen et al. 2005).

Table 4.1: List of variants reported in previous association studies for 5-FU efficacy or toxicity in CRC patients.

A Variants in the 5-FU PD pathways genes

Gene	rsNo	Gene Context	Phenotype	Association Found	No Association Found
TYMS	rs2853542 and rs34743033	E/1/G-58C and the 28 bp VNTR in 5UTR	Efficacy	(Kawakami and Watanabe 2003; Etienne-Grimaldi, Bennouna et al. 2012)	(Tsuji, Hidaka et al. 2003; Farina-Sarasqueta, van Lijnschoten et al. 2010; Vignoli, Nobili et al. 2011)
			Efficacy (TR)	(Graziano, Ruzzo et al. 2008; Etienne-Grimaldi, Bennouna et al. 2012)	(Farina-Sarasqueta, van Lijnschoten et al. 2010)
	rs16430	6 bp InDel in 3UTR	Efficacy (TR)	(Salgado, Zabalegui et al. 2007)	(Graziano, Ruzzo et al. 2008)
MTHFR	rs1801133	E/5/G677A, A222V	Efficacy	(De Mattia and Toffoli 2009)	(Marcuello, Altes et al. 2006; Ruzzo, Graziano et al. 2007)
			Efficacy (TR)	(Cohen, Panet-Raymond et al. 2003; Etienne, Formento et al. 2004; Jakobsen, Nielsen et al. 2005)	(Etienne-Grimaldi, Bennouna et al. 2012)
			Toxicity	(Gusella, Frigo et al. 2009)	(Cohen, Panet-Raymond et al. 2003)
	rs1801131	E/8/T1298G, E429A	Efficacy	(De Mattia and Toffoli 2009)	(Marcuello, Altes et al. 2006; Ruzzo, Graziano et al. 2007)
			Efficacy (TR)	--	(Etienne, Formento et al. 2004)

B Variants in the 5-FU PK pathway genes

Gene	rsNo	Gene Context	Phenotype	Association Found	No Association Found
UMPS	rs1801019	E/3/G638C, G213A	Toxicity	(Ichikawa, Takahashi et al. 2006)	--
			Efficacy (TR)	--	(Farina-Sarasqueta, van Lijnschoten et al. 2010)
DPYD	rs1801265	E/2/G85A, C29R	Toxicity	(Zhang, Li et al. 2007)	--
	rs2297595	E/6/T496C, M166V	Toxicity	(Gross, Busse et al. 2008)	--
	rs1801159	E/13/T1627C, I543V	Toxicity	(Zhang, Li et al. 2007)	--
			Efficacy (TR)	--	(Farina-Sarasqueta, van Lijnschoten et al. 2010)
	rs3918290	I/14/G1T	Toxicity	(Schwab, Zanger et al. 2008)	--

Note: "Efficacy (TR)" highlights that "Tumor Response" is used as the measurement of efficacy. "Efficacy" means patient survival is used as measurement of efficacy.

Thus far, only variants in the "5-FU PD" genes were reported to be significantly associated with efficacy in some studies. The more commonly studied SNPs in the "5-FU PD" genes (namely rs2853542 / rs34743033 and rs16430 in the UMPS gene) that have reported association with 5-FU are mainly non-coding, suggesting that non-coding SNPs may play an important role in determining 5-FU efficacy. There may also be locus heterogeneity in 5-FU response since different SNPs in different genes were reported to be associated. Unfortunately, replication of such reported association between "5-FU PD" gene variants and response remains challenging (Cohen, Panet-Raymond et al. 2003; Tsuji, Hidaka et al. 2003; Etienne, Formento et al. 2004; Marcuello, Altes et al. 2006; Ruzzo, Graziano et al. 2007; Graziano, Ruzzo et al. 2008; Farina-Sarasqueta, van Lijnschoten et al. 2010; Vignoli,

Nobili et al. 2011), suggesting the possible presence of other loci in determining 5-FU efficacy.

It was interesting to note that the variants in the "5-FU PK" genes were mainly investigated for their association with 5-FU toxicity, not efficacy (**Table 4.1 B**), despite the expression levels of these genes having been previously associated with 5-FU efficacy (Koopman, Venderbosch et al. 2009). Only one report (Farina-Sarasqueta, van Lijnschoten et al. 2010) tried to explore the possible association for efficacy, but failed. However, this study might have been underpowered to detect such an association, in that it also failed to replicate the association for the VNTR and SNP (**Table 4.1 A**) in the TYMS gene.

In summary, existing evidence suggests there can be more SNPs in both the "5-FU PD" and the "5-FU PK" pathway genes that can determine the efficacy of 5-FU. The function of such SNPs may also vary and may not be restricted as non-synonymous. The aim of this study was to explore as many potentially functional SNPs (pfSNPs) in both "5-FU PD" and "5-FU PK" pathway genes for possible association with 5-FU efficacy, as measured by tumor response. pfSNPs were determined to be good candidates for this study since SNPs with various functions are all covered.

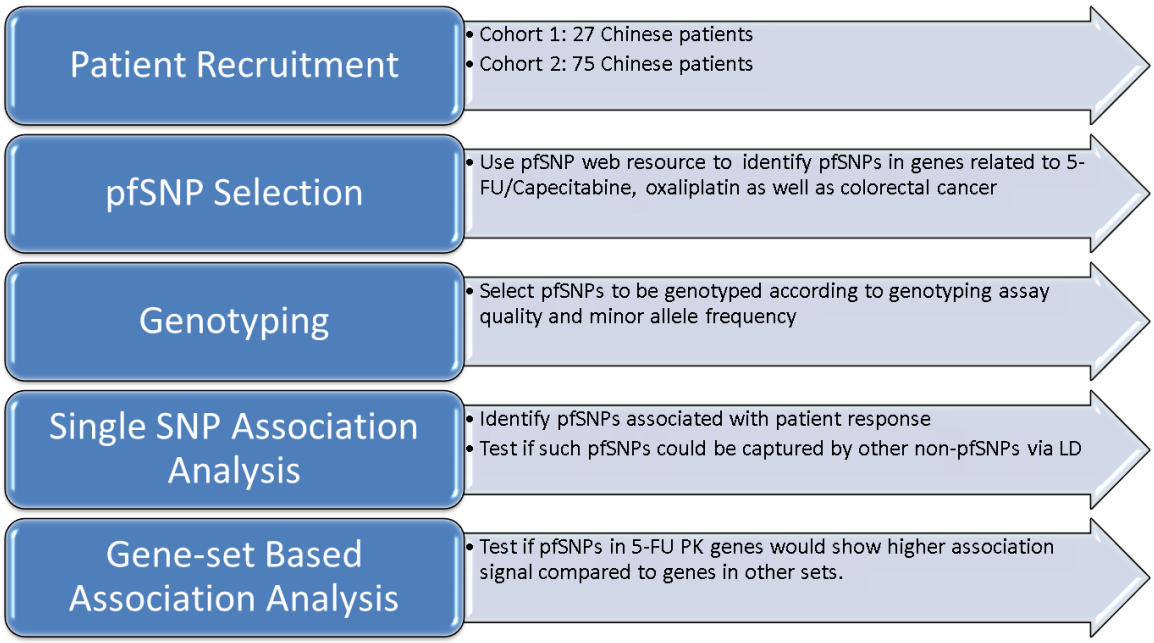
This study will also explore the suitability of using pfSNPs for pathway level, candidate gene-based association studies as compared with the traditional, non-synonymous SNP-centric approach. The 5-FU response is a good phenotype to examine, because it is known that expression levels of the genes involved are important in determining the phenotype. This study may reveal more expression-related SNPs due to locus heterogeneity, and evaluate the limitation of the

non-synonymous SNP-centric approach. Further, this study also aims to compare the tagging SNP-based approach versus the direct pfSNP approach.

4.2 Materials and Methods:

The workflow of the study is shown in **Figure 4.1**.

Figure 4.1: The workflow of the association study.



4.2.1 Patient samples and clinical parameters

The shrinkage of liver metastatic tumors was used as the clinical endpoint for measuring treatment efficacy. Response Evaluation Criteria In Solid Tumours (RECIST) (Sohaib 2012) is used to determine tumor response. According to the RECIST criteria, patient responses are classified into four categories, namely Complete Response (CR, disappearance of all target lesions), Partial Response (PR, at least a 30% decrease in the target lesions), Progressive Disease (PD, at least a 20%

increase in the target lesions or new lesions) and Stable Disease (SD, neither sufficient shrinkage to qualify for PR nor sufficient increase to qualify for PD). In our patient cohorts, there was no patient belonging to the 'Complete Response' category. Patients with 'Partial Response' were deemed as 'Responders' and patients with 'Progressive Disease' or 'Stable Disease' were classified as 'Non-Responders' in the association analysis.

There are two batches of unrelated Chinese metastatic CRC patients recruited for this study. The recruitment of both batches has been approved by Singhealth ethical reviewing committee. They are referred to as "Cohort 1" and "Cohort 2" respectively in the following text. All patients had liver metastasis and were given neo-adjuvant chemotherapy prior to operation for the liver lesion. **Table 4.2 A** shows the characteristics of the two cohorts. **Table 4.2 B** shows the age and gender distribution in responder and non-responder sub-groups.

"Cohort 1" comprised 27 unrelated Chinese liver metastatic CRC patients who were treated with 5-FU (a few had Capecitabine) alone (a few) or with oxaliplatin regime (most) as their neo-adjuvant chemotherapy. Some of these patients had also been previously exposed to these drugs. Of these 27 patients, 12 had partial response, seven had stable disease, and eight had progressive disease. Hence, the response rate was ~40-45%, which is typical for patients undergoing 2-drug combination therapy. Two-thirds of these patients are males, and Cohort 1 had a median age of 62 years. Age and gender were shown not to be confounding factors in this cohort (**Table 4.2 B**).

Table 4.2: The demographic characteristics of the two patient cohorts recruited.
A. The demographic characteristics in two cohorts. B. The age and gender distribution in responder and non-responders in each cohorts.

A

	Cohort 1	Cohort 2
Number of Patients	27	75
Ages (Median)	42-86 (62)	36-78 (59)
Males (Females)	18(9)	60(15)
Prior Drug Exposure		
5-FU alone	2	0
Capecitabine alone	3	0
5-FU + oxaliplatin	2	0
Capecitabine + oxaliplatin	0	0
5-FU + Radio Therapy	2	0
Drugs Treated		
5-FU alone	2	0
Capecitabine alone	1	75
5-FU + oxaliplatin	15	0
Capecitabine + oxaliplatin	8	0
5-FU + Irinotecan	1	0
Response		
Partial response	12	13
Stable disease	7	31
Progressive disease	8	18

B

		Responder	Non-Responder	P Value
Cohort 1	**Ages (Mean)**	62.3	63.5	0.30
	Males(Females)	8(4)	10(5)	0.39
Cohort 2	**Ages (Mean)**	65.7	59.0	0.03
	Males(Females)	5(8)	43(6)	0.0007

"Cohort 2" comprised 75 unrelated Stage IV CRC Chinese patients. These patients were treated with only Capecitabine and they have never been previously exposed to this drug. Of these 75 patients, 13 had partial response, 31 had stable disease, and 18 had progressive disease. The response of the remaining 13 patients remains unknown. Hence, the response rate of this group of patients is only ~17%,

which is typical for single-agent treatment. In "Cohort 2", ~80% of the patients were male, and the median age of the cohort was 59 years. Age and gender were shown to be confounding factors in "Cohort 2" (**Table 4.2 B**).

4.2.2 Selection of potentially functional SNPs (pfSNPs) for association study

I queried our pfSNP resource, as previously described (Wang, Ronaghi et al. 2011), for genes related to 5-FU/Capecitabine, oxaliplatin as well as colorectal cancer. A total of 214 genes and ~2800 pfSNPs were found to be associated with these drug pathways or CRC. As only 1,536 SNPs can be genotyped within a single customized GoldenGate Genotyping Array (Illumina, Inc), the following criteria were employed to select a subset of these 2800 pfSNPs: all SNPs within the promoter, coding, 5'/3' un-translated regions with has a GoldenGate Score (GGS: measure of assay quality by their platform) of greater than 0.5 were selected. For the introns, pfSNP with a GGS>0.7 were selected and monomorphic ones reported in HapMap CHB population were excluded, except for those previously reported as functional. For any adjacent SNPs that may interfere with each other in the assay, I selected the SNP according to the following order: "Previously reported → Non-Synonymous → Synonymous → UTR → Intron". A list of all of the SNPs included on the GoldenGate array is available as **Supplementary Table 1**.

Among the markers that were unsuitable to be genotyped by GoldenGate assay, I picked 14 important markers in 5-FU response prediction to be genotyped by other methods (listed in **Supplementary Table 2**). The 14 markers include the previously well studied VNTR with embedded SNP as well as the 6bps 3' UTR indel

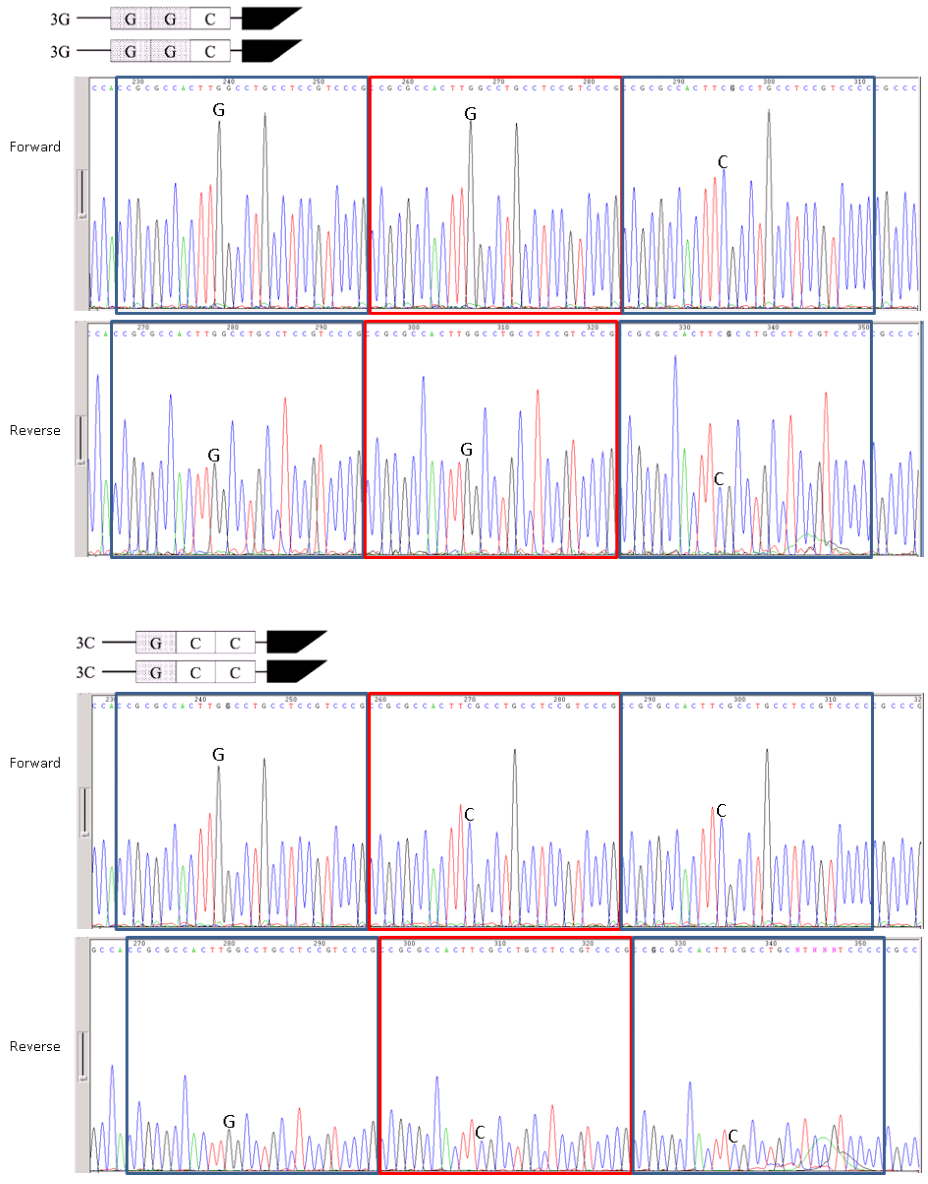
in the TYMS gene (**Table 4.1 A**) because GoldenGate technology could only genotype SNPs. Other markers are SNPs with low GoldenGate scores in the TYMS, TYMP and DPYD genes. A customized Sequenom MassARRAY® panel was used to genotype 11 of the 14 SNPs and 1 SNP was genotyped by ABI TaqMan. In addition, I developed a novel method to genotype the VNTR (rs2853542) and embedded SNP (rs34743033) in the TYMS gene. This VNTR can have either 2 or 3 repeats, and the SNP can occur in both the second and third repeat (Kawakami and Watanabe 2003). Previous papers used Restriction Fragment Length Polymorphism method to genotype them (Kawakami and Watanabe 2003), but the star activity of the restriction enzyme may make the results less reliable. I developed a robust method using Sanger Sequencing for this purpose. A fragment of 486 bp was amplified using the primers (F- CTGCTGGCTTAGAGAAGGCG and R- AGCGGAGGATGTGTTGGATC) and the amplicon was sequenced in both directions using the forward and reverse primers. Different genotypes would yield distinct patterns on the forward and reverse sequencing reads (As shown in **Figure 4.2**), allowing the genotype to be easily deduced.

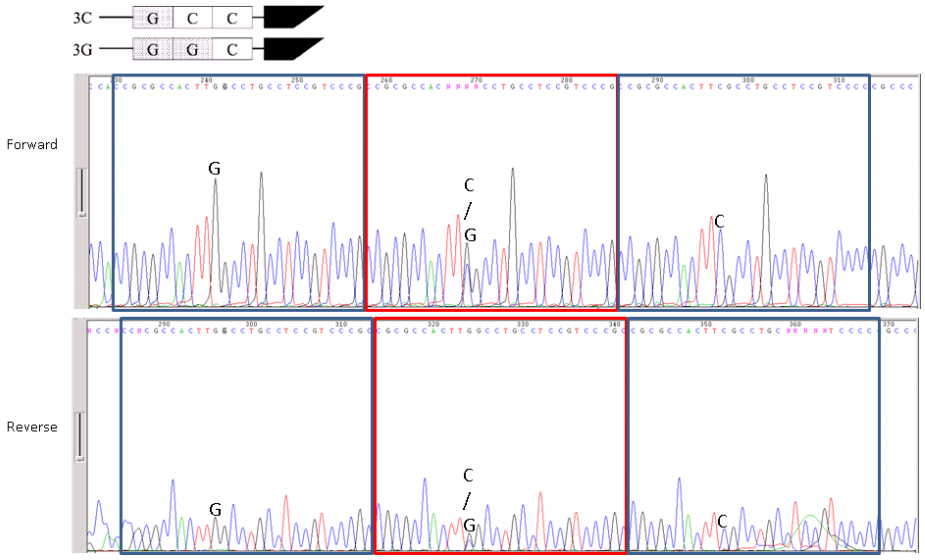
In summary, there were 1,536 markers (Listed in **Supplementary Table 1**) genotyped with a single customized Illumina GoldenGate SNP genotyping array, 11 (Listed in **Supplementary Table 2**) by Sequenom MassARRAY®, 1 (rs11479) by ABI TaqMan and the VNTR (rs2853542), with the embedded SNP (rs34743033) genotyped by Sanger Sequencing.

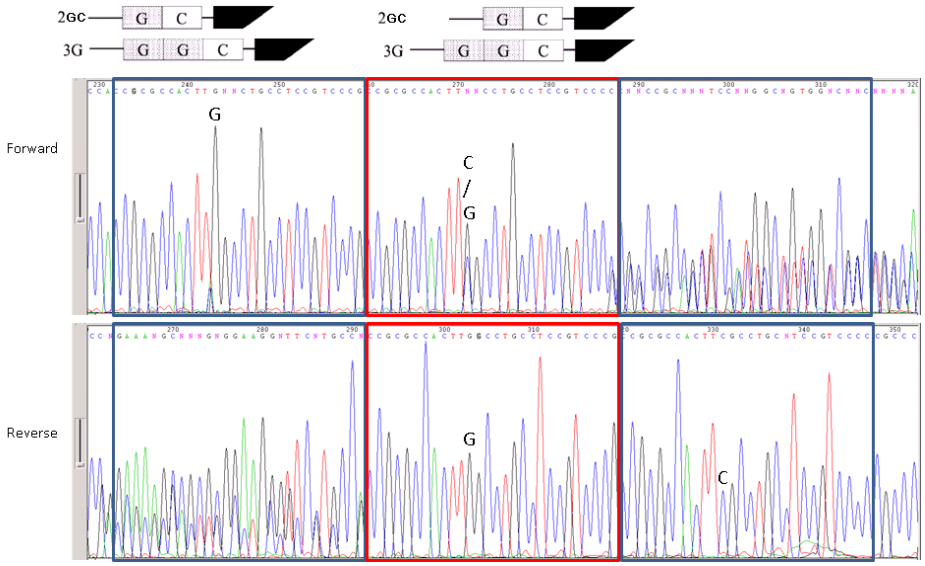
The distribution of the SNPs and genes selected in the three categories (CRC, Fluorouracil and Platinum related) is depicted in **Figure 4.3**. More than half of the SNPs (863) are from fluorouracil-related genes and 702 SNPs are from platinum-related genes. A considerable portion of the SNPs (656) are from

CRC-related genes, although 40% (32 out of 80) of these genes and 88% (579 out of 656) of the SNPs are related to fluorouracil and/or platinum as well. Nonetheless, this group of SNPs is valuable because they can be used in pathway-based analyses to form the background distribution to which a comparison can be made.

Figure 4.2: The different sequencing patterns generated by the different genotype of the VNTR and embedded SNP in the TYMS gene promoter region.




3C
G C C
3G
G G C
Forward
G
C
/
G
C
Reverse
G
C
/
G
C


2GC
G C
3G
G G C
2GC
G C
3G
G G C
Forward
G
C
/
G
Reverse
G
C

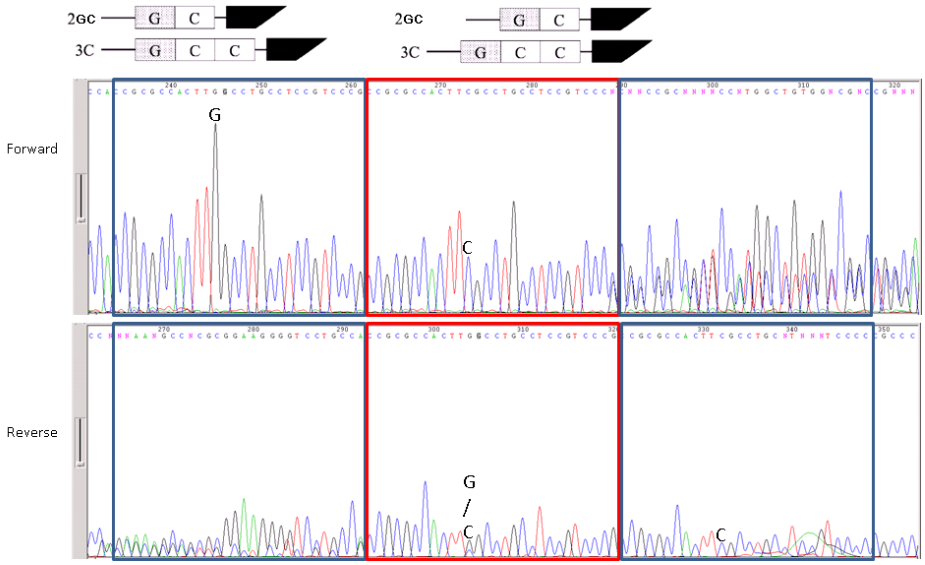
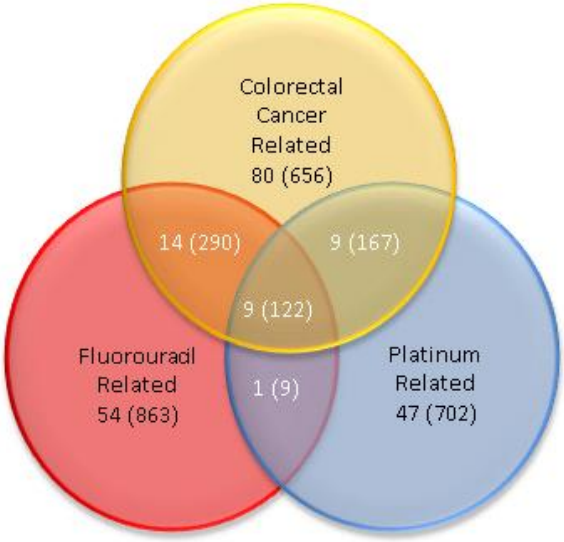

2GC
G C
3C
G C C
2GC
G C
3C
G C C
Forward
G
C
Reverse
G
/
C

Figure 4.3: The gene and pathway distribution of pfSNPs chosen for genotyping.



The numbers of SNPs in each gene region and function category covered in this study are shown in **Figure 4.4**. Each of the four gene regions, namely promoter, coding, intron and 3' UTR, are adequately covered in general. For promoter and 3'UTR, most of the SNPs genotyped are those that change TF binding sites. In the coding region, SNPs that change ESE/ESS sites are the most abundant, and non-synonymous SNPs that cause deleterious effects are the second most abundant. The intron region SNPs are enriched with those with a signature of recent positive selection.

The distribution of SNP minor allele frequency for the intronic versus non-intronic region is shown in **Figure 4.5**. For coding, 3'UTR, and promoter regions, a number of SNPs not previously genotyped by HapMap (designated by question mark in the figure) will be genotyped in this study. I also endeavoured to genotype a number of monomorphic SNPs reported by HapMap. For the intron region, since most of the SNPs are those with a signature of recent positive selection, they have MAF more than 5%. I did not genotype many monomorphic ones within introns.

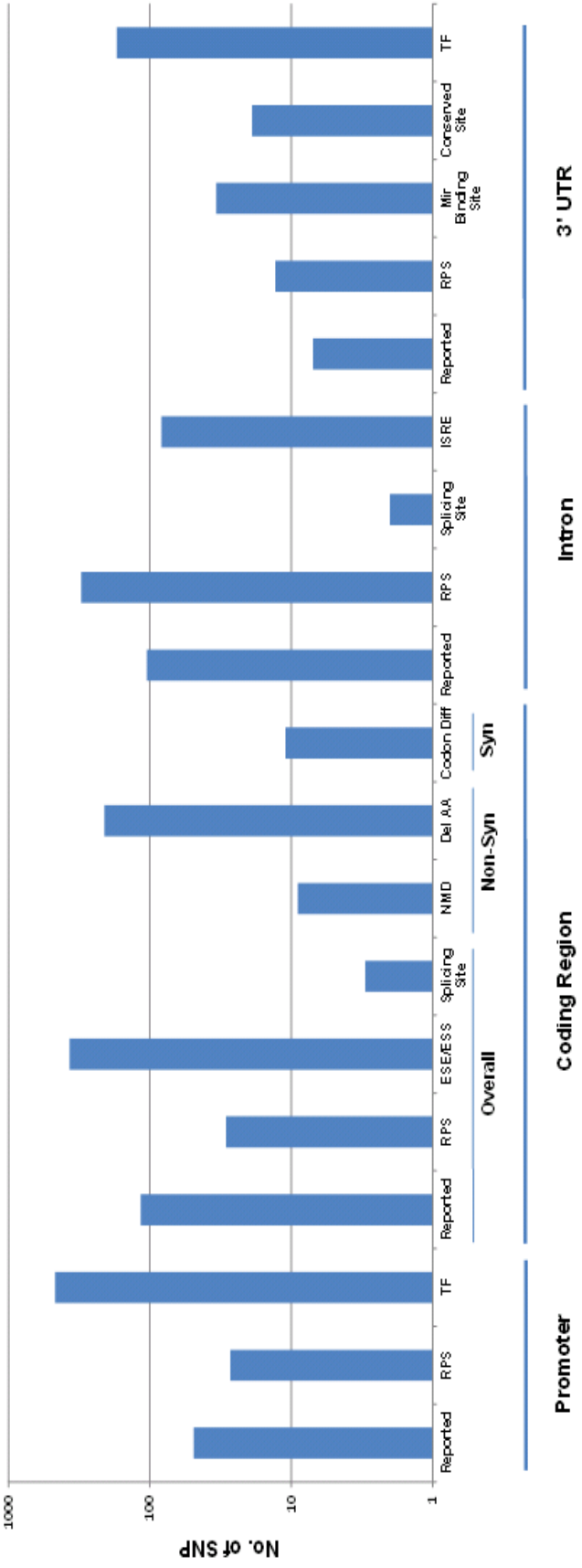


Figure 4.4: The number of SNPs selected for genotyping in each gene region and function category.

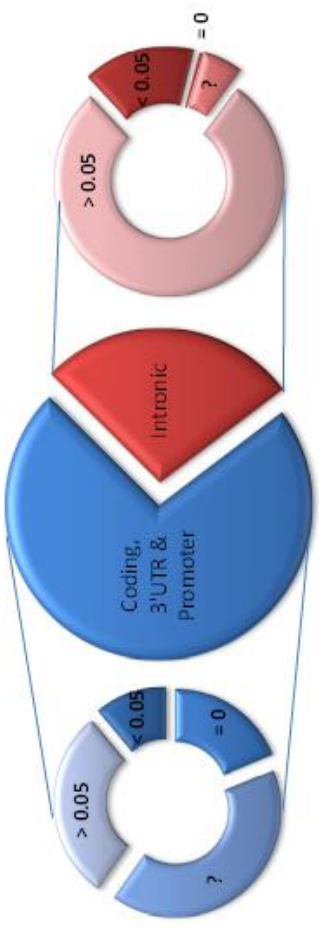


Figure 4.5: The MAF for SNPs selected for genotyping in each region.

4.2.3 Single marker association analysis

Hardy-Weinberg equilibrium was analysed using the Microsoft Excel-based SNP Statistics Calculator that I previously developed. Single marker association analysis was performed using PLINK (Purcell, Neale et al. 2007) . The P value was calculated by the permutation-based method without considering age and gender as confounding factors, and the Odds Ratio (OR) and corresponding 95% confidence interval were determined using regular allele-based association analyses, because the permutation-based method does not provide such information.

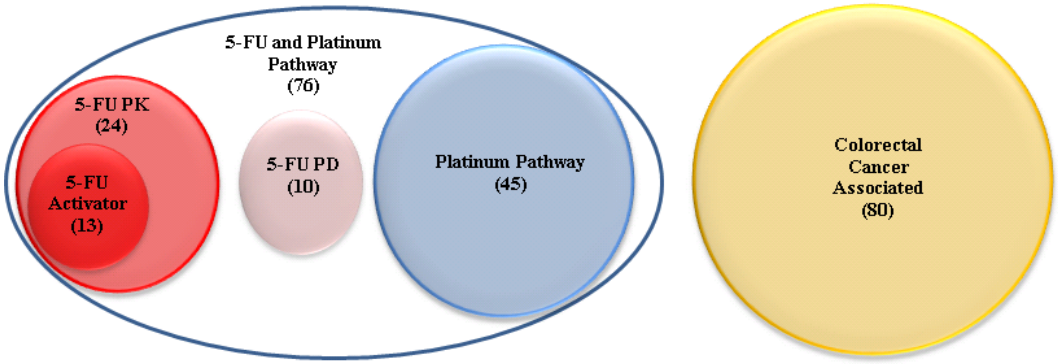
Haploview (Barrett, Fry et al. 2005), together with HapMap release 28 genotype data, were used to analyze the LD profile in genes of interest and to plot the pairwise LD diagram. LD diagrams were also generated using an in-house Microsoft Excel-based LD analysis tool written in VBA that I previously developed. SNAP (Johnson, Handsaker et al. 2008) was used to check for SNPs in LD with the SNP of interest in the 1 Mb flanking region.

4.2.4 Gene set analysis

Gene set analysis was performed using GSA-SNP (Nam, Kim et al. 2010) and Pathways of Distinction Analysis (PoDA) (Braun and Buetow 2011). The second best P value in each gene was used for the Z-score-based “parametric analysis of gene set enrichment” (PAGE) method (Kim and Volsky 2005), as it was suggested that using the best P value may increase the risk of a spurious result (Nam, Kim et al. 2010). PoDA results were obtained using the R package provided in the original paper.

The composition and relationship between the gene sets used are depicted in **Figure 4.6**. The detailed list of the genes in each gene set can be found in **Supplementary Table 3**.

Figure 4.6: The composition and relationship of gene sets used in the gene set analysis. The number in brackets is the number of genes in that gene set. The "5-FU Activator" is a subset of "5-FU PK". The "5-FU and Platinum Pathway" comprise "5-FU PK", "5-FU PD" and "Platinum Pathway".



Three 5-FU related pathways/gene sets, namely the "5-FU Activator" gene set, the "5-FU PK" pathway and the "5-FU PD" pathway, were included for gene set analysis. The "5-FU Activator" gene set is a subset of "5-FU PK" pathway and comprises 13 genes that code for enzymes converting 5-FU or Capecitabine into their active form. DPYD and related genes that catabolize 5-FU were excluded. Because a large number of patients in "Cohort 1" were also administered Oxaliplatin, I combined all the genes in "5-FU PK", "5-FU PD" and "Platinum pathway" to form the "5-FU and Platinum Pathway" set so that I could test if the genes in this set would show a higher association signal since they all may contribute to the final outcome of this cohort.

4.3 Results and Discussion:

4.3.1 Genotyping results

Most of the genotyped SNPs in the customized GoldenGate array were either monomorphic (36%) or had high minor allele frequency (MAF≥0.1, 46%) (**Figure 4.7**). When the MAF of our genotyping results were compared against the HapMap results, a high concordance (R^2=0.9494) was observed (**Figure 4.8**), which reassures the quality of our genotyping results. There were 72 SNPs in "Cohort 1" and 66 SNPs in "Cohort 2" have a Hardy-Weinberg Equilibrium P value less than 0.05 and these SNPs were excluded from further analysis.

Figure 4.7: The MAF distribution of GoldenGate genotyped SNPs.

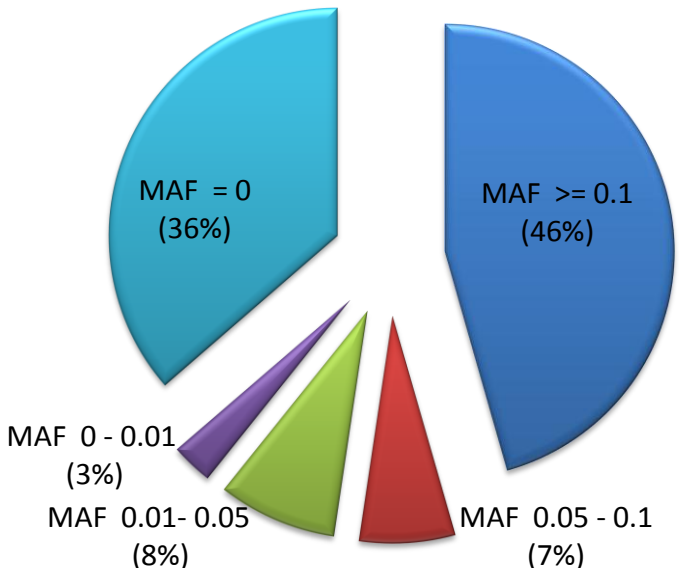
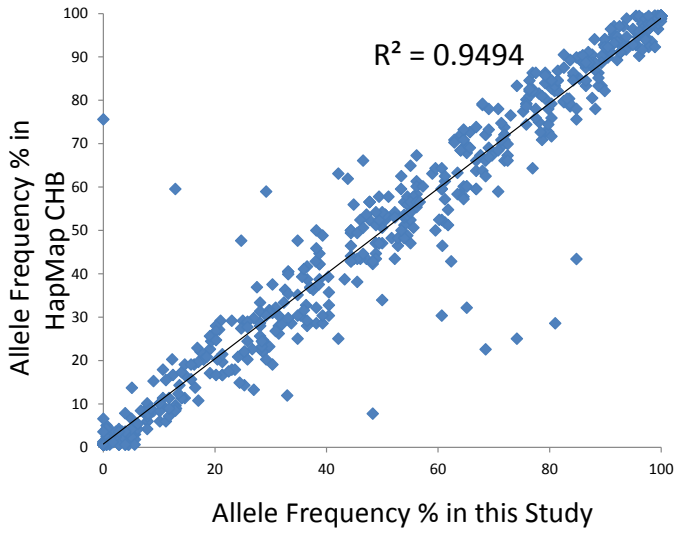


Figure 4.8: Comparing HapMap CHB reported MAF and MAF in this study.



Nine out of eleven SNPs genotyped by Sequenom MassARRAY® were successful, with the genotype calling rate ranging from 97% to 100% in all samples. The two SNPs genotyped by Sanger Sequencing and TaqMan were both successful. Sanger Sequencing successfully genotyped all of the samples while the TaqMan assay obtained a calling rate of 96%. All SNPs successfully genotyped by these methods passed the Hardy-Weinberg equilibrium.

4.3.2 Single SNP association analysis identified three SNPs in the UMPS gene whose minor allele occurs only in non-responders

Thirty-nine SNPs with un-corrected P-values less than 0.05 in "Cohort 1" are listed in **Table 4.3** and arranged in ascending order according to their P value. As the sample size was small (n=27), none of these markers are statistically significant after Bonferroni correction. Interestingly, there are four SNPs that had infinite OR because one allele is found only in either the Responders/Non-Responders groups (**Table 4.3**, SNP 25 and SNP 28-30).

Table 4.3: List of SNPs showing P<0.05 in "Cohort 1" ranked by P value in this cohort.

							Cohort 1			
							Allele Count			
SN	Gene	mRNA Location	AA Change	rs No.	Function Summary	P	NR (30)	Rsp (24)	OR	OR 95 CI
1	ABCC4	I/9/G-2464A	--	rs4148485	RPS	1.98E-03	7	16	6.57	1.98 21.78
2	ERCC4	5UR//G-3028A	--	rs6498485	TF	2.31E-03	5	13	5.91	1.69 20.66
3	ERCC4	5UR//G-4199A	--	rs12921111	RPS,TF	2.31E-03	5	13	5.91	1.69 20.66
4	ERCC4	5UR//G-4604C	--	rs11649492	TF	2.31E-03	5	13	5.91	1.69 20.66
5	SMARCA 2	I/28/G-9017A|I/29/G2082 A	--|--	rs10965088	RPS	2.87E-03	11	18	5.18	1.58 16.95
6	ERCC4	5UR//C-643T	--	rs3136038	TF	3.38E-03	6	13	4.73	1.42 15.73
7	ERCC4	5UR//T-29A	--	rs1799797	TF	3.38E-03	6	13	4.73	1.42 15.73
8	ABCC4	I/8/A728G	--	rs1751029	RPS	4.19E-03	11	1	13.32	1.57 112.70
9	ABCC4	I/8/G651A	--	rs1617844	RPS	4.19E-03	11	1	13.32	1.57 112.70
10	UPP2	E9/3UTR /T123G	--	rs2074954	TF	4.38E-03	14	2	8.75	1.73 44.25
11	MSH2	I/10/G12A	--	rs3732183	Reported	5.67E-03	17	4	6.54	1.79 23.84
12	**TYMS**	**Promoter**	--	**rs2853542 and rs34743033**	**Reported, TF**	1.46E-02	12	5	7.60	1.89 30.50
13	ABCC4	I/4/A-108G	--	rs899497	RPS,ISRE	1.54E-02	15	3	7.00	1.72 28.54
14	ABCC4	I/4/A-323G	--	rs4148469	RPS,ISRE	1.54E-02	15	3	7.00	1.72 28.54
15	ABCC4	I/4/C10763G	--	rs4303338	RPS	1.54E-02	15	3	7.00	1.72 28.54
16	ABCC4	I/4/G7454A	--	rs4773854	RPS	1.54E-02	15	3	7.00	1.72 28.54
17	ABCC4	I/4/T11908C	--	rs1926657	Reported	1.54E-02	15	3	7.00	1.72 28.54
18	ABCC4	I/4/G4340A	--	rs17300865	RPS	1.68E-02	9	1	9.86	1.15 84.54
19	MTHFR	I/2/T724C	--	rs9651118	Reported,TF	1.74E-02	6	13	4.73	1.42 15.73
20	UPP2	I/2/A-7469G	--	rs6437129	RPS	2.04E-02	19	6	5.18	1.58 16.95
21	ABCC4	I/3/T-1583C	--	rs9634642	RPS	2.74E-02	18	6	4.50	1.39 14.61
22	PTEN	I/2/C3284A	--	rs2299939	Reported	3.18E-02	2	7	5.76	1.07 31.03
23	PTEN	I/7/T-400C	--	rs17431184	RPS,ISRE	3.18E-02	2	7	5.76	1.07 31.03
24	SMARCA 2	E/2/G177A	T59T	rs10964471	ESE/ESS	3.30E-02	11	3	4.05	0.98 16.76
25	CYP1B1	E/3/C1294G	L432V	rs1056836	Reported,ESE/ESS	3.98E-02	0	4	Infinite	NA NA
26	ABCC4	I/4/T588C	--	rs9524856	RPS	4.33E-02	16	4	5.71	1.57 20.78
27	MSH2	5UR//T-364G	--	rs1863332	TF	4.57E-02	8	1	8.36	0.97 72.48
28	UMPS	E/4/T1050A	C350*	rs2291078	NMD, ProteinDomain, ESE/ESS	4.58E-02	5	0	Infinite	NA NA
29	UMPS	E/6/A1336G	H446Y	rs3772809	Non-Syn, ProteinDomain, ESE/ESS	4.58E-02	5	0	Infinite	NA NA
30	UMPS	E6/3UTR /A28G	--	rs3772810	miRNA	4.58E-02	5	0	Infinite	NA NA
31	ABCB1	I/14/G81A	--	rs2235035	Reported	4.63E-02	4	11	5.50	1.46 20.67
32	ABCB1	I/17/A-76T	--	rs1922242	Reported,ISRE	4.63E-02	4	11	5.50	1.46 20.67
33	ABCG2	E/5/G421T	Q141K	rs2231142	Reported,RPS,Non-Syn,Del AA	4.64E-02	14	6	2.63	0.81 8.45
34	UGT1A1	5UR//A-1336C	--	rs3755319	Reported,TF	4.71E-02	16	6	3.43	1.06 11.04
35	UGT1A1	5UR//T-2457G	--	rs4399719	TF	4.71E-02	16	6	3.43	1.06 11.04
36	UGT1A1	5UR//T-3259G	--	rs4124874	TF	4.71E-02	16	6	3.43	1.06 11.04
37	ABCC4	I/4/C3193T	--	rs9561811	RPS	4.90E-02	3	7	3.71	0.84 16.32
38	ABCC4	I/4/C4542T	--	rs17189481	RPS	4.90E-02	3	7	3.71	0.84 16.32
39	ABCC4	I/8/C-2406T	--	rs1751021	RPS	4.90E-02	3	7	3.71	0.84 16.32

Abbreviations: Reported: previously reported in the literature to be associated with disease/function; ESE/ESS: change Exon splice enhancer/silencer; NMD: mRNA nonsense mediated decay; Non-syn: non-synonymous SNP; ProteinDomain: residing in important protein domains; miRNA: change miRNA binding site; RPS: show signature of recent positive selection; TF: change transcription factor binding site; ISRE: change intron splice regulatory element

The C allele of the rs1056836 (CYP1B1 E/3/C1294G L432V, **Table 4.3** SNP 25) in CYP1B1 gene has been recently associated with a reduced sensitivity to various anti-metabolites, including 5-FU (Laroche-Clary, Le Morvan et al. 2010). Specifically, the haplotype containing the C allele was more responsive to anti-metabolite treatment by showing a higher -log (IC50) value as compared with other haplotypes containing the G allele. This is in concordance with our observation that the C allele is uniquely found in the responders.

The 3 SNPs in the UMPS gene (**Table 4.3**, SNP 28-30) whose minor allele only occurred in non-responders, namely rs2291078 (E/4/T1050A C350*), rs3772809 (E/6/A1336G H446Y) and rs3772810 (E6/3UTR/A28G), were in a perfect LD block within Cohorts 1 and 2. Although there were no published reports about SNPs within UMPS gene associated with drug response, the predicted functional significance of the SNPs is consistent with the observed association with 5-FU response. The UMPS gene is important in determining 5-FU response, as it converts 5-FU into its active metabolite, FUMP, which can participate in both the cell toxicity pathway as well as in the "anti-metabolite" pathway. In the cell toxicity pathway, FUMP can be further converted into FUTP and FdUTP and get incorporated into RNA and DNA respectively. In the "anti-metabolite" pathway, FUMP can be converted into FdUMP and inhibits TYMS. The molecular functions of the three alleles unique to the non-responders are all associated with the disruption of the UMPS gene function and support their unique presence in the non-responders. The A allele of rs2291078 (UMPS E/4/T1050A C350*) creates a stop codon in exon 4, likely resulting in non-sense mediated decay since the stop codon occur more than 50 bps from the last exon-exon junction would lead to a decrease in UMPS mRNA abundance. At the same time, translation from mRNA containing this allele may be further attenuated by

the G allele of the 3'UTR SNP rs3772810 (UMPS E6/3UTR/A28G). The G allele is predicted to create binding sites for miRNA 23a, 23b and 130a*. The miRNA 23a is shown to be up-regulated under hypoxic conditions (Kulshreshtha, Ferracin et al. 2007), which is common in tumors, and hence may possibly suppress the expression of UMPS mRNA containing the G allele. In addition to suppressing the expression level of the gene, the non-synonymous SNP in the LD block, rs3772809 (UMPS E/6/A1336G H446Y), may further enhance the suppression of gene function by disrupting the function of the protein product. It is predicted to reside in an important domain of the protein (Ribulose-phosphate binding barrel (superfamily SSF51366), pyrF: orotidine 5'-phosphate decarboxylase (HMMTigr TIGR01740)) even though it is not explicitly predicted to be deleterious by Polyphen and other tools.

In HapMap release 28, these non-responder specific alleles of the SNPs occur at ~5% in the CHB and JPT populations, 0.5% in the CEU and 0% in the YRI. This low frequency of these alleles especially in the CEU and YRI populations suggests that they may be under negative selection consistent with the predicted deleterious nature of these alleles.

In "Cohort 2", the three SNPs in the UMPS gene again showed the same allele carried uniquely by non-responders, even though the P values were still not statistically significant (**Table 4.4**). The observation that out of a total of 102 patients when both cohorts combined, no responders were found to carry this allele suggests that this allele is a "causal" allele for determining the response to 5-FU. The non-statistical significant data obtained suggests that this may not be the only "causal" alleles for 5-FU response and there are likely other alleles that also play a role in 5-FU response suggesting "locus heterogeneity" of response to 5-FU

Table 4.4: The OR and P value for the three SNPs showing one allele unique to non-responders in different cohorts.

						Cohort 1						Cohort 2						Combined					
							Allele Count						Allele Count						Allele Count				
SN	Gene	mRNA Location	AA Change	rs No.	Function Summary	P	NR (30)	Rsp (24)	OR	OR 95 CI		P	NR (98)	Rsp (26)	OR	OR 95 CI		P	NR (128)	Rsp (50)	OR	OR 95 CI	
1	UMPS	E/4/T1050A	C350*	rs2291078	NMD,Non-Syn, ProteinDomain, ESE/ESS	4.58E-02	5	0	Infinite	NA	NA	3.42E-01	6	0	Infinite	NA	NA	3.13E-02	11	0	Infinite	NA	NA
2	UMPS	E/6/A1336G	H446Y	rs3772809	Non-Syn, ProteinDomain, ESE/ESS	4.58E-02	5	0	Infinite	NA	NA	3.42E-01	6	0	Infinite	NA	NA	3.13E-02	11	0	Infinite	NA	NA
3	UMPS	E6/3UTR /A28G	--	rs3772810	miRNA	4.58E-02	5	0	Infinite	NA	NA	3.42E-01	6	0	Infinite	NA	NA	3.13E-02	11	0	Infinite	NA	NA

4.3.3 The three SNPs in the UMPS gene reside in a region of low LD and will not be "tagged" by "tagging SNP" of r^2>0.8

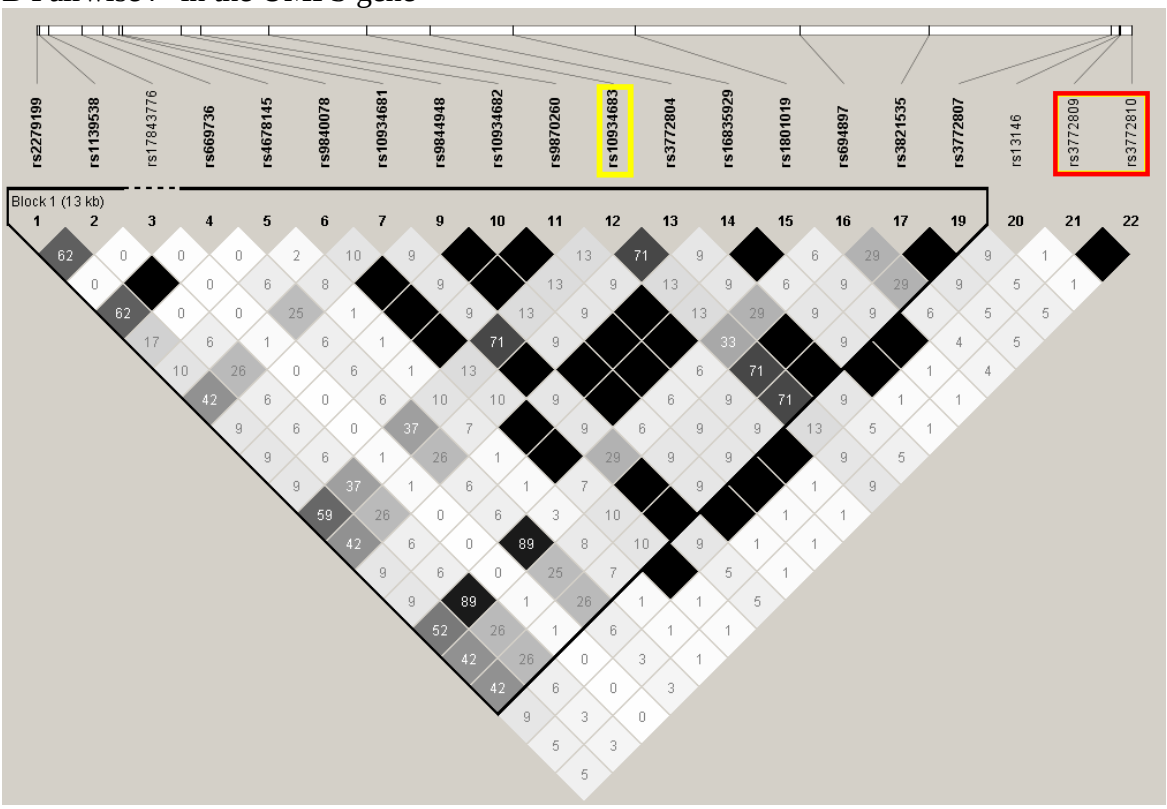
Since most current GWAS studies employ tagging SNP to identify SNPs associated with disease or drug response, it is worthwhile to examine if these three SNPs that I found to be associated with response to 5-FU/oxaliplatin can be "tagged" using pairwise tagging.

Only two of the three SNPs are genotyped (framed in red in **Figure 4.9**) in the HapMap CHB population (Release 27). The pairwise r^2 value between these SNPs and all the other 18 SNPs in the UMPS gene and even when the region is extended to 1Mb is very low. Within the gene, the best r^2 of 0.09 was obtained by a common SNP rs10934683 (MAF=0.44, framed in yellow in **Figure 4.9**), which was genotyped in our genotyping panel. This SNP is not associated with 5-FU response in any of the cohorts (P = 0.30 in "Cohort 1", P = 0.32 in "Cohort 2" and P = 0.27 when the two cohorts are combined). In the 1 Mb flanking region, the SNP with the highest r^2 is only 0.60.

Figure 4.9: The LD diagram for the SNPs in the UMPS gene region. The SNPs framed in red are the SNPs showing one allele unique to the non-responders. A. LD plot showing pairwise D'. B. LD plot showing pairwise r^2

A Pairwise D' in the UMPS gene

B Pairwise r^2 in the UMPS gene



4.3.4 Other interesting SNPs affecting 5-FU/Oxaliplatin drug response

In addition to the abovementioned SNPs that occur only in non-responders, several other interesting SNPs with lower P-values were also found to be associated with drug response in our patients before multiple test correction in Cohort 1. The VNTR (rs34743033) polymorphism containing the embedded SNP (rs2853542) was found to be significantly associated with drug response in our study with a P value of 0.015 and OR of 7.6 (**Table 4.3**, SNP 12). These 2 polymorphisms has been previously reported to be associated with 5-FU efficacy measured by tumor response (**Table 4.1**) which alludes to the robustness of our method without multiple test correction for identifying previously reported associated SNPs.

Another noteworthy observation from Cohort 1 is that the 39 SNPs that were associated with drug response before multiple test correction belong to only 14 genes suggesting that several SNPs within the same gene were found to have significant association. For example, 14 different SNPs within ABCC4 were found to be associated with drug response including SNP, rs4148485 (**Table 4.3**, SNP 1), within the intron 9 of the ABCC4 gene which showed signatures of recent positive selection and has the lowest P-value of 1.98x10-3 and a strong OR of 6.57. ABCC4 is a membrane transporter that resides at both the apical and luminal surface of the hepatocyte (Keppler 2005).Unlike ABCG2, another member of the ABC family of transporter, which localizes at one side of the cell, ABCC4 may plays roles in both in and out flux of 5-FU and may potentially have a more significant role in 5-FU pharmacokinetics (PK).

A total of 38 SNP in 13 genes in Cohort 2 are associated with drug response (**Table 4.5**). Interestingly, the SNP rs1801019 (**Table 4.5**, SNP 8) in the UMPS gene

Table 4.5: List of SNPs showing P<0.05 in "Cohort 2" ranked by P value in this cohort.

						Cohort 2					
							Allele Count				
SN	Gene	mRNA Location	AA Change	rs No.	Function Summary	P	NR (98)	Rsp (26)	OR	OR 95 CI	
1	ATP7A	E23/3UTR/T1960C	--	rs1062472	RPS	3.10E-03	22	14	4.03	1.63	9.97
2	RRM1	5UR//A-2723G	--	rs3750996	TF	4.18E-03	30	1	11.36	1.47	87.82
3	DLG5	E/23/G4442T	P1481Q	rs2289310	Reported,Non-Syn,ProteinDomain,DelAA,ESE/ESS	5.49E-03	15	11	4.06	1.56	10.52
4	RRM1	5UR//T-265G	--	rs1735068	TF	9.52E-03	19	12	3.56	1.42	8.94
5	RRM1	5UR//T-659C	--	rs1662162	TF	1.16E-02	20	12	3.34	1.34	8.34
6	RRM1	5UR//A-4023G	--	rs3794050	TF	1.16E-02	20	12	3.34	1.34	8.34
7	UMPS	5UR//T-1256A	--	rs12492095	TF	1.30E-02	11	9	4.19	1.51	11.64
8	**UMPS**	**E/3/G638C**	**G213A**	**rs1801019**	**Reported, Non-Syn, ESE/ESS**	1.30E-02	11	9	4.19	1.51	11.64
9	APC	I/2/C-230A	--	rs2464805	RPS,ISRE	1.84E-02	6	6	4.60	1.34	15.75
10	SMARCA2	I/28/G-8275T|I/29/G2824T	--|--	rs7048976	RPS	2.14E-02	39	4	3.64	1.16	11.36
11	RRM1	5UR//G-2528A	--	rs1561876	TF	2.25E-02	21	12	3.14	1.27	7.81
12	RRM1	E/19/G2232A	A744A	rs1042858	ESE/ESS	2.25E-02	21	12	3.14	1.27	7.81
13	RRM1	E19/3UTR/C316A	--	rs1042927	miRNA	2.25E-02	21	12	3.14	1.27	7.81
14	WDR7	I/13/A323G	--	rs11664579	Reported	2.36E-02	10	8	3.91	1.36	11.28
15	TFRC	E/4/C424T	S142G	rs3817672	Reported,Non-	2.47E-02	14	9	3.18	1.18	8.52
16	ATP7A	E/10/G2299C	V767L	rs2227291	Non-Syn,	2.49E-02	22	12	2.96	1.20	7.32
17	ATP7A	I/12/C-882A	--	rs17139617	RPS	2.49E-02	22	12	2.96	1.20	7.32
18	ABCC4	3DR//A75075G|E31/3UTR/A879G	--|--	rs1059751	Reported functional	2.52E-02	38	17	2.98	1.21	7.37
19	ABCB1	5UR//G-4254T	--	rs17160359	TF	2.88E-02	1	3	12.65	1.26	127.26
20	ABCC4	3DR//A74234C|E31/3UTR/A38C	--|--	rs3742106	miRNA	2.89E-02	40	17	2.74	1.11	6.76
21	ABCC4	3DR//T74507C|E31/3UTR/T311C	--|--	rs4148551	Reported functional	2.89E-02	40	17	2.74	1.11	6.76
22	APC	E/16/T5465A	V1822D	rs459552	Reported,Non-	3.31E-02	5	5	4.43	1.17	16.69
23	APC	I/6/T-3774C	--	rs2431238	RPS	3.31E-02	5	5	4.43	1.17	16.69
24	CDC2	5UR//C-3953T|I/1/C263T	--|--	rs2448341	TF,ISRE	3.95E-02	42	5	3.15	1.10	9.04
25	ATP7A	E23/3UTR/G2241C	--	rs17139614	TF	4.26E-02	0	2	Infinite	NA	NA
26	RRM1	5UR//C-3890T	--	rs7934581	TF	4.34E-02	13	8	2.91	1.05	8.03
27	ERCC6	I/6/T871G	--	rs4253101	RPS	4.51E-02	50	7	2.83	1.09	7.33
28	SLCO6A1	I/10/A-2493G	--	rs1562961	RPS	4.63E-02	38	16	2.53	1.04	6.14
29	SLCO6A1	I/11/C702T	--	rs10062613	RPS	4.63E-02	38	16	2.53	1.04	6.14
30	SLCO6A1	I/12/A7020C	--	rs6873738	RPS	4.63E-02	38	16	2.53	1.04	6.14
31	SLCO6A1	I/12/G-738T	--	rs6877722	RPS	4.63E-02	38	16	2.53	1.04	6.14
32	SLCO6A1	I/12/T517C	--	rs1901512	RPS	4.63E-02	38	16	2.53	1.04	6.14
33	SLCO6A1	I/3/T142C	--	rs10041525	RPS,ISRE	4.63E-02	38	16	2.53	1.04	6.14
34	SLCO6A1	I/3/T220C	--	rs10041507	RPS,ISRE	4.63E-02	38	16	2.53	1.04	6.14
35	SLCO6A1	I/4/A-9G	--	rs11746217	RPS,ISRE	4.63E-02	38	16	2.53	1.04	6.14
36	SLCO6A1	I/6/A4781G	--	rs1452057	RPS	4.63E-02	38	16	2.53	1.04	6.14
37	SLCO6A1	I/9/G1986T	--	rs1901521	RPS	4.63E-02	38	16	2.53	1.04	6.14
38	SLCO6A1	I/9/T5068G	--	rs1901522	RPS	4.63E-02	38	16	2.53	1.04	6.14

Abbreviations: Reported: previously reported in the literature to be associated with disease/function; ESE/ESS: change Exon splice enhancer/silencer; NMD: mRNA nonsense mediated decay; Non-syn: non-synonymous SNP; ProteinDomain: residing in important protein domains; miRNA: change miRNA binding site; RPS: show signature of recent positive selection; TF: change transcription factor binding site; ISRE: change intron splice regulatory element

which was significantly associated with capecitabine response in Cohort 2 was previously reported to be associated with UMPS mRNA level, UMPS enzyme activity and 5-FU toxicity (Ichikawa, Takahashi et al. 2006). Note that UMPS is the gene containing the 3 SNPs in both cohorts with infinite Odds Ratios mentioned preciously. Another noteworthy gene enriched with SNPs associated with drug response in Cohort 2 is the RRM1 gene where 8 SNPs in 3 LD blocks were associated with drug response. A 5'UTR SNP (5UTR/C-37A) in the RRM1 gene was reported to be associated with the transcription level of RRM1 (Dong, Guo et al. 2010; Rodriguez, Boni et al. 2011) and may determine the treatment outcome of gemcitabine, a nucleotide analogue like 5-FU, in metastatic CRC patients (Rodriguez, Boni et al. 2011) and other cancer patients (Dong, Guo et al. 2010). This SNP was not deposited into the dbSNP database and hence was not interrogated in this study. Nonetheless, similar to the reported promoter SNP, all three LD blocks in the RRM1 have at least one SNP associated with capecitabine response also localized to the promoter region. Curiously, SNPs associated with drug response in Cohort 1 is different from those in Cohort 2 (**Table 4.4** and **Table 4.5**). This is not surprising because these 2 cohorts comprise patients that have different drugs administered. Cohort 1 patients received 5-FU and oxaliplatin whereas Cohort 2 patients received only Capecitabine, precursor of 5-FU. Nonetheless, 4 genes, ABCC4, UMPS, ABCB1 and SMARCA2, albeit with different SNPs were associated with drug response in both cohorts. Three (ABCC4, UMPS, ABCB1) of these 4 genes are involved in the pharmacokinetics of 5-FU. This data suggests that although different SNPs are involved in drug response in the 2 cohorts, SNPs within the same 4 genes, 3 of which are in the 5-FU PK pathway, are associated with drug response in both cohorts. Additionally, a significant proportion (38-46%) of the genes that were associated with

drug response in Cohort1 as well as Cohort 2 is in the 5-FU PK pathway. Hence, it may be worthwhile to explore a Gene Set or pathway based approach to evaluate association with drug response in these patients.

4.3.5 Gene Set Analyses highlights the importance of the '5-FU PK' pathway

While single marker analyses has thus far been the most popular method to identify association of SNPs with disease, one limitation of this conventional approach for complex disease association studies is that there may not be sufficient statistical power to identify a true positive when many SNPs are interrogated simultaneously. Alternative complementary approaches have recently been developed to address the power limitations of single marker approaches including association analyses using haplotype based method or regression-based methods (Balding 2006) as well as the most recent pathway-based association approaches (Wang, Li et al. 2010).

In this study, although I observed several potentially functional SNPs that have potentially interesting association with drug response, none of these remained significant after multiple test correction since the sample size is small. These interesting SNPs include 3 SNPs within UMPS gene that has infinite odds ratio since the minor allele only occurs in the non-responders and a few SNPs that confirms previously reported association with the same drug. I thus employed pathway-based approach to evaluate if our observed association prior to multiple test corrections of multiple SNPs within a group of related genes have consistent but moderate deviation from chance suggesting that these genes and mechanisms are involved in drug response.

Currently there are several algorithms to perform pathway-based association analyses and these algorithms can be group into 2 categories, the 'competitive' and 'self-contained' approach of hypothesis testing. In the 'Competitive' approach, the statistics of genes in a selected pathway is compared with the statistics of all the other pathways to evaluate if the genes of the selected pathway are more associated with the disease/drug response. In the 'self-contained' approach, the test statistics of the selected pathway is compared against the expected test statistics generated by random sampling, for example.

I thus selected one 'competitive' approach, namely GSA-SNP (Nam, Kim et al. 2010) and a 'self-contained' approach, namely PODA (Braun and Buetow 2011) to further analyse our data.

Table 4.6 shows the results from the analyses using GSA-SNP (top) and PODA (bottom) for Cohort 1. In Cohort 1, ~85% (23 or 27) of the patients were treated with 5-FU/Capecitabine with oxaliplatin. GSA-SNP analyses revealed that although the 5-FU PK, platinum as well as the combined 5-FU and Platinum pathways were statistically significant before multiple test corrections, only the combined 5-FU and Platinum pathway remained statistically significant after multiple test correction. This is consistent with the drug treatment of Cohort 1 where majority of the patients received the 5-FU and Oxaliplatin combination treatment suggesting for this Cohort of patients, genes from both 5-FU as well as platinum pathways play roles in their drug response. These results were confirmed using PODA, a self-contained approach which is different from GSA-SNP which is a 'competitive' approach of analyses. As evident in the lower panel of **Table 4.6**, the 5-FU PK, Platinum as well as combined 5-FU and platinum pathways showed significant association with drug response for Cohort 1.

Table 4.6: The gene set analysis results for "Cohort 1". A. Results from GSA-SNP. The "Set size" column shows number of genes in the set and "Gene count" column shows the number of genes successfully genotyped in each set. The "P-value" column is derived based on the "Z-score" and the last column shows if the gene set with significant, the P-value would withstand Bonferroni correction. B. Results from PoDA. "DSp" is the distinction score for each gene set and "P-value" column shows the permutation-based P-value for DSp.

A.

Set name	Gene count	Set size	Z-score	P-value	Pass MTC
5-FU and Platinum Pathway	76	76	2.54	0.006	YES
5-FU PK	23	23	1.91	0.028	No
Platinum pathway	45	45	1.88	0.030	No
Colorectal Cancer Associated	79	80	0.92	0.179	--
5-FU Activator	13	13	0.82	0.206	--
5-FU PD	10	10	0.51	0.304	--

B.

Set name	DSp	P-value
5-FU PK	1.64	0.026
5-FU and Platinum Pathway	1.37	0.037
Platinum pathway	1.37	0.048
Colorectal Cancer Associated	0.98	0.160
5-FU Activator	0.25	0.457
5-FU PD	-0.37	0.716

Table 4.7 shows the results from the analyses using GSA-SNP (top) and PODA (bottom) for Cohort 2. All patients in Cohort 2 only received Capecitabine, a precursor of 5-FU. The GSA-SNP analyses revealed 5-FU Activator, 5-FU PK as well as combined 5-FU and Platinum pathways as statistically significant even after multiple testing corrections. PODA analyses showed that only the 5-FU activator is statistically significant. It is interesting to note that 5-FU activator is only statistically significant in Cohort 2 using both GSA-SNP as well as PoDA analyses but is not significant in Cohort 1 which has only very few patients receiving Capecitabine which is a precursor of 5-FU and has to be converted to 5-FU by genes within the '5-FU activator' pathways. Taken together, these data suggest that when

patients are given Capecitabine rather than 5-FU, genes within the 5-FU activation pathway will play a more significant role. Additionally the 5-FU PK pathway seemed to be more significant than the 5-FU PD pathway for these 2 Cohorts of patients.

Table 4.7: The gene set analysis results for "Cohort 2". A. Results from GSA-SNP. The "Set size" column shows the number of genes in the set, and "Gene count" column shows the number of genes successfully genotyped in each set. The "P-value" column is derived based on the "Z-score" and the last column shows if the gene set with a significant P-value could withstand Bonferroni correction. B. Results from PoDA. "DSp" is the distinction score for each gene set and "P-value" column shows the permutation-based P-value for DSp.

A.

Set name	Gene count	Set size	Z-score	P-value	Pass MTC
5-FU Activator	13	13	2.96	0.002	YES
5-FU and Platinum Pathway	76	76	2.92	0.002	YES
5-FU PK	23	23	2.83	0.002	YES
5-FU PD	10	10	1.35	0.089	--
Platinum pathway	45	45	1.24	0.108	--
Colorectal Cancer Associated	79	80	0.58	0.282	--

B

Set name	DSp	P-value
5-FU Activator	1.43	0.029
Platinum pathway	0.49	0.128
5-FU and Platinum Pathway	0.23	0.134
Colorectal Cancer Associated	-0.38	0.400
5-FU PD	-0.30	0.505
5-FU PK	-0.71	0.602

This represents the first report to highlight the significant role of genes within the 5-FU activator pathway for patients receiving the Capecitabine regime.

4.4 Summary

In this study, I found three pfSNPs in the UMPS gene, which is a "5-FU PK" gene, consistently showed one allele unique to the non-responders in both cohorts. The P values for these three pfSNPs in the LD block were less than 0.05 in "Cohort 1" as well as in the combined cohort. Although the P values were not statistical significant in "Cohort 2", this may be explained by the polygenic nature of the drug response phenotype, where these SNPs may cause the non-responsiveness; but, they only explain a small portion of all the non-responders who may carry other SNPs that will cause the non-responsiveness. Nevertheless, it is also possible that the low MAF of these SNPs and the small sample size contributed to the absence of such alleles in the responders. Follow up studies with larger sample size would be desirable to rule out this possibility.

When LD analysis was carried out, it was seen that the three pfSNPs with one allele unique to the non-responders were in perfect LD with each other but were in low LD (r^2 <0.09) with other gene region SNPs, as well as SNPs in the 1M bps flanking region (r^2 <0.6). This suggests that these SNPs will not be identified when tagging SNP approach is employed and highlights the limitations of the tagging SNP approach to identify the real causal pfSNP in association studies.

Besides these three pfSNPs in the UMPS gene, a number of other pfSNPs were also found to be interesting. The VNTR (rs2853542) polymorphism containing the embedded SNP (rs34743033) in the TYMS gene, which was previously associated with tumor response and patient survival, was found with a P value of 0.015 and OR of 7.6 in "Cohort 1". There were also 14 different SNPs within the ABCC4 gene found to be associated with drug response including the SNP (rs4148485) with the

lowest P value in cohort 1. The non-synonymous SNP in the UMPS gene (rs1801019, E/3/G638C, G213A) previously found to be associated with UMPS mRNA level, UMPS enzyme activity and 5-FU toxicity was found to be associated with drug response in "Cohort 2" as well as in the combined cohort. Eight SNPs in the RRM1 gene were found to be associated with drug response and 2 of them were close to the SNP (5UTR/C-37A) which had been associated with gemcitabine drug response in CRC patients. When gene set analysis was carried out, the "5-FU and Platinum pathway" as a whole was shown to be statistically significant by both GSA-SNP and PoDA in cohort 1 which is consistent with the drug regime used. In cohort 2, the "5-FU Activator" gene set, which is a subset of the "5-FU PK" pathway, was shown to be statistically significant by both methods and highlights the significant role of this gene set in determining Capcitabine efficacy.

CHAPTER 5: CONCLUSION

This thesis began with building a genome-scale potentially functional SNP (pfSNP) database, followed by characterization of the collection of pfSNPs in NCBI dbSNP129. It has been shown that these pfSNPs do not overlap with the SNP sets from the popular genotyping chips, and may also function as good tagging SNP. The tagging efficiency of the pfSNPs was comparable to the Illumina t-SNPs. The "pfSNP-centric tagging" SNP set, which comprised polymorphic pfSNPs and additional tagging SNPs identified from the genome, may cover the human genome more efficiently with a higher r^2 value compared with the existing Illumina t-SNP set. A biologist-friendly web-resource was then built to facilitate the pfSNPs to be used for different applications. A gene-based association study, using a pfSNP panel, was then carried out to identify pfSNPs associated with anti-cancer drug response in metastatic colorectal cancer patients. Three pfSNPs (one non-sense, one non-synonymous and one 3' UTR SNP that changes miRNA binding) in the UMPS gene were found, with one allele unique to the non-responders, and the potential molecular functions of these pfSNPs supports the association. These SNPs were in low LD with other SNPs in the UMPS gene as well as other SNPs in the 1 Mb flanking regions. This shows that using pfSNPs directly may enhance the power of a study because they may not be well-covered by other tagging SNPs. Pathway-based association analysis using pfSNPs also highlighted the relative importance of the "5-FU PK" pathway, as well as its subset "5-FU activator" pathway, in determining patient response in different treatment regimens used.

5.1 Current gaps in association study and using pfSNP dataset as a possible solution

With the abovementioned results, I will discuss the current gaps in association study and how the pfSNPs resource may be used to fill in these gaps.

5.1.1 Candidate gene-based association studies

Small sample size and an inadequate marker selection strategy that only targets non-synonymous SNPs in candidate gene-based association studies often result in difficulties in replicating the identified association (Hirschhorn, Lohmueller et al. 2002). In candidate gene-based association studies with a limited number of genes, the use of pfSNPs would provide coverage for SNPs in non-coding regions, which are proven to be important in determining gene functions. On the other hand, in candidate gene-based association studies targeting all the genes in one or more relevant of the pathways, the use of pfSNPs may help to reduce genotyping cost and multiple testing problems due to the elimination of non-functional SNPs in "function-rich" regions (e.g., not all promoter SNPs will change a transcription factor binding site). Furthermore, a gene set/pathway-based association analysis may be carried out with more power by targeting the causal SNPs directly and generating stronger association signals.

5.1.2 "Common Disease, Common Variants" hypothesis-based genome-wide association studies

The current CDCV hypothesis-based GWAS studies are often inundated with the problem of "Missing Heritability" (Manolio, Collins et al. 2009; Eichler, Flint et al. 2010) where the markers found to be associated with the phenotype only explain part of the heritability observed. The cause of the "Missing Heritability" remains unclear.

In my opinion, the problem of "Missing Heritability" may be partly caused by inadequate coverage of common functional SNPs by current genotyping platforms due to non-optimal marker selection. For the popular tagging SNP-based GWAS (here after referred to as "tagging SNP-centric" approach), the efficiency of a tagging SNP to represent the real causal one remain unclear since many factors may influence the strength of the marker (Terwilliger and Hiekkalinna 2006) and, under certain extreme conditions, the marker may never show sufficient evidence of disease association, even with infinite sample sizes (Terwilliger and Hiekkalinna 2006) (Please refer to Chapter 1 for more detail). Practically, the popular Illumina t-SNP set is capable of covering the entire human genome at an r^2 value of 0.1 to 0.2, which is too low to be considered useful (Pe'er, de Bakker et al. 2006) (Please refer to Chapter 2 for more detail).

Using an alternative to the "tagging SNP-centric" approaches in GWAS has been proposed by Risch (Risch 2000) in the form of "gene-centric, non-synonymous SNP favoured" approach at the beginning of the new millennium. This approach directly cover SNPs prioritized mainly according to the gene region in which they reside, with special emphasis on non-synonymous SNPs (Risch and Merikangas 1996). Risch foresees that the priority may need further refinement if new mechanism

of gene function/regulation is found. For example, in year 2000, SNPs in the 5'UTR was thought to be more important in influencing phenotype than the 3'UTR by Risch (Risch 2000) because 5' UTR SNPs were more conserved (Halushka, Fan et al. 1999). However, with the discovery of miRNAs in 2004 which can regulate several mRNAs via its binding to the 3'UTR, SNPs within the 3'UTR region is now recognized to play more important roles in modulating phenotype.

The "gene-centric, non-synonymous SNP favoured" strategy may need to be further enhanced. The emphasis on "non-synonymous SNPs" may no longer hold true, in that more than 80% of association signals in current GWAS are present outside of the coding region, despite the over-representation of coding SNPs on these chips (Hindorff, Sethupathy et al. 2009; Manolio, Collins et al. 2009). Practically, only one GWAS study (Webb, Broderick et al. 2009) actually implemented the "gene-centric, non-synonymous SNP only" approach, and it failed to produce any meaningful results. Furthermore, new discoveries in gene function and regulation also support the inclusion of more functional SNPs from other gene regions. Besides promoter SNPs and 3'UTR SNPs, which are now well-known for their importance in determining gene transcription and translation efficiency, synonymous SNPs previously thought to be "functionally neutral", and even intronic SNPs outside the mRNA splicing donor/receptor sites, are now recognized to be important in affecting splicing variant production and thus gene function. Even the inter-genic regions are no longer considered to be pure "junk", and the HAPMAP ENCODE project discovered that most of these "junk" regions are transcribed into RNA (Birney, Stamatoyannopoulos et al. 2007). All of these findings support the possibility that causal variants may well be present in other non-coding regions and extending the coverage outside the coding region may be a sound alternative.

As an extension to the "gene-centric, non-synonymous SNP favoured" approach, I propose a "pfSNP-centric" approach, which directly targets a set of pfSNPs anywhere in the human genome. As with the "gene-centric, non-synonymous SNP favoured" approach, the "pfSNP-centric" approach is a form of "direct association". However, the "pfSNP-centric" strategy proposed here attempts to cover all of the functional SNPs in different gene regions equally, and the coverage extends to inter-genic regions. In the meantime, stringent filters are employed where it is possible to control the total number of SNPs included. It may benefit GWAS in three ways. First, sufficient coverage in both gene and inter-genic regions will be provided so it will be more "genome-wide". Second, excessive genotyping costs and multiple testing problems will be avoided due to the elimination of non-functional SNPs in "function-rich" regions. Third, a gene set/pathway-based analysis may be carried out with more power by targeting the causal SNPs directly and generating stronger association signals.

Despite all these benefits, the "pfSNP-centric" approach proposed here may not be thorough enough to cover all the functional SNPs in the human genome in light that our current understanding of the human genome is far from complete. Therefore, a "pfSNP-centric tagging" approach, which uses the polymorphic pfSNPs as well as additional tagging SNPs identified from the genome, may be more practical for now. In Chapter 2, I showed that the polymorphic pfSNPs identified thus far can act as good tagging SNPs, with a tagging efficiency that is comparable to the Illumina t-SNP set. I also showed that the "pfSNP-centric tagging" SNP set would cover the human genome with a higher r^2 value as compared with the Illumina t-SNP set when the same number of markers is used. The increased r^2 value, ranging from 0.6 to 0.9, offered by the "pfSNP-centric tagging" SNP set would lead to higher power to

discover causal SNPs and thus provide a possible remedy to the "Missing Heritability" problem due to inadequate coverage.

5.1.3 "Common Disease, Rare Variants" hypothesis-based genome wide association studies

In a recent interview, experts (Eichler, Flint et al. 2010) acknowledged that rare SNPs could play a role in addressing the "Missing Heritability" problem for GWAS. The current focus of the field should be, therefore, to find new a statistical method to better delineate the additive/multiplicative effects of rare SNPs in the statistical analysis process.

A number of such tests have already been proposed (Morgenthaler and Thilly 2007; Li and Leal 2008; Madsen and Browning 2009; Hoffmann, Marini et al. 2010; Liu and Leal 2010; Makowsky, Pajewski et al. 2011; Cardin, Mefford et al. 2012). A popular approach is to combine or "collapse" rare variants by different criteria (Morgenthaler and Thilly 2007; Li and Leal 2008; Madsen and Browning 2009; Hoffmann, Marini et al. 2010; Cardin, Mefford et al. 2012). This approach is sensitive to the misclassification problem, where inclusion of non-functionally important SNPs and exclusion of functionally important SNPs would both cause a loss of statistical power (Li and Leal 2008; Madsen and Browning 2009; Hoffmann, Marini et al. 2010). Therefore, filtering for potentially functional SNPs before collapsing is crucial to its success. The Cohort Allelic Sums Test (CAST) utilises Human Gene Mutation Database (http://www.hgmd.org/), which stores previously identified disease-related SNPs; the Combined Multivariate and Collapsing (CMC) method (Li and Leal 2008), in comparison, recommends Polyphen or SIFT for this purpose. However, these tools all provide a limited SNP coverage. pfSNPs set may enhance the classification of the

SNPs by providing a more complete set of reported, predicted as well as inferred functional SNPs both inside and outside the coding region.

5.2 Current applications of the pfSNP web-resource and developing of the future pfSNP resource

As discussed in Chapter Three, the current pfSNP web-resource may be used for three purposes. The first is to select SNPs to investigate in a candidate gene-based association study (GBAS). The second purpose would be to analyze results obtained from a GWAS study. The third would be to predict, *in silico*, the functional consequences of a SNP to guide experimental investigations. Since the third function is straight forward, the following section will review the current features of the pfSNP web-resource in facilitating the first two applications and discuss the future directions of the pfSNP web-resource development to further enhance the functionalities. Moreover, as whole genome sequencing is replacing SNP-Chip based GWAS, I will discuss the challenges and future direction of pfSNP web-resource development to accommodate the analysis requirement of whole genome sequencing.

(i) SNP selection for candidate gene-based association studies (GBAS)

Selecting SNPs in candidate genes for study in genetic association studies is not easy. Selecting too few SNPs for study, whilst easier from a logistics perspective, may increase the risk of omitting the causative SNP. Conversely, selecting too many SNPs for a GBAS can have its own set of problems including increasing in type II error, where biologically significant

SNPs are inadvertently omitted from further study due to a more stringent threshold of statistical significance post multiple test correction (Balding 2006). Genotyping too many SNPs can also add to the cost of the study without contributing additional information towards the study question.

One possible solution to the problem of having too many SNPs for GBAS is to use the pfSNP web-resource to filter SNPs according to defined criteria, such as potential functionality and minor allele frequency.

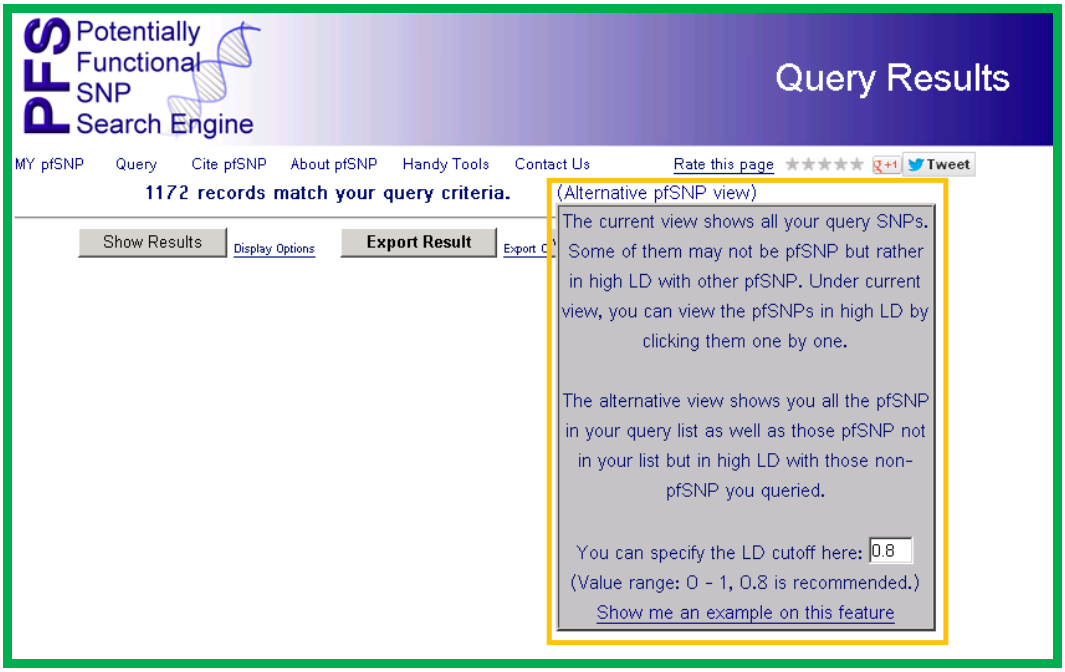
Reported functional SNPs from Online Mendelian Inheritance in Man (OMIM) and Genetic Association Database (GAD) are provided in the pfSNP web-resource so that researchers can be informed if the SNP has been previously studied. In addition, the pfSNP web-resource would allow the researcher to identify novel areas of SNP research (predicted to be functional using the pfSNP resource) that may be relevant but poorly investigated.

Furthermore, by using population-specific allele frequency data integrated into the pfSNP web-resource, researchers can also identify SNP allele frequencies in a population that is most closely related to their population of interest. This knowledge will allow researchers to plan aspects of the GBAS, such as sample size determination and determining the SNPs to genotype depending on whether the researcher is interested in rare or common variants.

The second possible solution to the problem of having too many SNPs for GBAS is to select tagging SNPs. Genotyping tagging SNPs for GBAS serves to reduce the genotyping costs as a result of genotyping fewer SNPs whilst minimizing the loss of statistical power (Balding 2006) for the genetic association study. The pfSNP web-resource currently provides a function

called "Alternative pfSNP view" (**Figure 5.1**) to help identifying pfSNPs to tag a list of candidate SNPs for GBAS. This function first distinguishes pfSNPs and non-pfSNPs from the list of SNPs. The pfSNPs in the list would require no further action. For the non-pfSNPs, pfSNPs in highest LD measured by r^2 are sought as tagging SNPs. The final output, therefore, is a list of pfSNPs in the original list as well as the pfSNPs that can tag the non-pfSNPs in the original list. The limitation of the current function is that the non-pfSNP would not be covered if there is no pfSNP in high LD with it.

Figure 5.1: The current function provided in pfSNP web-resource to help picking pfSNP as tagging SNPs.



In the future, the pfSNP web-resource may be further developed to enhance its utilization in GBAS. The first is to include more reported functional SNPs from the literature. Currently, OMIM is still the major source providing SNPs with confirmed biological function. As OMIM only includes confirmed causal SNPs, SNPs newly reported in the literature with limited

evidence of causality would not be covered by OMIM. With new emerging text mining tools, the pfSNP web-resource may use them to identify the newly reported SNPs from recently published literature. The second is to include SNP allele frequency information from resources such as 1000 Genomes Project to enhance the SNP allele frequency filter currently limited to SNPs genotyped by the HapMap project. The third is to enhance the identification of tagging SNP functionality to provide "pfSNP-centric tagging" SNPs, where SNPs outside the pfSNP set may be picked as tagging SNP if the pfSNPs alone cannot cover all of the SNPs of interest at the r^2 value required.

(ii) Analyses of results from genome wide association studies (GWAS)

Annotating tagging SNPs showing significant associations in GWAS may be misleading. This is because the tagging SNP showing association may not be the actual causative SNP but instead is in high LD with the causative SNP. To solve this problem, the "Alternative pfSNP view" function (**Figure 5.1**) previously mentioned would be useful. It provides the pfSNPs in high LD with the tagging SNP showing significant association so that the association may be explained by the pfSNP rather than the tagging SNP without an apparent biological function.

Besides annotating the GWAS results, the researcher may face further challenges depending on the threshold of statistical significance (alpha value) set and the P values of the SNPs tested for the association. The first challenge is that an extensive set of SNPs may be deemed statistically significant for the association. The second challenge, on the contrary, is that no SNPs are found

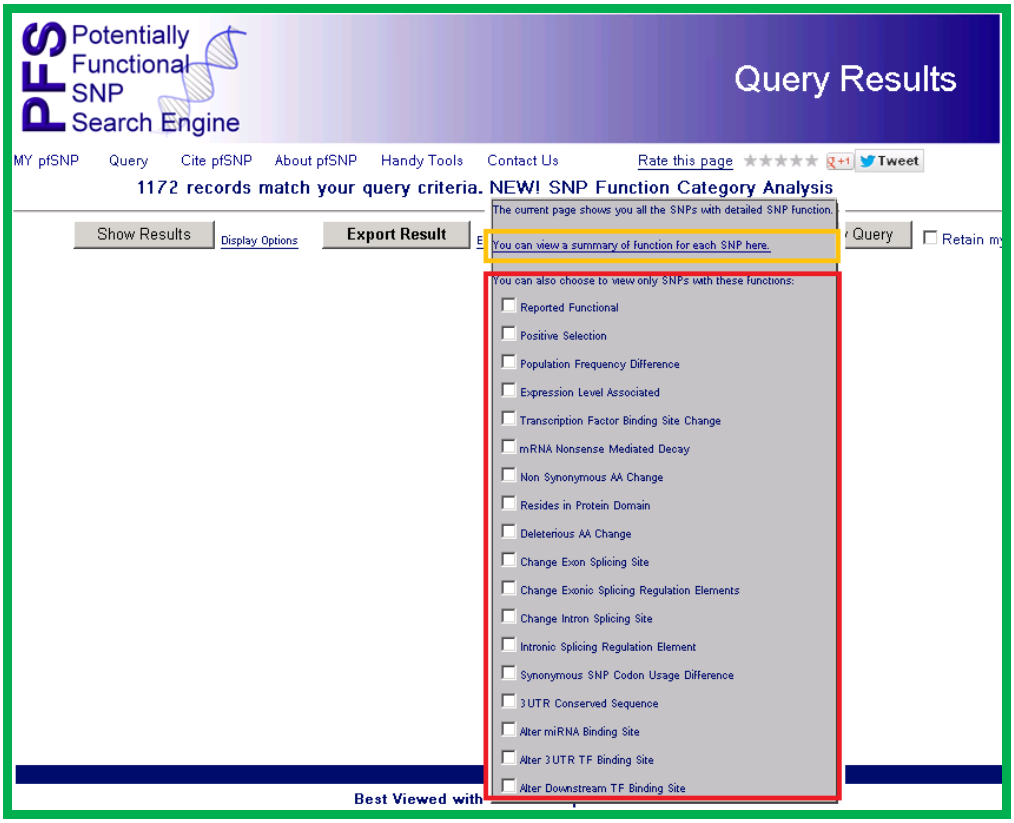
to be statistically significantly associated with the phenotype due to the stringent threshold set to account for multiple test correction.

With an extensive set of SNPs showing positive association with the phenotype, the pfSNP web-resource can be used to prioritize SNPs for further study. The pfSNP web-resource currently provides "SNP Function Category Analysis" (**Figure 5.2**), which displays a function summary for each SNP (**Figure 5.2**, framed in yellow). pfSNPs that are more likely to be functional (for example, those with predicted function and found to be associated with similar phenotypes in previous GWAS) may be easily prioritized by going through this summary. It can also facilitate prioritizing SNPs with specific function by providing a detailed filtering utility (**Figure 5.2**, framed in red). In the near future, more quantitative analytical features can be built into the "SNP Function Category Analysis". For example, Fisher's exact test can be used to detect if the SNPs in the list are enriched in a specific function category. A 2X2 contingency table can be used to tabulate the number of SNPs in four categories, namely (1) "In the SNP list with the specific function", (2) "In the SNP list without the specific function", (3) "Not in the SNP list but with the specific function" and (4) "Not in the SNP list and without the specific function". Fisher's exact test can then be applied to the contingency table and a P-value less than 0.05 in the Fisher's exact test can signify the SNPs in the list is enriched in the specific function category.

In contrast, if there were no SNPs found to be statistically significantly associated with the phenotype in a GWAS, the alpha value could be increased so as to yield a group of SNPs with small to moderate effect sizes. From this group of SNPs, pathway-based prioritization available in specialized web

resources, such as GesBaP (Medina, Montaner et al. 2009) or iGSEA4GWAS (Zhang, Cui et al. 2010), can be applied to determine if the group of SNPs generated after raising the alpha value is enriched in certain pathways and therefore the association observed is less likely to be spurious. Currently, the pfSNP and similar web-resources do not have any feature to support pathway-based prioritization as yet, and such a facility would be desirable in the future build.

Figure 5.2: The "SNP function category analysis" can help to prioritize SNPs that are more likely to be functional or to filter for SNPs belonging to specific function category. Framed in yellow is the link to display the function summary for the list of SNPs. Framed in red are the filters that can be applied individually or combined to filter for SNPs with specific functions.



(iii) Challenges and future directions in analyzing whole genome sequencing results

As whole genome sequencing technology improves, new challenges in the development of pfSNP web-resource emerge. These challenges are largely related to technical gaps, such as the capacity of the pfSNP web-resource to handle large volumes of data, and gaps in the development of tools to analyze whole genome sequencing results within a biologically meaningful context.

With the increasing amount of sequencing data generated, the demand for storing, analyzing and sharing SNP genotype data also grows. Unfortunately, the pfSNP web-resource and other similar resources are not able to handle the raw data without heavy investment in storage capacity. The large amount of data, together with the increased complexity of data analysis, would also demand a higher computing power to efficiently carry out the data analysis. One of the solutions to these problems might be cloud computing, which features essentially unlimited online storage and analytical power (2010; Stein 2010). Recent tools like "Cloud BioLinux" (Krampis, Booth et al. 2012) and "CloVR" (Angiuoli, Matalka et al. 2011) are already providing cloud enabled, virtual machine based solutions to the raw data handling and analysis challenges. I foresee similar solutions can be applicable to pfSNP web-resource where a bootable virtual machine image can be distributed in the cloud and used on-demand by the user. However, using cloud computing to store and analyze SNP data is still in its infancy with a limiting factor in slow data transfer rate provided by the current internet.

Besides these technical challenges faced by the pfSNP web resource, the complexity of the human genome may pose an even bigger challenge to the making of a successful SNP analytical resource.

We have a long way to go in our understanding of SNP architecture in the human genome. Many SNPs have yet to be discovered, as reflected in the observation that more than 2.7 million new SNPs are reported in the latest dbSNP 137 release (http://www.ncbi.nlm.nih.gov/projects/SNP/snp_summary.cgi).

We also know too little about the human genome in terms of pathways and/or regulation mechanisms that have yet to be discovered. Hirschhorn et al. (Hirschhorn 2009) pointed out that the real benefits of performing GWAS studies has been the illumination of novel pathways. The discovery of micro RNAs' impact on mRNA level (Zeng, Wagner et al. 2002) in 2002 further indicates that there may be other mechanisms affecting gene regulation yet to be discovered. However, current SNP function prediction tools were developed and trained on the basis of the current knowledge of known biological mechanisms. Hence, using current prediction tools may result in false negatives; that is, SNPs which have been deemed *in silico* as non-functional but are, in reality, functional *in vivo*. Therefore, there is a need for the pfSNP web-resource that can rapidly keep pace with current developments in biology and readily incorporate new tools.

One stark observation made since the advent of GWAS studies has been the relatively little success of GWAS in finding high-risk SNPs that can account for a large proportion of phenotypic differences (Moore 2003; Eichler, Flint et al. 2010). This observation has been particularly pertinent to complex

phenotypes. Current tools for GWAS analysis are not able to adequately analyze the genotype-complex phenotype relationships within a holistic biologically relevant framework that considers SNP-SNP and gene-gene interactions, as well as gene-environmental interactions. This problem is more severe in whole genome sequencing results analysis where the possible number of interactions is prohibitively large. Future analytical tools should consider incorporating prior biological knowledge about the SNP and gene pathway in which the SNP lies into the algorithm (Moore, Asselbergs et al. 2010).

Bibliography

(2010). "Gathering clouds and a sequencing storm: why cloud computing could broaden community access to next-generation sequencing." Nature biotechnology **28**(1): 1.

Akey, J. M., G. Zhang, et al. (2002). "Interrogating a high-density SNP map for signatures of natural selection." Genome research **12**(12): 1805-1814.

Altshuler, D. M., R. A. Gibbs, et al. (2010). "Integrating common and rare genetic variation in diverse human populations." Nature **467**(7311): 52-58.

Amigo, J., A. Salas, et al. (2008). "SPSmart: adapting population based SNP genotype databases for fast and comprehensive web access." BMC Bioinformatics **9**: 428.

Anderson, E. C. and J. Novembre (2003). "Finding haplotype block boundaries by using the minimum-description-length principle." American journal of human genetics **73**(2): 336-354.

Angiuoli, S. V., M. Matalka, et al. (2011). "CloVR: a virtual machine for automated and portable sequence analysis from the desktop using cloud computing." BMC Bioinformatics **12**: 356.

Aoki, K. F. and M. Kanehisa (2005). "Using the KEGG database resource." Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.] **Chapter 1**: Unit 1 12.

Ashburner, M., C. A. Ball, et al. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nature genetics **25**(1): 25-29.

Avi-Itzhak, H. I., X. Su, et al. (2003). "Selection of minimum subsets of single nucleotide polymorphisms to capture haplotype block diversity." Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing: 466-477.

Balding, D. J. (2006). "A tutorial on statistical methods for population association studies." Nature reviews. Genetics **7**(10): 781-791.

Barrett, J. C., B. Fry, et al. (2005). "Haploview: analysis and visualization of LD and haplotype maps." Bioinformatics **21**(2): 263-265.

Bartoszewski, R. A., M. Jablonsky, et al. (2010). "A synonymous single nucleotide polymorphism in DeltaF508 CFTR alters the secondary structure of the mRNA and the expression of the mutant protein." The Journal of biological chemistry **285**(37): 28741-28748.

Becker, K. G. (2004). "The common variants/multiple disease hypothesis of common complex genetic disorders." Medical hypotheses **62**(2): 309-317.

Benjamini, Y. and Y. Hochberg (1995). "Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing." Journal of the Royal Statistical Society Series B-Methodological **57**(1): 289-300.

Beroud, C., D. Hamroun, et al. (2005). "UMD (Universal Mutation Database): 2005 update." Human mutation **26**(3): 184-191.

Betel, D., M. Wilson, et al. (2008). "The microRNA.org resource: targets and expression." Nucleic acids research **36**(Database issue): D149-153.

Birney, E., J. A. Stamatoyannopoulos, et al. (2007). "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project." Nature **447**(7146): 799-816.

Blom, N., S. Gammeltoft, et al. (1999). "Sequence and structure-based prediction of eukaryotic protein phosphorylation sites." Journal of molecular biology **294**(5): 1351-1362.

Bobadilla, J. L., M. Macek, Jr., et al. (2002). "Cystic fibrosis: a worldwide analysis of CFTR mutations--correlation with incidence data and application to screening." Human mutation **19**(6): 575-606.

Bond, G. L., W. Hu, et al. (2005). "A single nucleotide polymorphism in the MDM2 gene: from a molecular and cellular explanation to clinical effect." Cancer research **65**(13): 5481-5484.

Braun, R. and K. Buetow (2011). "Pathways of distinction analysis: a new technique for multi-SNP analysis of GWAS data." PLoS genetics **7**(6): e1002101.

Bush, E. C. and B. T. Lahn (2008). "A genome-wide screen for noncoding elements important in primate evolution." BMC Evol Biol **8**: 17.

Cardin, N. J., J. A. Mefford, et al. (2012). "Joint association testing of common and rare genetic variants using hierarchical modeling." Genetic epidemiology **36**(6): 642-651.

Cargill, M., D. Altshuler, et al. (1999). "Characterization of single-nucleotide polymorphisms in coding regions of human genes." Nature genetics **22**(3): 231-238.

Carlson, C. S., M. A. Eberle, et al. (2004). "Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium." American journal of human genetics **74**(1): 106-120.

Cartegni, L., J. Wang, et al. (2003). "ESEfinder: A web resource to identify exonic splicing enhancers." Nucleic acids research **31**(13): 3568-3571.

Celli, J., R. Dalgleish, et al. (2012). "Curating gene variant databases (LSDBs): Toward a universal standard." Human mutation **33**(2): 291-297.

Chelala, C., A. Khan, et al. (2009). "SNPnexus: a web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms." Bioinformatics **25**(5): 655-661.

Chen, R., E. V. Davydov, et al. (2010). "Non-synonymous and synonymous coding SNPs show similar likelihood and effect size of human disease association." PloS one **5**(10): e13574.

Chorley, B. N., X. Wang, et al. (2008). "Discovery and verification of functional single nucleotide polymorphisms in regulatory genomic regions: current and developing technologies." Mutation research **659**(1-2): 147-157.

Chua, Y. J., D. Sargent, et al. (2005). "Definition of disease-free survival: this is my truth-show me yours." Annals of oncology : official journal of the European Society for Medical Oncology / ESMO **16**(11): 1719-1721.

Clark, A. G. and J. Li (2007). "Conjuring SNPs to detect associations." Nature genetics **39**(7): 815-816.

Claustres, M., O. Horaitis, et al. (2002). "Time for a unified system of mutation description and reporting: a review of locus-specific mutation databases." Genome research **12**(5): 680-688.

Cohen, V., V. Panet-Raymond, et al. (2003). "Methylenetetrahydrofolate reductase polymorphism in advanced colorectal cancer: a novel genomic predictor of clinical response to fluoropyrimidine-based chemotherapy." Clinical cancer research : an official journal of the American Association for Cancer Research **9**(5): 1611-1615.

Collins, F. S., M. S. Guyer, et al. (1997). "Variations on a theme: cataloging human DNA sequence variation." Science **278**(5343): 1580-1581.

Cordell, H. J. and D. G. Clayton (2005). "Genetic association studies." Lancet **366**(9491): 1121-1131.

Cotton, R. G., A. D. Auerbach, et al. (2008). "Recommendations for locus-specific databases and their curation." Human mutation **29**(1): 2-5.

Crawford, D. C., D. T. Akey, et al. (2005). "The patterns of natural variation in human genes." Annual review of genomics and human genetics **6**: 287-312.

Cunningham, D., W. Atkin, et al. (2010). "Colorectal cancer." Lancet **375**(9719): 1030-1047.

De Gobbi, M., V. Viprakasit, et al. (2006). "A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter." Science **312**(5777): 1215-1217.

De Mattia, E. and G. Toffoli (2009). "C677T and A1298C MTHFR polymorphisms, a challenge for antifolate and fluoropyrimidine-based therapy personalisation." Eur J Cancer **45**(8): 1333-1351.

den Dunnen, J. T., R. H. Sijmons, et al. (2009). "Sharing data between LSDBs and central repositories." Human mutation **30**(4): 493-495.

Dong, S., A. L. Guo, et al. (2010). "RRM1 single nucleotide polymorphism -37C-->A correlates with progression-free survival in NSCLC patients after gemcitabine-based chemotherapy." Journal of hematology & oncology **3**: 10.

Duan, S., W. Zhang, et al. (2008). "FstSNP-HapMap3: a database of SNPs with high population differentiation for HapMap3." Bioinformation **3**(3): 139-141.

Durbin, R. M., G. R. Abecasis, et al. (2010). "A map of human genome variation from population-scale sequencing." Nature **467**(7319): 1061-1073.

Edwards, A. O., R. Ritter, 3rd, et al. (2005). "Complement factor H polymorphism and age-related macular degeneration." Science **308**(5720): 421-424.

Eichler, E. E., J. Flint, et al. (2010). "Missing heritability and strategies for finding the underlying causes of complex disease." Nature reviews. Genetics **11**(6): 446-450.

Etienne-Grimaldi, M. C., J. Bennouna, et al. (2012). "Multifactorial pharmacogenetic analysis in colorectal cancer patients receiving 5-fluorouracil-based therapy together with cetuximab-irinotecan." British journal of clinical pharmacology **73**(5): 776-785.

Etienne, M. C., J. L. Formento, et al. (2004). "Methylenetetrahydrofolate reductase gene polymorphisms and response to fluorouracil-based treatment in advanced colorectal cancer patients." Pharmacogenetics **14**(12): 785-792.

Evans, D. M. and L. R. Cardon (2004). "Guidelines for genotyping in genomewide linkage studies: single-nucleotide-polymorphism maps versus microsatellite maps." American journal of human genetics **75**(4): 687-692.

Excoffier, L. and M. Slatkin (1995). "Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population." Molecular biology and evolution **12**(5): 921-927.

Fairbrother, W. G., R. F. Yeh, et al. (2002). "Predictive identification of exonic splicing enhancers in human genes." Science **297**(5583): 1007-1013.

Farina-Sarasqueta, A., G. van Lijnschoten, et al. (2010). "Value of gene polymorphisms as markers of 5-FU therapy response in stage III colon carcinoma: a pilot study." Cancer Chemother Pharmacol **66**(6): 1167-1171.

Fokkema, I. F., P. E. Taschner, et al. (2011). "LOVD v.2.0: the next generation in gene variant databases." Human mutation **32**(5): 557-563.

Fong, C., D. C. Ko, et al. (2010). "GWAS analyzer: integrating genotype, phenotype and public annotation data for genome-wide association study analysis." Bioinformatics **26**(4): 560-564.

Freimer, N. and C. Sabatti (2004). "The use of pedigree, sib-pair and association studies of common diseases for genetic mapping and epidemiology." Nature genetics **36**(10): 1045-1051.

Gabriel, S. B., S. F. Schaffner, et al. (2002). "The structure of haplotype blocks in the human genome." Science **296**(5576): 2225-2229.

Georges, M., A. Clop, et al. (2006). "Polymorphic microRNA-target interactions: a novel source of phenotypic variation." Cold Spring Harb Symp Quant Biol **71**: 343-350.

Goodswen, S. J., C. Gondro, et al. (2010). "FunctSNP: an R package to link SNPs to functional knowledge and dbAutoMaker: a suite of Perl scripts to build SNP databases." BMC Bioinformatics **11**: 311.

Gorlov, I. P., O. Y. Gorlova, et al. (2008). "Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms." American journal of human genetics **82**(1): 100-112.

Gosens, M. J., E. Moerland, et al. (2008). "Thymidylate synthase genotyping is more predictive for therapy response than immunohistochemistry in patients with colon cancer." Int J Cancer **123**(8): 1941-1949.

Graziano, F., A. Ruzzo, et al. (2008). "Liver-only metastatic colorectal cancer patients and thymidylate synthase polymorphisms for predicting response to 5-fluorouracil-based chemotherapy." British journal of cancer **99**(5): 716-721.

Grimson, A., K. K. Farh, et al. (2007). "MicroRNA targeting specificity in mammals: determinants beyond seed pairing." Molecular cell **27**(1): 91-105.

Gross, E., B. Busse, et al. (2008). "Strong association of a common dihydropyrimidine dehydrogenase gene polymorphism with fluoropyrimidine-related toxicity in cancer patients." PLoS One **3**(12): e4003.

Gusella, M., A. C. Frigo, et al. (2009). "Predictors of survival and toxicity in patients on adjuvant therapy with 5-fluorouracil for colorectal cancer." British journal of cancer **100**(10): 1549-1557.

Haines, J. L., M. A. Hauser, et al. (2005). "Complement factor H variant increases the risk of age-related macular degeneration." Science **308**(5720): 419-421.

Halldorsson, B. V., V. Bafna, et al. (2004). "Optimal haplotype block-free selection of tagging SNPs for genome-wide association studies." Genome research **14**(8): 1633-1640.

Halushka, M. K., J. B. Fan, et al. (1999). "Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis." Nature genetics **22**(3): 239-247.

Hamosh, A., B. J. Rosenstein, et al. (1992). "CFTR nonsense mutations G542X and W1282X associated with severe reduction of CFTR mRNA in nasal epithelial cells." Human molecular genetics **1**(7): 542-544.

Hamroun, D., S. Kato, et al. (2006). "The UMD TP53 database and website: update and revisions." Human mutation **27**(1): 14-20.

Hansen, J. E., O. Lund, et al. (1998). "NetOglyc: prediction of mucin type O-glycosylation sites based on sequence context and surface accessibility." Glycoconjugate journal **15**(2): 115-130.

Hartmaier, R. J., A. S. Richter, et al. (2012). "A SNP in steroid receptor coactivator-1 disrupts a GSK3beta phosphorylation site and is associated with altered tamoxifen response in bone." Molecular endocrinology **26**(2): 220-227.

Hastbacka, J., A. de la Chapelle, et al. (1992). "Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland." Nat Genet **2**(3): 204-211.

He, H., K. Jazdzewski, et al. (2005). "The role of microRNA genes in papillary thyroid carcinoma." Proceedings of the National Academy of Sciences of the United States of America **102**(52): 19075-19080.

Hindorff, L. A., P. Sethupathy, et al. (2009). "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits." Proceedings of the National Academy of Sciences of the United States of America **106**(23): 9362-9367.

Hirschhorn, J. N. (2009). "Genomewide association studies--illuminating biologic pathways." N Engl J Med **360**(17): 1699-1701.

Hirschhorn, J. N., K. Lohmueller, et al. (2002). "A comprehensive review of genetic association studies." Genetics in medicine : official journal of the American College of Medical Genetics **4**(2): 45-61.

Hoffmann, T. J., N. J. Marini, et al. (2010). "Comprehensive approach to analyzing rare genetic variants." PloS one **5**(11): e13584.

Hoggart, C. J., T. G. Clark, et al. (2008). "Genome-wide significance for dense SNP and resequencing data." Genetic epidemiology **32**(2): 179-185.

Ichikawa, W., T. Takahashi, et al. (2006). "Orotate phosphoribosyltransferase gene polymorphism predicts toxicity in patients treated with bolus 5-fluorouracil regimen." Clin Cancer Res **12**(13): 3928-3934.

Jakobsen, A., J. N. Nielsen, et al. (2005). "Thymidylate synthase and methylenetetrahydrofolate reductase gene polymorphism in normal tissue as predictors of fluorouracil sensitivity." J Clin Oncol **23**(7): 1365-1369.

Johnson, A. D., R. E. Handsaker, et al. (2008). "SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap." Bioinformatics **24**(24): 2938-2939.

Jorgenson, E. and J. S. Witte (2006). "A gene-centric approach to genome-wide association studies." Nat Rev Genet **7**(11): 885-891.

Karchin, R., M. Diekhans, et al. (2005). "LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources." Bioinformatics **21**(12): 2814-2820.

Kawakami, K. and G. Watanabe (2003). "Identification and functional analysis of single nucleotide polymorphism in the tandem repeat sequence of thymidylate synthase gene." Cancer Res **63**(18): 6004-6007.

Ke, X. and L. R. Cardon (2003). "Efficient selective screening of haplotype tag SNPs." Bioinformatics **19**(2): 287-288.

Keating, B. J., S. Tischfield, et al. (2008). "Concept, design and implementation of a cardiovascular gene-centric 50 k SNP array for large-scale genomic association studies." PLoS One **3**(10): e3583.

Keppler, D. (2005). "Uptake and efflux transporters for conjugates in human hepatocytes." Methods in enzymology **400**: 531-542.

Kerem, B., J. M. Rommens, et al. (1989). "Identification of the cystic fibrosis gene: genetic analysis." Science **245**(4922): 1073-1080.

Kerem, E. (2004). "Pharmacologic therapy for stop mutations: how much CFTR activity is enough?" Current opinion in pulmonary medicine **10**(6): 547-552.

Khoury, M. J. and Q. Yang (1998). "The future of genetic studies of complex human diseases: an epidemiologic perspective." Epidemiology **9**(3): 350-354.

Kim, B. C., W. Y. Kim, et al. (2008). "SNP@Promoter: a database of human SNPs (single nucleotide polymorphisms) within the putative promoter regions." BMC Bioinformatics **9 Suppl 1**: S2.

Kim, S. Y. and J. K. Pritchard (2007). "Adaptive evolution of conserved noncoding elements in mammals." PLoS genetics **3**(9): 1572-1586.

Kim, S. Y. and D. J. Volsky (2005). "PAGE: parametric analysis of gene set enrichment." BMC bioinformatics **6**: 144.

Kimchi-Sarfaty, C., J. M. Oh, et al. (2007). "A "silent" polymorphism in the MDR1 gene changes substrate specificity." Science **315**(5811): 525-528.

Kimura, M. (1979). "The neutral theory of molecular evolution." Scientific American **241**(5): 98-100, 102, 108 passim.

Klein, R. J., C. Zeiss, et al. (2005). "Complement factor H polymorphism in age-related macular degeneration." Science **308**(5720): 385-389.

Koopman, M., S. Venderbosch, et al. (2009). "A review on the use of molecular markers of cytotoxic therapy for colorectal cancer, what have we learned?" Eur J Cancer **45**(11): 1935-1949.

Krampis, K., T. Booth, et al. (2012). "Cloud BioLinux: pre-configured and on-demand bioinformatics computing for the genomics community." BMC Bioinformatics **13**: 42.

Krek, A., D. Grun, et al. (2005). "Combinatorial microRNA target predictions." Nature genetics **37**(5): 495-500.

Kruglyak, L. (1997). "The use of a genetic map of biallelic markers in linkage studies." Nature genetics **17**(1): 21-24.

Kruglyak, L. (1999). "Prospects for whole-genome linkage disequilibrium mapping of common disease genes." Nature genetics **22**(2): 139-144.

Kulshreshtha, R., M. Ferracin, et al. (2007). "A microRNA signature of hypoxia." Molecular and cellular biology **27**(5): 1859-1867.

Lai, E., C. Bowman, et al. (2002). "Medical applications of haplotype-based SNP maps: learning to walk before we run." Nature genetics **32**(3): 353.

Lander, E. S. and N. J. Schork (1994). "Genetic dissection of complex traits." Science **265**(5181): 2037-2048.

Laroche-Clary, A., V. Le Morvan, et al. (2010). "Cytochrome P450 1B1 gene polymorphisms as predictors of anticancer drug activity: studies with in vitro models." Molecular cancer therapeutics **9**(12): 3315-3321.

Lee, P. H. and H. Shatkay (2008). "F-SNP: computationally predicted functional SNPs for disease association studies." Nucleic acids research **36**(Database issue): D820-824.

Lees, C. W., J. C. Barrett, et al. (2011). "New IBD genetics: common pathways with other diseases." Gut **60**(12): 1739-1753.

Lewis, B. P., C. B. Burge, et al. (2005). "Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets." Cell **120**(1): 15-20.

Li, B. and S. M. Leal (2008). "Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data." American journal of human genetics **83**(3): 311-321.

Li, M. J., P. Wang, et al. (2012). "GWASdb: a database for human genetic variants identified by genome-wide association studies." Nucleic acids research **40**(Database issue): D1047-1054.

Li, Q. (2006). "A Melanesian alpha-thalassemia mutation suggests a novel mechanism for regulating gene expression." Genome biology **7**(10): 238.

Liberzon, A., A. Subramanian, et al. (2011). "Molecular signatures database (MSigDB) 3.0." Bioinformatics **27**(12): 1739-1740.

Lin, Z. and R. B. Altman (2004). "Finding haplotype tagging SNPs by use of principal components analysis." American journal of human genetics **75**(5): 850-861.

Liu, D. J. and S. M. Leal (2010). "A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions." PLoS genetics **6**(10): e1001156.

Liu, X., X. Jian, et al. (2011). "dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions." Human mutation **32**(8): 894-899.

Madsen, B. E. and S. R. Browning (2009). "A groupwise association test for rare mutations using a weighted sum statistic." PLoS genetics **5**(2): e1000384.

Mailman, M. D., M. Feolo, et al. (2007). "The NCBI dbGaP database of genotypes and phenotypes." Nature genetics **39**(10): 1181-1186.

Makowsky, R., N. M. Pajewski, et al. (2011). "Beyond missing heritability: prediction of complex traits." PLoS genetics **7**(4): e1002051.

Manolio, T. A., F. S. Collins, et al. (2009). "Finding the missing heritability of complex diseases." Nature **461**(7265): 747-753.

Marchini, J., B. Howie, et al. (2007). "A new multipoint method for genome-wide association studies by imputation of genotypes." Nat Genet **39**(7): 906-913.

Marcuello, E., A. Altes, et al. (2006). "Methylenetetrahydrofolate reductase gene polymorphisms: genomic predictors of clinical response to fluoropyrimidine-based chemotherapy?" Cancer chemotherapy and pharmacology **57**(6): 835-840.

Mason, P. J., J. M. Bautista, et al. (2007). "G6PD deficiency: the genotype-phenotype association." Blood reviews **21**(5): 267-283.

Matys, V., E. Fricke, et al. (2003). "TRANSFAC: transcriptional regulation, from patterns to profiles." Nucleic acids research **31**(1): 374-378.

McCarthy, M. I., G. R. Abecasis, et al. (2008). "Genome-wide association studies for complex traits: consensus, uncertainty and challenges." Nature reviews. Genetics **9**(5): 356-369.

Medina, I., D. Montaner, et al. (2009). "Gene set-based analysis of polymorphisms: finding pathways or biological processes associated to traits in genome-wide association studies." Nucleic Acids Res **37**(Web Server issue): W340-344.

Meng, Z., D. V. Zaykin, et al. (2003). "Selection of genetic markers for association analyses, using linkage disequilibrium and haplotypes." American journal of human genetics **73**(1): 115-130.

Miller, W., K. Rosenbloom, et al. (2007). "28-way vertebrate alignment and conservation track in the UCSC Genome Browser." Genome research **17**(12): 1797-1808.

Mishra, P. J., R. Humeniuk, et al. (2007). "A miR-24 microRNA binding-site polymorphism in dihydrofolate reductase gene leads to methotrexate resistance." Proceedings of the National Academy of Sciences of the United States of America **104**(33): 13513-13518.

Montgomery, G. W. (2011). "Genome-wide association studies and genetic architecture of common human diseases." BMC proceedings **5 Suppl 4**: S16.

Moore, J. H. (2003). "The ubiquitous nature of epistasis in determining susceptibility to common human diseases." Hum Hered **56**(1-3): 73-82.

Moore, J. H., F. W. Asselbergs, et al. (2010). "Bioinformatics challenges for genome-wide association studies." Bioinformatics **26**(4): 445-455.

Morgenthaler, S. and W. G. Thilly (2007). "A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST)." Mutation research **615**(1-2): 28-56.

Nagy, E. and L. E. Maquat (1998). "A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance." Trends Biochem Sci **23**(6): 198-199.

Nam, D., J. Kim, et al. (2010). "GSA-SNP: a general approach for gene set analysis of polymorphisms." Nucleic acids research **38**(Web Server issue): W749-754.

Ng, P. C. and S. Henikoff (2003). "SIFT: Predicting amino acid changes that affect protein function." Nucleic acids research **31**(13): 3812-3814.

Olivier, M., R. Eeles, et al. (2002). "The IARC TP53 database: new online mutation analysis and recommendations to users." Human mutation **19**(6): 607-614.

Osier, M. V., K. H. Cheung, et al. (2001). "ALFRED: an allele frequency database for diverse populations and DNA polymorphisms--an update." Nucleic Acids Res **29**(1): 317-319.

Pan, Q., O. Shai, et al. (2008). "Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing." Nature genetics **40**(12): 1413-1415.

Pe'er, I., P. I. de Bakker, et al. (2006). "Evaluating and improving power in whole-genome association studies using fixed marker sets." Nat Genet **38**(6): 663-667.

Perkel, J. (2008). "SNP genotyping: six technologies that keyed a revolution." Nat Methods **5**(5): 447-453.

Pickrell, J., F. Clerget-Darpoux, et al. (2007). "Power of genome-wide association studies in the presence of interacting loci." Genetic epidemiology **31**(7): 748-762.

Pico, A. R., I. V. Smirnov, et al. (2009). "SNPLogic: an interactive single nucleotide polymorphism selection, annotation, and prioritization system." Nucleic Acids Res **37**(Database issue): D803-809.

Pollard, K. S., S. R. Salama, et al. (2006). "Forces shaping the fastest evolving regions in the human genome." PLoS genetics **2**(10): e168.

Prabhakar, S., J. P. Noonan, et al. (2006). "Accelerated evolution of conserved noncoding sequences in humans." Science **314**(5800): 786.

Pritchard, J. K. and N. J. Cox (2002). "The allelic architecture of human disease genes: common disease-common variant...or not?" Human molecular genetics **11**(20): 2417-2423.

Purcell, S., B. Neale, et al. (2007). "PLINK: a tool set for whole-genome association and population-based linkage analyses." American journal of human genetics **81**(3): 559-575.

Quevillon, E., V. Silventoinen, et al. (2005). "InterProScan: protein domains identifier." Nucleic acids research **33**(Web Server issue): W116-120.

Ramensky, V., P. Bork, et al. (2002). "Human non-synonymous SNPs: server and survey." Nucleic Acids Res **30**(17): 3894-3900.

Reich, D. E. and E. S. Lander (2001). "On the allelic spectrum of human disease." Trends in genetics : TIG **17**(9): 502-510.

Rhead, B., D. Karolchik, et al. (2010). "The UCSC Genome Browser database: update 2010." Nucleic acids research **38**(Database issue): D613-619.

Rieder, M. J., R. J. Livingston, et al. (2008). "The environmental genome project: reference polymorphisms for drug metabolism genes and genome-wide association studies." Drug Metab Rev **40**(2): 241-261.

Risch, N. and K. Merikangas (1996). "The future of genetic studies of complex human diseases." Science **273**(5281): 1516-1517.

Risch, N. J. (2000). "Searching for genetic determinants in the new millennium." Nature **405**(6788): 847-856.

Rodriguez, J., V. Boni, et al. (2011). "Association of RRM1 -37A>C polymorphism with clinical outcome in colorectal cancer patients treated with gemcitabine-based chemotherapy." European journal of cancer **47**(6): 839-847.

Ruwende, C., S. C. Khoo, et al. (1995). "Natural selection of hemi- and heterozygotes for G6PD deficiency in Africa by resistance to severe malaria." Nature **376**(6537): 246-249.

Ruzzo, A., F. Graziano, et al. (2007). "Pharmacogenetic profiling in patients with advanced colorectal cancer treated with first-line FOLFOX-4 chemotherapy." Journal of clinical oncology : official journal of the American Society of Clinical Oncology **25**(10): 1247-1254.

Sahashi, K., A. Masuda, et al. (2007). "In vitro and in silico analysis reveals an efficient algorithm to predict the splicing consequences of mutations at the 5' splice sites." Nucleic Acids Res **35**(18): 5995-6003.

Salgado, J., N. Zabalegui, et al. (2007). "Polymorphisms in the thymidylate synthase and dihydropyrimidine dehydrogenase genes predict response and toxicity to capecitabine-raltitrexed in colorectal cancer." Oncology reports **17**(2): 325-328.

Sauna, Z. E. and C. Kimchi-Sarfaty (2011). "Understanding the contribution of synonymous mutations to human disease." Nature reviews. Genetics **12**(10): 683-691.

Sayers, E. W., T. Barrett, et al. (2011). "Database resources of the National Center for Biotechnology Information." Nucleic Acids Res **39**(Database issue): D38-51.

SC, R. (2011). "Singapore Cancer Registry Interim Annual Registry Report." Trends in Cancer Incidence in Singapore 2005-2009.

Schwab, M., U. M. Zanger, et al. (2008). "Role of genetic and nongenetic factors for fluorouracil treatment-related severe toxicity: a prospective clinical trial by the German 5-FU Toxicity Study Group." J Clin Oncol **26**(13): 2131-2138.

Scott, L. J., K. L. Mohlke, et al. (2007). "A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants." Science **316**(5829): 1341-1345.

Shen, T. H., C. S. Carlson, et al. (2009). "SNPit: a federated data integration system for the purpose of functional SNP annotation." Computer methods and programs in biomedicine **95**(2): 181-189.

Sivakumaran, S., F. Agakov, et al. (2011). "Abundant pleiotropy in human complex diseases and traits." American journal of human genetics **89**(5): 607-618.

Smith, A. V., D. J. Thomas, et al. (2005). "Sequence features in regions of weak and strong linkage disequilibrium." Genome research **15**(11): 1519-1534.

Sobrero, A. F., C. Aschele, et al. (1997). "Fluorouracil in colorectal cancer--a tale of two drugs: implications for biochemical modulation." J Clin Oncol **15**(1): 368-381.

Sohaib, A. (2012). "RECIST rules." Cancer imaging : the official publication of the International Cancer Imaging Society **12**(2): 345-346.

Soranzo, N., G. L. Cavalleri, et al. (2004). "Identifying candidate causal variants responsible for altered activity of the ABCB1 multidrug resistance gene." Genome research **14**(7): 1333-1344.

Stein, L. D. (2010). "The case for cloud computing in genome informatics." Genome Biol **11**(5): 207.

Stephens, M., N. J. Smith, et al. (2001). "A new statistical method for haplotype reconstruction from population data." American journal of human genetics **68**(4): 978-989.

Stoyanovich, J. and I. Pe'er (2008). "MutaGeneSys: estimating individual disease susceptibility based on genome-wide SNP array data." Bioinformatics **24**(3): 440-442.

Stram, D. O., C. A. Haiman, et al. (2003). "Choosing haplotype-tagging SNPS based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the Multiethnic Cohort Study." Human heredity **55**(1): 27-36.

Tang, K., L. P. Wong, et al. (2004). "Genomic evidence for recent positive selection at the human MDR1 gene locus." Human molecular genetics **13**(8): 783-797.

Terwilliger, J. D. and T. Hiekkalinna (2006). "An utter refutation of the "Fundamental Theorem of the HapMap"." Eur J Hum Genet **14**(4): 426-437.

Thomas, P. D. and A. Kejariwal (2004). "Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: evolutionary evidence for differences in molecular effects." Proceedings of the National Academy of Sciences of the United States of America **101**(43): 15398-15403.

Thorn, C. F., T. E. Klein, et al. (2010). "Pharmacogenomics and bioinformatics: PharmGKB." Pharmacogenomics **11**(4): 501-505.

Tishkoff, S. A. and B. C. Verrelli (2003). "Role of evolutionary history on haplotype block structure in the human genome: implications for disease mapping." Curr Opin Genet Dev **13**(6): 569-575.

Town, M., J. M. Bautista, et al. (1992). "Both mutations in G6PD A- are necessary to produce the G6PD deficient phenotype." Human molecular genetics **1**(3): 171-174.

Tsuji, T., S. Hidaka, et al. (2003). "Polymorphism in the thymidylate synthase promoter enhancer region is not an efficacious marker for tumor sensitivity to 5-fluorouracil-based oral adjuvant chemotherapy in colorectal cancer." Clinical cancer research : an official journal of the American Association for Cancer Research **9**(10 Pt 1): 3700-3704.

Vignoli, M., S. Nobili, et al. (2011). "Thymidylate synthase expression and genotype have no major impact on the clinical outcome of colorectal cancer patients treated with 5-fluorouracil." Pharmacological research : the official journal of the Italian Pharmacological Society **64**(3): 242-248.

Vihinen, M., J. T. den Dunnen, et al. (2012). "Guidelines for establishing locus specific databases." Human mutation **33**(2): 298-305.

Wang, E. T., R. Sandberg, et al. (2008). "Alternative isoform regulation in human tissue transcriptomes." Nature **456**(7221): 470-476.

Wang, G. S. and T. A. Cooper (2007). "Splicing in disease: disruption of the splicing code and the decoding machinery." Nature reviews. Genetics **8**(10): 749-761.

Wang, J., M. Ronaghi, et al. (2011). "pfSNP: An integrated potentially functional SNP resource that facilitates hypotheses generation through knowledge syntheses." Human mutation **32**(1): 19-24.

Wang, K., M. Li, et al. (2010). "Analysing biological pathways in genome-wide association studies." Nature reviews. Genetics **11**(12): 843-854.

Wang, L. and Y. Xu (2003). "Haplotype inference by maximum parsimony." Bioinformatics **19**(14): 1773-1780.

Wang, W. Y., B. J. Barratt, et al. (2005). "Genome-wide association studies: theoretical and practical concerns." Nature reviews. Genetics **6**(2): 109-118.

Wang, Z., M. E. Rolish, et al. (2004). "Systematic identification and analysis of exonic splicing silencers." Cell **119**(6): 831-845.

Wang, Z., B. Wang, et al. (2005). "A functional polymorphism within the MRP1 gene locus identified through its genomic signature of positive selection." Human molecular genetics **14**(14): 2075-2087.

Wang, Z., J. Wang, et al. (2007). "Signatures of recent positive selection at the ATP-binding cassette drug transporter superfamily gene loci." Human molecular genetics **16**(11): 1367-1380.

Weale, M. E., C. Depondt, et al. (2003). "Selection and evaluation of tagging SNPs in the neuronal-sodium-channel gene SCN1A: implications for linkage-disequilibrium gene mapping." American journal of human genetics **73**(3): 551-565.

Webb, E., P. Broderick, et al. (2009). "A genome-wide scan of 10 000 gene-centric variants and colorectal cancer risk." Eur J Hum Genet **17**(11): 1507-1514.

Xie, X., J. Lu, et al. (2005). "Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals." Nature **434**(7031): 338-345.

Xu, H., S. G. Gregory, et al. (2005). "SNPselector: a web tool for selecting SNPs for genetic association studies." Bioinformatics **21**(22): 4181-4186.

Xuan, W., P. Wang, et al. (2007). "Medline search engine for finding genetic markers with biological significance." Bioinformatics **23**(18): 2477-2484.

Yeo, G. W., E. L. Nostrand, et al. (2007). "Discovery and analysis of evolutionarily conserved intronic splicing regulatory elements." PLoS genetics **3**(5): e85.

Yuan, H. Y., J. J. Chiou, et al. (2006). "FASTSNP: an always up-to-date and extendable service for SNP function analysis and prioritization." Nucleic acids research **34**(Web Server issue): W635-641.

Yue, P., E. Melamud, et al. (2006). "SNPs3D: candidate gene and SNP selection for association studies." BMC Bioinformatics **7**: 166.

Zeggini, E., W. Rayner, et al. (2005). "An evaluation of HapMap sample size and tagging SNP performance in large-scale empirical and simulated data sets." Nature genetics **37**(12): 1320-1322.

Zeng, Y., E. J. Wagner, et al. (2002). "Both natural and designed micro RNAs can inhibit the expression of cognate mRNAs when expressed in human cells." Mol Cell **9**(6): 1327-1333.

Zhang, H., Y. M. Li, et al. (2007). "DPYD*5 gene mutation contributes to the reduced DPYD enzyme activity and chemotherapeutic toxicity of 5-FU: results from genotyping study on 75 gastric carcinoma and colon carcinoma patients." Med Oncol **24**(2): 251-258.

Zhang, H., L. Liu, et al. (2007). "Guideline for data analysis of genomewide association studies." Cancer genomics & proteomics **4**(1): 27-34.

Zhang, K., S. Cui, et al. (2010). "i-GSEA4GWAS: a web server for identification of pathways/gene sets associated with traits by applying an improved gene set enrichment analysis to genome-wide association study." Nucleic Acids Res **38 Suppl**: W90-95.

Zhang, K., Z. Qin, et al. (2005). "HapBlock: haplotype block partitioning and tag SNP selection software using a set of dynamic programming algorithms." Bioinformatics **21**(1): 131-134.

Zhang, M. Q. (1998). "Statistical features of human exons and their flanking regions." Human molecular genetics **7**(5): 919-932.

Zhang, X. H. and L. A. Chasin (2004). "Computational definition of sequence motifs governing constitutive exon splicing." Genes Dev **18**(11): 1241-1250.

Zheng, Z. M. (2004). "Regulation of alternative RNA splicing by exon definition and exon sequences in viral and mammalian gene expression." J Biomed Sci **11**(3): 278-294.

Zhernakova, A., C. C. van Diemen, et al. (2009). "Detecting shared pathogenesis from the shared genetics of immune-related diseases." Nature reviews. Genetics **10**(1): 43-55.

Zondervan, K. T. and L. R. Cardon (2004). "The complex interplay among factors that influence allelic association." Nature reviews. Genetics **5**(2): 89-100.

Supplementary Table 1: The list of all the SNPs included on the GoldenGate array

#	rsNo	Gene	mRNA Location	AA Change
1	rs10046	CYP19A1	E10/3UTR /G19A	--
2	rs1010167	GSTM4	E/1/C-285G	--
3	rs1041983	NAT2	E/2/C282T	Y94Y
4	rs1042522	TP53	E/4/G215C	P72R
5	rs1042636	CASR	E/7/A2968G	R990G
6	rs1047840	EXO1	E/12/G1765A	E589K
7	rs1048572	SMARCE1	I/4/A1465G	--
8	rs10493895	DPYD	I/3/A22444C	--
9	rs10502975	DCC	I/20/C1225G	--
10	rs10508022	ABCC4	I/19/T11361C	--
11	rs1051266	SLC19A1	E/2/T80C	H27R
12	rs10513348	HLTF	I/23/G975A	--
13	rs10515959	DCC	I/15/G-19191T	--
14	rs1052555	ERCC2	E/22/G2133A	D711D
15	rs1052748	PLD2	E/17/C1730T	T577I
16	rs1056836	CYP1B1	E/3/C1294G	--
17	rs1057910	CYP2C9	E/7/A1075C	I359L
18	rs1059234	CDKN1A	E3/3UTR /C20T	--
19	rs10875053	DPYD	I/20/C21659G	--
20	rs10903118	RUNX3	5UR//T-3376C	--
21	rs10934683	UMPS	I/1/C-557T	--
22	rs10947623	CDKN1A	5UR//G-3650A	--
23	rs11202592	PTEN	E/1/C-9G	--
24	rs1143627	IL1B	5UR//G-30A	--
25	rs1145231	PMS1	E/9/T1181C	M394T
26	rs11545078	GGH	E/5/G452A	T151I
27	rs11549467	HIF1A	E/12/G1762A	A588T
28	rs11569017	EGF	E/15/A2351T	D784V
29	rs11615	ERCC1	E/3/A354G	N118N
30	rs11664579	WDR7	I/13/A323G	--
31	rs12052058	SMARCA4	I/30/G7289T	--
32	rs12052201	SMARCA4	I/30/G6860T	--
33	rs1208	NAT2	E/2/G803A	R268K
34	rs1234213	PTEN	I/3/G-1482A	--
35	rs12458289	BCL2	3DR//G6601T	--
36	rs12917	MGMT	E/5/C250T	L84F
37	rs12954274	DCC	I/14/C-209T	--
38	rs1314	UPB1	E10/3UTR /T690G	--
39	rs131794	TYMP	5UR//A-3296C	--
40	rs13181	ERCC2	E/23/T2251G	K751Q
41	rs1367634	DCC	I/14/T-1415G	--
42	rs1381547	BCL2	3DR//T27848C	--
43	rs1402001	ABCC5	3DR//A58790G	--
44	rs16260	CDH1	5UR//C-160A	--
45	rs1650697	DHFR	E/1/A-473G	--
46	rs1657396	WDR7	E27/3UTR /T1969C	--

47	rs1657415	WDR7	I/25/G19951A	--
48	rs16950632	ABCC4	I/20/G6290A	--
49	rs17189467	ABCC4	I/4/A-1931T	--
50	rs17217716	MSH2	E/1/C23T	T8M
51	rs17217772	MSH2	E/3/A380G	N127S
52	rs17224367	MSH2	E/7/C1168T	L390F
53	rs17630758	SMARCB1	I/3/G667A	--
54	rs17655	ERCC5	E/15/G3310C	D1104H
55	rs17756073	BCL2	3DR//A175404G	--
56	rs1787479	WDR7	I/25/C-24771G	--
57	rs1799782	XRCC1	E/6/G580A	R194W
58	rs1799794	XRCC3	E/2/T-315C	--
59	rs1799930	NAT2	E/2/G590A	R197Q
60	rs1800371	TP53	E/4/G139A	P47S
61	rs1800566	NQO1	E/5/G445A	P149S
62	rs1800734	MLH1	5UR//G-32A	--
63	rs1800975	XPA	E/1/T-4C	--
64	rs1801126	OGG1	E/1/G-18T	--
65	rs1801131	MTHFR	E/8/T1286G	E429A
66	rs1801133	MTHFR	E/5/G665A	A222V
67	rs1801158	DPYD	E/13/C1601T	S534N
68	rs1801166	APC	E/16/G3949C	E1317Q
69	rs1801265	DPYD	E/2/G85A	C29R
70	rs1801268	DPYD	E/23/C2983A	V995F
71	rs1801270	CDKN1A	E/2/C93A	S31R
72	rs1804197	APC	E16/3UTR /C86A	--
73	rs1810132	ERBB2	I/4/C-61T	--
74	rs1811086	TK1	I/4/G-1323A	--
75	rs1860460	PMS2	5UR//G-4548A	--
76	rs1922242	ABCB1	I/17/A-76T	--
77	rs1926657	ABCC4	I/4/T11908C	--
78	rs1950902	MTHFD1	E/6/A401G	R134K
79	rs1979277	SHMT1	E/11/G1303A	L435F
80	rs2020873	MLH1	E/19/C2152T	H718Y
81	rs2020911	MSH6	I/5/A14T	--
82	rs2020912	MSH6	E/4/T2633C	V878A
83	rs2066109	SMARCA2	I/23/T-1115C	--
84	rs2066518	SMARCAL1	E/5/G1129C	G377R
85	rs2072671	CDA	E/1/A79C	K27Q
86	rs2073390	SMARCB1	I/1/G-373A	--
87	rs2078486	TP53	I/1/G-3143A	--
88	rs2083020	WDR7	I/20/A20363G	--
89	rs2227306	IL8	I/1/C-204T	--
90	rs2227311	RB1	I/17/A31453G	--
91	rs2228001	XPC	E/16/G2818T	L940M
92	rs2228527	ERCC6	E/18/T3637C	R1213G
93	rs2229080	DCC	E/3/C601G	R201G
94	rs2229992	APC	E/12/T1458C	Y486Y
95	rs2229995	APC	E/16/G7504A	G2502S
96	rs2231142	ABCG2	E/5/G421T	Q141K
97	rs2233919	SMUG1	E/3/G7A	Q3*

98	rs2234978	FAS	E/6/T579C	T193T
99	rs2235035	ABCB1	I/14/G81A	--
100	rs2236722	CYP19A1	E/2/A115G	W39R
101	rs2243115	IL12A	5UR//T-348G	--
102	rs2244500	TYMS	I/2/A-1141G	--
103	rs2250889	MMP9	E/10/G1721C	R574P
104	rs2270951	DCC	I/21/C127T	--
105	rs2274976	MTHFR	E/12/C1781T	R594Q
106	rs2277448	ATP7B	E/1/G-75T	--
107	rs2278495	WDR7	I/25/G-163A	--
108	rs2279115	BCL2	5UR//G-223T	--
109	rs2289310	DLG5	E/23/G4442T	P1481Q
110	rs2299939	PTEN	I/2/C3284A	--
111	rs2302323	PLD2	E/25/G2702A	G901D
112	rs2303428	MSH2	I/12/T-6C	--
113	rs2308321	MGMT	E/7/A427G	S143P
114	rs2308327	MGMT	E/7/A533G	G178E
115	rs2355164	UPP2	I/2/C27144T	--
116	rs2376311	SMARCA2	I/25/G666A	--
117	rs238406	ERCC2	E/6/T468G	R156R
118	rs240993	REV3L	I/18/T-720C	--
119	rs243865	MMP2	5UR//C-1281T	--
120	rs2517954	ERBB2	5UR//T-12703C	--
121	rs2517955	ERBB2	5UR//C-12572T	--
122	rs2517956	ERBB2	5UR//G-12394A	--
123	rs25487	XRCC1	E/10/T1196C	R399Q
124	rs25489	XRCC1	E/9/C839T	R280H
125	rs25496	XRCC1	E/3/A215G	V72A
126	rs2660744	PPAT	E/11/G1462A	Q488*
127	rs2665797	SMARCD2	5UR//G-2133C	--
128	rs2735343	PTEN	I/5/G-6446C	--
129	rs28366003	MT2A	5UR//A-19G	--
130	rs28399504	CYP2C19	E/1/A1G	M1V
131	rs2853741	TYMS	5UR//T-298C	--
132	rs28756987	MLH3	E/2/G1939A	R647C
133	rs3092856	ATM	E/2/C94T	H32Y
134	rs3136367	MSH6	I/8/C-40G	--
135	rs3136788	POLB	I/11/A-2634G	--
136	rs3136797	POLB	E/12/C725G	P242R
137	rs3176734	XPA	I/5/G-2522A	--
138	rs3212931	ERCC1	5UR//G-137T	--
139	rs3218599	REV3L	E/13/C5434G	D1812H
140	rs3219090	PARP1	I/13/T118C	--
141	rs34035085	UPB1	E/2/C254A	A85E
142	rs34136999	MSH2	E/5/C815T	A272V
143	rs34213726	MLH1	E/12/A1327C	K443Q
144	rs34330	CDKN1B	E/1/T-79C	--
145	rs34374438	MSH6	E/4/A2561T	K854M
146	rs34986638	MSH2	E/14/G2422T	E808*
147	rs35001569	MLH1	E/16/A1852G	K618E
148	rs35032294	MLH1	5UR//C-208G	--

149	rs35045067	MLH1	E/17/A1937G	Y646C
150	rs351855	FGFR4	E/9/G1162A	G388R
151	rs35917308	PTEN	E/4/C234T	P78P
152	rs3731249	CDKN2A	E/2/C442T	A148T
153	rs3732183	MSH2	I/10/G12A	--
154	rs3738888	BARD1	E/10/G1972A	R658C
155	rs3740066	ABCC2	E/28/C3972T	I1324I
156	rs3744951	BCL2	3DR//T59207C	--
157	rs3749441	ABCC5	3DR//C40735G	--
158	rs3755319	UGT1A1	5UR//A-1336C	--
159	rs3758149	GGH	5UR//G-341A	--
160	rs3758581	CYP2C19	E/7/G991A	I331V
161	rs3764496	DCC	I/15/T-61C	--
162	rs3789243	ABCB1	I/4/A4196G	--
163	rs3790674	UCK2	I/5/A1103G	--
164	rs3792582	ABCC5	3DR//A6939G	--
165	rs3793784	ERCC6	5UR//G-466C	--
166	rs3794917	DCC	I/16/G-1683T	--
167	rs3794922	DCC	I/16/G1739A	--
168	rs3803185	ARL11	E/2/T442C	C148R
169	rs3805112	ABCC5	3DR//C55879T	--
170	rs3808607	CYP7A1	5UR//G-202T	--
171	rs3817672	TFRC	E/4/C424T	S142G
172	rs3918290	DPYD	I/14/C1T	--
173	rs4105144	CYP2A6	5UR//T-2283C	--
174	rs41294980	MLH1	E/12/G1217A	S406N
175	rs41295278	MSH6	E/9/A3961G	R1321G
176	rs4150001	EXO1	E/14/G2276A	G759E
177	rs4150521	ERCC3	E/14/G2111A	S704L
178	rs41549115	ERCC2	E/2/G72A	Y24Y
179	rs42427	APC	E/16/G5034A	G1678G
180	rs4246514	DPYD	I/2/C-3979G	--
181	rs4253038	ERCC6	I/3/A-408G	--
182	rs4253208	ERCC6	E/18/G3284C	P1095R
183	rs4372296	DPYD	I/2/C-26740A	--
184	rs4379706	DPYD	I/2/C26441T	--
185	rs4421623	DPYD	I/2/T-22533G	--
186	rs4516035	VDR	5UR//T-1011C	--
187	rs459552	APC	E/16/T5465A	V1822D
188	rs4645878	BAX	5UR//A-178G	--
189	rs465899	APC	E/16/G5880A	P1960P
190	rs4673	CYBA	E/4/A214G	Y72H
191	rs4773866	ABCC4	I/1/C24169T	--
192	rs4775936	CYP19A1	I/1/C-875T	--
193	rs4986783	NAT1	E/3/T640G	S214A
194	rs4986909	CYP3A4	E/11/G1247A	P416L
195	rs4986913	CYP3A4	E/12/G1399A	P467S
196	rs4987188	MSH2	E/6/G965A	G322D
197	rs527912	CDA	I/2/G2751A	--
198	rs532545	CDA	5UR//C-271T	--
199	rs56131651	ABCC2	E/7/-842-	--

200	rs6413420	CYP2E1	5UR//G-37T	--
201	rs6413438	CYP2C19	E/5/C680T	P227L
202	rs6566846	WDR7	I/21/T-2912G	--
203	rs6593634	DPYD	I/20/A23413C	--
204	rs6665429	DPYD	I/14/C-5602T	--
205	rs671	ALDH2	E/12/G1510A	E504K
206	rs6759948	UPP2	I/2/A17195G	--
207	rs689466	PTGS2	5UR//T-1194C	--
208	rs717620	ABCC2	E/1/C-24T	--
209	rs7234941	BCL2	3DR//C62336T	--
210	rs7336051	ABCC4	I/1/C-26424T	--
211	rs7439366	UGT2B7	E/2/T802C	H268Y
212	rs762551	CYP1A2	I/1/C-154A	--
213	rs7646621	ABCC5	3DR//G60676T	--
214	rs7993619	ABCC4	I/19/A698C	--
215	rs7996205	ABCC4	3DR//G66320A	--
216	rs8082807	DCC	I/25/A-2265C	--
217	rs8094838	WDR7	I/21/A532G	--
218	rs8187699	ABCC2	E/27/A3817G	T1273A
219	rs8187710	ABCC2	E/32/G4544A	C1515Y
220	rs861539	XRCC3	E/7/G722A	T241M
221	rs869224	DCC	I/23/A766G	--
222	rs886205	ALDH2	E/1/A-360G	--
223	rs887829	UGT1A1	5UR//C-348T	--
224	rs945881	DPYD	I/13/G-29513A	--
225	rs955850	DPYD	I/14/T-9744C	--
226	rs9651118	MTHFR	I/2/T724C	--
227	rs965943	DCC	I/14/A-8653G	--
228	rs9807370	DCC	I/21/G719A	--
229	rs9946253	WDR7	I/18/A-2783C	--
230	rs9950970	DCC	I/14/T-2275C	--
231	rs10041507	SLCO6A1	I/3/T220C	--
232	rs10041525	SLCO6A1	I/3/T142C	--
233	rs1005793	BCL2	3DR//A95027G	--
234	rs10062613	SLCO6A1	I/11/C702T	--
235	rs10073892	SLCO6A1	I/10/T-21C	--
236	rs10081796	SLC31A1	I/1/C16220G	--
237	rs1011019	ERCC3	I/9/A463G	--
238	rs10124350	SMARCA2	I/23/A-2552G	--
239	rs10162199	ABCC4	I/11/C878T	--
240	rs10186677	UPP2	E/7/A462C	A154A
241	rs10204779	UPP2	I/2/G7261A	--
242	rs10276036	ABCB1	I/10/C-44T	--
243	rs1042040	PPAT	E11/3UTR /C1220G	--
244	rs10503079	BCL2	3DR//C72821T	--
245	rs10505058	DPYS	I/7/A4947G	--
246	rs10513190	SLC31A1	I/1/C9965T	--
247	rs1062472	ATP7A	E23/3UTR /T1960C	--
248	rs10747486	DPYD	I/3/A39932G	--
249	rs10757122	SMARCA2	I/1/T-4348C	--
250	rs10757185	SMARCA2	I/26/G1836T	--

251	rs10757188	SMARCA2	I/26/T-939C	--
252	rs10757211	SMARCA2	I/27/T12883G	--
253	rs10783069	DPYD	I/10/G-6470C	--
254	rs10811481	SMARCA2	I/27/C203T	--
255	rs10811504	SMARCA2	I/27/T5807C	--
256	rs10811515	SMARCA2	I/27/T8841C	--
257	rs10875098	DPYD	I/10/T-2194C	--
258	rs10964466	SMARCA2	I/1/A-45G	--
259	rs10964500	SMARCA2	I/3/A1359C	--
260	rs10964921	SMARCA2	I/27/G8243A	--
261	rs10965088	SMARCA2	I/28/G-9017A	--
262	rs10981699	SLC31A1	I/1/C11943T	--
263	rs11097408	SMARCAD1	I/6/T3505C	--
264	rs11202600	PTEN	I/2/G-2437C	--
265	rs11202607	PTEN	E9/3UTR /C2185T	--
266	rs11241185	APC	I/4/G3971A	--
267	rs1125205	SMARCA2	I/4/A-2508G	--
268	rs1132776	ABCC5	3DR//A5139G	--
269	rs11579252	CDA	I/2/A2681G	--
270	rs11665624	WDR7	I/25/G-27842A	--
271	rs11714840	SMARCC1	I/10/G800C	--
272	rs11746217	SLCO6A1	I/4/A-9G	--
273	rs118202	REV3L	I/21/G-1558T	--
274	rs11864810	CES1	5UR//T-2652G	--
275	rs11872907	WDR7	I/13/C4188T	--
276	rs11876256	WDR7	I/14/T9751A	--
277	rs11877604	WDR7	I/12/C-2229T	--
278	rs11879293	SMARCA4	I/1/G760A	--
279	rs1202168	ABCB1	I/7/G139A	--
280	rs12045999	DPYD	I/14/C7624T	--
281	rs12121543	MTHFR	I/7/C-76A	--
282	rs1234224	PTEN	I/2/A-9974G	--
283	rs1234225	PTEN	I/2/C-11701T	--
284	rs12345640	SMARCA2	I/27/T13816C	--
285	rs12380390	SMARCA2	I/5/A-1218G	--
286	rs12467193	UPP2	5UR//C-3548T	--
287	rs12511433	SMARCAD1	I/6/G268A	--
288	rs12571445	ERCC6	I/5/A-8125G	--
289	rs12610607	SMARCA4	I/1/A-9803G	--
290	rs12634398	ABCC5	I/2/A-1234G	--
291	rs12677953	GGH	I/7/C-100A	--
292	rs12921111	ERCC4	5UR//G-4199A	--
293	rs13090196	HLTF	I/23/C902G	--
294	rs13190449	SLCO6A1	E/1/G80A	A27V
295	rs1333717	DPYD	I/3/G29092A	--
296	rs13425206	MSH2	I/6/G-2037T	--
297	rs1370216	WDR7	I/26/G736A	--
298	rs1383596	BCL2	3DR//C76325A	--
299	rs1437069	WDR7	I/25/T-14715C	--
300	rs1437135	NQO1	I/1/A2508G	--
301	rs1452057	SLCO6A1	I/6/A4781G	--

302	rs1476413	MTHFR	I/10/C35T	--
303	rs1520025	PPAT	I/1/G1024A	--
304	rs1542005	WDR7	I/20/G8005A	--
305	rs1562961	SLCO6A1	I/10/A-2493G	--
306	rs1617844	ABCC4	I/8/G651A	--
307	rs1657410	WDR7	I/25/A-20153G	--
308	rs1657421	WDR7	I/25/C29064T	--
309	rs16851444	UCK2	I/1/T10643G	--
310	rs16861365	HLTF	E/15/T1506C	V502V
311	rs16930959	SLC31A1	I/1/A-6936G	--
312	rs16937154	SMARCA2	I/9/T-3581A	--
313	rs17017854	SMARCA5	I/14/G865C	--
314	rs17070841	BCL2	3DR//T87993C	--
315	rs17070861	BCL2	3DR//T78707G	--
316	rs17086746	PPAT	E11/3UTR /C1519T	--
317	rs17086758	PPAT	I/1/T-6672C	--
318	rs17107001	PTEN	I/3/G1195T	--
319	rs17139614	ATP7A	E23/3UTR /G2241C	--
320	rs17139617	ATP7A	I/12/C-882A	--
321	rs17189481	ABCC4	I/4/C4542T	--
322	rs17300865	ABCC4	I/4/G4340A	--
323	rs17431184	PTEN	I/7/T-400C	--
324	rs17471125	DPYD	I/21/T-5102C	--
325	rs1751021	ABCC4	I/8/C-2406T	--
326	rs1751029	ABCC4	I/8/A728G	--
327	rs17757541	BCL2	3DR//C105501G	--
328	rs17775180	ERCC6	I/8/C2846A	--
329	rs17785248	SMARCC1	I/24/A3765G	--
330	rs1787462	WDR7	I/25/A-28222G	--
331	rs1787463	WDR7	I/25/A-28121G	--
332	rs1787468	WDR7	I/25/C-27977T	--
333	rs1787475	WDR7	E27/3UTR /A1061G	--
334	rs1800668	GPX1	E/1/G-46A	--
335	rs1807999	BCL2	3DR//C83015G	--
336	rs184026	SMARCAL1	I/14/C642T	--
337	rs1871446	CDC2	E6/3UTR/A30G	--
338	rs1886261	SMARCA2	I/28/A-7981G	--
339	rs1893806	BCL2	3DR//C223A	--
340	rs1901512	SLCO6A1	I/12/T517C	--
341	rs1901521	SLCO6A1	I/9/G1986T	--
342	rs1901522	SLCO6A1	I/9/T5068G	--
343	rs1914	APC	I/8/A-6748T	--
344	rs1931063	DPYD	I/3/C30060T	--
345	rs1962292	SMARCA2	I/24/A164G	--
346	rs1962293	SMARCA2	I/24/T327C	--
347	rs2001776	SMARCA2	I/26/G1290A	--
348	rs2002042	ABCC2	I/19/C-2133T	--
349	rs2065943	DPYD	I/8/T-10704C	--
350	rs2066462	MTHFR	E/7/G1056A	S352S
351	rs2139512	PPAT	I/1/G-2737C	--
352	rs2214102	ABCB1	E/3/T-1C	--

353	rs2233913	SLC31A1	5UR//T-651C	--
354	rs2235046	ABCB1	I/17/T73C	--
355	rs2235048	ABCB1	I/27/G80A	--
356	rs2235074	ABCB1	I/4/G36A	--
357	rs2270860	SLC22A7	E/7/C1269T	S423S
358	rs2271937	ABCC5	3DR//G18581A	--
359	rs2273697	ABCC2	E/10/G1249A	V417I
360	rs2280392	ABCC5	3DR//A35799G	--
361	rs2281793	ERCC6	I/5/C4537T	--
362	rs2281794	ERCC6	I/5/T3659C	--
363	rs2282011	UCK1	I/2/G-444A	--
364	rs2293001	ABCC5	3DR//C1330T	--
365	rs2299941	PTEN	I/5/A-7156G	--
366	rs2306802	SMARCAD1	E/10/A1479G	Q493Q
367	rs232054	RRM1	I/2/C116G	--
368	rs2343614	REV3L	I/9/G-2511A	--
369	rs240954	REV3L	I/21/G-3009A	--
370	rs240955	REV3L	I/21/G-3830A	--
371	rs240991	REV3L	I/20/C2247G	--
372	rs240995	REV3L	I/18/T1704C	--
373	rs2431238	APC	I/6/T-3774C	--
374	rs2464803	APC	I/7/A3561G	--
375	rs2464805	APC	I/2/C-230A	--
376	rs2520464	ABCB1	I/5/C-1547T	--
377	rs2545162	APC	I/11/G743A	--
378	rs2546106	APC	I/8/C-5931A	--
379	rs2546108	APC	I/11/C-1336A	--
380	rs2546110	APC	I/11/A-392G	--
381	rs2576415	WDR7	I/19/T-30795A	--
382	rs2622604	ABCG2	I/1/T614C	--
383	rs2661681	TK1	I/5/C-4T	--
384	rs2707763	APC	I/4/T-2461C	--
385	rs2723877	TDG	I/1/T-1570C	--
386	rs2727280	SMARCD2	I/1/A1737G	--
387	rs2735691	RRM1	I/1/C1448G	--
388	rs2786498	DPYD	I/8/G-6425A	--
389	rs284564	SMARCAL1	I/4/T1088A	--
390	rs2848968	WDR7	I/25/G-20982A	--
391	rs2854510	XRCC1	I/2/A4850G	--
392	rs2909787	APC	I/14/G-2518C	--
393	rs2952615	APC	I/8/G1808C	--
394	rs3136748	POLB	I/5/C1214T	--
395	rs3176658	XPA	I/2/A-1942G	--
396	rs3176750	XPA	E/6/G754C	L252V
397	rs3212106	XRCC3	I/6/T-292C	--
398	rs3212121	XRCC3	E10/3UTR /T613C	--
399	rs3213255	XRCC1	I/2/G1555A	--
400	rs3213356	XRCC1	I/4/C-404T	--
401	rs351771	APC	E/14/G1635A	A545A
402	rs3740065	ABCC2	I/29/A154G	--
403	rs3745041	WDR7	I/13/C-199T	--

404	rs3749443	ABCC5	3DR//C62218T	--
405	rs3750417	SMARCA2	I/5/C-188T	--
406	rs3750748	ERCC6	I/5/C2665T	--
407	rs3750749	ERCC6	I/12/A33G	--
408	rs3758395	ABCC2	I/26/T154C	--
409	rs3774341	MLH1	I/3/A-659C	--
410	rs3780128	GGH	I/3/T-1214C	--
411	rs3782964	ABCC4	I/4/C-430T	--
412	rs3785911	ABCC3	I/30/A-1022C	--
413	rs3786362	TYMS	E/3/A381G	I127I
414	rs3789244	ABCB1	I/10/G1228T	--
415	rs3790661	UCK2	I/4/T2927C	--
416	rs3792584	ABCC5	3DR//T17631C	--
417	rs3793499	SMARCA2	I/27/C8053G	--
418	rs3805109	ABCC5	3DR//G11276A	--
419	rs3808830	UCK1	5UR//C-1778T	--
420	rs3810915	SLC31A1	5UR//A-2096G	--
421	rs3810944	ERCC6	I/19/A236G	--
422	rs3817403	ABCC5	3DR//A18396G	--
423	rs3829070	SMARCA2	I/26/T-80C	--
424	rs3829072	SMARCA2	I/27/G3859A	--
425	rs3829073	SMARCA2	I/27/T3946C	--
426	rs3829300	TDG	I/2/A12G	--
427	rs390092	APC	I/8/T-4995G	--
428	rs4135082	TDG	I/1/T-1409C	--
429	rs4135087	TDG	I/2/C387T	--
430	rs4135094	TDG	I/2/T-147C	--
431	rs4148328	UGT1A1	I/4/C574T	--
432	rs4148432	ABCC4	I/1/C-13075T	--
433	rs4148469	ABCC4	I/4/A-323G	--
434	rs4148485	ABCC4	I/9/G-2464A	--
435	rs4148568	ABCC5	I/2/C9031T	--
436	rs4148575	ABCC5	E7/3UTR /A447G	--
437	rs4148580	ABCC5	3DR//T16358C	--
438	rs4148584	ABCC5	3DR//A24040C	--
439	rs4148594	ABCC5	3DR//T61157C	--
440	rs4148738	ABCB1	I/21/C-2236T	--
441	rs4150402	ERCC3	I/3/T52C	--
442	rs4150456	ERCC3	E/9/C1485T	E495E
443	rs4150471	ERCC3	I/10/A2210G	--
444	rs4253055	ERCC6	I/5/A3752G	--
445	rs4253060	ERCC6	I/5/C4292T	--
446	rs4253095	ERCC6	I/5/A-1137G	--
447	rs4253101	ERCC6	I/6/T871G	--
448	rs4253121	ERCC6	I/7/G1597A	--
449	rs4253126	ERCC6	I/7/G-892T	--
450	rs4253138	ERCC6	I/8/C828T	--
451	rs4253165	ERCC6	I/9/C-128T	--
452	rs4253166	ERCC6	I/10/T716G	--
453	rs4303338	ABCC4	I/4/C10763G	--
454	rs4373572	SMARCA2	I/27/A632G	--

455	rs4537601	DPYD	I/3/G-27495C	--
456	rs453776	REV3L	I/21/T1665C	--
457	rs454886	APC	I/8/A-5075G	--
458	rs455645	REV3L	E/13/C4485T	S1495S
459	rs455650	REV3L	I/13/C2276T	--
460	rs460594	REV3L	I/20/C616G	--
461	rs461646	REV3L	E/13/G3301A	L1101L
462	rs4641140	SLC31A1	I/1/A1629G	--
463	rs4664236	UPP2	I/2/T17318G	--
464	rs4664920	UPP2	I/2/G7182A	--
465	rs4741637	SMARCA2	I/5/C3248G	--
466	rs4741638	SMARCA2	I/5/A3360T	--
467	rs4741640	SMARCA2	I/6/A-780C	--
468	rs4773854	ABCC4	I/4/G7454A	--
469	rs4940574	BCL2	3DR//G147343C	--
470	rs4941189	BCL2	3DR//C77260T	--
471	rs4970728	DPYD	I/4/G3787C	--
472	rs4987770	BCL2	3DR//C80832T	--
473	rs4987802	BCL2	3DR//C140464G	--
474	rs562	ABCC5	3DR//T63696C	--
475	rs614080	GSTP1	5UR//G-3998A	--
476	rs628959	DPYD	I/18/C31955T	--
477	rs6437129	UPP2	I/2/A-7469G	--
478	rs6475506	SMARCA2	I/27/G2301A	--
479	rs6475507	SMARCA2	I/27/A3532T	--
480	rs6475511	SMARCA2	I/27/T4991G	--
481	rs6475514	SMARCA2	I/27/G5170C	--
482	rs6475520	SMARCA2	I/27/A6891G	--
483	rs6475524	SMARCA2	I/27/A7597G	--
484	rs6567334	BCL2	3DR//A67807C	--
485	rs6683957	DPYD	I/4/A5229G	--
486	rs6768699	SMARCC1	I/8/A-450G	--
487	rs6790814	ABCC5	3DR//C44735G	--
488	rs6806313	ABCC5	3DR//G59389C	--
489	rs6854326	PPAT	I/1/T-2027C	--
490	rs6873738	SLCO6A1	I/12/A7020C	--
491	rs6877722	SLCO6A1	I/12/G-738T	--
492	rs6884141	SLCO6A1	E/2/G393A	G131G
493	rs6915753	REV3L	I/1/T-4004C	--
494	rs6922226	REV3L	I/4/A2189G	--
495	rs6935759	REV3L	I/1/G25233A	--
496	rs6942129	REV3L	I/1/T-25945C	--
497	rs6949448	ABCB1	I/26/T2733C	--
498	rs6961665	ABCB1	I/10/C-1264A	--
499	rs699937	POLH	5UR//G-3018A	--
500	rs699937	XPO5	I/2/G291A	--
501	rs7021817	SMARCA2	I/27/A16224G	--
502	rs7033529	SMARCA2	I/27/A1164G	--
503	rs7035071	SMARCA2	I/14/T1474C	--
504	rs7035608	SMARCA2	I/9/C3521T	--
505	rs7035898	SMARCA2	I/9/C3809T	--

506	rs7040174	SMARCA2	I/24/A-1116C	--
507	rs7040968	SMARCA2	I/27/G-12720A	--
508	rs7048496	SMARCA2	I/27/C1294A	--
509	rs7048976	SMARCA2	I/28/G-8275T	--
510	rs7073830	ERCC6	I/7/G-1932A	--
511	rs719717	ABCC3	I/1/C-3695G	--
512	rs720159	UPP2	5UR//G-2886A	--
513	rs7243516	WDR7	I/12/C-7239T	--
514	rs726511	BCL2	3DR//A72295T	--
515	rs731420	XRCC1	I/4/G-137A	--
516	rs748766	MLH1	I/14/T-885C	--
517	rs7506986	WDR7	I/14/G5730A	--
518	rs7618758	SMARCC1	I/6/C-258T	--
519	rs7654543	PPAT	I/1/A-8415C	--
520	rs7751272	REV3L	I/1/A-2596G	--
521	rs7764543	REV3L	I/1/T-6644C	--
522	rs7766610	REV3L	I/9/C1147A	--
523	rs7847382	SMARCA2	I/27/T4325G	--
524	rs7851702	SMARCA2	I/23/A-2352G	--
525	rs7853086	SMARCA2	I/26/A-299T	--
526	rs7858597	SMARCA2	I/27/C4461G	--
527	rs7862805	SMARCA2	I/27/T13414A	--
528	rs7864080	SMARCA2	I/11/A-52T	--
529	rs7864452	SLC31A1	I/1/C14630T	--
530	rs7869436	SMARCA2	I/27/A3378G	--
531	rs7920256	ERCC6	I/7/A-2848G	--
532	rs7927381	GSTP1	5UR//T-4542C	--
533	rs8001444	ABCC4	I/1/T1058C	--
534	rs8060569	MT1A	5UR//T-1437C	--
535	rs8073898	MPO	E12/3UTR /A561T	--
536	rs8086404	BCL2	3DR//G68063C	--
537	rs8093133	WDR7	I/12/T-444G	--
538	rs8096380	BCL2	3DR//G64706A	--
539	rs8187692	ABCC2	E/25/G3542T	R1181L
540	rs8192729	CYP2A6	I/7/C203T	--
541	rs861531	XRCC3	I/5/C533A	--
542	rs869951	ABCC4	5UR//C-522G	--
543	rs899497	ABCC4	I/4/A-108G	--
544	rs9400476	REV3L	I/1/C19578T	--
545	rs9472084	POLH	I/3/A-1658G	--
546	rs9487639	REV3L	I/1/G-10319A	--
547	rs9487645	REV3L	I/1/T-26593C	--
548	rs9524856	ABCC4	I/4/T588C	--
549	rs9524873	ABCC4	I/1/G26109A	--
550	rs9524885	ABCC4	I/1/T17906C	--
551	rs9561811	ABCC4	I/4/C3193T	--
552	rs956275	PPAT	I/7/C90T	--
553	rs9568682	ATP7B	I/1/C-15456G	--
554	rs9590228	ABCC4	I/1/C-23383T	--
555	rs9634642	ABCC4	I/3/T-1583C	--
556	rs971667	ERCC6	I/7/G3490A	--

557	rs9812777	ABCC5	3DR//C38368T	--
558	rs981988	SLCO6A1	I/1/T-3843C	--
559	rs9864549	HLTF	I/18/A-1738C	--
560	rs991791	WDR7	I/25/T-22204C	--
561	rs9936741	MT1A	5UR//T-4792C	--
562	rs9972996	BCL2	3DR//A63297G	--
563	rs2695784	SMARCC2	5UR//T-3166C	--
564	rs291592	DPYD	E23/3UTR /C768T	--
565	rs2930788	SMARCC2	5UR//T-2971C	--
566	rs3136717	POLB	I/1/C-89T	--
567	rs4801043	WDR7	5UR//T-1512C	--
568	rs2071487	GSTM1	E/7/T462C	T154T
569	rs4846049	MTHFR	E12/3UTR /T372G	--
570	rs733590	CDKN1A	5UR//T-1283C	--
571	rs10075210	ATOX1	5UR//G-771A	--
572	rs1008767	MT2A	5UR//T-2567C	--
573	rs10118903	FPGS	5UR//C-106T	--
574	rs10120113	SMARCA2	5UR//A-2574G	--
575	rs10121175	SMARCA2	5UR//G-4580A	--
576	rs1023159	SLC19A1	5UR//G-3828A	--
577	rs10239964	POLM	5UR//A-2978G	--
578	rs10252226	ABCB1	5UR//A-4823C	--
579	rs10261685	ABCB1	5UR//A-1853C	--
580	rs10265836	POLM	5UR//C-2395A	--
581	rs10269664	POLM	5UR//C-3062T	--
582	rs10424731	XRCC1	5UR//T-2582C	--
583	rs10458900	RRM1	5UR//C-4328G	--
584	rs1046512	MLH1	5UR//A-2682C	--
585	rs1057985	SLC29A1	5UR//T-1340C	--
586	rs10857503	ERCC6	5UR//C-2667T	--
587	rs10857795	GSTM1	I/1/G-25A	--
588	rs10929302	UGT1A1	5UR//G-3136A	--
589	rs10947622	CDKN1A	5UR//C-4084T	--
590	rs10951974	PMS2	5UR//C-2630T	--
591	rs10987740	FPGS	5UR//G-2689A	--
592	rs11101152	ERCC6	5UR//C-1488A	--
593	rs11111858	TDG	5UR//T-1934A	--
594	rs11133442	PPAT	5UR//A-4621T	--
595	rs1114357	SLC22A7	5UR//C-3068T	--
596	rs1130609	RRM2	E/1/T-6G	--
597	rs11400410	MSH6	5UR//--3687G	--
598	rs11445091	UMPS	5UR//--3982T	--
599	rs11594945	ERCC6	5UR//T-4153C	--
600	rs11649492	ERCC4	5UR//G-4604C	--
601	rs11653440	SMARCD2	5UR//T-2821C	--
602	rs11657432	TK1	5UR//C-3010A	--
603	rs11665663	FXYD3	5UR//T-4460C	--
604	rs11671498	XRCC1	5UR//G-4956C	--
605	rs11673726	UGT1A1	5UR//G-4858T	--
606	rs11681986	RRM2	5UR//G-1507A	--
607	rs11705654	GSTT1	5UR//A-2542G	--

608	rs11769882	POLM	5UR//G-2000A	--
609	rs11779233	DPYS	5UR//T-4702A	--
610	rs11790511	SLC31A1	5UR//C-3396T	--
611	rs11809445	GPX7	E/1/G-27T	--
612	rs11816547	ABCC2	5UR//T-3037C	--
613	rs11846751	XRCC3	5UR//C-2833G	--
614	rs11846751	ZFYVE21	I/1/C2341G	--
615	rs11846838	XRCC3	5UR//G-2913A	--
616	rs11846838	ZFYVE21	I/1/G2421A	--
617	rs11848956	XRCC3	5UR//G-4444T	--
618	rs11848956	ZFYVE21	I/1/G3952T	--
619	rs11865517	MT2A	5UR//G-1835A	--
620	rs11878644	ERCC2	5UR//T-3137C	--
621	rs11880627	ERCC1	5UR//G-3793A	--
622	rs11909852	SOD1	5UR//C-4004T	--
623	rs11933670	SMARCA5	E/1/G-183A	--
624	rs11948127	FGFR4	5UR//G-1395A	--
625	rs11971012	UPP1	5UR//T-3523C	--
626	rs12068997	GSTM1	E/2/C81T	S27S
627	rs12192827	CDKN1A	5UR//C-3515T	--
628	rs12199346	CDKN1A	5UR//C-4940A	--
629	rs12214686	CDKN1A	5UR//A-3935G	--
630	rs12379987	FPGS	5UR//T-3465C	--
631	rs12445357	CES1	5UR//T-1752C	--
632	rs12463451	ERCC3	5UR//C-2130T	--
633	rs12471171	ERCC3	5UR//G-1851A	--
634	rs12492095	UMPS	5UR//T-1256A	--
635	rs12526616	POLH	I/1/A16C	--
636	rs12526616	XPO5	5UR//A-528C	--
637	rs12536587	POLM	5UR//G-709C	--
638	rs12597222	MT2A	5UR//T-2922C	--
639	rs12621805	SMARCAL1	5UR//T-3123C	--
640	rs12632590	UMPS	I/1/C30T	--
641	rs12659155	ATOX1	5UR//C-1641T	--
642	rs12806698	RRM1	E/1/C-269A	--
643	rs12949848	MPO	5UR//A-3770G	--
644	rs12978764	FXYD3	5UR//C-4585T	--
645	rs13079924	SMARCC1	5UR//A-1323C	--
646	rs1319052	ERCC1	5UR//G-4043A	--
647	rs13206175	CDKN1A	5UR//A-1452T	--
648	rs13306560	MTHFR	5UR//C-67T	--
649	rs1331700	HMGB1	5UR//C-1834T	--
650	rs13330389	CES1	5UR//A-662G	--
651	rs13401024	UPP2	5UR//C-2098T	--
652	rs13405649	SMARCAL1	5UR//G-4805T	--
653	rs13414112	UPP2	I/1/G32C	--
654	rs1364362	ERCC4	5UR//T-3128C	--
655	rs1382539	DHFR	5UR//G-1353A	--
656	rs1382541	DHFR	5UR//T-2683C	--
657	rs1382542	DHFR	5UR//T-2780C	--
658	rs140313	GSTT1	I/1/C166T	--

659	rs140315	GSTT1	5UR//G-3437A	--
660	rs140316	GSTT1	5UR//C-4042T	--
661	rs140317	GSTT1	5UR//A-4410G	--
662	rs1465952	RRM1	5UR//G-1384A	--
663	rs1469908	NQO1	5UR//C-3878T	--
664	rs1473418	BCL2	E/1/C-428G	--
665	rs1510841	SMARCAL1	I/1/G115A	--
666	rs1520195	ABCC5	5UR//G-1154A	--
667	rs1526603	SMARCE1	5UR//C-1164T	--
668	rs1544105	FPGS	5UR//C-2428T	--
669	rs1549920	ATOX1	5UR//A-2633T	--
670	rs1554494	UPP1	5UR//G-165A	--
671	rs1561876	RRM1	5UR//G-2528A	--
672	rs1611028	DHFR	5UR//TT-2880-	--
673	rs1634252	GSTM1	5UR//T-1751C	--
674	rs1643639	DHFR	5UR//T-1589C	--
675	rs1662162	RRM1	5UR//T-659C	--
676	rs1677667	DHFR	5UR//C-1740G	--
677	rs16835902	UMPS	5UR//C-3448G	--
678	rs16835912	UMPS	5UR//C-3308A	--
679	rs16856038	SMARCAL1	5UR//C-3302G	--
680	rs1688038	FXYD3	5UR//G-1799C	--
681	rs1688039	FXYD3	5UR//A-1575T	--
682	rs16896398	SLC22A7	5UR//A-3293T	--
683	rs16929410	RRM1	5UR//G-382A	--
684	rs17037425	MTHFR	5UR//G-4267A	--
685	rs17160359	ABCB1	5UR//G-4254T	--
686	rs172814	TDG	5UR//T-1802C	--
687	rs1735068	RRM1	5UR//T-265G	--
688	rs17510346	REV3L	5UR//G-1937A	--
689	rs17511525	REV3L	E/31/G9045T	I3015I
690	rs1764416	ABCC4	5UR//T-3360C	--
691	rs1766902	ABCC4	5UR//T-3275C	--
692	rs17767961	CES2	5UR//G-4155A	--
693	rs17779585	SMUG1	5UR//C-2175G	--
694	rs17843768	UMPS	5UR//C-827A	--
695	rs17880282	TP53	5UR//C-1079T	--
696	rs17882503	SOD1	5UR//T-649C	--
697	rs17883184	TP53	5UR//G-636A	--
698	rs17883908	TP53	I/1/A236G	--
699	rs17884410	TP53	E/1/T-110C	--
700	rs17886079	TP53	5UR//C-342T	--
701	rs1799797	ERCC4	5UR//T-29A	--
702	rs1811322	MT2A	5UR//A-2805T	--
703	rs1827211	MT1A	5UR//C-4321T	--
704	rs1827212	MT1A	5UR//T-2674C	--
705	rs1827213	MT1A	5UR//G-736A	--
706	rs184239	XRCC1	5UR//G-4439A	--
707	rs1862849	MT2A	5UR//G-3016A	--
708	rs1863332	MSH2	5UR//T-364G	--
709	rs1863333	MSH2	5UR//T-658C	--

710	rs1917800	ERCC6	5UR//G-3844C	--
711	rs1944423	BCL2	5UR//A-3904G	--
712	rs1977172	CDKN1A	5UR//A-4732C	--
713	rs2009115	PMS2	5UR//G-3818A	--
714	rs2014704	SMARCE1	5UR//T-4145G	--
715	rs2020872	MLH1	E/1/A94G	I32V
716	rs2066461	MTHFR	E/3/G345T	T115T
717	rs2066466	MTHFR	E/3/C417T	T139T
718	rs2070473	UPB1	5UR//G-356A	--
719	rs2070474	UPB1	E/1/C-80G	--
720	rs2073387	SMARCB1	5UR//T-144G	--
721	rs2107545	MPO	5UR//A-1821G	--
722	rs2108811	UPP2	5UR//A-3048G	--
723	rs2144078	XRCC3	5UR//T-2523C	--
724	rs2144078	ZFYVE21	I/1/T2031C	--
725	rs2145851	ABCC2	5UR//C-4676A	--
726	rs2161737	MT2A	5UR//A-3276G	--
727	rs2180989	ABCC2	5UR//T-4761G	--
728	rs219240	SMARCD3	5UR//C-32793T	--
729	rs2231135	ABCG2	E/1/A-475G	--
730	rs2232861	UPB1	5UR//G-96A	--
731	rs2233914	SLC31A1	5UR//G-327A	--
732	rs2266635	GSTT1	E/1/C61T	A21T
733	rs2270836	MT1A	5UR//C-4963T	--
734	rs2276665	SMARCAL1	5UR//T-1005G	--
735	rs2276910	SMARCAD1	I/1/T17C	--
736	rs2287497	TP53	5UR//G-1862A	--
737	rs2287498	TP53	5UR//C-1642T	--
738	rs2287499	TP53	5UR//C-1250G	--
739	rs2297393	SLC29A1	5UR//T-2522C	--
740	rs2298840	DPYS	E/1/G216A	F72F
741	rs2307160	POLB	E/1/A-62G	--
742	rs2313211	ABCC5	5UR//T-2898A	--
743	rs2334102	TYMP	5UR//G-1296C	--
744	rs2395655	CDKN1A	5UR//A-790G	--
745	rs2429247	SMARCD3	5UR//T-28911C	--
746	rs244711	FGFR4	5UR//C-4727T	--
747	rs2448341	CDC2	5UR//C-3953T	--
748	rs2448342	CDC2	5UR//A-3883G	--
749	rs2463365	SMARCC2	5UR//C-4298T	--
750	rs2467573	TK1	5UR//A-286T	--
751	rs2517953	ERBB2	5UR//G-15042C	--
752	rs2546095	FGFR4	5UR//T-4336G	--
753	rs2584618	SMARCD2	5UR//T-3259C	--
754	rs2584879	RRM1	E/1/G17A	R6Q
755	rs2607997	HLTF	5UR//G-4655T	--
756	rs2665796	SMARCD2	5UR//G-3364C	--
757	rs2682585	XRCC1	5UR//A-1572G	--
758	rs2727326	SMARCD2	5UR//G-2664A	--
759	rs2727327	SMARCD2	5UR//G-3070A	--
760	rs2804403	ABCC2	5UR//C-3357T	--

761	rs2805832	XPA	5UR//G-3188A	--
762	rs28372687	POLM	5UR//G-841A	--
763	rs28381715	ABCB1	5UR//T-334G	--
764	rs28382572	SMARCB1	5UR//C-3836T	--
765	rs28382580	SMARCB1	5UR//A-2457-	--
766	rs28382617	POLM	5UR//T-1584C	--
767	rs28382618	POLM	5UR//A-1410G	--
768	rs28382628	POLM	5UR//G-216A	--
769	rs28382644	POLM	E/5/C659G	G220A
770	rs28382812	CES2	E/1/C111T	I37I
771	rs28382815	CES2	E/2/C406T	R136*
772	rs28419190	SMARCAD1	5UR//T-2796C	--
773	rs2853533	TYMS	I/1/G117C	--
774	rs2853742	TYMS	5UR//T-176C	--
775	rs2854495	XRCC1	5UR//C-2834A	--
776	rs28565268	POLB	5UR//T-4791C	--
777	rs28570299	FPGS	5UR//A-1540T	--
778	rs28712867	TK1	5UR//C-3689T	--
779	rs2877173	SMARCB1	5UR//G-3584A	--
780	rs2892547	HMGB1	5UR//C-3393A	--
781	rs2917667	NQO1	5UR//A-3244G	--
782	rs2942570	DPYS	5UR//T-2427C	--
783	rs2959024	DPYS	5UR//T-493G	--
784	rs2959025	DPYS	5UR//A-902G	--
785	rs2959026	DPYS	5UR//A-3374G	--
786	rs2959027	DPYS	5UR//A-4439G	--
787	rs2964584	ATOX1	5UR//T-4667C	--
788	rs2992904	ABCC4	5UR//A-4485C	--
789	rs304731	XRCC1	5UR//T-3229C	--
790	rs3057854	SMARCA2	5UR//--745TATTTTT	--
791	rs3106134	SMARCAD1	5UR//A-1102G	--
792	rs3118106	SMARCA1	5UR//G-2060C	--
793	rs3131275	SMARCA1	5UR//A-2539G	--
794	rs3136038	ERCC4	5UR//C-643T	--
795	rs3136227	MSH6	5UR//C-613A	--
796	rs3136716	POLB	I/1/C159G	--
797	rs3172297	MLH1	5UR//T-2608C	--
798	rs3176320	CDKN1A	I/1/A213G	--
799	rs3176323	CDKN1A	I/1/T133C	--
800	rs3176628	XPA	5UR//G-845A	--
801	rs3176629	XPA	5UR//G-383A	--
802	rs3204953	REV3L	E/31/C9190T	V3064I
803	rs3212929	ERCC1	5UR//C-450A	--
804	rs3212935	ERCC1	E/1/T-143C	--
805	rs3212936	ERCC1	E/1/--78C	--
806	rs3213138	E2F1	5UR//A-322G	--
807	rs3213143	E2F1	E/1/G111A	S37S
808	rs3213174	E2F1	E/6/G932T	T311N
809	rs3213177	E2F1	E7/3UTR /C5A	--
810	rs3213180	E2F1	E7/3UTR /G914C	--
811	rs3213236	XRCC1	5UR//--1739G	--

812	rs3213246	XRCC1	E/1/G-64A	--
813	rs3218602	REV3L	E/31/A9108G	Y3036Y
814	rs3218655	POLM	E/1/C102A	L34L
815	rs3218657	POLM	E/2/C351T	V117V
816	rs322102	TDG	I/1/C94G	--
817	rs322104	TDG	5UR//G-472A	--
818	rs330792	MSH6	5UR//A-1646C	--
819	rs33934538	SMARCA2	5UR//A-2071-	--
820	rs34009709	TDG	5UR//G-2327A	--
821	rs34067256	TP53	5UR//C-1455G	--
822	rs34144509	WDR7	5UR//T-1451C	--
823	rs34213602	FDXR	I/1/C147G	--
824	rs34248325	SMARCB1	5UR//G-711A	--
825	rs34428341	CES1	5UR//C-1005T	--
826	rs34484367	SMARCE1	5UR//C-2083G	--
827	rs34547608	UGT1A1	5UR//T-90C	--
828	rs34547779	FDXR	5UR//G-667C	--
829	rs34800257	MT2A	5UR//T-1478C	--
830	rs34976170	SMARCB1	5UR//C-482G	--
831	rs34994762	MTHFR	5UR//A-4163G	--
832	rs35033646	FDXR	E12/3UTR /G139T	--
833	rs35072974	FDXR	E/8/G743A	P248L
834	rs35263175	SLC22A7	E/1/T222C	D74D
835	rs35395489	MT2A	5UR//A-561G	--
836	rs35448124	ERCC4	5UR//TGTG-4441-	--
837	rs35466868	ERBB2	5UR//C-13786T	--
838	rs35634719	FDXR	5UR//TG-1250-	--
839	rs35665780	UGT1A1	5UR//C-620T	--
840	rs35880761	FDXR	5UR//A-1824G	--
841	rs35891829	TP53	5UR//A-2059C	--
842	rs35960304	RRM1	5UR//C-3341T	--
843	rs35988004	MT1A	5UR//T-1668-/A	--
844	rs36010696	FDXR	5UR//A-1108T	--
845	rs36106739	FDXR	I/1/T292C	--
846	rs36124867	TYMS	5UR//A-4424C	--
847	rs36204705	FPGS	5UR//G-1011T	--
848	rs36209093	GSTM1	5UR//C-654T	--
849	rs36211400	GSTT1	5UR//C-688G	--
850	rs36230817	SLC19A1	5UR//A-1948G	--
851	rs36233090	SOD1	5UR//C-1659G	--
852	rs3734701	SLC29A1	5UR//T-1987C	--
853	rs3735295	PMS2	I/1/G72A	--
854	rs3735296	PMS2	5UR//G-66C	--
855	rs3737965	MTHFR	5UR//G-335A	--
856	rs3747802	ABCB1	5UR//A-21G	--
857	rs3750187	DPYS	5UR//G-3197A	--
858	rs3750747	ERCC6	I/1/G156C	--
859	rs3750994	RRM1	5UR//T-2453G	--
860	rs3750996	RRM1	5UR//A-2723G	--
861	rs3755141	SMARCAL1	5UR//C-1325T	--
862	rs3755142	SMARCAL1	5UR//T-1647C	--

863	rs376184	SMARCD1	5UR//T-4942C	--
864	rs3763505	UPP1	E/1/G-363A	--
865	rs3787037	FXYD3	5UR//A-1925G	--
866	rs3787039	FXYD3	5UR//A-2261G	--
867	rs3794050	RRM1	5UR//A-4023G	--
868	rs3806573	SMARCAL1	5UR//A-551G	--
869	rs3809159	SMUG1	5UR//G-1270C	--
870	rs3814270	ABCC4	5UR//G-639A	--
871	rs3826317	FDXR	5UR//C-491T	--
872	rs3826992	FXYD3	5UR//G-2347C	--
873	rs3829965	CDKN1A	5UR//A-1976G	--
874	rs3829967	CDKN1A	5UR//T-1528C	--
875	rs3831222	SLC29A1	5UR//G-2445-	--
876	rs406113	GPX6	E/1/A39C	F13L
877	rs4124874	UGT1A1	5UR//T-3259G	--
878	rs412543	GSTM1	5UR//G-497C	--
879	rs41275676	SMARCAD1	E/2/A-6G	--
880	rs41294988	MSH6	E/1/A38C	K13T
881	rs4135036	TDG	5UR//T-562C	--
882	rs4145763	UCK2	5UR//C-1645T	--
883	rs4147563	GSTM1	5UR//C-338T	--
884	rs4147581	GSTP1	I/1/C-20G	--
885	rs4148727	ABCB1	5UR//A-201G	--
886	rs4151702	CDKN1A	5UR//G-498C	--
887	rs423143	FGFR4	5UR//G-4265A	--
888	rs4253003	ERCC6	5UR//C-137T	--
889	rs4253006	ERCC6	E/1/C-21T	--
890	rs4253046	ERCC6	E/5/T1274G	D425A
891	rs4399719	UGT1A1	5UR//T-2457G	--
892	rs4401102	MPO	5UR//C-2540T	--
893	rs442767	DHFR	5UR//G-695T	--
894	rs453544	DHFR	5UR//G-1008A	--
895	rs45498791	MTHFR	5UR//C-479A	--
896	rs45546035	MTHFR	E/3/G276A	D92D
897	rs45557639	FGFR4	5UR//T-3458C	--
898	rs45585938	FGFR4	5UR//A-3570G	--
899	rs4647201	MLH1	5UR//C-1487A	--
900	rs4647203	MLH1	5UR//G-794A	--
901	rs4681481	HLTF	5UR//T-2502C	--
902	rs4711460	CDKN1A	5UR//C-4358A	--
903	rs4714003	CDKN1A	5UR//C-4226T	--
904	rs4724280	POLM	5UR//G-2496A	--
905	rs4741636	SMARCA2	5UR//T-1761C	--
906	rs4759344	SMUG1	5UR//A-357G	--
907	rs477415	WDR7	I/1/A292T	--
908	rs4784701	MT1A	5UR//T-1910G	--
909	rs4806085	FXYD3	5UR//T-4784C	--
910	rs4806087	FXYD3	5UR//G-3524A	--
911	rs4806088	FXYD3	5UR//G-3458A	--
912	rs4846054	MTHFR	5UR//C-3114T	--
913	rs485365	WDR7	5UR//G-3183A	--

914	rs487140	WDR7	5UR//G-3339A	--
915	rs492095	FDXR	E/1/A-9G	--
916	rs4987706	BCL2	I/1/T42C	--
917	rs501415	WDR7	I/1/A35G	--
918	rs507964	SLC29A1	5UR//T-4905G	--
919	rs535437	WDR7	5UR//G-1238A	--
920	rs559998	WDR7	5UR//G-1592T	--
921	rs561762	WDR7	5UR//C-1764T	--
922	rs572362	FDXR	5UR//T-870G	--
923	rs6003880	SMARCB1	5UR//C-219T	--
924	rs6120343	E2F1	5UR//G-3228A	--
925	rs613653	WDR7	5UR//T-4204C	--
926	rs6141997	E2F1	5UR//T-2621C	--
927	rs6151599	DHFR	5UR//A-231G	--
928	rs6151600	DHFR	5UR//T-907C	--
929	rs621018	WDR7	5UR//A-443G	--
930	rs628047	WDR7	5UR//C-3276G	--
931	rs632427	WDR7	5UR//C-703T	--
932	rs6440583	HLTF	5UR//C-857T	--
933	rs6498485	ERCC4	5UR//G-3028A	--
934	rs6499786	CES1	5UR//A-2159G	--
935	rs6499788	CES1	5UR//T-4761A	--
936	rs6511716	SMARCA4	5UR//A-3770G	--
937	rs6580978	SMUG1	5UR//A-779G	--
938	rs662855	WDR7	5UR//G-2907T	--
939	rs6672420	RUNX3	5UR//A-34239T	--
940	rs6689902	GSTM1	5UR//A-1643C	--
941	rs6714634	UGT1A1	5UR//T-4153C	--
942	rs6798870	ABCC5	5UR//G-4622A	--
943	rs6830518	PPAT	5UR//C-4949T	--
944	rs6883528	SLCO6A1	5UR//T-1419C	--
945	rs688890	WDR7	5UR//G-288A	--
946	rs689384	WDR7	5UR//T-161C	--
947	rs689456	NQO1	5UR//C-1029T	--
948	rs6898458	SLCO6A1	5UR//G-1944A	--
949	rs690115	FDXR	5UR//C-4912T	--
950	rs6976251	PMS2	5UR//G-543C	--
951	rs7040790	SMARCA2	5UR//T-1379G	--
952	rs717378	UPP2	5UR//G-3222C	--
953	rs7196890	MT1A	5UR//C-1557A	--
954	rs7202530	CES2	5UR//T-4928A	--
955	rs7219483	SMARCE1	5UR//A-3853G	--
956	rs7277748	SOD1	E/1/A-109G	--
957	rs7289159	UPB1	5UR//G-901T	--
958	rs7292735	TYMP	5UR//A-2347G	--
959	rs7328090	ATP7B	5UR//G-2175T	--
960	rs7459020	UPP1	5UR//G-149A	--
961	rs7498748	CES1	5UR//T-163C	--
962	rs7534738	UCK2	5UR//A-1501C	--
963	rs7556417	UCK2	5UR//C-2312T	--
964	rs7582263	RRM2	5UR//C-1246A	--

965	rs7619819	ABCC5	5UR//A-3831G	--
966	rs762623	CDKN1A	5UR//G-1020A	--
967	rs762624	CDKN1A	5UR//A-898C	--
968	rs7628248	HLTF	5UR//C-3217A	--
969	rs7688482	PPAT	5UR//T-2585C	--
970	rs7753792	SLC29A1	5UR//C-1718A	--
971	rs7793905	UPP1	5UR//C-1160T	--
972	rs7927657	GSTP1	5UR//T-4297C	--
973	rs7934581	RRM1	5UR//C-3890T	--
974	rs7941648	GSTP1	5UR//C-3690T	--
975	rs7945035	GSTP1	5UR//A-3123G	--
976	rs796736	TDG	5UR//T-2448A	--
977	rs7991067	HMGB1	5UR//C-4553A	--
978	rs8068725	FDXR	5UR//A-3081G	--
979	rs8071253	TK1	5UR//G-1181A	--
980	rs8073029	FDXR	5UR//T-3265C	--
981	rs812498	TDG	5UR//T-4817C	--
982	rs8132524	SOD1	5UR//C-2443T	--
983	rs8137555	SMARCB1	5UR//T-2842C	--
984	rs8177412	GPX3	E/1/T-129C	--
985	rs8177413	GPX3	E/1/G39C	L13L
986	rs8191439	GSTP1	E/1/G-18A	--
987	rs8192443	SMARCD1	5UR//C-3479A	--
988	rs868853	ABCC4	5UR//C-1388T	--
989	rs879000	WDR7	5UR//A-4398G	--
990	rs897761	FXYD3	5UR//T-326A	--
991	rs903501	ERBB2	5UR//T-16760C	--
992	rs9311149	MLH1	5UR//C-4803A	--
993	rs9333500	POLH	E/1/G-200T	--
994	rs9333500	XPO5	5UR//G-316T	--
995	rs9357436	SLC29A1	5UR//G-4521A	--
996	rs961077	UMPS	5UR//C-4613T	--
997	rs9630729	SMARCE1	5UR//T-3757C	--
998	rs9673491	MT1A	5UR//G-3655T	--
999	rs9674227	MT1A	5UR//A-3590C	--
1000	rs9890046	ABCC3	I/1/C369G	--
1001	rs9927585	CES1	5UR//G-4382A	--
1002	rs9947507	TYMS	5UR//C-555T	--
1003	rs17222547	ABCC2	E/22/C2901A	Y967*
1004	rs2066523	SMARCAL1	E/11/G1945A	R649*
1005	rs2291078	UMPS	E/4/T1050A	C350*
1006	rs3201997	ABCG2	E/9/C1000A	E334*
1007	rs5023780	CES1	E/3/G310A	R104*
1008	rs596909	TYMS	E/4/G470T	L157*
1009	rs1003355	ABCC3	E/12/C1583G	A528G
1010	rs10055840	SLCO6A1	E/12/G1961C	T654R
1011	rs10073333	SLCO6A1	E/9/G1579C	P527A
1012	rs10091081	POLB	E/8/G431C	G144A
1013	rs1042709	UGT1A1	E/5/G1531C	A511P
1014	rs1050101	SMARCD3	E/5/G508A	P170S
1015	rs1061017	ABCG2	E/5/G496C	Q166E

1016	rs1061018	ABCG2	E/6/A623G	F208S
1017	rs10964468	SMARCA2	E/2/G-5A	--
1018	rs10983315	XPA	E/3/C289T	V97I
1019	rs1132543	TK1	E/5/C442T	V148M
1020	rs1138272	GSTP1	E/6/C341T	A114V
1021	rs11553892	GSTP1	E/7/C526A	L176M
1022	rs11555797	SMARCAL1	E/2/G341A	T114M
1023	rs11568587	ABCC3	E/22/-3051-	--
1024	rs11568591	ABCC3	E/27/G3890A	R1297H
1025	rs11568599	ABCC3	E/15/-1926-	--
1026	rs11568658	ABCC4	E/5/C559A	G187W
1027	rs11568669	ABCC4	E/11/T1492C	K498E
1028	rs11568705	ABCC4	E/9/G1208A	P403L
1029	rs11656685	ABCC3	E/31/C4538A	A1513D
1030	rs11708427	ABCC5	3DR//C1158G	--
1031	rs11722476	SMARCAD1	E/7/G740A	S247N
1032	rs11723410	SMARCAD1	E/3/C197T	S66F
1033	rs11954456	FGFR4	E/6/C825G	S275R
1034	rs11971829	UPP1	E/3/C11T	T4M
1035	rs1202183	ABCB1	E/5/T131C	N44S
1036	rs12103928	SMARCE1	E/5/T183G	K61N
1037	rs12367528	TDG	E/10/C1136A	P379H
1038	rs12388502	SMARCA1	E/17/C2113G	E705Q
1039	rs12678588	POLB	E/7/G410A	R137Q
1040	rs12682945	UCK1	E/7/A785G	L262P
1041	rs12686275	FPGS	E/14/T1160A	V387D
1042	rs12808005	RRM1	E/7/A536C	H179P
1043	rs12928616	ERCC4	E/11/C2240T	S747F
1044	rs12928650	ERCC4	E/11/C2303T	S768F
1045	rs13079661	SMARCC1	E/14/T1357A	S453C
1046	rs13091100	HLTF	E/3/A248C	V83G
1047	rs13308178	POLM	E/10/C1357G	G453R
1048	rs1332018	GSTM3	E/1/G-63T	--
1049	rs13400205	RRM2	E/4/T320G	G107V
1050	rs1695	GSTP1	E/5/A313G	I105V
1051	rs17150488	SLCO6A1	E/7/T1142C	K381R
1052	rs17216317	ABCC2	E/28/C3872T	P1291L
1053	rs17222674	ABCC2	E/8/A998G	D333G
1054	rs17510963	REV3L	E/14/A6225C	I2075M
1055	rs17539588	REV3L	E/13/C2885T	R962Q
1056	rs17539616	REV3L	E/13/G4015T	P1339T
1057	rs17539692	REV3L	E/14/T6044A	E2015V
1058	rs17843776	UMPS	E/1/A88G	R30W
1059	rs17843819	UMPS	E/3/T859C	E287K
1060	rs17880492	GPX2	E/2/G436A	R146C
1061	rs17882252	TP53	E/10/C1015T	E339K
1062	rs1799792	ERCC2	E/8/G601A	H201Y
1063	rs1799802	ERCC4	E/7/C1135T	P379S
1064	rs1800067	ERCC4	E/8/G1244A	R415Q
1065	rs1800068	ERCC4	E/8/G1727C	R576T
1066	rs1800124	ERCC4	E/11/A2624G	E875G

1067	rs1800152	MSH2	E/12/T1917G	H639Q
1068	rs1800938	MSH6	E/4/A660C	E220D
1069	rs1801201	ERBB2	E/17/A1960G	I654V
1070	rs1801244	ATP7B	E/3/C1366G	V456L
1071	rs1801266	DPYD	E/7/G703A	R235W
1072	rs1803687	GSTM3	E/6/C384G	K128N
1073	rs1805318	PMS2	E/11/T1789A	T597S
1074	rs1805322	PMS2	E/8/G830T	T277K
1075	rs1805323	PMS2	E/11/G1454T	T485K
1076	rs1805324	PMS2	E/11/C1866T	M622I
1077	rs1966265	FGFR4	E/1/G28A	V10I
1078	rs2020908	MSH6	E/4/C1186G	L396V
1079	rs2020955	ERCC4	E/10/T1984C	S662P
1080	rs2020956	ERCC4	E/11/G2735A	G912E
1081	rs2020959	ERCC4	E/11/C2169A	C723*
1082	rs2020961	ERCC4	E/3/C503T	A168V
1083	rs2066472	MTHFR	E/2/C203T	R68Q
1084	rs2066522	SMARCAL1	E/4/C945G	Q315H
1085	rs2066524	SMARCAL1	E/2/G127A	C43R
1086	rs2227291	ATP7A	E/10/G2299C	V767L
1087	rs2227963	GSTM5	E/7/G536A	L179P
1088	rs2228006	PMS2	E/11/T1621C	E541K
1089	rs2228529	ERCC6	E/21/T4238C	Q1413R
1090	rs2229107	ABCB1	E/27/A3421T	S1141T
1091	rs2229361	HLTF	E/21/C2456T	R819H
1092	rs2229996	APC	E/16/C4487G	T1496S
1093	rs2231137	ABCG2	E/2/C34T	V12M
1094	rs2233915	SLC31A1	E/2/C73G	P25A
1095	rs2233916	SLC31A1	E/4/C365G	T122S
1096	rs2234935	ATP7A	E/9/T2006C	I669T
1097	rs2234953	GSTT1	E/4/C517T	E173K
1098	rs2235036	ABCB1	E/16/C1795T	A599T
1099	rs2235039	ABCB1	E/21/C2401T	V801M
1100	rs2266633	GSTT1	E/4/C421T	D141N
1101	rs2271336	SMARCAL1	E/13/C2225T	Y742C
1102	rs2274974	MTHFR	E/11/C1697T	G566E
1103	rs2277447	ATP7B	E/7/C2029T	E677K
1104	rs2296212	SMARCA2	E/32/C4584G	D1528E
1105	rs2305868	HLTF	E/8/T932C	N311S
1106	rs2307167	XRCC1	E/15/C1676T	R559Q
1107	rs2307184	XRCC1	E/13/G1454T	S485Y
1108	rs2307191	XRCC1	E/5/G482A	P161L
1109	rs2307227	CES1	E/5/G609T	D203E
1110	rs2307240	CES1	E/2/C224T	S75N
1111	rs2307456	POLH	E/5/G626T	G209V
1112	rs2388544	HMGB1	E/3/G293A	P98L
1113	rs25474	XRCC1	E/14/G1541A	P514L
1114	rs25491	XRCC1	E/9/G925A	P309S
1115	rs2632398	SMARCAD1	E/4/C418T	R140C
1116	rs2682557	XRCC1	E/16/T1726A	Y576N
1117	rs28364274	ABCB1	E/29/C3751T	V1251I

1118	rs28381801	ABCB1	E/2/T-42C	--
1119	rs28381902	ABCB1	E/15/C1696T	E566K
1120	rs28381967	ABCB1	E/22/T2506C	I836V
1121	rs28382653	POLM	E/6/C736A	V246F
1122	rs28382661	POLM	E/11/G1450A	L484F
1123	rs28401798	ABCB1	E/26/G3151C	P1051A
1124	rs28563878	CES1	E/1/A34C	S12A
1125	rs3103135	SMARCAD1	E/14/A1765C	N589H
1126	rs3113842	SMARCAD1	E/12/T1612C	Y538H
1127	rs3116448	ABCG2	E/7/A742G	S248P
1128	rs3136334	MSH6	E/4/C1867G	P623A
1129	rs3136389	SMUG1	E/4/G313A	R105W
1130	rs3188420	ERCC1	E/2/G230T	P77H
1131	rs3212057	XRCC3	E/5/C281T	R94H
1132	rs3212977	ERCC1	E/8/C796T	A266T
1133	rs3213172	E2F1	E/5/C755T	R252H
1134	rs3213173	E2F1	E/5/C826T	V276M
1135	rs3213176	E2F1	E/7/C1177T	G393S
1136	rs3218572	REV3L	E/13/T4406G	Q1469P
1137	rs3218578	REV3L	E/13/T3850G	T1284P
1138	rs3218579	REV3L	E/10/T1190G	Q397P
1139	rs3218582	REV3L	E/13/G4727A	S1576L
1140	rs3218585	REV3L	E/13/C5137T	D1713N
1141	rs3218592	REV3L	E/26/C8285T	R2762Q
1142	rs3218593	REV3L	E/13/A2078G	M693T
1143	rs3218595	REV3L	E/13/C3927A	Q1309H
1144	rs3218600	REV3L	E/13/G3659A	S1220L
1145	rs3218604	REV3L	E/14/C5767G	G1923R
1146	rs3218606	REV3L	E/14/C5909T	R1970H
1147	rs33974176	APC	E/16/C2608T	P870S
1148	rs34110964	UPB1	E/9/C1019A	A340D
1149	rs34138361	FGFR4	E/11/C1652T	S551F
1150	rs34157245	APC	E/16/G5645C	R1882T
1151	rs34284947	FGFR4	E/10/G1466A	R489Q
1152	rs34354111	FPGS	E/15/G1433C	S478T
1153	rs34447156	NQO1	E/5/C693G	Q231H
1154	rs34622270	HLTF	E/20/T2284C	I762V
1155	rs34628871	MPO	E/1/A-65G	--
1156	rs34825130	GPX6	E/2/A157G	Y53H
1157	rs35003977	UGT1A1	E/1/T674G	V225G
1158	rs35102176	FDXR	E/7/G624C	C208W
1159	rs35163653	TP53	E/6/C649T	V217M
1160	rs35338630	MLH1	E/9/C790G	H264D
1161	rs35394555	GPX6	E/4/C408G	E136D
1162	rs35578165	FXYD3	E/6/G118A	G40S
1163	rs35658392	GPX6	E/5/G469A	P157S
1164	rs35660143	FDXR	E/10/G1034A	T345M
1165	rs35667202	UPP2	E/8/G674T	R225L
1166	rs35670089	MPO	E/11/G1810A	R604C
1167	rs35671174	CDC2	E/6/A665G	K222R
1168	rs35675573	POLH	E/8/C986T	T329I

1169	rs35702888	MPO	E/12/C2047G	E683Q
1170	rs35717727	MSH6	E/8/G3700C	E1234Q
1171	rs35737219	MTHFR	E/12/G1958A	T653M
1172	rs35993958	TP53	E/10/C1079G	G360A
1173	rs36027551	DPYS	E/3/G541A	R181W
1174	rs36040909	SLC22A7	E/6/C973T	R325W
1175	rs3740071	ABCC2	E/20/G2677C	E893Q
1176	rs3740072	ABCC2	E/17/A2153G	N718S
1177	rs3765534	ABCC4	E/18/C2269T	E757K
1178	rs376618	FGFR4	E/3/C407T	P136L
1179	rs3772406	SMARCC1	E/28/G3224T	P1075H
1180	rs3772809	UMPS	E/6/A1336G	H446Y
1181	rs377860	APC	E/16/A4108C	K1370Q
1182	rs3793510	SMARCA2	E/29/G4247C	G1416A
1183	rs3826192	CES1	E/2/C112T	V38I
1184	rs41295268	MSH6	E/4/G1403A	R468H
1185	rs41295270	MSH6	E/4/C1739T	S580L
1186	rs41306702	FPGS	E/2/C103T	R35W
1187	rs41318029	ABCC2	E/21/G2761A	G921S
1188	rs4135038	TDG	E/1/G-169A	--
1189	rs4135113	TDG	E/5/G595A	G199S
1190	rs4148323	UGT1A1	E/1/G211A	G71R
1191	rs4148460	ABCC4	E/4/A511C	C171G
1192	rs4150522	ERCC3	E/14/A2203G	S735P
1193	rs41540513	ERCC1	5UR//G-320A	--
1194	rs41542214	MLH1	E/18/C2065A	Q689K
1195	rs41549213	ERCC6	E/7/C1670T	R557H
1196	rs41552412	ERCC4	E/8/C1563G	S521R
1197	rs41557814	ERCC4	E/8/C1429T	R477W
1198	rs41557921	ERCC6	E/15/C2803G	D935H
1199	rs41559922	ERCC2	E/14/A1343G	F448S
1200	rs4253047	ERCC6	E/5/C1337T	G446D
1201	rs4253206	ERCC6	E/17/T3005C	Y1002C
1202	rs4253219	ERCC6	E/19/C3965A	G1322V
1203	rs4253227	ERCC6	E/21/C4114T	G1372R
1204	rs4253230	ERCC6	E/21/G4322A	T1441I
1205	rs4338942	SMARCAL1	E/2/T569G	V190G
1206	rs45441199	ABCC2	E/23/T3107C	I1036T
1207	rs45458701	SLC29A1	E/12/G1171A	E391K
1208	rs45462493	ABCC2	E/7/A736C	M246L
1209	rs45477596	ABCC4	3DR//C12523T	--
1210	rs45496998	MTHFR	E/10/G1555A	R519C
1211	rs45504892	ABCC4	3DR//C20231A	--
1212	rs45573936	SLC29A1	E/7/T647C	I216T
1213	rs45589337	DPYD	E/8/T775C	K259E
1214	rs45617731	ABCC3	E/8/G941T	S314I
1215	rs458017	REV3L	E/13/T3467C	Y1156C
1216	rs4705693	APC	E/16/G7567T	A2523S
1217	rs4826245	ATP7A	E/21/G4048A	E1350K
1218	rs4851	GPX4	E/7/C680T	S227L
1219	rs4986866	CDKN1A	E/2/C11T	P4L

1220	rs4986867	CDKN1A	E/2/C189A	F63L
1221	rs4986949	GSTP1	E/6/G439T	D147Y
1222	rs4997557	CYP2A6	E/6/G881C	T294S
1223	rs520611	FDXR	E/2/C120A	Q40H
1224	rs5959130	ATP7A	E/22/G4201C	V1401L
1225	rs598078	WDR7	E/3/A235C	K79Q
1226	rs6446261	GPX1	E/2/C580T	A194T
1227	rs6710480	UPP2	E/3/G30T	R10S
1228	rs6941583	POLH	E/11/A1939T	M647L
1229	rs7020514	SMARCA2	E/8/T1362G	S454R
1230	rs7080681	ABCC2	E/9/G1058A	R353H
1231	rs728619	MSH6	E/4/A1613C	Y538S
1232	rs7356934	REV3L	E/25/G8036A	P2679L
1233	rs7439869	SMARCAD1	E/9/T902C	A301V
1234	rs7483	GSTM3	E/8/C670T	V224I
1235	rs7561584	UPP2	E/5/A232C	M78L
1236	rs769188	GPX5	E/3/C253G	L85V
1237	rs8177445	GPX3	E/4/T382C	F128L
1238	rs8192730	CYP2A6	E/8/C1257G	E419D
1239	rs8192924	CES2	E/5/G809A	R270H
1240	rs927344	ABCC2	E/2/A116T	F39Y
1241	rs9282571	ABCG2	E/14/A1711T	F571I
1242	rs9296419	POLH	E/11/C1433T	T478M
1243	rs9333555	POLH	E/11/A1783G	M595V
1244	rs1042858	RRM1	E/19/G2232A	A744A
1245	rs1048977	CDA	E/4/C435T	T145T
1246	rs1050102	SMARCD3	E/8/G828A	H276H
1247	rs1056806	GSTM1	E/7/C528T	D176D
1248	rs1065767	TK1	E/7/G720A	A240A
1249	rs10964471	SMARCA2	E/2/G177A	T59T
1250	rs11100790	SMARCA5	E/3/T282C	Y94Y
1251	rs11553301	UCK2	E/4/G408A	G136G
1252	rs11568695	ABCC4	3DR//C51485T	--
1253	rs11568704	ABCC4	3DR//C61070T	--
1254	rs11682453	UGT1A1	E/4/C1279T	L427L
1255	rs11786893	GGH	E/2/C174T	A58A
1256	rs11840224	ATP7B	E/13/G2967A	D989D
1257	rs1189466	ABCC4	3DR//A21484G	--
1258	rs1200937	CES2	E/1/G-140C	--
1259	rs12532895	PMS2	E/4/G288A	A96A
1260	rs13288443	SMARCA2	E/11/A1827G	P609P
1261	rs13306555	MTHFR	E/6/G906A	A302A
1262	rs13427563	ERCC3	E/5/C615T	E205E
1263	rs13428173	ERCC3	E/11/G1740A	I580I
1264	rs17090346	WDR7	E/15/C2542T	L848L
1265	rs17216296	ABCC2	E/30/C4242T	H1414H
1266	rs17510914	REV3L	E/13/T4650C	P1550P
1267	rs17849090	PTEN	E/7/T723C	L241L
1268	rs1800144	MLH1	E/4/A375G	A125A
1269	rs1800150	MSH2	E/2/G219A	K73K
1270	rs1800369	TP53	E/2/G63A	D21D

1271	rs1800932	MSH6	E/2/A276G	P92P
1272	rs1800935	MSH6	E/3/T540C	D180D
1273	rs1801018	BCL2	E/2/T21C	T7T
1274	rs1801019	UMPS	E/3/G638C	--
1275	rs1801248	ATP7B	E/13/C3045T	L1015L
1276	rs1970951	GPX7	E/2/T237C	F79F
1277	rs2020913	MSH6	E/4/T2253C	N751N
1278	rs2020953	ERCC4	E/11/A2463G	P821P
1279	rs2020958	ERCC4	E/9/A1884G	E628E
1280	rs2066527	SMARCAL1	E/11/T2070C	N690N
1281	rs2228544	ERCC3	E/8/C1119T	Q373Q
1282	rs2229993	APC	E/16/G6921A	S2307S
1283	rs2229997	APC	E/16/C5250G	V1750V
1284	rs2230761	UPP1	E/7/A606G	T202T
1285	rs2232867	UPB1	E/7/C846T	F282F
1286	rs2232870	UPB1	E/10/T1086C	Y362Y
1287	rs2288845	SMARCA4	E/9/C1557T	N519N
1288	rs2307174	XRCC1	E/3/C150T	E50E
1289	rs2307189	XRCC1	E/2/G126T	T42T
1290	rs2307460	POLH	E/6/C678T	A226A
1291	rs28381867	ABCB1	E/9/C738T	A246A
1292	rs28382610	GPX5	E/5/C633T	I211I
1293	rs28382827	CES2	E/12/C1791T	L597L
1294	rs28997580	SMARCA4	E/16/C2388T	L796L
1295	rs28997582	SMARCA4	E/29/C4053T	D1351D
1296	rs3136804	POLB	E/13/C888T	Y296Y
1297	rs3212045	XRCC3	E/4/G132A	P44P
1298	rs3218577	REV3L	E/27/C8367A	L2789L
1299	rs34219015	WDR7	E/22/C3714A	I1238I
1300	rs34237683	RRM2	E/1/C87T	R29R
1301	rs34308410	SMARCC2	E/13/T1158C	E386E
1302	rs34312619	MSH2	E/2/C336A	S112S
1303	rs34474865	HLTF	E/16/T1740C	R580R
1304	rs34566456	MLH1	5UR//G-532C	--
1305	rs35013010	DPYS	E/6/A1062G	D354D
1306	rs35043160	APC	E/16/A7704G	G2568G
1307	rs35051203	SMARCD3	E/13/C1422T	L474L
1308	rs35182583	ERCC6	E/10/C2082T	P694P
1309	rs35225190	MLH1	E/7/A552T	S184S
1310	rs35420817	PPAT	E/9/T1026C	P342P
1311	rs35464006	ERBB2	5UR//G-15393C	--
1312	rs35642130	MSH6	E/5/G3354A	E1118E
1313	rs35653697	MTHFR	E/9/C1476T	P492P
1314	rs35756610	ERCC6	E/18/T3774C	K1258K
1315	rs35908749	MLH1	E/9/G702A	E234E
1316	rs3737967	MTHFR	E12/3UTR /G3288A	--
1317	rs3916876	ERCC2	E/18/G1737A	V579V
1318	rs41280126	ABCC3	E/16/A2043G	L681L
1319	rs4135119	TDG	E/8/C795T	L265L
1320	rs4135120	TDG	E/8/C867T	Y289Y
1321	rs41557516	REV3L	E/30/G8874A	P2958P

1322	rs4252610	ERBB2	E/2/G99A	L33L
1323	rs4252655	ERBB2	E/25/C3078T	P1026P
1324	rs4252656	ERBB2	E/27/G3531A	K1177K
1325	rs4253013	ERCC6	E/2/C411T	L137L
1326	rs4253027	ERCC6	E/3/T528C	R176R
1327	rs4253044	ERCC6	E/5/A885G	A295A
1328	rs4253210	ERCC6	E/18/A3534G	F1178F
1329	rs446382	FGFR4	E/2/T162G	R54R
1330	rs452885	FGFR4	E/5/C702T	R234R
1331	rs455732	REV3L	E/13/C4290T	V1430V
1332	rs458486	REV3L	E/13/T2706C	G902G
1333	rs4647256	MLH1	E/6/C474T	N158N
1334	rs4807542	GPX4	5UR//G-570A	--
1335	rs507082	SMARCC1	E/5/G532A	L178L
1336	rs6670886	DPYD	E/6/C525T	S175S
1337	rs6823404	SMARCAD1	E/15/C1839T	D613D
1338	rs689453	NQO1	E/2/C72T	E24E
1339	rs6947955	UPP1	E/2/C-43T	--
1340	rs6972869	PMS2	E/11/A1557G	Y519Y
1341	rs7136420	SMARCC2	E/5/T438C	P146P
1342	rs7275	SMARCA4	E/34/T4887C	D1629D
1343	rs7636910	ABCC5	3DR//T2025C	--
1344	rs7899457	ABCC2	E/29/C4110T	L1370L
1345	rs8182267	ERBB2	5UR//G-4942A	--
1346	rs8187630	SLC29A1	E/3/G84A	P28P
1347	rs8187706	ABCC2	E/31/G4410A	E1470E
1348	rs8187707	ABCC2	E/31/C4488T	H1496H
1349	rs9105	SMARCA4	E/32/C4584T	D1528D
1350	rs939336	ABCC5	3DR//A16007G	--
1351	rs1537516	MTHFR	E12/3UTR /G2876A	--
1352	rs17885803	TP53	5UR//C-1564T	--
1353	rs17886250	TP53	I/1/C148T	--
1354	rs2119342	HLTF	E25/3UTR /T386C	--
1355	rs28399438	CYP2A6	I/1/A-34C	--
1356	rs2856857	MPO	I/1/G-13A	--
1357	rs3176626	XPA	5UR//A-951C	--
1358	rs34957864	HLTF	I/1/C235T	--
1359	rs3829963	CDKN1A	5UR//C-2100A	--
1360	rs3890213	CES2	5UR//C-1547T	--
1361	rs408626	DHFR	5UR//T-332C	--
1362	rs4987707	BCL2	I/1/C-87T	--
1363	rs11568589	ABCC3	E/28/G4008A	L1336L
1364	rs2307083	SMARCD1	E/4/A423G	V141V
1365	rs3547	XRCC1	E/17/T1896C	Q632Q
1366	rs1042482	DPYD	E23/3UTR /C573T	--
1367	rs1061388	REV3L	E32/3UTR /A925C	--
1368	rs11591427	PTEN	E9/3UTR /A2821T	--
1369	rs1803541	ERCC3	E15/3UTR /C259T	--
1370	rs2584622	SMARCD2	E13/3UTR /G131A	--
1371	rs35180794	GPX7	E3/3UTR /T162-	--
1372	rs41436046	FPGS	E15/3UTR /TT442-	--

1373	rs4150525	ERCC3	E15/3UTR /T177C	--
1374	rs4252661	ERBB2	E27/3UTR /T591C	--
1375	rs45616134	SMARCA2	E33/3UTR /T698C	--
1376	rs4987864	BCL2	3DR//C193453A	--
1377	rs4987865	BCL2	3DR//C193630T	--
1378	rs4987868	BCL2	3DR//G194378T	--
1379	rs5030760	RRM2	E10/3UTR /T893C	--
1380	rs5031031	GSTP1	E7/3UTR /A56G	--
1381	rs7337469	HMGB1	E5/3UTR /T2156C	--
1382	rs975922	HLTF	E25/3UTR /A1326G	--
1383	rs9955129	WDR7	E27/3UTR /T943C	--
1384	rs1042927	RRM1	E19/3UTR /C316A	--
1385	rs1047619	GPX7	E3/3UTR /A315G	--
1386	rs1047635	GPX7	E3/3UTR /C435A	--
1387	rs10981708	SLC31A1	E5/3UTR /C3287T	--
1388	rs11151988	WDR7	E27/3UTR /T1272G	--
1389	rs11610906	TDG	E10/3UTR /C1585G	--
1390	rs12106470	SLC19A1	E6/3UTR /G701T	--
1391	rs12329692	SLC19A1	E6/3UTR /G172A	--
1392	rs16948421	TYMS	E7/3UTR /G398A	--
1393	rs16950472	ABCC4	3DR//A75520G	--
1394	rs16956880	TP53	E11/3UTR /C205T	--
1395	rs17135042	APC	E16/3UTR /T1050C	--
1396	rs17225060	MSH2	E16/3UTR /A226G	--
1397	rs2230303	GPX3	E5/3UTR /T138G	--
1398	rs2735347	PTEN	E9/3UTR /T957G	--
1399	rs28382578	SMARCB1	5UR//G-3038A	--
1400	rs3176359	CDKN1A	E3/3UTR /G1165A	--
1401	rs3733326	PPAT	E11/3UTR /G284C	--
1402	rs3742106	ABCC4	3DR//A74234C	--
1403	rs3772810	UMPS	E6/3UTR /A28G	--
1404	rs397768	APC	E16/3UTR /G1753A	--
1405	rs4835	FXYD3	E9/3UTR /A173T	--
1406	rs699517	TYMS	E7/3UTR /C19T	--
1407	rs8177450	GPX3	E5/3UTR /A255G	--
1408	rs8177452	GPX3	E5/3UTR /A608G	--
1409	rs8192925	CES2	E12/3UTR/A69G	--
1410	rs873652	FGFR4	E16/3UTR /A26T	--
1411	rs9468385	GPX6	E5/3UTR /C457T	--
1412	rs9516521	ABCC4	3DR//T74903C	--
1413	rs9835477	SMARCC1	E28/3UTR /T412G	--
1414	rs10106	FPGS	E15/3UTR /T192C	--
1415	rs1016860	BCL2	3DR//C190113T	--
1416	rs1042710	UGT1A1	E5/3UTR /A3G	--
1417	rs1045411	HMGB1	E5/3UTR /C2262T	--
1418	rs1050569	GPX1	E1/3UTR /T695G	--
1419	rs10513202	SLC31A1	E5/3UTR /A2206G	--
1420	rs1051332	ATP7B	E17/3UTR /C1172T	--
1421	rs10513347	HLTF	E25/3UTR /T1491C	--
1422	rs10517	NQO1	E5/3UTR /A1119G	--
1423	rs1059316	GPX3	E5/3UTR /C713T	--

1424	rs10636	MT2A	E3/3UTR /G77C	--
1425	rs1065769	TK1	E7/3UTR /C105T	--
1426	rs10929303	UGT1A1	E5/3UTR /T211C	--
1427	rs11375183	SMARCC1	E28/3UTR /-898G	--
1428	rs1143245	SLC31A1	E5/3UTR /G3502C	--
1429	rs12010382	ATP7A	E23/3UTR /T1819C	--
1430	rs12189	APC	E16/3UTR /C434T	--
1431	rs12852	FXYD3	E9/3UTR /A818G	--
1432	rs13048427	SLC19A1	E6/3UTR /G23A	--
1433	rs1537514	MTHFR	E12/3UTR /G2669C	--
1434	rs16842633	UPP2	E9/3UTR /C962T	--
1435	rs16861315	HLTF	E25/3UTR /C1641T	--
1436	rs16948409	TYMS	E7/3UTR /G214T	--
1437	rs17225053	MSH2	E16/3UTR /T141G	--
1438	rs17387924	SMARCA2	E33/3UTR /G431A	--
1439	rs17511602	REV3L	E32/3UTR /T293C	--
1440	rs1787474	WDR7	E27/3UTR /T275C	--
1441	rs17880487	SOD1	E5/3UTR /C339T	--
1442	rs17881366	TP53	E11/3UTR /C328T	--
1443	rs1804447	SOD1	E5/3UTR /C2T	--
1444	rs2020906	MSH6	E10/3UTR /T85A	--
1445	rs2074954	UPP2	E9/3UTR /T123G	--
1446	rs2077360	MTHFR	E12/3UTR /A1858G	--
1447	rs2290736	SMARCD3	E13/3UTR /C65G	--
1448	rs2635723	DPYD	E23/3UTR /G466T	--
1449	rs2736630	PTEN	E9/3UTR /T1133C	--
1450	rs2790	TYMS	E7/3UTR /A89G	--
1451	rs28364275	ABCB1	E29/3UTR /A21G	--
1452	rs28364277	ABCB1	E29/3UTR /C146T	--
1453	rs28364279	ABCB1	E29/3UTR /T252G	--
1454	rs28364280	ABCB1	E29/3UTR /C316T	--
1455	rs28364281	ABCB1	E29/3UTR /A562G	--
1456	rs28364610	SMARCA5	5UR//G-130A	--
1457	rs28382662	POLM	E11/3UTR /T139C	--
1458	rs28382664	POLM	E11/3UTR /C676T	--
1459	rs3136391	SMUG1	E4/3UTR /A260G	--
1460	rs3136392	SMUG1	E4/3UTR /C334T	--
1461	rs3176358	CDKN1A	E3/3UTR /G385A	--
1462	rs3176753	XPA	E6/3UTR /A278G	--
1463	rs3176754	XPA	E6/3UTR /T464C	--
1464	rs3177111	GPX4	E7/3UTR /C140A	--
1465	rs3212116	XRCC3	E10/3UTR /A145G	--
1466	rs3212117	XRCC3	E10/3UTR /G448T	--
1467	rs3212125	XRCC3	E10/3UTR /A868C	--
1468	rs3212126	XRCC3	E10/3UTR /C1094T	--
1469	rs34874603	GPX6	E5/3UTR /A908G	--
1470	rs35091626	GPX7	E3/3UTR /A549T	--
1471	rs35481105	PTEN	E9/3UTR /A1697G	--
1472	rs35919705	TP53	E11/3UTR /G105A	--
1473	rs3744935	BCL2	3DR//C190507T	--
1474	rs3745030	WDR7	E27/3UTR /C2176G	--

1475	rs3745032	WDR7	E27/3UTR /C112T	--
1476	rs3745033	WDR7	E27/3UTR /C14T	--
1477	rs3749440	ABCC5	E7/3UTR /A633G	--
1478	rs3749444	ABCC5	3DR//C62512T	--
1479	rs3805114	ABCC5	3DR//T63819G	--
1480	rs3895070	PTEN	E9/3UTR /T320G	--
1481	rs4097504	HMGB1	E5/3UTR /T1167C	--
1482	rs41275468	MTHFR	E12/3UTR /G2594A	--
1483	rs41292780	ATP7B	E17/3UTR /C1385T	--
1484	rs41297348	ABCB1	E29/3UTR /A393G	--
1485	rs41483150	FPGS	E15/3UTR /A629C	--
1486	rs4148551	ABCC4	3DR//T74507C	--
1487	rs4148555	ABCC4	3DR//A75760T	--
1488	rs4148595	ABCC5	3DR//G63298A	--
1489	rs4150523	ERCC3	E15/3UTR /G29A	--
1490	rs4252658	ERBB2	E27/3UTR /C66T	--
1491	rs4253231	ERCC6	E21/3UTR /A53G	--
1492	rs448475	APC	E16/3UTR /C1556G	--
1493	rs45574135	MTHFR	E12/3UTR /C1070T	--
1494	rs45593641	NQO1	E5/3UTR /C380T	--
1495	rs45625835	MTHFR	E12/3UTR /C543T	--
1496	rs4968187	TP53	E11/3UTR /C485T	--
1497	rs4987843	BCL2	3DR//C189646T	--
1498	rs4987844	BCL2	3DR//T189914G	--
1499	rs4987845	BCL2	3DR//C189999T	--
1500	rs4987848	BCL2	3DR//C190271T	--
1501	rs4987850	BCL2	3DR//C190671T	--
1502	rs4987851	BCL2	3DR//C190928T	--
1503	rs4987852	BCL2	3DR//T191266C	--
1504	rs4987854	BCL2	3DR//C191561T	--
1505	rs4987856	BCL2	3DR//C191693T	--
1506	rs4987858	BCL2	3DR//A192393G	--
1507	rs4987860	BCL2	3DR//C192600T	--
1508	rs4987866	BCL2	3DR//T193866C	--
1509	rs4987867	BCL2	3DR//C194115T	--
1510	rs4987869	BCL2	3DR//T194457G	--
1511	rs6594650	APC	E16/3UTR /A1203C	--
1512	rs6650282	ABCC4	E21/3UTR /T168C	--
1513	rs6899628	POLH	E11/3UTR /C290T	--
1514	rs6919	SMARCD2	E13/3UTR /T813A	--
1515	rs701848	PTEN	E9/3UTR /T1516C	--
1516	rs7032466	SLC31A1	E5/3UTR /T2808G	--
1517	rs7048532	SMARCA2	E33/3UTR /T620C	--
1518	rs757412	SMARCE1	E11/3UTR /G110A	--
1519	rs7778745	POLM	E11/3UTR /A361G	--
1520	rs8026	SMARCAD1	E24/3UTR /A1500G	--
1521	rs8056468	NQO1	E5/3UTR /G776A	--
1522	rs8065799	TP53	E11/3UTR /T9G	--
1523	rs8330	UGT1A1	E5/3UTR /G440C	--
1524	rs8336	SMARCAD1	E24/3UTR /T925C	--
1525	rs868014	MTHFR	E12/3UTR /A1290G	--

1526	rs9516520	ABCC4	3DR//T75167C	--
1527	rs9535794	ATP7B	E17/3UTR /G1744A	--
1528	rs9590161	ABCC4	3DR//A75889G	--
1529	rs9680103	SLC19A1	E6/3UTR /C421A	--
1530	rs971	SMUG1	E4/3UTR /T422C	--
1531	rs10609062	UCK1	E7/3UTR/GTGA426-	--
1532	rs2266636	GSTT1	E/4/C354T	V118V
1533	rs28364609	SMARCA5	5UR//T-740C	--
1534	rs28606552	SMARCA5	5UR//T-3655A	--
1535	rs3218658	POLM	E/8/C1032T	A344A
1536	rs6836313	SMARCA5	5UR//A-811C	--

Supplementary Table 2: The 14 markers not suitable to be genotyped by GoldenGate array and genotyped by other methods

#	rs No	Gene	mRNA Location	AA Change
1	rs2853542	TYMS	E/1/G-58C	--
2	rs34743033	TYMS	28 bp VNTR in 5UTR	--
3	rs16430	TYMS	6 bp InDel in 3UTR	--
4	rs3786362	TYMS	E/3/A381G	I127I
5	rs1801159	DPYD	E/13/T1627C	I543V
6	rs2297595	DPYD	E/6/T496C	M166V
7	rs11479	TYMP	E/10/G1412A	S471L
8	rs9628204	TYMP	E/7/C787T	G263R
9	rs17851631	TYMP	E/5/G585T	D195E
10	rs28931613	TYMP	E/2/C131T	R44Q
11	rs3210145	TYMP	I/6/T-2A	--
12	rs470119	TYMP	I/4/T27C	--
13	rs1061205	TYMP	E/10/G1401A	F467F
14	rs131804	TYMP	E/8/G972A	A324A

Supplementary Table 3: The genes in each gene set for gene set enrichment analysis.

Gene Group	Gene
5-FU Activator	RRM1,UMPS,UPP2,CDA,CES2,UCK2,CES1,UPP1, RRM2,PPAT,UCK1,TYMP,TK1
5-FU PK	RRM1,UMPS,SLC29A1,ABCC4,UPP2,CDA,DPYS,TK1,DPYD,CES2, UCK2,ABCG2,CES1,UPP1,ABCC5,UPB1,RRM2,TYMP, ABCC3,PPAT,CYP2A6,UCK1,SLC22A7
5-FU PD	MTHFR,ERCC2,GGH,FPGS,TYMS,SMUG1,TP53, XRCC3,TDG,DHFR
Platinum pathway	POLH,ERCC6,SMARCA1,SMARCA2,SLC31A1,NQO1,ATP7A,POLB, ATP7B,SMARCD1,ABCC2,REV3L,MT2A,ERCC4,MT1A,XRCC1, ERCC2,SOD1,ATOX1,PMS2,SLCO6A1,GSTM1,SMARCAD1, MSH6,SMARCD3,ABCG2,GSTP1,ERCC1,SMARCA5,SMARCD2, SMARCAL1,XPA,POLM,SMARCB1,MSH2,SMARCA4, HLTF,MLH1,HMGB1,SMARCE1,SMARCC1,GSTT1, ERCC3,MPO,SMARCC2
5-FU and Platinum Pathway	POLH,ERCC6,SMARCA1,SMARCA2,SLC31A1,NQO1,ATP7A, POLB,ATP7B,SMARCD1,ABCC2,REV3L,MT2A,ERCC4,MT1A, XRCC1,ERCC2,SOD1,ATOX1,PMS2,SLCO6A1,GSTM1,SMARCAD1, MSH6,SMARCD3,ABCG2,GSTP1,ERCC1,SMARCA5,SMARCD2, SMARCAL1,XPA,POLM,SMARCB1,MSH2,SMARCA4,HLTF, MLH1,HMGB1,SMARCE1,SMARCC1,GSTT1,ERCC3,MPO, SMARCC2,RRM1,UMPS,SLC29A1,ABCC4,UPP2,CDA,DPYS, TK1,DPYD,CES2,UCK2,TYMS,CES1,UPP1,ABCC5,UPB1, RRM2,TYMP,ABCC3,PPAT,CYP2A6,UCK1,SLC22A7, MTHFR,GGH,FPGS,SMUG1,TP53,XRCC3,TDG,DHFR
Colorectal Cancer Associated	DLG5,ERCC6,UMPS,ERBB2,PTEN,CYP19A1,EXO1,ABCB1,APC,NQO1, MGMT,POLB,HIF1A,EGF,MTHFR,ERCC4,DPYD,XRCC1,CES2,CYP1B1, ERCC2,CDKN1A,IL1B,PMS2,GSTM1,TYMS,MSH6,CES1,TP53,GSTP1, FAS,UGT1A1,IL8,CDKN1B,ERCC1,TFRC,XPA,SHMT1,XRCC3,MSH2, RB1,GSTM3,PLD2,PTGS2,MLH1,XPC,ALDH2,NAT2,CASR,ARL11, CYP1A2,GSTT1,IL12A,CYP2A6,ERCC3,ERCC5,FGFR4 ,MMP9,ATM,BARD1,CDH1,CYBA,CYP2C9,CYP2E1,CYP7A1, MMP2,MTHFD1,NAT1,PARP1,VDR,UGT2B7,PMS1,OGG1 ,MLH3,GPX1,DCC,CYP3A4,CYP2C19,CDKN2A,BAX