University of Reading

Predictive tools for the study of variations in ADP platelet responses: Implications for personalised CVD risk and prevention strategies

A thesis submitted for the degree of Doctor of Philosophy

by

Bajuna Rashid Salehe

School of Biological Sciences
University of Reading

June 2017

Declaration

I confirm that this is my own work and the use of all materials from other sources have been properly and fully acknowledged.

Bajuna Rashid Salehe:

Date:

Abstract

The major aim of this project is to develop novel computational approaches for rapid identification of key omic variations, particularly SNPs that are likely to be associated with the variability of the ADP (Adenosine diphosphate) activated platelet responses. The ADP platelet response was chosen as a model system due to its distinct role during the platelet amplification and aggregation, and it is the main therapeutic target for cardiovascular disease (CVD) antiplatelet drug treatments. Based on recent studies, CVD is currently the second lethal non-communicable disease after cancer in both developed and developing countries. Inter-individual variability of the ADP platelet responses was previously reported in genetic association studies, and susceptible SNPs were identified. However, most of the standard biostatistical methods that were previously employed were found to be suboptimal, and it is assumed that other crucial SNPs might have been potentially missed. In genetics, this phenomenon is known as 'missing heritability' problem. Therefore, to address this issue, this study aims to employ alternative computational approaches in an integrated manner in order to identify previously unidentified key SNPs, which may underlie the ADP platelet responses variability. Additionally, the project aims to develop predictive approaches to unveil the molecular mechanisms of the identified key SNPs, which are likely to underpin the inter-individual variability in the ADP platelet responses and aggregation. The molecular mechanisms underpinning these SNPs, or 'omic variations are rarely addressed in standard genetic mapping or association studies. This may be due to the experimental hurdles related to the costs and labour that are required in pursuing such undertakings, hence our predictive approach seeks to address such inefficiencies in closing these knowledge gaps. Moreover, the project culminates in the development of a method for predicting an individuals' ADP platelet response levels with a focus on determining the extreme cases, i.e., individuals showing high and low responses to ADP platelet activation. Predicting ADP responses levels might be

suitable for determining which allelic features will contribute most to the extreme ADP platelet responses. This understanding may be useful for suggesting new drug targets or individualised treatments in the targeted CVD therapeutics or personalised medical settings for the next generation of medical practice.

Contents

List of Figures

List of Tables

Table of Abbreviations

3D – Three-dimension

ANN – Artificial neural network

ACS – Acute coronary syndrome

ADP – Adenosine diphosphate

CA domain – Catalytic domain

CASP – Critical assessment of protein structure prediction

CNV – Copy number variations

CPU – Central processing unit

CRP-XL - Collagen-related peptide

cSNPs – Coding SNPs

CVD – Cardiovascular disease

EMA – Estimate Model Accuracy

eQTL – Expressed quantitative trait loci

FA – Fibrinogen binding in response to ADP

FC – Fibrinogen binding in response to CRP-XL

FM – Free modelling

GAS – Genetic association studies

GWAS – Genome-wide association study

HLH – Helix-loop-helix

IC domain – intracellular

LASSO – Least absolute shrinkage and selection operator

LD – Linkage disequilibrium

LCR – Locus control regions

MAF – Minor allele frequency

MI – Myocardial Infarction

ML – Machine learning

MLP – Multilayer Perceptron

MQA – Model Quality Assessment

MQAPs – Model Quality Assessment Programs

OOB – Out-of-bag

OLS – Ordinary least square

PA – P-selectin expression in response to ADP

PC – P-selectin in response to CRP-XL

PCAST – President's Council of Advisors on Science Technology

PG – Proteoglycan

PM – Personalised Medicine

pQTL – Proteomic quantitative trait loci

PTM - Post-translational modification

QTL – Quantitative trait loci

RBP – RNA binding protein

RBPS – RNA binding protein site

RCSB/PDB – Research Collaboratory for Structural Bioinformatics/Protein Data Bank

RF – Random Forests

RR – Ridge regression

RC – R Core

rSNP – Regulatory SNP

SNP – Single nucleotide polymorphism

SD – Randomly SNPs

SR – RAPIDSNPs Set

SJ – Jones et al. SNPs

TBM – Template based modelling

TF – Transcription factor

TFBSs – Transcription factor binding sites

VI – Variable importance

WHO – World Health Organisation

Acknowledgements

I thank my Lord the Almighty creator of everything who helped me to do this work, mom for her moral support and encouragement, and supervisory committee lead by Liam McGuffin with other co-supervisors: Chris Jones and Giuseppe Di Fatta for their time and professional support. The special thank goes to the special person, my lovely wife Zamradi Salum for her great patience. Lastly, I thank all members of the McGuffin group.

This work was supported by Institute of Finance Management (IFM) under the Ministry of Finance of the Tanzanian Government.

Chapter 1 – Introduction and background

1.0 Abstract

The major aim of this project is to develop novel computational approaches for rapid identification of key omic variations, particularly SNPs that are likely to be associated with the variability of the ADP (Adenosine diphosphate) activated platelet responses. The ADP platelet response was chosen as a model system due to its distinct role during the platelet amplification and aggregation, and it is the main therapeutic target for cardiovascular disease (CVD) antiplatelet drug treatments. Based on recent studies, CVD is currently the second lethal non-communicable disease after cancer in both developed and developing countries. Inter-individual variability of the ADP platelet responses was previously reported in genetic association studies, and susceptible SNPs were identified. However, most of the standard biostatistical methods that were previously employed were found to be suboptimal, and it is assumed that other crucial SNPs might have been potentially missed. In genetics, this phenomenon is known as 'missing heritability' problem. Therefore, to further address this issue, this study aims to employ alternative computational approaches in an integrated manner in order to identify previously unidentified key SNPs, which may underlie the ADP platelet responses variability.

1.1 Introduction

This introductory chapter begins by giving an overview of CVD and the ADP platelet activation mechanism, which is one of the key physiological processes underlying major CVD events. The section also provides an outline of the genetic aspects of the ADP platelet responses by explicitly looking at the associated single nucleotide polymorphisms or SNPs. SNPs are considerable variations of single DNA bases among individuals and they are widely spread (i.e. ~ 90%) in the human genome (Collins et al., 1998). A significant number of these SNPs are disease/trait associated (Shastry, 2002). Furthermore, the section briefly describes other SNPs

associated ‘omic variations and highlights different approaches that are applied in the identification of the disease/trait associated SNPs. In this case, the focus is on the standard methods used in genetic association studies (GASs). Following this, the integrated computational approaches developed in this study, which are an alternative to the standard biostatistical methods, are discussed. Additionally, the concept of personalised medicine, which utilises individual genomic information for disease treatment is briefly discussed. Finally, this section describes the overall objectives and conceptual detail of the study using the computational framework and integrated predictive pipelines for ‘omic variation analyses.

1.2 CVDs and their types

CVDs are a variety of disorders affecting the heart and blood vessels. These include: coronary heart disease, cerebrovascular disease, peripheral arterial disease, rheumatic heart disease, congenital heart disease, deep vein thrombosis, pulmonary embolism, myocardial infarction and strokes (“WHO | Cardiovascular diseases (CVDs),” 2016). This study broadly entails the alternative approaches in further addressing CVDs and their associated risks.

1.2.1 CVD in developing and developed countries

Cardiovascular diseases are the leading cause of death worldwide. There are an estimated 18 million deaths from CVDs each year, which accounts for 33% of the 55 million total deaths, and 75% of these are from coronary heart disease and stroke (Stanner, 2008; “WHO | Cardiovascular diseases (CVDs),” 2016). Recently published data from the World Health Organisation (WHO) estimated that there are 7.5 million deaths worldwide due to the hypertension (“WHO | Raised blood pressure,” 2016). In the United Kingdom CVDs are the leading cause of death after cancer (British Heart Foundation, 2015; Stanner, 2008). Figures 1.1 and 1.2 shows the CVD death statistics in the UK relative to other disease problems for men and women under 75 respectively.

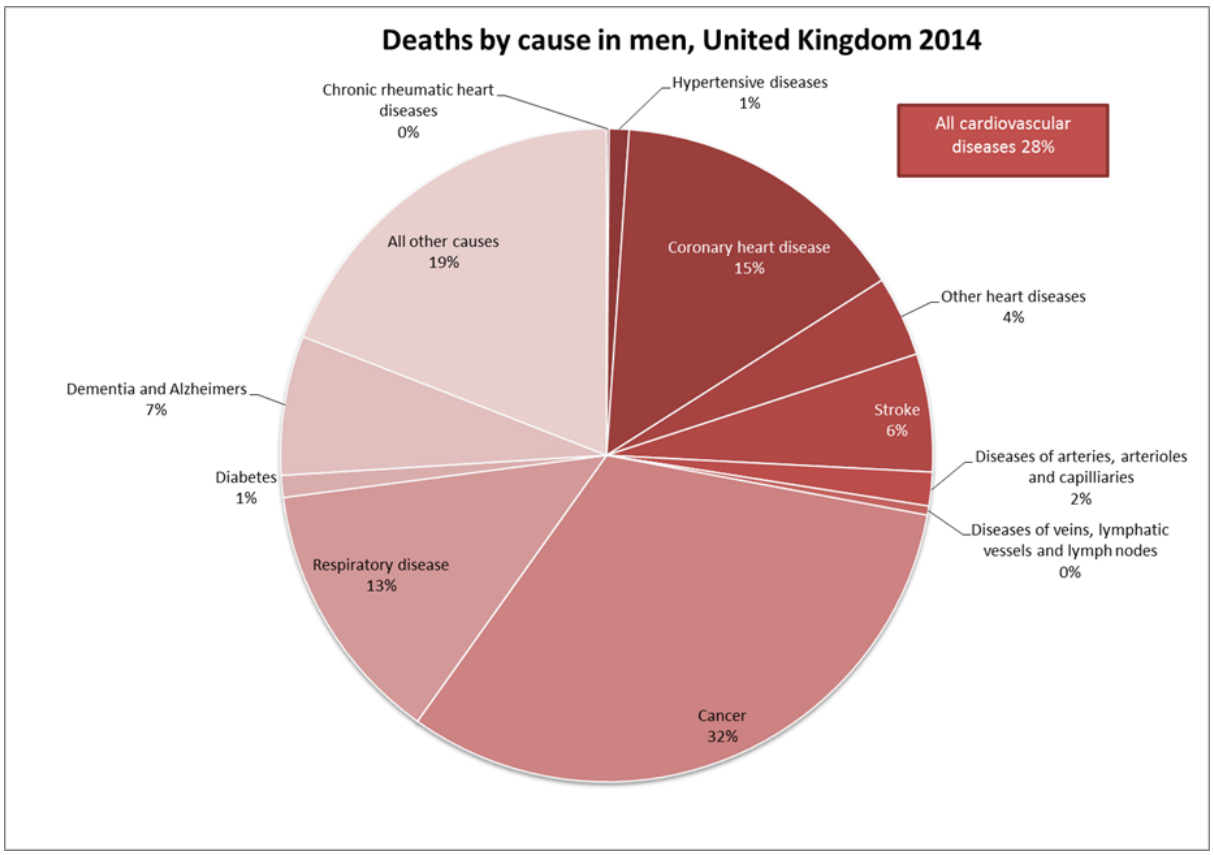


Figure 1.1 The recent statistical data showing the mortality rate caused by the CVD related problems for men in the UK. It is clear that CVD is the second deadly non-communicable disease after cancer, which kills men. (The Figure was taken from British heart foundation (British Heart Foundation, 2015))

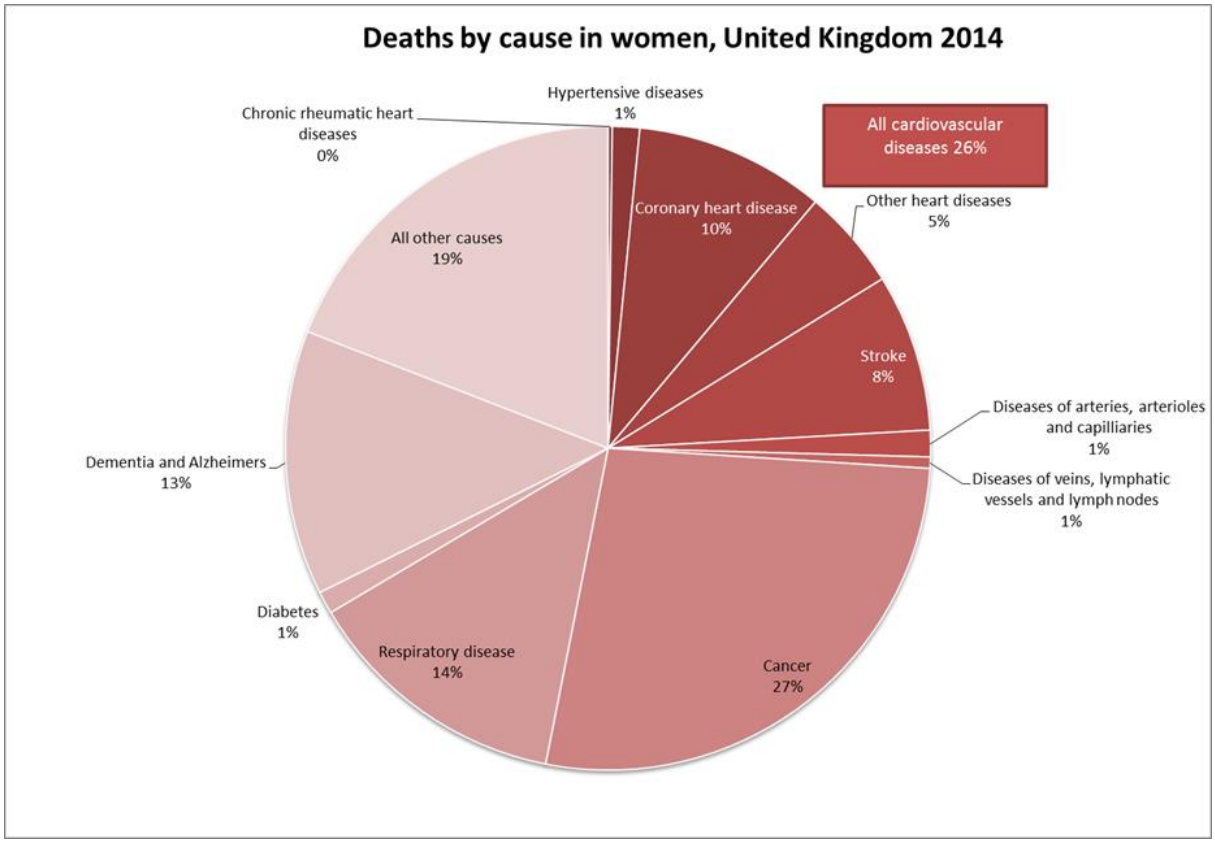


Figure 1.2 The recent statistical data showing the mortality rate caused by the CVD related problems for women in the UK. A similar trend of mortality rate for men can be also observed in the women's fatality rate caused by CVD. Nevertheless, it appears that women have less mortality rate due to CVD than men. (The Figure was taken from British heart foundation (British Heart Foundation, 2015))

In the developing countries particularly in Sub-Saharan Africa, CVDs are predominant emerging non-communicable diseases (NCD), termed as 'silent killers' with the combination of risks factors (Ouyang, 2014). It has been estimated that one in two people whose age is 25 years and above has undiagnosed hypertension in Sub-Saharan Africa and three-quarters of the CVD deaths are from the low- and middle-income countries (Ouyang, 2014; "WHO | Cardiovascular diseases (CVDs)," 2016). In developing countries, CVD is reported to account for nearly 40 percent of all deaths, which is higher than in the UK (Mbewu and Mbanya, 2006). Figure 1.3 shows the estimated increase of CVD deaths for male and female in Africa.

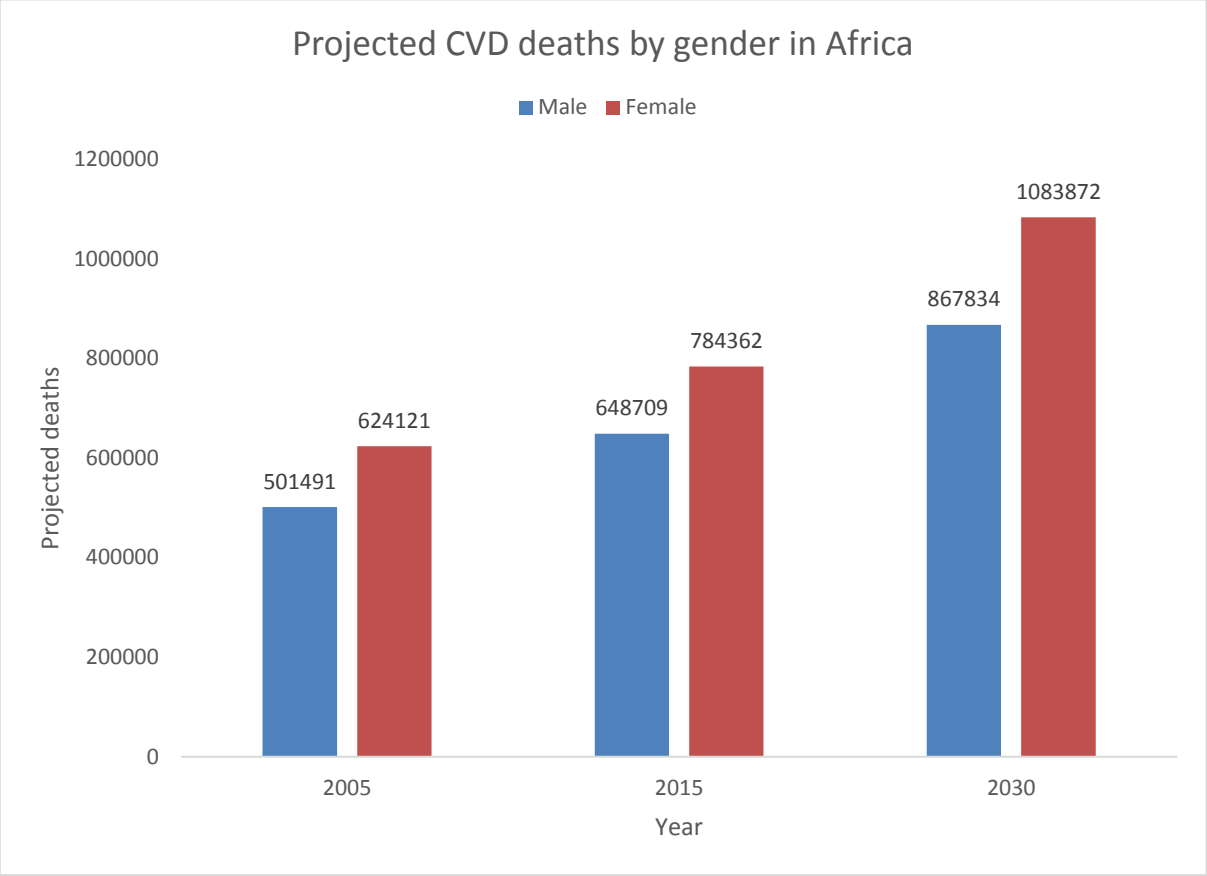


Figure 1.3 The estimated deaths caused by CVD in Africa by 2030. The trend as it can be observed is that the mortality rate due to CVD is yearly increasing for both genders with higher rate in women. (The image was taken from (WHO | Department of Measurement and Health Information, 2006))

Therefore, from the above data, the deaths contributed from CVD seems to be relative high for both developed and developing countries in Africa. Thus, efforts have been stepped up further, particularly in scientific research aimed at understanding, identifying, and minimising the risks associated with CVD problems (Craddock et al., 2010; Keating et al., 2016; Ouyang, 2014; Reddy and Yusuf, 1998; Stanner, 2008; Vizioli et al., 2009; "WHO | Cardiovascular diseases (CVDs)," 2016).

1.2.2 Major risk factors for CVDs

Many of the CVD risk factors are non-genetic and are related lifestyle factors such as diet, tobacco and alcohol use (Banerjee, 2012; Berry et al., 2012). One major risk factor includes the nature of blood clots within the blood vessels, which may block the flow of the blood to the

heart and brain. (Stanner, 2008; “WHO | Cardiovascular diseases (CVDs),” 2016). The mechanism of blood clotting is controlled by the platelet activity in the blood, which has a strong genetic association (Lewis et al., 2013; Williams et al., 2010). Therefore, understanding the genetic basis underpinning platelet function is vital for further elucidating the genetic risk factors that are -associated with CVD (Kvasnicka et al., 2015; Lewis et al., 2013).

1.3 Platelet activation and responses

Platelets are small anucleate cells packed with complex signalling machinery that enables them to react rapidly to damage in blood vessels to prevent blood loss. Platelets are formed from the megakaryocytes in the bone marrow. There are approximately one thousand billion platelets circulating in the human body, continually screening the vascular endothelium for biochemical signals of injury.

Platelets are multifunctional in nature and are associated with several pathophysiological processes including; thrombosis, clot retraction, vessel constriction and repair, inflammation including promotion of atherosclerosis, host defence and even tumour growth (metastasis) (Harrison, 2005). Figure 1.4 depicts the multifunctional nature of the blood platelet.

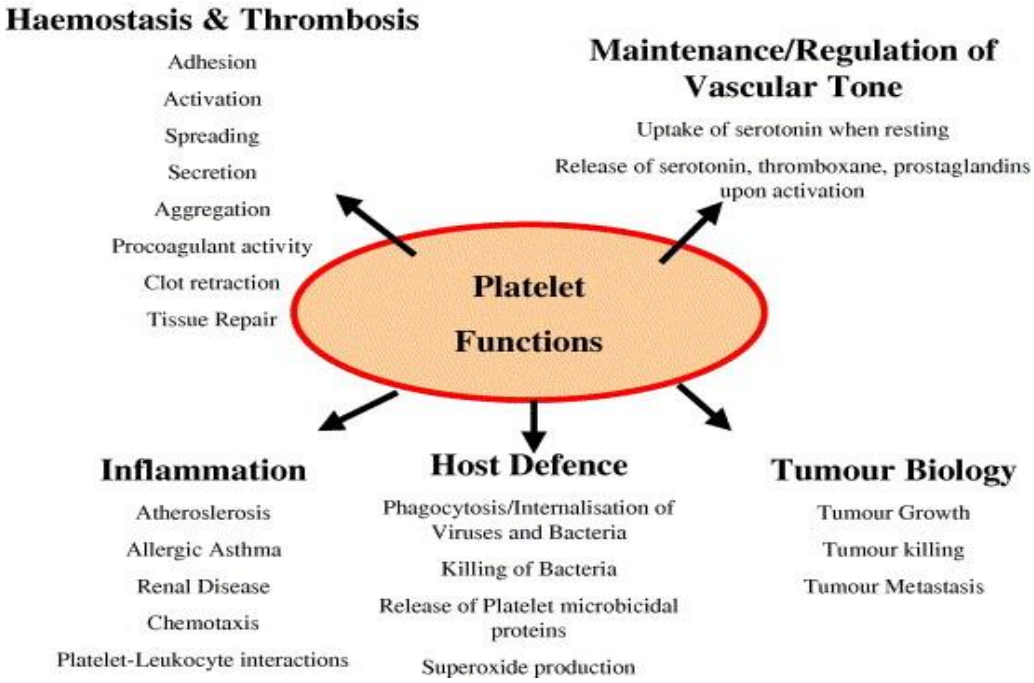


Figure 1.4 The multifunctional nature of the platelet in different pathophysiological processes. The haemostasis & thrombosis and inflammation, which are more inclined to CVD problems are the focus of this study. (The image was taken from (Harrison, 2005)

This study mainly focuses on haemostasis and thrombus formation, and in particular the latter, which has been associated with several CVDs (Rauch et al., 2001; Stanner, 2008).

Platelets must be activated to form a clot, or thrombus, and there are many agonists (small molecules that activate receptors) involved in the activation process. These include thrombin, collagen, ADP (Adenosine Diphosphate), Thromboxane A_2, adrenaline and serotonin, some of which act together *in vivo* with other agonists (Gibbins, 2004; Jackson et al., 2003; Rivera et al., 2009). Platelet activation is a result of the attachment of these agonists to the receptors of the platelet's plasma membrane and von Willebrand factor (VWF) and collagen in the subendothelium. Figure 1.5 shows the key agonists involved for activating platelets.

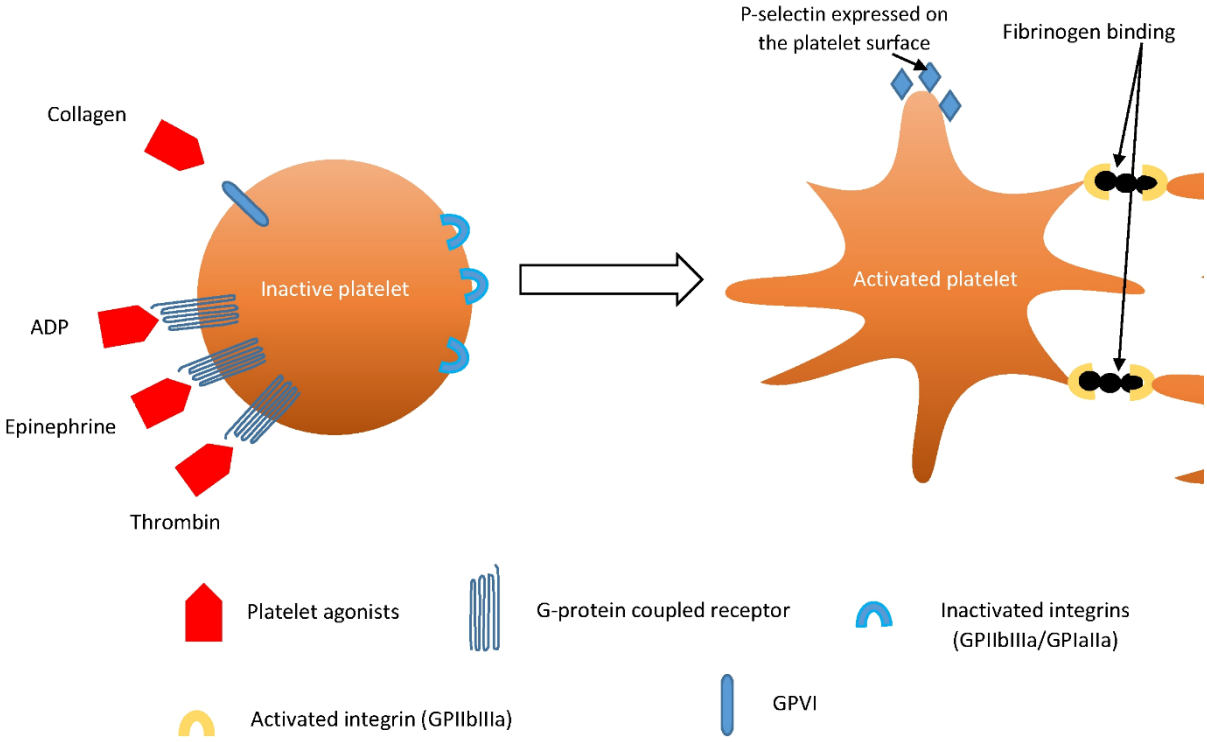


Figure 1.5 Platelet before and after activation by different agonists, with their involved receptors. There are many agonists, which are required in activating the platelet depending on the involved activation stage. ADP which is the core agonist in the secondary activation stage of the platelet is the main focus of this study. Similarly, different receptors are involved in different stages of activation. The most important of these receptors are glycoproteins or integrins, which mediate collagen and fibrinogen binding during the primary adhesion and the follow-up platelet aggregation for eventual thrombus formation.

1.3.1 A brief description of platelet activation, responses and thrombus formation

The general platelet activation mechanism for thrombus formation can be explained as follows. At sites of vascular damage, platelets adhere to the sub-endothelial matrix (Alevriadou et al., 1993; Ruggeri, 2003; Wu et al., 2000). These first adherent platelets are stabilised and activated by the binding of GPVI and integrin $\alpha_2\beta_1$ to exposed collagens (Siljander et al., 2004), Figure 1.6. Following this initial deposition, subsequent encountering platelets are activated by a host of other agonists, which are generated and secreted by the activated platelets. Such agonists are ADP released upon platelet degranulation (Gachet et al., 1997), Thromboxane A_2 (Siess et al., 1983a, 1983b), synthesised on the platelet surface from arachidonic acid, or thrombin (Bevers et al., 1982; Coughlin, 2000), activated from prothrombin on the negatively charged surface of activated platelets.

Intracellular signalling cascades, which are initiated by the interaction of agonists with specific cell surface receptors lead to: calcium mobilisation (from both internal stores and the extracellular space into the cytoplasm), platelet shape change, degranulation and a change in the affinity of integrin $\alpha_{IIb}\beta_3$ for VWF and fibrinogen (Fg) binding (Coppinger et al., 2004; Hartwig, 2006; Italiano Jr et al., 2008; Ma et al., 2007; Varga-Szabo et al., 2009). The binding of fibrinogen to integrin $\alpha_{IIb}\beta_3$ on different platelets supports their aggregation and thrombus formation (Bennett, 2001; Pytela et al., 1986). Figure 1.6 further shows the underlying activation process.

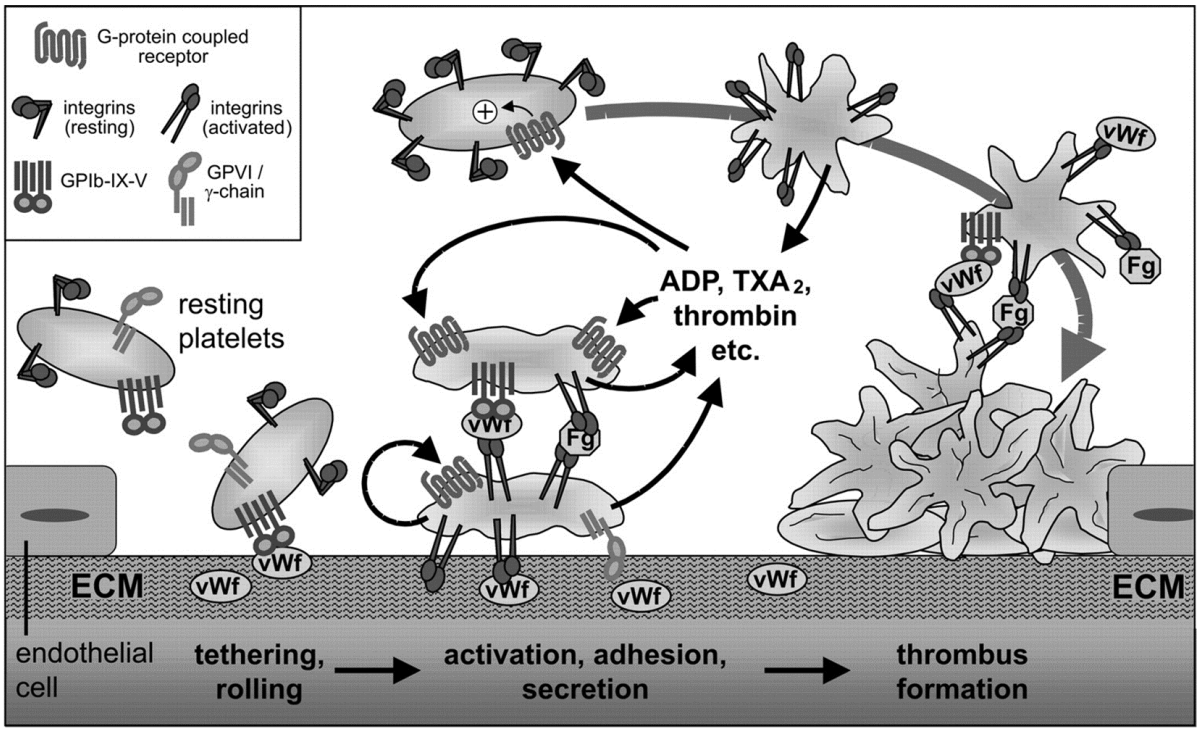


Figure 1.6 The underlying platelet aggregation process and thrombus formation. Three main phases are involved. The tethering and rolling are when the platelets encounter the exposed collagen and VWF in the extracellular matrix (ECM), which results in the primary activation through GPVI channel. The secondary adhesion, secretion, and aggregation is the next phase before the thrombus is formed. The latter involve rapid platelet aggregation through ADP channel in the amplification process to increase the size of the thrombus. TXA_2 is Thromboxane A_2 (Image was taken from Spiel et al., 2008).

1.3.2 Platelet activities and CVDs

The process of platelet aggregation and thrombus formation is highly regulated. Abnormal platelet functions have been associated with many blood and CVD related problems including

stroke, peripheral vascular disease, and diabetic mellitus (Michelson, 2004). Reduced or slow thrombus formation results in bleeding, whereas larger thrombi that occlude vessels and block the flow of blood can lead to CVD events, such as myocardial infarction or stroke (Zee et al., 2008). In addition, the growth in thrombus size due to platelet aggregation is one of the principal causes of ischemic stroke (Viles-Gonzalez et al., 2004).

1.3.3 Why focusing on the ADP platelet responses?

1.3.3.1 ***ADP ultimately governs platelet aggregation and thrombus formation***

Studies have found that ADP plays a distinct role in the amplification of platelet activation and subsequent aggregation *in vivo* (Fontana et al., 2003; Jin et al., 2002). ADP is involved in the activation of platelet through two major G protein-coupled receptors (GPCRs): P2Y1 and P2Y12 (Hollopeter et al., 2001; Zhang et al., 2001). The P2Y1 receptor activates phospholipase C, which is involved with changes in platelet shape and an increase of the intracellular Ca^{2+} store required for platelet aggregation. P2Y12 suppresses cAMP formation, which is necessary for activation to occur since it inhibits platelet aggregation (Noé et al., 2010; Woulfe et al., 2001). Moreover, the platelet activation by ADP mediates the function and release of other agonists, particularly TXA_2, which further increases the platelet aggregation (Jin et al., 2002).

1.3.3.2 ***ADP platelet responses are the key focus for CVD clinical treatments***

ADP is one of the key molecules, which is widely used in clinical investigations and platelet responses to ADP are a treatment focus for CVD patients. For instance, the severity of CVD conditions was assessed by examining the increase of the platelet aggregation before and after acute myocardial infarction attack, which was monitored through ADP concentration (Miyamoto et al., 2000). Moreover, the effectiveness of acute coronary syndrome (ACS) treatment among patients was assessed by examining platelet aggregation through fibrinogen

binding in response to ADP platelet activation. Patients treated with low concentration of orbofiban (GPIIbIIIa antagonist) (suboptimal dose) enhances platelet aggregation in the presence of ADP, while those treated with high concentration of orbofiban, platelet aggregation appeared to be decreasing (Cox et al., 2000).

Furthermore, several anti-platelet aggregation agents such as aspirin, are widely used as CVD treatments and tend to work efficiently in dose combination with other agents such as Ticlopidine and Clopidogrel (Albers et al., 2001). These are used to target and inhibit the activity of the P2Y12, which is an ADP receptor (Woulfe et al., 2001). In principle, aspirin blocks COX-1 and thus, prevent synthesis of thromboxane, while thienopyridine drugs inhibits the activity of P2Y12. However, in most cases, these drugs tend only to work effectively with particular groups of patients while showing low efficacy in others (Hollopeter et al., 2001). It was suggested that this was likely to be due to genetic factors, which may underlie the variability of the disease pathophysiology and responses to treatment amongst individual patients (Fontana et al., 2003).

1.4 'Omic' variations

1.4.1 Types of 'omic' variations

1.4.1.1 ***DNA sequence variations (genomic) level***

Since the completion of the human genome project (Consortium, 2004), a central focus of researchers in biology and medicine has been to understand the existing relationship between genotype and phenotype (Consortium, 2005). Although the human reference genome sequencing has been completed, the need for systematic investigating and understanding of human DNA sequence variations in the genome has received much attention through different projects such as HapMap (Consortium, 2005), and 1000 genomes (Consortium, 2012, 2010).

This is because human genome variations play a major role in the differences between individuals in terms of their disease states and other traits (Pevsner, 2009). Despite some successes of GASs in identifying thousands of genomic regions in the DNA associated with disease states and other traits using the data from these projects, a deeper understanding of the genomic variations and their association with phenotype is still needed (Consortium, 2010).

At the genome level there are different categories of variations, which include non-functional and functional associated variations, ranging from, but not limited to, SNPs, small insertions and deletions, copy number variations (CNVs) and non-coding RNAs (Bhartiya and Scaria, 2016; Collins et al., 1998; Djebali et al., 2012; Iafrate et al., 2004; McCarroll et al., 2008; Sebat et al., 2004; Wong et al., 2007). Among these variations, SNPs are arguably the most prevalent and associated with many complex diseases and traits, which include platelet responses and CVD in general (Brookes, 1999; Burton et al., 2007; Johnson et al., 2010; Wang and Moult, 2001). The next section gives details of the underlying representation of SNPs in GASs and then highlights how GASs infer the SNPs' association to the diseases based on this representation.

1.4.1.1.1 Single nucleotide polymorphisms (SNPs)

SNPs represent most of the human genome variations (Pevsner, 2009), and are attributed to the most complex diseases and traits (Craddock et al., 2010). From the genomic perspective, a SNP refers to the DNA sequence variation where one base changes to another, and there are reportedly 15 million recognised SNPs in the human genome (Riancho, 2012). Consider the following SNP (rs41306982) of the *GP6* gene (from dbSNP (Sherry et al., 2001), an Entrez database of SNPs):

CTGGGGAGGTCCCCACACCTGCCTA[A/G]GAGCTGGGGAGCTTTTTGGCTGTAT

This means that the DNA base Adenine (A) is substituted by Guanine (G), which conventionally means there are two variants (A or G) in the population in this red-coloured position of the genome. Both may also be represented as A or B for the major (dominant), or minor (recessive) alleles in the population respectively, and therefore, for a diploid human individual a genotype call would be AA or BB (homozygous) or AB (heterozygous). Understanding the individual SNPs genotype may provide a clue for assessing susceptibility to a particular disease or trait (Masood, 1999).

1.4.1.1.1.1 Data structure for the SNPs

SNPs can be represented in different structures and formats using different platforms (Danecek et al., 2011; Riva and Kohane, 2004; Sherry et al., 2001). However, for inferring genotype-phenotype in the association studies, SNPs are generally represented in the matrix form using tabular structure (Zhang et al., 2006). Individual's samples are represented in rows while the columns are divided in two parts; one part stores the SNPs genotypes and the other part, the phenotype.

Table 1.1 shows this representation of the data:

	SNP1	SNP2	…	SNP*p*	Phenotype1	Phenotype2
Subject 1	A/T	C/C	…	G/A	0.444	Y
…	…	…	…	…	…	
Subject n	A/A	G/C	…	G/G	0.1234	N

Table 1.1 Representation of SNP data sets. The genotype call for each SNP is represented in the column. For example, SNP1 for each subject in a population may be either A/T, A/A, or T/T for heterozygote, homozygous dominant, or recessive respectively. A certain combination of SNPs across columns involving one subject and a single base may form a haplotype in the genomic sequence of an individual subject. Example, a haplotype 1 for subject1 may contain SNP1, SNP2, and SNPp with nucleotides A, C, and G. Phenotype1 is for quantitative trait (QT) and Phenotype2 for case control studies.

Each individual SNP genotype for particular subject is derived from the two sequence reads of a pair of chromosomes:

SNP1

Chromosome1: ACCGTTT**A**GGGTTA

Chromosome2: ACCGTTT**T**GGGTTA

Subject 1 genotype: {A/T}

Subject 1 genotype: {A/T} (heterozygous allele)

Chromosome1: ACCGTTT**A**GGGTTA

Chromosome2: ACCGTTT**A**GGGTTA

Subject 2 genotype: {A/A} (major allele)

Chromosome1: ACCGTTT**T**GGGTTA

Chromosome2: ACCGTTT**T**GGGTTA

Subject 3 genotype: {T/T} (minor allele)

For smooth processing in the statistical and computational methods used, these genotypes are transformed into a numeric representation. For each SNP, major (dominant), heterozygous, or minor (recessive) alleles, are conventionally represented as 1, 2, or 3 respectively, or 0, 1, or 2 respectively (Zhang et al., 2006).

Thus, Table 1.1 would now become represented as follows, Table 1.2:

	SNP1	SNP2	…	SNP*p*	Phenotype1	Phenotype2
Subject1	3	1	…	3	0.444	Y
…	…	…	…	…	…	
Subject n	1	3	…	2	0.1234	N

Table 1.2. Numeric representation of SNP data sets. An alternative way of representing SNPs by genotyping using dummy variables in numeric form.

From the above tables, a typical genotype-phenotype association task will then be as follows: given genotypes and phenotypes of individuals in a population, identify which genomic positions or SNP combinations are associated with a particular phenotype. For identification

and mapping of these genomic positions to the phenotype, a quantitative trait loci (QTLs) analysis is applied (Miles et al., 2008). In QTL analysis, multiple SNPs may come from multiple loci (genes) and the association study may aim to quantify candidate genes (multiple loci) contributing to the overall continuous phenotypic effect, e.g. phenotype1 in the Table 1.2 above.

In most GASs, the number of SNPs (p) is larger than the number of subjects (n), since few individuals in a sample may give thousands to tens of thousands of SNPs. This phenomenon typically leads to the large p and small n feature selection problem, i.e. $p > n$ (Ayers and Cordell, 2010; Hastie et al., 2005; Saeys et al., 2007; Touw et al., 2013).

Based on the above representation, the central focus of the initial part of the study (Chapter 2) is to design an alternative approach, which rapidly identifies potential and previously unidentified key SNPs that may contribute quantitatively to the variability of the ADP activated platelet responses phenotype.

1.4.1.1.2 Other types of genomic variations

There are other types of variations at the genome level, which are associated with diseases/complex traits and subject to ongoing research. These are structural variants, which mainly include, but are not limited to, copy number variations (CNV) (Sebat et al., 2004; Wong et al., 2007), and DNA methylation (Jones and Takai, 2001; Mikeska and Craig, 2014; Moore et al., 2013). These variants were elsewhere reported to be associated with complex diseases, including CVD (Mikeska and Craig, 2014; Myocardial Infarction Genetics Consortium et al., 2009; Stankiewicz and Lupski, 2010). However, they are not well characterised and extensively studied in comparison with SNPs (Stankiewicz and Lupski, 2010). Nevertheless, DNA methylation has gained considerable attention in understanding its disease associated epigenetic

mechanisms, and there are several emerging computational approaches for curating and identifying these epigenetic variations (Balaur et al., 2016; Pancione et al., 2012; Roznovăţ and Ruskin, 2013). For future related studies, it would be worth examining the epigenetic features, such as methylation, that may underlie ADP platelet responses and CVD.

1.4.1.2 ***Protein sequence (Proteomic) variations (Amino acid substitutions)***

These are variations, which are related to SNPs, that occur in the exon region of the gene or coding regions (formally known as cSNPs) and they are more likely to affect gene function and cause individual phenotypic variations and disease (Collins et al., 1998). It has been reported that nearly 50% of cSNPs are missense mutations (missense or non-synonymous SNPs), meaning they lead to the codons that code for different amino acids (aa). These genomic changes may result in the allele-specific variations in the corresponding protein structure (Cargill et al., 1999; Collins et al., 1998; Wang and Moult, 2001). In this case, the cSNPs are regarded as deleterious (damaging), which may result in the possible functional differences of the associated proteins (Chen et al., 2010; Fazel-Najafabadi et al., 2015; Shi et al., 2012; Shukla and Mishra, 2011). Several of these missense mutations have been suggested to cause the variability of the individual traits, responses to drugs, and susceptibility to disease (Cargill et al., 1999; Flaherty, 2007; Pal and Moult, 2015; Shi and Moult, 2011; Wang and Moult, 2001). Moreover, there are several of these cSNPs that have been reported to be associated with CVD (Ohnishi et al., 2000; Okuda et al., 2002).

Therefore, identification of these deleterious protein sequence variations is of high importance (Burke et al., 2007). Thus, a compendium of computational methods to predictably identify these missense mutations are under continuous development (Gnad et al., 2013; Teng et al., 2008). Furthermore, to understand the potential molecular mechanisms of the identified deleterious SNPs (cSNPs) that are associated with the ADP platelet responses, new predictive

approaches will be described, which exploit several cutting-edge structural and functional protein bioinformatics tools (Chapter 3).

1.4.1.3 ***Transcriptomic variations***

Transcriptomic variations involve inter-individual variations in differentially expressed genes, in specific cells, or cell lines in which the mRNA transcripts are produced (Djebali et al., 2012). In this case, the variations might be due to the presence of SNPs in the regulatory regions of genes, which can significantly lead to differences in the cellular mRNA transcript levels (Stepanova et al., 2006). For instance, the SNPs might be localised in the binding sites (BS) of various transcription factors (TFs; TFBS), and lead to a possible functional effect on gene transcription regulation (Bryzgalov et al., 2013). The regulatory effect might be as a result of the increase or decrease in the binding specificity of the TFs, which may lead to allele-specific gene expression (Alj et al., 2004). In turn, this may lead to the differential trait/disease phenotypic states and drug responses among individuals (Chhibber et al., 2016; Drachkova et al., 2011).

Therefore, identification of the SNPs, which lead to the individual transcriptomic variations is vital in understanding the underlying molecular mechanisms of complex traits and diseases. Hence, efforts for designing experimental and computational methods for determining and predicting these regulatory SNPs have been stepped up (Andersen et al., 2008; Djebali et al., 2012; Hanson et al., 2015; Meyniel et al., 2010; Wan et al., 2014). Thus, in realising the potential of the transcriptomic variation, a subsequent objective of this study is to design a computational methodology or protocol, which predictably identifies how likely the SNPs may lead to the differential transcriptomic variations (Chapter 4). Doing this may further provide a clue on the variability of the ADP platelet responses and aggregation.

1.4.1.4 ***Phenotype variations (Phenomic) level***

This may refer to the observed variability of the complex trait/disease phenotype among individuals, which might be due to the genomic, proteomic, and transcriptomic variations (Moyra, 2011; Weischenfeldt et al., 2013). The key phenotypic variation, which is the focus of this study, is the variability of the ADP platelet responses and aggregation among individuals.

1.4.2 'Omic' variations, platelet activities, and CVDs

An individual's platelet response to agonists shows a high level of heritability and is highly variable within the population. Several experimental studies have sought to understand the role of 'omic variations in the underlying platelet functioning, and considerable success has been achieved.

In this case, for the genomic variations, numerous studies were able to explain in detail and illuminate the genetic determinants that underpin platelet responses. These studies include those which examined and identified genetic variant(s) in a single gene, as well as those which performed genetic mapping involving many genes and genome-wide association (GWAS) (Herrera-Galeano et al., 2008; Johnson et al., 2010; Jones et al., 2009; Soranzo et al., 2009). However, as it is later described, there are reported inherent flaws in the underlying methods in these studies, among other things, which have led to unaccounted crucial genetic variants (Moore et al., 2010).

Furthermore, notable studies have been performed in examining the effects of proteomic variations, i.e. the underlying effects caused by the missense mutations leading to individual protein structural and functional perturbations that may alter platelet activities. For instance, the point mutation Gly > Ser at position 233 (G233S) in GPIbalpha protein was identified to be deleterious by affecting the protein's binding affinity to VWF in allele-specific manner. This

was suggested to be the likely cause of platelet-von Willebrand factor disease (VWD) (Matsubara et al., 2003). In addition, a missense mutation (4115T>G) in the *VWF* gene, which leads to the substitution of an isoleucine with serine at position 1372 (I1732S) of VWF protein, was linked with the increased platelet aggregation (Casonato et al., 2007). In a similar study, a missense mutation (4263C > G) in the VWF, which causes N1421K mutation in the VWF domain was reported to be affecting the binding affinity of VWF with GPIb leading to the differential platelet aggregation (Lanke et al., 2008). Another related study that examined the effects of these mutations on the different proteins associated with the platelet dysfunction was performed by Lozano et al. (Lozano et al., 2016). However, each of these experimental studies are expensive in terms of equipment and consumables, which is compounded with the costs in labour and time, leading to inefficiencies in identifying novel molecular mechanisms associated with genetic changes (Yue and Moult, 2006). Therefore, as the platelet contains >5000 proteins, alternative/parallel computational efforts to rapidly and confidently identify each of the structural and functional variations in these proteins will be crucial to further illuminate our understanding of platelet responses and aggregation, and CVD pathogenesis (Boyanova et al., 2012; Burkhart et al., 2014; Vélez and García, 2015).

Furthermore, in the case of transcriptomic variation, numerous studies have been undertaken to analyse the effect of differential transcript levels on platelet activation and responses. For instance, Goodall et al. performed transcriptomic analyses to examine the SNPs that are involved in the regulation of variation of gene expression levels, which were associated with the ADP and collagen-related peptide CRP-XL. They identified crucial regulatory genomic regions, which influence the variability of the gene expression levels and are also associated with the myocardial infarction (Goodall et al., 2010). Moreover, other platelet dysfunctions were found to be associated with variability in regulatory transcriptomic regions among

individuals with risk of cardiovascular disease (Glembotsky et al., 2014; Lood et al., 2010; Raghavachari et al., 2007; Sun et al., 2007).

Therefore, it is clear that we must develop a deeper understanding of the regulatory SNPs underpinning the differential transcription and gene expression profiles that likely contribute to the variability of the platelet responses and aggregation. However, the inherent labour costs accompanied with the required resources make these studies expensive. Hence, alternative bioinformatics-based approaches which exploit the current computational advances and 'omic data availability, will likely drive forward our understanding of the differential transcriptomic variations underpinning inter-individual ADP platelet responses.

1.4.2.1 ***Genetic association studies (GASs) involving ADP platelet responses***

Genetic association studies (GASs) allow scientists to study and analyse SNPs that are associated with complex traits/diseases. Essentially, GASs aim at identifying the SNPs that are likely to be predisposing the trait or disease at a single gene or many genes, either in the specific genomic region, or on a genome-wide scale (Cordell and Clayton, 2005; Riancho, 2012). The SNPs determined from these studies are thought to be the basis of the associated variability at the molecular levels, i.e. proteomic and transcriptomic levels, which in turn may bring about the phenotypic variability (Pal and Moult, 2015; Wang and Moult, 2001).

The inter-individual variability of platelet responses and aggregation in the population is mainly heritable and associated with the predisposing genetic variant(s) or SNPs, which might be determined through genetic mapping or GASs (Herrera-Galeano et al., 2008; Soranzo et al., 2009). Previous GASs on the ADP platelet response were gene specific, such as those focused on *P2Y12* (Cavallari et al., 2007; Fontana et al., 2003). Cavallari et al. further identified key variants in this gene that were susceptible to low response to antiplatelet drugs and potential

risk of the atherothrombosis. Other studies investigated the effect of SNPs in the *PEAR1* gene on ADP platelet aggregation (Faraday et al., 2011; Herrera-Galeano et al., 2008; Lewis et al., 2013). However, the prevailing theory is that for the complex disease/trait, the predisposing associated SNPs in a single gene are interdependent with others in different genes or loci (i.e., polygenic) (Riancho, 2012; Robinson et al., 2014).

Thus, from polygenic aspects of the ADP platelet responses, there are GASs, which examined linkage disequilibrium (LD) of many genes or loci spanning either a particular region or at the genome-wide level (Goodall et al., 2010; Johnson et al., 2010; Jones et al., 2009). In this regard, Jones et al. analysed two ADP platelet responses, which were P-selectin exposure (a marker of degranulation) in response to ADP agonist (PA), and fibrinogen binding to integrin in response to ADP (FA) (Jones et al., 2009). They identified several significant SNPs that are associated with PA and FA platelet responses. However, it is argued that the employed biostatistical methods in most of these GASs are suboptimal and are not able to explain all genetic variants (Moore et al., 2010; Robinson et al., 2014). Thus, holistic approaches incorporating current advances in computational methods and bioinformatics are required for further explanation of unaccounted genetic variability (Moore et al., 2010).

Therefore, this study extends Jones *et al*., (2009) work further by alternatively investigating potentially unaccounted genetic variability (SNPs) and their roles underpinning ADP platelet responses variability using integrated computational approaches. The premise is that these approaches might become useful additional tools to aid future CVD personalised medicine.

1.5 Approaches for identifying trait/disease associated SNPs

The traditional approach for the most polygenic GASs is to analyse one SNP at a time, which normally fails to account for the effects of other causal SNPs (Hoggart et al., 2008). Besides,

as the number of SNPs keeps increasing, the single SNP analysis approach, using individual test statistics is prone to the risk of selecting false positive SNPs due to the multiple testing correction problem (Balding, 2006). The difficulty is compounded with testing large number of SNPs (p) containing a small number of observations (n), i.e. $p > n$, which is typically a feature selection problem, where different methods may be applied (Saeys et al., 2007). Due to these problems, in most cases, the traditional approach might be leading to the 'missing heritability' issue (Manolio et al., 2009).

The methods that adopt simultaneous inclusion of all SNPs, for a while have become the standard approach for GAS in dealing with the problem (Hoggart et al., 2008). These include the widely used standard *forward stepwise* method for multiple SNP analysis (Cordell and Clayton, 2002), which was adopted in the Jones et al. work to analyse nearly 1553 SNPs from approximately 500 subjects (Jones et al., 2009). The stepwise approach is argued to be statistically sub-optimal (Harrell, 2001) and tends to omit key genetic variants, particularly those with high linkage (Malo et al., 2008). Other multiple-SNP approaches include variants of penalised regression methods (Ayers and Cordell, 2010; Wu et al., 2009).

Besides these improvements, different analytical techniques are still needed that might be able to further identify unaccounted genetic variants that are associated with the complex traits/diseases and which address the 'missing heritability' (Manolio et al., 2009; Moore et al., 2010; Robinson et al., 2014). Thus, several authors have suggested approaches, which involve machine learning and data mining, in addition to bio-statistical methods, for efficient identification of novel SNPs associated with complex traits (Ao, 2008; Fernald et al., 2011; Moore et al., 2010; Touw et al., 2013).

Therefore, in addressing the above problems, this study is proposing a novel computational pipeline (RAPIDSNPs) (Salehe et al., 2017), which rapidly identifies previously unaccounted

potential key SNPs that are associated with ADP platelet responses. The pipeline involves an integrated, heuristic-based, hybrid strategy, incorporating data mining and machine learning methodologies to identify potential key and significant unaccounted SNPs. The project further incorporates the bioinformatics and predictive approaches for unravelling the underlying molecular mechanistic effects of the identified key SNPs', which likely contribute to the variability of the ADP platelet responses and CVD risk.

1.6 Integrated computational and predictive approach

1.6.1 Integrated approaches for analysing omic variations

In addressing the above analytical problems related to GASs, the integrated analytical approach appears to be a more pragmatic strategy (Ritchie et al., 2015). To implement this approach, different data sets and results are integrated in multi-phased analyses rather than focusing only on each SNP's genotype and its association with the concerned phenotype (Sieberts and Schadt, 2007; Hamid *et al.*, 2009).

In this regard, the integrated analytical approach entails generating models that may describe and predict the potential interactions and complexity between SNPs, and other omic variations, which may further explain the variability of the ADP platelet responses phenotype (Ritchie et al., 2015).

1.6.2 Integrated approach using data mining and machine learning for 'omic variations analyses

Several different integrated approaches for analysing omic data have been developed by different research groups. Each depend on, or are guided by, the homogeneity or heterogeneity of the data that are to be integrated (Hamid *et al.*, 2009). However, systems genomics approaches, which incorporate meta-dimensional and multi-staged analyses have gained higher

attention (Hawkins et al., 2011; Holzinger and Ritchie, 2012). In the interests of this study, the multi-staged approach is appropriate, where the omic variations that underlie the trait or disease are examined in hierarchical or linear manner. For instance, the DNA sequence variation or SNP is examined and if it leads to changes in the protein sequence, then it is further investigated to examine its effects on the structure/function, and resulting changes in phenotype (Ritchie et al., 2015).

Data mining and machine learning methods have been extensively used in order to implement such approaches (Ao, 2008; Hamid et al., 2009; Nicodemus and Malley, 2009). Data mining and machine learning have been found to be effective, as they incorporate prior biological knowledge such as different omic interactions and are capable of finding useful patterns from the multivariate data (Moore et al., 2010; Holzinger and Ritchie, 2012).

Since the data and results in this study might come from the different omic levels, then, a hybrid strategy for integrating data and results is designed and implemented. In the initial phase, the RAPIDSNPs approach, which is a hybrid and integrated approach for key SNPs identification, is developed. The underlying mechanism of RAPIDSNPs is explained in detail in Chapter 2. The method combines and integrates multiple approaches, which are grounded on data mining and machine learning algorithms to form a consensus selection of the key SNPs associated with ADP platelet responses variation. It has been suggested that combining multiple individual approaches in downstream analyses is a useful strategy for minimising the selection of false positive SNPs in genotype – phenotype studies (Holzinger and Ritchie, 2012).

In the intermediate phase of the pipeline, key SNPs are investigated using integrated bioinformatics approaches. The results from this phase provide a further understanding of the likely molecular mechanisms that may lead to differential ADP platelet responses and aggregation. In the long term, this knowledge might be useful for guiding the interpretation of

the results for CVD personalised medical decisions in clinical settings. Figure 1.7 shows the information flow in a simplified integrated approach for investigating the underpinning variability of ADP platelet responses for CVD personalised healthcare.

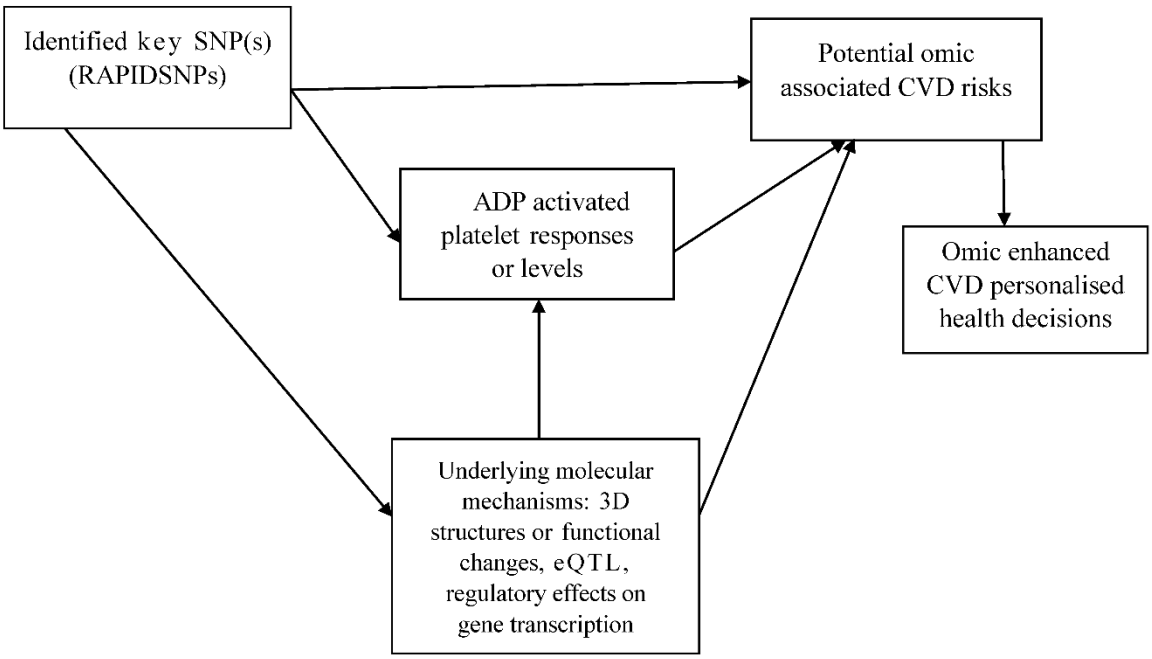


Figure 1.7 The linear relationship showing the analytical flow of the data and results from different phases in the integrated manner. The research's primary hypothesis is that there are potential previously unidentified key SNPs, which are associated with the ADP platelet responses and that may underpin CVD risks. However, the broader hypothesis is that these SNPs are related to other omic variations, which further may underpin the variability of the ADP platelet responses and may also underlie the CVD risks. Hence, there are several intermediate hypotheses between the identification of key SNPs phase and the omic enhanced CVD personalised healthcare phase.

1.6.2.1 ***Brief description of the integrated approach of the project***

In Figure 1.7 the identified SNPs or key SNPs, are obtained using RAPIDSNPs, which incorporates feature selection tools for ultimately selecting significant and key SNPs patterns associated with ADP platelet responses.

Several of these significant SNPs cause missense and deleterious mutations to the related proteins, which are shown to have likely structural and functional effects that may underlie the differences in ADP platelet responses and CVD. The identification of the deleteriousness of the SNPs and their potential structural/functional effects on the related proteins, which likely

underpins the molecular mechanisms, are performed using protein bioinformatics analyses. Numerous 3D models of tertiary structures are generated using a protein structure prediction protocol stack. Moreover, for further understanding of the molecular mechanisms, the identified significant intronic, non-coding, non-deleterious missense and synonymous SNPs are examined their possible involvement in the regulatory roles. The aim is to investigate how likely it is that the identified SNPs are localised in the regulatory regions and involved in the differential gene transcription (transcriptomic variations). The presence of SNPs in the regulatory regions, in turn, may likely affect the regulation of the individuals' ADP platelet responses or levels.

Furthermore, another predictive approach, which is based on supervised ML algorithm has been developed for predicting the individuals' ADP platelet responses levels based on the identified SNPs' allelic features in phase 1 (Chapter 5). This new predictive approach has been used to generate models, which can help to determine how likely an individual's ADP platelet response levels would be, based on examination of the population's SNPs genotypes (alleles). These models may be useful for identifying potential individuals with extreme cases of ADP platelet responses, i.e. high and low responders. Ultimately, the predictive models may be used for therapeutic or clinical purposes, as basis for predicting the individuals' increased (or decreased) platelet aggregation and potential risk of CVD (Chapter 6).

Furthermore, apart from the specific ADP platelet responses, the likelihood of the identified SNPs to be associated with CVD in general is determined using a literature search and following a meta-analysis technique. This task is aided by exploiting computational tools such as SNPedia (Cariaso and Lennon, 2012) and SNPNexus (Ullah et al., 2012). The results will help us to generate hypotheses for future studies on how likely the identified omic variations, that are associated with the ADP platelet responses, participate in CVD pathogenesis. In the

long term, the results will also be useful in assessing the individual's genetic determinants, CVD risk levels (Chapter 6), and help to develop appropriate targeted therapeutics in clinical settings.

The approach as demonstrated in Figure 1.7 entails modelling complex interactions, which may underlie complex disease (CVD) by integrating information from the different aspects of the omic variations underpinning ADP platelet responses variability.

1.7 Personalised healthcare using identified omic variations associated with ADP platelet response levels for CVDs?

1.7.1 What is personalised healthcare?

Personalised healthcare is a modern approach to healthcare delivery, which is based on the examining the variations in an individual's inherited characteristics that might contribute to the disease outcome (Collins and Varmus, 2015). This has been made possible by the recent advancement of key biotechnologies, computational tools and methods, which are capable of characterising and generating large-scale omic and other biological data for individual patients (Collins and Varmus, 2015). Personalised healthcare is mainly derived from the term personalised medicine, which was originally defined as providing medicines based on individual genetic makeup (Langreth and Waldholz, 1999). It might also be more simply defined as the ability of practitioners to provide the right drugs for the right patients (Bates, 2010). However, this definition bears controversy, as no doctor intentionally prescribes the wrong medicine (Bates, 2010)! A standard definition is provided by PCAST (US President's Council of Advisors on Science Technology):

'Personalized medicine refers to the tailoring of medical treatment to the individual characteristics of each patient. It does not literally mean the creation of drugs or medical

devices that are unique to a patient but rather the ability to classify individuals into subpopulations that differ in their susceptibility to a particular disease or their response to a specific treatment. Preventive or therapeutic interventions can then be concentrated on those who will benefit, sparing expense and side effects for those who will not' (President's Council of Advisors on Science Technology, 2008).

The key component of personalised healthcare is the use of information implicit in the genome and its expressed products, such as RNAs and proteins, to guide medical decisions. Thus, personalised medicine is alternatively termed *genomic medicine* (Ginsburg and Willard, 2009). Another alternative name is *precision medicine*, which means tailoring medical treatment by incorporating individual variability in genes, lifestyle, and environment (Precision Medicine Initiative, 2015).

Furthermore, personalised healthcare might be further and broadly defined based on its features, i.e. it is a model of healthcare, which is predictive, personalised, preventive, and participatory (P4 medicine) (Hood and Flores, 2012). This healthcare paradigm is characterised by the development and application of computational tools to biomedical research data through bioinformatics, which hopefully may guide healthcare providers and consumers in making an informed decision for improving human health (Overby and Tarczy-Hornoch, 2013).

At the core of personalised healthcare is the use of data analytics and emergent technologies in identifying and understanding the underlying cause of disease (NHS England, 2015). These technologies are embedded with or rely on the advances in research in the molecular biology, genomics, and bioinformatics (Collins and Varmus, 2015). Data on omic variations is at the heart of personalised healthcare, exemplified through the advancement of GWAS and its related technologies (Peterson et al., 2013). These data are being generated at an ever increasing rate (Consortium, 2012, 2010; Djebali et al., 2012; Reva et al., 2011; Wilhelm et al., 2014).

This big data phenomenon urgently requires new methods and optimised computation tools for integrating data and results in order to understand the underlying genetic interactions and molecular mechanisms governing disease, which can be eventually translated into personalised healthcare practice (Fernald et al., 2011).

1.7.2 Why do we need personalised healthcare?

Conventional medicine and clinical practice are based on the philosophy of 'one size fit all', which results in treatment designed for the average patient. This can often work well for the majority of patients but not for others (Precision Medicine Initiative, 2015). Personalised healthcare is tailored to proactively deliver healthcare services to meet patient's individual health needs (diagnosis, treatment, prevention). There is clearly a need for the development tools, using bioinformatics and systems biology approaches, for predicting individual responses to treatments, which will allow us to reduce healthcare costs, while improving and maintaining the wellbeing of patients (Burnette et al., 2012; Hood and Flores, 2012; Overby and Tarczy-Hornoch, 2013; Snyderman R and Dinan MA, 2010).

1.7.3 Omic variations and personalised healthcare for ADP platelet responses and CVD

SNPs play a significant role in the genetic heritability of platelet responses through their influences on the hypo and hyper-reactiveness of the platelet responses among individuals. It has been suggested that understanding the genetic contribution to platelet functioning might have a clinical impact on personalising platelet focused CVD therapeutics (Williams et al., 2010). This is due to the fact current anti-platelet drugs that target the ADP receptor (P2Y12) have been reported to have reduced efficacy in many patients (Offermanns, 2006; Woulfe et al., 2001). Therefore, we are focusing our attention on ADP platelet responses as they have high potential for directing personalised therapeutics using antiplatelet drug combinations.

1.8 General aim and hypothesis

In general, the study aims at the development of computational methods for the identification of novel or key SNPs and other omic variations that might be associated with variability of complex traits/diseases, for potential application in personalised medicine.

The study specifically aims to investigate the effect of omic variations on ADP activated platelet responses that are likely to contribute to the inter-individual variations in platelet aggregation and thrombus formation, and CVD disease risks. The study further aims to elucidate the genotype – phenotype relationship underpinning high or low ADP platelet response levels, which may be used in the long term for the development of personalised approaches to the treatment of CVD.

The primary hypothesis that drives this study is that there are previously unaccounted genetic variants (SNPs) or omic variations, which are likely to be associated with the variability of the ADP platelet responses. These unaccounted variants might play further key roles in the molecular mechanisms, which may underlie inter-individual variability in platelet aggregations (or thrombus formation) that will lead to differing CVD prognoses.

In vitro experiment to measure thrombus formation over time among individuals was performed by C.I Jones. Initial rate of platelet calcium flux and fibrinogen binding were then measured and associated with the rate and size of thrombus formation. The plot in Figure 1.8 (data from this preliminary study) shows the differing thrombus sizes of 45 individuals, which are potentially due to the effects of key genetic variants indicated in previous studies (Jones et al., 2007; Pruissen et al., 2009; Rauch et al., 2001).

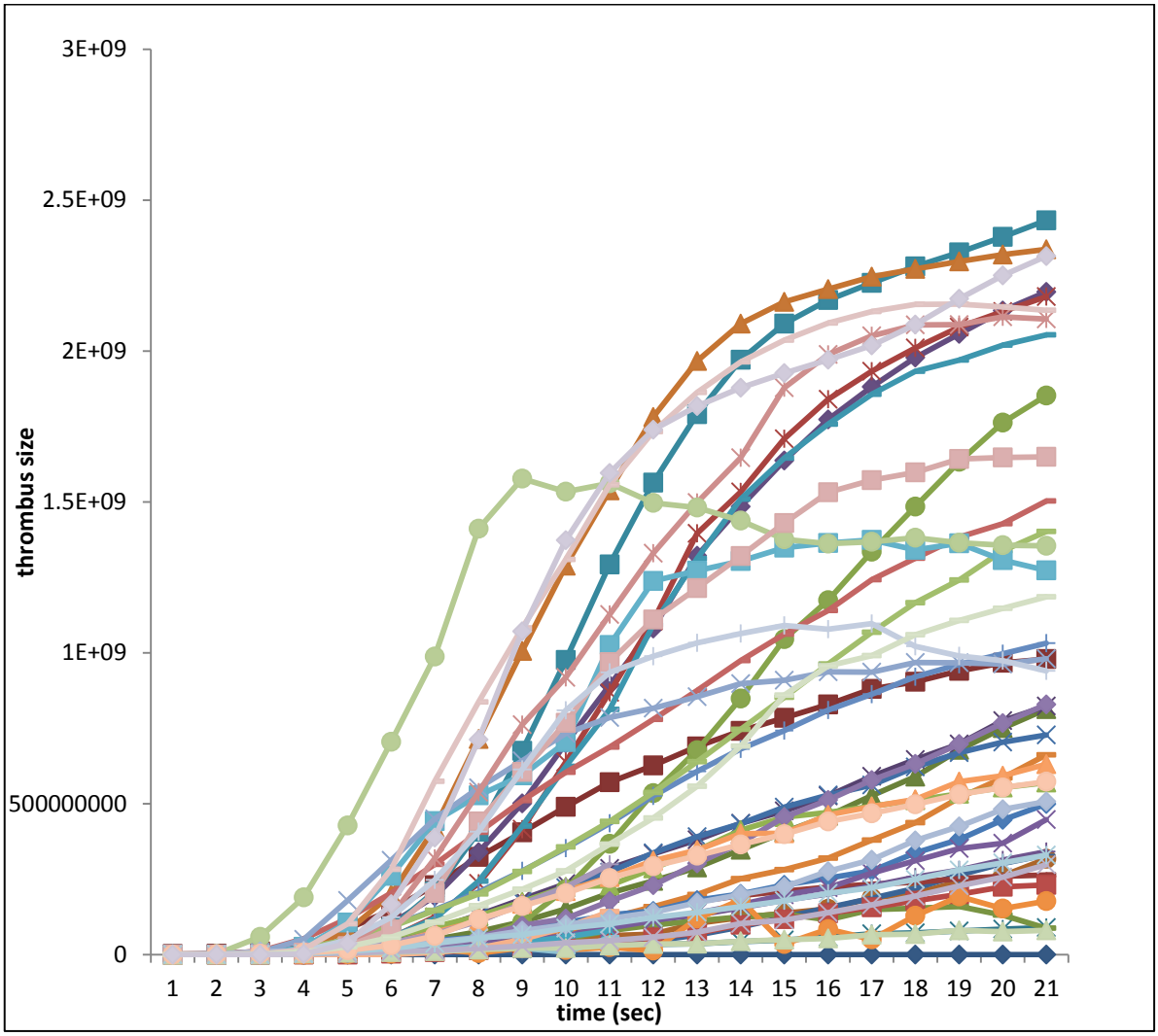


Figure 1.8 The variation of the *in vitro* thrombus formation (Hounsfield unit) of 45 individuals at the different time (in seconds). Each line represents the intra-individual variation of the thrombus formation rate for each subject out of the 45 subjects' blood samples. It might be seen clearly that individuals' thrombus formation varies relative to each other (inter-individual variations). The rate of thrombus formation data for 45 subjects was acquired from C.I. Jones, University of Reading.

This inter-individual variation was further examined through the in vitro measurement of P-selectin expression and fibrinogen binding platelet responses as the result of ADP activation (measured by flow cytometry). Platelet responses to ADP were analysed *in vitro* using flow cytometer, in which the binding of fibrinogen to platelets and expression of P-selectin by platelets was measured. The resulting data are plotted in Figure 1.9, which shows the correlation between two ADP platelet response measures (PA and FA) and Figures 1.10 and 1.11, which show their distribution.

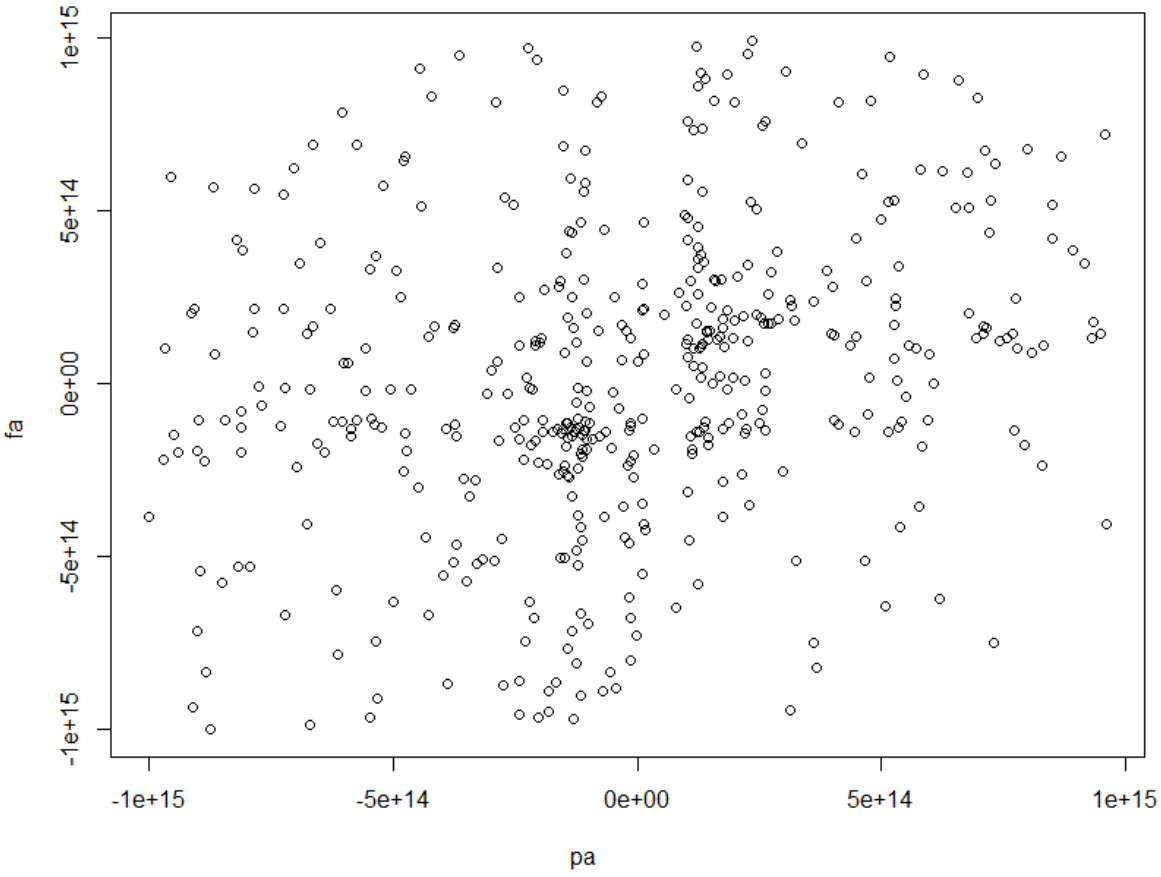


Figure 1.9 The correlation between PA and FA platelet response measures. The x and y-axes are P-selectin expression and fibrinogen binding in response to ADP platelet activations (pa & fa) respectively.

From Figure 1.9, PA and FA appear to be positively correlated. And based on the standard Pearson correlation test, their correlation appears to be significant (**p-value = 9.697e-08**). This might mean the increase or decrease in degranulation, which results in P-selectin release may signify the increase or decrease in fibrinogen binding in response to ADP activation.

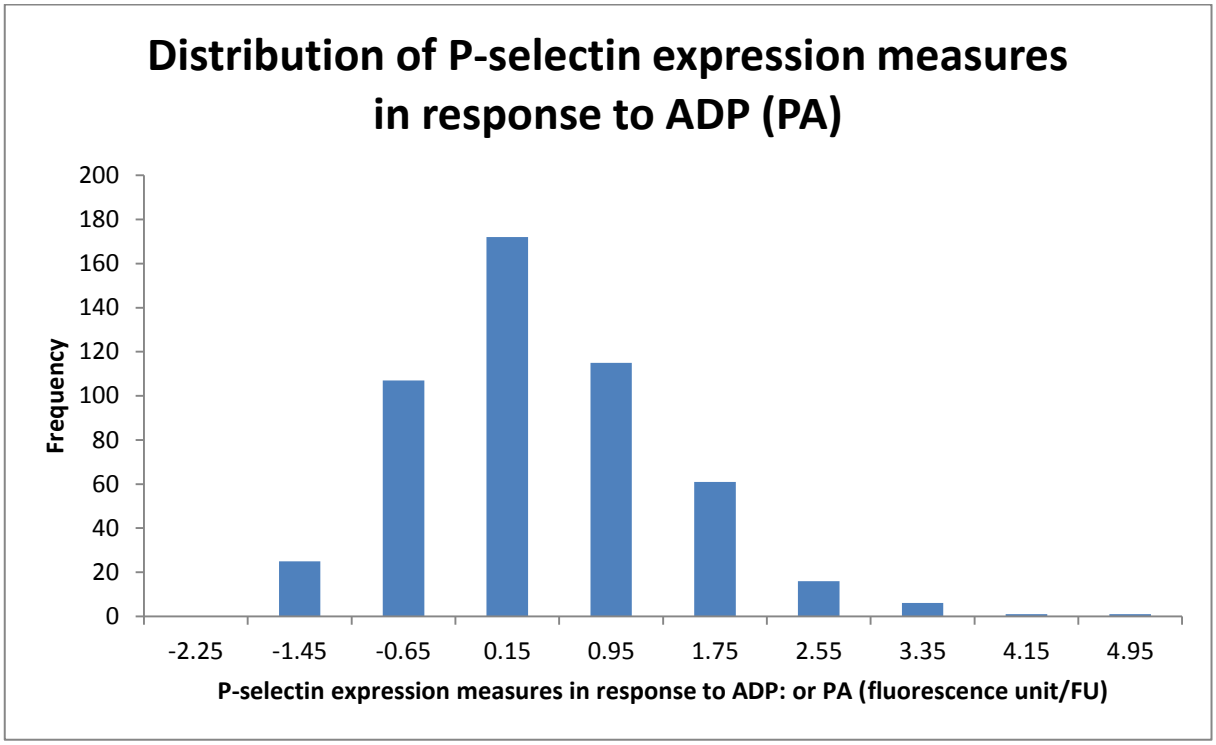


Figure 1.10 The distribution of the PA platelet response among different subjects. There are total of 497 subjects who were measured for their PA platelet responses. The plot further reflects Figure 1.8, which shows the inter-individuals variability of the ADP activated platelet responses and thrombus formation. The interesting subjects are those with extreme values, i.e., those with high and low responses. The data was acquired from (Jones et al., 2007).

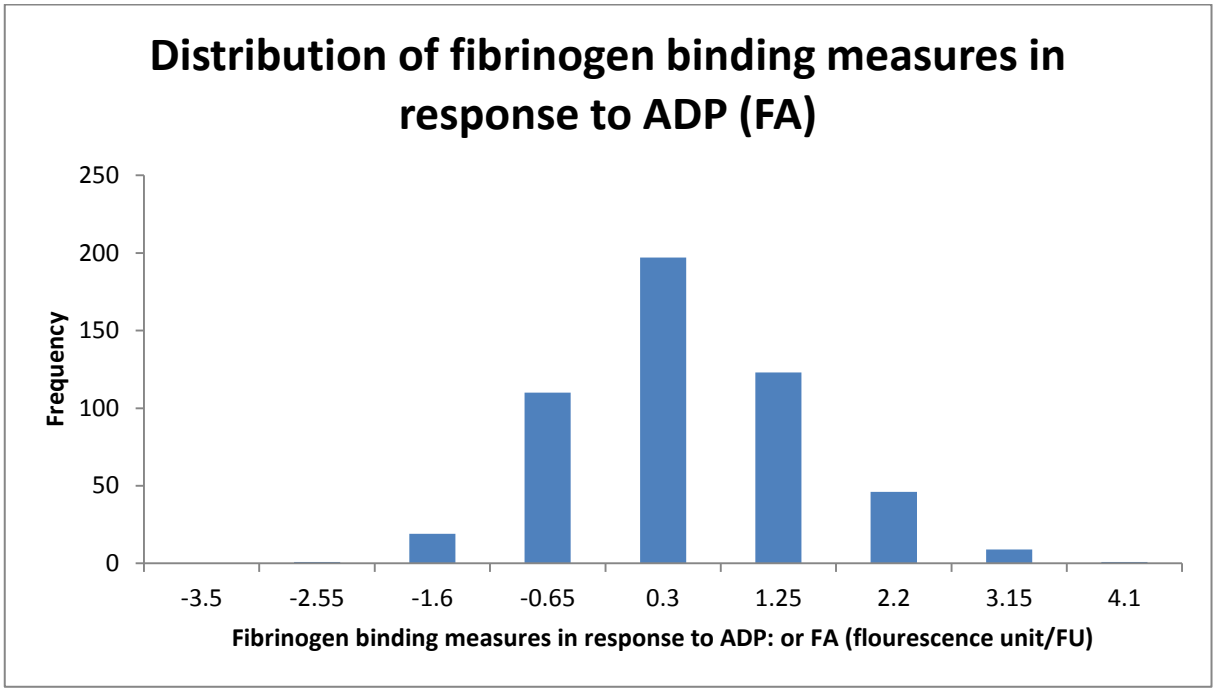


Figure 1.11 The distribution of FA platelet responses among different subjects. Similarly, the total number of subjects were 497. Again, here the interesting subjects are those with the extreme responses, i.e. tails of the distribution. The data was acquired from (Jones et al., 2007).

The plots in Figures 1.8, 1.10 and 1.11 show the possible effects of the omic variations to the overall inter-individual variations in thrombus formation, P-selectin ADP responses and fibrinogen ADP responses, respectively. Nevertheless, some of the variations are due to environment. The most interesting individuals, from clinical and biological perspectives, are those showing extreme variations in the thrombus formations, and ADP platelet response levels.

Therefore, a further aim of this study is to determine the predictive ability of the identified key SNPs in predicting the likelihood of an individual to have extreme ADP platelet responses (high or low PA or FA) levels based on their alleles. In addition, the related aim is to elucidate which of the alleles (allelic-patterns) are more likely to be associated with these extreme variations of high or low PA/FA levels (Chapter 5). The approach would highlight the important associated individual SNPs and their alleles, which could be used to guide future CVD personalised

healthcare decisions or strategies (for instances in CVD diagnosis or/and treatments) (Chapter 6).

1.9 Study objectives

To achieve the above aims, the specific objectives are to:

1) Design a computational approach for rapid identification of the previously unidentified key SNPs that might be associated with the inter-individual ADP platelet responses phenotype.
2) Investigate the roles of the identified SNPs that are likely to underpin the mechanisms of ADP platelet responses and aggregation at the molecular level. In this case, the objectives are to:
    a. Design the predictive approaches that may identify the deleterious missense SNPs and investigate their effects on the structures/functions of the related proteins at the proteomic level.
    b. Design the computational approach for investigating the potential regulatory mechanisms of the identified non-damaging missense, non-coding and intronic SNPs.
3) Design a method for predicting the ADP platelet response levels (high or low, i.e. the extreme cases), which may underpin increases or decreases of platelet aggregation, thrombus formation, and CVDs risks for personalised healthcare.

Generally, it is hoped that the study will further contribute to the current knowledge of the genetic basis underpinning the ADP activated platelet responses and CVD prognosis. In addition, the results obtained might have wider implications in overall platelet systems biology and potential to affect personalised CVDs medical decisions.

The key research questions of the study are: (1) What are the key previously unidentified genomic variations (SNPs), which are likely to be further involved with the variation of ADP platelet responses (PA and FA)? (2) Are the identified key SNPs damaging to the structures and/or functions of the proteins associated with ADP platelet functions? (3) Can we confidently predict the structures/functions of the identified interesting proteins with damaging mutations (SNPs)? (4) Do the identified key SNPs found in the regulatory regions affect the regulation of genes transcription associated with ADP platelet functions? (6) How are the identified SNPs involved in the regulation of genes transcription associated with the ADP functions? (7) Can we predict ADP activated platelet responses levels in an individual?

The research will focus particularly on ADP platelet activation pathways, which play an important role in the amplification of platelet aggregation and thrombus formation, and CVD prognosis. However, these computational and predictive approaches may also generally be applied for the study of other genetic diseases/traits involve continuous phenotypic variation.

1.10 Conceptual framework

To elucidate the aims and objectives of this study further, the designed conceptual framework underpinning the genotype-phenotype aspects of the study is shown in the Figure 1.12.

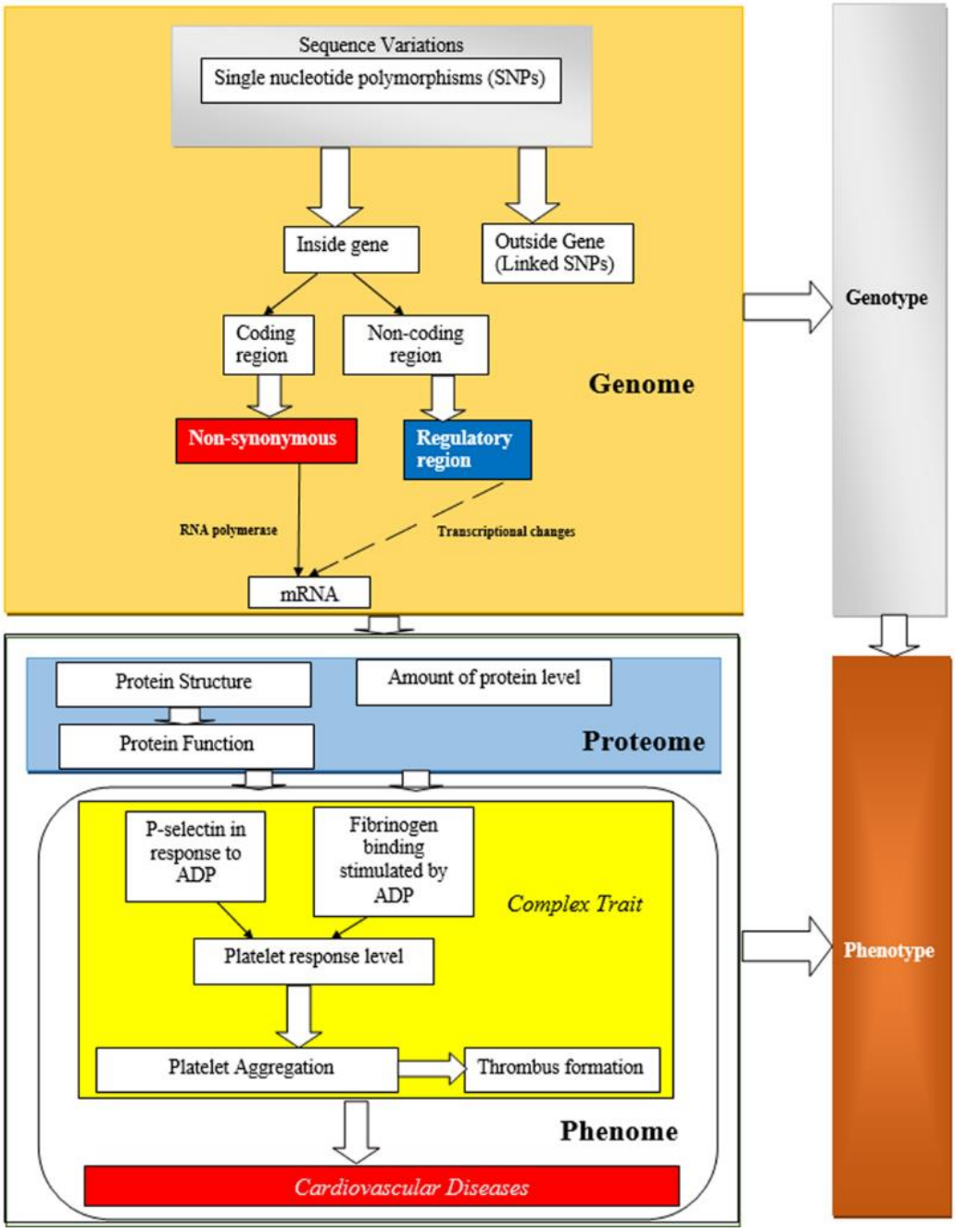


Figure 1.12 The general framework for studying and analysing omic variation data for understanding their likely association with complex traits and disease prognosis. At the genome level, the main focus is on the identification and understanding of the key SNPs that might initially associate with or contribute to the intermediate changes. For instance, SNPs in the genome level might lead to the regulatory changes and result in the transcriptomic variations. These changes may have possible effects on the proteome levels, which may include structural and functional changes of the proteins or the amount of protein produced. These molecular changes might lead to changes or variability associated with the complex traits that may progress to the disease status at the phenome level. At the phenome level, the framework emphasises the study of ADP platelet responses and its effect on the thrombus formation as a complex trait, which may underpin the CVD prognosis. However, the general framework is adaptable to other complex traits/diseases.

Holistically, the integrated approach described earlier in the section 1.6.2 (Figure 1.7) is applied to interpret and integrate the results from one phase of the conceptual framework to another (Figure 1.12). Moreover, the project pipeline is based on the described integrated framework,

comprising an integrated computational approach for analysing the omic variations in ADP platelet responses.

1.10.1 General project pipeline for omic variation analyses

Functional genomics partly involves finding and understanding the relationship between genotype and phenotype for complex traits and diseases (Pevsner, 2009). To further understand the genotype-phenotype association, researchers have described the urgent need for genomic analysis pipelines (Bromberg, 2013; Morris and Zeggini, 2010). Therefore, the computational pipeline is designed for integration and smooth interpretation of the data and results from the various phases of the described conceptual framework. This pipeline is configured in order to unveil the relationship between complex genetic or molecular interactions and phenotypes, for future applications in biomedical science and potentially clinical practice.

The anatomy of this pipeline, which in this case is applied for investigating the relationship between SNPs and ADP platelet activated responses, has the following phases; phase A – SNP *screening and discovery*, phase B – bioinformatics analyses (for investigating the molecular aspects of the SNPs, i.e. damaging missense, regulatory, etc.), phase C – *phenotype identification (intermediate, including 3D structural models)*, phase D – *phenotype prediction* (*ADP platelet response levels/CVD potential risks*), phase E – *evaluation (iterative in some phases)*. Figure 1.13 depicts this pipeline as a flow chart:

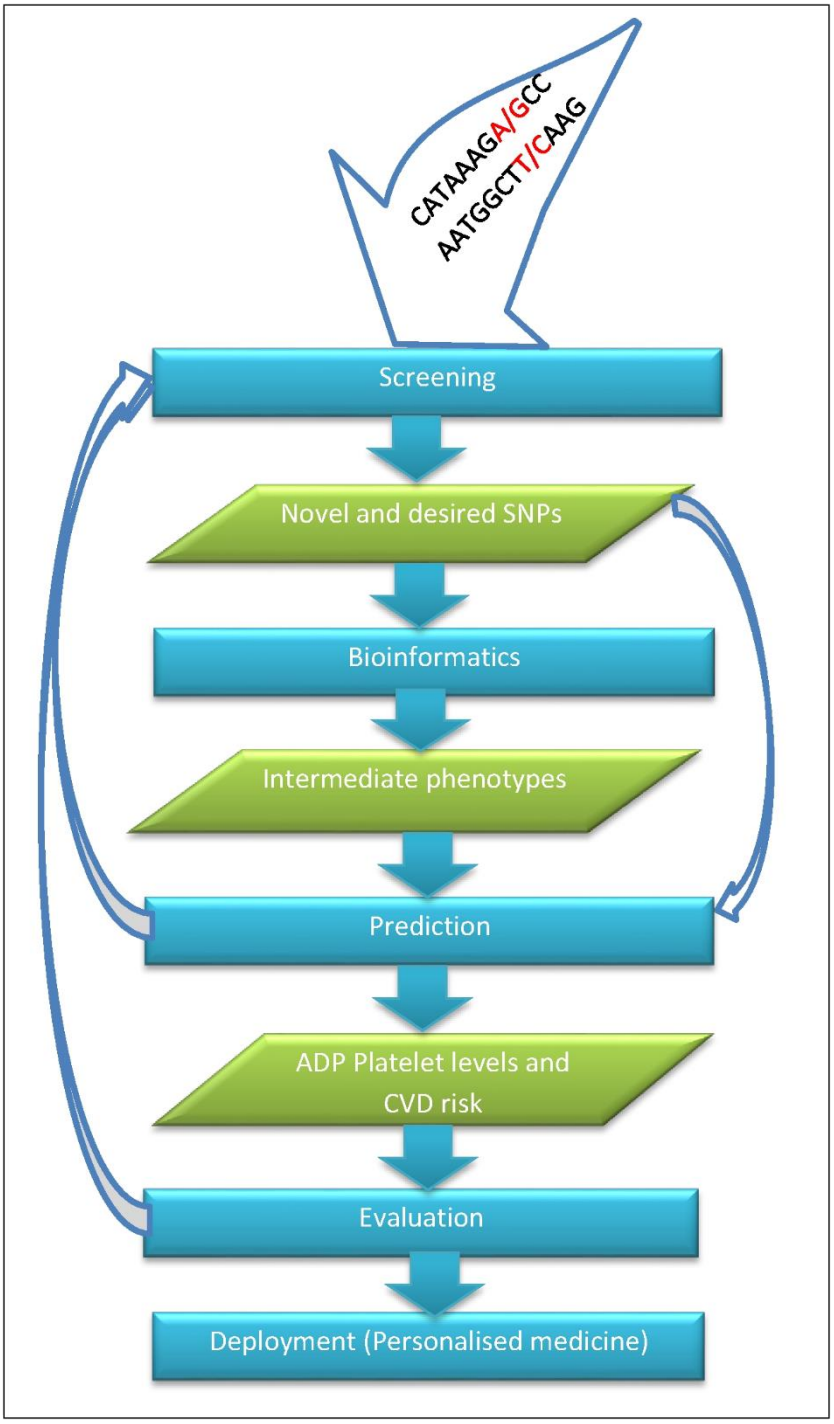


Figure 1.13 The high-level flow of data and results in the pipeline for the integrated computational approach, which is needed to implement the framework shown in Figure 1.11. This pipeline shows critical stages and necessary outputs/inputs during the analyses of the various omic variations under the study. The filtering (screening) phase describes an approach for performing screening of the SNPs obtained from GASs. The aim is to reduce the model space of the SNPs for finding those that likely underpin the ADP platelet responses phenotype. In polygenic traits, different SNPs have varying degrees of effect i.e. some are more/less contributing to the phenotypic effect. Therefore, this phase is crucial for obtaining useful key SNPs for further analyses. The bioinformatics phase aims to investigate the underlying molecular aspects of the obtained key SNPs underpinning the ADP platelet response. The prediction phase aims to generate models for predicting ADP platelet response levels and potential CVD disease risk status. Thus, there are three key outputs expected: 1) Previously unidentified key SNPs, which are identified through a rapid computational approach (RAPIDSNPs). 2) Intermediate phenotypes, such as protein 3D structural models for investigating the damaging missense SNPs, and transcriptomic variation related data such as transcription factor binding sites, RNA binding sites, and eQTL that might be due to the regulatory roles of the identified key SNPs. 3) Predictive models for predicting ADP response levels and potential CVD risks.

This pipeline adheres to an integrated analytical approach strategy, which was designed by Ritchie et al. for modelling of 'omic' variations associated with complex phenotypes. The integrated approach is an effective means for identifying predictive models for complex phenotypic traits outcomes (Moore et al., 2010; Ritchie et al., 2015) (Figure 1.14).

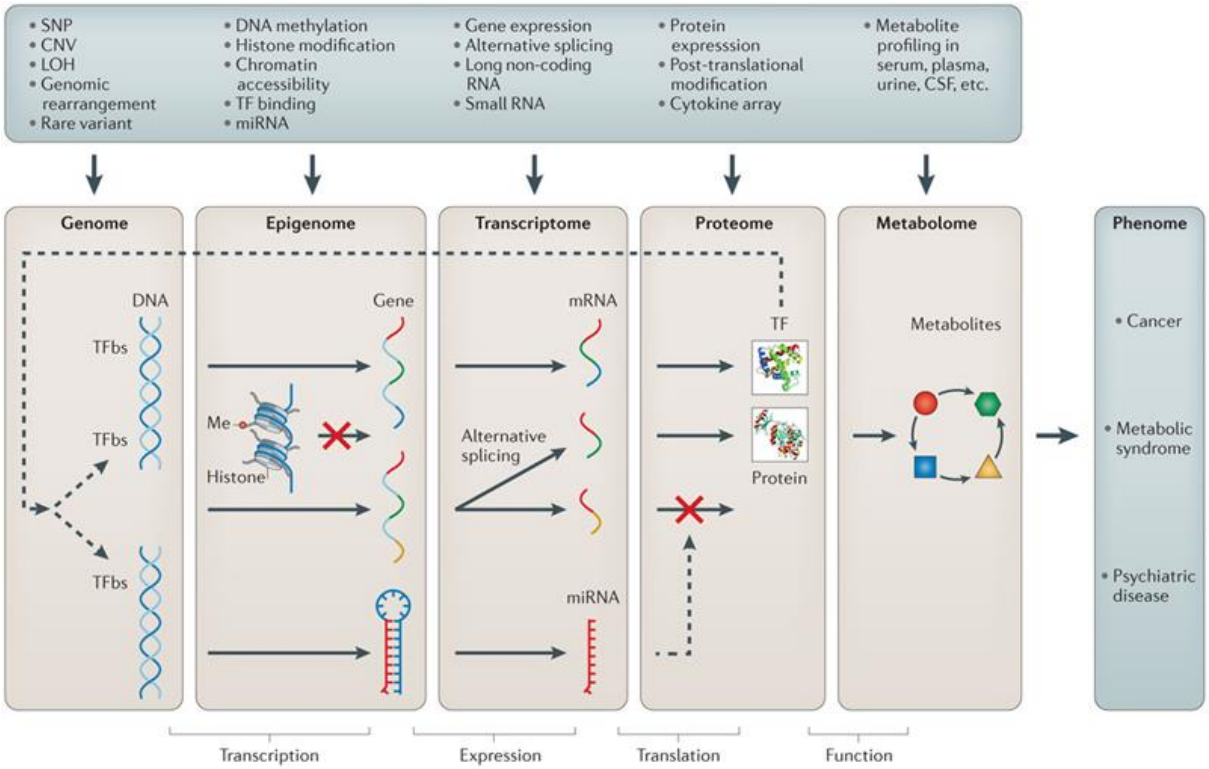


Figure 1.14 An integrated framework involving multistage analyses underlying genotype-phenotype association for 'omic' variation data proposed by Ritchie et al. This framework complements our designed conceptual framework for this study, shown in Figure 1.11. The developed theoretical framework for this study involved the key SNPs at the genome level, eQTL or transcription factors, RNA binding sites, etc. at the transcriptome level, and protein 3D structures/functions at the proteome level. Holistically, all levels may be contributing to the underlying variability of the ADP platelet responses, thrombus formation and CVDs risks at the phenome level. (Figure was taken from (Ritchie et al., 2015).

The following sections briefly describe the individual phases of the pipeline in the Figure 1.13 and further reflect the designed integrated approach in the sections 1.6.2 and 1.6.2.1.

1.10.1.1 ***Screening phase***

90% of human genome variations in DNA are SNPs (Collins et al., 1998). Nevertheless, only a small proportion of SNPs that occur around and in the coding regions are of biological

importance (Pevsner, 2009). Therefore, the screening phase is important to remove from the sample data those SNPs, which are likely to be less significant to the overall variability of the ADP platelet functions. Thus, a rapid computational methodological approach (RAPIDSNPs) has been developed and employed for identifying key SNPs, which were previously unidentified.

1.10.1.2 ***Bioinformatics analyses***

This phase aims to investigate the possible molecular effects of the identified key SNPs from the screening phase. In general, bioinformatics approaches can help us to determine whether the identified SNPs are likely to play any role at the molecular level (Hutchins, 2014). For instance, in the case of the identified SNPs in the coding regions, the interest is to find their effect on the sequences, structures and functions of the resulting proteins. Thus, the major focus is to identify whether the SNP is likely to be a deleterious missense mutation and then predict the possible structural/functional effects of the related proteins (Cavallo and Martin, 2005; Wang and Moult, 2001).

Conversely, for the identified intronic, non-coding, and synonymous SNPs, the purpose is to find whether they are involved with the regulatory activities. These are likely to contribute to the variations at the transcriptome level that may also underpin the variability of the ADP platelet responses (Gerasimova et al., 2013; Gibson et al., 2001; Hull et al., 2007).

Figure 1.15 is a flowchart, which illustrates a detailed flow of information for the molecular analyses of the identified key SNPs using the Bioinformatics approaches.

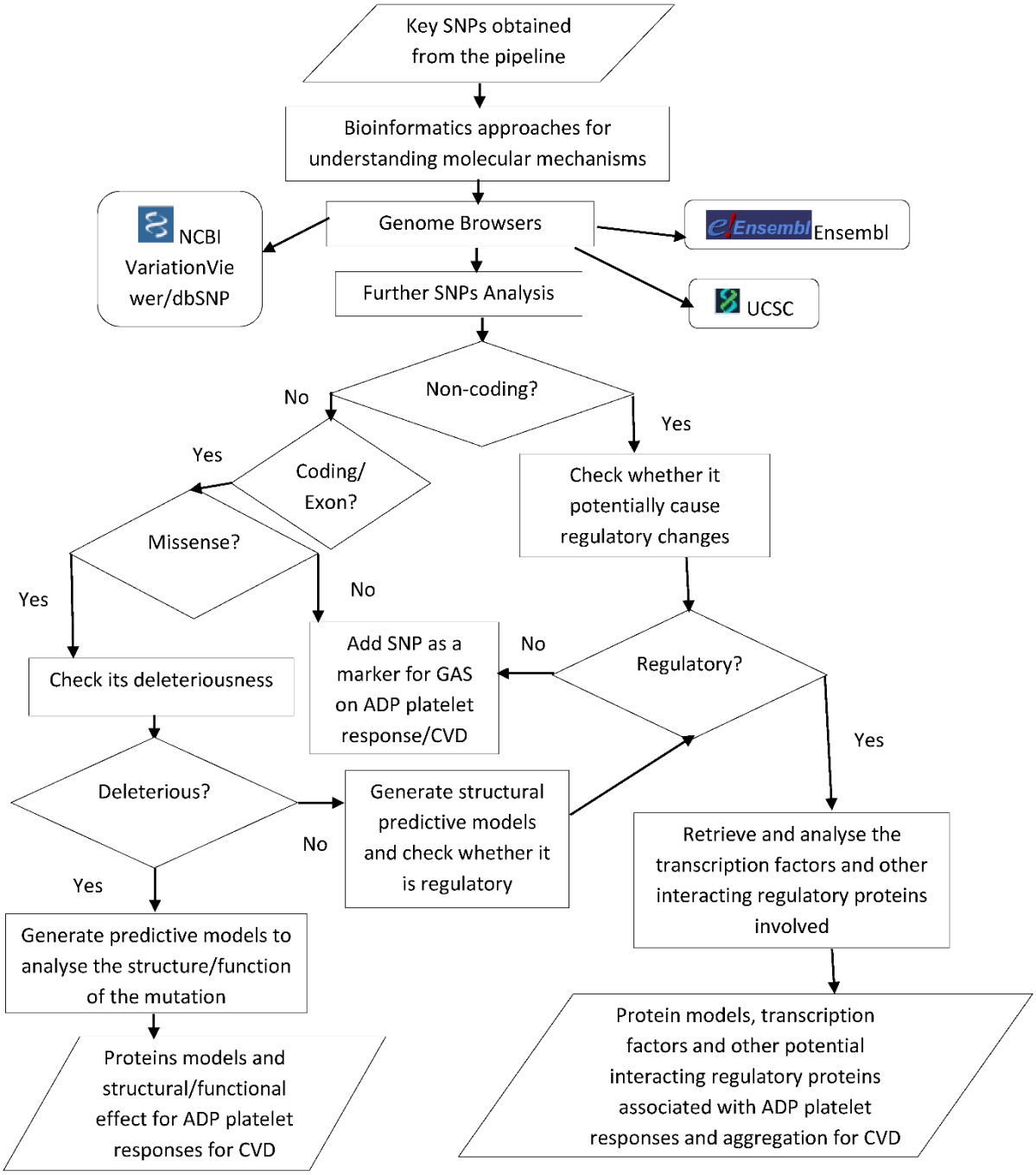


Figure 1.15 The schematic flowchart of the bioinformatics analysis pipeline for investigating the molecular aspects of the key SNPs obtained in the filtering phase. The interest is to examine the structural/functional effects due to the identified damaging missense SNPs in addition to the regulatory mechanisms of the intron, non-damaging and non-coding, and synonymous SNPs that likely underpin the ADP platelet responses variability. The identified SNPs from the filtering phase are first analysed using genome browsers. Depending on the SNPs category, different computational and bioinformatics methods are then applied to determine their effect on the structural and/or functional, or regulatory activities of the related proteins/genes.

1.10.1.3 ***Prediction***

The data on significant SNPs, which are obtained from the filtering phase, may be further utilised to predict whether individuals are likely to have low or high level ADP platelet responses. Such predictions would be useful for determining whether platelet aggregation/thrombus formation is likely to be decreased or increased. In turn, the results might be used to predictively determine the individuals' CVD risk, informing therapeutic and clinical decisions, or interventions, depending on the confidence in the model and the SNPs allelic features involved in making the prediction.

1.10.1.4 ***Evaluation***

Various predictive models are evaluated using the standard approaches to determine their accuracy and quality for potential application in the personalised medicine or targeted therapeutics. In the case of the generated protein 3D structural models, the state of the art Model Quality Assessment Programs (MQAPs) are used. For the predictive models of ADP platelet response levels, the standard approaches such as confusion matrices are used to evaluate prediction quality.

1.11 Organisation of the thesis

Based on the framework and pipeline, this thesis is organised as follows: Chapter 2 describes the RAPIDSNPs approach for the rapid identification of the previously unidentified key SNPs, which are likely to be associated with variability of the ADP platelet responses. Chapter 3 describes the predictive approaches for identification of the potential structural/functional effects of the identified key SNPs (missense SNPs) identified in Chapter 2. Chapter 4 describes the computational approaches, to identify or predict the potential regulatory roles of the key SNPs identified in Chapter 2. Chapter 5 describes a method for predicting the ADP platelet

response levels by exploiting the allelic features of the identified SNPs in Chapter 2. Lastly, Chapter 6 provides the synopsis and future direction of the research based on the results from different chapters.

1.12 Summary

The ability to generate high throughput genomic sequences has enabled many genetic association studies to be performed to understand complex trait/disease causing variants (e.g. SNPs). However, the methods employed by these studies have not fully accounted for key genomic/post-genomic variations nor have they explained the remaining or missing heritability. In addition, these studies rarely offer insight on the molecular mechanisms of the variants, which is vital for our deep understanding of the variability of complex traits/diseases and their pathophysiology for personalised healthcare/medical decisions.

This project provides an alternative integrated computational and predictive approach for further elucidating the remaining unexplained genetic aetiology of complex traits/diseases. The ADP platelet responses, which play the significant role in the platelet aggregation, underpinning various CVD problems, has been used in this study as a key “model system”, for testing our methodology. The results provide some new insights into platelet biology and suggest new directions for targeted antiplatelet therapy for personalised healthcare. In future, our approaches might be applied to investigate other complex traits/diseases.

Chapter 2 – RAPIDSNPs: Rapid computational pipeline for identifying key SNPs associated with ADP platelet responses

2.0 Abstract

This chapter describes a novel computational pipeline for identification of the previously unidentified SNPs, which are associated with continuous phenotype from the genetic association studies (GASs). Advances in omics particularly genotyping technologies have led to the discovery of genetic markers, or single nucleotide polymorphisms (SNPs), that are associated with complex diseases/traits using GASs approaches. Although there have been significant improvements in the GASs approaches used to analyse associations of SNPs with the disease, further optimised and rapid techniques are needed to keep up with the rate of SNP discovery, which has exacerbated the 'missing heritability' problem. Herein, a novel, integrated, heuristic-based, hybrid analytical computational pipeline, for rapidly detecting novel or key genetic variants (SNPs) that are associated with diseases or complex traits is described. The pipeline (RAPIDSNPs) is more efficient for investigating small sets of genotyped SNPs defined in high dimensional spaces that may be associated with continuous phenotypes, rather than for the investigation of whole genome variants. The RAPIDSNPs employs a consensus approach to rapidly identify previously unseen key SNPs. It is able to identify SNPs, which are significantly associated with the ADP platelet response that is used as a complex trait case study (phenotype). Several of these SNPs, such as rs6141803 of *COMMD7* and rs12953 in *PECAM1*, have independently confirmed associations with cardiovascular diseases (CVDs) according to other unrelated studies, suggesting that the RAPIDSNPs is robust in identifying key genetic variants. This approach provides an important step towards addressing the problem of 'missing heritability' through enhanced detection of key genetic variants (SNPs) that are associated with continuous complex traits/disease phenotypes. Moreover, the identified key SNPs might be indicating novel biological loci, which may require further attention and worth pursuing.

2.1 Introduction

Genetic association studies (GASs) allow scientists to study and analyse SNPs associated with complex traits or diseases. The traditional approach for genetic association (GA) analysis is to analyse one SNP at a time. However, multiple SNP analysis approaches have recently received much attention, and different strategies have been designed and adopted (Hoggart et al., 2008). For instance, the widely used standard multiple SNP analysis approach is the forward stepwise method (Cordell and Clayton, 2002). Other approaches include variants of penalised regression methods (Ayers and Cordell, 2010; Wu et al., 2009) and a compendium of the burden tests methods for analysing and detecting rare variants (Han and Pan, 2010; Hoffmann et al., 2010; Li and Leal, 2008; Liu and Leal, 2010; Morgenthaler and Thilly, 2007). Besides these improvements, approaches that are computational and bioinformatics-based, are likely to complement the biostatistical methods and further improve crucial SNPs identification, and hence, further addressing missing heritability (Eichler et al., 2010; Manolio et al., 2009; Moore et al., 2010). Here a novel, integrated, heuristic-based, hybrid analytical computational pipeline (RAPIDSNPs), for rapidly detecting novel or key genetic variants that are associated with complex traits continuous phenotype is described. The pipeline combines the power of random forests (RF) (Breiman, 2001) and regularised regression methods, using ridge and least absolute shrinkage and selection operator (lasso) (Hoerl and Kennard, 1970; Tibshirani, 1996) for the analysis of SNPs in GASs, in addition to the stepwise method. The RAPIDSNPs has also been coupled with an additional feature selection layer containing Boruta method (Kursa and Rudnicki, 2010), for further improving the SNPs identification. In brief, this pipeline describes a consensus model based on the RF for identifying key genetic variants (SNPs) for further biological interpretation or predictive purposes.

The RAPIDSNPs is able to select key SNPs associated with continuous phenotypic responses and has been applied to analyse the effect of multiple SNPs and loci associated with ADP

platelet responses. The RAPIDSNPs has identified several novel genetic variants significantly associated with platelet responses that were previously unidentified when only the standard forward stepwise method was used, yet it is also generally applicable for studying other continuous phenotypes.

The previous study by Jones et al. (2009) analysed genotyped SNPs, which were associated with four platelet responses: 1. P-selectin exposure (a marker of degranulation) in response to adenosine diphosphate (ADP) agonist (denoted by PA), 2. Fibrinogen binding in response to ADP (FA), 3. P-selectin in response to the GPVI specific agonist cross-linked collagen-related peptide (CRP-XL) (PC), and 4. Fibrinogen binding in response to CRP-XL (FC) (Jones et al., 2009). The genotyped SNPs data was obtained from the previous platelet responses functional genomic study (Jones et al., 2007). The key analytical method that was deployed was based on the forward stepwise method (Cordell and Clayton, 2002), which is argued to be statistically suboptimal (Harrell, 2001) and tends to omit key genetic variants, particularly those with strong linkage disequilibrium (Malo et al., 2008).

Here, the RAPIDSNPs is critically evaluated against the previous method using the same data (Jones et al., 2007), focusing on the ADP platelet responses (i.e. PA and FA). Furthermore, it is shown that using RAPIDSNPs, enhances the ability to identify key significant SNPs that are associated with ADP platelet responses phenotypes while also assessing their confidence level. Several of these SNPs were missed when previously analysed using the standard biostatistical forward stepwise method. Several of these SNPs such as rs6141803 of *COMMD7* and rs12953 in *PECAM1* have been also independently confirmed associations with cardiovascular diseases (CVDs) in other unrelated studies, suggesting the approach's robustness in identifying key genetic SNPs.

Additionally, the RAPIDSNPs was tested with the age covariate and demonstrated that it has promising potential in accounting for further heritability of platelet responses and other continuous complex traits phenotypes, and in reducing the difficulty in finding 'missing heritability'. Moreover, the RAPIDSNPs is useful in the genetic association studies where the genotyped SNPs data are highly dimensional (p > n).

2.2 Methods

2.2.1 Data acquisition and pre-processing

2.2.1.1 ***Data acquisition***

The data containing SNPs was acquired from the Jones et al. arising from the Bloodomics project (Jones et al., 2007). The data consisted of nearly 1553 SNPs from 512 individuals. The data is therefore highly dimensional with the number of all SNPs 'p' greater than the number of observations 'n' (i.e. p > n). The data was represented based on the structure described in the section 1.4.1.1 in the previous chapter. The phenotypes contained is the standardised logit transformed of two ADP platelet responses, i.e. PA, and FA. These were quantitative and essentially continuous trait phenotypes and measured in a previous study by flow cytometry through the expression level of the two released molecules, i.e. fibrinogen (F) and P-selectin (P) after the platelet has been activated by agonists ADP (Adenosine diphosphate). Each SNP in a column is genotyped by using numeric factor variables 1, 2, and 3 for major homozygous (dominant allele), heterozygous, and minor homozygous (recessive allele) respectively, for each individual in a population, Table 2.1.

individual	snp1	snp2	snp3	snp4	snp5	snp6	snp7	snp8	snp9	snp10	snp11	snp12	snp13	snp14	snp15	snp16	snp17	snp18	snp19	snp20
	rs4127084	NT_00448	rs4127085	rs4130621	rs4126426	rs4127321	rs4127321	rs4129959	NT_00448	NT_00448	rs4127321	rs4126400	rs4126400	rs4126509	rs4130417	rs4130821	rs4130559	NT_00540	NT_00540	rs4131
	FCER1G	FCER1G	FCER1G	SHC1	SHC1	PEAR1	PEAR1	PEAR1	PEAR1	PEAR1	PEAR1	LCK	LCK	IRS1	PIP5K3	PIP5K3	PIP5K3	PIP5K3	PIP5K3	PIP5K
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	
2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
4	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	
5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
6	1	1	1	1	1	1	1	1	1	3	1	1	1	1	1	1	1	1	1	
7	2	1	1	1	1	1	1	1	1	1	2	1	1	1	1	2	1	1	1	
8	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	
9	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
10	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
11	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
12	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
13	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	
14	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
15	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
16	1	1	1	1	1	1	2	2	1	1	1	1	1	1	1	1	1	1	1	
17	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	

Table 2.1 Screenshot showing a spreadsheet table containing individual subjects and their genotyped SNPs. Each SNP in the column is named using SNP1, SNP2, …, SNP1553 and SNP's dbSNP reference id with its respective gene/locus.

2.2.1.2 ***Data pre-processing***

The SNPs dataset was pre-processed mainly to remove missing genotypes between the SNPs (column-wide) and observations or subjects (row-wide) (Figure 2.2 and Figure 2.3).

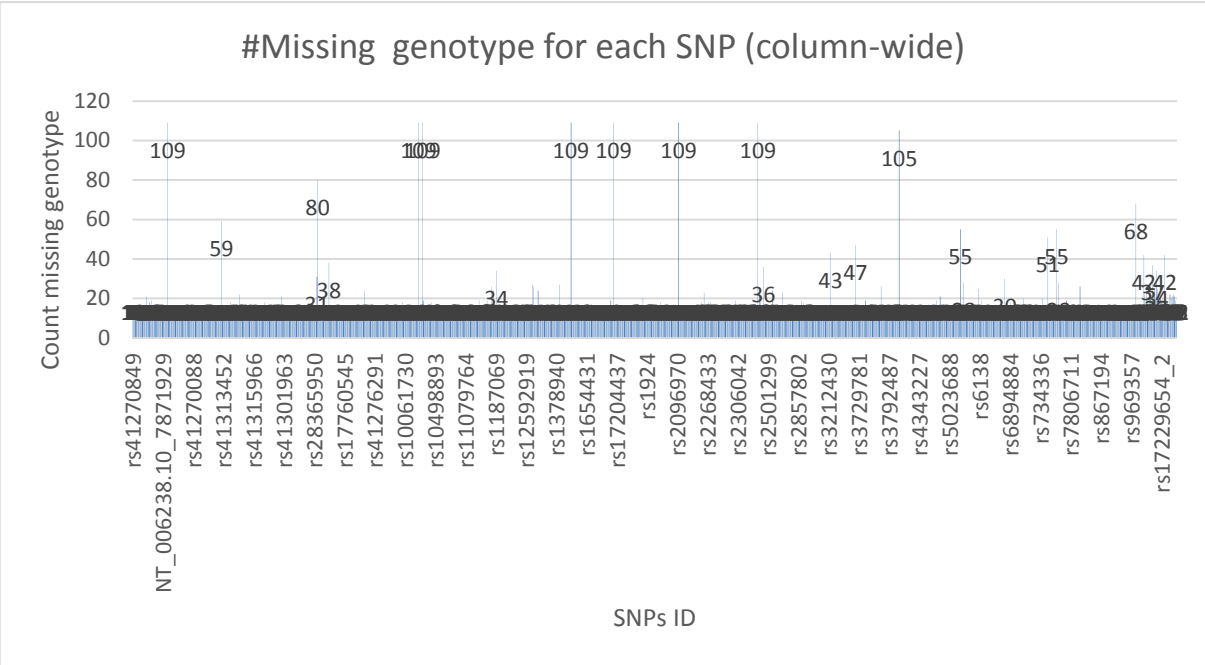


Figure 2.1 The missing genotype counts of few selected SNPs. The SNPs are represented horizontally with their dbSNP id, and the total number of missing genotypes for each SNP is represented on the vertical axis. The total missing genotype count for most SNPs are below 20. There are few observed SNPs with extreme missing genotypes.

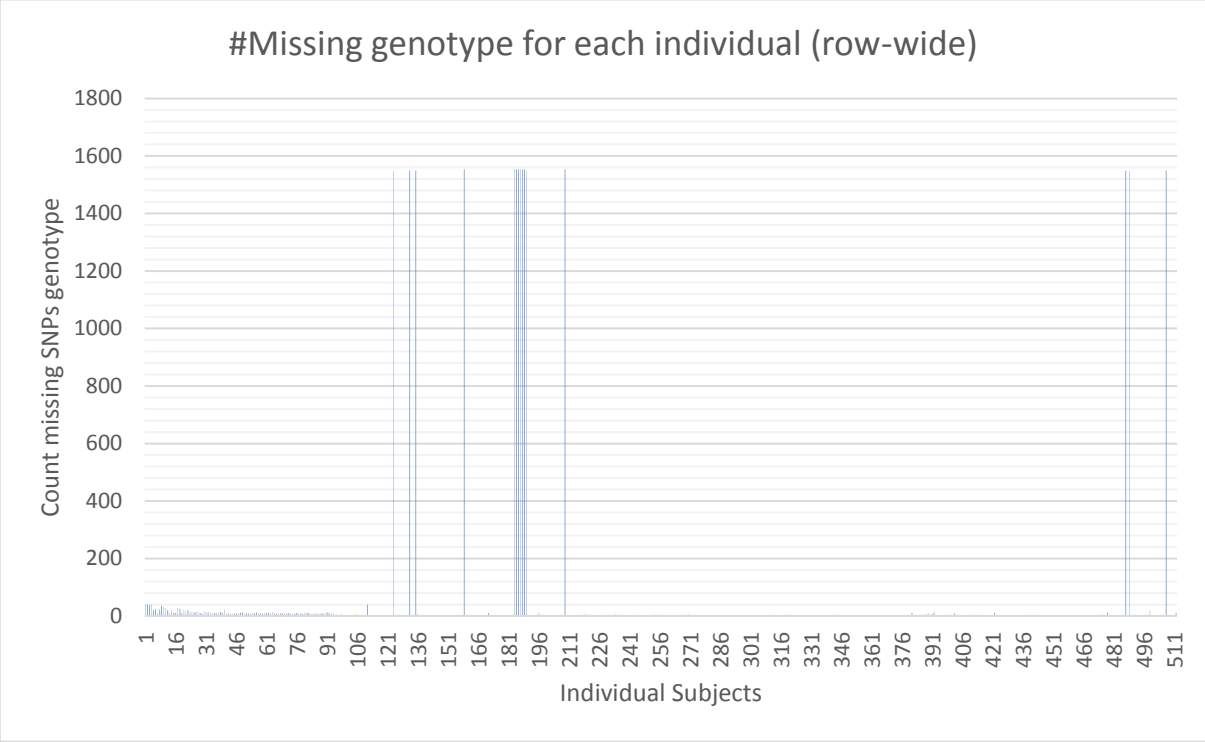


Figure 2.2 The distribution of missing genotypes among the individual subjects. Many individuals have relatively small number of missing genotype in each row with some few exceptions, whom their genotypes seem to be not recorded in entirety. These exceptional subjects were then completely removed from the table. The meaningful strategies were then applied to deal with the remaining missing genotypes.

In dealing with the missing genotypes, three different strategies were applied based on the presence of missing genotypes in SNPs (column-wide) and individual cases (row-wide) as described below:

In the first approach, since 72.33% of subjects possessed missing genotype values across SNPs, then all missing genotypes among subjects and SNPs were removed crosswise. This was performed after first omitting subjects and SNPs with a large number of missing genotypes (Figures 2.1 and 2.2). Then, the resulting dataset (named dataset 1) contained 462 subjects, i.e. 93% of all 497 subjects, with 1430 SNPs.

In the second approach, the dataset (dataset 2) was obtained after imposing threshold value based on the purity of each SNP genotype in the column. The SNP genotype with above 97% purity was removed from the set, i.e. the SNP with 97% threshold or above, of either 1, 2, or 3 genotypes was removed from the dataset. The major assumption is that the particular SNP's

genotype is conserved across the population, which genetically means does not exhibit any variability (Foulkes, 2009). Then random imputation method (Zhu, 2014) was applied to replace the remaining missing values with the most frequent genotype across the SNPs based on the distribution of the SNPs genotypes (1, 2, or 3). Eventually, the resulting dataset for analyses contained 497 subjects and 1270 SNPs.

In the third strategy, each row containing case with missing genotypes was completely removed by using default handling based on the 'complete case analysis' theory (Briggs et al., 2003). After doing this, 27.77% of the subjects remained as complete cases (138 subjects with 1553 SNPs). This dataset was named as dataset 3.

The entire pre-processing was carried out using R scripts (RC Team, 2014), in additional to other software tools such Microsoft Excel.

2.2.2 The computational pipeline (RAPIDSNPs)

The RAPIDSNPs aims to address the screening phase of the integrated omic analyses pipeline, which was discussed in the section 1.10.1 in the previous chapter.

2.2.2.1 ***The general method***

The RAPIDSNPs is based on the random forests (RF) (Breiman, 2001). The use of RF as an efficient tool for dealing with high dimensional data in the biomedical and life science has been elucidated in this previous review (Touw et al., 2013). The RAPIDSNPs follows a two staged analyses involving RF based on the work of (Schwarz et al., 2007), which is a standard for SNP discovery, as further explained by Goldstein et al. (Goldstein et al., 2011). The detailed description of the RAPIDSNPs approach is as follows

The random forests (RF) models were iteratively trained by using the above data subsets (Section 2.2.1). For each subset, the models were used to select the useful k SNPs from p. In this case, for each iteration based on the *ntree* (the number of trees used in generating RF model), an RF regression model was trained for both ADP platelet responses (PA and FA) in the dataset using all p SNPs. Then, the top 40 (k) among the overall ranked SNPs were selected using the permutation variable importance (VI) feature score measure (Breiman, 2001). An approximation of $\sqrt{p}$ as a cut-off value was used for selecting the top ranked k SNPs in each platelet responses. The k SNPs were used as a baseline for the downstream selection of the key significant SNPs in the pipeline.

For each iteration, the RF model was retrained using the k SNPs to examine whether the model has improved. The performance improvement was observed with the increase in the value of *ntree*, starting from 500, up to 3000 trees (i.e. 500, 1000, 2000, 3000 for iterations 1, 2, 3, and 4 respectively) where the models exhibited a stable performance. The relative increase of *ntree* was shown to significantly increase the performance, and proven to enhance the selection of the relevant variables (Strobl and Zeileis, 2008). The performance of the RF models was evaluated using equation (1).

$$R^2 = 1 - \frac{\sum(P_{observed} - P_{predicted})^2}{\sum(P_{observed} - \bar{P}_{observed})^2} \quad (1)$$

Where:

R^2 is the root mean squared, $P_{observed}$ and $P_{predicted}$ are observed and predicted platelet responses respectively for each of the FA, and PA. $\bar{P}_{observed}$ is the mean platelet responses for each of the FA, and PA.

For each iteration, the k SNPs were further passed through the designed layer of (regularised) regression methods ensemble, which was used to find highly significant SNPs associated with platelet responses. The rationale was that devising this layer would potentially increase the likelihood of identifying many significant SNPs based on the varying performances of the individual methods (Hastie et al., 2005). An additional aim in applying this layer was to increase the power of detecting significant SNPs that are likely to be missed by any of the other methods.

In the implementation of this layer, ridge and lasso were used in additional to the stepwise forward regression methods. The forward stepwise method was initially used to examine the number of SNPs that would have been selected relative to the previous study using the same data (Jones et al., 2009). Lasso was included to retain potentially sparse interactions among the genetic variants (Wu et al., 2010). Ridge regression was applied to take into account potential multicollinearity among SNPs, particularly those with strong linkage (Malo et al., 2008).

The SNPs resulting from each model, which was generated from the different selected regression methods were collated and tested to find those that were significantly associated with PA and FA platelet responses. The significant SNPs from each method were parametrically tested and selected based on the cut-off p-value of <= 0.01.

Tables 2.2, 2.3, and 2.4 demonstrate the effect of relatively increasing the RF's *ntree* parameter on the variance of k SNPs for PA platelet response and significance of regression models for the datasets 1, 2, and 3 respectively.

Random Forests (RF) Run				RF+Stepwise	RF+Ridge regression	RF+Lasso
#Iterations	**Number of trees (ntree)**	**% Variance all (p) SNPs**	**% Variance k SNPs**	**Model significance (r-squared & p-values)**	**Model significance (%Variance)**	**Model significance (r-squared & p-values)**
1	500	-0.67	8.58	0.088 & 4.965e-09	14.7	0.096 & 1.83e-09
2	1000	0.43	13.7	0.11 & 4.771e-11	17.83	0.10 & 8.1e-10
3	2000	0.23	16.66	0.111 & 5.4e-11	18.77	0.138 & 7.108e-13
4	3000	0.51	17.94	0.13 & 1.49e-12	17.5	0.16 & 6.61e-15

Table 2.2 The performance evaluation of the models for PA in the pipeline (RAPIDSNPs) using dataset 1. For each iteration, the RF model performance was evaluated based on the increase of % variance when the model was run using all p SNPs and selected k important SNPs. This was further followed up by passing the k SNPs into the multiple regression methods, where the confidence of their models was subsequently evaluated. This was repeated until the models showed convergence. Note, the % variance is a negative number (-0.67) in the first iteration, i.e. ntree=500. The negative number indicates that the prediction is very poor due to incorporating all SNPs (p) in the full model, a situation where many bad variables (SNPs) might be included *(Genuer et al., 2010; Strobl et al., 2008)*.

Random Forests (RF) Run				RF+Stepwise	RF+Ridge regression	RF+Lasso
#Iterations	**Number of trees (ntree)**	**% Variance all (p) SNPs**	**% Variance k SNPs**	**Model significance (r-squared & p-values)**	**Model significance (%Variance)**	**Model significance (r-squared & p-values)**
1	500	-0.5	10.05	0.1016 & 1.761e-10	19.44	0.1085 & 6.363e-11
2	1000	-0.5	15.06	0.09728 & 2.485e-10	20.94	0.1315 & < 3.787e-13
3	2000	-0.19	15.18	0.09766 & 2.256e-10	18.41	0.1517 & 8.44e-15
4	3000	-0.08	16.69	0.16 & 3.353e-11	20.68	0.1512 & 9.774e-15

Table 2.3 The performance evaluation of the models for PA in the pipeline (RAPIDSNPs) using dataset 2. The relative similar performance can be observed as it was in the dataset 1.

Random Forests (RF) Run				RF+Stepwise	RF+Ridge regression	RF+Lasso
#Iterations	**Number of trees (ntree)**	**% Variance all (p) SNPs**	**% Variance k SNPs**	**Model significance (r-squared & p-values)**	**Model significance (%Variance)**	**Model significance (r-squared & p-values)**
1	500	-4.72	21.07	0.323 & 1.256e-10	17.93	0.3348 & 9.909e-11
2	1000	-6.27	22.81	0.2791 & 2.634e-09	17.89	0.335 & 9.742e-11
3	2000	-5.32	27.65	0.2097 & 1.462e-07	19.02	0.2314 & 6.242e-08
4	3000	-5.9	24.71	0.3211 & 3.436e-10	15.92	0.2558 & 1.923e-08

Table 2.4 The performance evaluation of the models for PA in the pipeline (RAPIDSNPs) using dataset 3. The relative similar performance can be observed as it was in the dataset 1 and 2.

Thus, from Tables 2.2, 2.3, and 2.4, the increase in the variance explained by the RF, and confidence of the intermediate regression models might be an indicative feature of the importance of the selected k and highly significant SNPs respectively. A similar pattern is observed when the pipeline is validated using the simulated data (see a section titled 'Validation of the pipeline'). This further supports the work of Paul et al. (2013) and (Strobl and Zeileis, 2008) who showed that the variables selected using the VI measures are likely to be statistically significant, and the increase in the value of ntree plays a significant role in the selection of the relevant variables respectively. The k SNPs from the optimal or converged RF model (i.e. when the ntree=3000) were thus used to find the most significant and key SNPs in the final consensus approach.

Furthermore, for each iteration the k SNPs were alternatively passed through Boruta method (Kursa and Rudnicki, 2010) layer, which is an RF-based method normally used to select all relevant important features from the RF model. The Boruta method has previously shown the relative robustness in selecting potentially important genes (Kursa, 2014). In this approach, it was applied to enhance the consensus during the identification of the most significant SNPs by

independently examining the significant SNPs relatively to those selected by the regression methods layer in addition to new ones. The Boruta method finds k′ relevant (important) SNPs from k. The significance (or the importance) of SNPs in the Boruta method is measured using the Z-score. Once the optimal state of the pipeline was determined (i.e. in convergence), the SNPs from each method in the regression layer and Boruta were then extracted and compared to discover which of those were found to be the most significant by consensus, indicating to be key genetic variants (Figure 2.3).

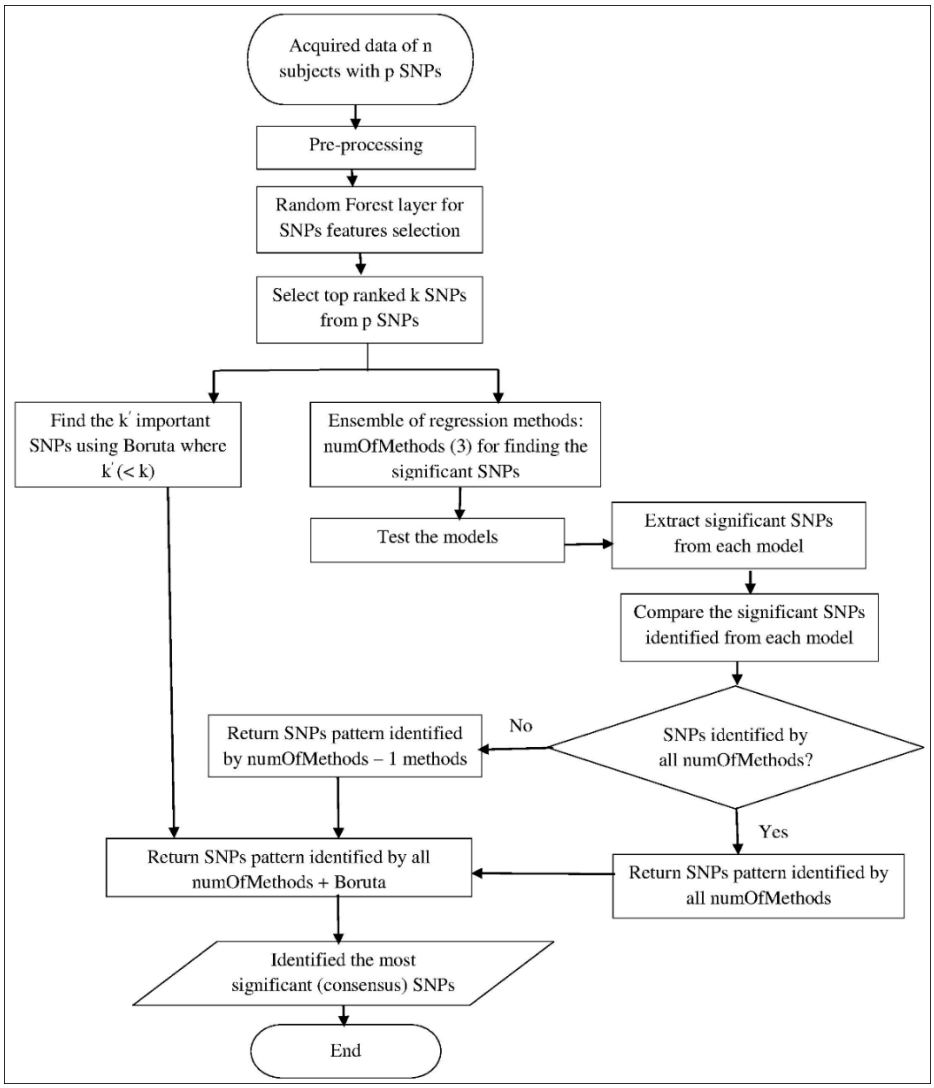


Figure 2.3 Flowchart showing the general methodological approach underpinning the RAPIDSNPs. In high dimensional genetic data of n samples with p genotyped SNPs, the number of SNPs was first reduced from p to k by means of the RF layer. The selected k SNPs were further reduced by means of two alternative methods, the ensemble of three regression methods and the Boruta method. The most significant SNPs are those that are selected by both methods, i.e. in consensus during the final iteration.

Based on the examined significant SNPs in different intermediate models in each iteration (i.e. ntree) of the pipeline, a confidence level was then assigned during the final iteration to verify that the selected key significant SNPs were not false positives. The confidence level also allows to ensure that true key significant SNPs (true positives) are not rejected, due to either being selected by a single method or being completely missed out in the final iteration when the pipeline converges. In order to assign a confidence score, a plot was created showing the frequency of the selected significant SNPs in the intermediate models in every iteration (see the Results section). The higher the frequency of appearance in the intermediate models, the greater the confidence score, or higher likelihood of being true key significant SNPs (true positives), i.e. during the observed convergence, if the same SNP appears in different intermediate models and in several iterations, then it is more likely to be a true positive. The confidence level of the selected significant SNP was then determined by taking the ratio of the frequency of appearance of a SNP (p_m) in the intermediate models in all iterations (i.e. ntree=500, ntree=1000, ntree=2000, and ntree=3000) to the normalised total number of the models multiplied by total number of iterations. Equation (2)

$$Confidence\ for\ SNP\ (P_m) = \frac{frequency\ of\ SNP\ (P_m)\ in\ the\ models\ in\ all\ iterations}{total\ number\ of\ models \times total\ number\ of\ iterations} \qquad (2)$$

From equation 2, a minimum threshold confidence level can be set, for instance, any score greater than 0.5 is more likely to be a true positive significant SNP.

2.2.2.2 ***Detail theoretical underpinnings and implementation of the RAPIDSNPs***

2.2.2.2.1 Random Forests (RF)

The RF is a tree-based ensemble machine learning algorithm, which encapsulates the data resampling idea. It is a variant of resampling method, known as bagging, which was proposed by Breiman (Breiman, 1996). The RF tends to construct several ensembles of tree-based models from different drawn bootstrap samples from the original sample data taken with/without replacement for the purpose of improving learning through aggregating all models (regression) or majority selected class votes (prediction). This might be determined based on the response variable type, i.e. the response variable is whether a categorical or continuous (Breiman, 2001; Liaw and Wiener, 2002)..

The mechanism of RF

The underlying mechanism of RF algorithm generally works as follow:

1) For b = 1 to B (Total number of trees)
    a. Draw a training set or 'bootstrap' **Z** of size k from the original data set (n).
    b. Grow the random forest tree T_b to the bootstrapped data by recursively repeating the following steps for each terminal node of the tree, until the minimum node size k_{min} is reached.
        i. Select m variables at random from the p variables.
        ii. Pick the best variable/split-point among the m.
        iii. Split the nodes into two daughter nodes.
2) Produce the output of trees ensemble.
3) Aggregate information from the B trees such as majority voting for classification.

Therefore, for regression the following function might be used:

$$\hat{\mathcal{F}}_{rf}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T(x) \quad (3)$$

where: $T(x)$ is the total number of forests (tree ensemble) generated from the data.

The key thing about RF is that trees are not pruned like the classical decision trees, since, the random selection of candidate variables ensures a low correlation between trees and avoiding excessive training of forests (Breiman, 2001). The unselected data during the bootstrapping process of RF is called out-of-bag (OOB). This is used as test data in the built-in cross-validation mechanism and for finding variable importance using the permutation score.

RF and variable importance (VI) measure: Ranking

RF can estimate the importance of variables for classification or prediction (Breiman, 2001). Variable importance (VI) estimates might be useful in understanding the relevance of variables in the given data set through the importance scores (Breiman, 2001; Breiman and Cutler, 2004; Liaw and Wiener, 2002). In the RAPIDSNPs, the RF is used as a method for selecting initial important SNPs for further downstream analyses using VI measure. And this is a gold standard approach for SNPs discovery in the genetic association studies (Goldstein et al., 2011; Schwarz et al., 2007). Different reviews have further discussed the suitability of this approach for high-dimensional data in the genomic and genetic association analyses (Chen and Ishwaran, 2012; Szymczak et al., 2009; Touw et al., 2013; Verikas et al., 2011).

Gini index and permutation importance scores are the two most common used VI measures in RF (Breiman, 2001; Breiman and Cutler, 2004; Liaw and Wiener, 2002). For the regression purpose, permutation importance is the used VI measure (Breiman, 2001). In this case, to

calculate VI, the variable is permuted in the OOB, which is the original data and that is unused during bootstrapping or training. The prediction error estimate is calculated using the OOB. The difference between this error involving permutation and the OOB error without permutation is computed for each tree. Finally, the average for all trees is calculated and normalised by its standard deviation for the permutation scores of a variable, and hence its importance. The variable becomes much more important if it contains far larger permutation importance (Breiman, 2001).

VI measures using permutation scores can be described using Figure 2.4

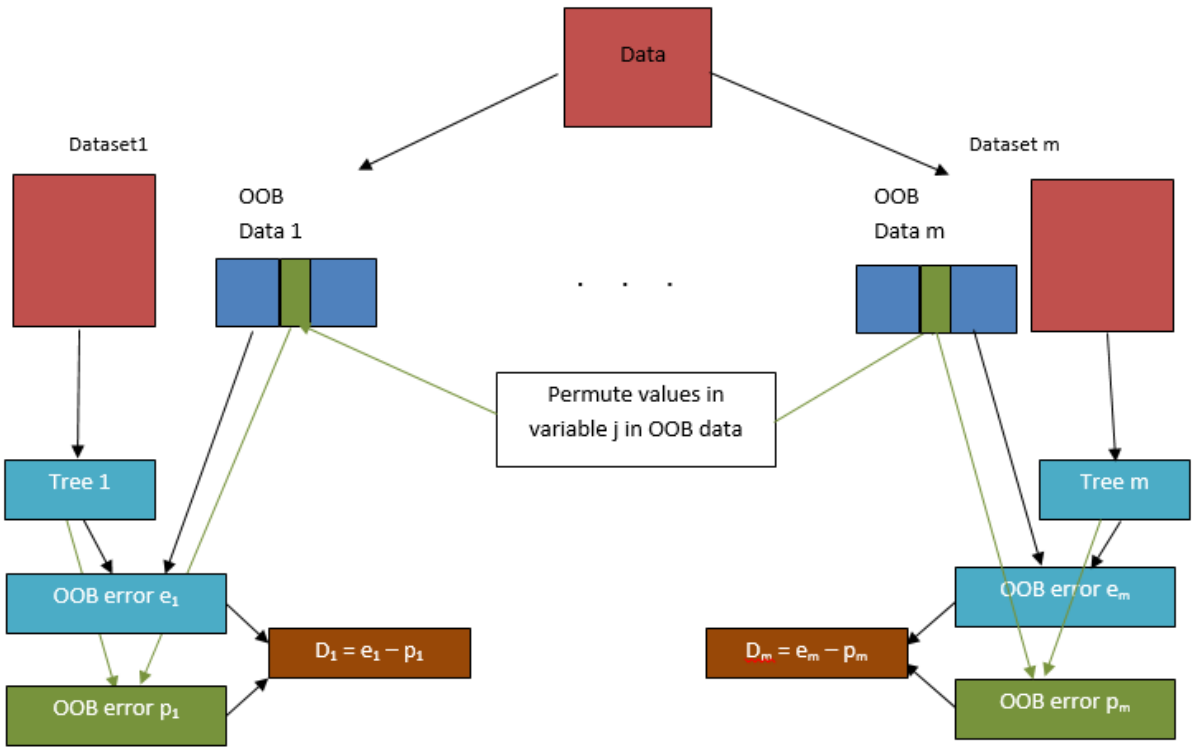


Figure 2.4 The mechanism of VI measure using permutation score. The OOB data is used for testing prediction accuracy of the training data in dataset i and when permuting with variable j. The average of the differences in the prediction error is used to compute the VI. The diagram was adapted from[1].

Therefore, from the diagram above VI can mathematically be computed as follows:

[1] http://stat.ethz.ch/education/semesters/ss2012/ams/slides/v10.2.pdf

$$\bar{D} = \frac{1}{m}\sum_{j=1}^{m} D_i \quad (4)$$

where, $\bar{D}$ is an average of individual difference of prediction error (D_i) for each OOB data. Then to obtain VI, the variance for each mean squared error prediction is computed as follows:

$$s_D^2 = \frac{1}{m-1}\sum_{j=1}^{m}(D - \overline{D})^2 \quad (5)$$

where: s_D^2 is the variance due to all permutations of OOB dataset (1,..., m).

Hence:

$$VI = \frac{\bar{D}}{s_D} \quad (6)$$

For GASs, the recommended (VI) measure for selecting and ranking important SNPs is permutation score rather than Gini index (Chen and Ishwaran, 2012). This is because Gini index tends to favour SNPs with large MAFs (Nicodemus et al., 2010; Nicodemus and Malley, 2009).

The randomForest (Liaw and Wiener, 2002) package in R language (Ihaka and Gentleman, 1996) is used to run the RF layer.

2.2.2.2.2 Regression methods (Regression layer)

Generally, the regression methods, which were used in the RAPIDSNPs pipeline, are based on the standard linear regression model given by equation (7).

$$Y = X\beta + \varepsilon \tag{7}$$

where:

Y is the response phenotype of concern (i.e. ADP platelet responses PA, and FA for individuals), which is (*n x 1*) vector of dependent variables; X is an (*n x p*) design matrix, in this case, are the SNPs genotype-coded with 1 for major homozygous, 2 for heterozygous, and 3 for minor homozygous; β is a (*p* x 1) vector of regression coefficients β_j, $j = (1,\ldots,p)$; and ε is an assumed vector of normally distributed random errors with mean 0 and unit standard deviation. So the model is a relationship between the continuous phenotype Y (i.e. ADP platelet responses) determined by weighted SNPs X_p of *n* individuals.

2.2.2.2.2.1 Stepwise forward regression

The stepwise regression was applied with the forward selection method, after filtering the SNPs using RF. Generally, the forward stepwise selection method starts with a null model and allows one SNP at a time to enter the model, based on which SNP is most correlated with each of the platelet responses, i.e. the addition of the SNP in the model depends on the SNP that gives the highest significant improvement in fit (Cordell and Clayton, 2002). The selected SNPs in the stepwise model were tested for significance using the Wald test. The stepwise regression was implemented using the LEAPS package (Lumley, 2015) in R.

2.2.2.2.2.2 Shrinkage (regularised or penalised) regression methods

Shrinkage methods (Hastie et al., 2005), use a regularisation strategy to penalise SNPs from k SNPs from the RF layer, assuming that the underlying RF functioning might select SNPs that

are not significant. Thus, the shrinkage methods further simplified and enhanced the selection of highly significant SNPs. The shrinkage methods were applied using the ridge regression and lasso with R packages 'ridge' (Cule, 2015) and 'glmnet' (Friedman et al., 2010) respectively. In applying the glmnet package, the family option is set to "gaussian" as the response phenotypes (platelet responses) are quantitative and assumed to follow the Gaussian distribution.

Ridge regression (RR)

Ridge regression (Hoerl and Kennard, 1970) is the method that shrinks regression parameters by penalising their size and reduced towards zero. The ridge regression was applied to ensure that potential collinear SNPs were kept in the models, particularly those with strong linkage (Malo et al., 2008). Thus, based on the model given by equation (6) above, the regression coefficients estimates could be determined using the ordinary least square method (OLS), which is the standard approach and is given by equation (8).

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{8}$$

However, this equation does not work particularly in the context of genetic data where collinearity is common among SNPs due to the high linkage (Malo et al., 2008). RR shrinks regression coefficients by penalising their size and reducing towards zero using the computed ridge shrinkage parameter (lambda). The optimal shrinkage parameter helps to identify the regions where the model parameters are stable and control the classical trade-off between the high bias and variances, which commonly occur when there are a large number of parameters and collinearity among SNPs. Equation (9) shows the ridge regression model:

$$\hat{\beta} = (X'X + \lambda I)^{-1}X'Y \quad (9)$$

where, the lambda (λ) is a ridge parameter, which determines the degree of shrinkage. I is a p x p identity matrix. Adding the term λI in the model, reduces the coefficient estimates toward each other, potential collinearity among SNPs, and eliminates the possibility of matrix $X'X$ being singular. The parameter λ is selected between 0 and ∞ values. If λ = 0 then the RR model is turned to be OLS solution, and if λ = ∞ then the model would behave as if no parameters have been estimated, and the solution would be the mean of the response variable, i.e. $\bar{Y} = \beta_0$. An automatic lambda selection method proposed by (Cule et al., 2011) and the Wald test for testing the significant SNPs from the RR were used. In implementing RR, the ridge package (Cule, 2015) from the R language was used to generate RR models.

Lasso (least absolute shrinkage and selection operator)

Lasso (Tibshirani, 1996) was applied to the selected important SNPs from the RF model to reduce further the SNPs with insignificant low coefficients (Hastie et al., 2005), which might be selected from the RF filtering procedure. Additionally, lasso may possibly retains sparse interactions among the SNPs (Wu et al., 2010). The lasso model is given in equation (10).

$$\hat{\beta}^{lasso} = \underset{\beta}{argmin} \sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\,\beta_j\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \quad (10)$$

where:

y_i is the vector of particular response phenotype (platelet activation pathway among PA and FA) for observation i; X is a design matrix of SNPs and $\hat{\beta}^{lasso}$ are the coefficient estimates of the SNPs; the lambda term is the weight given for the regularisation term (L1 norm), which

sparsely picks the SNPs entering the model when the tuning parameter is very small or exactly zero.

The best SNPs from the lasso models were extracted through the selection of the smallest optimal lambda (or tuning parameter) using 10-fold cross validation (Motyer et al., 2011). The lasso models generated the sparse matrix of SNPs coefficients estimates. The SNPs with relative large coefficient estimates from the sparse matrix were selected and tested in a stepwise manner using the partial F-test (Kohannim et al., 2012) to determine the individual SNP's significance level in the model. For these analyses, the glmnet package (Friedman et al., 2010) implemented in R was used.

2.2.2.2.3 Boruta method

Boruta is all-relevant feature selection method, which provides an improved mechanism for selecting an important feature or variable from the RF using Z-score. It is a wrapper algorithm, which ranks the features from the RF through improved Z-score. The applied Z-score within Boruta provides the statistical significance, and hence the relevance of the selected important variable or feature. This is performed by comparing the relevance of the true feature to that of the random probe or permutation in the RF (Kursa and Rudnicki, 2010). In doing this, the method employs 'shadow attributes' whose importance are used to determine the true important attribute. Shadow attributes provide demarcation for examining whether the selected important features from the RF are truly important or are just due to the random fluctuations as the results of the underlying RF mechanism (Kursa and Rudnicki, 2010). Thus, adding Boruta layer to the RAPIDSNPs pipeline, potentially provides further enhanced consensus selection of the key SNPs, in addition to the regression layer.

Boruta was run using Boruta package in R (Kursa and Rudnicki, 2010).

2.2.3 The performance of the pipeline with the inclusion of covariates

The pipeline is specifically designed for analysing predetermined, genotyped SNPs to identify the most significant SNPs (key SNPs) that are associated with continuous complex trait phenotypes and would have been likely to be missed by other approaches such as stepwise. The pipeline was initially applied to alternatively analyse the combined effect of the SNPs and benchmarks the results against those obtained from the stepwise forward approach (Jones et al., 2009), which did not need to take into account the covariates, such as age, gender, height, weight, ethnicity, aspirin taking, medication, smoker, contraceptive pill, because they were already treated separately during the data pre-processing stage of the Bloodomics project (Jones et al., 2007).

Nevertheless, the pipeline have been re-tested to demonstrate the incorporation of an example key covariate for CVD: age. The approaches for handling covariates in determining the effect of SNPs on the phenotype using RF have been well elucidated by Nonyane and Foulkes (Nonyane and Foulkes, 2008). In running the pipeline, the age was included as a numeric type and potential predictor together with SNPs under the additive model.

Tables 2.5 shows the performance of the RF models when the pipeline is run with age as a covariate in identifying the most significant SNPs associated with PA platelet response.

Random Forests (RF) – SNPs *without* age incorporated as a covariate				RF - SNPs *with* age incorporated as a covariate	
#Iterations	Number of trees (*ntree*)	% Variance all (p) SNPs	% Variance k SNPs	% Variance all (p) SNPs	% Variance k SNPs
1	500	-0.16	*11.85*	0.17	*13.95*
2	1000	-0.5	*14.29*	-0.69	*14.46*
3	2000	-0.06	*16.86*	0.15	*18.54*
4	3000	0.33	*15.92*	0.12	*16.36*

Table 2.5 The performance of the RF with and without age as a covariate in determining the PA platelet response.

From Table 2.5, there are an observed marginal increase in the variation explained by the RF models when age is included as a covariate. The residuals plots are shown in the Figures 2.5

and 2.6. The significance of the regression models due to the covariate in the intermediate regression models are shown in the Table 2.6. Few intermediate models have higher significance in the early iterations when age is included as a covariate comparing than when it is excluded.

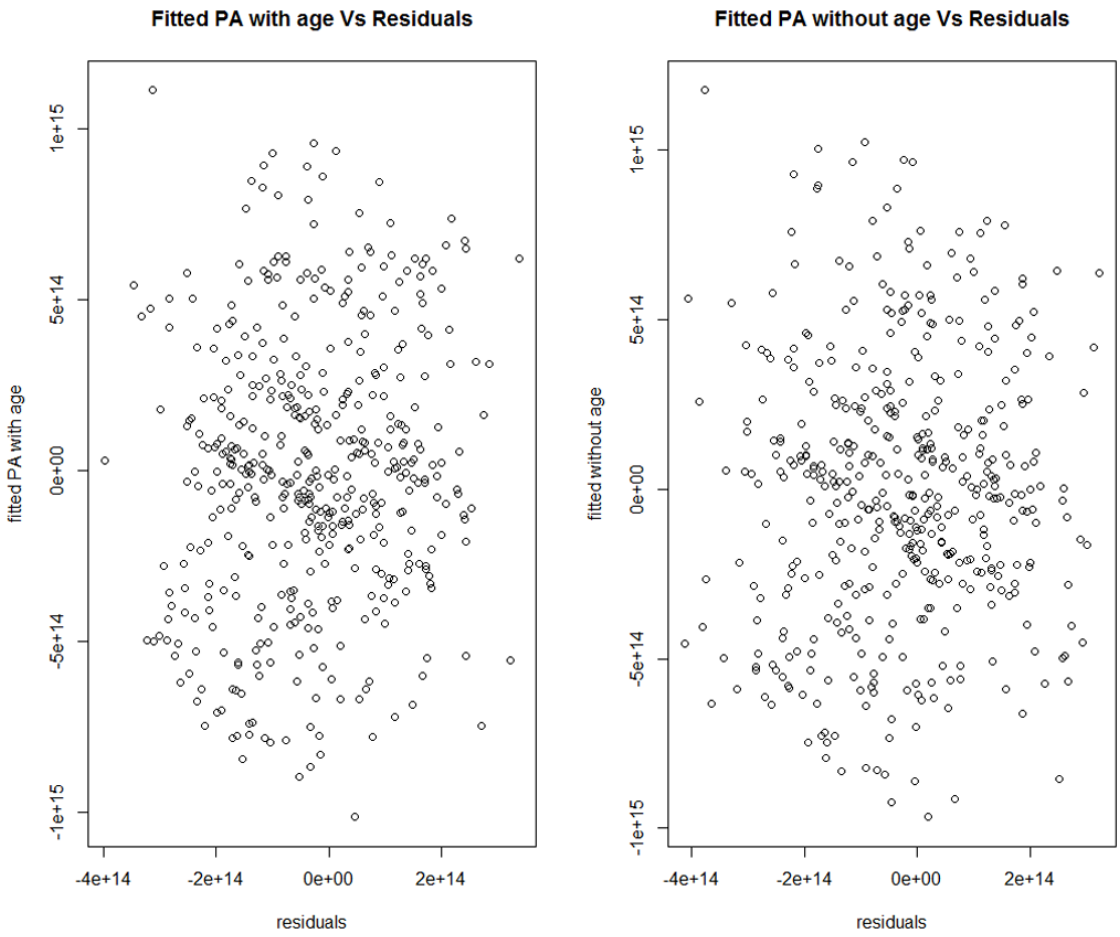


Figure 2.5 The residual plot when fitting the PA response using SNPs with or without age as a covariate.

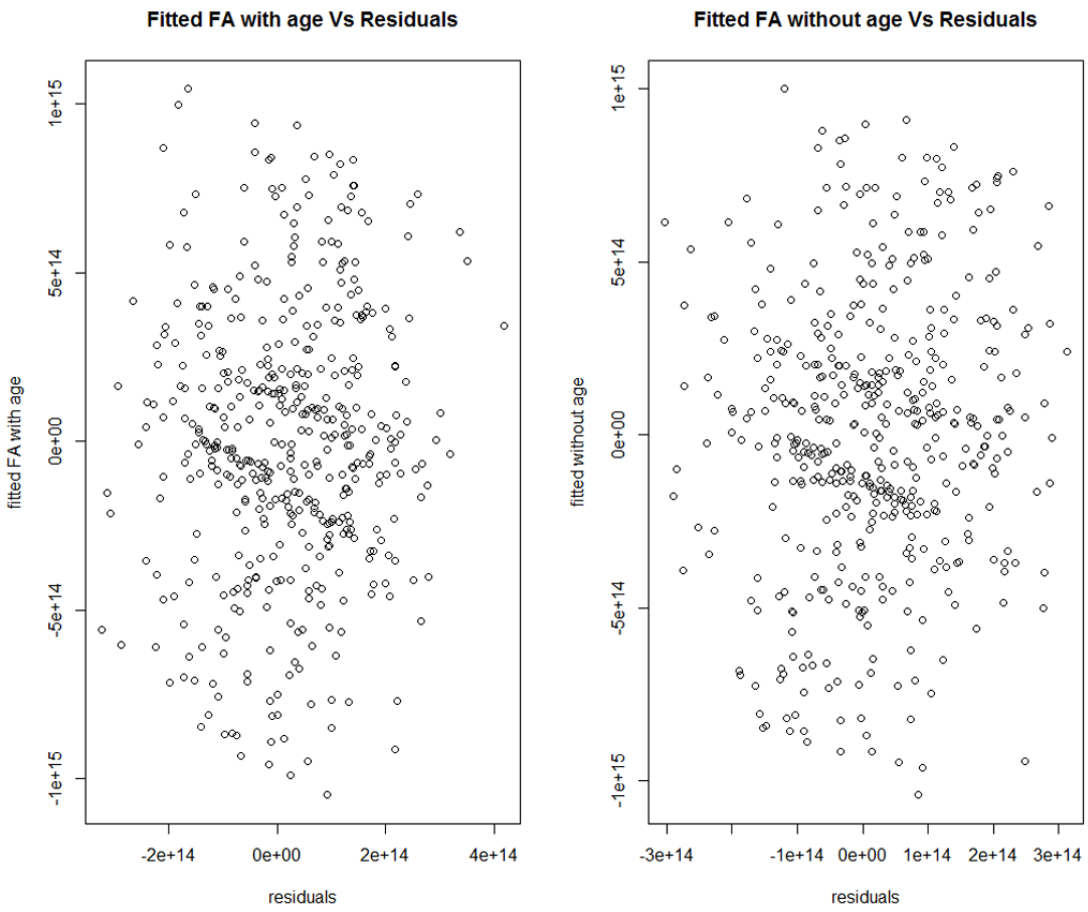


Figure 2.6 The residual plot when fitting the FA response using SNPs with or without age as a covariate

#Iterations	Intermediate models with no age covariate			Intermediate models with age covariate		
	RF+Stepwise	RF+Ridge regression	RF+Lasso	RF+Stepwise	RF+Ridge regression	RF+Lasso
	Model significance (r-squared & p-values)	Model significance (%Variance)	Model significance (r-squared & p-values)	Model significance (r-squared & p-values)	Model significance (%Variance)	Model significance (r-squared & p-values)
1	0.088 & 4.965e-09	14.7	0.096 & 1.83e-09	0.09 & 4.06e-09	18.21	0.097 & 1.38e-09
2	0.11 & 4.771e-11	17.83	0.10 & 8.1e-10	0.089 & 4.136e-09	18.01	0.097 & 2.63e-09
3	0.111 & 5.4e-11	18.77	0.138 & 7.108e-13	0.123 & 6.982e-12	15.92	0.12 & 2.98e-11
4	0.13 & 1.49e-12	17.5	0.16 & 6.61e-15	0.113 & 3.286e-11	13.53	0.096 & 8.64e-10

Table 2.6 The significance of the intermediate models due to the exclusion and inclusion of age as covariate to the PA platelet response.

2.2.4 Validation of the RAPIDSNPs pipeline

2.2.4.1 ***Validation of the pipeline***

To validate the pipeline, 460 subjects containing 1400 artificially genotyped SNPs with their associated continuous phenotype were randomly simulated. The R code snippet for reproducing the data is shown in Figure 2.7.

```
# Set seed for reproducing the data
set.seed(12345)

# Simulating artificial SNPs with their genotypes (independent variables) 1400 SNPs, and 460 subjects
df_SNPs <- data.frame(replicate(1400, sample(1:3, 460, rep=TRUE)))

# Simulating response variable (phenotype); normally distributed.
yres_phen <- rnorm(460)

# Combining the response variable and SNPs, i.e, adding a new column to the above created data frame of SNPs (df_SNPs)
df_SNPs["phenotype"] <- yres_phen
df_SNPs_yres_phen <- df_SNPs[, 1:ncol(df_SNPs)]

# Viewing the few first rows and columns of the created data frame
head(df_SNPs_yres_phen[,c(1:4,1401)])
```

Figure 2.7. The R code snippet for reproducing the simulated artificial SNPs and phenotype.

The simulated phenotype is a univariate normal distribution with n(0,1). The genotypes of these artificial SNPs follow the standard representation consisting of 1, 2, and 3, which represents major homozygous, heterozygous and minor homozygous respectively. This simulated data set was applied to the RAPIDSNPs. The RF and the multiple regression models using k SNPs were observed to improve as *ntree* was increased in each iteration starting from 500, 1000, 2000, until 3000 trees, where the variance and confidence of the models started to converge (Table 2.7).

Random Forests (RF) Run				RF+Stepwise	RF+Ridge regression	RF+Lasso
#Iterations	**Number of trees (ntree)**	**% Variance all (p) SNPs**	**% Variance k SNPs**	**Model significance (r-squared & p-values)**	**Model significance (%Variance)**	**Model significance (r-squared & p-values)**
1	500	1.14	11.84	0.12 & 7.342e-12	24.36	0.14 & 2.482e-13
2	1000	1.71	16.36	0.14 & 1.042e-13	27.51	0.20 & 2.2e-16
3	2000	2.55	21.31	0.15 & 1.082e-14	28.97	0.22 & 2.2e-16
4	3000	1.6	19.34	0.13 & 1.604e-12	28.2	0.19 & 2.2e-16

Table 2.7 The performance evaluation of the pipeline for the simulated SNPs. As *ntree* is increased in each iteration the resultant RF models for both full and simplified models with p and k SNPs respectively, the significance of the subsequent regression models also relatively increases, which may imply that selected SNPs are likely to be significant.

The models' patterns observed using the artificial SNP data are shown to reflect those observed with the real SNP data (Table 2.2, 2.3, and 2.4), even though the convergence, in this case, seems to be in the third iteration when the ntree was 2000.

2.2.4.2 ***Further validation of the RAPIDSNPs using second simulated data set***

The second simulated data set containing similar size with that of section 2.2.4.1 was used to further validate the pipeline. The simulated artificial genotypes within this set were not identical as were the previous set and thus, its artificial allele genotypes were more reflecting the real genotypes in the population based manner. In this regard, there were fewer minor alleles comparing to major and heterozygous alleles. The simulated phenotype used followed the same univariate distribution n(0,1). Table 2.8 shows the performance of the pipeline with this second simulated set.

Random Forests (RF) Run				RF+Stepwise	RF+Ridge regression	RF+Lasso
#Iterations	**Number of trees (ntree)**	**% Variance all (p) SNPs**	**% Variance k SNPs**	**Model significance (r-squared & p-values)**	**Model significance (%Variance)**	**Model significance (r-squared & p-values)**
1	500	-0.03	13.27	0.12 & 1.998e-12	25.82	0.10 & 2.601e-10
2	1000	-0.44	18	0.12 & 2.724e-12	28.61	0.16 & 3.3e-15
3	2000	0.05	20	0.15 & 1.69e-14	29.08	0.17 & 1.31e-15
4	3000	1.09	19.99	0.15 & 8.755e-15	27.67	0.20 & < 2.2e-16

Table 2.8 The performance evaluation of the pipeline for the second simulated SNPs with fewer minor alleles.

From Table 2.8, the models' performance pattern appears to be similar with the previous Table 2.7. Nevertheless, there is slightly decrease in significance of some intermediate models in some iterations.

Therefore, the similarity in the performance of the pipeline using both the real and simulated genotyped SNPs data sets, indicates that the pipeline is likely to be robust when applied to other continuous phenotypes.

Additionally, the identified key SNPs (Section 2.3.1.1) could be seen to be randomly associated with the phenotype due to randomly selection of the artificial simulated SNPs. Hence, the reasearcher further validated the robustness of the RAPIDSNPs by alternatively testing whether these key SNPs are truly associated with the simulated phenotype. In this regard, the key SNPs were initially identified (See section 2.3.1.2) and the simulated phenotype values were alternatively increased and decreased. Based on the identified artificially simulated key SNPs' coefficient estimates, the phenotype values were increased or decreased by the scale of (+/-0.2) under the same gaussian distribution. The identified key SNPs were then checked whether they are still associated with the phenotype under these alterations.

2.3 Results

2.3.1 The consensus approach for identifying key SNPs

Firstly, the new approach has identified several significant SNPs that are associated with both ADP platelet responses and are consistent with the previous study (Jones et al., 2009). Importantly, the method has also discovered numerous additional SNPs that are significantly associated with platelet responses and were not previously identified, or previously found to be insignificantly associated with platelet responses using the forward stepwise method.

Tables 2.9 and 2.10 show the overall significant and key SNPs identified by this pipeline and the previous method that are associated with PA, and FA platelet responses respectively using *dataset 1*. From the results, a consensus approach for the key SNPs identification is established, in which a SNP is identified as key if it has been selected by the three out of four methods within the pipeline.

			Stepwise (Jones et al 2009)	RF with Stepwise	RF with Ridge regression	RF with LASSO	RF with Boruta (P=0.01)	Consensus (3/4)
Platelet response type		Beta (+ve/-ve)	PA	PA	PA	PA	PA	PA
SNPs ID	Gene/Location							
rs17229705	*VAV3*	+ve	✓ (0.0009)	×	×	×	×	
rs3788337	*GNAZ*	-ve	✓ (0.0009)	×	×	×	×	
rs5227	*PTGS2*	+ve	✓ (0.01)	×	×	×	×	
rs1778614	*ITPR1*	+ve	✓ (0.003)	×	×	×	×	
rs246406	*ITGA2*	-ve	✓ (0.002)	×	×	×	×	
rs11631474	*MAP2K5*	+ve	✓ (0.007)	×	×	×	×	

rs851007	*MAPK 14*	+ve	✓ (0.003)	×	×	×	×	
rs6141803	*COMM D7*	-ve	×	✓ (0.0033)	×	✓ (0.000 6)	✓	✓
rs6442896	*ITPR1*	-ve	× (0.049)	✓ (0.0006)	✓(0.000 2)	✓(0.0 021)	✓	✓
rs3730051	*AKT2*	+ve	× (0.031)	✓ (0.0002)	✓ (0.0031)	✓ (0.000 2)	×	✓
rs1527480	*CD36*	-ve	× (0.449)	✓ (0.0021)	✓(0.000 8)	✓ (0.003 6)	✓	✓
rs8033381	*CSK*	-ve	× (0.792)	✓ (0.0018)	✓ (0.0082)	✓ (0.003 8)	×	✓
rs1006173 0	*ITGA2*	-ve	× (0.517)	×	×	✓ (0.000 5)	×	
rs2292867	*ITGB3*	-ve	×(0.039)	×	✓ (0.017)	✓ (0.008 0)	×	
rs2300065	*SKP1*	+ve		×	✓(0.013 8)	✓(0.0 164)	×	
rs3212391	*ITGA2*	-ve	×	✓ (0.0002)	×	×	✓	
rs6433658	*ITPR1*	+ve	×	×	×	×	✓	
rs6442895	*ITPR1*	-ve	×(0.029)	×	×	×	✓	
rs1704140 1	*ITPR1*	+ve	✓(0.003)	×	×	×	✓	
rs3212386	*ITGA2*	-ve	×(0.378)	×	×	×	✓	
rs33443	*ITGA2*	-ve	×(0.547)	×	×	×	✓	
rs26682	*ITGA2*	-ve	×(0.126)	×	×	×	✓	
rs3212418	*ITGA2*	-ve	✓(0.013)	×	×	×	✓	
rs1174255 8	*ITGA2*	-ve	×(0.713)	×	×	×	✓	
rs7568033	*NFE2L 2*	+ve	×	×	×	×	✓	

Table 2.9 Consensus identification of the most significant SNPs associated with PA platelet response in dataset 1. The consensus SNP is selected if it has been identified by at least three methods, which means it has higher significance and hence is more likely to be a key genetic variant. × indicates either the SNP was not identified by the method or previously identified as insignificant. ✓ indicates the SNP was identified by the method. Numbers inside the brackets after ✓ or × indicate p values

of the SNPs calculated using Wald test, or partial F-test. +ve/-ve indicate the sign of coefficient estimates of the SNPs from the regression models.

			Stepwise (Jones et al 2009)	RF with Stepwise	RF with Ridge regression	RF with LASSO	RF with Boruta (P=0.01)	Consensus (3/4)
Platelet response type		Beta (+ve/-ve)	FA	FA	FA	FA	FA	FA
SNPs ID	Gene/Location							
rs11637556	*MAP2K1*	+ve	✓ (0.005)	✓ (0.0007)	✓ (0.0083)	✓ (0.0008)	✓	✓
rs10429491	*JAK2*	-ve	✓ (0.0006)	×	×	×	×	
rs3729931	*RAF1*	+ve	✓ (0.0001)	×	×	×	×	
rs41305896	*ITGA2*	+ve	✓ (0.001)	×	×	×	×	
rs350916	*MAP2K2*	+ve	✓ (0.001)	×	×	×	×	
rs17786144	*ITPR1*	+ve	✓ (0.002)	×	×	×	×	
rs11264579	*PEAR1*	-ve	✓ (0.004)	×	×	×	×	
rs41304345	*MADD*	+ve	✓ (0.003)	×	×	×	×	
rs1388622	*P2RY12*	+ve	× (0.058)	✓ (0.0001)	×	×	✓	
rs2071676	*CA9*	-ve	×	✓ (0.0122)	✓(0.0058)	✓(0.0098)	✓	✓
rs1491978	*P2RY12*	+ve	× (0.092)	×	×	✓(0.0003)	✓	
rs1537593	*CD36*	-ve	× (0.731)	×	×	✓(0.0058)	✓	
rs9895150	*ITGB3*	-ve	× (0.177)	×	✓(0.0193)	✓(0.0141)	×	
rs1038639	*ITPR1*	+ve	× (0.138)	×	✓(0.0019)	✓(0.0006)	✓	✓
rs10499858	*CD36*	-ve	× (0.129)	✓ (0.0012)	×	×	✓	
rs7034539	*JAK2*	+ve	× (0.061)	✓ (0.0053)	✓(0.0058)	✓(0.0077)	×	✓
rs3742633	*PRKCH*	-ve	× (0.985)	✓ (0.0172)	×	✓(0.0075)	×	
rs41282607	*MAPK1*	+ve	× (0.2)	✓(0.0113)	✓(0.0034)	✓(0.0087)	✓	✓
rs41305272	*MAP2K5*	+ve	× (0.955)	×	✓(0.0127)	✓(0.0100)	✓	✓

rs7180408	*GTF2A2*	+ve	×	×	✓(0.0191)	×	×	
rs3736101	*MADD*	-ve	✓(0.015)	×	×	✓(0.0076)	×	
rs304076	*ITPR1*	-ve	× (0.395)	×	×	✓(0.0083)	×	
rs17204437	*P2Y12*	+ve	× (0.499)	×	×	✓(0.0010)	✓	
rs6787801	*P2Y12*	+ve	× (0.448)	×	×	×	✓	
rs3173798	*CD36*	-ve	× (0.085)	×	×	×	✓	

Table 2.10 Consensus identification of the most significant SNPs associated with FA platelet response in dataset 1. The consensus SNP is selected if it has been identified by at least three methods, which means it has higher significance and hence is more likely to be a key genetic variant. × indicates either the SNP was not identified by the method or previously identified as insignificant. ✓ indicates the SNP was identified by the method. Numbers inside the brackets after ✓ or × indicate p values of the SNPs calculated using Wald test, or partial F-test. +ve/-ve indicate the sign of coefficient estimates of the SNPs from the regression models.

In Figure 2.8 A – B Venn diagrams are provided for the overall significant and key SNPs identified by the multiple regression methods layer within the pipeline using dataset 1. These diagrams provide an alternative way of observing the key SNPs lying within the intersection regions.

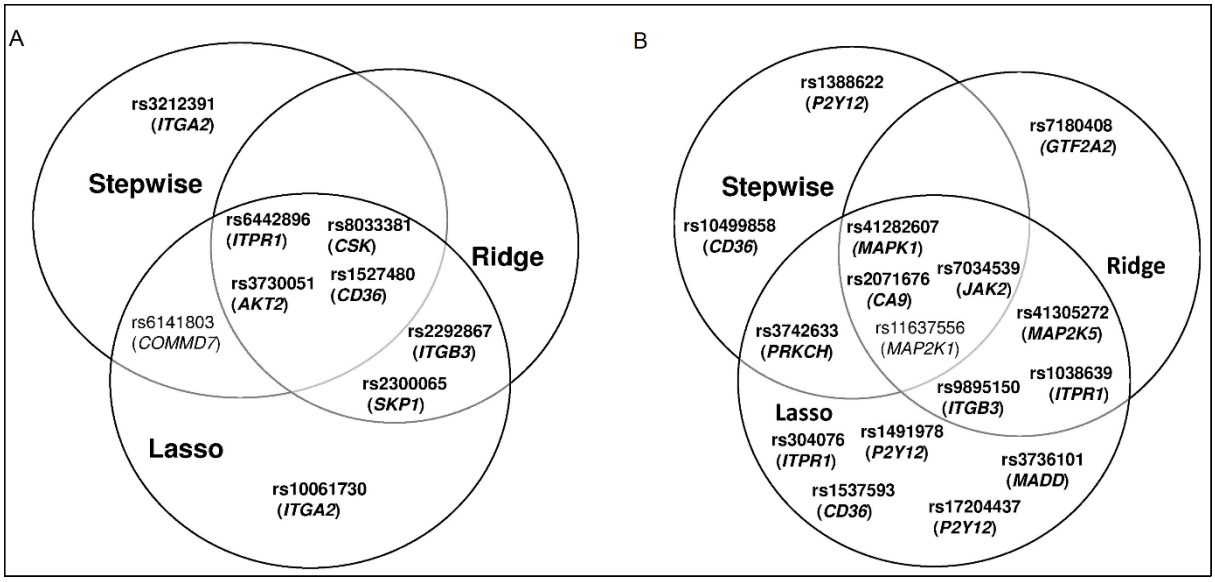


Figure 2.8 Venn diagrams showing the identified significant and key SNPs from the regression layer in the pipeline using dataset 1. The identified significant SNPs in (A) and (B) are associated with PA (p-selectin in response to adenosine diphosphate) and FA (fibrinogen binding in response to adenosine diphosphate) respectively. The newly detected SNPs or those reported as insignificant in the previous study are shown in bold. The key SNPs are found in the intersection regions and detected by a consensus of the three methods

Using the Boruta method layer, several of the identified significant SNPs that were found to be associated with both ADP platelet responses, were also closely similar to those identified by the regression methods layer (Figure 2.9 A – B). Thus, this similarity further improved the consensus selection of the most significant SNPs associated with the ADP platelet responses and strengthens the confidence of their association with each ADP platelet response phenotype.

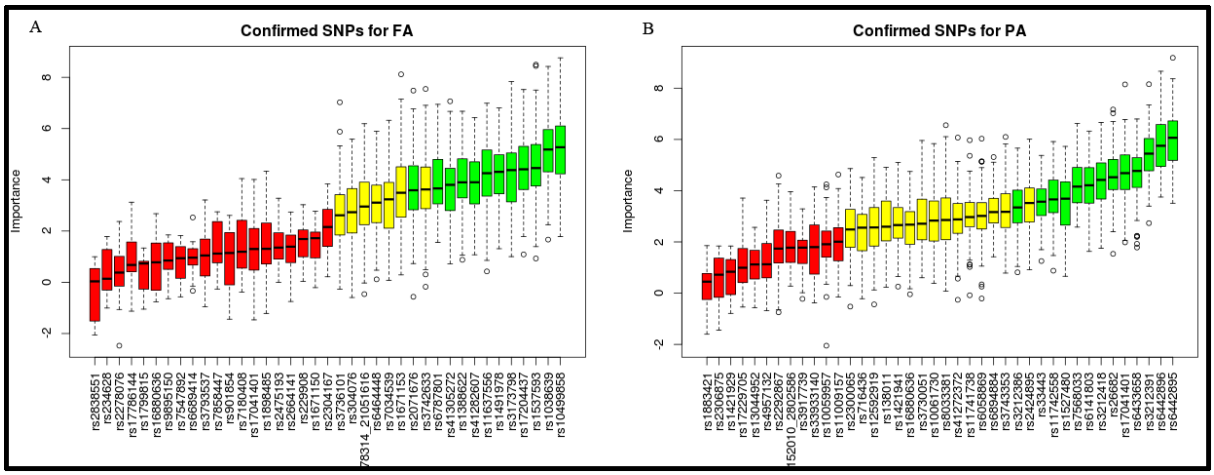


Figure 2.9 The Boruta method plot shows SNPs that are associated with (A) FA, and (B) PA platelet responses. The green, yellow and red boxplots are the confirmed important, tentative, and rejected SNPs respectively. The confirmed important SNPs are the significant SNPs associated with platelet responses. The selected significant SNPs here add more weight to the already identified SNPs from other methods, which may improve the consensus identification of the key SNPs. Also, it highlights other significant SNPs that might potentially be missed by other methods in the pipeline.

Moreover, using Boruta as an additional layer in the pipeline enhances further the discovery of significantly associated SNPs that may be missed by other methods in the pipeline.

Furthermore, the results of the RAPIDSNPs analyses for the dataset 2 are shown in Table 2.11 and 2.12, which include the significant and key SNPs that are associated with PA and FA respectively.

			Stepwise (Jones et al 2009)	RF with Stepwise	RF with Ridge regression	RF with LASSO	RF with Boruta (P=0.01)	Consensus (3/4)
Platelet response type		Beta (+ve/-ve)	PA	PA	PA	PA	PA	PA
SNPs ID	Gene/Location							
rs17229705	*VAV3*	+ve	✓(0.0009)	✓(0.0021)	✓(0.01003)	✓(0.00312)	✓	✓
rs246406	*ITGA2*	-ve	✓(0.002)	✓(0.0045)	✓(0.0117)	✓(0.004)	✓	✓
rs11631474	*MAP2K5*	+ve	✓(0.007)	×	✓(0.01003)	✓(0.005)	×	✓
rs6057638	*ch20:32751526*	-ve	×	×	×	✓(0.012)	✓	
rs1472122	*P2Y12/MED12L*	-ve	✓(0.0007)	✓(0.00034)	✓(0.0153)	✓(0.0009)	✓	✓
rs950365	*ch21:18840644*	+ve	×	×	×	✓(0.006)	✓	
rs2815805	*MAPK14*	+ve	✓(0.007)	✓(0.0032)	✓(0.008)	✓(0.004)	✓	✓
rs5277	*PTGS2*	+ve	✓(0.006)	✓(0.0011)	✓(0.00367)	✓(0.002)	×	✓
rs41307147	*CD109*	-ve	✓(0.013)	×	✓(0.0092)	✓(0.0105)	×	✓
rs2228671	*LDLR*	+ve	✓(0.017)	✓(0.01)	✓(0.0026)	✓(0.0065)	×	✓
rs17041401	*ITPR1*	+ve	✓(0.003)	×	×	×	✓	
rs2769668	*VAV3*	+ve	✓(0.005)	×	×	×	✓	
rs6141803	*ch20:32752550*	-ve	×	×	×	×	✓	
rs12410842	*VAV3*	+ve	✓(0.005)	×	×	×	✓	
rs2825207	*ch21:18855073*	+ve	×	×	×	×	✓	
rs3788337	*GNAZ*	-ve	✓(0.0009)	×	×	×	✓	

rs9612234	*GNAZ*	+ve	×(0.024)	×	×	×	✓	
rs3745406	*PRKCG*	-ve	×(0.085)	×	✓(0.014)	×	×	
rs2633717	*ITPR1*	+ve	×	✓(0.008)	×	×	×	

Table 2.11 Consensus identification of the most significant SNPs associated with PA platelet response using the imputed dataset 2. The consensus SNP is selected if it has been identified by at least three methods, which means it has higher significance and hence is more likely to be a key genetic variant. × indicates either the SNP was not identified by the method or previously identified as insignificant. ✓ indicates the SNP was identified by the method. Numbers inside the brackets after ✓ or × indicate p values of the SNPs calculated using Wald test, or partial F-test. The string characters started with 'ch' represent the chromosomal location of the SNP in the genome. +ve/-ve indicate the sign of coefficient estimates of the SNPs from the regression models.

			Stepwise (Jones et al 2009)	RF with Stepwise	RF with Ridge regression	RF with LASSO	RF with Boruta (P=0.01)	Consensus (3/4)
Platelet response type		Beta (+ve/-ve)	FA	FA	FA	FA	FA	FA
SNPs ID	Gene/Location							
rs12485738	*ARHGEF3*	+ve	×	×	✓(0.0104)	✓(0.0085)	✓	✓
rs17786144	*ITPR1*	+ve	✓(0.003)	✓(0.0024)	✓(0.00663)	✓(0.003)	×	✓
rs5746223	*RAF1*	+ve	✓(0.0005)	×	×	✓(0.00054)	×	
rs41307142	*GAS6*	-ve	×	×	✓(0.0035)	✓(0.0062)	×	
rs6450105	*ITGA2*	+ve	✓(0.001)	×	×	✓(0.007)	✓	✓
rs822442	*PEAR1*	+ve	✓(0.017)	✓(8.73e-05)	✓(0.00066)	✓(5.52e-05)	×	✓
rs11637556	*MAP2K1*	+ve	✓(0.005)	✓(3.96e-05)	✓(0.00047)	✓(0.00046)	✓	✓
rs11264579	*PEAR1*	-ve	✓(0.004)	✓(8.17e-06)	✓(0.00016)	✓(6.46e-06)	✓	✓
rs10429491	*JAK2*	-ve	✓(0.0006)	✓(0.00039)	✓(0.00022)	✓(9.63e-05)	✓	✓
rs7180408	*GTF2A2*	+ve	×	×	✓(0.0089)	×	✓	
rs1552031	*BNIP2*	-ve	×	×	×	×	✓	

rs12910751	*ch15:59630966*	+ve	×	×	×	×	✓	
rs17041401	*ITPR1*	+ve	× (0.023)	×	×	×	✓	
rs2838551	*ch21:44344514*	+ve	×	×	×	×	✓	
rs41305896	*ITGA2*	+ve	✓(0.001)	×	×	×	✓	
rs1979422	*ch15:59636012*	-ve	×	×	×	×	✓	
rs7858447	*TLN1*	+ve	×(0.329)	×	×	×	✓	
rs7739455	*CD109*	-ve	×(0.063)	×	✓(0.0049)	×	✓	
rs350916	*MAP2K2*	+ve	✓(0.001)	✓(0.00197)	×	×	×	
rs3729931	*RAF1*	+ve	✓(0.0001)	✓(0.00062)	×	×	×	

Table 2.12 Consensus identification of the most significant (key) SNPs associated with FA platelet response using entropy based SNPs' genotypes' imputed dataset 2. The string characters started with 'ch' represent the chromosomal location of the SNP in the genome. The consensus SNP is selected if it has been identified by at least three methods, which means it has higher significance and hence is more likely to be a key genetic variant. × indicates either the SNP was not identified by the method or previously identified as insignificant. ✓ indicates the SNP was identified by the method. Numbers inside the brackets after ✓ or × indicate p values of the SNPs calculated using Wald test, or partial F-test. The string characters started with 'ch' represent the chromosomal location of the SNP in the genome. +ve/-ve indicate the sign of coefficient estimates of the SNPs from the regression models.

Figure 2.10 A - B shows the key SNPs that have been selected by the multiple regression methods using dataset 2

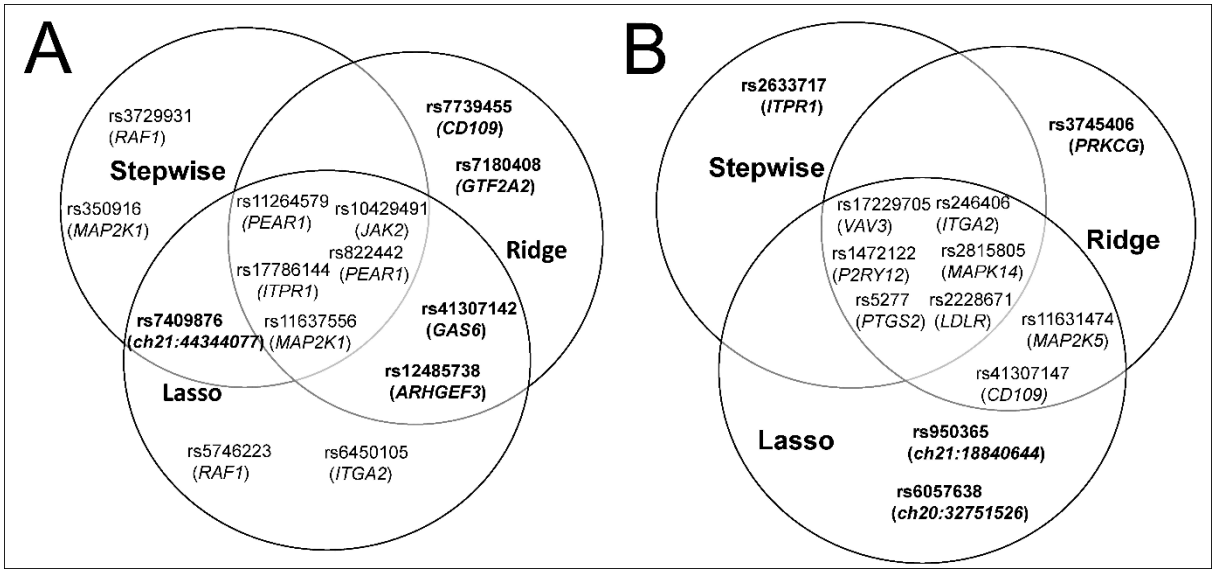


Figure 2.10 Venn diagrams showing the identified significant and key SNPs from the regression layer in the pipeline using dataset 2. The identified significant SNPs in (A) and (B) are associated with FA (fibrinogen binding in response to adenosine diphosphate) and PA (p-selectin in response to adenosine diphosphate) respectively. The newly detected SNPs or those reported as insignificant in the previous study are shown in bold.

Furthermore, the Boruta method layer identified several significant SNPs, which are associated with both ADP platelet responses using dataset 2 and closely similar to those identified by the regression methods layer. These can be visualised by using the Boruta plots (Figure 2.11 A - B).

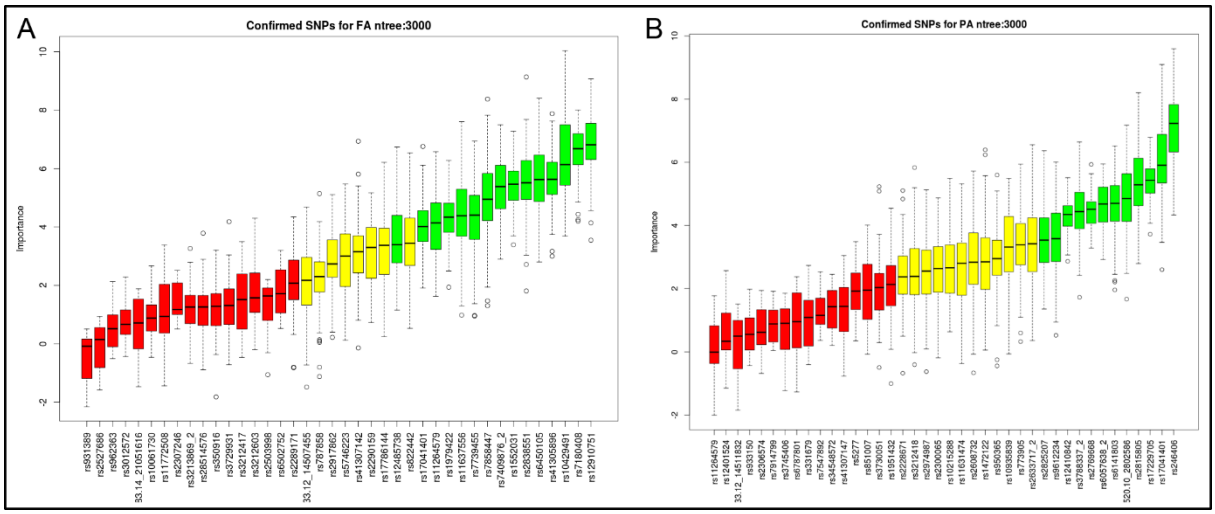


Figure 2.11 The Boruta method plot showing SNPs that are associated with ADP platelet responses using dataset 2. These SNPs are associated with (A) FA (fibrinogen binding in response to adenosine diphosphate), and (B) PA (p-selectin in response to adenosine diphosphate) platelet responses. The green, yellow and red boxplots are the confirmed important, tentative, and rejected SNPs respectively. The confirmed important SNPs are the significant SNPs associated with platelet responses.

Additionally, using the consensus identification approach, the following key significant SNPs were obtained when the pipeline was applied to the *dataset 3*, and are shown in the Tables 2.13 and 2.14. These SNPs are associated with PA and FA respectively.

			Stepwise (Jones et al 2009)	RF with Stepwise	RF with Ridge regression	RF with LASSO	RF with Boruta (P=0.01)	Consensus (3/4)
Platelet response type		Beta (+ve/-ve)	PA	PA	PA	PA	PA	PA
SNPs ID	Gene/Location							
rs906766	*MED12L*	+ve	×	✓(0.004)	×	✓(0.012)	✓	✓
rs246406	*ITGA2*	-ve	✓(0.002)	✓ (0.002)	✓ (0.001)	✓(0.004)	×	✓
rs2633717	*ITPR1*	+ve	×	×	✓(0.004)	×	✓	
rs17786144	*ITPR1*	+ve	✓(0.003)	✓(0.0011)	✓ (0.005)	✓(0.002)	✓	
rs2276829	*VIPR1*	+ve	×	✓ (0.0067)	×	×	✓	
rs12953	*PECAM1*	-ve	× (0.421)	✓(0.0018)	✓(0.004)	✓(0.007)	✓	✓
rs17729525	*ITPR1*	+ve	×(0.024)	×	×	×	✓	
rs41305276	*THBS1*	+ve	×(0.202)	×	✓(0.001)	✓(0.004)	✓	✓
rs12973968	*GP6*	+ve	×(0.111)	×	×	×	✓	
rs9612234	*GNAZ*	+ve	×(0.024)	×	×	×	✓	
rs7853785	*TLN1*	+ve	×(0.611)	×	×	×	✓	
rs2289171	*PIP5K3*	+ve	×(0.062)	×	✓ (0.0083)	✓ (00462)	×	

Table 2.13 Consensus identification of the most significant (key) SNPs associated with PA platelet response using dataset 3. The consensus SNP is selected if it has been identified by at least three methods, which means it has higher significance and hence is a key genetic variant. × indicates either the SNP was not identified by the method or previously identified as insignificant. ✓ indicates the SNP was identified by the method. Numbers inside the brackets after ✓ or × indicate p values of the SNPs calculated using Wald test, or partial F-test. The string characters started with 'ch' represent the chromosomal location of the SNP in the genome. +ve/-ve indicate the sign of coefficient estimates of the SNPs from the regression models.

			Stepwise (Jones et al 2009)	RF with Stepwise	RF with Ridge regression	RF with LASSO	RF with Boruta (P=0.01)	Consensus (3/4)
Platelet response type		Beta (+ve/-ve)	FA	FA	FA	FA	FA	FA
SNPs ID	Gene/Location							
rs2596831	*RAF1*	+ve	×(0.146)	×	×	✓(0.005353)	✓	
rs17204376	*P2RY12*	-ve	×(0.059)	✓(0.0095)	×	✓(0.001146)	✓	✓
rs13135667	*ch4:1765340*	+ve	×	✓(0.00022)	×	✓(0.007)	×	
rs3212603	*ITGA2*	+ve	×(0.025)	✓(0.0035)	×	✓(0.0009)	×	
rs4792219	*MAP2K4*	+ve	×(0.039)	×	✓(0.0104)	✓(0.0011)	×	
rs17786144	*ITPR1*	+ve	✓(0.003)	✓(0.000302)	×	✓(0.0193)	×	✓
rs12609974	*GP6/NLRP2*	-ve	×(0.746)	×	×	✓(0.001)	×	
rs2276829	*VIPR1*	+ve	×(0.456)	×	✓(0.0120)	✓(0.0029)	✓	✓
rs906766	*MED12L*	+ve	×	✓(0.00034)	✓(7.65e-05)	✓(9.05e-05)	✓	✓
rs722432	*ITGB1*	-ve	×(0.584)	×	✓(0.0011)	✓(1.45e-05)	×	
rs2290159	*RAF1*	+ve	✓(0.001)	×	×	×	✓	
rs158687	*SYK*	+ve	×(0.668)	×	×	×	✓	
rs6502752	*P2RX1*	-ve	×(0.079)	×	×	×	✓	
rs6086714	*PLCB1*	-ve	×(0.201)	×	×	×	✓	
rs158688	*SYK*	-ve	×(0.785)	×	×	×	✓	
rs2290149	*MADD*	-ve	×(0.113)	×	×	×	✓	
rs2206266	*PLCB1*	-ve	×(0.252)	×	×	×	✓	
rs17296289	*ITGB1*	-ve	×(0.448)	✓(0.006784)	✓(0.0011)	×	×	
rs2306875	*ITPR1*	+ve	×(0.075)	×	✓(0.003)	×	×	
rs3739038	*LRRFIP1*	+ve	×	✓(0.00352)	×	×	×	

Table 2.14 Consensus identification of the most significant (key) SNPs associated with FA platelet response using dataset 3. The consensus SNP is selected if it has been identified by at least three methods, which means it has higher significance and hence is a key genetic variant. × indicates either the SNP was not identified by the method or previously identified as insignificant. ✓ indicates the SNP was identified by the method. Numbers inside the brackets after ✓ or × indicate p values of the SNPs calculated using Wald test, or partial F-test. The string characters started with 'ch' represent the chromosomal location of the SNP in the genome. The string characters started with 'ch' represent the chromosomal location of the SNP in the genome. +ve/-ve indicate the sign of coefficient estimates of the SNPs from the regression models.

The key SNPs observed and selected by the multiple regression methods from the dataset 3 might be similarly visualised by Venn diagram in Figure 2.12 A-B.

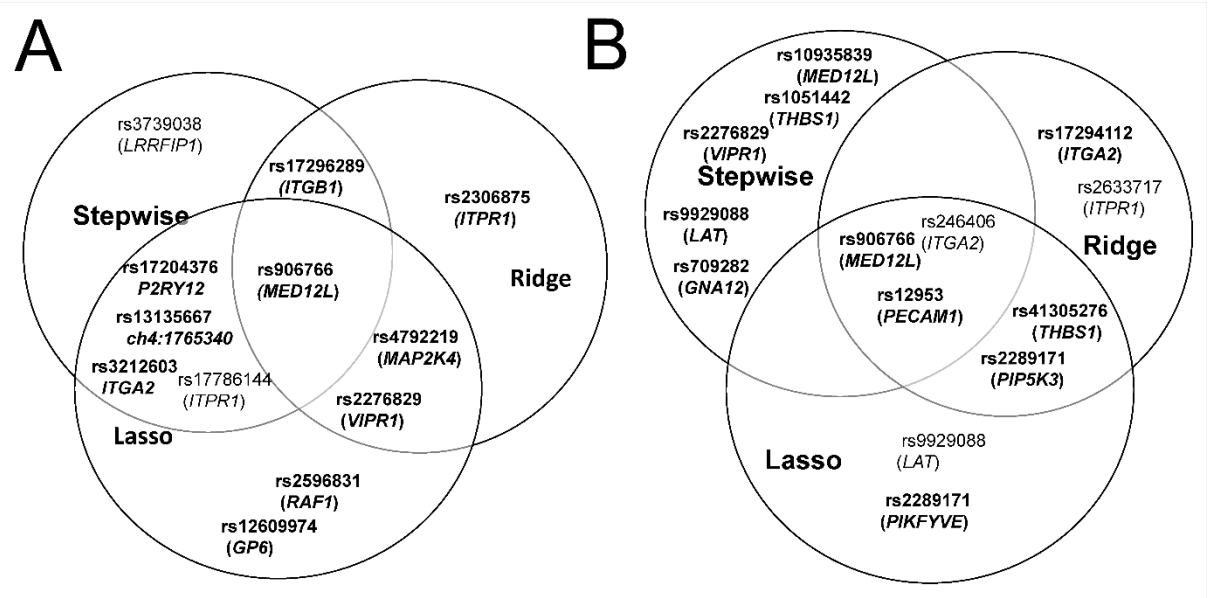


Figure 2.12 Venn diagram for identifying significant and key SNPs associated with the ADP platelet responses, which were identified by the regression layer in the pipeline using dataset 3. The newly detected SNPs or those reported as insignificant in the previous study are shown in bold. In this dataset the key significant SNPs is rs906766 in *MED12L*, which is significantly associated with A) FA using all three multiple regression methods and also selected by the two regression methods in B) PA platelet response.

Similarly, the Boruta method layer identified several significant SNPs, which are associated with both ADP platelet responses using dataset 3 and also closely similar to those identified by the regression methods layer. These can be visualised by using the Boruta plots (Figure 2.13 A – B).

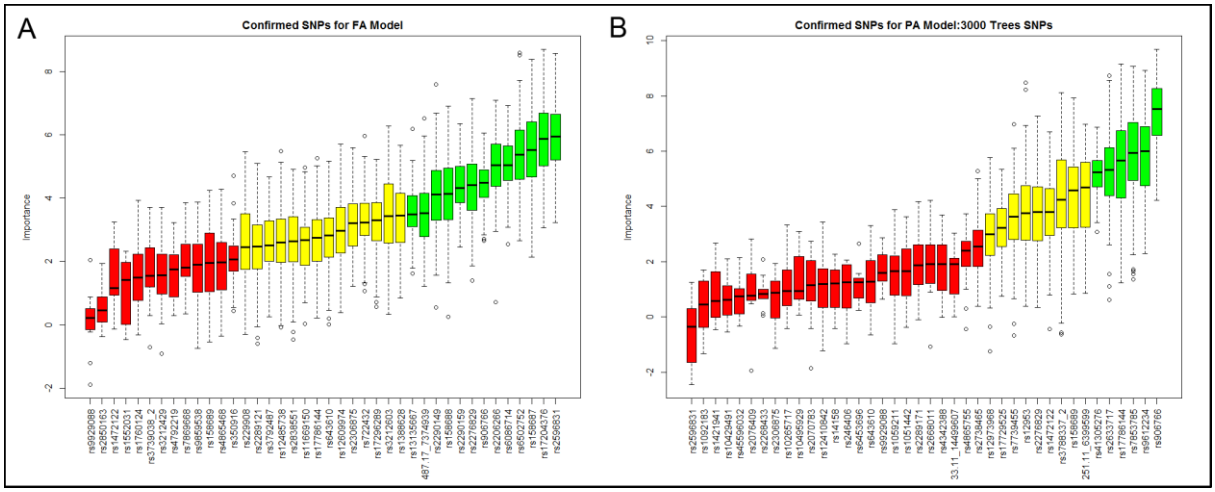


Figure 2.13 The Boruta method plot showing SNPs that are associated with ADP platelet responses using dataset 3. These are SNPs associated with (A) FA (fibrinogen binding in response to adenosine diphosphate), and (B) PA (p-selectin in response to adenosine diphosphate) platelet responses. The green, yellow and red boxplots are the confirmed important, tentative, and rejected SNPs respectively. The confirmed important SNPs are the significant SNPs associated with platelet responses.

For verifying the selected significant key SNPs in the final iteration for each of the three datasets whether are true positives, the confidence level equation (equation 2) is applied. The identified significant SNPs in all iterations are initially visualised using the frequency plot. For instance, to assess the confidence of PA associated significant key SNPs for datasets (or subsets) 1, 2, and 3, the plots showing the frequency of all significant SNPs in all iterations for each of the above datasets are initially created (Figures 2.14, 2.15, and 2.16 respectively). Tables 2.15, 2.16, and 2.17 show the frequencies of each selected SNP in each iteration for each of the datasets 1, 2, and 3 respectively associated with PA platelet response.

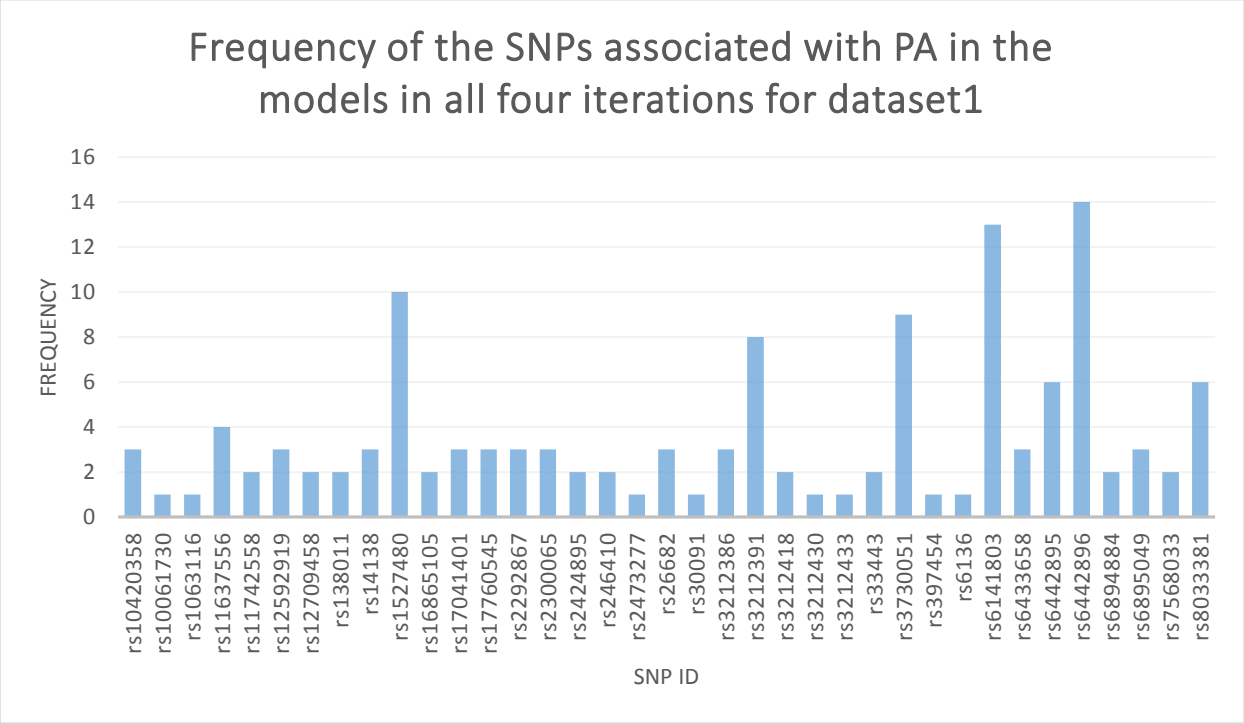


Figure 2.14 The frequency of the selected significant SNPs, which are associated with PA platelet responses in all iterations within the intermediate models for the dataset 1. It can be seen clearly that some SNPs have relatively low or high frequencies, which mean they are more likely to be false or true positive key significant SNPs respectively. The maximum frequency is 16, which means the SNP appears in the four models in each of the four iterations

SNP's frequency of appearance in the models	SNP's Id	Iteration number	RF + Model Name
1	rs3212391	1	Stepwise
2	rs6141803	1	Stepwise
3	rs2292867	1	Stepwise
4	rs6442895	1	Stepwise
5	rs12592919	1	Stepwise
6	rs2292867	1	Ridge
7	rs6442895	1	Ridge
8	rs12592919	1	Ridge
9	rs6442896	1	Ridge
10	rs3212386	1	Lasso
11	rs6141803	1	Lasso
12	rs2292867	1	Lasso
13	rs12592919	1	Lasso
14	rs6442896	1	Lasso
15	rs30091	1	Lasso
16	rs3212391	1	Boruta
17	rs6141803	1	Boruta
18	rs26682	1	Boruta
19	rs246410	1	Boruta
20	rs6442895	1	Boruta

21	rs138011	1	Boruta
22	rs6894884	1	Boruta
23	rs6442896	1	Boruta
24	rs3212430	1	Boruta
25	rs3212418	1	Boruta
26	rs6895049	1	Boruta
27	rs3212433	1	Boruta
28	rs3212391	2	Stepwise
29	rs3730051	2	Stepwise
30	rs6442896	2	Stepwise
31	rs6141803	2	Stepwise
32	rs1527480	2	Stepwise
33	rs2300065	2	Stepwise
34	rs6442896	2	Ridge
35	rs3730051	2	Ridge
36	rs2300065	2	Ridge
37	rs1527480	2	Ridge
38	rs6141803	2	Lasso
39	rs3730051	2	Lasso
40	rs2300065	2	Lasso
41	rs1527480	2	Lasso
42	rs6136	2	Lasso
43	rs6895049	2	Lasso
44	rs6141803	2	Boruta
45	rs6442895	2	Boruta
46	rs6442896	2	Boruta
47	rs6433658	2	Boruta
48	rs17041401	2	Boruta
49	rs138011	2	Boruta
50	rs3212391	2	Boruta
51	rs17760545	2	Boruta
52	rs3212386	2	Boruta
53	rs2424905	2	Boruta
54	rs2424895	2	Boruta
55	rs1527480	2	Boruta
56	rs3212391	3	Stepwise
57	rs3730051	3	Stepwise
58	rs6442896	3	Stepwise
59	rs1527480	3	Stepwise
60	rs8033381	3	Stepwise
61	rs11637556	3	Stepwise
62	rs6141803	3	Ridge
63	rs3730051	3	Ridge
64	rs6442896	3	Ridge
65	rs1527480	3	Ridge

66	rs11637556	3	Ridge
67	rs8033381	3	Ridge
68	rs10420358	3	Ridge
69	rs6141803	3	Lasso
70	rs3730051	3	Lasso
71	rs6442896	3	Lasso
72	rs1527480	3	Lasso
73	rs8033381	3	Lasso
74	rs11742558	3	Lasso
75	rs2473277	3	Lasso
76	rs1063116	3	Lasso
77	rs11637556	3	Lasso
78	rs6141803	3	Boruta
79	rs6442895	3	Boruta
80	rs6442896	3	Boruta
81	rs6433658	3	Boruta
82	rs17041401	3	Boruta
83	rs3212391	3	Boruta
84	rs17760545	3	Boruta
85	rs11742558	3	Boruta
86	rs7568033	3	Boruta
87	rs397454	3	Boruta
88	rs26682	3	Boruta
89	rs33443	3	Boruta
90	rs6894884	3	Boruta
91	rs16865105	3	Boruta
92	rs6442896	4	Stepwise
94	rs3212391	4	Stepwise
95	rs3730051	4	Stepwise
96	rs6141803	4	Stepwise
97	rs1527480	4	Stepwise
98	rs8033381	4	Stepwise
99	rs14138	4	Stepwise
100	rs6141803	4	Ridge
101	rs6442896	4	Ridge
102	rs3730051	4	Ridge
103	rs10420358	4	Ridge
104	rs14138	4	Ridge
105	rs1527480	4	Ridge
106	rs12709458	4	Ridge
107	rs8033381	4	Ridge
108	rs6141803	4	Lasso
109	rs6442896	4	Lasso
110	rs3730051	4	Lasso
111	rs1527480	4	Lasso

112	rs14138	4	Lasso
113	rs10061730	4	Lasso
114	rs12709458	4	Lasso
115	rs11637556	4	Lasso
116	rs8033381	4	Lasso
117	rs10420358	4	Lasso
118	rs6141803	4	Boruta
119	rs6442896	4	Boruta
120	rs6433658	4	Boruta
121	rs6442895	4	Boruta
122	rs3212391	4	Boruta
123	rs17041401	4	Boruta
124	rs3212386	4	Boruta
125	rs33443	4	Boruta
126	rs6895049	4	Boruta
127	rs2424895	4	Boruta
128	rs26682	4	Boruta
129	rs3212418	4	Boruta
130	rs2424905	4	Boruta
131	rs7568033	4	Boruta
132	rs17760545	4	Boruta
133	rs246410	4	Boruta
134	rs16865105	4	Boruta

Table 2.15 The frequency table for SNPs, which are associated with PA response for the dataset 1 and that wereselected in all iterations of the pipeline.

For instance, from the data in Figure 2.14 and Table 2.14, the SNP rs6141803 has appeared in the intermediate models 13 times in all iterations. The total number of models (methods) within the pipeline is 4. The total number of iterations are 4, i.e. four different RF run *ntree* sizes (*ntree* = 500, *ntree* = 1000, *ntree* = 2000, and *ntree* = 3000), thus, the confidence level of SNP would be 13/4*4 = 0.8125. This confidence score exceeds 0.5, and therefore, the selected SNP is more likely to be a true positive.

Applying equation 2 to the data in Figure 2.14 and Table 2.14, 7 key SNPs (rs1527480, rs3212391, rs3730051, rs6141803, rs6442896, rs6442895, and rs8033381) have been identified with high confidence to be significantly associated with PA platelet response.

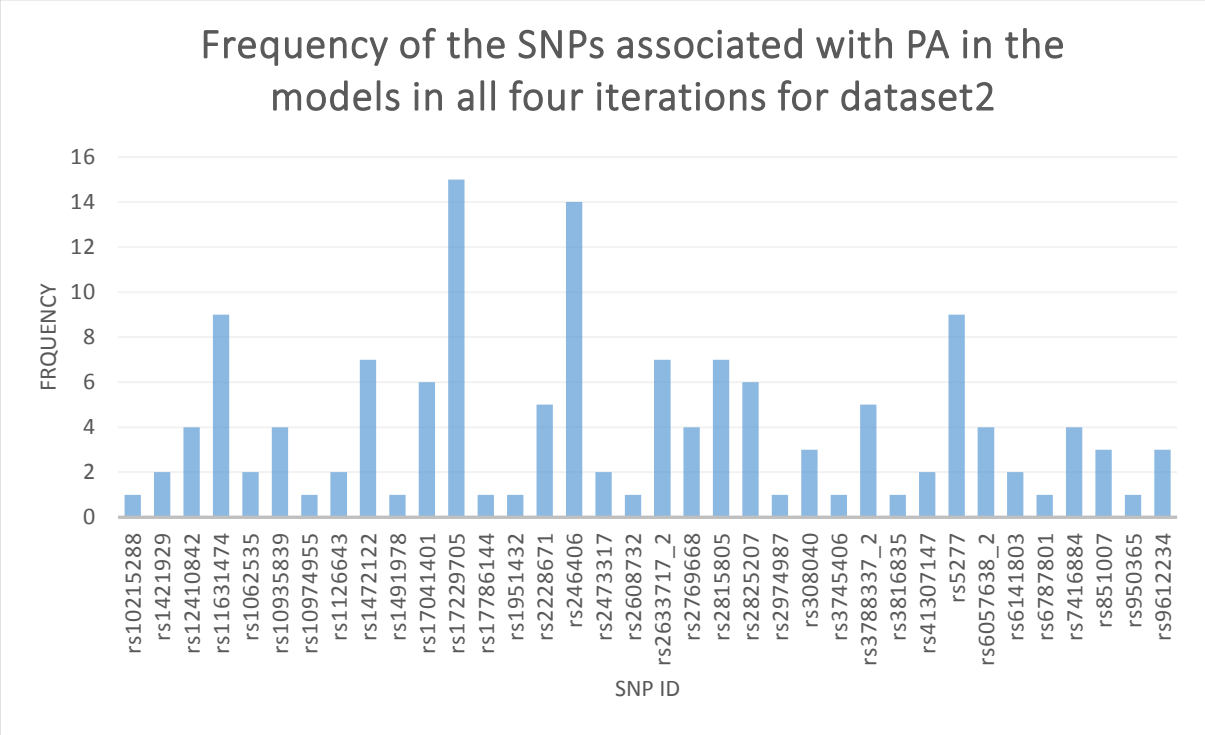


Figure 2.15 The frequency of the selected significant SNPs, which are associated with PA platelet responses in all iterations within the intermediate models for the dataset 2.

SNP's frequency of appearance in the models	SNP's Id	Iteration number	RF + Model Name
1	rs17229705	1	Stepwise
2	rs17041401	1	Stepwise
3	rs246406	1	Stepwise
4	rs11631474	1	Stepwise
5	rs3788337_2	1	Stepwise
6	rs2825207	1	Stepwise
7	rs851007	1	Stepwise
8	rs17229705	1	Ridge
9	rs10935839	1	Ridge
10	rs11631474	1	Ridge
11	rs3788337_2	1	Ridge
12	rs2473317	1	Ridge
13	rs2825207	1	Ridge
14	rs851007	1	Ridge
15	rs17229705	1	Lasso
16	rs17041401	1	Lasso
17	rs246406	1	Lasso
18	rs1491978	1	Lasso
19	rs10935839	1	Lasso
20	rs11631474	1	Lasso
21	rs3788337_2	1	Lasso

22	rs851007	1	Lasso
23	rs2473317	1	Lasso
24	rs17229705	1	Boruta
25	rs17041401	1	Boruta
26	rs12410842	1	Boruta
27	rs246406	1	Boruta
28	rs2769668	1	Boruta
29	rs10935839	1	Boruta
30	rs11631474	1	Boruta
31	rs3788337_2	1	Boruta
32	rs1126643	1	Boruta
33	rs1062535	1	Boruta
34	rs1421929	1	Boruta
35	rs9612234	1	Boruta
36	rs1472122	2	Stepwise
37	rs17229705	2	Stepwise
38	rs246406	2	Stepwise
39	rs5277	2	Stepwise
40	rs2633717_2	2	Stepwise
41	rs7416884	2	Stepwise
42	rs246406	2	Ridge
43	rs11631474	2	Ridge
44	rs308040	2	Ridge
45	rs17229705	2	Ridge
46	rs5277	2	Ridge
47	rs2633717_2	2	Ridge
48	rs7416884	2	Ridge
49	rs2228671	2	Ridge
50	rs1951432	2	Ridge
51	rs2825207	2	Lasso
52	rs11631474	2	Lasso
53	rs1472122	2	Lasso
54	rs17229705	2	Lasso
55	rs5277	2	Lasso
56	rs2633717_2	2	Lasso
57	rs7416884	2	Lasso
58	rs2228671	2	Lasso
59	rs17041401	2	Boruta
60	rs246406	2	Boruta
61	rs10935839	2	Boruta
62	rs2769668	2	Boruta
63	rs17229705	2	Boruta
64	rs2608732	2	Boruta
65	rs2633717_2	2	Boruta
66	rs9612234	2	Boruta

67	rs12410842	2	Boruta
68	rs246406	3	Stepwise
69	rs17229705	3	Stepwise
70	rs10974955	3	Stepwise
71	rs1472122	3	Stepwise
72	rs2815805	3	Stepwise
73	rs5277	3	Stepwise
74	rs2633717_2	3	Stepwise
75	rs246406	3	Ridge
76	rs17229705	3	Ridge
77	rs2815805	3	Ridge
78	rs5277	3	Ridge
79	rs3816835	3	Ridge
80	rs10215288	3	Ridge
81	rs308040	3	Ridge
82	rs246406	3	Lasso
83	rs17229705	3	Lasso
84	rs11631474	3	Lasso
85	rs6057638_2	3	Lasso
86	rs1472122	3	Lasso
87	rs2815805	3	Lasso
88	rs2825207	3	Lasso
89	rs5277	3	Lasso
90	rs17786144	3	Lasso
91	rs7416884	3	Lasso
92	rs308040	3	Lasso
93	rs246406	3	Boruta
94	rs17041401	3	Boruta
95	rs17229705	3	Boruta
96	rs2769668	3	Boruta
97	rs6057638_2	3	Boruta
98	rs2815805	3	Boruta
99	rs2825207	3	Boruta
100	rs1126643	3	Boruta
101	rs12410842	3	Boruta
102	rs1062535	3	Boruta
103	rs2974987	3	Boruta
104	rs1421929	3	Boruta
105	rs6141803	3	Boruta
106	rs246406	4	Stepwise
107	rs17229705	4	Stepwise
108	rs1472122	4	Stepwise
109	rs2815805	4	Stepwise
110	rs5277	4	Stepwise
111	rs2633717_2	4	Stepwise

112	rs2228671	4	Stepwise
113	rs246406	4	Ridge
114	rs17229705	4	Ridge
115	rs11631474	4	Ridge
116	rs1472122	4	Ridge
117	rs5277	4	Ridge
118	rs3745406	4	Ridge
119	rs41307147	4	Ridge
120	rs2228671	4	Ridge
121	rs246406	4	Lasso
122	rs17229705	4	Lasso
123	rs11631474	4	Lasso
124	rs6057638_2	4	Lasso
125	rs1472122	4	Lasso
126	rs950365	4	Lasso
127	rs2815805	4	Lasso
128	rs5277	4	Lasso
129	rs2633717_2	4	Lasso
130	rs41307147	4	Lasso
131	rs2228671	4	Lasso
132	rs17041401	4	Boruta
133	rs246406	4	Boruta
134	rs17229705	4	Boruta
135	rs6787801	4	Boruta
136	rs2769668	4	Boruta
137	rs6141803	4	Boruta
138	rs12410842	4	Boruta
139	rs6057638_2	4	Boruta
140	rs2825207	4	Boruta
141	rs3788337_2	4	Boruta
142	rs2815805	4	Boruta
143	rs9612234	4	Boruta

Table 2.16. The frequency table showing the selected significant SNPs in the intermediate models in each iteration of the pipeline. These SNPs are associated with PA platelet response from the dataset 2.

Furthermore, in applying equation (2) on the data in Figure 2.15 and Table 2.15, nearly 7 key SNPs (rs11631474, rs1472122, rs17229705, rs2633717, rs2815805, rs246406, and rs5277) were identified with high confidence to be significantly associated with PA platelet response.

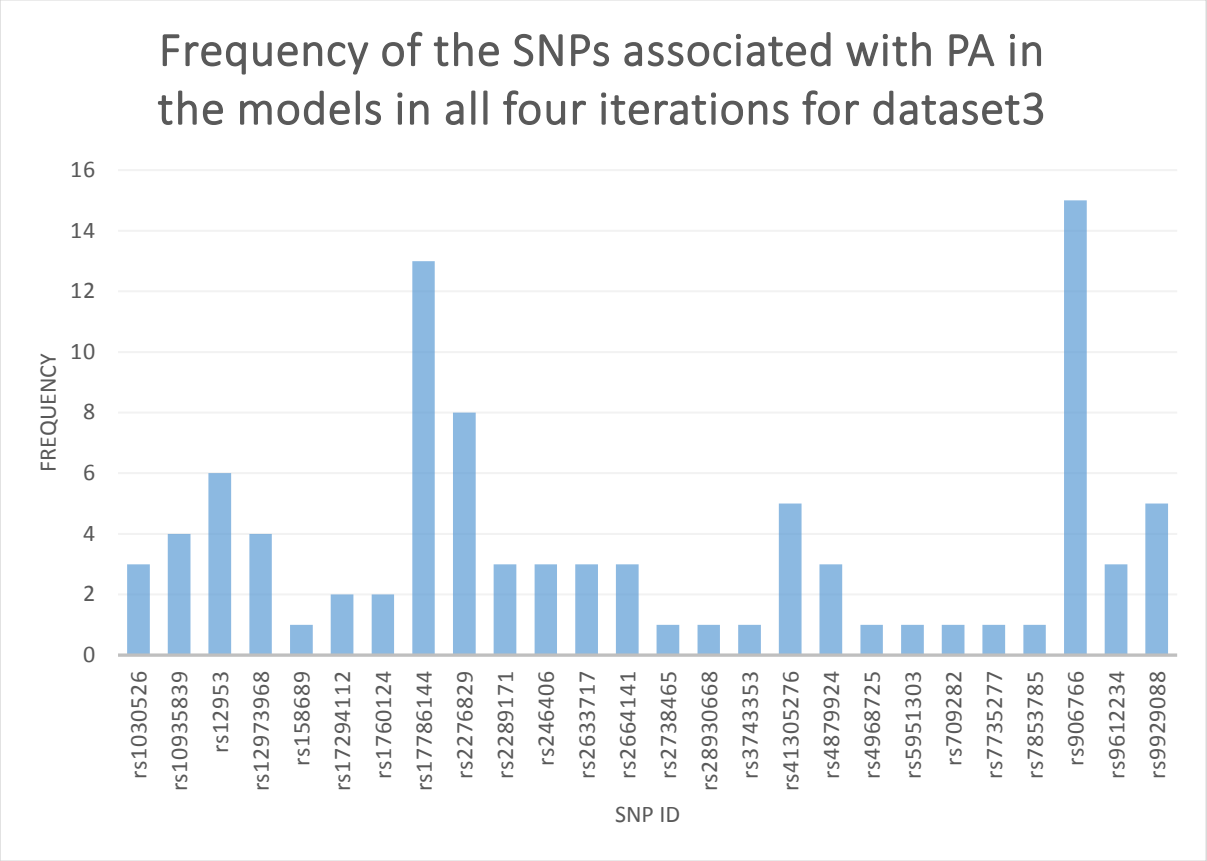


Figure 2.16 The frequency of the selected significant SNPs, which are associated with PA platelet responses in all iterations within the intermediate models for the dataset 3.

SNP's frequency of appearance in the models	SNP's Id	Iteration number	RF + Model Name
1	rs17786144	1	Stepwise
2	rs246406	1	Stepwise
3	rs2289171	1	Stepwise
4	rs12953	1	Stepwise
5	rs2664141	1	Stepwise
6	rs906766	1	Stepwise
7	rs246406	1	Ridge
8	rs2289171	1	Ridge
9	rs12953	1	Ridge
10	rs2664141	1	Ridge
11	rs906766	1	Ridge
12	rs1030526	1	Ridge
13	rs12973968	1	Ridge
14	rs906766	1	Lasso
15	rs1030526	1	Lasso
16	rs28930668	1	Lasso
17	rs17786144	1	Lasso
18	rs246406	1	Lasso

19	rs12953	1	Lasso
20	rs2664141	1	Lasso
21	rs17786144	1	Boruta
22	rs2738465	1	Boruta
23	rs2276829	1	Boruta
24	rs10935839	1	Boruta
25	rs9612234	1	Boruta
26	rs906766	1	Boruta
27	rs5951303	1	Boruta
28	rs1030526	1	Boruta
29	rs2633717	1	Boruta
30	rs12973968	1	Boruta
31	rs7735277	1	Boruta
32	rs906766	2	Stepwise
33	rs17786144	2	Stepwise
34	rs2276829	2	Stepwise
35	rs12953	2	Stepwise
36	rs9929088	2	Stepwise
37	rs17786144	2	Ridge
38	rs906766	2	Ridge
39	rs2276829	2	Ridge
40	rs4968725	2	Ridge
41	rs12953	2	Ridge
42	rs3743353	2	Ridge
43	rs4879924	2	Ridge
44	rs1760124	2	Ridge
45	rs9929088	2	Ridge
46	rs41305276	2	Ridge
47	rs709282	2	Ridge
48	rs906766	2	Lasso
49	rs17786144	2	Lasso
50	rs2276829	2	Lasso
51	rs12953	2	Lasso
52	rs4879924	2	Lasso
53	rs1760124	2	Lasso
54	rs9929088	2	Lasso
55	rs17786144	2	Boruta
56	rs906766	2	Boruta
57	rs2276829	2	Boruta
58	rs2633717	2	Boruta
59	rs9612234	2	Boruta
60	rs12973968	2	Boruta
61	rs41305276	2	Boruta
62	rs906766	3	Stepwise
63	rs17786144	3	Stepwise

64	rs10935839	3	Stepwise
65	rs906766	3	Ridge
66	rs17294112	3	Ridge
67	rs41305276	3	Ridge
68	rs906766	3	Lasso
69	rs17786144	3	Lasso
70	rs2276829	3	Lasso
71	rs17786144	3	Boruta
72	rs2633717	3	Boruta
73	rs906766	3	Boruta
74	rs9612234	3	Boruta
75	rs2276829	3	Boruta
76	rs10935839	3	Boruta
77	rs12973968	3	Boruta
78	rs158689	3	Boruta
79	rs7853785	3	Boruta
80	rs41305276	3	Boruta
81	rs906766	4	Stepwise
82	rs17786144	4	Stepwise
83	rs10935839	4	Stepwise
84	rs2276829	4	Stepwise
85	rs4879924	4	Stepwise
86	rs9929088	4	Stepwise
87	rs17786144	4	Ridge
88	rs906766	4	Ridge
89	rs17294112	4	Ridge
90	rs906766	4	Lasso
91	rs17786144	4	Lasso
92	rs41305276	4	Lasso
93	rs9929088	4	Lasso
94	rs2289171	4	Lasso
95	rs17786144	4	Boruta
96	rs906766	4	Boruta
97	rs2633717	4	Boruta
98	rs10935839	4	Boruta
99	rs9612234	4	Boruta
100	rs2276829	4	Boruta
101	rs41305276	4	Boruta
102	rs17294112	4	Boruta
103	rs4879924	4	Boruta

Table 2.17. The frequency table showing the selected significant SNPs in the intermediate models in each iteration of the pipeline. These SNPs are associated with PA platelet response from the dataset 3.

Lastly, applying equation 2 to the data in Figure 2.16 and Table 2.17, 4 SNPs (rs12953, rs17786144, rs2276829, and rs906766) have high confidence score and hence, are true (key) positive and significantly associated with PA platelet response.

Furthermore, for the FA platelet response, to assess the likely true positives key SNPs, the frequency plots showing the frequencies of appearance for each significant SNPs in each iteration are initially created. Figures 2.17 – 2.19 show these plots. Additionally, Tables 2.18 – 2.20 show the frequencies of each SNP in each iteration in the intermediate models. The Tables are used to generate the plots.

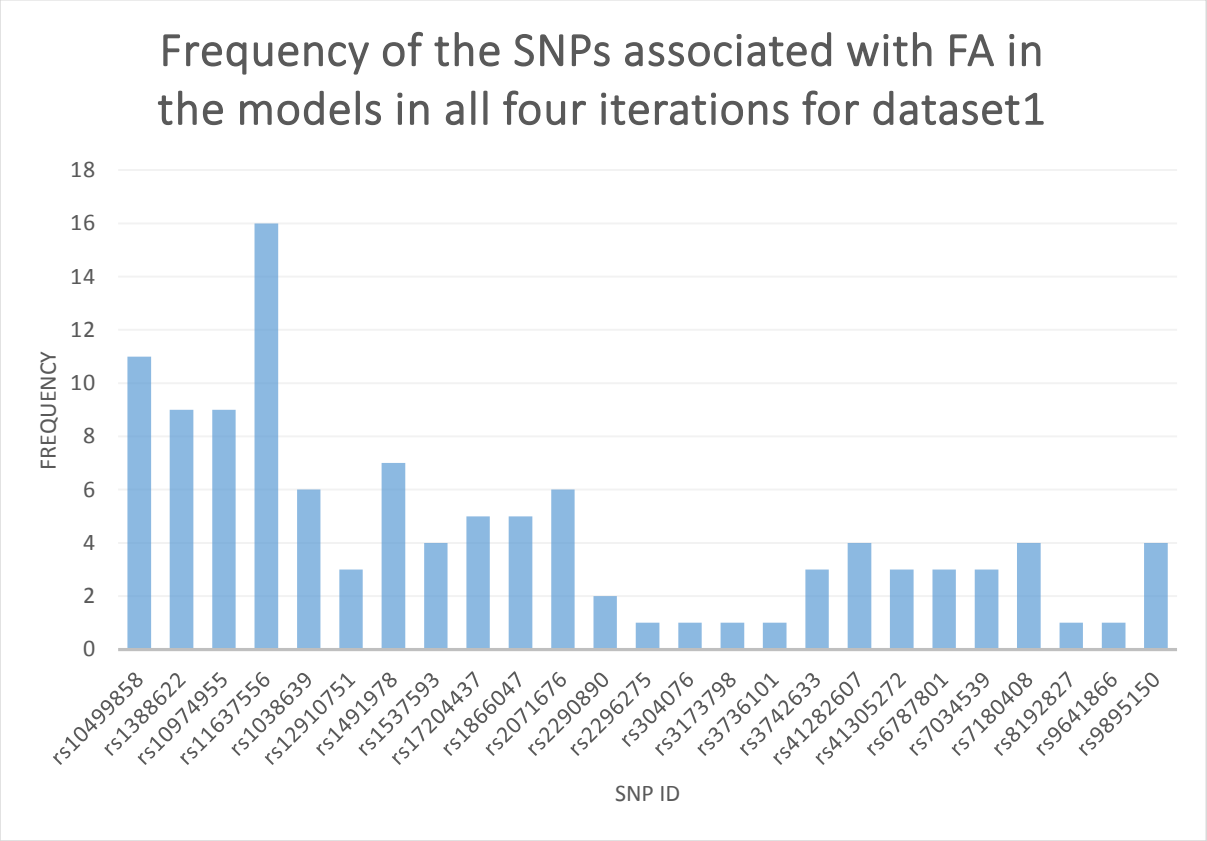


Figure 2.17 The frequency of the selected significant SNPs, which are associated with FA platelet responses in all iterations within the intermediate models for the dataset 1. Few SNPs are shown to be highly significant. For instance, rs11637556 in *MAPK1* has been selected in each iteration

SNP's frequency of appearance in the models	SNP's Id	Iteration number	RF + Model Name
1	rs1388622	1	Stepwise
2	rs11637556	1	Stepwise
3	rs10499858	1	Stepwise
4	rs10974955	1	Stepwise
5	rs1866047	1	Stepwise
6	rs1491978	1	Stepwise
7	rs11637556	1	Ridge
8	rs10974955	1	Ridge
9	rs1866047	1	Ridge
10	rs11637556	1	Lasso
11	rs10499858	1	Lasso
12	rs1388622	1	Lasso
13	rs1491978		Lasso
14	rs10974955	1	Lasso
15	rs1866047	1	Lasso
16	rs10499858	1	Boruta
17	rs1491978	1	Boruta
18	rs11637556	1	Boruta
19	rs1388622	1	Boruta
20	rs17204437	1	Boruta
21	rs1866047	1	Boruta
22	rs2296275	1	Boruta
23	rs1388622	2	Stepwise
24	rs11637556	2	Stepwise
25	rs2071676	2	Stepwise
26	rs10499858	2	Stepwise
27	rs10974955	2	Stepwise
28	rs8192827	2	Stepwise
29	rs11637556	2	Ridge
30	rs10974955	2	Ridge
31	rs9895150	2	Ridge
32	rs1388622	2	Ridge
33	rs10499858	2	Ridge
34	rs11637556	2	Lasso
35	rs1388622	2	Lasso
36	rs2071676	2	Lasso
37	rs10499858	2	Lasso
38	rs10974955	2	Lasso
39	rs9895150	2	Lasso
40	rs1388622	2	Boruta
41	rs11637556	2	Boruta
42	rs6787801	2	Boruta
43	rs17204437	2	Boruta

44	rs2071676	2	Boruta
45	rs1537593	2	Boruta
46	rs3742633	2	Boruta
47	rs1866047	2	Boruta
48	rs1038639	2	Boruta
49	rs1491978	3	Stepwise
50	rs11637556	3	Stepwise
51	rs10974955	3	Stepwise
52	rs10499858	3	Stepwise
53	rs2290890	3	Stepwise
54	rs12910751	3	Stepwise
55	rs1038639	3	Ridge
56	rs11637556	3	Ridge
57	rs10974955	3	Ridge
58	rs12910751	3	Ridge
59	rs11637556	3	Lasso
60	rs1491978	3	Lasso
61	rs10974955	3	Lasso
62	rs2290890	3	Lasso
63	rs12910751	3	Lasso
64	rs10499858	3	Lasso
65	rs1491978	3	Boruta
66	rs1038639	3	Boruta
67	rs11637556	3	Boruta
68	rs6787801	3	Boruta
69	rs1537593	3	Boruta
70	rs17204437	3	Boruta
71	rs9641866	3	Boruta
72	rs10499858	3	Boruta
73	rs1388622	4	Stepwise
74	rs11637556	4	Stepwise
75	rs10499858	4	Stepwise
76	rs2071676	4	Stepwise
77	rs7034539	4	Stepwise
78	rs3742633	4	Stepwise
79	rs41282607	4	Stepwise
80	rs11637556	4	Ridge
81	rs7034539	4	Ridge
82	rs1038639	4	Ridge
83	rs41305272	4	Ridge
84	rs41282607	4	Ridge
85	rs9895150	4	Ridge
86	rs7180408	4	Ridge
87	rs11637556	4	Lasso
88	rs3736101	4	Lasso

89	rs2071676	4	Lasso
90	rs9895150	4	Lasso
91	rs1038639	4	Lasso
92	rs41305272	4	Lasso
93	rs41282607	4	Lasso
94	rs304076	4	Lasso
95	rs7034539	4	Lasso
96	rs17204437	4	Lasso
97	rs1537593	4	Lasso
98	rs3742633	4	Lasso
99	rs11637556	4	Boruta
100	rs1491978	4	Boruta
101	rs1537593	4	Boruta
102	rs6787801	4	Boruta
103	rs1038639	4	Boruta
104	rs2071676	4	Boruta
105	rs17204437	4	Boruta
106	rs41305272	4	Boruta
107	rs41282607	4	Boruta
108	rs10499858	4	Boruta
109	rs1388622	4	Boruta
110	rs3173798	4	Boruta

Table 2.18. The frequency table showing the selected significant SNPs in the intermediate models in each iteration of the pipeline. These SNPs are associated with FA platelet response from the dataset 1.

From the data in Figure 2.17 and Table 2.17, the rs11637556 SNP *in MAPK1* has a confidence level of 16/4*4 = 1, (the highest confidence level for an FA platelet response associated SNP). Moreover, 7 key SNPs (rs10499858, rs11637556, rs1388622, rs10974955, rs1038639, rs1491978, and rs2071676) have been identified to be confidently associated with FA platelet responses.

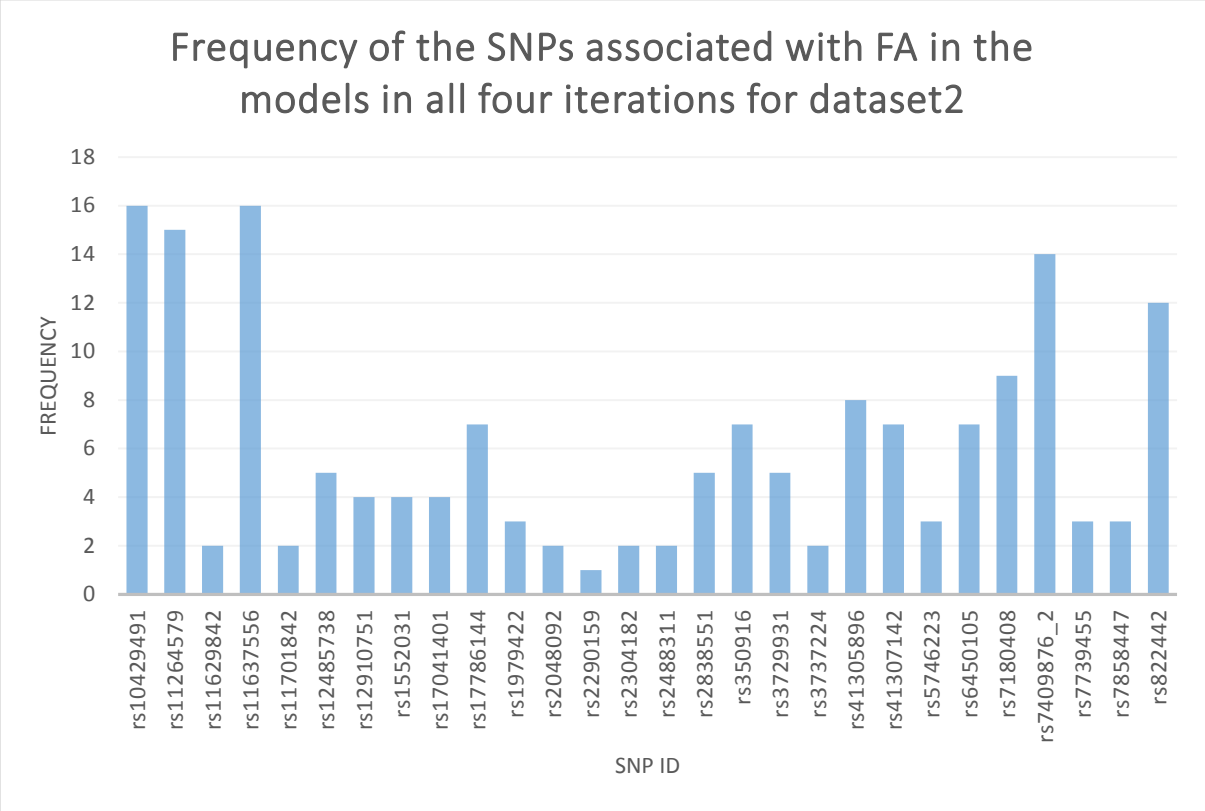


Figure 2.18 The frequency of the selected significant SNPs, which are associated with FA platelet responses in all iterations within the intermediate models for the dataset 2.

SNP's frequency of appearance in the models	SNP's Id	Iteration number	RF + Model Name
1	rs10429491	1	Stepwise
2	rs7409876_2	1	Stepwise
3	rs11637556	1	Stepwise
4	rs11264579	1	Stepwise
5	rs3737224	1	Stepwise
6	rs41305896	1	Stepwise
7	rs7180408	1	Ridge
8	rs10429491	1	Ridge
9	rs2838551	1	Ridge
10	rs7409876_2	1	Ridge
11	rs11637556	1	Ridge
12	rs2488311	1	Ridge
13	rs41307142	1	Ridge
14	rs11264579	1	Ridge
15	rs6450105	1	Ridge
16	rs17786144	1	Ridge
17	rs3737224	1	Ridge
18	rs41305896	1	Ridge
19	rs7180408	1	Lasso

20	rs10429491	1	Lasso
21	rs7409876_2	1	Lasso
22	rs11637556	1	Lasso
23	rs12910751	1	Boruta
24	rs7180408	1	Boruta
25	rs10429491	1	Boruta
26	rs2838551	1	Boruta
27	rs7409876_2	1	Boruta
28	rs11637556	1	Boruta
29	rs1552031	1	Boruta
30	rs11701842	1	Boruta
31	rs822442	1	Boruta
32	rs11264579	1	Boruta
33	rs6450105	1	Boruta
34	rs1979422	1	Boruta
35	rs17786144	1	Boruta
36	rs7739455	1	Boruta
37	rs41305896	1	Boruta
38	rs17041401	1	Boruta
39	rs2290159	1	Boruta
40	rs10429491	2	Stepwise
41	rs3729931	2	Stepwise
42	rs7409876_2	2	Stepwise
43	rs11637556	2	Stepwise
44	rs11264579	2	Stepwise
45	rs350916	2	Stepwise
46	rs822442	2	Stepwise
47	rs10429491	2	Ridge
48	rs7180408	2	Ridge
49	rs3729931	2	Ridge
50	rs7409876_2	2	Ridge
51	rs11637556	2	Ridge
52	rs11264579	2	Ridge
53	rs41305896	2	Ridge
54	rs6450105	2	Ridge
55	rs822442	2	Ridge
56	rs12485738	2	Ridge
57	rs350916	2	Ridge
58	rs2304182	2	Ridge
59	rs10429491	2	Lasso
60	rs3729931	2	Lasso
61	rs7409876_2	2	Lasso
62	rs41307142	2	Lasso
63	rs11637556	2	Lasso
64	rs11264579	2	Lasso

65	rs41305896	2	Lasso
66	rs822442	2	Lasso
67	rs350916	2	Lasso
68	rs2304182	2	Lasso
69	rs17786144	2	Lasso
70	rs10429491	2	Boruta
71	rs7180408	2	Boruta
72	rs17041401	2	Boruta
73	rs7409876_2	2	Boruta
74	rs12910751	2	Boruta
75	rs6450105	2	Boruta
76	rs822442	2	Boruta
77	rs41307142	2	Boruta
78	rs2838551	2	Boruta
79	rs11629842	2	Boruta
80	rs7858447	2	Boruta
81	rs1552031	2	Boruta
82	rs11637556	2	Boruta
83	rs11264579	2	Boruta
84	rs41305896	2	Boruta
85	rs10429491	3	Stepwise
86	rs11637556	3	Stepwise
87	rs7409876_2	3	Stepwise
88	rs11264579	3	Stepwise
89	rs3729931	3	Stepwise
90	rs350916	3	Stepwise
91	rs822442	3	Stepwise
92	rs10429491	3	Ridge
93	rs7180408	3	Ridge
94	rs11637556	3	Ridge
95	rs5746223	3	Ridge
96	rs350916	3	Ridge
97	rs2488311	3	Ridge
98	rs41307142	3	Ridge
99	rs2048092	3	Ridge
100	rs11264579	3	Ridge
101	rs822442	3	Ridge
102	rs12485738	3	Ridge
103	rs10429491	3	Lasso
104	rs11637556	3	Lasso
105	rs5746223	3	Lasso
106	rs7409876_2	3	Lasso
107	rs11264579	3	Lasso
108	rs822442	3	Lasso
109	rs12485738	3	Lasso

110	rs17786144	3	Lasso
111	rs41307142	3	Lasso
112	rs350916	3	Lasso
113	rs2048092	3	Lasso
114	rs10429491	3	Boruta
115	rs7180408	3	Boruta
116	rs12910751	3	Boruta
117	rs1552031	3	Boruta
118	rs11637556	3	Boruta
119	rs1979422	3	Boruta
120	rs7409876_2	3	Boruta
121	rs17041401	3	Boruta
122	rs11264579	3	Boruta
123	rs7858447	3	Boruta
124	rs6450105	3	Boruta
125	rs12485738	3	Boruta
126	rs41305896	3	Boruta
127	rs11629842	3	Boruta
128	rs2838551	3	Boruta
129	rs822442	3	Boruta
130	rs11701842	3	Boruta
131	rs10429491	4	Stepwise
132	rs11264579	4	Stepwise
133	rs11637556	4	Stepwise
134	rs3729931	4	Stepwise
135	rs7409876_2	4	Stepwise
136	rs822442	4	Stepwise
137	rs17786144	4	Stepwise
138	rs350916	4	Stepwise
139	rs7180408	4	Ridge
140	rs10429491	4	Ridge
141	rs11264579	4	Ridge
142	rs11637556	4	Ridge
143	rs822442	4	Ridge
144	rs41307142	4	Ridge
145	rs17786144	4	Ridge
146	rs7739455	4	Ridge
147	rs10429491	4	Lasso
148	rs11264579	4	Lasso
149	rs11637556	4	Lasso
150	rs822442	4	Lasso
151	rs6450105	4	Lasso
152	rs41307142	4	Lasso
153	rs5746223	4	Lasso
154	rs7409876_2	4	Lasso

155	rs17786144	4	Lasso
156	rs12485738	4	Lasso
157	rs7180408	4	Boruta
158	rs10429491	4	Boruta
159	rs1552031	4	Boruta
160	rs12910751	4	Boruta
161	rs17041401	4	Boruta
162	rs11264579	4	Boruta
163	rs2838551	4	Boruta
164	rs11637556	4	Boruta
165	rs6450105	4	Boruta
166	rs7409876_2	4	Boruta
167	rs41305896	4	Boruta
168	rs1979422	4	Boruta
169	rs7858447	4	Boruta
170	rs7739455	5	Boruta

Table 2.19. The frequency table showing the selected significant SNPs in the intermediate models in each iteration of the pipeline. These SNPs are associated with FA platelet response from the dataset 2.

Furthermore, applying equation (2) to the data in Figure 2.18 and Table 2.18, 7 key SNPs (rs10429491, rs11637556, rs11264579, rs41305896, rs7180408, rs7409876_2, and rs822442, which are in *JAK2*, *MAP2K1*, *PEAR1*, *ch5:52979888*, *GTF2A2*, *ch21:44344077*, *and PEAR1* genomic regions) have been identified to be confidently associated with FA platelet responses. In addition, rs10429491 and rs11637556 SNPs obtained the highest confidence level score of 1, meaning that they might be of biological interest for further experiments.

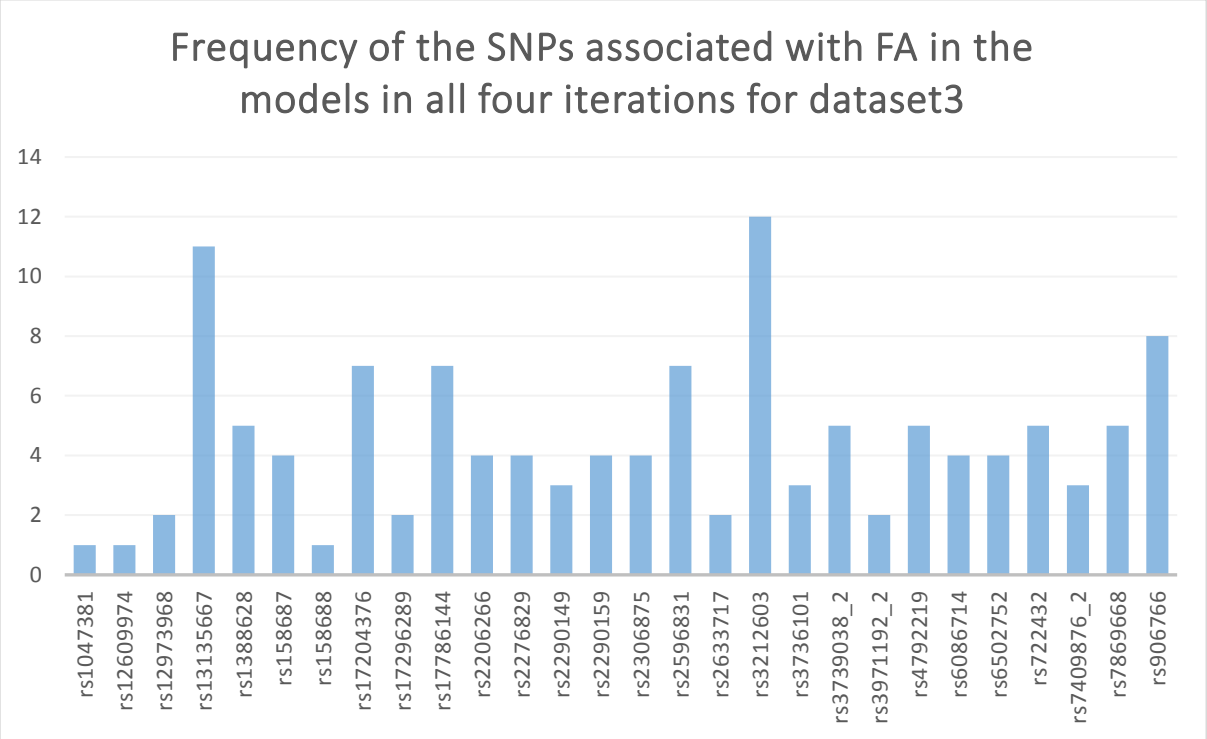


Figure 2.19 The frequency of the selected significant SNPs, which are associated with PA platelet responses in all iterations within the intermediate models for the dataset 3.

SNP's frequency of appearance in the models	SNP's Id	Iteration number	RF + Model Name
1	rs17786144	1	Stepwise
2	rs13135667	1	Stepwise
3	rs3212603	1	Stepwise
4	rs3739038_2	1	Stepwise
5	rs1388628	1	Stepwise
6	rs17786144	1	Ridge
7	rs3971192_2	1	Ridge
8	rs1388628	1	Ridge
9	rs17786144	1	Lasso
10	rs3971192_2	1	Lasso
11	rs3212603	1	Lasso
12	rs3739038_2	1	Lasso
13	rs13135667	1	Lasso
14	rs2290159	1	Boruta
15	rs2596831	1	Boruta
16	rs158687	1	Boruta
17	rs13135667	1	Boruta
18	rs3736101	1	Boruta
19	rs2206266	1	Boruta
20	rs6502752	1	Boruta
21	rs17204376	1	Boruta

22	rs6086714	1	Boruta
23	rs17786144	2	Stepwise
24	rs13135667	2	Stepwise
25	rs3739038_2	2	Stepwise
26	rs1388628	2	Stepwise
27	rs3212603	2	Stepwise
28	rs4792219	2	Ridge
29	rs3212603	2	Ridge
30	rs7869668	2	Ridge
31	rs17786144	2	Lasso
32	rs3739038_2	2	Lasso
33	rs13135667	2	Lasso
34	rs7869668	2	Lasso
35	rs3212603	2	Lasso
36	rs1388628	2	Lasso
37	rs2290149	2	Boruta
38	rs2290159	2	Boruta
39	rs2596831	2	Boruta
40	rs158687	2	Boruta
41	rs13135667	2	Boruta
42	rs1047381	2	Boruta
43	rs1388628	2	Boruta
44	rs2206266	2	Boruta
45	rs3212603	2	Boruta
46	rs3736101	2	Boruta
47	rs6502752	2	Boruta
48	rs2306875	2	Boruta
49	rs17204376	2	Boruta
50	rs6086714	2	Boruta
51	rs13135667	3	Stepwise
52	rs7869668	3	Stepwise
53	rs2633717	3	Stepwise
54	rs3212603	3	Stepwise
55	rs906766	3	Stepwise
56	rs7409876_2	3	Stepwise
57	rs722432	3	Stepwise
58	rs2276829	3	Ridge
59	rs906766	3	Ridge
60	rs722432	3	Ridge
61	rs12973968	3	Ridge
62	rs7869668	3	Ridge
63	rs2306875	3	Ridge
64	rs2633717	3	Ridge
65	rs2596831	3	Ridge
66	rs3212603	3	Ridge

67	rs4792219	3	Ridge
68	rs2596831	3	Lasso
69	rs17204376	3	Lasso
70	rs13135667	3	Lasso
71	rs3212603	3	Lasso
72	rs4792219	3	Lasso
73	rs7869668	3	Lasso
74	rs7409876_2	3	Lasso
75	rs2276829	3	Lasso
76	rs906766	3	Lasso
77	rs722432	3	Lasso
78	rs12973968	3	Lasso
79	rs2290159	3	Boruta
80	rs2596831	3	Boruta
81	rs906766	3	Boruta
82	rs158687	3	Boruta
83	rs17204376	3	Boruta
84	rs6502752	3	Boruta
85	rs3212603	3	Boruta
86	rs2206266	3	Boruta
87	rs2290149	3	Boruta
88	rs3736101	3	Boruta
89	rs2306875	3	Boruta
90	rs6086714	3	Boruta
91	rs7409876_2	3	Boruta
92	rs906766	4	Stepwise
93	rs17786144	4	Stepwise
94	rs13135667	4	Stepwise
95	rs17204376	4	Stepwise
96	rs3212603	4	Stepwise
97	rs17296289	4	Stepwise
98	rs3739038_2	4	Stepwise
99	rs2306875	4	Ridge
100	rs4792219	4	Ridge
101	rs722432	4	Ridge
102	rs906766	4	Ridge
103	rs17296289	4	Ridge
104	rs906766	4	Lasso
105	rs17786144	4	Lasso
106	rs4792219	4	Lasso
107	rs12609974	4	Lasso
108	rs2276829	4	Lasso
109	rs722432	4	Lasso
110	rs2596831	4	Lasso
111	rs17204376	4	Lasso

112	rs13135667	4	Lasso
113	rs3212603	4	Lasso
114	rs2290159	4	Boruta
115	rs2596831	4	Boruta
116	rs158687	4	Boruta
117	rs17204376	4	Boruta
118	rs6502752	4	Boruta
119	rs6086714	4	Boruta
120	rs158688	4	Boruta
121	rs2290149	4	Boruta
122	rs2206266	4	Boruta
123	rs13135667	4	Boruta
124	rs2276829	4	Boruta
125	rs906766	4	Boruta

Table 2.20. The frequency table showing the selected significant SNPs in the intermediate models in each iteration of the pipeline. These SNPs are associated with FA platelet response from the dataset 3.

For the case of dataset 3 (Figure 2.19 and Table 2.19), which is associated with FA, when the equation (2) is applied, 6 key SNPs (rs13135667, rs17204376, rs17786144, rs2596831, rs3212603, and rs906766) were identified with high confidence to be true (key) positive and significantly associated with FA.

2.3.1.1 ***Validation of the pipeline – Identified consensus or key artificial SNPs***

Furthermore, several of the artificially simulated genotyped SNPs were consistently identified across the methods (i.e. in consensus) in the final optimised RF model and were significantly associated with the simulated continuous phenotype. Table 2.21 and Figure 2.20 show the selected key and significant artificial simulated genotyped SNPs in a consensus manner.

Artificial SNPs ID	SNP's significance in the models				
	RF with Stepwise	**RF with Ridge regression**	**RF with LASSO**	**RF with Boruta (P=0.01)**	**Consensus (3/4)**
X1306	✓(4.15e-05)	✓(0.0003)	✓(1.73e-05)	✓	✓
X92	✓(0.0004)	✓(0.003)	✓(0.0014)	×	✓
X1112	✓(0.0017)	✓(0.00204)	✓(0.0013)	✓	✓
X808	✓(0.0013)	×	✓(0.0073)	×	
X859	✓(0.0034)	✓(0. 0061)	✓(0.001)	×	✓
X263	✓(0.0021)	✓(0.0151)	✓(0.0061)	✓	✓
X829	×	✓(0.0065)	✓(0.009)	×	
X1203	×	✓(0.0171)	×	✓	
X242	×	✓(0.0075)	✓(0.003)	✓	✓
X56	×	✓(0.0135)	✓(0.0071)	✓	✓
X1051	×	✓(0.0122)	✓(0.005)	×	
X877	×	✓(0.0151)	×	×	
X512	×	✓(0.0019)	✓(0.0131)	×	
X847	×	×	✓(0.01)	×	
X760	×	×	×	✓	

Table 2.21 The selected consensus artificial SNPs from the simulated data set. X_m represents an identifier of the genotyped SNP m in the simulated data set. Several of the significant SNPs associated with phenotype were selected across the methods meaning that they are likely to be key significant SNPs associated with the complex phenotype. × indicates either the SNP was not identified by the method. ✓ indicates the SNP was identified by the method. Numbers inside the brackets after ✓ indicate p values of the SNPs calculated using Wald test, or partial F-test.

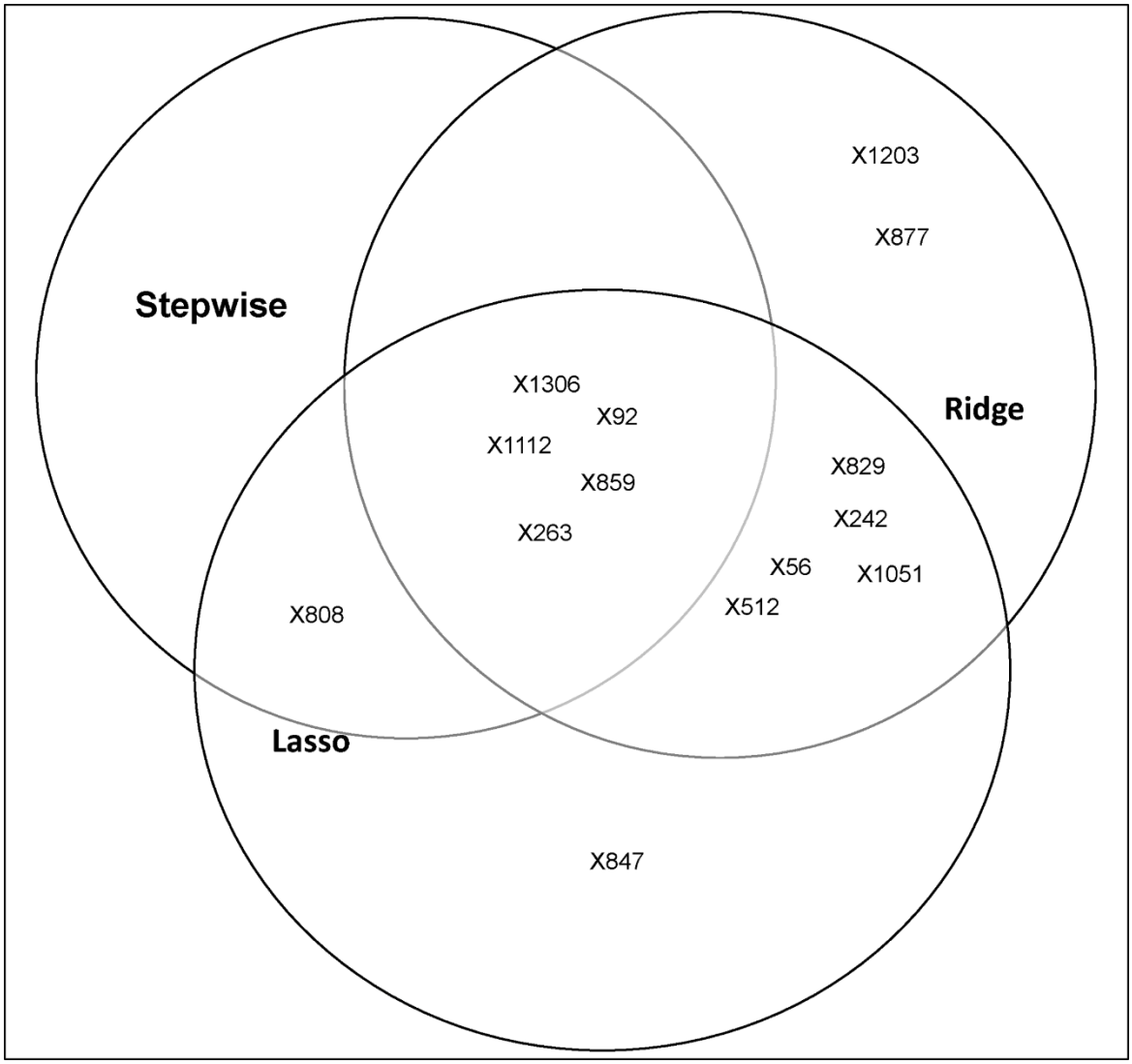


Figure 2.20 The visualisation of the detected key significant artificial SNPs (intersection regions). X_m represents the identifier of the simulated genotyped artificial SNP m. Several simulated artificial SNPs were consistently identified to be significant by the multiple methods as occurred in the actual SNPs data set.

Furthermore, Figure 2.21 shows the significant artificial SNPs, which were detected by the Boruta layer, which further enhances the key SNP identification in addition to the regression layer.

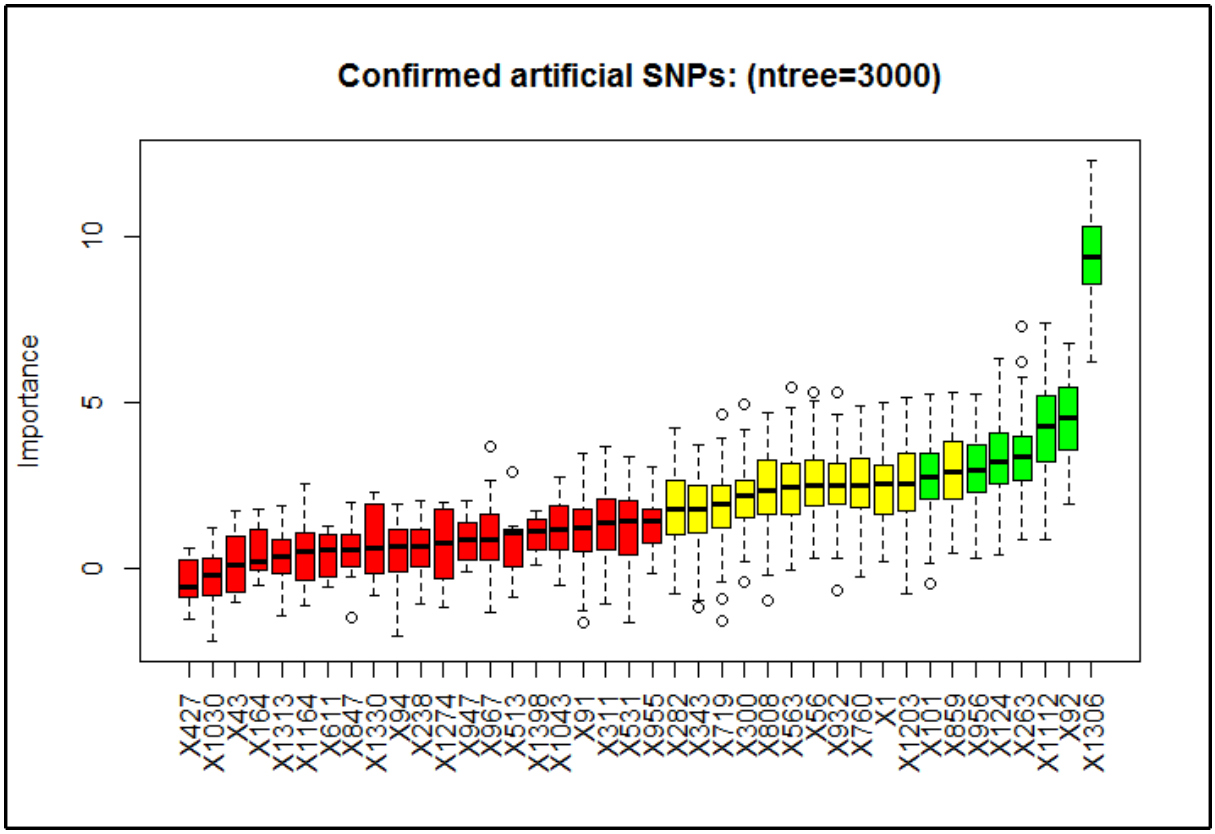


Figure 2.21 The confirmed selected artificially simulated key SNPs by the Boruta. It can be clearly seen that most of the selected SNPs in the regression layer have been also selected by Boruta, which further enhanced key SNPs selection.

For the identified significant simulated artificial SNPs, the frequency plot (Figure 2.22) was initially created, and the equation 2 was applied to determine true positive key SNPs. Table 2.22 shows the frequency of significant simulated artificial SNPs, which were selected in every iteration.

For instance, applying the equation 2 to the simulated significant SNP X56, the confidence level will be 0.6875, which has surpassed the minimum threshold confidence level and hence, is more likely to be a true positive key SNP. In total 8 artificially simulated SNPs were identified to be confidently associated with the simulated phenotype.

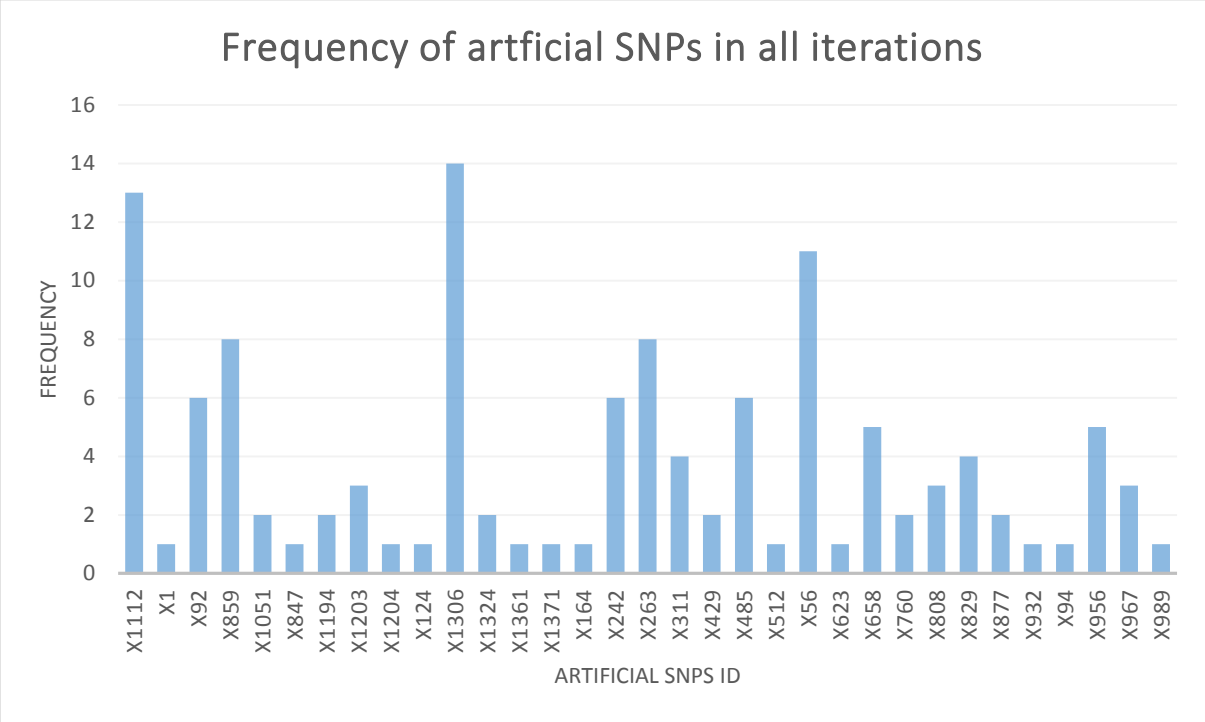


Figure 2.22 The frequency plot showing the overall selected significant artificially simulated SNPs in the intermediate models in all four iterations. The highly 'enriched' simulated SNPs can be easily identified.

Artificial SNP's frequency of appearing in the models	Artificial SNP's Id	Iteration number	RF + Model Name
1	X1306	1	Stepwise
2	X1112	1	Stepwise
3	X485	1	Stepwise
4	X967	1	Stepwise
5	X56	1	Stepwise
6	X956	1	Stepwise
7	X1306	1	Ridge
8	X1112	1	Ridge
9	X56	1	Ridge
10	X967	1	Ridge
11	X989	1	Ridge
12	X859	1	Ridge
13	X485	1	Ridge
14	X1204	1	Ridge
15	X956	1	Ridge
16	X1306	1	Lasso
17	X1112	1	Lasso
18	X56	1	Lasso
19	X967	1	Lasso
20	X859	1	Lasso
21	X485	1	Lasso
22	X956	1	Lasso

23	X1361	1	Lasso
24	X1306	1	Boruta
25	X1203	1	Boruta
26	X760	1	Boruta
27	X1112	1	Boruta
28	X56	1	Boruta
29	X485	1	Boruta
30	X94	1	Boruta
31	X956	1	Boruta
32	X1306	2	Stepwise
33	X263	2	Stepwise
34	X485	2	Stepwise
35	X311	2	Stepwise
36	X56	2	Stepwise
37	X658	2	Stepwise
38	X242	2	Stepwise
39	X1306	2	Ridge
40	X1112	2	Ridge
41	X1203	2	Ridge
42	X429	2	Ridge
43	X263	2	Ridge
44	X1324	2	Ridge
45	X658	2	Ridge
46	X242	2	Ridge
47	X311	2	Ridge
48	X1306	2	Lasso
49	X1112	2	Lasso
50	X56	2	Lasso
51	X263	2	Lasso
52	X429	2	Lasso
53	X1324	2	Lasso
54	X658	2	Lasso
55	X877	2	Lasso
56	X242	2	Lasso
57	X485	2	Lasso
58	X311	2	Lasso
59	X1306	2	Boruta
60	X1112	2	Boruta
61	X760	2	Boruta
62	X263	2	Boruta
63	X1306	3	Stepwise
64	X92	3	Stepwise
65	X1112	3	Stepwise
66	X1	3	Stepwise
67	X56	3	Stepwise

68	X658	3	Stepwise
69	X859	3	Stepwise
70	X808	3	Stepwise
71	X1306	3	Ridge
72	X92	3	Ridge
73	X1112	3	Ridge
74	X56	3	Ridge
75	X829	3	Ridge
76	X859	3	Ridge
77	X956	3	Ridge
78	X1	3	Ridge
79	X311	3	Ridge
80	X242	3	Ridge
81	X164	3	Ridge
82	X1194	3	Ridge
83	X658	3	Ridge
84	X1306	3	Lasso
85	X92	3	Lasso
86	X263	3	Lasso
87	X1	3	Lasso
88	X1371	3	Lasso
89	X1112	3	Lasso
90	X56	3	Lasso
91	X829	3	Lasso
92	X859	3	Lasso
93	X124	3	Lasso
94	X623	3	Lasso
95	X1194	3	Lasso
96	X932	3	Lasso
97	X1306	4	Stepwise
98	X92	4	Stepwise
99	X1112	4	Stepwise
100	X808	4	Stepwise
101	X859	4	Stepwise
102	X263	4	Stepwise
103	X1306	4	Ridge
104	X1112	4	Ridge
105	X92	4	Ridge
106	X1203	4	Ridge
107	X829	4	Ridge
108	X242	4	Ridge
109	X56	4	Ridge
110	X859	4	Ridge
111	X263	4	Ridge
112	X1051	4	Ridge

113	X877	4	Ridge
114	X1306	4	Lasso
115	X92	4	Lasso
116	X808	4	Lasso
117	X1112	4	Lasso
118	X56	4	Lasso
119	X829	4	Lasso
120	X859	4	Lasso
121	X242	4	Lasso
122	X847	4	Lasso
123	X263	4	Lasso
124	X1051	4	Lasso
125	X512	4	Lasso

Table 2.22. The frequency table showing the selected significant artificially simulated SNPs in the intermediate models in each iteration of the pipeline. These SNPs are associated with simulated continuous phenotype.

2.3.1.2 ***Validation with second simulated set***

With second simulated set, which contains fewer minor alleles among artificial SNPs, the results appear to be similar with those in Figure 2.22. Several artificially genotyped SNPs, which were detected to be key in the previous section 2.3.1.1 were again identfied to be key. Few of the SNPs from the new simulated set were not detected signifying the effect of taking into account the true representation of the minor alleles. Figure 2.23 shows the frequencies of these identified artificial SNPs in the intermediate models.

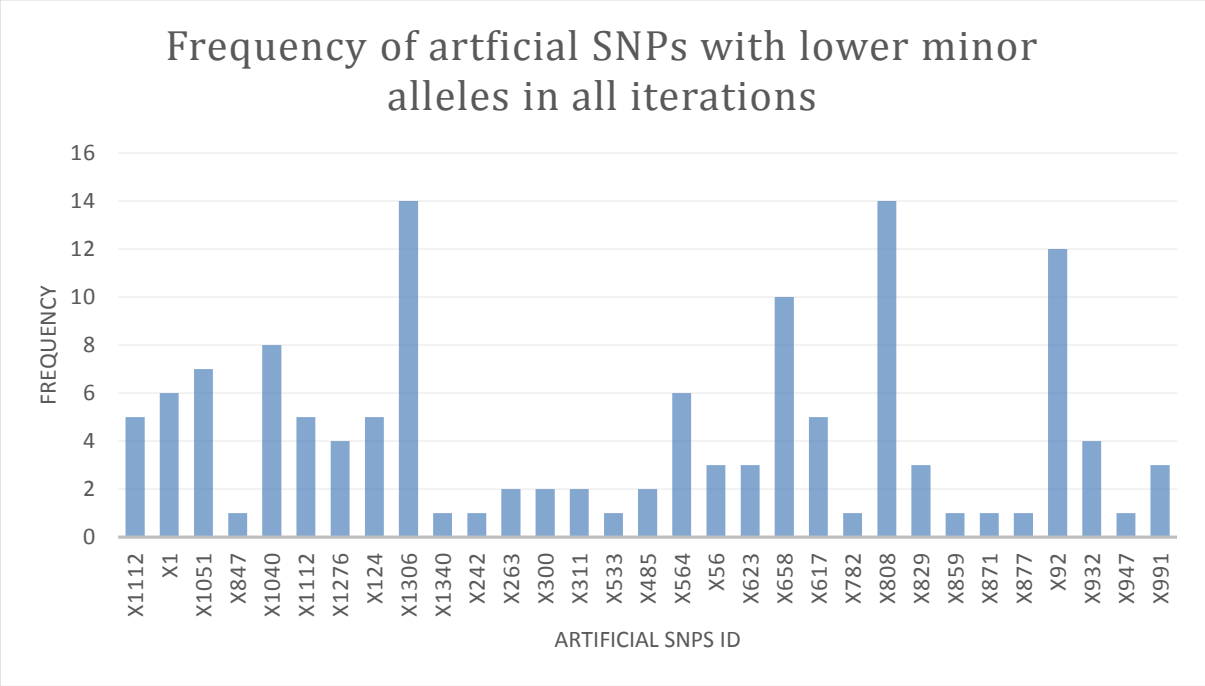


Figure 2.23 The frequency plot showing the overall selected significant second artificially simulated SNPs in the intermediate models in all four iterations. The second simulated SNPs contain low number of minor alleles.

From Figure 2.23, the artificially SNPs such as X1112, X859, and several others, which were identified to be key are now not confidently identified to be key SNPs contrasting the results from the first simulated set. On other hand, the artificial SNPs such as X92, X808, and several others, which were not confidently identified to be key are now significantly detected to be key. The artificial SNP X1306 was confidently identified to be the most significant in both simulated sets. The results further show that RAPIDSNPs appears to be capable of identifying the most significant SNPs when the validation set reflects true genotypic states using the same model assumptions.

Additionally, the similar artificial key SNPs as those in Figure 2.23 were being picked by the RAPIDSNPs when the simulated phenotype values were being altered (Figure 2.24). This may indicate that these are true disease/trait associated artificial key SNPs even under the alteration of phenotypic values. There is notable variation of confidence levels amongsts trait associated

SNPs relative to those in Figure 2.23. For instance, the X112 and X1, which have low confidence scores (Figure 2.23) are now associated with the phenotype with high scores.

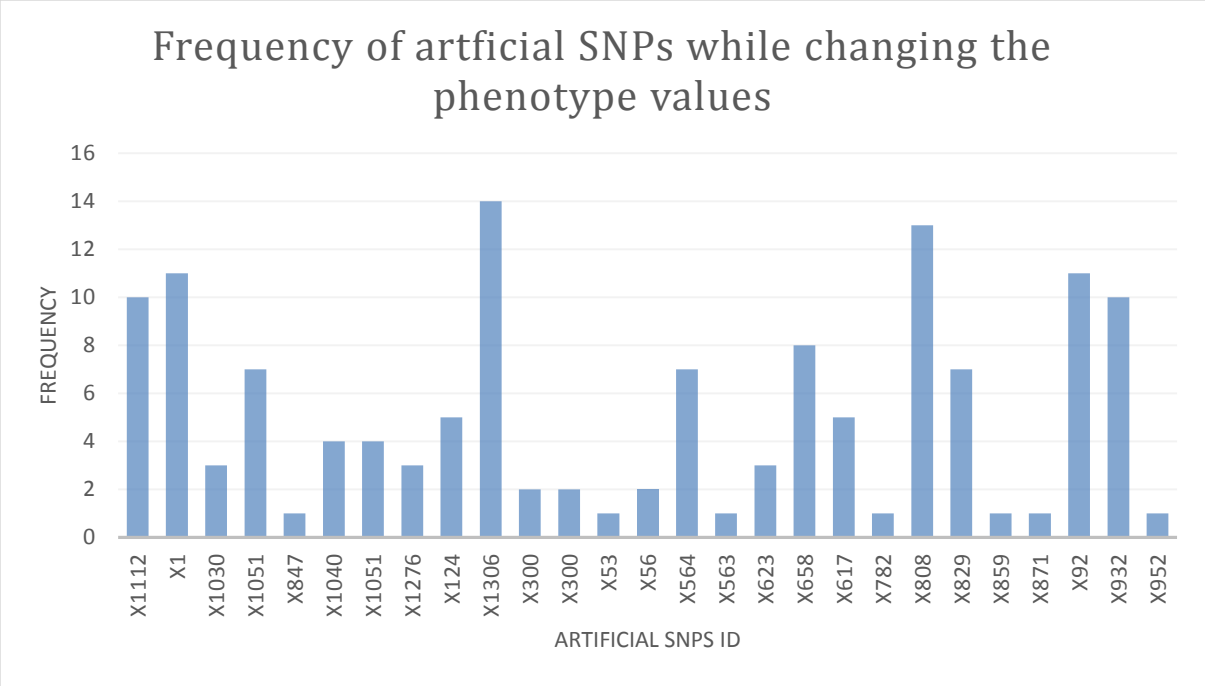


Figure 2.24 The frequency plot showing the overall selected significant second artificially simulated SNPs in the intermediate models in all four iterations when the phenotype values were being perturbed.

Thus, the selection of the same artificial key SNPs patterns even after being altering the phenotype values means that there are real and not *random* association between these selected artificial key SNPs and the phenotype. Hence, these results further indicate and validate the robustness of the RAPIDSNPs in detecting the disease/trait associated key SNPs.

2.3.2 Effects of age as an example covariate and the selection of key SNPs

It has been found that in most cases, the key SNPs, which were significantly identified to be associated with the platelet responses when the pipeline is run age incorporated as a covariate are the same as those when age is not incorporated. For instance, for PA platelet response, most of the SNPs were identical to those selected when age is not included, signifying that the age

might have a less significant effect when it is combined with SNPs in explaining the PA variation. Table 2.23 shows the frequencies of the SNPs selection in the intermediate models associated with the PA platelet response in each iteration, when age is included as a covariate. Figure 2.24 shows the plot, which illustrates the most frequent selected significant SNPs that are associated with PA for all iterations of the pipeline. All of the selected key SNPs are the same, except rs8033381, which was not selected under the presence of age as a covariate.

SNP's frequency of appearance in the models	SNP's Id	Iteration number	RF + Model Name
1	rs3212391	1	Stepwise
2	rs6141803	1	Stepwise
3	rs2300065	1	Stepwise
4	rs6442895	1	Stepwise
5	rs12592919	1	Stepwise
6	rs12709458	1	Ridge
7	rs12592919	1	Ridge
8	rs6442896	1	Ridge
9	rs6136	1	Lasso
10	rs2300065	1	Lasso
11	rs927239	1	Lasso
12	rs12592919	1	Lasso
13	rs6442896	1	Lasso
14	rs3212391	1	Lasso
15	rs3212391	1	Boruta
16	rs6141803	1	Boruta
17	rs3212386	1	Boruta
18	rs16865105	1	Boruta
19	rs6442895	1	Boruta
20	rs2424895	1	Boruta
21	rs2424905	1	Boruta
22	rs6442896	1	Boruta
23	rs6433658	1	Boruta
24	rs17041401	1	Boruta
25	rs13316843	1	Boruta
26	rs12592919	1	Boruta
27	rs26682	2	Stepwise
28	rs2292867	2	Stepwise
29	rs6141803	2	Stepwise
30	rs6442895	2	Stepwise
31	rs1527480	2	Stepwise

32	rs6442896	2	Ridge
33	rs6442895	2	Ridge
34	rs2292867	2	Ridge
35	rs3745406	2	Ridge
36	rs1527480	2	Ridge
37	rs6141803	2	Lasso
38	rs12592919	2	Lasso
39	rs7187863	2	Lasso
40	rs1527480	2	Lasso
41	rs10061730	2	Lasso
42	rs3745406	2	Lasso
43	rs2292867	2	Lasso
44	rs6141803	2	Boruta
45	rs6442895	2	Boruta
46	rs6442896	2	Boruta
47	rs6433658	2	Boruta
48	rs17041401	2	Boruta
49	rs12592919	2	Boruta
50	rs26682	2	Boruta
51	rs10061730	2	Boruta
52	rs3212386	2	Boruta
53	rs2424905	2	Boruta
54	rs2424895	2	Boruta
55	rs1527480	3	Stepwise
56	rs3212391	3	Stepwise
57	rs3730051	3	Stepwise
58	rs6442896	3	Stepwise
59	rs6141803	3	Stepwise
60	rs2300065	3	Stepwise
61	rs11637556	3	Stepwise
62	rs6141803	3	Ridge
63	rs3730051	3	Ridge
64	rs6442896	3	Ridge
65	rs1527480	3	Ridge
66	rs11637556	3	Ridge
67	rs6442895	3	Ridge
68	rs6141803	3	Lasso
69	rs3730051	3	Lasso
70	rs6442896	3	Lasso
71	rs1527480	3	Lasso
72	rs2300065	3	Lasso
73	rs3212386	3	Lasso
74	rs11637556	3	Lasso
75	rs6141803	3	Boruta
76	rs6442895	3	Boruta

77	rs6442896	3	Boruta
78	rs6433658	3	Boruta
79	rs17041401	3	Boruta
80	rs3212391	3	Boruta
81	rs17760545	3	Boruta
82	rs2424895	3	Boruta
83	rs7568033	3	Boruta
84	rs3212418	3	Boruta
85	rs26682	3	Boruta
86	rs33443	3	Boruta
87	rs246410	3	Boruta
88	rs397454	3	Boruta
89	rs2424905	3	Boruta
90	rs16865105	3	Boruta
91	rs6442896	4	Stepwise
92	rs3212391	4	Stepwise
94	rs3730051	4	Stepwise
95	rs6141803	4	Stepwise
96	rs1527480	4	Stepwise
97	rs11637556	4	Stepwise
98	rs6442896	4	Ridge
99	rs3730051	4	Ridge
100	rs1527480	4	Ridge
101	rs11637556	4	Ridge
102	rs6442896	4	Lasso
103	rs3730051	4	Lasso
104	rs11637556	4	Lasso
105	rs3212418	4	Lasso
106	rs6141803	4	Lasso
107	rs6442896	4	Boruta
108	rs6433658	4	Boruta
109	rs6442895	4	Boruta
110	rs3212391	4	Boruta
111	rs17041401	4	Boruta
112	rs6058869	4	Boruta
113	rs33443	4	Boruta
114	rs6895049	4	Boruta
115	rs2424895	4	Boruta
116	rs26682	4	Boruta
117	rs3212418	4	Boruta
118	rs2424905	4	Boruta
119	rs7568033	4	Boruta
120	rs16865105	4	Boruta

Table 2.23 The frequency of each selected significant SNP associated with PA platelet response in each iteration

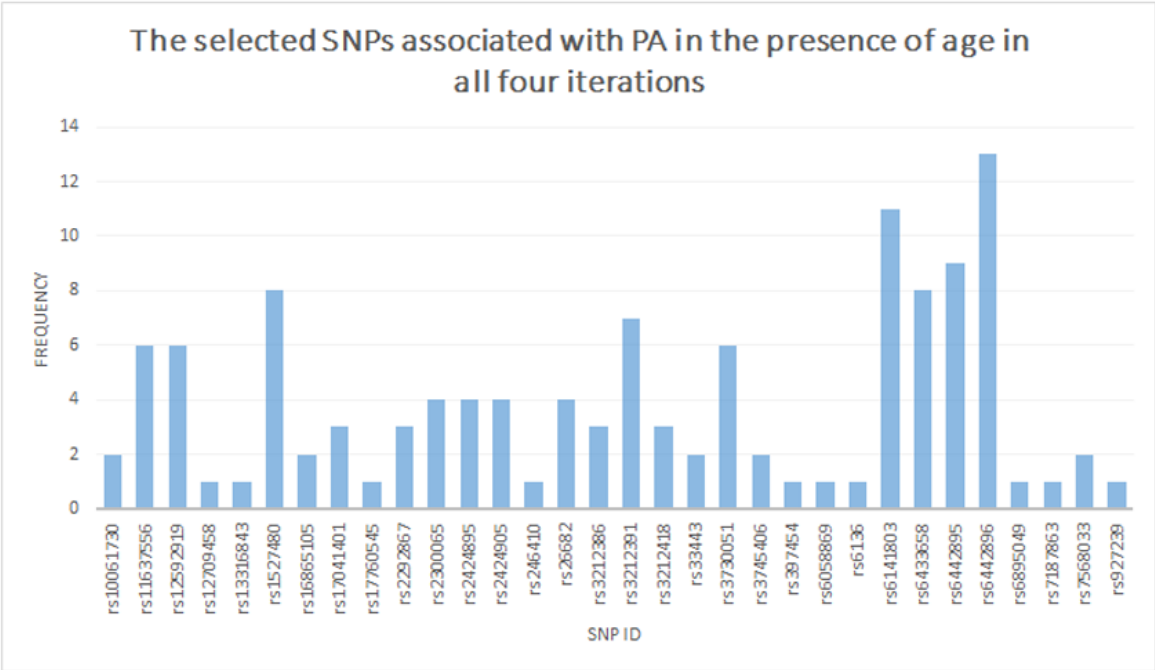


Figure 2.25 The frequency plot showing the overall selected significant SNPs that are associated with the PA platelet response in the intermediate models in all four iterations when age is included as a covariate. Most of the selected SNPs are similar to those selected when age is not included as a covariate.

Furthermore, for the FA platelet response, nearly all the SNPs, which were identified to be significantly associated with FA when age is excluded are the same with those under the inclusion of age. However, in some stages of the pipeline run, age appears to have a likely association with FA platelet response, but in addition to other key SNPs. The plot in Figure 2.25 with its related table in Table 2.24 shows the different SNPs that are selected in every iteration; age is selected in the fourth iteration by the stepwise method with a p-value of 0.016.

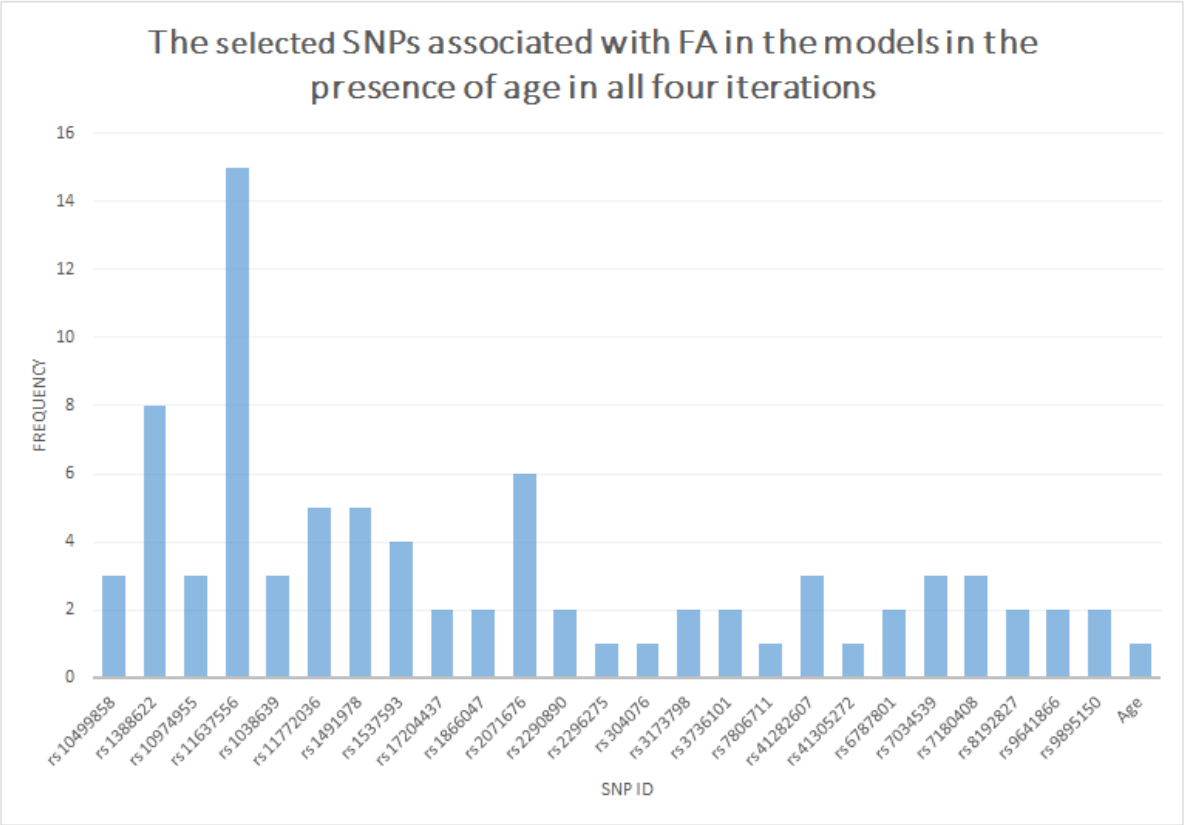


Figure 2.26 The frequency plot showing the overall selected significant SNPs that are associated with the FA platelet response in the intermediate models in all four iterations when age is included as a covariate.

SNP's frequency of appearance in the models	SNP's Id	Iteration number	RF + Model Name
1	rs1388622	1	Stepwise
2	rs11637556	1	Stepwise
3	rs8192827	1	Stepwise
4	rs1537593	1	Stepwise
5	rs11637556	1	Ridge
6	rs7180408	1	Ridge
7	rs1038639	1	Ridge
8	rs11637556	1	Lasso
9	rs1537593	1	Lasso
10	rs1388622	1	Lasso
11	rs3173798	1	Boruta
12	rs1491978	1	Boruta
13	rs11637556	1	Boruta
14	rs1388622	1	Boruta
15	rs17204437	1	Boruta
16	rs1537593	1	Boruta
17	rs9641866	1	Boruta

18	rs1388622	2	Stepwise
19	rs11637556	2	Stepwise
20	rs7180408	2	Stepwise
21	rs2290890	2	Stepwise
22	rs11637556	2	Ridge
23	rs7806711	2	Ridge
24	rs9895150	2	Ridge
25	rs7180408	2	Ridge
26	rs11637556	2	Lasso
27	rs1388622	2	Lasso
28	rs2290890	2	Lasso
29	rs11772036	2	Lasso
30	rs41305272	2	Lasso
31	rs9895150	2	Lasso
32	rs1388622	2	Boruta
33	rs11637556	2	Boruta
34	rs6787801	2	Boruta
35	rs1491978	2	Boruta
36	rs1491978	3	Stepwise
37	rs11637556	3	Stepwise
38	rs10974955	3	Stepwise
39	rs10499858	3	Stepwise
40	rs2071676	3	Stepwise
41	rs41282607	3	Stepwise
42	rs1866047	3	Stepwise
43	rs2071676	3	Ridge
44	rs11637556	3	Ridge
45	rs41282607	3	Ridge
46	rs3212417	3	Ridge
47	rs11637556	3	Lasso
48	rs2071676	3	Lasso
49	rs10974955	3	Lasso
50	rs10499858	3	Lasso
51	rs41282607	3	Lasso
52	rs1866047	3	Lasso
53	rs8192827	3	Lasso
54	rs3736101	3	Lasso
55	rs1491978	3	Boruta
56	rs7034539	3	Boruta
57	rs10974955	3	Boruta
58	rs3173798	3	Boruta
59	rs1537593	3	Boruta
60	rs17204437	3	Boruta
61	rs10499858	3	Boruta
62	rs1388622	4	Stepwise

63	rs11637556	4	Stepwise
64	Age	4	Stepwise
65	rs2071676	4	Stepwise
66	rs7034539	4	Stepwise
67	rs2296275	4	Stepwise
68	rs11772036	4	Stepwise
69	rs11637556	4	Ridge
70	rs7034539	4	Ridge
71	rs1038639	4	Ridge
72	rs2071676	4	Ridge
73	rs11637556	4	Lasso
74	rs3736101	4	Lasso
75	rs2071676	4	Lasso
76	rs11772036	4	Lasso
77	rs1038639	4	Lasso
78	rs304076	4	Lasso
79	rs11637556	4	Boruta
80	rs1491978	4	Boruta
81	rs11772036	4	Boruta
82	rs6787801	4	Boruta
83	rs1388622	4	Boruta
84	rs9641866	4	Boruta

Table 2.24 The frequency of each selected significant SNP associated with the FA response in each iteration.

The age was separately tested with the key SNPs (rs11637556, rs1388622, and rs2071676) and found that it has a likely significance with FA platelet response (p-value = 0.05) along with rs1388622 and rs11637556 of *P2RY12* and *MAP2K1* respectively. Moreover, in almost every iteration of the RF, age was among the top ranked predictors, in addition to other SNPs, S17 Figure 2.26.

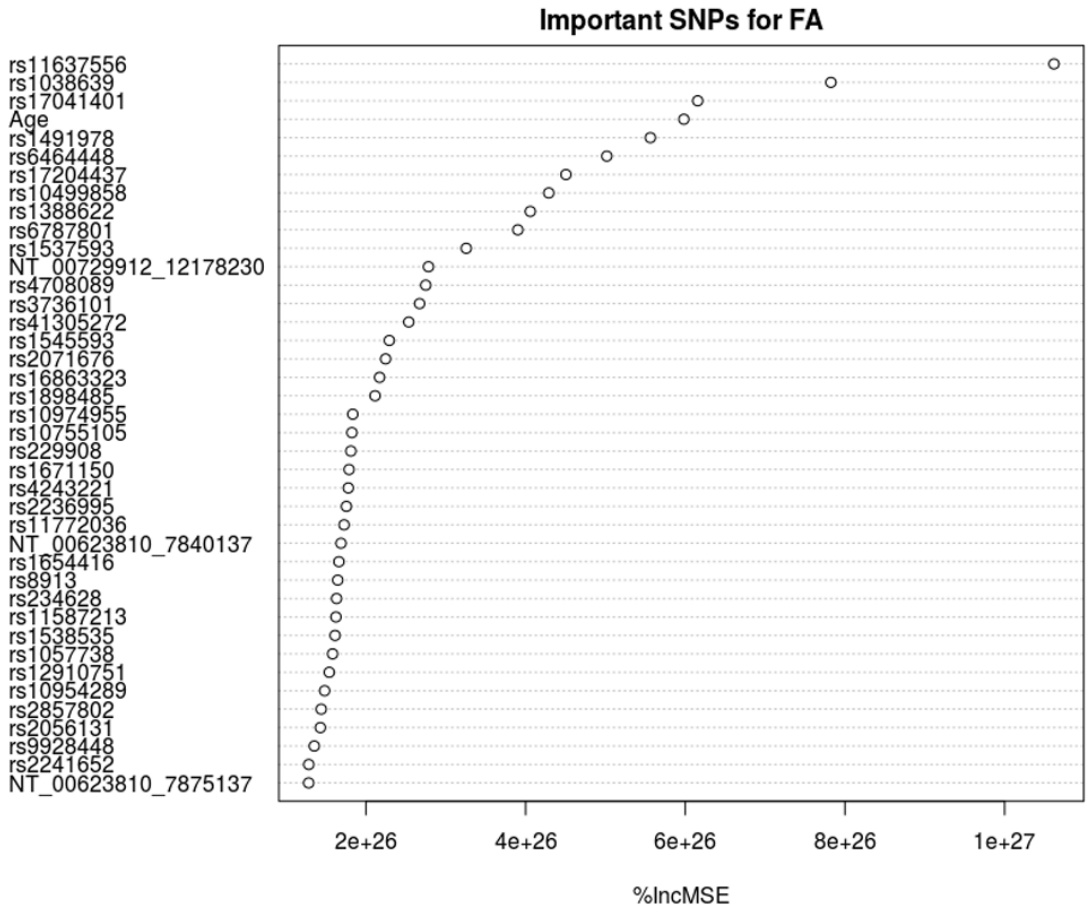


Figure 2.27 The importance of the variables (SNPs and age) in y-axis, which have been selected by the RF based on their ranks and that are associated with FA platelet response. The x-axis shows the increase in mean squared estimate as a function of the individual variable's importance. The higher %IncMSE the higher the variable to be likely important or significant in affecting the phenotype, which in this case is FA platelet response.

Therefore, the similarity in the performance of the pipeline and the resultant key SNPs selection patterns using both, the real and simulated genotyped SNPs datasets, indicates that the pipeline is likely to be workable and robust when is applied to other continuous phenotypes. In addition, the pipeline has shown to be promising in simultaneously handling or analysing the SNPs in the presence of covariates while investigating the association of the SNPs and continuous complex trait/disease.

2.3.3 The identified missense (non-synonymous SNPs)

Several significant SNPs across the datasets were identified to be non-synonymous, or missense mutations. Thus, they are likely to have structural/functional effects on the underlying *in vivo* related proteins, which are likely to further contribute to the variability of PA and FA platelet responses and hence, differential individual response to treatments and CVD risks. Table 2.25 shows the missense SNPs that were identified by the pipeline from the three multiple datasets used.

SNP	Dataset	ADP platelet responses	Gene/Locus
rs2071676	1	FA	*CA9*
rs12953	3	PA	*PECAM1*
rs822442	2	FA	*PEAR1*
rs3739038	3	FA	*LRRFIP1*
rs3736101	1	FA	*MADD*

Table 2.25 The overall identified missense SNPs, which are significantly associated with ADP platelet responses from the different three datasets applied to the pipeline. The bolded SNPs were unidentified in the previous study. These SNPs might be worth considering investigation of their structural/functional effects related to their proteins and that may further underlie the variability of the ADP platelet responses.

Therefore, in the next chapter, these identified missense SNPs are investigated their likely damaging level to the structural/function roles of related proteins that may likely to explain the molecular mechanisms underpinning the ADP platelet responses variability.

2.4 Discussion and Conclusion

2.4.1 Advantages of this approach (RAPIDSNPs)

There are several advantages of the RAPIDSNPs approach. Firstly, the RF layer plays a crucial role in ensuring that potentially highly important SNPs are selected and passed through to the regression ensemble and Boruta layers. This use of the RF as an initial filtering stage is a well described standard approach for SNP discovery and plays a crucial role in selecting potentially highly important SNPs, using the appropriate *ntree* and VI parameters (Goldstein et al., 2011;

Nicodemus et al., 2010; Strobl and Zeileis, 2008). The selection bias introduced by the VI measure with the ranking approach has been shown to mainly affect predictors with different categories and scale of measurements (Strobl et al., 2007), which is not the case in this study. In addition, the use of the VI measure with a ranking approach is still regarded as a useful strategy for selecting important SNPs for downstream analyses (Braga-Neto et al., 2004; Díaz-Uriarte and Andrés, 2006).

Secondly, it is possible to rapidly identify the key genetic variants, or markers, using a consensus of multiple alternate methods. Additionally, by introducing the multiple alternate methods layers, the likelihood of identifying other significant SNPs that might have been missed in one or more of the methods increases. This combination of methods in an integrated manner is a good approach for reducing false positives as multiple methods might be pointing to the same SNPs (Ritchie et al., 2015). This potentially increases the chance of keeping functional SNPs associated with the phenotype, minimising the risk of 'missing heritability' (Manolio et al., 2009), which is one of the thorny issues in GASs (Moore et al., 2010). Moreover, based on this approach, the identified true complex trait associated key SNPs are more likely to be indicating the significantly overexpressed loci, which are likely to be proper candidates for follow-up experiments.

Furthermore, the pipeline is computationally adaptable and scalable to different implementations, particularly in the regression methods ensemble layer. It is possible to increase the number of (regularised) regression methods for optimising the detection of the key SNPs through consensus identification.

Furthermore, the computational speed of the pipeline means that is practical to implement as an additional tool. For the dataset 1, the time taken to run the entire pipeline was 229.77012 secs on a modest quad core system running Ubuntu 14.04. The RAPIDSNPs does not

necessarily aim to replace existing methods such as EMMAX (Kang et al., 2010) and PLINK (Purcell et al., 2007), rather it may be used to supplement and further enhance the identification of key SNPs associated with continuous response phenotypes, with little additional computational overhead.

In addition, the pipeline may have an observed advantage over existing RF based methods in terms of its ability to identify other true trait associated SNPs. For instance, Boruta is the RF based method for relevant feature selection. After comparing the SNPs that were obtained after running the pipeline with those from the Boruta method, it was found that the pipeline has the potential edge in identifying key SNPs, which might be missed by using only Boruta. For example, in the case of the PA associated significant SNPs for the dataset1, it was found that the pipeline is able to identify rs3730051 in the *AKT2* locus as a key SNP, which was not recognised as a relevant important feature by the Boruta.

2.4.2 Limitations/caveats of the approach (RAPIDSNPs)

The limitations of the approach are discussed below.

2.4.2.1 ***Sample size of the SNPs data***

Furthermore, the RAPIDSNPs is likely to be most suitable for genetic association studies with relatively small SNP datasets (Reif et al., 2006), and it appears to perform well when applied to the platelet responses data. However, this approach has not yet been tested or applied to genome-wide scale data e.g. with several million SNPs for association mapping. In such cases, the subspace SNPs selection methods could be initially employed (Nguyen et al., 2015; Wu et al., 2012), for selecting a subspace of informative SNPs and minimising the computational cost in generating trees, prior to using this approach.

2.4.2.2 ***Missing genotypes***

For large numbers of missing genotypes, several established methods and tools, such as IMPUTE (Howie et al., 2009; Marchini and Howie, 2010), Beagle (Browning and Browning, 2007) and PLINK could be used.

2.4.2.3 ***Long range LD and rare variants***

The pipeline is solely generic in use for the identification of key significant SNPs within candidate genes associated with continuous phenotypic traits. For examining whether the identified SNPs are in long range LD (Koch et al., 2013), the pipeline could be supplemented with other methods or tools such as GLIDERS (Lawrence et al., 2009) and GWAS3D (Li et al., 2013). Furthermore, the pipeline has not been tested whether it is able to detect the rare variants. Instead, other approaches such as those proposed by Hoffmann et al. (Hoffmann et al., 2010), sequence kernel association test (SKAT) (Wu et al., 2011), and kernel-based adaptive cluster (KBAC) (Liu and Leal, 2010) might be used accordingly for detection of rare variants.

2.4.3 Newly identified SNPs and their biological and clinical significance

This approach was able to discover numerous and previously undetected SNPs, which are significantly associated with the ADP platelet response phenotype. Several of these SNPs have also been highlighted in other independent studies as being implicated in CVD. The following examples underpin the results and serve to further strengthen the confidence on the ability of the approach for identify key genetic variants.

For example, the identified intergenic SNP rs6141803 in *COMMD7*, which is associated with PA was also identified in another platelet functional study (Goodall et al., 2010) to be a likely

risk factor for myocardial infarction. In addition, the SNPs rs1491978 and rs1388622 in *P2Y12*, which were previously found to be insignificant, have been identified by this new pipeline to be significantly associated with FA. Interestingly, *P2Y12* is the main receptor of ADP in platelets and a target of antiplatelet drugs prescribed to CVD patients (Offermanns, 2006). *P2Y12* has been widely studied in order to understand its associated risks and devise better treatment strategies for CVDs (Fontana et al., 2003; Offermanns, 2006; Woulfe et al., 2001), suggesting that these SNPs in this gene also have potential biological and clinical significance.

Moreover, this pipeline significantly identified non-synonymous SNP rs2071676 in *CA9*, which is associated with FA and was previously unidentified. The *CA9* product (CA IX) is one of the isoforms of the carbonic anhydrases which have been linked with several disease problems (Frost and McKenna, 2013) in addition to the platelet and CVD (Woodman et al., 2010). Moreover, several *CA9* polymorphisms have been identified to be associated with oncological problems (Chien et al., 2012; de Martino et al., 2009). Thus, it might be worth pursuing the effects of the rs2071676 SNP that may underlie *CA9* with its product and platelet functions.

The identified new key non-synonymous SNP rs12953 of *PECAM1* has been a subject of interest for clinical studies related to *PECAM1* polymorphisms, which are associated with myocardial infarction and other CVD problems (Pamuk et al., 2014; Ye et al., 2013). Hence, it might be of clinical importance to investigate whether its identified ADP platelet responses association is likely to lead to the related CVD prognosis. Thus, the reported association of the SNP with MI may further indicate the robustness of the RAPIDSNPs in identifying these key SNPs. Also, the key SNPs rs11264579 and rs822442 in the PEAR1 were identified to be significantly associated with FA in the dataset 2. Suffice to note that there are several polymorphisms in this *PEAR1*, which have been found to be associated with the increase of the

platelet aggregation and that have been reported to have an inhibitory role to the aspirin among CVD patients (Herrera-Galeano et al., 2008). Therefore, it might be clinically interesting to examine whether both identified key SNPs in this gene have a similar role, which also could be vital for targeted therapeutics. In particular, the non-synonymous (ns) SNP rs822442 might be of further interest since, in this study has been found to be significantly associated with FA. This is in contrast with previous study, which was found to be associated with the CRP platelet responses (Jones et al., 2009).

Furthermore, RAPIDSNPs identified rs12485738 in the upstream of *ARHGEF3*, which was not identified using the previous approach by Jones et. al (Jones et al., 2009). This SNP is known to be associated with the increase in the mean platelet volume (MPV), which is a feature observed in the morbidity and mortality cases among MI and cerebral infarction (ischemic stroke) patients (Meisinger et al., 2009). The fact that it has been identified by the pipeline to be associated with FA, this SNP might be a potential marker for the genomic studies that are intended to find the link between ADP responses (particularly FA) and MI. This identification strongly and further suggests the robustness of the pipeline in identifying crucial markers, which are likely to be unidentified by the standard biostatistical approaches such as forward stepwise.

The RAPIDSNPs was also able to identify a SNP rs2228671 in *LDLR* to be significantly associated with PA platelet response. This SNP has been also identified in several unrelated studies, which investigated the polymorphisms in *LDLR* and their association with CVDs (coronary heart and MI) (Franceschini et al., 2009; Myocardial Infarction Genetics Consortium et al., 2009). The fact that it has been significantly identified to be associated with PA platelet response, it might be clinically interesting to find the association or correlation between PA and MI using this SNP as a marker. Also, the pipeline identified the common SNP rs5277 in

COX-2 (*PTGS2*) to be significantly associated with PA. This gene produces the cyclooxygenase-2 enzyme, which catalyses prostaglandins that are involved with atherosclerosis (Belton et al., 2000). Type-2 diabetic individuals with the SNP rs5277 in this gene were found to be associated with the high risk of developing CVD (Rudock et al., 2009). Based on these results, there is a likely association between the PA and atherosclerosis that might be linked through this SNP. Hence, this warrants its further investigation.

In general, the above results further suggest and strengthen the belief that the devised pipeline (RAPIDSNPs) is capable of identifying crucial complex trait/disease-associated key SNPs.

2.4.4 Summary of ADP platelet responses and CVD associated SNPs

Based on the results and discussion, Table 2.26 summarises the identified ADP platelet responses SNPs and their likely CVD type association.

S/N	Identified SNPs	Gene/Locus	Associated ADP platelet responses	High response/ Increase Low response /Decrease	Likely associated CVD type based on the discussion
1	**rs6141803**	*COMMD7*	PA	low	Myocardial Infarction (MI) (Goodall *et al.*, 2010)
2	**rs1491978**	*P2Y12*	FA	high	Has a potential for atherothrombosis (Simon et al., 2009; Zee et al., 2008)
4	**rs2071676**	*CA9*	FA	low	Has a potential for hypertension (Woodman et al., 2010)
5	**rs12953**	*PECAM1*	PA	low	MI/atherosclerosis (Listì et al., 2004; Pamuk et al., 2014)
7	rs822442	*PEAR1*	FA	high	Has a potential for MI (Herrera-Galeano et al., 2008)
8	**rs12485738**	*ARHGEF3*	FA	high	MI & ischemic stroke (Meisinger et al., 2009)
9	rs2228671	*LDLR*	PA	high	Coronary artery disease (Linsel-Nitschke et al., 2008)
10	rs5277	*PTGS2*	PA	high	Atherosclerosis (Rudock et al., 2009)
11	rs1472122	*P2Y12*	PA	low	Potential for Ischemic Stroke (Zee et al., 2008)

Table 2.26. The ADP platelet responses associated SNPs that were identified by the RAPIDSNPs pipeline and have association or potential association with CVD. The bolded SNPs were previously unidentified or found to have insignificant association with the ADP platelet responses. Some of these SNPs were not directly reported to have association with CVD but other SNPs in the same genomic position or LD were found to have association. Thus, termed to be 'potential' for a particular CVD type. High or low response means the SNP is likely to increase or decrease platelet aggregation and is based on the estimated coefficient value.

2.4.5 Conclusion

The RAPIDSNPs is a developed robust computational tool for rapid discovery of key bio-markers associated with complex phenotypes. In this study the approach has been applied to reveal previously unidentified SNPs associated with ADP platelet response phenotypes, which have been independently implicated in CVDs. This strongly suggests that the approach is robust in identifying key genetic variants or SNPs that are likely to be missed by following only the standard stepwise forward or single method. Moreover, it may be generally applied in other

disease contexts for the discovery of multiple genetic variations that may better account for the heritability of continuous phenotypes. Thus, the approach has strong potential to become a useful additional tool for rapid discovery of key critical biomarkers prior to performing complex analyses in the GASs.

In the coming two chapters, the molecular mechanisms underpinning the roles of these identified variations (key SNPs) are examined through predictive approaches. In Chapter 3, the focus will be on examining the role of the identified missense (non-synonymous) SNPs that likely underlies ADP induced platelet responses variability. To the best of the researcher's knowledge, there is still a gap in the understanding of the effect of the missense SNPs to the proteins' structures and functions associated with ADP platelet response. The follow-up Chapter 4 aims to examine the likely regulatory roles of these identified SNPs that may further contribute to the variability of the ADP platelet responses and potentially CVD prognosis.

Chapter 3 – Predicting structural & functional effects of the ADP platelet responses associated with missense SNPs using structural bioinformatics approaches

3.0 Abstract

The RAPIDSNPs method developed in the previous chapter that rapidly identifies key significant SNPs is like other genetic association analysis tools. In most cases, they might be able to identify the SNPs that are susceptible to complex disease/traits. However, such analysis methods do not provide an indication of the underlying molecular mechanisms that are likely to contribute to the phenotypic variations. In this chapter, the molecular mechanisms of the previously identified missense SNPs are elucidated using predictive structural bioinformatics. Normally, the missense SNPs are likely to affect the structures of *in-vivo* related proteins and their functions, which may likely to lead to the changes in the individual phenotypic effect. Thus, the objective of this chapter is to design a structural bioinformatics approach for investigating the likely structural and functional effects (roles) of the identified missense SNPs on their related proteins. The driving hypothesis is that the identified missense SNPs from the RAPIDSNPs are likely to cause harmful mutations and change structures and/or functions of the associated proteins, which may further explain the ADP platelet responses variability. Furthermore, predicting these likely changes in structures and functions of the related proteins might indicate new avenues of investigation for targeted CVD therapeutics.

3.1 Introduction

The RAPIDSNPs pipeline, like other GAS analysis tools, are used to identify SNPs in the genome that are susceptible to the disease/trait being studied. The pipeline identified SNPs that are significantly associated with FA and PA platelet responses. However, the pipeline, on its own, is similar to many other GASs approaches, which do not allow for further elucidation of the mechanisms by which these SNPs lead to the disease/trait (Pal and Moult, 2015). There are

number of ways by which the presence of these SNPs may be linked to the gene product's functional effect and hence disease or complex trait risk. One of these ways involves missense SNPs, which cause the amino acid substitution. The missense SNPs might play a substantial role in common human disease and traits by altering *in vivo* function of the associated protein in many ways. For instance, they may change protein's fold (3-dimensional structure), ligand binding ability, or catalysis (Pal and Moult, 2015).

Thus, the identified missense SNPs from the RAPIDSNPs are further examined to investigate their structural and functional role(s), which may contribute to the differential FA and PA platelet responses among individuals. To achieve this aim, a sequence-structure-function prediction protocol is designed in order to build 3D protein models for determining potential structural or/and functional effects of the identified missense SNPs.

3.2 Why build 3D models to investigate the identified missense SNPs?

A missense SNP, which leads to an amino acid substitution in the translated protein, may cause changes in the protein's fold and hence, alter its intended function. This study is based on the prevailing theory that sequence determines structure – also known as Anfinsen's dogma - and in turn that the structure determines function (Anfinsen et al., 1961). Understanding structural and functional impacts of the identified missense SNPs is vital as they have been widely reported to play key roles in the several diseases or traits including CVD (Ohnishi et al., 2000; Okuda et al., 2002; Porto et al., 2015; Shi et al., 2012).

Thus, this study aims to investigate the roles of identified missense SNPs and their likely effect on the related proteins structures that may underlie ADP platelet responses and potentially contribute to the CVD.

In undertaking this objective, public databases such as the Research Collaboratory for Structural Bioinformatics/Protein Data Bank (RCSB/PDB) (Kouranov et al., 2006; Rose et al., 2013) and Protein Data Bank in Europe (PDBe) (Gutmanas et al., 2014) are essential for the structure prediction approaches. There has been a yearly exponential growth of experimentally solved structures particularly in RCSB/PDB (Figure 3.1), which serve as reliable templates for predicting yet to be solved structures of known sequences.

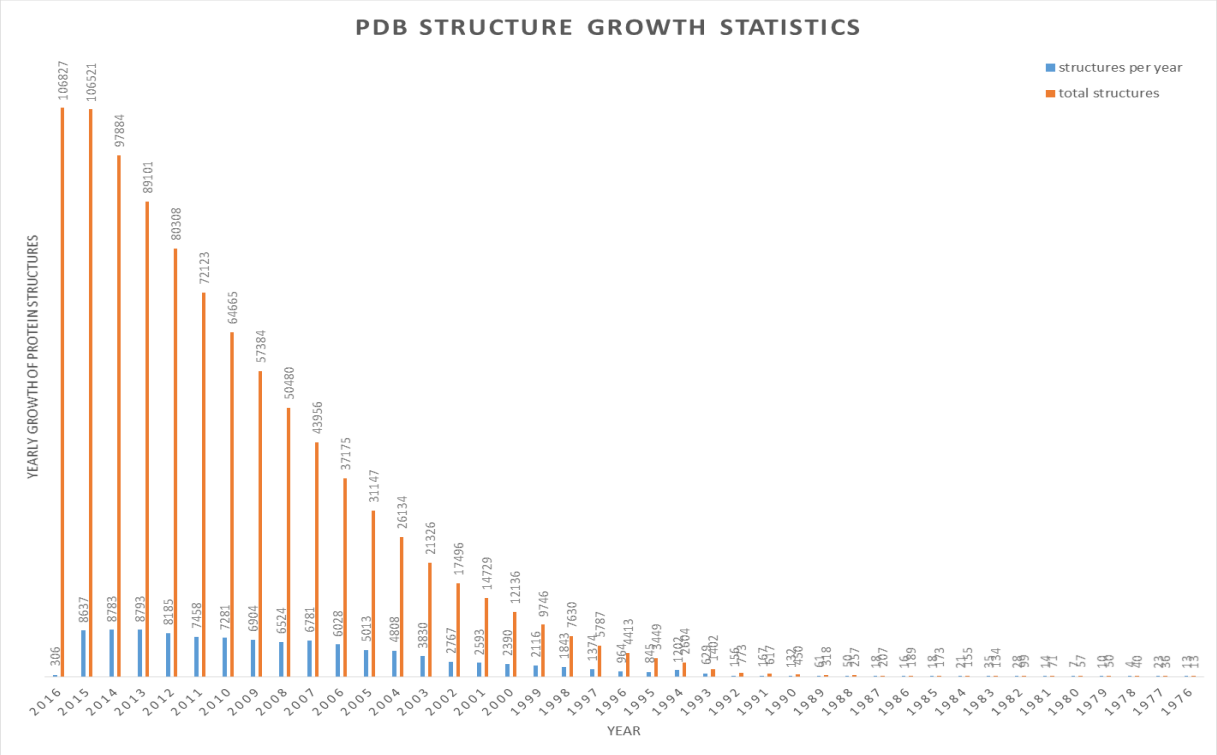


Figure 3.1. The exponential growth of the number of structure per year in the RCSB. The public availability of these structures may play substantial role in performing the predictive task for investigating the potential or likely structural change due to missense SNPs on the proteins associated with ADP platelet responses. (The Figure was adapted from *(Agüero et al., 2007)*.

While this growth in known structures is impressive, it is far outstripped by the growth in sequence databases, Figure 3.2

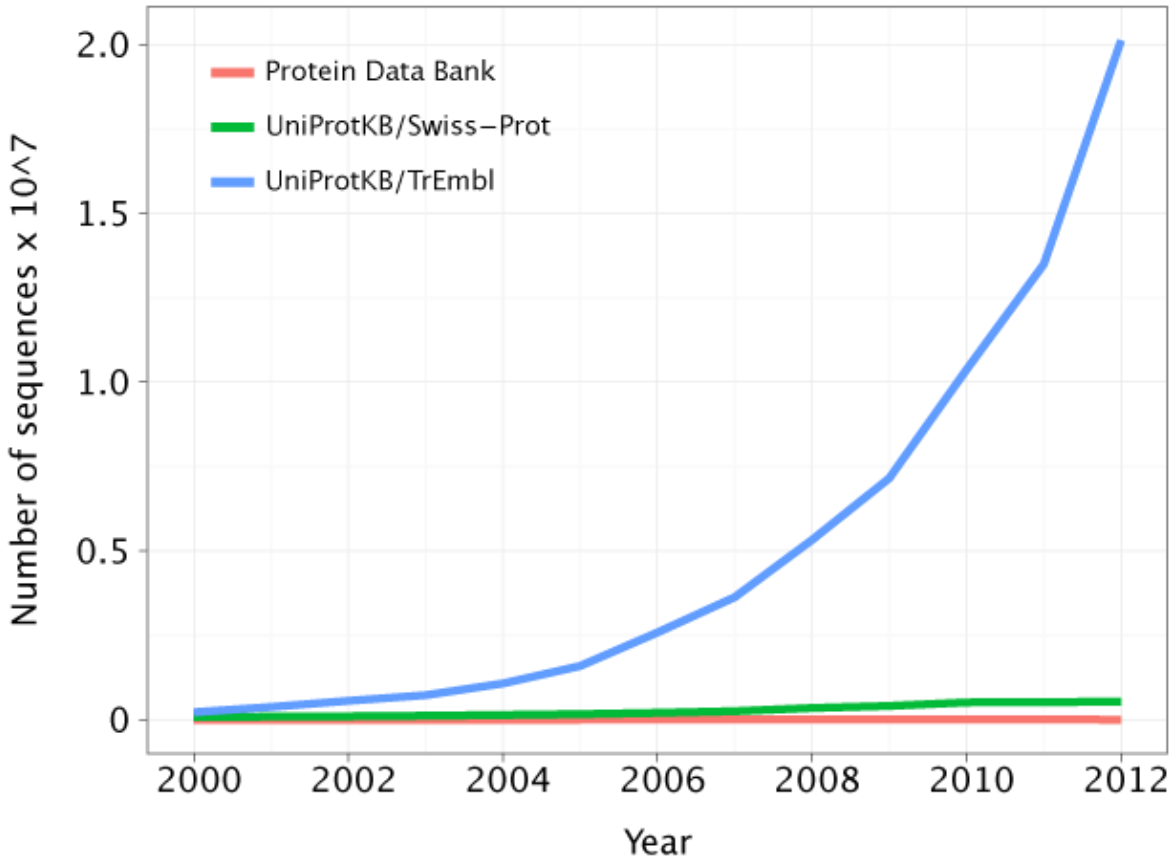


Figure 3.2 The growth of sequences in different databases which outstrips the number of experimental structures deposited in PDB (Figure 3.1). (Image was taken from http://gorbi.irb.hr/en/method/growth-of-sequence-databases/).

Thus, structure prediction methods are essential for bridging the sequence-structure gap and therefore supplementing our knowledge of protein functions.

3.3 What are the identified protein mutations?

Table 3.1 shows a summary of the overall identified missense SNPs from the RAPIDSNPS method, which are significantly associated with ADP platelet responses. The detail on their selection basis can be found in the Chapter 2.

SNP ID	SNP's selected flanking DNA sequence	mRNA accession number	Allele change and position in the mRNA	Residue change and protein position	Protein name and accession numbers (NCBI and UniProt)
rs2071676	ACTGCTGCTGTCACTGCTGCTTCTG**[A/G]**TGCCTGTCCATCCCCAGAGGTTGCC	NM_001216.2	**G**TG ⇒ **A**TG (139)	Val33Met	carbonic anhydrase 9 [precursor] NP_001207.2 & Q16790
rs822442	GCCCGCTCTTTGCCAGCCTGCAGAA[**A/C/T**]CCTGAGCGGCCAGGTGGGGCCCAAG	NM_001080471.1	AA**C** ⇒ AA**A** (2660)	Asn848Lys	platelet endothelial aggregation receptor 1 [precursor] NP_001073940.1 & Q5VY43
rs12953	GCATTTTGGACCAAGCAGAAGGCTA**[A/G/T]**CAAGGAACAGGAGGGAGAGTATTAC	NM_000442.4	A**G**C ⇒ A**A**C	Ser563Asn	platelet endothelial cell adhesion molecule1 [precursor] NP_000433.4 & P16284
rs3739038	TCTTCTCTGCTTTCTGGATTGAAGT**[C/G]**CCCTGGCTCTCTTCCTGGTGCCGAC	NM_001137553.1	**C**AC ⇒ **G**AC	His727Asp	leucine-rich repeat flightless-interacting protein 1 isoform 5. NP_001131025.1 & Q32MZ4
rs3736101	GAAGTATGGATTGTCATAGATTCGC**[C/T]**GGCGCACTGACCCCTCTCCAATTTC *(C/T REV)*	NM_130470.2	C**G**G ⇒ C**A**G	Arg765Gln	MAP kinase-activating death domain protein isoform a. NP_569826.2 & Q8WXG6

Table 3.1 The identified missense SNPs from the RAPIDSNPs that are associated with both FA and PA platelet responses. The bolded nucleotide letters are SNPs' allele changes in the second and fourth columns respectively. The corresponding mutation (wildtype-mutant residues and its position) is in the fifth column, and the involved wildtype protein name is in the last column.

The above SNPs in the Table 3.1 are examined initially for their likely deleteriousness levels in the associated proteins before performing the structure prediction approaches.

3.4 Structure prediction approaches

Tertiary structural prediction approaches aim to produce accurate three-dimensional (3D) models of a protein from its amino acid sequence. Determining protein's 3D structure is vital for understanding its underlying molecular functions, which govern physiological processes. Moreover, from the theory, the protein's function is derived from its 3D structure, which is determined from the amino acid sequence - Anfinsen's dogma. Thus, changes in sequence may likely to cause changes in structure and maybe function. However, the experimental procedures for determining the proteins 3D structures are expensive, laborious, and time consuming. Therefore, predicting 3D structure is indispensable for understanding any underlying functional changes, which are caused by the changes in sequence. In turn, that knowledge may have a potential impact on biomedical studies, including those involving personalised medicine.

In this regard, different approaches have been developed for the purpose of predicting high accuracy 3D structures, which may be used for biomedical applications. These approaches are described beneath.

3.4.1 Homology modelling approaches

Template based homology/comparative modelling (TBM) is a protein structure prediction approach, which is used to recognise the protein fold of a given sequence based on the evolutionary information of the experimentally determined structures available at the RCSB/PDB. This is likely to be achieved when there is a good similarity between sequence of unknown structure "*target*" and the experimentally determined structures from the PDB "*template*" (Krieger et al., 2003). Two important theories underpinning the homology modelling are: 1) the information necessary for a particular protein to uniquely fold into three dimension structure is deciphered in its amino acid sequence (Anfinsen et al., 1961), 2) Throughout evolution, structures are more highly preserved than the sequences, which

practically means, two proteins with very similar sequences are likely to have near identical structures, and even low homology sequences (~25-30% identity) will still fold into similar structures (Chothia and Lesk, 1986).

Chothia and Lesk's theory was further demonstrated by (Rost, 1999) who objectively showed that the theory holds only when the percentage of the residues are within the 'safe' zone, i.e. two protein sequences are likely to have the same fold if their percentage identity is above 25%, depending on the length of their sequences (Godzik, 2003). This might be achieved through the alignment and alignment correction steps of TBM approaches (Krieger et al., 2003). Obtaining the optimal alignment is crucial for achieving the desired accuracy of homology models and providing a solid groundwork for subsequent functional analyses, including investigating protein-ligand interactions in the missense SNPs mutation studies (Torrent et al., 2004; Wilson et al., 2000). TBM methods are widely applied in modelling of proteins in different diseases studies and, in particular, for modelling of missense mutations (Lino Cardenas et al., 2011; Moghaddasian et al., 2014). Thus, majority of the incorporated methods in the designed bioinformatics pipeline for this study are under this category.

3.4.2 Fold recognition and threading approaches

Homology modelling is a better approach for 3D modelling for the protein structure when there is a good homology or significant identity between the target and template sequences. But, when there is very low percentage identity, particularly if the percentage identity is beyond the twilight zone (<25-30%), then fold recognition methods are essential for identifying the correct fold of the target sequence (Godzik, 2003; McGuffin, 2008b). The terms 'fold recognition' and 'threading' are used interchangeably, although the latter term has started to fade in use. They literally mean finding the 3D structure that best fits the sequence as opposed to homology modelling, which aims to find the sequence that best fits the 3D structure (McGuffin, 2008b).

The underlying theory behind this approach is that the structure will often remain more highly conserved than the sequence and many proteins with very unrelated sequences tend to have similar folds, which implies that the "fold space" for the most proteins is limited (Orengo et al., 1994; Sander and Schneider, 1991; Zhang and Skolnick, 2005a). The original threading method (Jones et al., 1992) worked by trying to fit (or 'thread') the target sequence to the backbone of the fold of each template and then use knowledge-based statistical energy and solvent potentials to evaluate and obtain the desired fold with minimum free energy. The traditional threading approach contained several drawbacks: i) they were CPU intensive due to the applied double dynamic programming algorithm, ii) they were not reliable for producing accurate coordinates for 3D models due to gaps & loop modelling requirements, iii) they were hard to automate due to the generation of many alternative energy potentials rather than a single score. These drawbacks led to the development of the hybrid and fully automated methods for fold recognition which are now routinely used (McGuffin, 2008b). The hybrid fully automated approach has been elsewhere successfully used to model the effect of missense SNPs (Monteagudo et al., 2015; Shukla and Mishra, 2011). Thus, we have also incorporated this approach in the designed bioinformatics pipeline for investigating the role of the identified missense SNPs that likely underpin the ADP responses variability.

3.4.3 *Ab initio* or free modelling (FM) approaches

Ab initio or template free modelling (FM) approaches seek to model the protein's 3-dimensional structure solely from its amino sequence without the prior knowledge of any other similar sequences or structures. The fundamental assumption is that the protein's true native 3-dimensional structure (protein's active form) becomes into its folded state under the lowest free energy. The prediction methods require a free energy function that is close to this native state and an ability to search the vast conformational space (due to the combinatorial number of

residues) for the lowest energy conformation (Chivian et al., 2003). The underlying theory is based on the (Anfinsen et al., 1961) who showed that the information necessary for a protein to fold into its native 3-dimensional structure is entirely encrypted in its amino acid sequence. The *ab initio* modelling approaches are only usually applied when the underlying assumptions for the sequence-based searches or structure-based fold recognition methods fail, i.e., the protein's fold has to be completely modelled in the absence of any homologous/analogous structural information. This is because there are still many protein targets, which have completely novel sequences and folds. In addition, some proteins might have some sequence similarities but fold differently (Helles, 2008).

There are current successful methods that have shown very promising results in achieving near crystal structure folds, but these are highly computationally intensive and only perform well in predicting single domain proteins with a small number of residues (<100) (Helles, 2008; Kryshtafovych et al., 2010). Nevertheless, these methods may be used for modelling missense SNPs, when the mutation is located in a specific protein region of interest (for instance, a short binding domain) (Taylor et al., 2003). Thus, some of the methods, which are in the designed bioinformatics pipeline for this chapter, inherently employ, or are entirely based on this structure prediction category.

3.4.4 The CASP experiment

The critical assessment of protein structure prediction (CASP) is the pre-eminent biennial experiment for evaluating the performance of numerous protein structure prediction methods (Kryshtafovych et al., 2010; Moult et al., 1995). The CASP experiment is a blind competition in which different independent research groups around the world attempt to make accurate predictions of the 3D-structures of proteins prior to the release of their experimentally determined atom coordinates. The only data available to the predictors are the amino acid

sequences of the proteins or ‘targets’. Researchers use various computational approaches, as described above, in order to predict the structures of the targets.

Furthermore, different categories of predictions have been gradually added since the start of the challenge. The more interesting categories for biomedical applications are TBM and model quality assessment (MQA). TBM is by far the most successful of the tertiary structure prediction approaches, producing models that are often close to near native structures. Additionally, MQA is important for evaluating the predicted models and assessing how close they are to the experimentally determined structures (Kryshtafovych et al., 2010). The predictors and assessors gather in biennial conferences for intense discussions of the top performing methods in each of these categories.

3.4.4.1 ***Why CASP experiment is important***

The CASP experiment has been a major driving force in the field of structure prediction. Despite major advances there are still bottlenecks, and challenges in the underlying prediction process (Bourne, 2003; Kryshtafovych et al., 2010; Moult, 2005). Achieving near native structure models is vital for the advancement of biological research, particularly biomedicine, where molecular aspects of the diseases are now closely investigated for the personalised medicine (Collins and Varmus, 2015; de Bono and Ashworth, 2010). In our case, for investigating the role of missense SNPs that may underlie ADP platelet responses, we use the CASP recommended top performing methods, in both structural prediction and model quality assessment.

The next section describes methods for modelling the effects of identified missense SNPs.

3.5 Methods

3.5.1 The procedure used to identify the deleteriousness of missense SNPs (damaging mutations)

The identified significant SNPs in the coding region (missense/ns/cSNPs) (Table 3.1) were initially analysed to determine their potential damaging level to the *in vivo* proteins using different computational methods. The missense SNPs were judged to be deleterious if they were identified by the selected sequence profile and evolutionary conservation, and machine learning methods, which are SIFT (Ng and Henikoff, 2003), Polyphen-2 (Adzhubei et al., 2010), and Fathmm (Shihab et al., 2013)), and (SuSpect) (Yates et al., 2014) respectively. The selection of these methods was based on their performance assessment reported by Gnad et al. (Gnad et al., 2013). The flowchart in the Figure 3.3 shows the general approach, which was used to determine the deleteriousness of the identified missense SNPs.

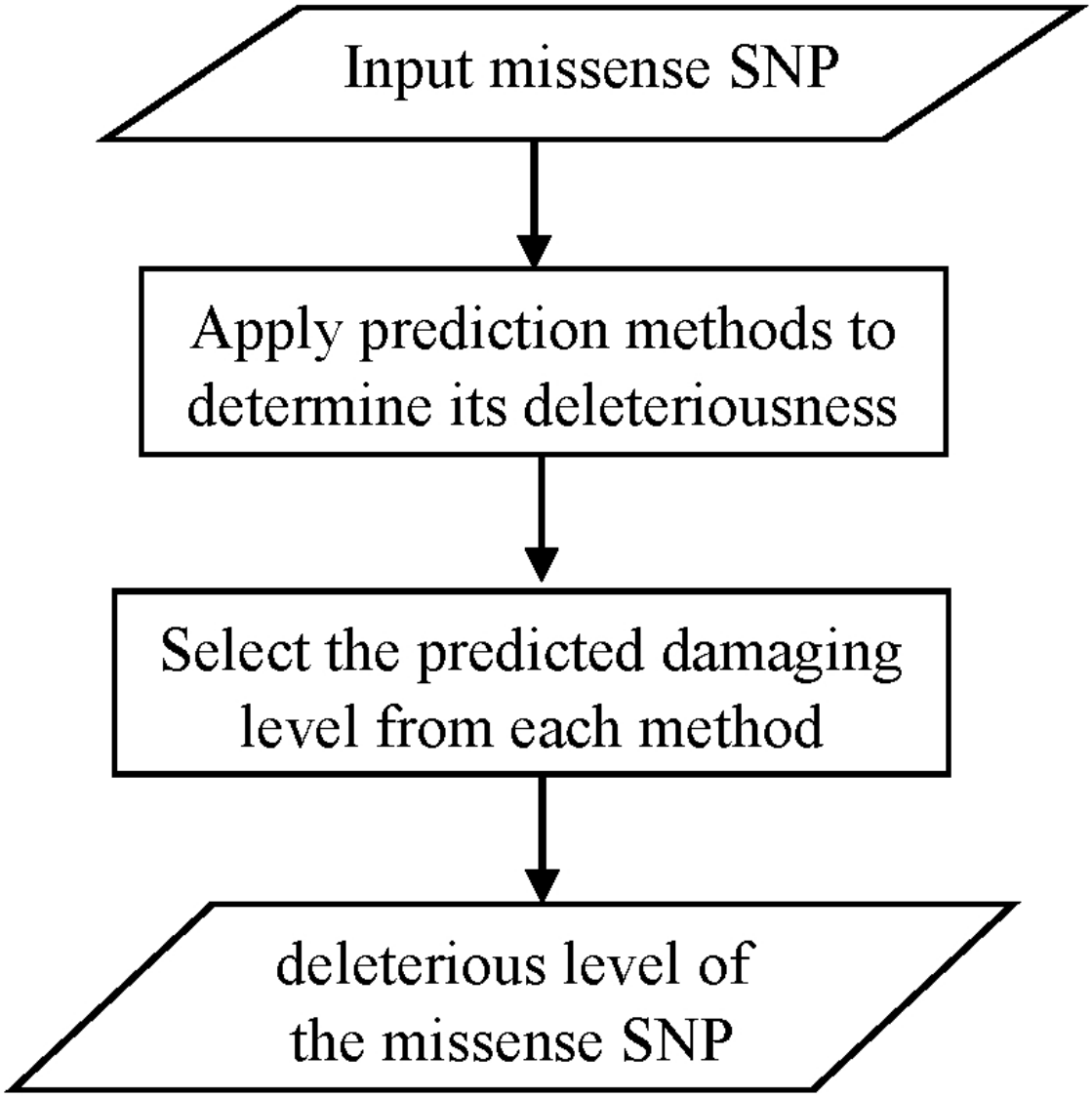


Figure 3.3 The general approach used to identify the deleteriousness of the nsSNPs identified by the RAPIDSNPs. The damaging level of the SNPs were judged based on the agreement of the results from varying prediction methods.

In using Fathmm, two algorithms (weighted and unweighted) were alternatively applied with human phenotype ontology being generic phenotype association.

3.5.2 Predictive modelling of proteins related to the identified deleterious missense SNPs

The structural bioinformatics pipeline that incorporates different methods (TBM and FM based) was developed in order to model the effect of the identified damaging missense SNPs

on the related proteins. The initial structural modelling approach involved TBM methods to generate the initial 3D models for full sequence protein. Subsequently, modelling of short sequences, particularly in the disordered regions and/or loops, was followed, if the full sequence produced low accuracy protein models.

In applying TBM, template recognition and initial alignments were carried out using Clustal omega (Sievers et al., 2011), HHalign (Söding, 2005) and PSI-BLAST (Altschul et al., 1997), which allowed us to examine the homologous sequence similarity between targets (both wildtype and mutant proteins) and templates. However, in most cases TBM was carried out in a fully automated mode. The methods deployed include IntFOLD3-TS (McGuffin et al., 2015), HHpred (Söding, 2005), GenThreader (McGuffin and Jones, 2003), SwissModel (Biasini et al., 2014), I-TASSER (Zhang, 2008), RaptorX (Källberg et al., 2012), and SparksX (Yang et al., 2011). The IntFOLD3 server was also used to examine per residue accuracy, which helps to identify the well-folded and badly predicted regions. PSIPRED (McGuffin *et al.*, 2000) was used to provide the secondary structure comparisons.

The selection of the above methods was also guided by the CAMEO project data (Haas et al., 2013), which provides continuous performance evaluation of public methods, supplementary to the top performers identified in the CASP experiment. Moreover, using a combination of different methods from TBM and FM based approaches, enables to observe the variations likely model quality, as a result of underlying parameter settings and/or varying target-template alignments (Kryshtafovych and Fidelis, 2009).

The workflow diagram showing the general approach used to model these identified mutations is shown in the Figure 3.4

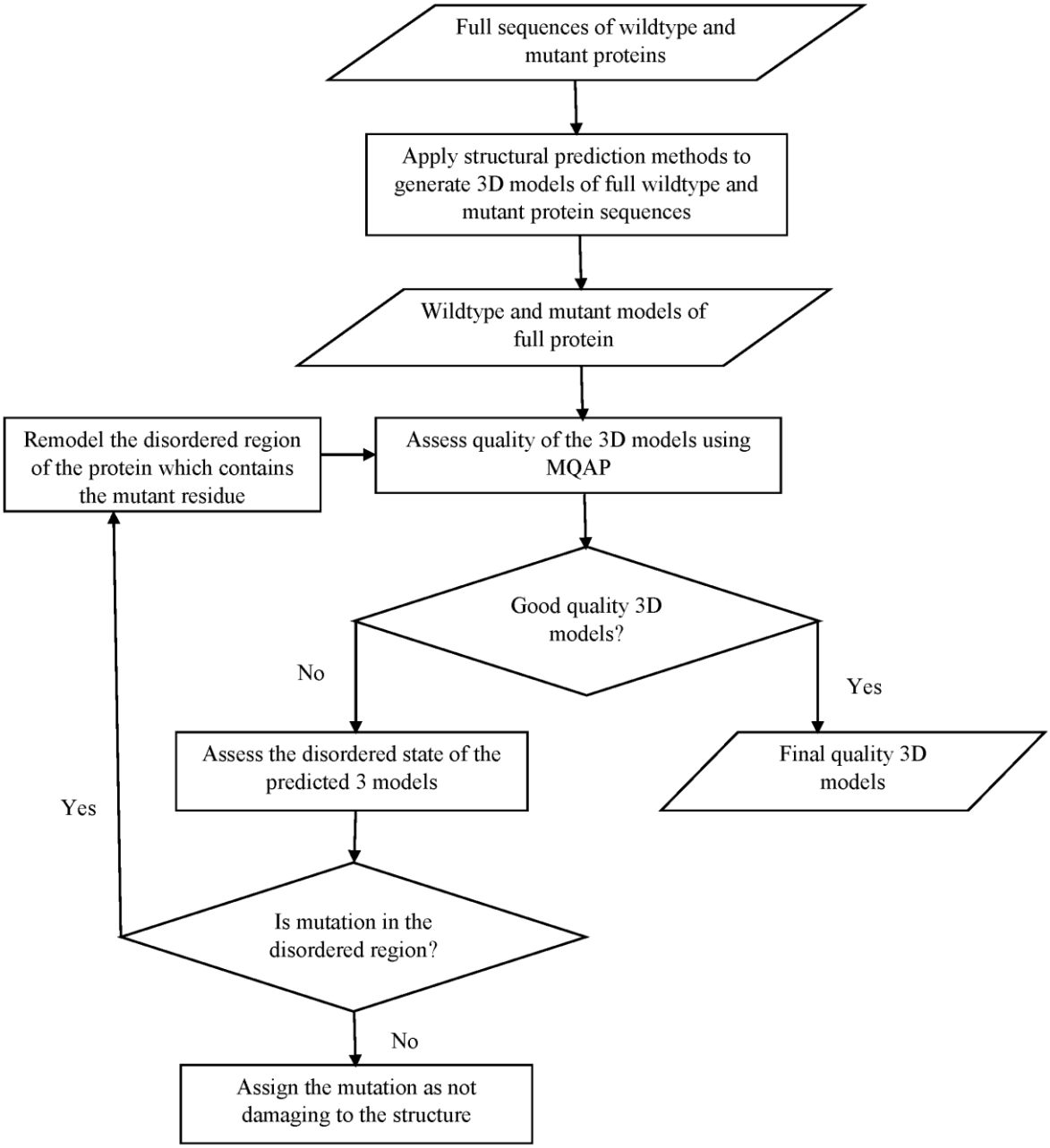


Figure 3.4. The general workflow showing the overall methods used for generating the 3D structural protein models to analyse the effect of the identified missense SNPs. The applied structural prediction methods are among the top performing in the previous 2014 and recent 2016 CASP challenges, and according to CAMEO.org. The models are then assessed their quality using model quality assessment programs (MQAP) before selecting the final model(s).

For further investigating mutations in the termini regions or short or unstructured sequences, FM and disorder prediction methods were used. For FM, Quark (Xu and Zhang, 2012) and Robetta (Kim et al., 2004) were used. The disordered regions were predicted using DISOPRED (Ward et al., 2004) and DISOclust (McGuffin, 2008a).

3.5.3 Model quality assessment (MQA)

The top performing Model Quality Assessment Programs (MQAPs) were applied to Estimate the likely Model Accuracy (EMA) of each of the generated models. In particular, the ModFOLD4 (McGuffin et al., 2013) was applied in addition to ModFOLD6 (Maghrabi and McGuffin, 2017), which were among the top performers in the latest CASP challenges (CASP10-12). In addition, ModFOLD4 and ModFOLD6 are continuously validated as top performing methods according to the CAMEO project.

Furthermore, in evaluating the quality of the models, the TM score was initially used to align the models to each other, and the comparison matrices containing TM scores were then produced to examine the closely correlated models, i.e. most similar in structure. The closely correlated models were then compared and run through ModFOLD4/6 to select the final models for downstream analyses.

3.5.4 Functional prediction of the predicted models

FunFOLD2 server (Roche et al., 2013) was used for functional prediction to identify any putative ligand binding sites in the selected models.

3.5.5 Further models analyses

Further molecular analyses, visualisation and rendering of the individual predicted models was performed using PyMOL (DeLano, 2002).

3.5.6 Overall prediction protocol

The sequences for the wildtype proteins were obtained from the NCBI based on their accession numbers, CA IX = NP_001207.2, PECAM1 = NP_000433.4, and PEAR1 = NP_001073940.1

with their full length containing 459, 738 and 1037 amino acids respectively. These proteins were selected since their related missense SNPs are the most damaging among the identified missense SNPs, refer to the "Results" section on deleteriousness prediction (section 3.6). Then TBM models for each full-length individual protein, for both wildtype and mutant, were generated using the pipeline described above. TM-align method (Zhang and Skolnick, 2005b) was later used to superpose the models and template, providing structural alignments.

For the CA IX protein, DISOPRED and DISOclust were further used to examine the PG domain in the N terminus region, which proved to be difficult to fold up and was where the V33M mutation located. Hence, part of this domain region, consisting of residues numbered from 1-50, was later re-modelled using the above pipeline (figure 3.4) with both TBM and FM methods applied. The sequence of this region was further analysed using the alignment programs, as previously described (section 3.5.2), in order to examine the conserved regions in which the mutation appears to be localised. Moreover, the predicted topology within part of 1-50 residues shows similarity across methods (see Results). Following this similarity, the models were aligned or superposed to each other using TM-score program (Zhang and Skolnick, 2004). The aim was to compare and determine the closely correlated models, which have the closest fold similarities for further quality assessment and analyses. The comparison matrices containing pairwise TM-scores were generated for each combination.

Furthermore, based on the CA IX annotation, in the Uniprot (Consortium, 2008) (accession ID = Q16790) within PTM/Processing, the mutation (V33M) appeared to be predicted as part of signal peptide (1 – 37 residues). Thus, SignalP (Petersen et al., 2011) with signalP HMM (Bendtsen et al., 2004), was then applied to predict the selected residues from 1-26 positions. This region was selected as initial signal peptide prediction using the same methods in the 1 – 37 residues appeared to have low score and in addition to indicating that the cleavage site to be

likely within residues 1 – 26. Additionally, it was in our interest to examine whether the highly conserved region within residues 22 – 45, where the V33M mutation was located, was part of the signal peptide or in the transmembrane region within the PG domain. Therefore, in this case, MEMSAT SVM (Nugent and Jones, 2009) was also applied.

For PECAM1, the region from 497 – 596 of the protein was additionally modelled as the full-length protein using the above pipeline failed to produce high quality models (see the Results section). The selection of this region is based on the DomFOLD predictions (McGuffin and Roche, 2011) (see the Results section). The mutation S563N was located within this predicted domain. The mutant and wildtype proteins of this domain were subsequently modelled using the TBM and FM methods in the pipeline previously described. Furthermore, the TM-score comparisons were applied for further quality assessment, as the models were similar in topology. The obtained high quality wildtype and mutant models were further analysed by comparing their structures with templates found using PSI-BLAST. The sequence (PECAM1 domain 6 from Uniprot – Family & Domains (499 – 591 residues) was run against the PDB structures, and two iterations of PSI-BLAST were used. The aim was to examine the likely structural and functional effects when the wildtype residue Ser at 563 is changed to Asn as the predicted domain type for wildtype in Uniprot is of C2 Ig domain, while the mutant appeared to adopt V-type Ig domain (See the Results section).

For PEAR1, the modelling process using the above pipeline was repeated three times with different subsequences. The full sequence protein was computationally hard to predict, as it contains 1037 amino acids. Only two methods IntFOLD3 and SPARKS-X managed to produce models but with poor quality. As the mutation of interest is in the positon N848K, the protein was repeatedly re-modelled focusing on the 801 – 900 and 801 – 850 residues sub regions. The selection of these regions was based on the disordered predictions and secondary structure

prediction using PSIPRED, which was recalled to check the likely secondary structure of the wildtype/mutant residues.

3.6 Results

3.6.1 Predicted deleteriousness missense SNPs

Firstly, the rs2071676 SNP is predicted to be the most deleterious comparing to other SNPs. The prediction of damage level of the missense SNPs associated with the ADP associated proteins are summarised in Table 3.2 below:

SNP ID	Protein	Method	Score	Overall score
rs2071676	Carbonic anhydrase IX (CA IX)	Polyphen-2	0.718	Damaging
		SuSpect	44	Near damaging
		fathmm	-4.54	Damaging
rs12953	Platelet endothelial cell adhesion molecule (PECAM1)	Polyphen-2	0.995	Damaging
		SuSpect	20	Neutral
		fathmm	-5.09	Damaging
rs822442	Platelet endothelial aggregation receptor 1 (PEAR1)	Polyphen-2	0.057	Neutral
		SuSpect	22	Neutral
		fathmm	-2.46	Damaging
rs3739038	Leucine-rich repeat flightless-interacting protein 1 isoform 5 (LRRFIP1)	Polyphen-2	0.000	Neutral
		SuSpect	14	Neutral
		fathmm	2.13	Neutral
rs3736101	MAP kinase-activating death domain protein isoform a (MADD)	Polyphen-2	0.000	Neutral
		SuSpect	7	Neutral
		fathmm	-0.83	Neutral

Table 3.2 Predictions of deleteriousness of the identified missense SNPs associated with ADP platelet responses. Polyphen-2 uses the value between 0 and 1 with the highest score for deleteriousness of the SNPs being 1. For SuSpect, the damage score is between 1 and 100. The cut-off for "damaging" is 50, a score between 1 and 50 means that it is more tolerated, and the values above 50 represent deleteriousness of the SNPs. Fathmm's deleterious score ranges between -ve and +ve values with the larger –ve values implying higher likelihood of deleteriousness.

The damage level of the identified missense SNPs can be further visualised using Figure 3.5, which is based on Polyphen-2:

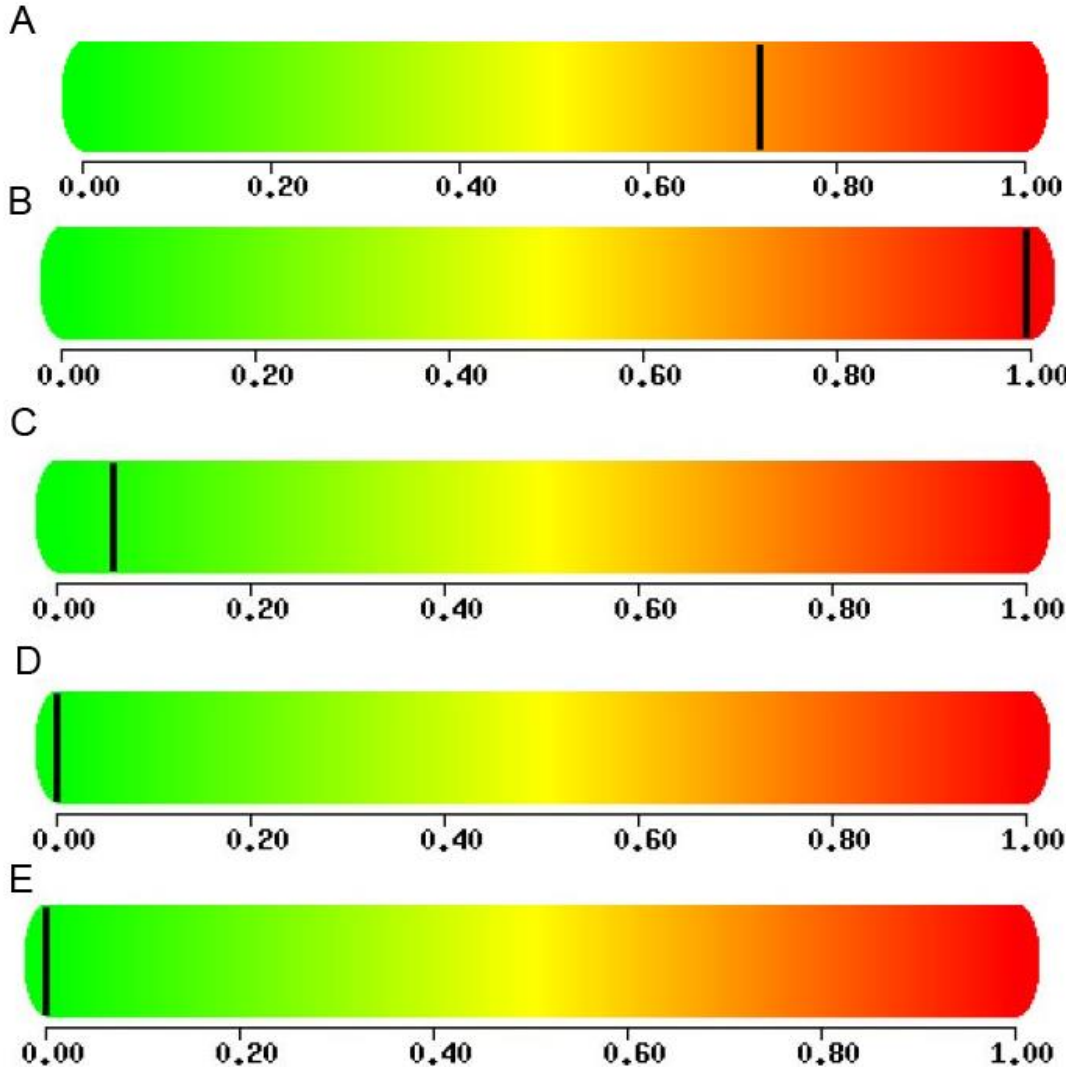


Figure 3.5 The predicted damage level of all identified missense SNPs based on Polyphen-2. (A) rs2071676 of CA9 (B) rs12953 of PECAM1, (c) rs822442 of PEAR1 (d) rs3739038 of LRRFIP1 and (e) rs3736101 of MADD. The green, yellow, orange, and red colour legends indicate the damage levels starting from: neutral, maybe damaging, slightly damaging and damaging respectively. The black vertical bar represents the score for the SNP.

From the results in the Tables 3.2 and Figure 3.5, it is clear that three missense SNPs (mutations) rs2071676 (V33M), rs12953 (S563N), and rs822442 (N848K) in the *CA9* (CA IX), *PECAM1* (PECAM1), and *PEAR1* (PEAR1) genes (proteins) respectively are more likely to be damaging to their respective proteins structures/functions.

3.6.2 3D models and function predictions characterising the CA IX The V33M mutation

3.6.2.1 ***Predicted structures of the full-length wildtype and mutant CA IX***

Several 3D models of the full-length CA IX were produced, but the models are predicted to be of poor quality. Based on the UniProt accession annotation family & domain database reports, the protein contains three main sequence domains, which are the N-terminus proteoglycan (PG) which includes signal peptide and transmembrane region, CA catalytic, and C-terminus intracellular (IC) cytoplasmic domains. The only domain that appears to be predicted well by a majority of the models is the catalytic (CA) domain. In addition, this domain is the only region of the protein, which has been experimentally solved; the PDB entry 3iai (Alterio et al., 2009) containing 257 residues, which was selected as a homologous template by nearly all of the TBM methods. The models align well when superposed with CA domain crystal structure using TM-align. Table 3.3 and 3.4 show different models scores (both wildtype and mutant) when the template-models are superposed using TM-align. As the template used by the methods is similar, these tables aim to examine the models with relative higher scores.

		TM-Align	**ModFOLD4 global score**	**Sequence length coverage**
METHOD	**TEMPLATE (PDB ID)**	**RMSD (TM-Score)**		
IntFOLD-TS	3iaiA	**0.16 (0.55967)**	**0.643**	**257**
RAPTORX	3iaiA	**0.07 (0.55981)**	**0.531**	**257**
SPARKS-X	3iaiA	0.28 (0.55920)	0.527	257
Bioserf/GenThreader	3iaiA	1.54 (0.54215)	0.538	255
SwissModel	3iaiA	0.06 (0.99991)	0.513	251
I-TASSER	3iaiA	0.222 (0.55673)	0.536	255
HHpred	3iaiA	**0.16 (0.55968)**	0.533	253

Table 3.3 TM-align scores of different methods when the wildtype model was aligned with the template. The gray highlighted score from the Swissmodel truncated both termini of the full length CA IX.

		TM-Align	ModFOLD4 global score	Sequence length Coverage
METHOD	**TEMPLATE (PDB ID)**	**RMSD (TM-Score)**		
IntFOLD-TS	3iaiA	**0.12 (0.55977)**	**0.667**	**257**
RAPTORX	3iaiA	**0.08 (0.55986)**	**0.5297**	**257**
SPARKS-X	3iaiA	0.27 (0.55928)	0.5270	257
GenThreader	3iaiA	0.94 (0.55001)	0.5312	255
SwissModel	3iaiA	0.06 (0.99989)	0.5242	255
I-TASSER	3iaiA	0.30 (0.55908)	0.5292	257
HHpred	5fl4	0.15 (0.54664)	0.633	257

Table 3.4 TM-align scores of different methods when the mutant model was aligned with the template (s).

The predicted structures vary in the termini regions of CA IX, as no good template was obtained to fit the models well in these regions, particularly the N-terminus, which contains the PG domain and where the mutation is located. Thus, most generated models are of low estimated quality. A few are of good quality, but this is only achieved by truncating the termini and then comparing the trimmed models with the well-aligned CA domain. This is illustrated in Figures 3.6 A – B, 3.7 A – B, 3.8 A – B, and 3.9 A – B, which show the models from IntFOLD-TS and RaptorX and their predicted per-residue error plots according to ModFOLD4. The RaptorX method truncates the entire N-terminus region of the mutant model, while the SwissModel method truncates both termini of the proteins.

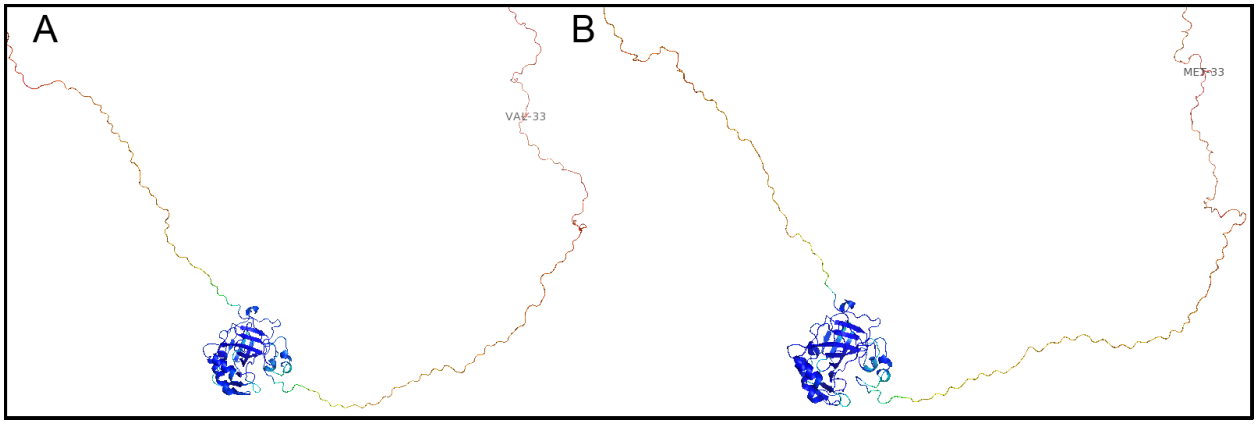


Figure 3.6 An IntFOLD-TS model (A) wildtype and (B) mutant showing the hard-to-model N and C termini (red) and the catalytic domain (blue), which was well modelled. The mutation of interest V33M is in the N-terminus labelled Val-33 for wildtype residue and Met-33 for mutant residue.

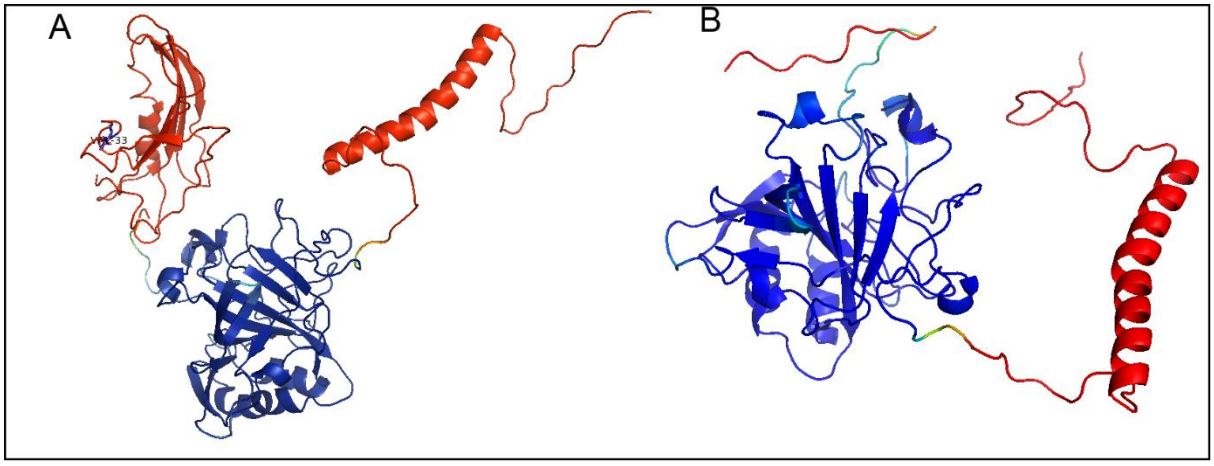


Figure 3.7. The RaptorX full-length CA IX models (A) wildtype and (B) mutant. Again the catalytic domain is well-modelled in both cases (blue region). However, the mutant model failed to fold up the N-terminus, which is where the mutation occurs.

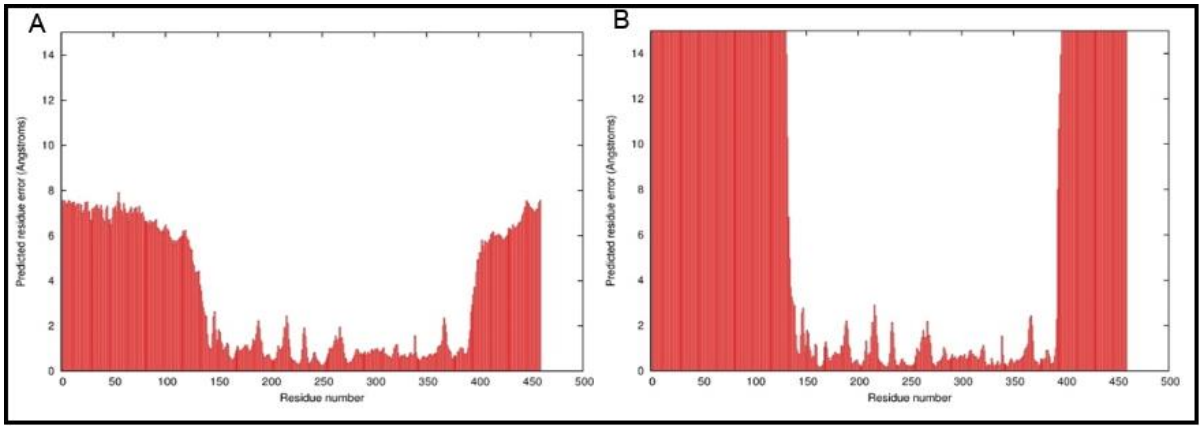


Figure 3.8. The per residue accuracy based on the ModFOLD4 for the full length CA IX wildtype between A) IntFOLD and B) RaptorX models. It can be clearly seen that the core of the protein has a good prediction accuracy for most of the predicted residues. However, the N-terminus has been poorly predicted with very poor accuracy in the RaptorX model.

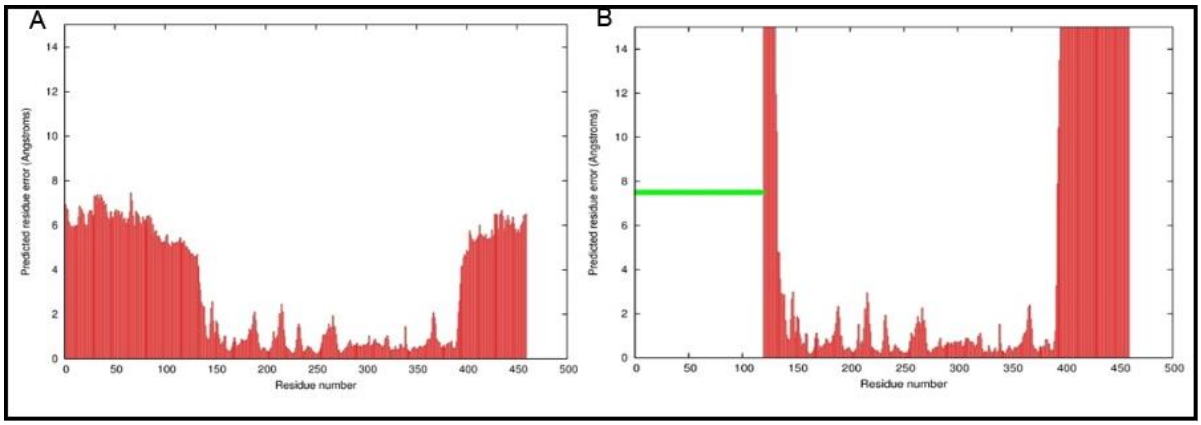


Figure 3.9. The per residue accuracy based on the ModFOLD4 for the full length mutant CA IX models, A) IntFOLD and B) RaptorX. The latter failed to fold in the N-terminus where the mutation of our interest is located. The green crosses indicate the truncated/missing residues.

3.6.2.1.1 Multiple sequence alignment (MSA) for CA IX wildtype PG domain

Based on the full sequence alignment of the protein, only the CA domain appears to align well with the crystal structure (3iaiA), which reflects the results of the models in Figures 3.6 A – B, 3.7 A – B, 3.8 A – B, and 3.9 A – B. Moreover, the entirety of the N-terminus PG domain (1-134 residues), where the V33M mutation occurs, is unaligned (Figure 3.10 and Figure 3.11).

```
CA_IX_Homo_sapiens          MAPLCPSPWLPLLIPAPAPGLTVQLLLSLLLLVPVHPQRLPRMQEDSPLGGGSSGEDDPL   60
3IAI:A|PDBID|CHAIN|SEQUENCE ------------------------------------------------------------   0

CA_IX_Homo_sapiens          GEEDLPSEEDSPREEDPPGEEDLPGEEDLPGEEDLPEVKPKSEEEGSLKLEDLPTVEAPG   120
3IAI:A|PDBID|CHAIN|SEQUENCE ------------------------------------------------------------   0

CA_IX_Homo_sapiens          DPQEPQNNAHRDKEGDDQSHWRYGGDPPWPRVSPACAGRFQSPVDIRPQLAAFCPALRPL   180
3IAI:A|PDBID|CHAIN|SEQUENCE --------------GPDQSHWRYGGDPPWPRVSPACAGRFQSPVDIRPQLAAFSPALRPL   46
                            * ************************************.******

CA_IX_Homo_sapiens          ELLGFQLPPLPELRLRNNGHSVQLTLPPGLEMALGPGREYRALQLHLHWGAAGRPGSEHT   240
3IAI:A|PDBID|CHAIN|SEQUENCE ELLGFQLPPLPELRLRNNGHSVQLTLPPGLEMALGPGREYRALQLHLHWGAAGRPGSEHT   106
                            ************************************************************

CA_IX_Homo_sapiens          VEGHRFPAEIHVVHLSTAFARVDEALGRPGGLAVLAAFLEEGPEENSAYEQLLSRLEEIA   300
3IAI:A|PDBID|CHAIN|SEQUENCE VEGHRFPAEIHVVHLSTAFARVDEALGRPGGLAVLAAFLEEGPEENSAYEQLLSRLEEIA   166
                            ************************************************************

CA_IX_Homo_sapiens          EEGSETQVPGLDISALLPSDFSRYFQYEGSLTTPPCAQGVIWTVFNQTVMLSAKQLHTLS   360
3IAI:A|PDBID|CHAIN|SEQUENCE EEGSETQVPGLDISALLPSDFSRYFQYEGSLTTPPCAQGVIWTVFNQTVMLSAKQLHTLS   226
                            ************************************************************

CA_IX_Homo_sapiens          DTLWGPGDSRLQLNFRATQPLNGRVIEASFPAGVDSSPRAAEPVQLNSCLAAGDILALVF   420
3IAI:A|PDBID|CHAIN|SEQUENCE DTLWGPGDSRLQLNFRATQPLNGRVIEASFP----------------------------   257
                            ******************************

CA_IX_Homo_sapiens          GLLFAVTSVAFLVQMRRQHRRGTKGGVSYRPAEVAETGA459
3IAI:A|PDBID|CHAIN|SEQUENCE ---------------------------------------        257
```

Figure 3.10. The pairwise sequence alignment of the CA IX and the crystal structure of the catalytic domain of CA IX, PDB 3iaiA. The active catalytic domain of the crystal structure is well-aligned. The mutation of the interest Val33Met is in the flanking N-terminus of the wildtype (CA_IX_Homo_sapiens) in the position V33.

```
CA_IX_Homo_sapiens_mutant        MAPLCPSPWLPLLIPAPAPGLTVQLLLSLLLLMPVHPQRLPRMQEDSPLGGGSSGEDDPL   60
3IAI:A|PDBID|CHAIN|SEQUENCE      ------------------------------------------------------------   0

CA_IX_Homo_sapiens_mutant        GEEDLPSEEDSPREEDPPGEEDLPGEEDLPGEEDLPEVKPKSEEEGSLKLEDLPTVEAPG   120
3IAI:A|PDBID|CHAIN|SEQUENCE      ------------------------------------------------------------   0

CA_IX_Homo_sapiens_mutant        DPQEPQNNAHRDKEGDDQSHWRYGGDPPWPRVSPACAGRFQSPVDIRPQLAAFCPALRPL   180
3IAI:A|PDBID|CHAIN|SEQUENCE      --------------GPDQSHWRYGGDPPWPRVSPACAGRFQSPVDIRPQLAAFSPALRPL   46
                                               * ************************************.******

CA_IX_Homo_sapiens_mutant        ELLGFQLPPLPELRLRNNGHSVQLTLPPGLEMALGPGREYRALQLHLHWGAAGRPGSEHT   240
3IAI:A|PDBID|CHAIN|SEQUENCE      ELLGFQLPPLPELRLRNNGHSVQLTLPPGLEMALGPGREYRALQLHLHWGAAGRPGSEHT   106
                                 ************************************************************

CA_IX_Homo_sapiens_mutant        VEGHRFPAEIHVVHLSTAFARVDEALGRPGGLAVLAAFLEEGPEENSAYEQLLSRLEEIA   300
3IAI:A|PDBID|CHAIN|SEQUENCE      VEGHRFPAEIHVVHLSTAFARVDEALGRPGGLAVLAAFLEEGPEENSAYEQLLSRLEEIA   166
                                 ************************************************************

CA_IX_Homo_sapiens_mutant        EEGSETQVPGLDISALLPSDFSRYFQYEGSLTTPPCAQGVIWTVFNQTVMLSAKQLHTLS   360
3IAI:A|PDBID|CHAIN|SEQUENCE      EEGSETQVPGLDISALLPSDFSRYFQYEGSLTTPPCAQGVIWTVFNQTVMLSAKQLHTLS   226
                                 ************************************************************

CA_IX_Homo_sapiens_mutant        DTLWGPGDSRLQLNFRATQPLNGRVIEASFPAGVDSSPRAAEPVQLNSCLAAGDILALVF   420
3IAI:A|PDBID|CHAIN|SEQUENCE      DTLWGPGDSRLQLNFRATQPLNGRVIEASFP----------------------------   257
                                 ******************************

CA_IX_Homo_sapiens_mutant        GLLFAVTSVAFLVQMRRQHRRGTKGGVSYRPAEVAETGA        459
3IAI:A|PDBID|CHAIN|SEQUENCE      --------------------------------------        257
```

Figure 3.11 The pairwise alignment of the CA IX with mutant and the crystal structure of the catalytic domain of CA IX, PDB 3iaiA. The active catalytic domain of the crystal structure is well-aligned. The mutation of the interest Val33Met is in the flanking N-terminus of the wildtype (CA_IX_Homo_sapiens) in the position M33.

Furthermore, in comparing the multiple sequence alignment (MSA) results involving the N-terminus PG domain with different homologous sequences, there are sequence patterns and tandem repeat motifs, which are conserved and can be clearly observed (Figures 3.12 and 3.13). These include the sparsely conserved six-fold tandem repeat of GEELDP peptides, which is associated with the cell adhesion (Závada et al., 2000). Additionally, there is relatively highly conserved homologous region between 22 and 45 residues, where the V33M mutation resides.

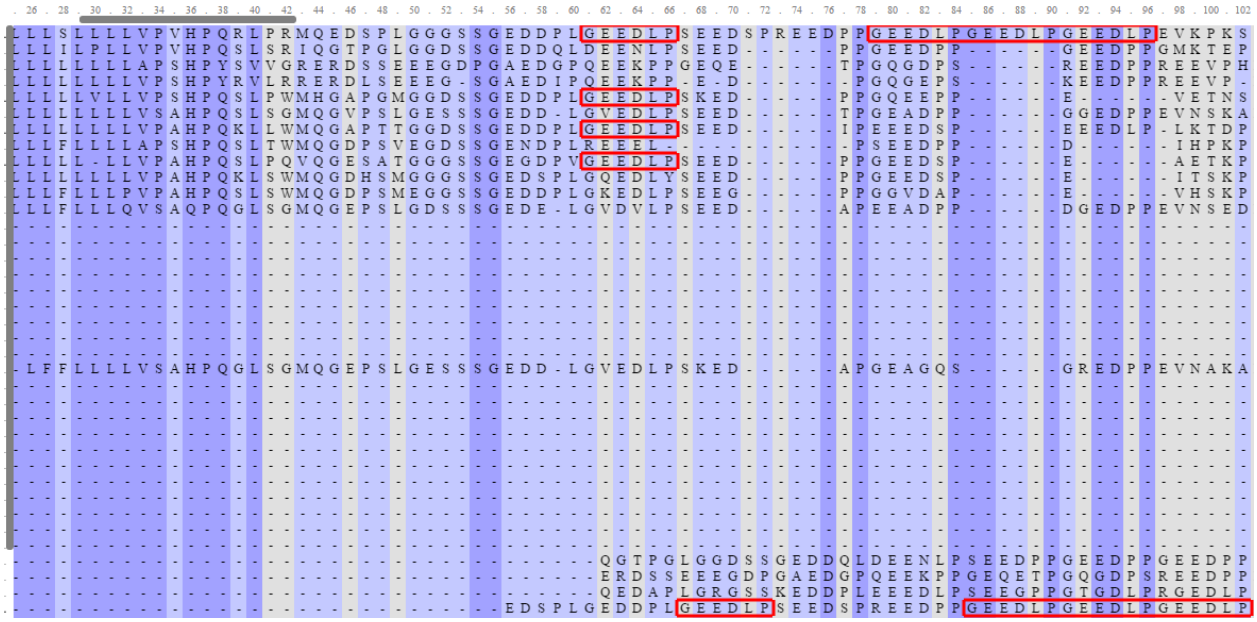


Figure 3.12. The spuriously conserved region of six-fold tandem repeat of peptide GEEDLP across different carbonic anhydrase isoforms. The first sequence is CA IX, which has a number of these tandem repeat. This N-terminus region of PG is believed to be involved with cell adhesion.

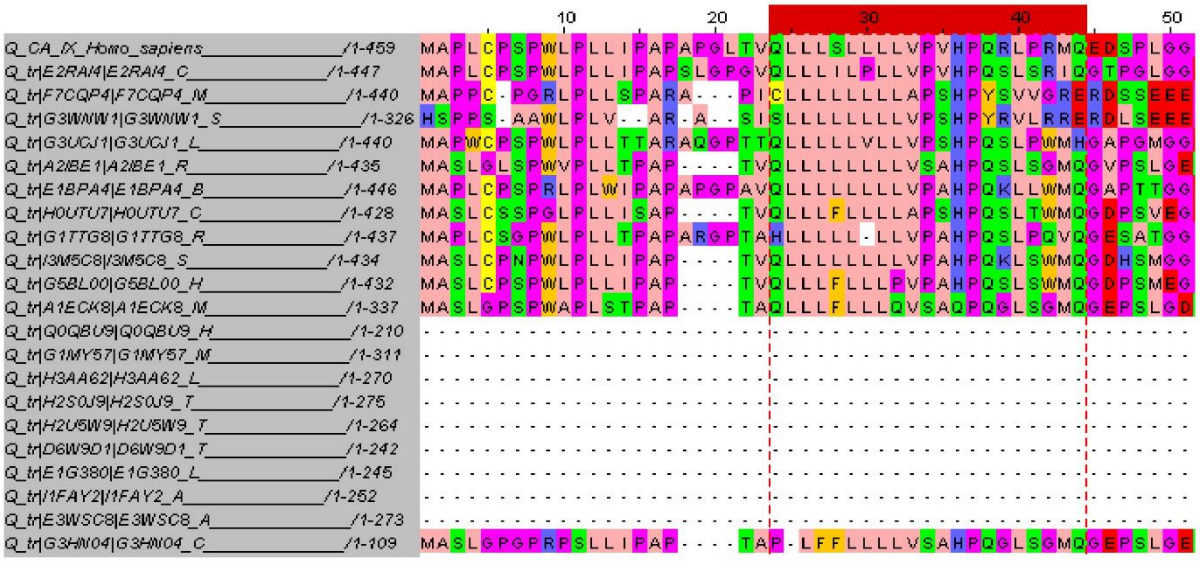


Figure 3.13. The relative highly conserved regions between 22-45 residues (the red-highlighted numbers). This region is abundant with highly hydrophobic residues and is close to signal peptides.

3.6.2.2 ***3D models of the extracellular PG domain (1 – 50 residues) region***

Based on the disorder prediction of the full protein, the mutation appears to be within the intrinsically disordered region of the N-terminus PG domain (1-50 residues). Figures 3.14 and 3.15 show the disordered plots from DISOclust and DISOPRED respectively, with relatively lower disorder probability in the region of V33M mutation, which is also likely to be protein binding.

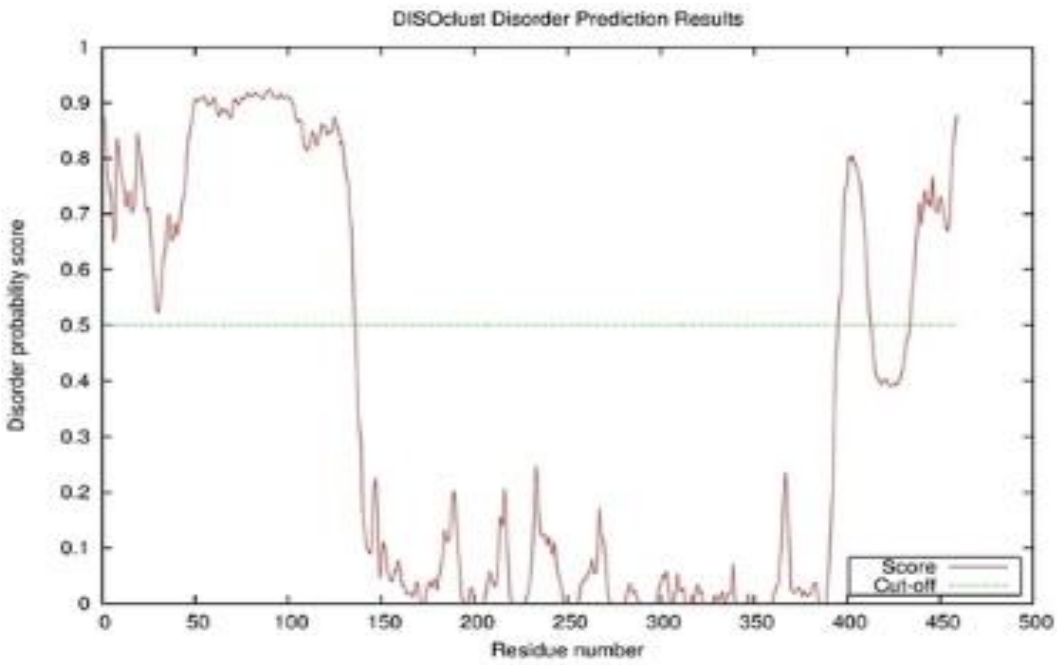


Figure 3.14. The DISOclust plot showing the probability distribution of disordered state for each residues. The Val33Met mutation is likely to be in the disordered regions but with low probability, which means it could fold up.

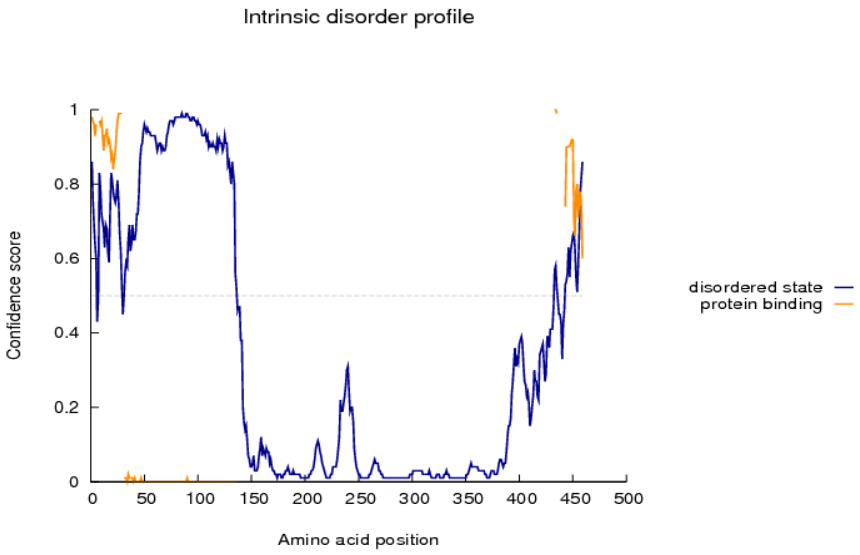


Figure 3.15. The disordered plot from DisoPRED showing the probability distribution of the disordered state for each residue. The Val33Met mutation likely is in the relatively low disordered and protein binding region.

In separately modelling the 3D structures for the disordered region from residues 1-50, the prediction results show that the highly conserved segment (from 22 – 40/45 residues) can be folded, reflecting the above MSA results (Figure 3.13). The majority of the models produced a similar fold, which potentially resembles the helix-loop-helix (HLH) motif (Figure 3.16).

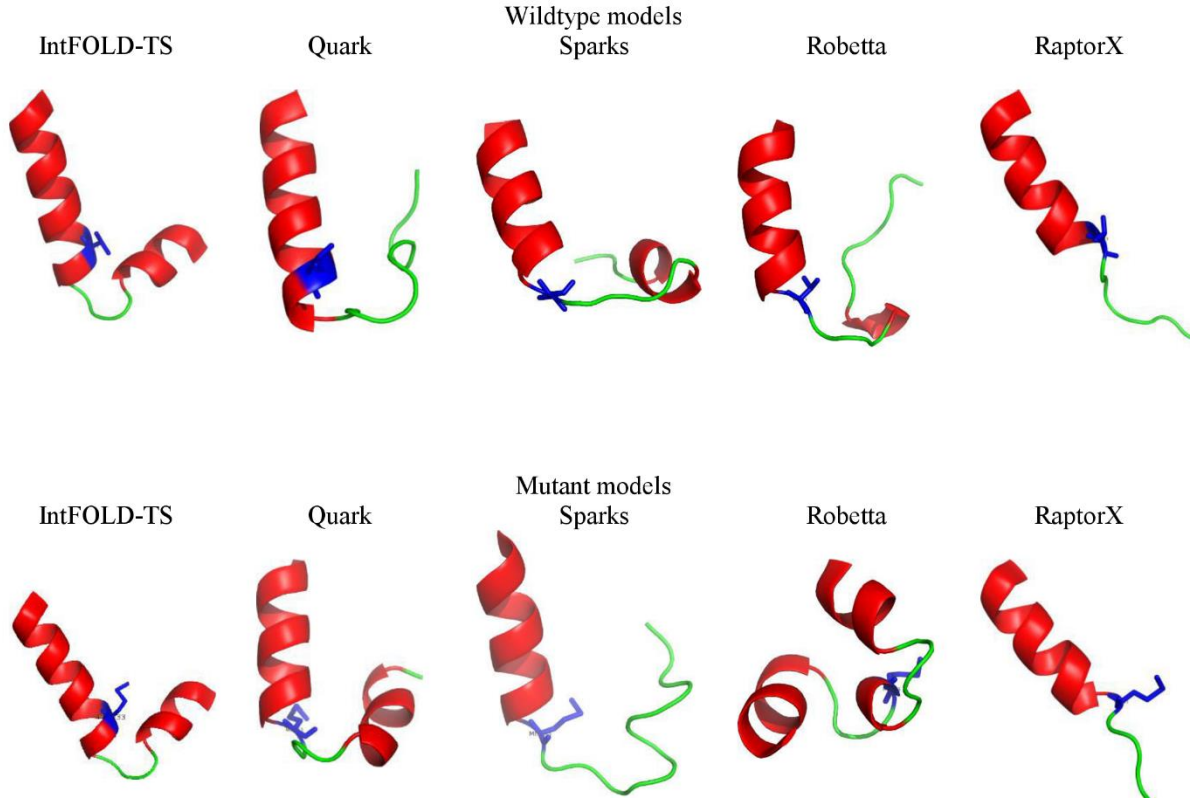


Figure 3.16 Wildtype (top panel) and mutant (bottom panel) 3D models of the N-terminus PG subdomain of the CA IX protein starting from residues 1-50. These are selected models from the methods showing relative similarity in their folds in the region 22 – 40/45. They all have long single helix with potentially similar turns and slightly vary on what to be likely another helix.

3.6.2.2.1 Model comparison, scores and quality assessment

Further assessment of the models from Figure 3.16, using the TM-score pairwise structural comparison method, resulted in the score matrices as shown in the Tables 3.5 and 3.6, for all five wildtype and mutant models respectively.

	IntFOLD	Quark	Robetta	SparksX	RaptorX	Bioserf	Sum	Mean
IntFOLD	-	0.3492	0.2812	0.2968	0.3019	0.2057	**1.4348**	**0.239133**
Quark	0.3492	-	0.3054	0.2813	0.3576	0.2181	**1.5116**	**0.251933**
Robetta	0.2812	0.3054	-	0.2944	0.2717	0.2308	1.3835	0.230583
SparksX	0.2968	0.2813	0.2944	-	0.2498	0.1876	1.3099	0.218317
RaptorX	0.3019	0.3576	0.2717	0.2498	-	0.188	1.369	0.228167
Bioserf	0.2057	0.2181	0.2308	0.1876	0.188	-	1.0302	0.1717

Table 3.5. The comparison matrix showing TM scores among wildtype models (CA IX 1-50 residues). IntFOLD and Quark models have relatively higher scores with Quark being the highest.

	IntFOLD	Quark	Robetta	SparksX	RaptorX	Bioserf	Sum	Mean
IntFOLD	-	0.3127	0.2776	0.2828	0.3171	0.234	**1.4242**	**0.237367**
Quark	0.3127	-	0.2827	0.3556	0.2847	0.2511	**1.4868**	**0.2478**
Robetta	0.2776	0.2827	-	0.2855	0.275	0.2384	1.3592	0.226533
SparksX	0.2828	0.3556	0.2855	-	0.2645	0.2118	1.4002	0.233367
RaptorX	0.3171	0.2847	0.275	0.2645	-	0.2164	1.3577	0.226283
Bioserf	0.234	0.2511	0.2384	0.2118	0.2164	-	1.1517	0.19195

Table 3.6. The comparison matrix showing TM-scores among mutant models (CA IX 1-50 residues). Again, IntFOLD and Quark models have relatively higher scores with Quark being the highest.

Based on the ModFOLD4 Estimation of Model Accuracy (EMA) scores, the IntFOLD and Quark models have a relatively higher significance scores. The global model quality scores for the IntFOLD and Quark are 0.3530 and 0.3458 respectively, with “medium” significance statistical confidence p-values of 2.298E-2 and 2.477E-2 respectively (less than a 1/20 chance that the models are incorrect). Conversely, the better performing version of the EMA method, ModFOLD6, predicts the Quark model to be of higher quality than that of IntFOLD. The global scores for the wildtype Quark and IntFOLD models are 0.4708 and 0.3344 respectively, with

"high" and "medium" significance statistical confidence p-values of 3.05E-4 (less than a 1/100 chance that the models are incorrect) and 7.71E-2 (less than a 1/20 chance that the models are incorrect) respectively, while the mutant Quark and IntFOLD models are 0.4550 and 0.3496 respectively, with "high" and "medium" significance p-values 6.045E-4 and 5.026E-2 respectively. Thus, from Figure 3.16, and based on the EMA method, the highest scoring 3D model is from Quark followed by the model from the IntFOLD reflecting the results of the comparison matrices in the Tables 3.5 and 3.6.

3.6.2.2.2 The likelihood of the mutation being part of the signal peptide

The signal peptide probability score for residues 1-26 is significantly high with a score of 0.885. The cleavage site probability is also significantly high, with a score of 0.832, and the predicted site is likely to occur between the positions G20 and L21 (Figure 3.17).

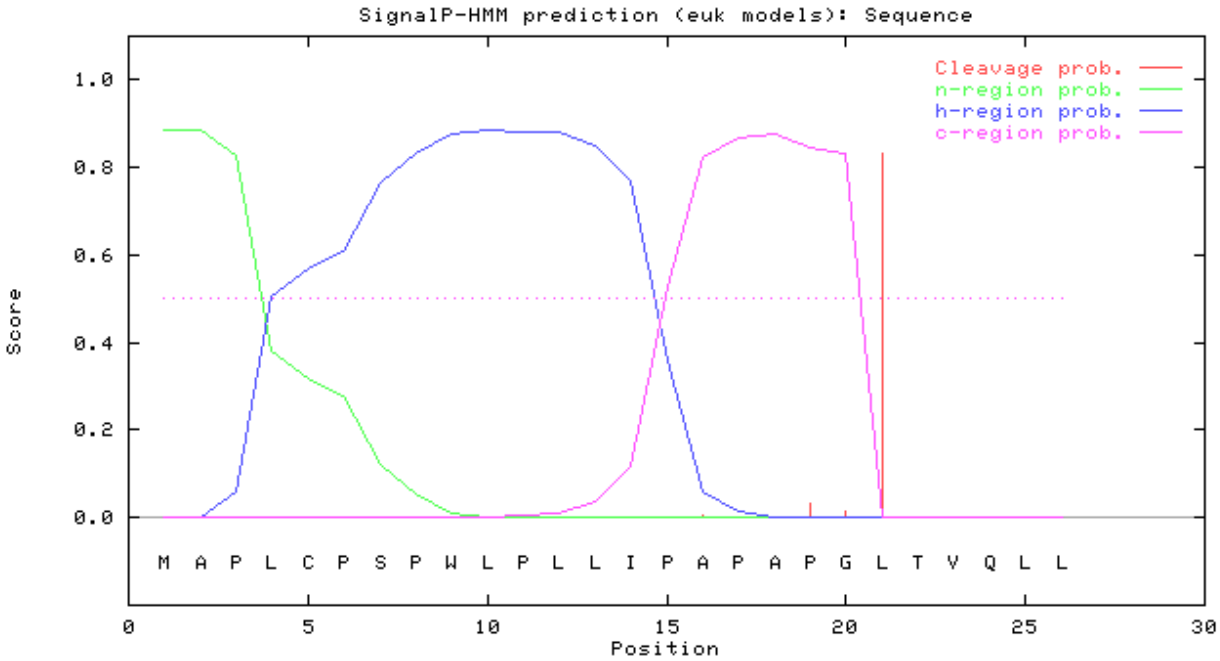


Figure 3.17. The signal peptide prediction score for 1 – 26 residues of N-terminus CA IX by signalP HMM. It can be seen here that the cleavage site probability score has sharply and significantly increased at position L21 (score > 0.8)

Therefore, the signal peptide is likely to end at the position L21, which means that the mutation does not occur within the signal peptide region of CA IX. Moreover, the results from the

transmembrane prediction using MEMSAT SVM is largely consistent with the signal results with slight variation. In this case, with MEMSAT SVM, the results indicate that the signal peptide is likely to end at the position S28.

3.6.2.2.3 The confidence of the predicted model to form an HLH motif

Based on the signal peptide predictions, the residues from T22 or V23 are likely to be a start of the mature CA IX in the PG domain. And based on the several generated models, its predicted 3D structure is potentially an HLH-like motif, Figure 3.19A-B.

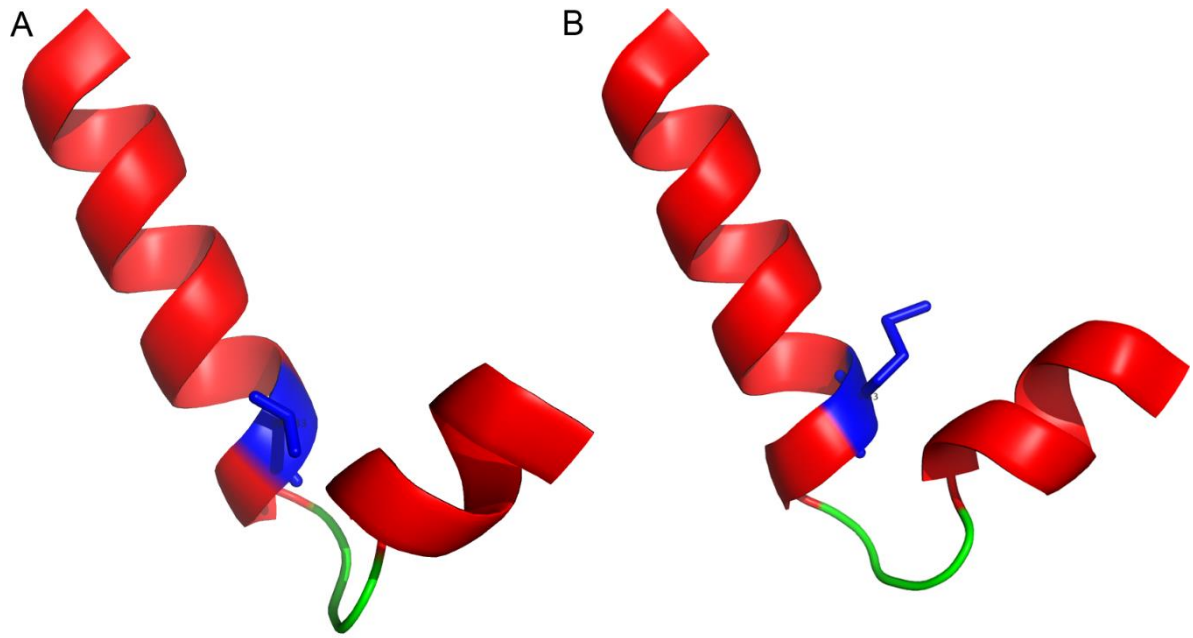


Figure 3.18 The A) wildtype and B) mutant models from IntFOLD for the region starting from residues 22 – 45 within the PG domain. The mature protein is likely to start at position T22 or V23.

The structural alignment results across models appears to be reflecting the sequence alignments, which show the conserved sequences to be within the same the region, i.e. residues 22 – 40 (Figures 3.10 and 3.11). Hence, it is more likely that the N – terminus PG domain contains a potential HLH motif, which seems to be starting at position T22 or V23. Moreover,

several of the conserved sequence patterns in this region pertain to those of HLH (Atchley and Fitch, 1997), which may further suggest the potential of this region to be HLH like subdomain.

Additionally, majority of the templates used by several of the methods to predict the folds of this HLH-like segment are transcription factors or have some transcription regulatory roles. For instance, the template PDB IDs include 2G7G & 3FIW (IntFOLD-TS), 3ZQ7 (Robetta), and 2I7X (RaptorX). This indicates that the folded region is likely to be involved with certain regulatory activity, if not DNA binding. Based on the initial function predicting using FunFOLD, the segment is predicted to bind with CLA ligand, which contains Mg^{2+}. Furthermore, alternative potential structural motif with HLH could be an EF-hand or calcium ion binding proteins, which are associated with cellular signalling or signal transduction (Gifford et al., 2007; Yáñez et al., 2012). Hence, this predicted HLH is likely to be metal ion binding such as Mg^{2+} or Ca^{2+}. In this regard, several of the residues in the likely EF loop region of this predicted HLH-like structure are closely similar to those that frequently appear in the most of the experimentally analysed EF-hand motifs and calcium binding proteins with some variations (Gifford et al., 2007; Haeseleer et al., 2002). The variations might be due to the prediction errors, which could not rule out the possibility of this segment to be an EF-hand, as irregularities and conformational space of the EF-hand and calcium binding residues are vast (Grabarek, 2006).

The final selected high quality models from IntFOLD and Quark methods (wildtype and mutant) are shown in Figure 3.19.

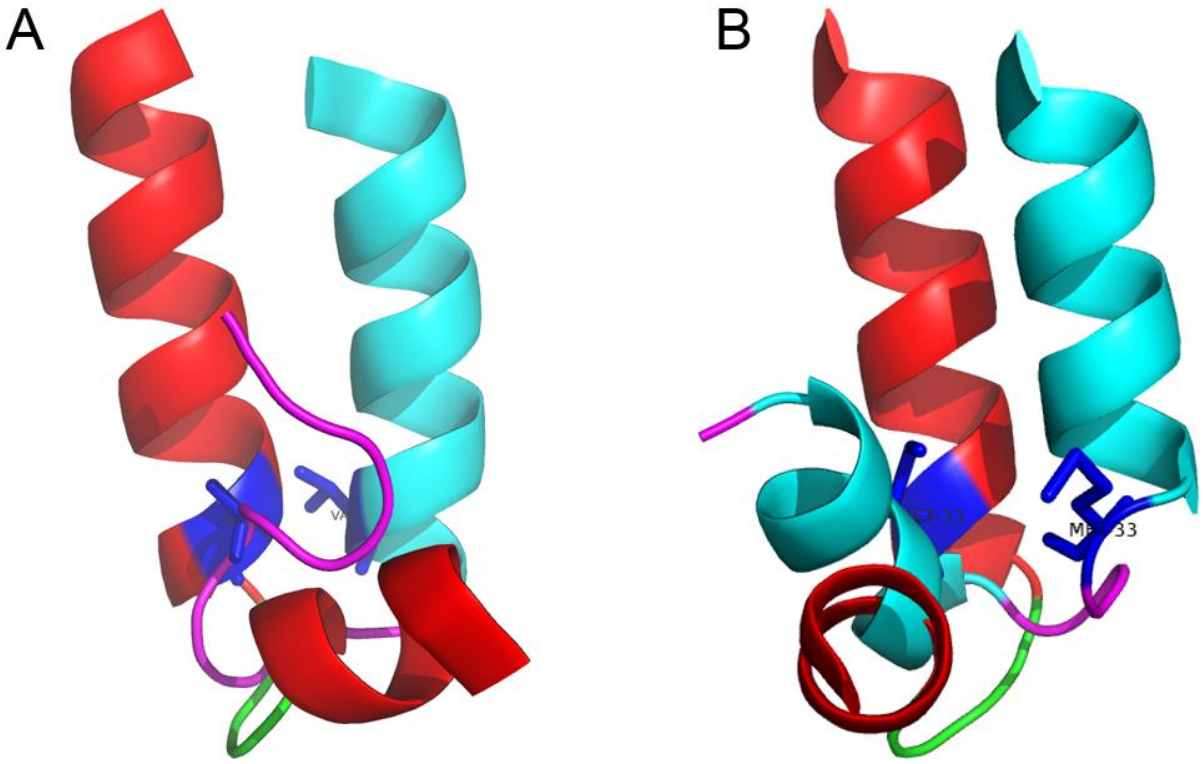


Figure 3.19. The cartoon representation of predicted potential HLH like subdomain within the N-terminus PG domain with A) wildtype and B) mutant superposed models from the IntFOLD and Quark, which interestingly appear to be relatively similar. The wildtype Val-33 and mutant Met-33 residues are in blue sticks. The IntFOLD and Quark models are coloured with red and green, and cyan and magenta in the helices and loops respectively.

Based on the selected structural 3D high quality models in the above Figure 3.19 and sequence alignments, the V33M mutation due to the rs2071676 missense SNP, is likely to be occurring in the putative HLH-like segment within the PG domain in the N-terminus. The similarity of the models from two different methods increases the belief that there is likely presence of HLH like structure in the domain. Hence, the occurrence of this missense mutation in this region is likely to affect the binding activities (due to likely perturbation of binding sites) of the associated domain as is further discussed in the section 3.7.1. Besides, the presence of sulphur in the side chain of Met33 may further affect the ligand interactions or specificity of the molecule.

3.6.3 3D models and function predictions characterising the PECAM1 S563N mutation

3.6.3.1 ***Predicted structures of the full-length wildtype and mutant PECAM1 protein***

The PECAM1 structures for the full length protein are poorly predicted with low global quality score estimates. Figure 3.20 shows the results from the ModFOLD4, which contains the top selected models and their global quality scores.

Graphical ModFOLD4 results for pecam1_wt_full_length_all_models | Help

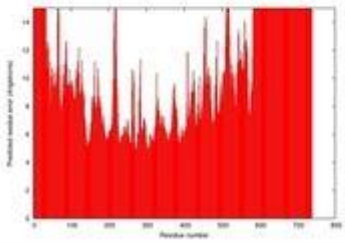
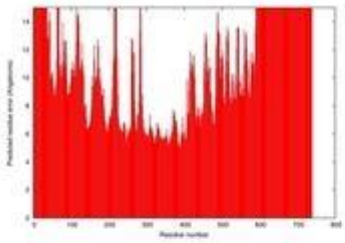
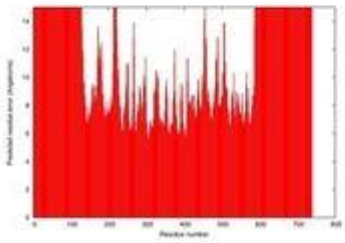
Model name	Confidence and P-value	Global model quality score	Residue error plot (click image for large version)	3D view of residue error (click image for large version)
sparksX_prot1.TS1.pdb	POOR: 3.829E-1	0.0824		
I-tasser_model1.pdb	POOR: 4.086E-1	0.0761		
IntFOLD3_wPPAS_multi8_TS1.bfact.pdb	POOR: 5.226E-1	0.0525		

Figure 3.20 The ModFOLD4 result showing the accuracy of the top 3 predicted models for the full length protein. It is clear that all models have very low confidence values and global scores.

3.6.3.2 ***3D models for the domain with residues 497-596***

Based on the local per residue accuracy of the full length structure models, several regions show a likelihood of being disordered. Figures 3.21 and 3.22 show the predicted disordered regions with mutation (S563N) likely to occur within a disordered region.

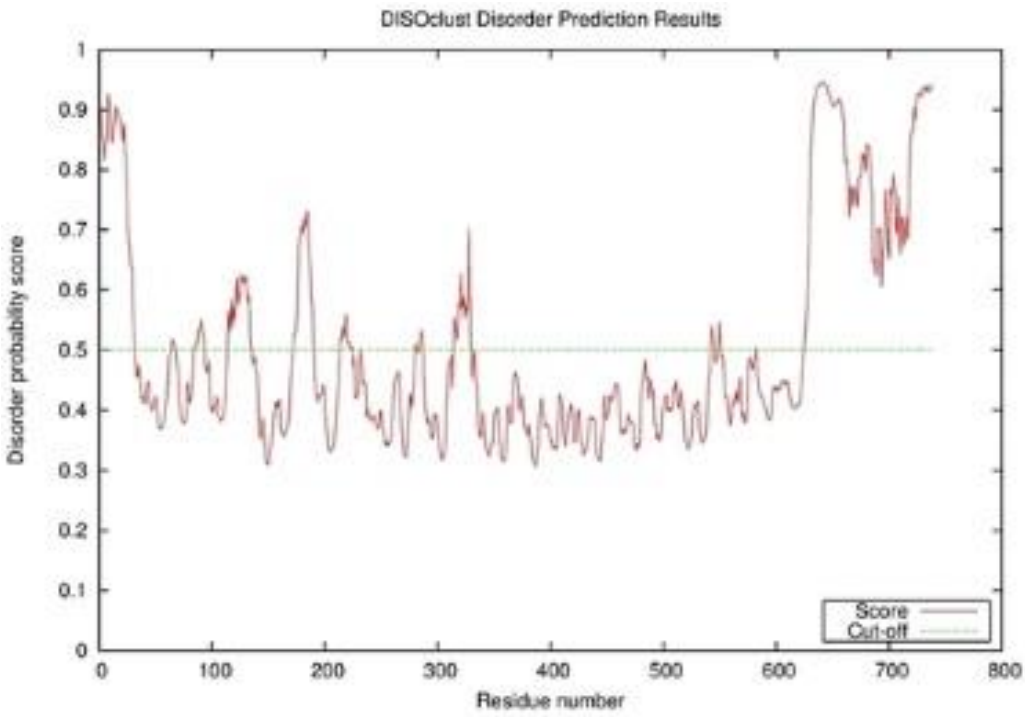


Figure 3.21 The DISOclust plot showing the disordered state of each residue in the wildtype full-length PECAM-1 protein. The mutation S563N is likely to be in a disordered region

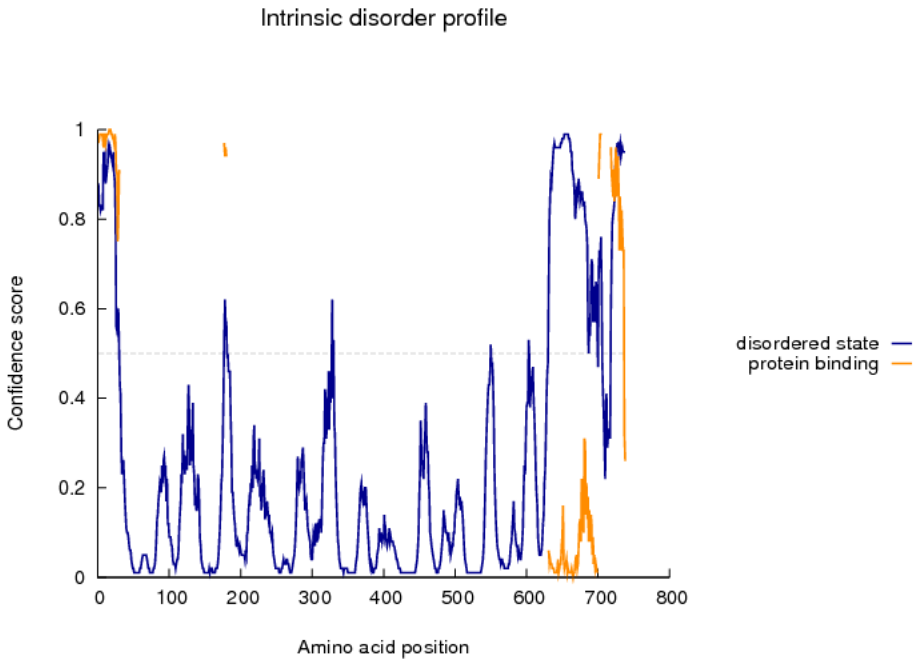
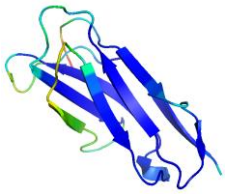
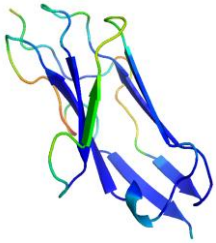
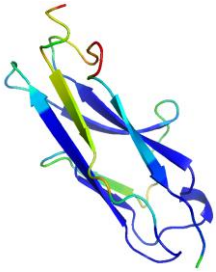


Figure 3.22 The DISOPRED disorder profile showing the disorder probability for each individual residue in the full-length PECAM-1 protein. Again, here the mutation is likely to be located in a disordered region though not a protein binding.

Furthermore, the domain prediction indicates the mutation S563N is likely to be in structural domain 6, which ranges from 497-596 residues. Re-modelling the protein, focussing only on this domain 6 region, resulted in 9 alternative 3D models for each of the wildtype and mutant sequences. The predicted wildtype and mutant models of this key domain have higher quality scores with the structures to be likely to have an Ig (Immunoglobulin) like fold (Table 3.7). Moreover, the predicted Ig domain appears to be relatively similar in each model with the same number of sheets.

Template	**Method**	**Model**	**Reliability Score based on method**	**ModFOLD 4 Score**
1e07_A, 1f97_A2	IntFOLD3		0.7791	0.7812
1im9_D	RaptorX		72	0.7425
4n8v_A	SwissModel		57	0.7013
3p2t_A,	HHpred		4646.1411	0.6683
Igl4b_A	SparksX		10.27	0.6339

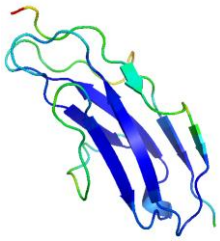
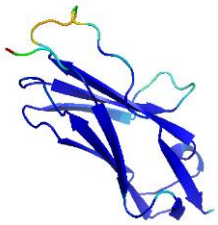
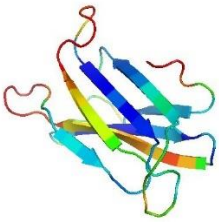
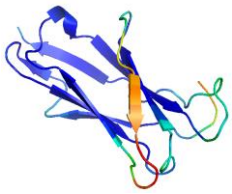
1nbq_A	I-TASSER		0.81	0.7035
4zne_A	Robetta		0.55	0.7476
No template	Quark		0.394	0.3378
2d3vA	Bioserf		49.796	0.7357

Table 3.7 The different models for the wildtype PECAM-1 segment starting from 497-596 positions. The models seem to be relatively significant. Most of the templates used are related or contain Ig domain. which is known to be involved with cell surface recognition, cell adhesion or other immune system-related roles.

3.6.3.3 ***Model quality assessment***

In examining Table 3.7, it appears that the models have closely related structures and share the same general fold. This can be further observed through the pairwise comparison matrices containing the TM scores. Tables 3.8 and 3.9 show the comparison matrices for the mutant and wildtype models respectively.

	IntFOLD	HHpred	Robetta	SwissModel	RaptorX	Bioserf	I-tasser	Sparksx	Quark
IntFOLD		0.662	0.768	0.686	0.719	0.747	**0.789**	0.652	0.263
Hhpred	0.662		0.676	0.605	0.636	0.643	0.749	**0.741**	0.301
Robetta	0.768	0.676		0.732	0.791	0.769	0.754	0.634	0.31
SwissModel	0.686	0.605	0.732		0.819	0.834	0.644	0.581	0.283
RaptorX	0.719	0.636	**0.791**	0.819		**0.854**	0.691	0.62	0.305
Bioserf	0.747	0.643	0.769	**0.834**	**0.854**		0.717	0.615	0.279
I-tasser	**0.789**	**0.749**	0.754	0.644	0.691	0.717		0.671	0.273
SparksX	0.652	0.741	0.634	0.581	0.62	0.615	0.671		0.29
Quark									
Sum	5.023	4.712	5.124	4.901	5.13	5.179	5.015	4.514	2.304
Mean TM score	0.718	0.673	0.732	0.7	0.733	0.74	0.716	0.645	**0.288**

Table 3.8 The TM-scores for the predicted PECAM1 domain 6 mutant models (residues 497-596). Based on the scores the models appear to be structurally similar.

	IntFOLD	Hhpred	Robetta	SwissModel	RaptorX	Bioserf	I-tasser	SparksX	Quark
IntFOLD		0.651	0.799	0.7	0.731	0.757	0.745	0.623	**0.343**
Hhpred	0.651		0.724	0.678	0.723	0.688	**0.774**	0.646	0.261
Robetta	**0.799**	0.724		0.761	0.826	0.811	0.735	0.601	0.317
SwissModel	0.7	0.678	0.761		0.832	0.825	0.648	0.579	0.31
RaptorX	0.731	0.723	0.826	**0.832**		**0.856**	0.697	0.614	0.316
Bioserf	0.757	0.688	**0.811**	0.825	**0.856**		0.712	0.602	0.307
I-tasser	0.745	**0.774**	0.735	0.648	0.697	0.712		**0.657**	0.285
SparksX	0.623	0.646	0.601	0.579	0.614	0.602	0.657		0.291
Quark									
Sum	5.006	4.884	5.257	5.023	5.279	5.251	4.968	4.322	2.43
Mean TM Score	0.715	0.698	0.751	0.717	0.754	0.75	0.71	0.617	0.304

Table 3.9 The TM-scores for the PECAM1 domain 6 wildtype models (residues 497-596).

The above Tables 3.8 and 3.9 provides an overview of the potential quality of the models. The relatively high pairwise TM-scores might imply high model quality (as most models agree) and hence they are suited for further quality assessment using MQAPs. When evaluating the models using ModFOLD6, the results agree with those shown in the Tables 3.8 and 3.9, (Figures 3.23 and 3.24). Additionally, the Quark model is the lowest ranked and thus, reflecting the pairwise TM-scores as shown in Tables 3.8 and 3.9.

Model name	Confidence and P-value	Global model quality score	Residue error plot (click image for large version)	3D view of residue error (click image for large version)
Robetta_model1_wt_PECAM1_497-596res.pdb	CERT: 7.392E-4	0.6837		
IntFOLD3__wt_PECAM1_497-596res.bfact.pdb	HIGH: 1.012E-3	0.6535		
RaptorX_wt_PECAM1_497-596res.pdb	HIGH: 1.209E-3	0.6363		
Bioserf_wt_PECAM1_497-596res.pdb	HIGH: 1.243E-3	0.6336		
I-tasser_model1_wt_PECAM1_497-596res.pdb	HIGH: 1.678E-3	0.6048		
Swiss_model_02_wt_PECAM1_497-596res.pdb	HIGH: 1.738E-3	0.6014		
hhpred_pecam1_wt_497-596res.pdb	HIGH: 1.758E-3	0.6003		
SparksXmodel1_wt_PECAM1_497-596res.pdb	HIGH: 1.831E-3	0.5964		
Quarks_model1_wt_PECAM1_497-596res.pdb	HIGH: 8.696E-3	0.4465		

Figure 3.23, The ModFOLD6 assessment of the wildtype models (residues 497-596). Most of the models' scores appear to be similar with those in Table 3.8. Based on these scores, the Robetta model was ranked at the top.

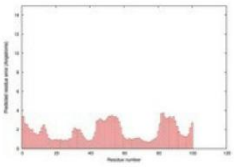
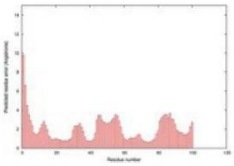
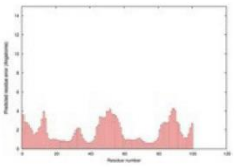
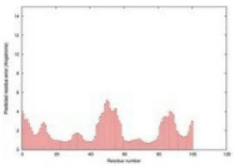
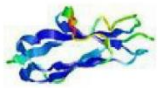
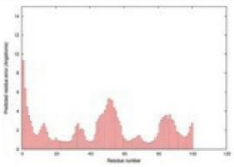
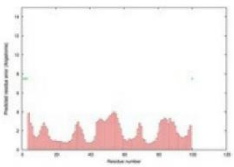
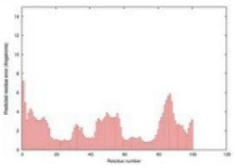
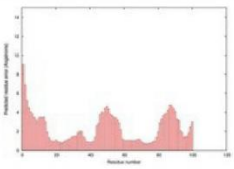
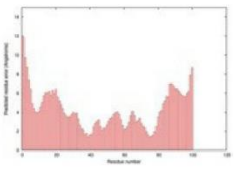
Model name	Confidence and P-value	Global model quality score	Residue error plot (click image for large version)	3D view of residue error (click image for large version)
Robetta_model1_PECAM1_mt_497-596res.pdb	CERT: 8.033E-4	0.6757		
RaptorX_model_PECAM1_mt_497-596res.pdb	CERT: 8.71E-4	0.6679		
I-tasser_model1_PECAM1_mt_497-596res.pdb	CERT: 8.756E-4	0.6674		
IntFOLD_TS1_PECAM1_mt_497-596res.bfact.pdb	HIGH: 1.013E-3	0.6534		
Bioserf_model_pecam1_497_596res_mut.pdb	HIGH: 1.019E-3	0.6528		
Swiss_model02_PECAM1_mt_497-596res.pdb	HIGH: 1.226E-3	0.6350		
SparksX_.TS1_PECAM1_mt_497-596res.pdb	HIGH: 1.671E-3	0.6052		
hhpred_model_PECAM1_mt_497-596res.pdb	HIGH: 1.722E-3	0.6023		
Quark_model1_PECAM1_mt_497-596res.pdb	HIGH: 7.27E-3	0.4637		

Figure 3.24 The ModFOLD6 assessment of the mutant models (residues 497-596). Again, the Robetta mutant model was ranked at the top.

The 3D models (wildtype and mutant) from Robetta are predicted to be the highest quality and thus, are of higher significance, Figures 3.25 and 3.26.

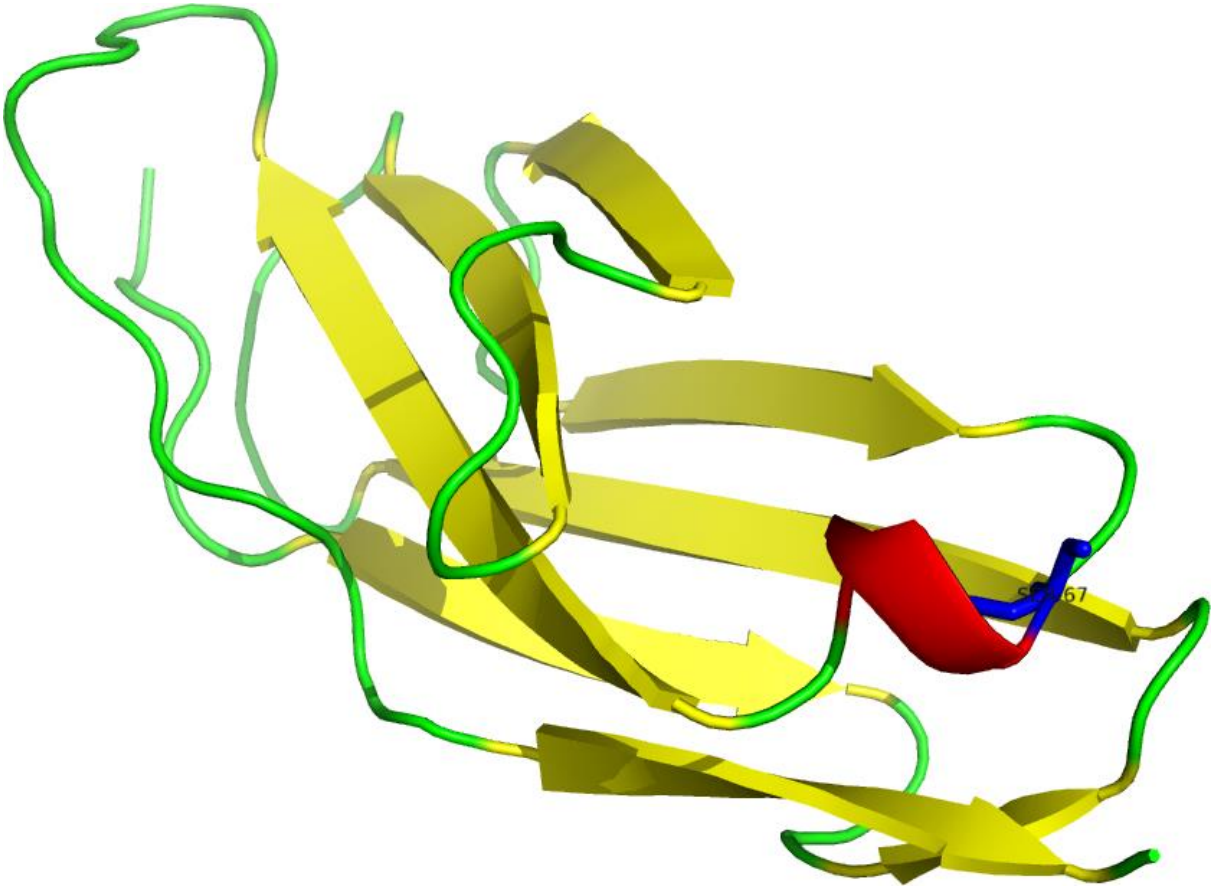


Figure 3.25 The wildtype Robetta model for the PECAM1 domain 6 (residues 497-596). The model was created based on the template 4ZNE. The wildtype residue (Ser) is in the position 67, since the domain was separately modelled. The Ig domain type of the wildtype appears to be C-type.

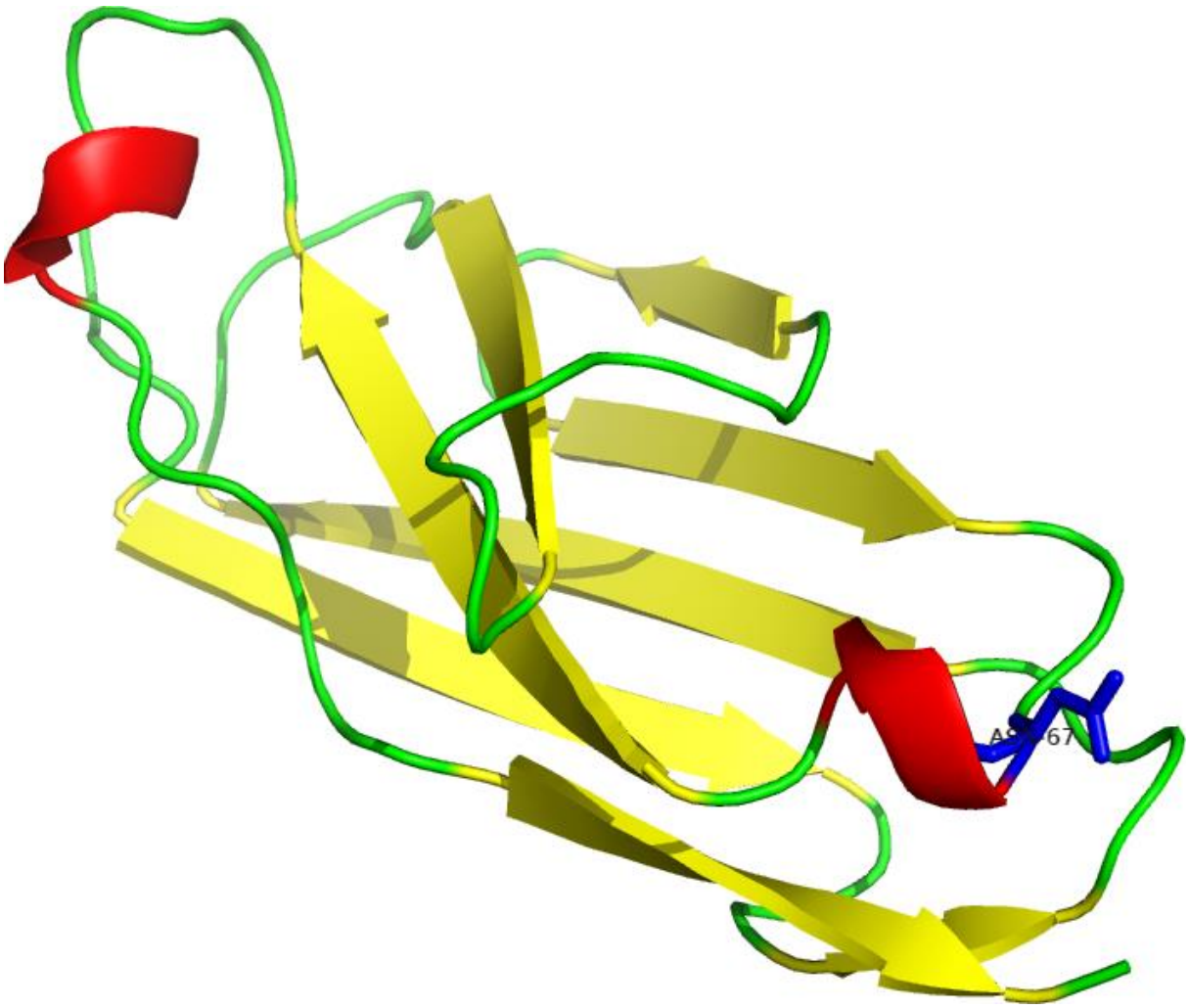


Figure 3.26 The mutant Robetta model for the PECAM1 domain 6 (residues 497-596). The model was created based on the template 4PBV. The mutant residue (Asn) is in the position 67. The Ig domain type of the wildtype appears to be V-type.

Based on the templates used to generate the selected models, the mutant's domain 6 appears to be adopting a V-Ig domain type similar to the template 4PBV (Coles et al., 2014), while the wildtype's domain 6 is of C-Ig domain type similar to 4ZNE (Oganesyan et al., 2015).

3.6.3.4 ***Further model analyses to investigate the potential structural impact of the S563N mutation***

3.6.3.4.1 PSI-BLAST results for the related domain 6 Ig -type from the Uniprot -Family & Domains (499 – 591 residues)

The PECAM1 domain 6 from Uniprot's Family & Domains is of Ig-like C2- type, which contains the residues 499 – 591. This domain appears to be similar with residues 497 – 596 in the predicted domain 6 of the wildtype 3D models excluding the additional residues 497 and 498, and 592 – 596 in N and C termini respectively. The PSI-BLAST results for the PECAM1 domain 6 from UniProt Family & Domains (499 – 591 residues) show that there is significant alignment with the sequence from the PDB structure 2NPL_X, Figures 3.27 and 3.28. Moreover, the sequence identity with the aligned structures is just beyond the "twilight zone" sequence identity threshold (~25%), where it is possible to perform further comparative structural analyses with the predicted 3D models.

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

	Description	Max score	Total score	Query cover	E value	Ident	Accession	Select for PSI blast	Used to build PSSM
☑	Chain X, Nmr Structure Of Card D2 Domain	36.3	36.3	68%	5e-04	29%	2NPL_X	☑	
☑	Chain A, Drosophila Robo Ig1-2 (Monoclinic Form)	35.9	35.9	77%	0.002	32%	2VRA_A	☑	
☑	Chain A, Drosophila Robo Ig1-2 (Tetragonal Form)	35.9	35.9	77%	0.002	32%	2VR9_A	☑	
☑	Chain A, I-band Fragment I65-i70 From Titin	35.5	66.4	91%	0.004	30%	3B43_A	☑	
☑	Chain A, Two Membrane Distal Igsf Domains Of Cd166	35.1	35.1	91%	0.004	26%	5A2F_A	☑	

Figure 3.27 The PSI BLAST results after aligning the Uniprot Ig domain C2 type 6 (residues 499 – 591) against the PDB sequences. It can be seen that the sequence significantly aligned with five structures including those of Ig domains. The sequence identity is above the threshold (>25%), where it might be possible to infer function.

Chain X, Nmr Structure Of Card D2 Domain

Sequence ID: 2NPL_X **Length:** 96 **Number of Matches:** 1

Range 1: 14 to 91 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
33.5 bits(75)	0.008	Compositional matrix adjust.	20/78(26%)	35/78(44%)	1/78(1%)

```
Query  16  ESGEDIVLQCAVNEGSGPITYKFYREKEGKPF-YQMTSNATQAFWTKQKASKEQEGEYYC  74
           E G D  ++C    EGS P+ Y++ +  + +       +  T +  + + AS E  G Y C
Sbjct  14  EIGSDFKIKCEPKEGSLPLQYEWQKLSDSQKMPTSWLAEMTSSVISVKNASSEYSGTYSC  73

Query  75  TAFNRANHASSVPRSKIL  92
           T  NR      + R  ++
Sbjct  74  TVRNRVGSDQCLLRLNVV  91
```

Figure 3.28 Sequence alignment between the Uniprot wildtype Ig domain C2 type 6 (499 – 591 residues) and the top hit 2NPL_X structure. The wildtype residue Ser (S) at position 65.

Furthermore, the PSI-BLAST results involving the mutant residue with the same sequence from the Uniprot, shows a significant alignment with 4X4M_E, which was a top hit (Figures 3.29 and 3.30).

Select: All None Selected:0

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

	Description	Max score	Total score	Query cover	E value	Ident	Accession	Select for PSI blast	Used to build PSSM
☐	Chain E, Structure Of Fcgammari In Complex With Fc Reveals The Importance Of Glycan Recognition For	35.0	35.0	81%	0.005	25%	4X4M_E	☑	

Run PSI-Blast iteration 2 with max 500 Go

Sequences with E-value WORSE than threshold

Select: All None Selected:0

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

	Description	Max score	Total score	Query cover	E value	Ident	Accession	Select for PSI blast	Used to build PSSM
☐	Chain A, Igg1 Fc-fcgammari Ecd Complex	34.3	34.3	89%	0.009	24%	4ZNE_A	☐	
☐	Chain A, Crystal Structure Of Fc Ri And Its Implication To High Affinity Immunoglobulin G Binding	34.3	34.3	89%	0.009	24%	3RJD_A	☐	
☐	Chain A, Human Fc Gamma Receptor Iib Ectodomain (Cd32)	33.5	33.5	79%	0.014	26%	2FCB_A	☐	

Figure 3.29 The PSI BLAST alignment results after inserting the mutant residue N in the C-type domain 6 of the PECAM1, which was retrieved from the Uniprot-Family & Domains. It can be clearly seen that the top hit is 4X4M_E structure. But, also other structures have relatively high significant as well though below the threshold set during PSI-BLAST parameters' set up.

Chain E, Structure Of Fcgammari In Complex With Fc Reveals The Importance Of Glycan Recognition

Sequence ID: 4X4M_E **Length:** 275 **Number of Matches:** 1

▶ See 1 more title(s)

Range 1: 89 to 159 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
35.0 bits(79)	0.005	Compositional matrix adjust.	19/76(25%)	33/76(43%)	5/76(6%)

```
Query  10   LSSKVVESGEDIVLQCAVNEGSGPITYKFYREKEGKPFYQMTSNATQAFWTKQKANKEQE  69
            +SS+V+  GE + L+C   +        +YR  +   F+   SN T       K N
Sbjct  89   VSSRVLTEGEPLALRCHAWKDKLVYNVLYYRNGKAFKFFHWNSNLTIL-----KTNMSHS  143

Query  70   GEYYCTAFNRANHASS  85
            G Y+C+   +  + S+
Sbjct  144  GTYHCSGMGKHRYTSA  159
```

Figure 3.30 Sequence alignment between the Uniprot mutant Ig domain 6 (residues 499 – 591) and the top hit 4X4M_E_A structure. The mutant residue Asn (N) is at position 65.

3.6.3.4.2 Further structural analyses

From the PSI-BLAST alignments results, the first structure (2NPL_X) (Jiang and Caffrey, 2007), which significantly aligned with the Uniprot domain (499-591 residues) (E-value 5e-04) was examined together with the wildtype model. It appears that the published 2NPL_X structure is of Ig-domain C-type and hence, may confirm the wildtype domain fold to be of C-type as shown above and similar to the UniProt's Family & Domains annotation. Figure 3.31 compares the folds of 2NPL_X and the predicted wildtype model for PECAM1 domain 6.

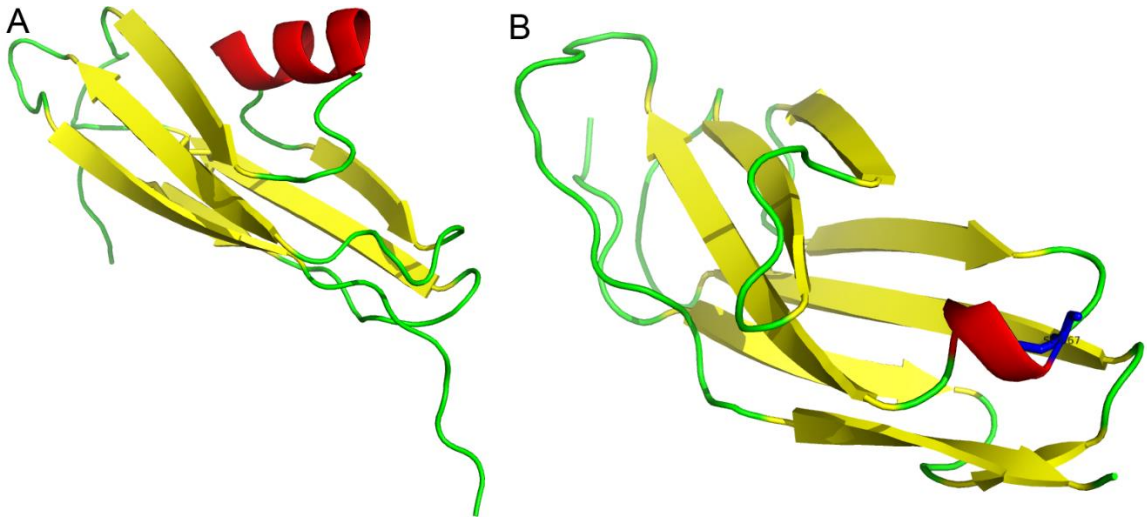


Figure 3.31 The 3D structures of A) 2NPL_X and B) wildtype PECAM1 Ig – domain 6 model (497-596 residues), which is similar to the Uniprot annotated domain (499-591 residues). The wildtype appears to adopt a C-2 type fold as in the case for the annotated Uniprot domain and similar to the 2NPL_X structure. The two beta sheets are stabilised by the disulphide bonds.

Furthermore, it appears that the fold of the mutant model resembles that of the 4X4M_E (Sondermann et al., 1999), which is likely to be of V-type. Figures 3.32 shows the folds of 4X4M_E and the mutant model.

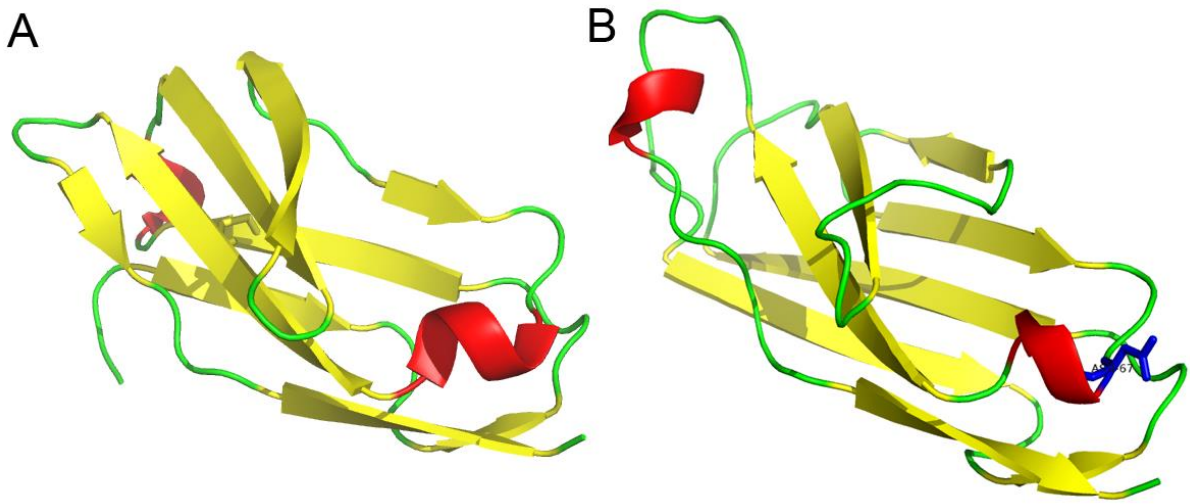


Figure 3.32 The 3D structures of the A) 4X4M_E Ig domain 1 (101 – 186 residues) and mutant model PECAM1 (497-596 residues). The mutant model of the PECAM1 domain appears to adopt the Ig V-type similar to that of aligned 4X4M_E. The two beta sheets are stabilised by the disulphide bonds (yellow stick in 4X4M).

Therefore, from the top model templates used to generate the 3D structures and the PSI-BLAST sequence alignment results, the S563N replacement due to missense SNP might be potentially changing the related Ig-domain of PECAM1 from C- to V-type.

3.6.3.4.3 Overview of Ig – domain types

The above structures all have classical Ig like fold, which is characterised by a sandwich of two beta-sheets containing antiparallel beta strands featuring a Greek key topology. The two sheets are normally stabilised by a conserved disulphide bond (Williams and Barclay, 1988).

In order to further understand the structural impact of mutation, it is vital to examine the underlying features or characteristics of the Ig or Ig superfamily (IgSF) domains. The IgSF domains can be grouped into four major classes namely V, C1, C2, and I with C1 and C2 representing the C (constant) domain (Williams and Barclay, 1988). The beta strands in the two sheets of both the C and V classes are given the letters starting from A in the N-terminus to G in the C-terminus (Figure 3.26). The V (variable) domain is longer and may contain two extra strands in the core of the domain named C' and C" as an extension of the C strand in the core of the domain and hence, leading to the variability of the V domain. In addition, the variability is enhanced using three loops, which connect B with C, C' with C", and F with G strands. These are coloured red as shown in the Figure 3.33 (Barclay, 1999; Teichmann and Chothia, 2000; Williams and Barclay, 1988).

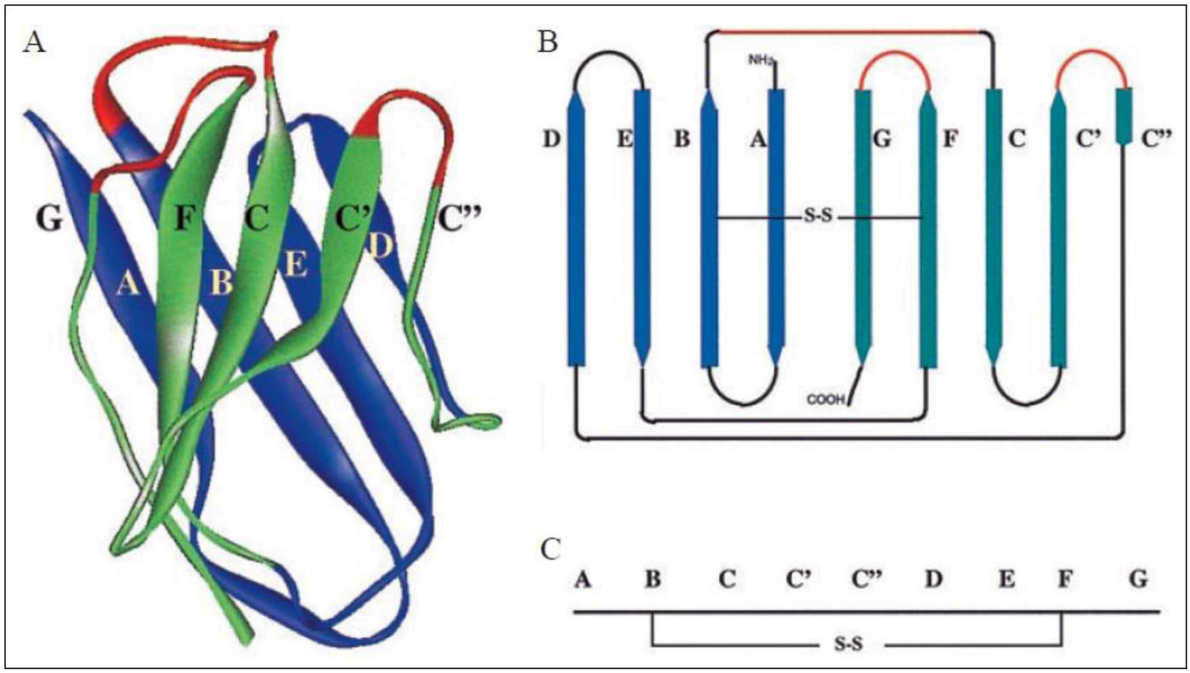


Figure 3.33 The basic Ig fold structure of the V type domains. A) Example of ribbon diagram from human myeloma (PDB 7fab, B) and C) are schematics representing the linear relationship between the two sheets and loops. The hypervariable loops are coloured in red. The distinguishing feature between Ig C and V type domains, is the extension of the C loop through C' and C''. The Ig C2 type domain does not normally contain the C''. The wildtype model has a short loop and resembles the C2 type domain. On other hand the mutant seems to have an extended loop, which resemble with the V-type Ig domain. The image was taken from rom (Barclay, 1999).

Thus, based on the above Ig domain characteristics, the wildtype model when is closely observed, it is likely to belong to the C- type, which is similar to the Uniprot annotation, and the mutant model appears to be closer to the V type.

3.6.3.5 ***Functional analyses of Ig domains and the potential functional effect of the S563N mutation***

The Serine (Ser) at position 67 (or 563) in the wild type domain 6, is one of the amino acids, which is heavily involved in the proteins function (Holm and Sander, 1996). The presence of the OH group enables it to form strong hydrogen bond with other substrates, particularly in the protein active sites. Additionally, it could be further interacting with other proteins or protein kinases in the phosphorylation process for enhancing signal transduction. The replacement with Asn, although often favourable, may likely affect the binding specificity versus a Ser at the same position, as Asn is more specific to negatively charged residues.

Thus, from the functional perspective the mutation may potentially decrease the protein's (or domain's) stability and affect the interaction with other proteins, or potentially other Ig domains (Valencia and Pazos, 2003; Williams and Barclay, 1988). And this potential functional effect of the mutation is further explained elsewhere (Baldwin et al., 1994; Wollscheid et al., 2009).

3.6.4 3D models and function predictions characterising the PEAR1 N848K mutation

3.6.4.1 ***Predicted structures of the full-length wildtype and mutant PEAR1 protein***

Predicted 3D models of the full-length PEAR1

The generated 3D models for the full-length PEAR1 are poor as the protein sequence is very long. Thus, nearly all methods were unable to generate convincing full-length 3D models. For instance, the IntFOLD3 predicted wildtype and mutant models are essentially random over the full chain. Figures 3.34 shows the overall IntFOLD3 result of the top two wildtype models.

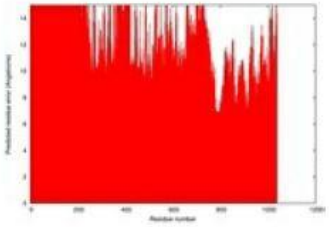
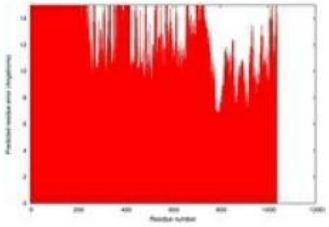
Model ID and PDBsum links for all templates used	Confidence and P-value	Global model quality score	Local model quality plot (click images to download plots)	Model coloured by local quality (click images to view models, local errors and target coverage interactively)
IntFOLD3 wPPAS multi7 TS1 4xbmB	POOR: 5.476E-1	0.0480		
IntFOLD3 wPPAS multi6 TS1 4xbmB	POOR: 5.476E-1	0.0480		

Figure 3.34 The IntFOLD results for the predicted models of the full-length wildtype PEAR1 containing the target point mutation of interest (N848). The models are poorly predicted with high frequency of likely per-residue errors over the full protein chain.

Disordered state of the models

Furthermore, the disorder prediction from DISOclust shows that the mutation region has high probability of being disordered (Figure 3.35), meaning that modelling a fold for this region is likely to be difficult. However, the PSIPRED secondary structure map (Figure 3.36) and DISOPRED results indicate that the wildtype/mutant residue(s) are not predicted to be in a disordered region.

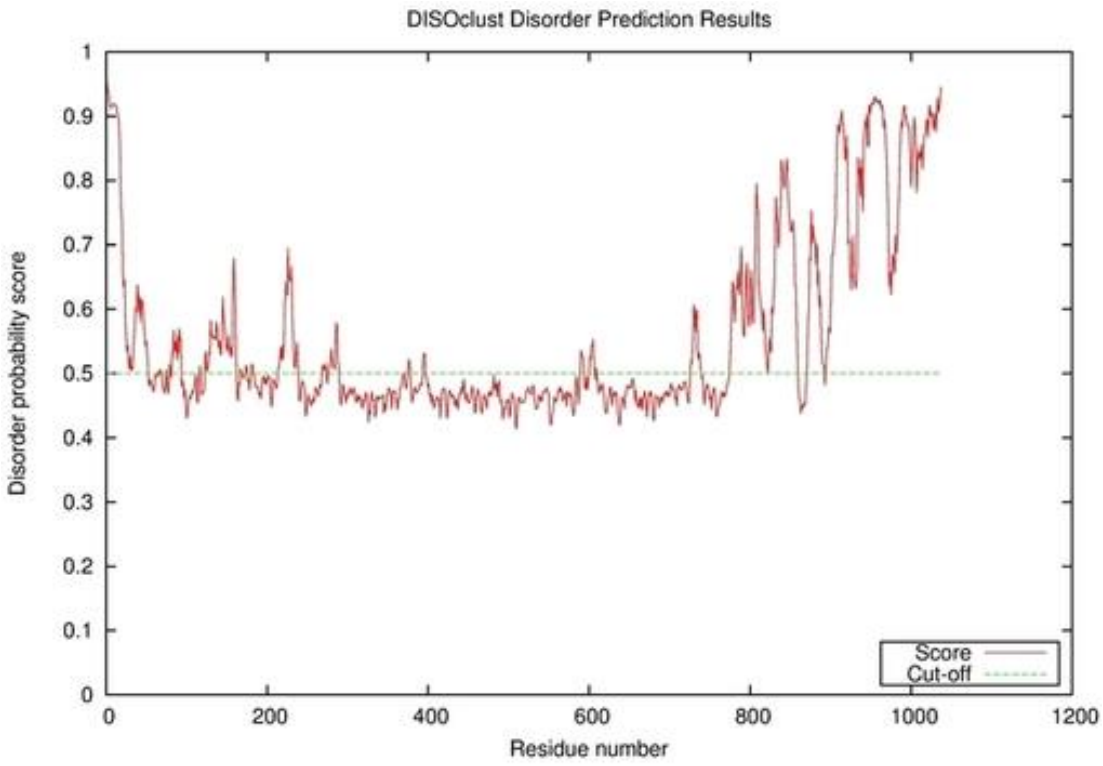


Figure 3.35 The DISOclust plot showing the disordered prediction of the full-length PEAR1. The mutation region has high disordered probability indicating that it might be difficult to fold.

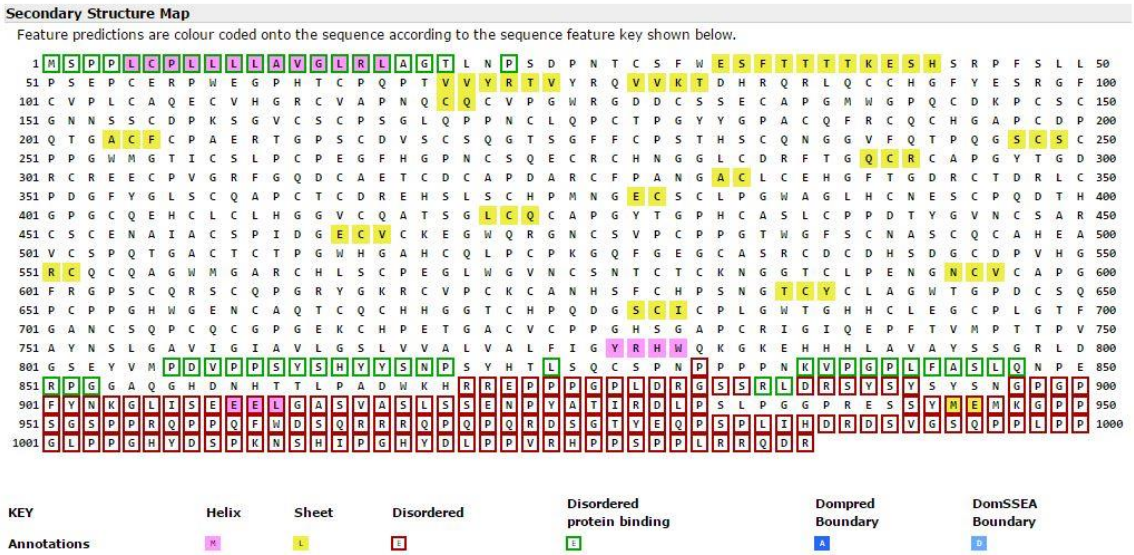


Figure 3.36 The secondary structure and disorder map of the full length PEAR1 with the wildtype N848 residue. It can be seen the residue is near, but not within, the disordered region.

Therefore, based on the disorder predictions and PSIPRED's secondary structure map, the protein was remodelled by reducing its length and focusing only on the 100 residues, from 801 to 900.

Predicted 3D models for the wildtype sequence, residues 801 – 900

The predicted 3D models for this region are of medium significance. Figure 3.37 shows the results of the evaluated quality of the models using ModFOLD6 (Maghrabi and McGuffin, 2017).

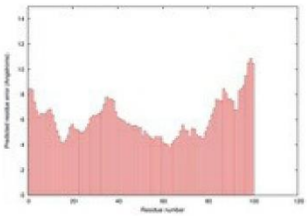
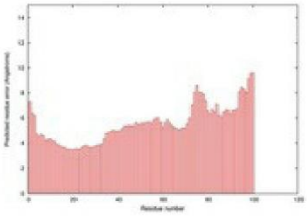
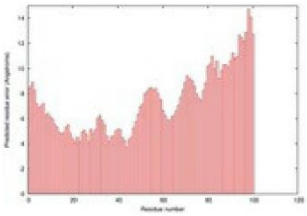
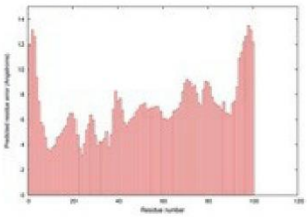
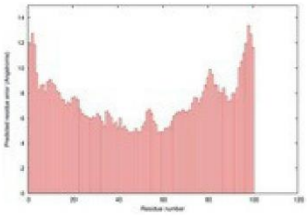
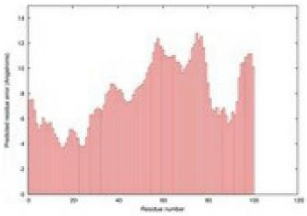
Graphical ModFOLD6 results for PEAR1-wt_801-900res | Help				
Model name	**Confidence and P-value**	**Global model quality score**	**Residue error plot (click image for large version)**	**3D view of residue error (click image for large version)**
Quarks_PEAR1_wt_801-900res.pdb	**MEDIUM: 1.545E-2**	**0.3912**		
IntFOLD3_PEAR1_wt_801-900res.pdb	**MEDIUM: 1.702E-2**	**0.3819**		
Bioserf_PEAR1_wt_801-900res.pdb	**MEDIUM: 2.381E-2**	**0.3496**		
SparksX_PEAR1_wt_801-900res.pdb	**MEDIUM: 2.599E-2**	**0.3412**		
i-tasser_PEAR1_wt_801-900res.pdb	**MEDIUM: 2.775E-2**	**0.3349**		
RaptorX_PEAR1_wt_801-900res.pdb	**MEDIUM: 3.066E-2**	**0.3253**		

Figure 3.37 The quality of the 3D models for the PEAR1 region covering residues 801 – 900. All generated models are of medium significance.

In addition, the DISOclust results for this region show that the mutation N848K is in the relatively low disordered region (i.e. from residues 840-850, or 40-50 on the plot), thus, it may or not fold. Figure 3.38 shows the plot for the disordered regions.

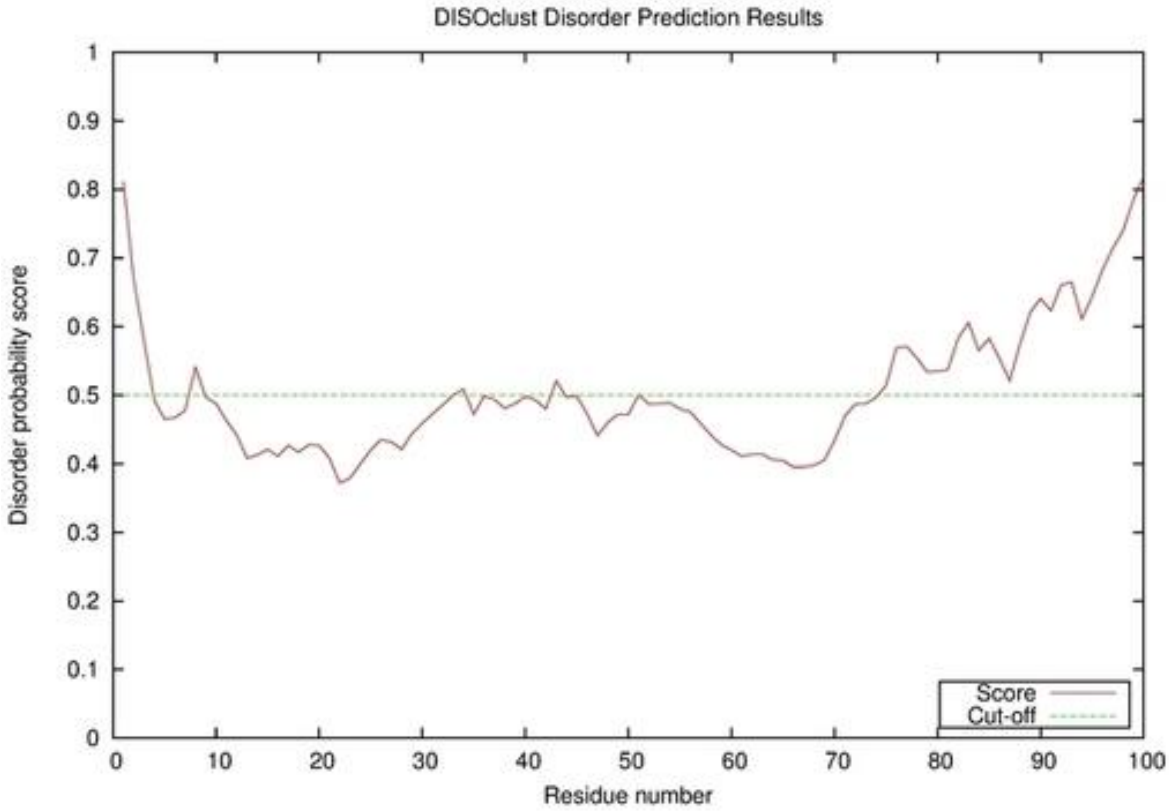


Figure 3.38 The disordered plot for residues 801-900 of the wildtype PEAR1. The mutation of interest (N848K) is likely to be in the relatively low disordered region between residues 840-850 residues (40-50 on the plot).

Attempting to remodel the 40 – 50 regions (Figure 3.38) failed to produce any significant and reasonable 3D models and hence from the structural aspect the mutation is yet uncharacterised.

3.6.4.2 ***Functional effect due to N848K mutation***

Replacement of Asn by Lys is neutrally disfavoured comparing to other polar amino acids such as Aspartate (Asp). Thus, substitution may likely decrease the pocket binding specificity, particularly with other polar amino acids. Therefore, from functional aspect this mutation has

theoretical potential of decreasing or changing the binding specificity of the PEAR1 binding involving this particular domain.

3.7 Discussion

3.7.1 Remarks on the V33M CA IX mutation

3.7.1.1 ***The function of CA IX in relation to platelet activity and CVDs***

The carbonic anhydrase (CA) IX (CA IX) (Pastorek et al., 1994; Závada et al., 2000) is a transmembrane protein, which is associated with several oncological problems and is a target of different cancer therapies (Robertson et al., 2004; Thiry et al., 2006). Its expression is regulated by two main transcription factors, i.e., hypoxia-inducible factor-1 (HIF-1) and SP-1 that bind to the promoter region of the *CA9* gene (Kaluz et al., 2003; Kopacek et al., 2005). CA IX is one of the 12 CA isozymes/isoforms, which catalyse the reversible conversion of CO_2 to bicarbonate ion and proton mainly under hypoxic and non-hypoxic conditions in different cell lines and tissue types (Alterio et al., 2012; Berchner-Pfannschmidt et al., 2004; Ihnatko et al., 2006; Yu et al., 2011). This process occurs in the different cellular compartments depending on the involved isoform (Scheibe et al., 2006; Supuran, 2008).

This protein is associated with many different physiological roles, which include changes in the intracellular calcium levels, cell adhesion, cell proliferation, and intracellular and extracellular pH regulation (Berchner-Pfannschmidt et al., 2004; Ihnatko et al., 2006; Svastová et al., 2003; Swietach et al., 2010). Moreover, CA IX and other CA isoforms are associated with different diseases and subjects of multiple proposed inhibitors/activators for therapeutic purposes (Alterio et al., 2012; Supuran, 2008).

With regard to CVD, high levels of CA IX was spotted in cancer patients with thrombosis and hypertension, and who exhibited reduced efficacy to the CA inhibitors compared to others

(Reardon et al., 2009). The same effect was observed when these inhibitors were clinically applied to assess platelet carbonic anhydrase activity targeting cytosolic CA II isoform for cardiovascular disease patients (Woodman et al., 2010). Thus, based on the above evidence, there is likely to be an unknown association between CA IX, platelet activity and CVD. Hence, the likely effect of the V33M mutation pertaining to this association is further explored below based on the results from the above generated 3D models.

3.7.1.2 ***Potential effect of the CA IX V33M mutation pertaining to ADP platelet activated responses and CVDs***

Based on the model in Figures 3.18 & 3.19, and from the structural/function perspectives, the V33M mutation is in the predicted HLH-like segment, which could be a DNA binding involved with transcription (Pastorek et al., 1994). Moreover, the predicted structure motif, which harbour the mutation could be an EF-hand motif, and thus have affinity to magnesium or calcium ion binding supporting signal transduction/signalling mechanisms among cells (Gifford et al., 2007; Grabarek, 2006). Therefore, the mutation is also more likely to affect the binding affinity of CA IX within the PG domain, which is involved with cell-cell adhesion and calcium ion transport (Svastová et al., 2003), and likely to lead to low FA platelet response activation and aggregation.

However, it is hard to establish any plausible relationship with the mechanism underpinning Ca^{2+} signalling/transport (influx/efflux) in the platelet, which is a well-known and an established process in the platelets (Dionisio et al., 2012; Jardin et al., 2012; Varga-Szabo et al., 2009). Nevertheless, it might open up further investigation on whether CA IX (and in particular this potentially HLH-like segment) has any role in the Ca2+ signalling in the platelet. This is due to the fact that among the CA IX interacting proteins are linked with nucleo-cytoplasmic ATP synthase subunits (Buanne et al., 2013), of which have been reported to be

involved with control of cytoplasmic Ca2+ in cardiomyocytes sarcolemma or sarcoplasmic reticulum (Kettlewell et al., 2009). Interestingly, CA IX has been also reported to be expressed in sarcoplasmic reticulum/t-tubules in the cardiomyocytes (Scheibe et al., 2006) and has been found to be associated with other protein complex in the heart using a mouse model (Orlowski et al., 2012). Moreover, the CA VIII isoform was reported to be expressed in the cerebellar cells and thereby modulates calcium signalling after stimulating the binding of ITP to its receptor ITPR1 within endoplasmic reticulum (Türkmen et al., 2009). With this regard, it might be worth examining whether this potentially damaging mutation has any role pertaining to Ca2+ signalling in the platelet.

Taken together, the wildtype CA IX (with V33) is likely to be involved with the increase of platelet aggregation, while the mutant CA IX decreases platelet aggregation. This suggestion may likely to tie with the results shown in the previous chapter 2, which identified rs2071676 SNP in *CA9* to be significantly associated with the decrease or low fibrinogen binding due to ADP platelet response (FA). Moreover, as the related CA II has been previously reported association with CVD, hence, it is worth investigating the likely therapeutic and clinical impact of this mutation that may underlie CVD individual health decisions.

3.7.2 Remarks on the structural and functional effect of S563N PECAM1 mutation associated with ADP platelet and CVD

3.7.2.1 ***PECAM1's functions***

PECAM1/CD31 is a glycosylated transmembrane protein of the Ig superfamily, which is expressed in the circulating platelets and other cells (Kirschbaum et al., 1994; Sun et al., 1996). Structurally, the large part of the protein is extracellularly located with 6 Ig homology units

containing 574 residues. The domain 6 in which the mutation S563N occurs is within the extracellular region.

Generally, this protein participates in several functions, which include or are related with cell-cell adhesion and migrations, cellular signalling and signal transduction, mediation of cellular interactions particularly its cytoplasmic domain using both homophilic and heterophilic interactions (Brown et al., 2002; Gong and Chatterjee, 2003).

3.7.2.2 ***Potential effect of the S563N to the ADP platelet responses and CVD***

In principle, there are three mutations of PECAM1, which have been previously shown to occur as a homozygous haplotype blocks, L125V, S563N, and R670G (LSR and VNG) in domains 1 and 6, and cytoplasmic domain respectively (Novinska et al., 2006). Furthermore, two isoforms have been identified to be occurring as heterozygous block (LSR/VNG) underpinning neutrophils cell adhesion. Based on these alleles, heterophilic binding of PECAM1 through Ig-6 domain with CD177 enhanced faster migration of neutrophils for the individuals with the LSR block comparing with those with VNG or LSR/VNG block suggesting the role of wildtype S563 in enhancing adhesion. The same study further showed that the stimulation of phosphorylation event through S563 is affected when it is mutated to N563 and hence, decrease the neutrophil migration (Bayat et al., 2010). This data may suggest the molecular mechanism underpinning the mutation in decreasing PA response.

Therefore, based on the 3D models and functional analyses, we have shown that the mutation has potential structural impact on changing the PECAM1 Ig-domain type, which may affect binding activities of its partners, leading to decrease in PA platelet response and aggregation. Hence, it might be worth further investigating the potential implications due to this mutation

underpinning targeted antiplatelet and CVD therapeutics as it was previously reported to be associated with myocardial infarction (MI) (Listì et al., 2004).

3.7.3 N848K PEAR1 mutation and its structural and functional effects associated with the platelet responses

3.7.3.1 ***The role of PEAR1***

Platelet endothelial aggregation receptor 1 (PEAR1; also known as MEGF12 or JEDI) is a membrane receptor protein, which is highly expressed in platelets and endothelial cells (Nanda et al., 2005). It is present in the resting platelets and also released from the alpha-granules during platelet activation, which increases its membrane expression (Kauskot et al., 2012). Under agonists activation and platelet aggregation, it becomes phosphorylated through Tyr-925 and Ser-953/1029, depending on the oligomerisation with αIIbβ3, which sustains the platelet aggregation (Kauskot et al., 2012; Nanda et al., 2005). The PEAR1 phosphorylation enhances signalling cascades, which culminates in binding of PI3K and form complex with, and activates αIIbβ3 for the stabilisation of platelet aggregation (Cosemans et al., 2006; Kauskot et al., 2012).

3.7.3.2 ***The structural and functional effect of N848K mutation***

Based on the generated 3D models, the substitution of Asn (N) to Lys (K) at the position 848 appears to be in the loop regions and hence somewhat little or less damaging. However, as the mutation occurs in the vicinity of the active C-terminus where the protein forms a complex with PI3K, in its signal transduction role during platelet aggregation (Kauskot et al., 2012), the mutation is likely to have a stabilising role during the formation of complexes. Since, from the functional aspect, K is positively charged and is more likely to be involved with salt bridges and form strong hydrogen bonds, which may stabilise the formed complex.

3.7.3.3 ***PEAR1 polymorphisms and potential effect of N848K mutation on the ADP platelet responses and CVD***

It is likely that the N848K PEAR1 mutation plays important molecular role by potentially stabilising PEAR1 when it forms complex with the PI3K and therefore upregulating the FA response. This in consequence activates binding of αIIbβ3, which maintains the platelet-platelet contacts and increases or sustains platelet aggregation as previously reported (Cosemans et al., 2006; Kauskot et al., 2012). Thus, this mutation merits further investigation for understanding its potential as antiplatelet therapy target.

3.8 Conclusions

This chapter investigated the molecular mechanisms of the damaging missense SNPs, which were identified by the RAPIDSNPs pipeline developed in Chapter 2. The investigation was carried out using the structural predictive approaches, which generated several alternative 3D models of the full length and subsequences of the proteins containing the mutations. These mutations were hypothesised to be affecting the related protein structures and functions that may underpin the variability of ADP platelet responses and aggregation. Three key nsSNPs were investigated, which were identified by computational predictions to be damaging. These are rs2071676 (Val33Met) of CA IX, rs12953 (Ser563Asn) of PECAM1, and rs822442 (Asn848Lys) of PEAR1 proteins. The results of the 3D models and predicted functional effects are reflective their predicted deleteriousness.

The Val33Met mutation in CA IX (carbonic anhydrase isoform IX), which was predicted to be highly damaging, is likely to affect the N-terminus CA IX PG domain, which is yet to be experimentally solved. Based on the 3D models, the mutation was identified to be likely occurring in the HLH-like structural motif (presumably EF-hand), which is more likely to be metal (presumably Ca^{2+}) binding, which may underpin its role in the regulation of signal

transduction (ion transport), in and out of the membrane. The mutation was found to potentially decrease the platelet aggregation in response to fibrinogen binding mediated by ADP (PA). The mutant is likely to affect the ion binding affinity of CA IX in the potential HLH motif region and affect the ion transport with its interacting partner(s), and thus, likely to decrease the platelet aggregation. Moreover, CA IX is likely to associate with CVD, and thus, the mutation merits further investigation for understanding its potential for targeted antiplatelet therapeutics for the CVD.

Furthermore, the S563N mutation of PECAM1 appeared to be changing its Ig-domain 6 structure from C- to V-type. Based on the follow up analyses, we have shown that this mutation may likely to affect the crucial binding partners of PECAM1 which may likely to lead to downregulation of PA platelet response and aggregation. This mutation has been found to be associated with CVD (Listì et al., 2004; Pamuk et al., 2014; Sahebkar et al., 2013). Thus, based on the 3D structural results and functional analyses, it might also be worth investigating the potential of this mutation for CVD targeted antiplatelet therapy.

On other hand, the rs822442 (Asn848Lys) mutation of PEAR1 appears to be structurally less damaging. However, in theory, the substitution to Lysine (i.e. to 848K) from asparagine (N) has a potential stabilising role to the structure in the vicinity of the C-terminus where other PEAR1 mutations participate in the phosphorylation events, which are important in sustaining platelet aggregation. Moreover, several PEAR1 polymorphisms are investigated their association with the inter-individual variability of the antiplatelet therapy (Lewis et al., 2013; Würtz et al., 2014).

Taken together, the results above demonstrate the similarity of the results between the predicted potential molecular consequences due to missense SNPs underpinning the ADP platelet response and identified effect of the same SNPs on the ADP platelet responses in Chapter 2.

This may further strengthen our confidence in use of the RAPIDSNPs in identifying crucial disease/trait associated SNPs. Moreover, these SNPs have shown association or potential association with CVD as discussed in the previous chapter and have been further explored in this chapter. Thus, the predicted molecular consequences from structural/functional perspectives could be a marker for genetic risks associated with CVD and potentially applied in targeted therapeutics and clinical personalised medicine (PM).

Chapter 4 - Predicting the regulatory roles of key SNPs from the RAPIDSNPs

4.0 Abstract

Determining the functional consequences of the complex trait/disease-associated non-coding, intronic, and synonymous SNPs identified by genetic association studies (GASs) presents a challenge. Nevertheless, it is an indispensable task for a more complete understanding of molecular mechanisms underlying the complex trait/disease. Many of these non-coding SNPs may have potential regulatory effects at the genome level, such as perturbing the binding regions of the transcription factors, which regulate gene transcription. In this chapter, the identified non-coding, intronic, and synonymous key SNPs from the described RAPIDSNPs approach are hypothesised to be potentially involved in the regulation of the molecular functions, which may further underlie the variability of the ADP platelet response. Experimental procedures to determine the SNPs' regulatory roles are laborious and expensive and initially theoretical approaches should always be explored. Thus, this chapter describes an alternative bioinformatics driven approach for analysing the potential of these key SNPs to be regulatory (i.e. rSNPs). The predicted regulatory mechanisms include: expression quantitative loci (eQTLs), regulation of transcription factors and RNA binding affinity with bound proteins, long range interactions, and chromatins and histone modifications.

Based on the bioinformatics approach, several key SNPs, were predicted to be involved in several regulatory roles. To mention afew, the regulatory SNPs include: rs3730051 in the intron of *AKT2* and rs6141803 in the intergenic region and proximal to *COMMD7*. These rSNPs are likely to be involved with: eQTL of several related genes, transcription regulation (transcription factor binding affinity) and long range interactions targeting other genes/loci, and RNA binding affinity. Thus, these molecular roles, which are likely to be regulated by or involving the rSNPs, may further contribute to the individual ADP platelet responses variability (increase or decrease

of FA/PA responses). Moreover, several of the identified rSNPs appears to have potential associations with CVD.

4.1 Introduction

4.1.1 Why regulatory genomic variants?

As explored in Chapter 3, changes in the coding regions of the genes may result in structural and functional changes of the resulting proteins at the molecular level, which may further elucidate the underlying cause of the inter-individual phenotypic variation of complex traits/diseases. The previous chapter described the effect of the missense SNPs in the coding region, which are likely to be associated with structural and functional changes of the key proteins involved with ADP platelet responses. Furthermore, these molecular changes were likely to contribute to the inter-individual variations, which underpin the ADP platelet responses and CVD risks

Despite the vital role of the missense SNPs, the individual phenotypic variations susceptible to diseases and drug responses have been further associated with changes in the individual's gene expression (GuhaThakurta et al., 2006). These changes of gene expression are largely caused by the SNPs present in the regulatory regions of the genes, which may include cis-regulatory elements or cis-expression quantity loci (cis-eQTL) (Franke and Jansen, 2009; GuhaThakurta et al., 2006; Wang et al., 2005) and several of these occur in the promoter regions of the genes (Stepanova et al., 2006). Additionally, SNPs are present in the transcription factor binding sites, such as distal regulatory elements (Heintzman and Ren, 2009). These SNPs are known as regulatory SNPs (rSNPs), and have been reported to be common in 50% of all genes and it is highly likely that all genes within human population have rSNPs (Buckland, 2006). A majority of these rSNPs are suggested to lead to the inter-individual responses to exposure and susceptibility to disease/complex trait in allele specific manner (Wang et al., 2005).

Therefore, studying and understanding these regulatory variants are vital for uncovering their regulatory effects on the individual phenotypic traits/diseases and drug responsiveness (Buckland, 2006; Franke and Jansen, 2009; Prokunina and Alarcón-Riquelme, 2004). However, the key problem that arises is the ability to identify these crucial rSNPs, and thus, the design of different computational methods is an ongoing task (Hudson, 2003; Wang et al., 2005).

In light of the above, it is postulated that some of the identified key SNPs from the RAPIDSNPs approach in the Chapter 2, are potentially rSNPs and are thus likely to be involved with regulatory mechanisms that may underlie ADP platelet response variability. Furthermore, this chapter aims to examine the regulatory mechanisms by designing bioinformatics analyses pipeline, which may predict potential regulatory roles that are likely to be associated with the key SNPs identified in Chapter 2.

This chapter begins by describing different regulatory mechanisms in which the rSNPs are likely to be involved. This may include the regulation of: the eQTL, transcription factors binding affinity in the proximal and distal regions of the targeted genes with long range interactions, RNA binding affinity of the related proteins, and chromatin and histone modifications.

4.2 Understanding the regulatory mechanisms of the rSNPs

4.2.1 SNPs in the eQTL genomic regions

Expression quantitative trait loci (eQTLs) are parts of the genome accommodating DNA sequence variants that influence the expression levels of one or more genes. The eQTLs may help us to understand the biological mechanisms in which the causal variants, or SNPs identified from the GASs, influence the population individuals' traits or diseases (Albert and

Kruglyak, 2015). The regulatory SNPs are able to affect individuals by interfering at any point in time during the entire gene expression process. This influence could be measured through the differential mRNA, or protein abundance of individuals given gene(s) (pQTL) (Albert and Kruglyak, 2015; Liu et al., 2016). Figure 4.1 illustrates the underlying mechanism in which the regulatory variants or SNPs influence the gene expression levels and pQTL, leading to high/low disease or trait risks in individuals.

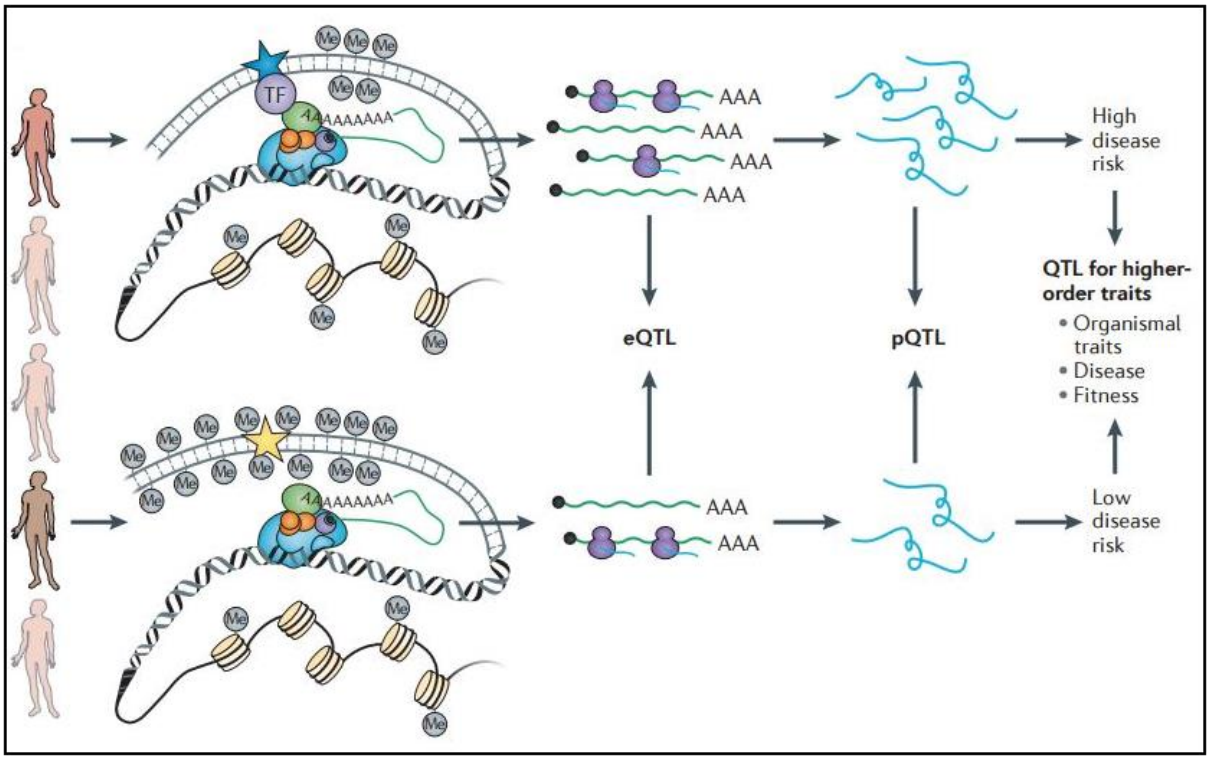


Figure 4.1 The results for the eQTL or pQTL study for the two individuals in the population that differ gene expression. Depending on the study interest, the mRNA or protein abundance is measured in each individual. The loci involved is marked by the star. This molecular variation is compared to the genetic variation among the individuals to determine the significance of the mRNA abundance or eQTL using the association or linkage analysis. The allele with high mRNA transcript levels determine the individual with high risk of the disease or trait in case of the particular model system. (The image was taken from Albert and Kruglyak, 2015).

Recently, eQTLs have been reported to be heritable across different tissues and model systems (Albert and Kruglyak, 2015; Powell et al., 2012). As ADP platelet responses are heritable, it might be essential to identify and understand the rSNPs involved with eQTLs, which are likely to contribute to the ADP platelet responses variability. However, the identification of these crucial variants is still a major challenge (Albert and Kruglyak, 2015; Hudson, 2003).

4.2.2 rSNPs in the Transcription Factor (TF) binding sites (TFBS)/Cis-regulatory elements (CREs) and involved with TF binding regulation

The influence of rSNPs may also be examined through the differential TF binding sites regulation, (Wittkopp and Kalay, 2012). TF binding sites (TFBS) are more likely candidate regions as they have potential to harbour functional rSNPs. TFBS or *cis*-regulatory elements (CREs) are genomic regions or sequences in close proximity (100 – 1000 bps) to a target gene and consist of short nucleotides (<30 bps) where a TF protein binds to regulate the transcription of the target genes. The mutations or SNPs that affect the function of these sequences may lead to phenotypic variability (Brown and Feder, 2005; Savinkova et al., 2009; Wittkopp and Kalay, 2012). Moreover, the TF binding regulation in CREs could be affected by the rSNPs or mutation in the intron of genes (Bianchi et al., 2009; Liao et al., 2013). Hence, identification of the rSNPs in these regions is vital in understanding their participation in the regulation of the TFs that lead to phenotypic changes. Thus, a further aim is to find whether the identified SNPs from the RAPIDSNPs are potentially located in the TFBS or intron regions in regulating transcription, and hence, likely to contribute to the ADP platelet response variability.

For instance, there are number of possible scenarios that might occur at the molecular level due to the presence of the rSNPs in the TFBS, as explained by Chorley et al. (2008): 1) The rSNP in the TFBS might not have any effect on the binding interaction with the TF and hence cannot change the expression. 2) The rSNP may increase, or decrease binding, which may result to allele-specific gene expression. 3) The rSNP may completely remove the existing natural binding site or produce a novel binding site and thus the gene regulation will no longer supported by its original TF. Each of these scenarios indicate that rSNPs in TFBS may have key functional roles with potential to be instrumental in driving differential gene expression

and resultant phenotypic effects (Chorley et al., 2008). Figure 4.2 demonstrates different effects that the rSNPs may exert on gene expression.

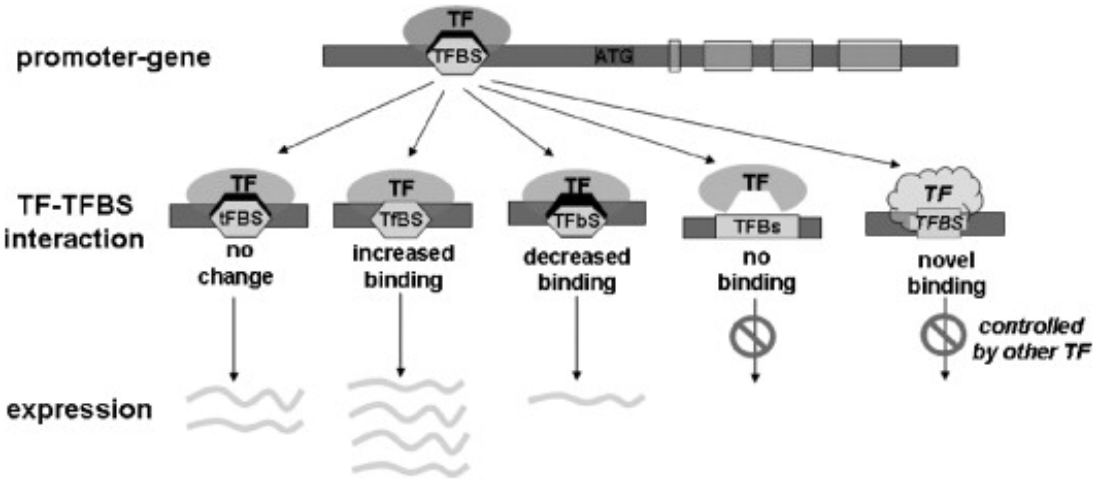


Figure 4.2 The possible consequence of rSNP in TF binding site (TFBS). On most occasions the rSNP is unlikely to disrupt the TF binding ability or gene expression, because there is allowable variation in the consensus sequence of the binding site. It is also possible that, in some occasions, allele-specific gene expression will occur as result of the presence of the rSNP in the binding site, which may increase or decrease the TF binding. In rare occasions the rSNP may cause complete removal of the original binding site or lead to the generation of a novel binding site, and hence the gene will not be regulated by the original TF. (The image was taken from Chorley et al. 2008).

Thus, it could be argued that the rSNPs in the TF binding sites are the key factors, which lead to the differential TF binding that eventually direct the differential mRNA expression and eQTL (Albert and Kruglyak, 2015).

Therefore, identifying rSNPs that have potentially to be located within these sites might be vital in complimenting the understanding of the eQTL, which may ultimately contribute to the variability/heritability of the disease/trait risk (Albert and Kruglyak, 2015; Chorley et al., 2008). Indeed, the presence of the rSNPs in TF binding sites has been previously reported to be associated with myocardial infarction (MI) and other cardiovascular diseases (Nakamura et al., 2002; Savinkova et al., 2009).

Although some methods and tools for identifying the rSNPs in TFBS exist (Li et al., 2015), there remains a need for further development of techniques for identification of potential regulatory SNPs in TFBS (Chorley et al., 2008; Ward and Kellis, 2012a).

4.2.2.1 ***rSNPs in the proximal and distal regulatory regions***

The rSNPs may also be altering and affecting the nearby (proximal) or remote (distal, i.e. enhancer, silencer, locus control regions (LCR)) regions of promoters (Bryzgalov et al., 2013; Bulger and Groudine, 2010; Chorley et al., 2008). The rSNPs in these regions may change their TFs binding and result in differential transcription regulation (Bryzgalov et al., 2013; GuhaThakurta et al., 2006; Heintzman and Ren, 2009; Hoogendoorn et al., 2003; Li et al., 2015; Maurano et al., 2012). In addition, the presence of rSNPs in these regions has been previously reported to be associated with the disease including CVD events (Drachkova et al., 2011; Koivisto et al., 1994; Musunuru et al., 2010). Therefore, identification of the rSNPs associated with these regions is of high importance (Bryzgalov et al., 2013). Several experimental and computational approaches have been designed for the identification of rSNPs in these regions (Gerstein et al., 2012; González et al., 2012; Hallikas et al., 2006; Maurano et al., 2012; Stepanova et al., 2006). However, these efforts are yet to be able to fully identify these elusive genetic variants in these regions on a genomic scale (Albert and Kruglyak, 2015; González et al., 2012; Li et al., 2015).

4.2.3 rSNPs involved with chromatin and histone modifications

The rSNPs might be involved in the chromatin DNA structure leading to chromatin marks or histone modifications, in which case they are likely to be affecting gene transcription and associated with the complex trait/disease in allele-specific manner (Parker et al., 2013). These rSNPs might be part of the chromatin state, which are profiles used for detecting the regulatory

activity of non-coding parts of the genome responsible for intermediate cellular phenotypes and diseases (Ernst et al., 2011). Moreover, such profiles enable us to elucidate possible interactions between promoters and enhancers that might be influenced by 3D chromatin configuration affected by the presence of non-coding rSNPs (Mora et al., 2015). The 3D DNA chromatin conformation may enhance the spatial interactions or long range interactions between different *cis*-regulatory elements, such as enhancers and promoters, which dictate the gene transcription in cell-specific manner (Ernst et al., 2011). These interactions may involve chromatin state and histone modifications, and affect the binding affinity of various TFs, and so they are likely to regulate transcription of variety of genes (Schierding et al., 2014; van Steensel, 2011). Therefore, the presence of non-coding SNPs or loci from different enhancers and promoters might come together as a result of the 3D packing of chromatin, influencing differential gene transcription regulation (Ernst et al., 2011).

As platelets contain only RNA, which are originated from the megakaryocytes cell types (Gieger et al., 2011; McRedmond et al., 2004), the chromatin states inside the megakaryocytes, might have possible effect on the genes transcription underlying the platelet functions (Weyrich et al., 2009). Hence, identifying and understanding the potential non-coding SNPs, which may associate chromatin states and histone modifications might be vital in understanding and predicting the differential gene transcription that is likely to affect the platelet functions.

4.2.4 rSNPs in the RNA-binding protein sites (RBPS) or motifs

RNA binding proteins (RBPs) form complexes with the RNA molecules in the cell to regulate their structures and interactions. RBPS are vital in mediating and altering various regulatory processes that are involved during post-transcription, such as alternative splicing, and polyadenylation (Glisovic et al., 2008). RBPS are specific binding sites whose main function is to give instructions to or control RBPs (Glisovic et al., 2008; Paz et al., 2014). It has been

found that these sites may contain SNPs in either exon or intron, which may affect the splicing events or differential gene expression, resulting in abnormal transcript levels (mRNA levels (Li et al., 2015; Zhao et al., 2013). In addition, several of these SNPs have been reported to be disease associated variants, including those in the non-coding RNAs (Fraser and Xie, 2009; Makrythanasis and Antonarakis, 2013). In theory, the presence of the rSNPs in the RBPS may lead to the abnormalities in the transcript abundance levels of the mRNAs among individuals and may help to explain differential responses to the particular complex trait/disease or treatment. However, identifying or predicting these variants and their sites of occurrence still remains a major problem (Li et al., 2015; Paz et al., 2014; Zhao et al., 2013).

Therefore, the major aims of the designed bioinformatics pipeline is: - to identify the potential key SNPs (rSNPs) from RAPIDSNPs that are likely to i) contribute to the heritability of eQTLs, ii) be in TFBS of the genes and whether they are in the proximal or distal of the core regulatory regions, iii) be involved in the chromatin and histone modifications including long-range interactions, and iv) be occurring in RNA binding sites. In the latter aim, the approach intends to identify certain motifs, which are likely to be in the RBPS and that may contain rSNPs where the proteins bind. Thus, these regulatory roles predictions may further explain the underlying molecular mechanisms and perturbations leading to ADP platelet responses variability and provide improved strategies for targeted CVD therapeutics.

4.3 Methods

4.3.1 Inputs used: key SNPs from three subsets

The major inputs to the developed bioinformatics pipeline are three different key SNPs sets, which were obtained from the RAPIDSNPs run on each of the three subsets of dataset named as dataset 1, dataset 2, and dataset 3 as described in the Chapter 2. In principle, these datasets came from the original sample cohort from the study by Jones et al (2009). The imputation

appraoch employed in Chapter 2 in dealing with large number of missing genotypes led to have the above three subsets. Further details are in Chapter 2.

4.3.2 Bioinformatics pipeline for regulatory SNPs identification

A bioinformatics pipeline was designed for analysis and identification of regulatory SNPs. Different regulatory roles were examined, including: eQTL, transcription factor binding sites in the *cis*-regulatory regions in the promoters or enhancers (proximal/distal) regulatory regions, RNA binding sites affinity, and chromatin and histone marks or modifications. Figure 4.3 shows the flow of methods and data in the pipeline.

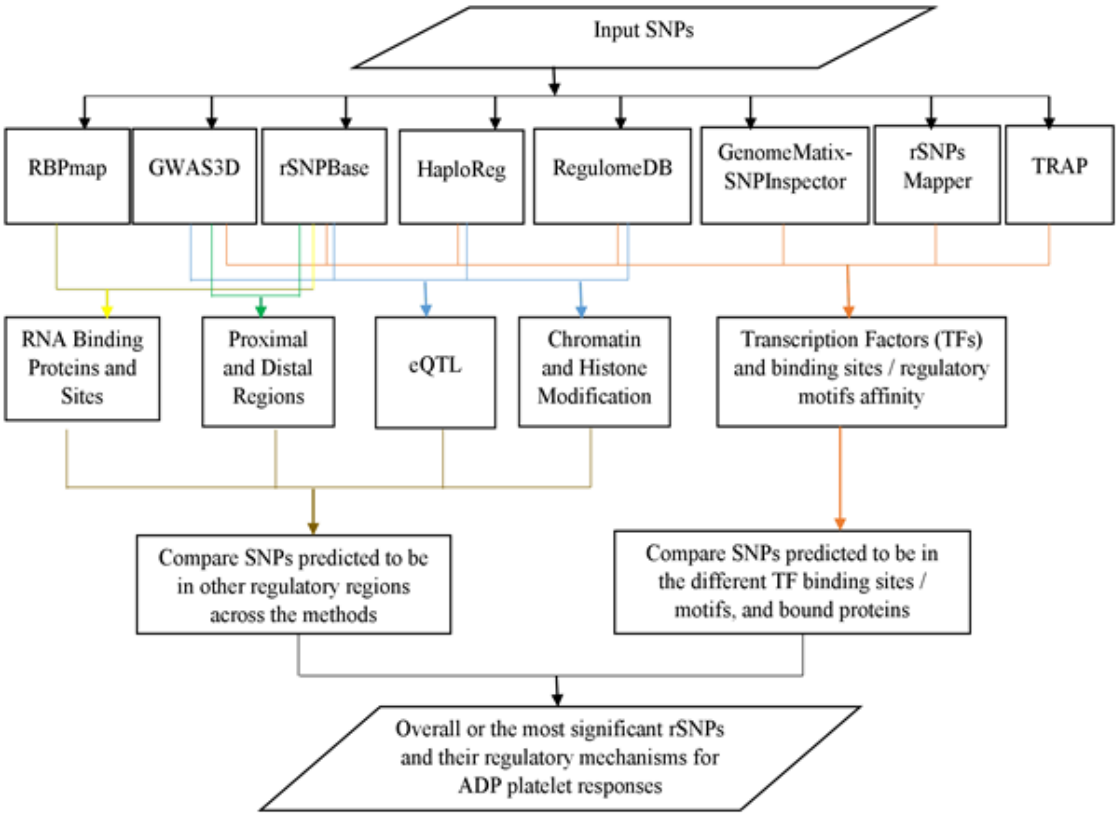


Figure 4.3 Bioinformatics analyses pipeline for investigating and identifying the potential rSNPs from the RAPIDSNPs that are likely to contribute to the differential ADP platelet responses. The input SNPs formats might be either in the dbSNP (i.e. with 'rs') or DNA sequence. The top layer after the input SNPs are the regulatory roles to be examined. The next layer contains selected methods, which are applied to the input SNPs. Different methods may commonly predict a SNP to have the same regulatory role(s). The SNPs, which have been predicted by many methods in the same role may indicate their increasing likelihood of being highly regulatory. The arrows colour legends indicate different regulatory roles, which were predicted by the different selected methods. The yellow, green, blue, and red arrows signify predictions of RNA binding sites, proximal and/or distal regulatory regions, chromatin and histone marks/modifications, and transcription factors binding sites/motifs binding affinity regulatory mechanisms respectively.

4.3.3 Detail description of the pipeline

For examining the involvement of the SNPs in the regulation of the differential eQTL and chromatin or histone marks, the rSNPBase (Guo et al., 2014), HaploReg (Ward and Kellis, 2012b), RegulomeDB (Boyle et al., 2012), and GWAS3D (Li et al., 2013) methods were used.

For predicting the likelihood of SNPs of co-localising and influencing the binding affinity of the TFs or regulatory motif and bound TF, a compendium of methods was used. These methods included HaploReg, RegulomeDB, GWAS3D, and rSNPBase in addition to TRAP (Thomas-Chollier et al., 2011), GenomeMatix-SNPInspector (Cartharius et al., 2005), and rSNPMapper

(Riva, 2012). Moreover, for predicting proximal/distal regulatory regions the rSNPBase and GWAS3D methods were used. In addition, GWAS3D was used to identify rSNPs that were likely to be participating in the long range interactions with other loci or chromosomal regions. In this case, GWAS associated SNPs, in either of HapMap and 1000 Genomes LD were alternatively referenced. Hence, in predicting likely distal or long range interactions, there were likely to be two alternatively predicted rSNPs for FA or PA platelet responses. Moreover, the TRAP method was more appropriate, since the predicted binding sites where the SNP might be occurring was indicated in a sequence specific manner, which was useful when maping the actual genomic location of the SNP.

Furthermore, in order to predict whether the SNP was within an RBPS, the rSNPBase method was used in addition to RBPmap (Paz et al., 2014), which also indicated the likely binding proteins. The stringency level for selecting the significant binding sites was set to medium (i.e. p-values <0.005 and <0.01). Both the rSNPBase and RBPmap methods use different public databases; the former uses ENCODE (Consortium, 2004) and the latter uses motifs and bound proteins derived from the literature and different *in vivo* experimental data (Akerman et al., 2009; Paz et al., 2010). All of the above methods were selected based on their performance benchmarks, which were reported by Li et al. (Li et al., 2015).

4.3.3.1 ***Statistical significance of the identified regulatory SNPs***

The underlying significance of the identified SNPs to be regulatory was determined by each of the individual method's scoring measures. Generally, in determining the rSNPs significance for TFBS affinity, many methods used a position-weight matrix (PWM) scoring mechanism (Stormo and Fields, 1998). Furthermore, chi square with Fisher's exact tests were used to determine the significance of the rSNPs identified by the TRAP method. For the GWAS3D method, the combined p-value (cp) is used. Cp includes the calculated p-values for the GWAS

signal, the binding affinity effect on TF, and the genomic evolutionary conservation, in order to determine the significance of overall regulatory role of each rSNP (Li et al., 2013).

4.3.4 Identifying the target genes of the predicted TFs that are likely associated with rSNPs and ADP platelet responses

Several approaches were taken in order to identify the likely target genes of the predicted TFs, which are associated with the identified rSNPs. Firstly, the PAZAR method (Portales-Casamar et al., 2009) was used with the downloadable file of the target genes (http://www.pazar.info/cgi-bin/downloads_csv.pl), secondly, the enrichment analyses tool 'Enrichr' (Kuleshov et al., 2016), thirdly the TRED (Jiang et al., 2007) method, and fourthly the FANTOM5 gene sets (Marbach et al., 2016) were used. Additionally, the R package 'tftarget' (https://github.com/slowkow/tftargets) was used to extract the likely related TF's target genes from FANTOM5 gene sets. The related TF target genes were selected based on the genes' likely association with the platelet functions from the literature.

4.4 Results

Several of the intronic, and other non-coding key SNPs, which were identified by the RAPIDSNPs method in all three datasets are likely to be regulatory (i.e. rSNPs). These predicted rSNPs are categorised in the follow sections based on their regulatory functions.

4.4.1 Predicted rSNPs that are likely to be involved with eQTL

Several PA and FA responses associated SNPs from RAPIDSNPs were identified to be associated with eQTL and hence, are likely to regulate expression levels of other related genes, which potentially, may lead to the variability of the ADP platelet responses. Table 4.1 shows the selected predicted rSNPs and their likely differential expressed genes (i.e. eQTL genes) from different datasets and their associated ADP platelet responses.

Data	ADP platelet responses	SNP ID	Associated locus /gene	Method	Related eQTL genes
Dataset 1	PA	**rs6141803**	*COMMD 7*	rSNPBase	*CPNE1*
				HaploReg	*COMMD7*
Dataset 2	PA	**rs6057638**	*chr:20,31 339332*	rSNPBase	*CPNE1*
Dataset 3		rs246406	*ITGA2*	HaploReg	*ITGA2*
Dataset 1	FA	**rs1491978**	*P2RY12*		*P2RY12*
		rs7034539	*JAK2*		*INSL6*, *JAK2*
Dataset 2	FA	**rs12485738**	*ARHGEF 3*	rSNPBase	*CD226*, *ADCY3*, *ADCY6*, *ARHGEF12*, *PEAR1*, *ARHGAP21* *TLN1*, *VCL* ,*VWF*
				Regulome DB	*ADCY3*, *ADCY6*, *ARHGAP21*, *ARHGEF12*, *CD226*,*CD9*,*CETP*,*TLN1*, *VWF*
		rs3729931	*RAF1*	HaploReg	*PPARG*, *RAF1*
Dataset 3	FA	**rs17204376**	*GPR87/ MED12L*		*MED12L*

Table 4.1 The most interesting SNPs from the RAPIDSNPs method, which are associated with the ADP platelet responses and that have been identified to be likely regulatory (rSNPs) and involved with eQTL. Green highlighted cells contain SNPs, which have been predicted by more than one method and are likely to be regulating several differentially expressed genes. The yellow highlighted eQTL genes are those which are known to be associated with ADP platelet responses. The red highlighted are the eQTL genes, which are associated with CVD. In bold are the newly discovered significant SNPs, which were not previously identified in the study by Jones et al. (2009).

From Table 4.1 some of the predicted eQTL genes such as *PPARG* are also transcription factors (next section) and have been associated with CVD as are discussed further in section 4.5.

4.4.2 Predicted rSNPs involved with transcription factor binding sites (*cis*-regulatory elements) regulation

Several SNPs are identified to be likely located in the TFBS or introns and potentially involved in the regulation of the transcription factors, or proteins that bind to the core transcription machinery of the various genes. In addition, several of these identified transcription factors are likely to be involved in the underlying regulation of ADP platelet responses.

Tables 4.2 and 4.3 show the selected identified interesting rSNPs that are likely to be involved with the regulation of the TFBS or *cis*-regulatory elements for PA and FA platelet responses respectively.

Dataset	rSNP	Gene/locus	Method	Identified TF or/and regulatory motifs	Significance if applicable
Dataset 1	**rs3212391**	*ITGA2*	SNPInspector	GATA1	0.955, 0.900
			rSNPs Mapper	PPARG	2.4
			TRAP	PARP	0.00872
			GWAS3D (HapMap)	STAT1	3.1665E-05
	rs6141803	*COMMD7*	RegulomeDB	GATA1, GATA2, CEBPB, STAT1, STAT5A	N/A
			HaploReg	GATA2	N/A
			GWAS3D (1000 Genomes)	GATA-1,	5.7800E-03
	rs8033381	*CSK*	RegulomeDB	CREBBP	N/A
			rSNPBase	Pol2, STAT5A,	N/A
	rs2300065	*SKP1*	SNPInspector	PPARG,	0.862
			rSNPBase	Pol2, GATA2	N/A
	rs3730051	*AKT2*	SNPInspector	PAX5	0.807
	rs6442896	*ITPR1*	TRAP	STAT5A	0.000691
Dataset 2	rs1472122	*P2Y12/GPR87*	RegulomeDB	CEBPB	N/A
			rSNPBase	Pol2, SP1, ELF1, STAT3	N/A
	rs246406	*ITGA2*	GWAS3D	STAT5A	2.2608E-05
	rs3788337	*GNAZ*	rSNPBase	Pol2, SP1, SP2, CEBPB	N/A
			GWAS3D	STAT5A	2.5912E-05
	rs6057638	*20:32751526*	RegulomeDB	GATA2, GATA1	N/A
	rs2228671	*LDLR*	RegulomeDB	POLR2A	N/A
	rs2815805	*MAPK14*	RegulomeDB	GATA1	N/A
	rs17229705	*VAV3*	GWAS3D	STAT5A	2.5796E-05
	rs5277	*PTGS2*	rSNPBase	Pol2	N/A
Dataset 3	**rs41305276**	*THBS1*	rSNPBase	Pol2	N/A
	rs2289171	*PIKFYVE*	HaploReg	MAFK	N/A
			RegulomeDB	MAFK	N/A
	rs906766	*MED12L*	RegulmeDB	MAFK	N/A

Table 4.2 The most interesting identified rSNPs, which are associated with the PA response and are likely to be involved with the binding affinity of TFs and hence transcription regulation. The green highlighted SNPs are those which have been predicted by more than one method to be highly regulatory in TF binding. Several of these TFs were commonly predicted across methods. In bold are the newly discovered significant SNPs, which were not previously identified in the study by Jones et al. (2009).

Dataset	rSNP	Gene/locus	Method	Identified TF or/and regulatory motifs	Significance if applicable
Dataset 1	**rs1491978**	*P2RY12*	SNPInspector	NFAT	0.954
	rs41282607	*MAPK1*	SNPInspector	CEBPE	0.746
			RegulomeDB	ELF1	N/A
			GWAS3D	CEBPB	9.6400E-03
	rs3736101	*MADD*	rSNPBase	Pol2	N/A
Dataset 2	**rs12485738**	*ARHGEF3*	RegulomeDB	FOXA1	N/A
	rs6450105	*ch5:52980479*	SNPInspector	PPARG, CEBPB	0.836, 0.951
	rs11264579	*PEAR1*	TRAP	PAX6	0.00683
	rs7180408	*GTF2A2*	SNPInspector	GATA1	0.973
			HaploReg	FOXA1	N/A
			TRAP	GATA1	0.00585
	rs3729931	*RAF1*	HaploReg	POL2	N/A
			GWAS3D	STAT1	4.9268E-05
Dataset 3	**rs17296289**	*10:32971771*	HaploReg	GATA1	N/A
			RegulomeDB	GATA1, GATA2	
			TRAP	PPARG, PPAR, GATA2,	0.000133, 0.0024, 0.00732,
	rs2290159	*RAF1*	rSNPBase	Pol2	N/A

Table 4.3 The identified rSNPs, which are associated with the FA response and likely to be involved with the binding affinity of TFs and hence transcription regulation. The green highlighted SNPs are those which have been predicted by more than one method to be highly regulatory in TF binding. In bold are the newly discovered significant SNPs, which were not previously identified in the study by Jones et al. (2009).

From Tables 4.2 and 4.3 the selected identified rSNPs are most likely to be involved with TF binding regulation in which the bound proteins are more likely to be associated with ADP platelet responses related genes and CVD diseases. (See Discussion section). Moreover, some of the predicted TFs such as GATA1 and PPARG are common across different methods and found to be related with different rSNPs.

4.4.2.1 ***Identified rSNPs that are in the binding sites based on the TRAP method***

In contrast to other methods, the TRAP method also provides the results in the sequence specific manner in which it is possible to further examine the SNPs that may likely to occur in the TFBS. Thus, several predicted SNPs, which appear to be occurring in the TFBS and that are associated with the PA and FA responses were closely examined as they are likely to regulate the TFs and potentially contribute to the variability of ADP platelet responses.

The following are rSNPs, which were found to be likely to occur in the TFBS. Each SNP is shown with its corresponding sequence in FASTA format, which contains the SNP location, indicated in bold. This is followed up with the similar sequence from TRAP, which contains the predicted binding sites (coloured red and in bold). The SNP is shown in black bold font with an upper case letter where it is near the predicted binding site. The SNP is shown in bold red, if it is found within the predicted binding site. Each SNP ID is followed by its corresponding locus or gene name. The italicised SNPs are those which were previously identified by Jones et al. (Jones et al., 2009).

Dataset 1-PA

>rs8033381(*CSK*)

CCCACCTGTCCATTTTCAGGGTGTC**[A/G]**TCTGTCCTGCACAAGGAAAGGTGGG

cccaCCTGTCCAtttcagggtgtc**G**TCTGTCCTgcacaaggaaaggtggg

>rs3212391(*ITGA2*)

GAGGGAAAGAAAGCAGAGGTATGGA**[A/T]**AAAGGTACCTCCCATCCTCCAGAGT

gagggaaagaaagcagaggtatGGA**T**AAAggtacctcccatcctccagagt

Dataset 2-PA

>*rs3788337*(*GNAZ*)

GAAGGATTGGTGAGACAATCCAGGA**[A/C/G]**AGCAGGTGCCACTAGCCCCACTTTG

gaaggattggtgagacaatccagga**G**AGCAGGTgccactagccccactttg

Dataset 3-PA

>rs2289171(*PIKFYVE*)

CACATCTAAAATGAAAAATAGTCTA**[C/T]**TTATATATTAATATTCACAGAGAAT

caCATCTAAAATGAAAAATAGTCTA**T**TTATATATTAATATTCACAGAgaat

Dataset 2-FA

>rs7180408(*GTF2A2*)

AAAAGCATTTGTATTGGCTTTCCTA**[A/T]**CTGCCTGAATGCTCTTCCTGATATT

aaaagcatttgtattgGCTTTCCTa**A**ctgcctgaatgCTCTTCCTgatatt

>rs7739455(*CD109*)

CAGCTGTAGACGGTTCATAAACAAA**[A/T]**GAGCATGGTTGTGTACCAGCAGAAC

cagctgtagacggttcATAAACAAA**A**GAgcatggttgtgtaccagcagaac

>*rs11264579*(*PEAR1*)

GGCACCTGAACTAGACCTTGAAAGA**[C/T]**GGGAACTCCAGGTGAAGAATGAGAC

ggcaccTGAACTAGACCTtgaaaga**A**gGGAACTccAGGTGAagaatgagac

>*rs41305896(ITGA2)*

CCTACCTAGCATGAAGAAAGAACCA**[C/T]**TTCTTACCGCACAGGGTTCGAAAGT

cctacctagcatgaagaaagAACCA**T**TTCTtaccgcacagggttcgaaagt

>rs41307142(*GAS6*)

CCCAGATCTAACCTGGACAGGCTGG**[C/G]**GTTTCTGGTAGTGAATGCGGAAGAG

cccagatctaacctggacaggctgG**G**GTTTCTGgtagtgaatgcggaagag

Dataset 3-FA

>rs17296289(*chr10:32971771*)

ACCACAGAGATCAAAGGGCAAAAGA[A/G]TAAAAGCTAAGAGAAAAAACTAAAG

ccacaGAGATCAAAGGGCAaaaga**A**taaaagctaagagaaaaaactaaag

>rs17204376(*MED12L*)

TTACATTCTGCTTTGGGTGGATTAA[**G/T**]TATGTTACGGAAAAGTAGCTAGTCC

ttacattctgctttgggtgGATTAA**T**Tatgttacggaaaagtagctagtcc

4.4.2.2 ***Predicted rSNPs that are in the proximal and distal regulatory regions***

Several SNPs were predicted to be involved with the regulation of the proximal or distal regions related to several genes. Tables 4.4 and 4.5 show the identified rSNPs, which are associated with the PA and FA platelet responses and are likely to be located in the distal and proximal regions as detected by the rSNPBase method.

Dataset	SNP	Locus	Proximal/Distal Regulation	Related Genes
Dataset 1	**rs2300065**	*SKP1*	Proximal	*SKP1*
			Distal	*SKP1*
	rs3730051	*AKT2*	Proximal	*AKT2*
			Distal	*AKT2*
	rs6141803	*COMMD7*	Distal	*COMMD7, DNMT3B*
	rs8033381	*CSK*	Proximal	*CSK*
			Distal	*CSK*
Dataset 2	rs5277	*PTGS2*	Proximal	*PTGS2*
	rs1472122	*P2YR12*	Proximal	*P2YR12*
	rs2769668	*VAV3*	Proximal	*VAV3*
	rs2815805	*MAPK14*	Distal	*MAPK14*
	rs3788337	*GNAZ*	Proximal	*GNAZ*
	rs17229705	*VAV3*	Proximal	*VAV3*
			Distal	*VAV3*
Dataset 3	**rs906766**	*MED12L*	Distal	*MED12L*
	rs41305276	*THBS1*	Proximal	*THBS1*
			Distal	*THBS1*

Table 4.4 The identified rSNPs which are associated with the PA platelet response and are likely to be involved with proximal or distal regulation. The proximal and distal regulation might be involved with the chromatin/histone modifications in the enhancer or promoter regions. In bold are the newly discovered significant SNPs, which were not previously identified in the study by Jones et al. (2009).

Dataset	SNP	Locus	Proximal/Distal Regulation	Related Genes
Dataset 1	**rs3736101**	*MADD*	Proximal	*MADD*
	rs41282607	*MAPK1*	Proximal	*MAPK1*
			Distal	*MAPK1*
	rs41305272	*MAP2K5*	Proximal	*MAP2K5*
Dataset 2	rs3729931	*RAF1*	Distal	*RAF1*
	rs11264579	*PEAR1*	Proximal	*PEAR1*
Dataset 3	rs2290159	*RAF1*	Proximal	*RAF1*
			Distal	*RAF1*
	rs4792219	*MAP2K4*	Proximal	*MAP2K4*

Table 4.5 The identified rSNPs which are associated with the FA platelet response and are likely to be involved with proximal or distal regulation. The new identified SNPs are in bold.

4.4.2.3 ***Predicted rSNPs associated with distal long range interactions and other transcription regulatory elements based on the GWAS3D method***

In contrast to other methods, the GWAS3D enables the researcher to closely analyse the rSNPs, which are likely to be involved in the long range interactions in a 3D manner using circos-like plot. Based on the method or tool, the plot displays variants in a genome-wide manner. Thus, the detected rSNPs from the RAPIDSNPs may be displayed with other genomic or chromosomal interactions related with other rSNPs from different LDs as per 1000 Genomes or HapMap data. In addition, the detailed results of other regulatory functions associated with the key SNPs are shown in the tabular manner with their significance. Therefore, a couple of the rSNPs in each dataset were predicted to be likely involved with distal and long-range interactions. These SNPs are represented using the circos-like graphs in Figures 4.4, 4.5, and 4.6, which show the PA platelet response associated SNPs for the dataset 1, 2, and 3 respectively. For the FA platelet response, the predicted rSNPs are shown in Figures 4.7, 4.8, and 4.9 with respect to the dataset 1, 2, and 3. In both ADP platelet responses, the predicted rSNPs are based on the selected HapMap GWAS LDs reference SNPs.

For the 1000 Genomes GWAS LD reference SNPs, the predicted rSNPs, which are involved with distal and long range interactions and associated with PA platelet response are shown in the Figures 4.10, 4.11, and 4.12. For the FA platelet response, the predicted rSNPs are shown in the Figures 4.13, 4.14, and 4.15. In addition, each of the above mentioned figures are followed by tables indicating the detected significant rSNPs with the p-values and other associated regulatory functional elements.

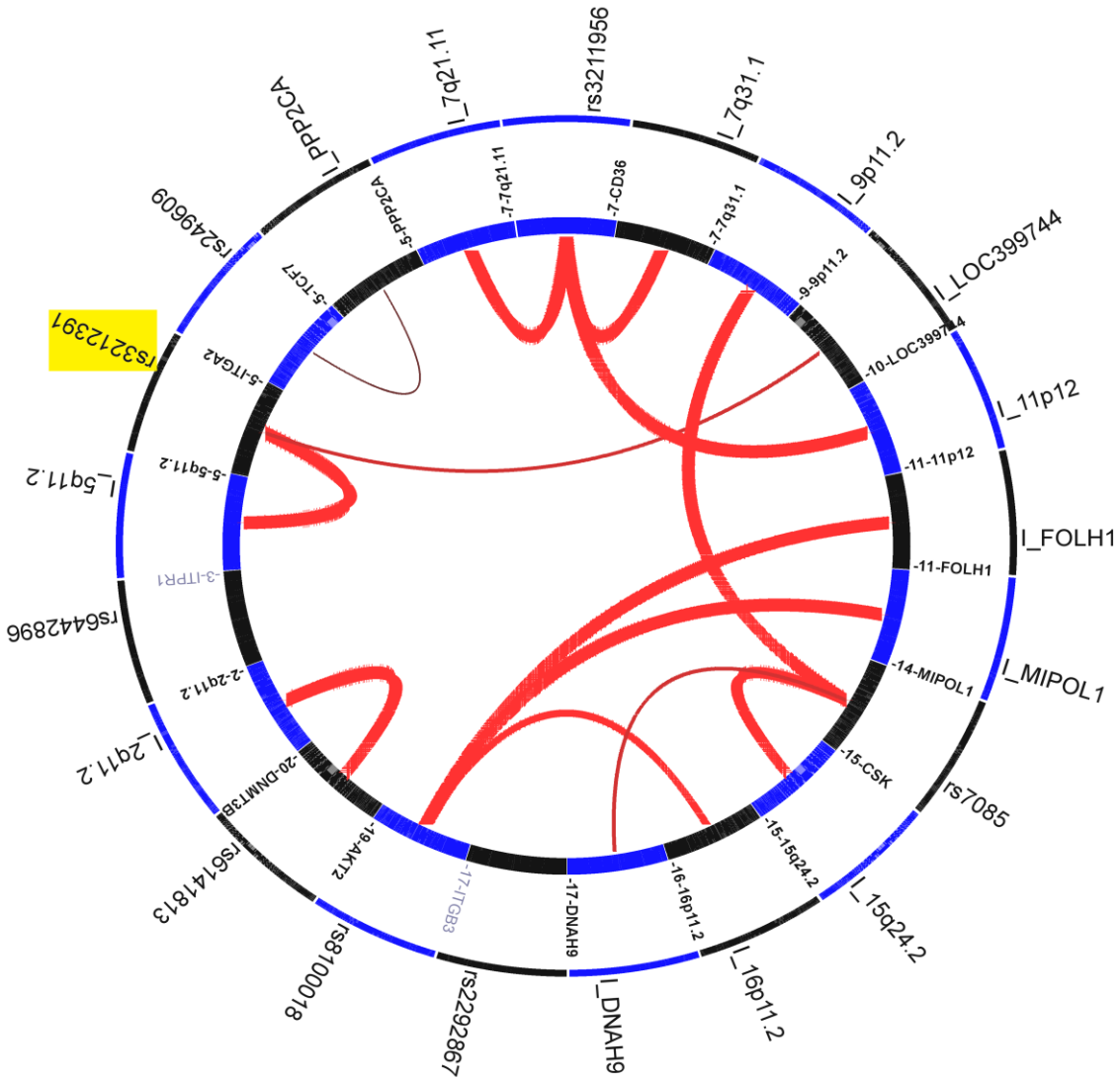


Figure 4.4 The representation of the long range interactions involving the predicted rSNPs associated with PA response from dataset 1 or regulatory loci based on GWAS related HapMap LD SNPs. Top regulatory variants (rSNPs) and distal interaction regions are displayed in the outer circle. The loci or genomic locations connected to respective rSNPs are shown in the inner circle. The interesting significant rSNP from the key SNP, which is involved with distal long range interactions with other genomic regions is highlighted in yellow. The thicker the width of the red line, the higher the intensity of the interaction. Other detected significant rSNPs are not related to RAPIDSNPs, but they are in the same HapMap LD(s) and are shown to compare the significance of those from the RAPIDSNPs pipeline.

From Figure 4.4, there is only one key SNP detected to be significant regulatory, which is rs3212391 and located in the region of *ITGA2* in chromosome 5. This key variant has a long range interaction with *LOC399744* and *5q11.2* loci in the chromosomes 10 and 5 respectively.

CHRPOS	SNPID	GENOTYPE	*LOCUS*	FINALP	LeadSNP	LEADSNP_P	RSQUARE	STATUS
15:75095483	rs7085	T|C	*CSK*	1.89E-04	rs8033381	1.00E+00	1	td, bda, enhancer
5:133472102	rs249609	C|T	*TCF7*	6.47E-04	rs2300065	1.00E+00	0.825	td, bda, enhancer
5:52292405	rs3212391	A|T	*ITGA2*	8.50E-04	rs3212391	1.00E+00	1	td, bda, enhancer
19:40752023	rs8100018	C|G	*AKT2*	2.82E-03	rs3730051	1.00E+00	0.861	td, bda, enhancer
7:80303762	rs3211956	T|G	*CD36*	1.61E-02	rs1527480	1.00E+00	0.869	td, bda
20:31366243	rs6141813	A|G	*DNMT3B*	2.29E-02	rs6141803	1.00E+00	0.89	td, bda, enhancer
3:4694010	rs6442896	A|C	*ITPR1*	8.21E-02	rs6442896	1.00E+00	1	bda, enhancer
17:45357489	rs2292867	C|T	*ITGB3*	8.80E-02	rs2292867	1.00E+00	1	bda, enhancer
3:4809969	rs17041401	T|C	*ITPR1*	9.09E-02	rs17041401	1.00E+00	1	td, bda, enhancer

Table 4.6 The identified rSNPs from dataset 1 based on the HapMap GWAS related SNPs, which are used to compare with those identified by RAPIDSNPs as associated with the PA response. The CHRPOS shows the chromosomal position or co-ordinates of the HapMap/1000 Genomes variant, which is shown in the SNPID column. 'GENOTYPE' is the column with the SNPs alleles in genomic locus. The 'LOCUS' column shows the genomic position of the SNP. The 'FINALP' column shows the final p-values The 'LeadSNP' column shows the related submitted RAPIDSNPs key SNPs. The 'LEADSNP_P' column shows the LeadSNP's p-value. The rSNP from RAPIDSNPs is regarded as significant if it has high p-value and appears in both the 'LeadSNP' and SNPID columns. The RSQUARE column shows the LD (linkage disequilibrium) strength between variants from HapMap/1000 Genomes and that of the RAPIDSNPs. The STATUS column shows the predicted regulatory functional elements such as td (distal regulation), bda (binding affinity to TF) or whether the SNP is located in enhancer. The significant rSNP, which is a key SNP from the RAPIDSNPs is highlighted in green.

From Table 4.6, the only predicted significant rSNP from RAPIDSNPs is rs3212391 in *ITGA2* and is predicted to be involved with different regulatory functional elements.

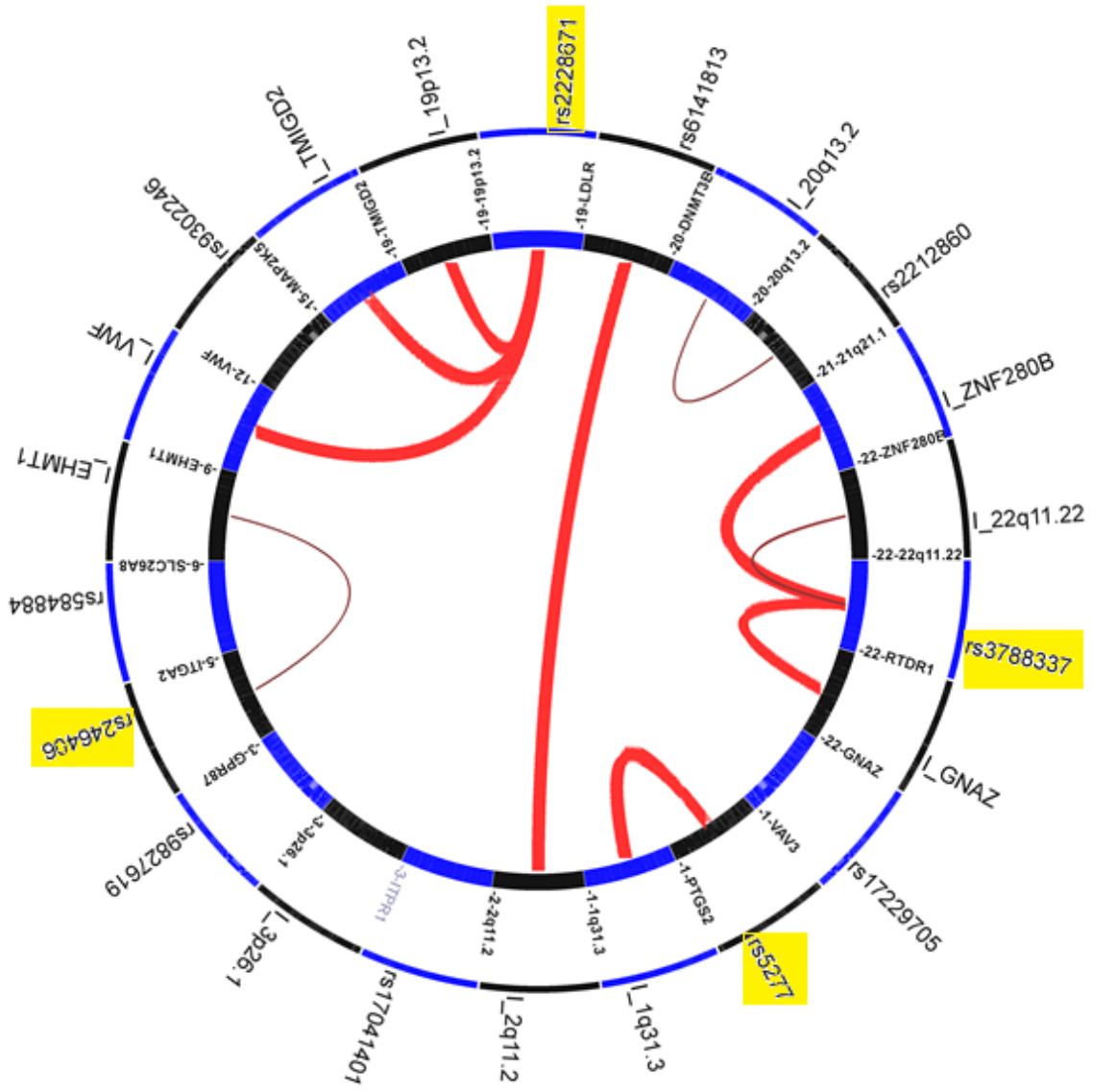


Figure 4.5 The representation of the long range interactions involving the predicted rSNPs associated with PA associated from dataset 2 or regulatory loci based on GWAS related HapMap LD SNPs. The significant rSNPs from RAPIDSNPs, which are likely to be involved with distal long range interactions with other genomic regions are highlighted in yellow.

From Figure 4.5, the most interacting SNP is rs2228671 in the *LDLR* locus in the chromosome 19, which has been predicted to have three distal interactions that include loci *VWF*, *TMIGD2*, and *19p13.2* in the chromosome 12, 19, and 19 respectively. Another rSNP with several

potential distal and long range interactions is rs3788337 in the *GNAZ* locus, which is in the chromosome 22. This locus has several distal regulations with other loci within the same chromosome, which include *RTDR1*, *22q11.22*, and *ZNF280B*.

CHRPOS	SNPID	GENOTYPE	LOCUS	FINALP	LeadSNP	LEADSNP_P	RSQUARE	STATUS
1:108138961	rs17229705	T|C	VAV3	1.48E-05	rs17229705	1.00E+00	1	bda, enhancer
15:68021280	rs9302246	A|G	MAP2K5	3.42E-04	rs11631474	1.00E+00	0.961	bda, enhancer,
5:52319076	rs246406	C|T	ITGA2	6.47E-04	rs246406	1.00E+00	1	td, bda, enhancer
22:23412017	rs3788337	G|A	RTDR1	7.23E-04	rs3788337	1.00E+00	1	td, bda, enhancer
3:151015872	rs9827619	T|C	GPR87	8.07E-04	rs1472122	1.00E+00	1	bda, enhancer
6:35987507	rs584884	C|A	SLC26A8	1.11E-03	rs2815805	1.00E+00	1	td, bda, enhancer
1:186648197	rs5277	C|G	PTGS2	1.37E-03	rs5277	1.00E+00	1	td, bda, enhancer
21:20203619	rs2212860	A|T	21q21.1	1.36E-02	rs950365	1.00E+00	0.959	td, bda
20:31366243	rs6141813	A|G	DNMT3B	2.29E-02	rs6057638	1.00E+00	0.89	td, bda, enhancer
19:11210912	rs2228671	C|T	LDLR	4.99E-02	rs2228671	1.00E+00	1	td, bda, enhancer
3:4809969	rs17041401	T|C	ITPR1	9.09E-02	rs17041401	1.00E+00	1	td, bda, enhancer
1:108115145	rs2769668	T|C	VAV3	1.23E-01	rs2769668	1.00E+00	1	bda, enhancer

Table 4.7 The identified rSNPs from dataset 2 based on the HapMap GWAS related SNPs, which are used to compare with those submitted from the RAPIDSNPs and are associated with PA platelet response. Four key SNPs have been predicted to be highly significant and involved with several transcriptional regulatory elements and are green highlighted. These are rs17229705, rs246406, rs3788337, and rs5277.

From Table 4.7 the top regulatory key SNP is rs17229705 in *VAV3*, which was not found to be involved with long range interactions in Figure 4.5 but, was predicted to be likely involved with other regulatory functional elements such as binding affinity to TF (likely STAT5A) and enhancer.

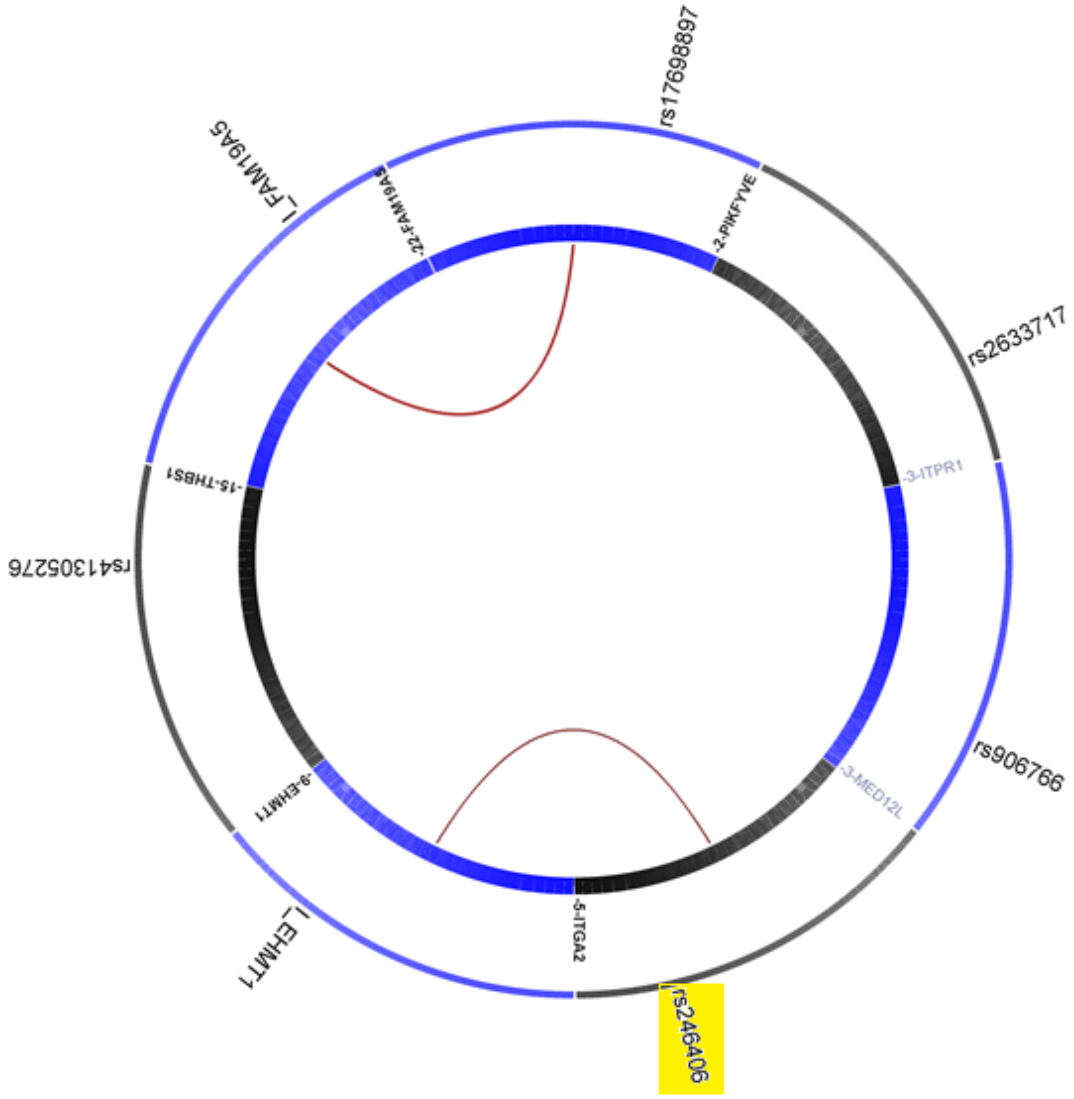


Figure 4.6 The representation of the long range interactions involving the predicted rSNPs associated with the PA response from dataset 3 or regulatory loci based on GWAS related HapMap LD SNPs. A significant rSNP from the RAPIDSNPs is highlighted in yellow and is likely to be involved with distal long range interactions with other genomic regions.

CHRPOS	SNPID	GENOTYPE	LOCUS	FINALP	LeadSNP	LEADSNP_P	RSQUARE	STATUS
5:52319076	rs246406	C|T	ITGA2	6.47E-04	rs246406	1.00E+00	1	td, bda, enhancer
2:209167476	rs17698897	A|G	PIKFYVE	1.27E-02	rs2289171	1.00E+00	0.819	td, bda
15:39878578	rs41305276	C|T	THBS1	2.68E-02	rs41305276	1.00E+00	1	bda, enhancer
3:150811294	rs906766	C|T	MED12L	8.11E-02	rs906766	1.00E+00	1	bda, enhancer
3:4828674	rs2633717	T|C	ITPR1	8.53E-02	rs2633717	1.00E+00	1	bda, enhancer

Table 4.8 The identified rSNP from dataset 3 based on the HapMap GWAS related SNPs, which are used to compare with those submitted from the RAPIDSNPs and are associated with PA response. Only rs246406 key SNP was found to be involved with regulatory function.

From Figure 4.6 and Table 4.8, there is only one predicted rSNP rs246406 in the *ITGA2* locus, which is predicted to be likely involved with the long range interaction with *EHMT1* chromosomal region.

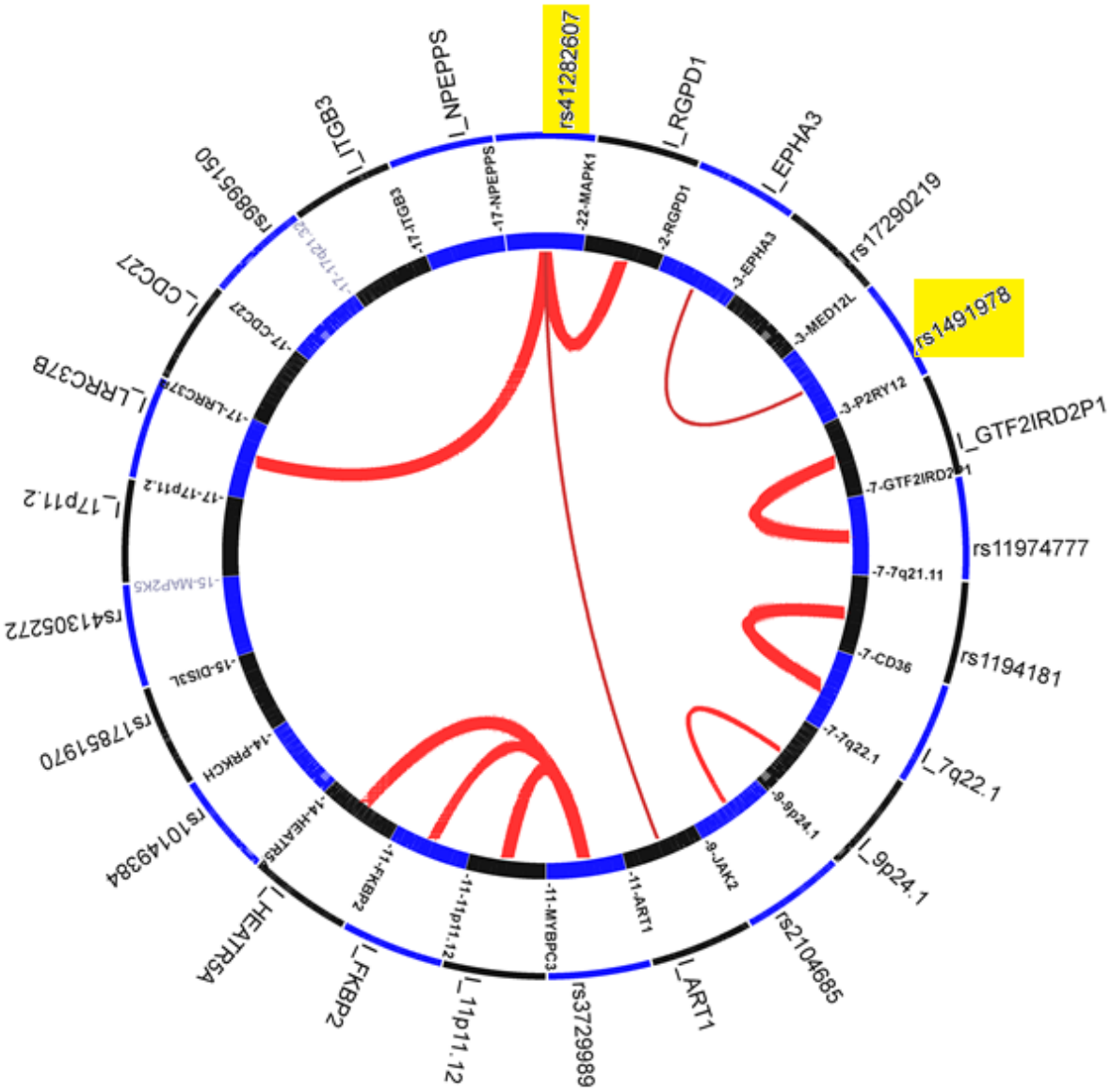


Figure 4.7 The representation of the long range interactions involving the predicted rSNPs associated with the FA response from dataset 1 or regulatory loci based on GWAS related HapMap LD SNPs. The SNPs from RAPIDSNPs, which are significant and likely to be involved with distal long range interactions with other genomic regions are yellow highlighted.

CHRPOS	SNPID	GENOTYPE	LOCUS	FINALP	LeadSNP	LEADSNP_P	RSQUARE	STATUS
3:150881774	rs17290219	A|C	MED12L	6.16E-04	rs17204437	1.00E+00	0.883	bda, enhancer
3:151080070	rs1491978	C|T	P2RY12	9.66E-04	rs1491978	1.00E+00	1	td, bda, enhancer
7:80231793	rs1194181	G|A	CD36	1.11E-03	rs1537593	1.00E+00	0.838	td, bda
15:66612965	rs17851970	T|C	DIS3L	1.65E-03	rs11637556	1.00E+00	1	bda, enhancer
22:22118229	rs41282607	C|T	MAPK1	2.06E-03	rs41282607	1.00E+00	1	td, bda, enhancer
11:47370041	rs3729989	T|C	MYBPC3	2.11E-03	rs3736101	1.00E+00	0.906	td, bda, enhancer
14:61846130	rs10149384	C|A	PRKCH	2.13E-03	rs3742633	1.00E+00	0.874	bda, enhancer
9:5106023	rs2104685	A|T	JAK2	2.23E-03	rs7034539	1.00E+00	1	td, bda
7:80213056	rs11974777	T|G	7q21.11	2.82E-03	rs10499858	1.00E+00	0.801	td, bda
17:45307928	rs9895150	A|G	17q21.32	1.27E-01	rs9895150	1.00E+00	1	td, bda, enhancer
15:68099443	rs41305272	C|T	MAP2K5	1.28E-01	rs41305272	1.00E+00	1	td, bda, enhancer

Table 4.9 The identified rSNPs from dataset 1 based on the HapMap GWAS related SNPs, which are used to compare with those submitted from the RAPIDSNPs and are associated with FA response in the dataset 1. The significant rSNPs from the RAPIDSNPs are highlighted in green.

From Figure 4.7 and Table 4.9, there are two predicted rSNPs, which seem to be involved with many interactions. These rSNPs are rs1491978 in *P2YR12*, which has 3 potential distal and long range interactions with the *EPHA3*, and rs41282607, which has 3 potential significant distal interactions with the RGPD1, *ART1*, and *LRRC378* loci.

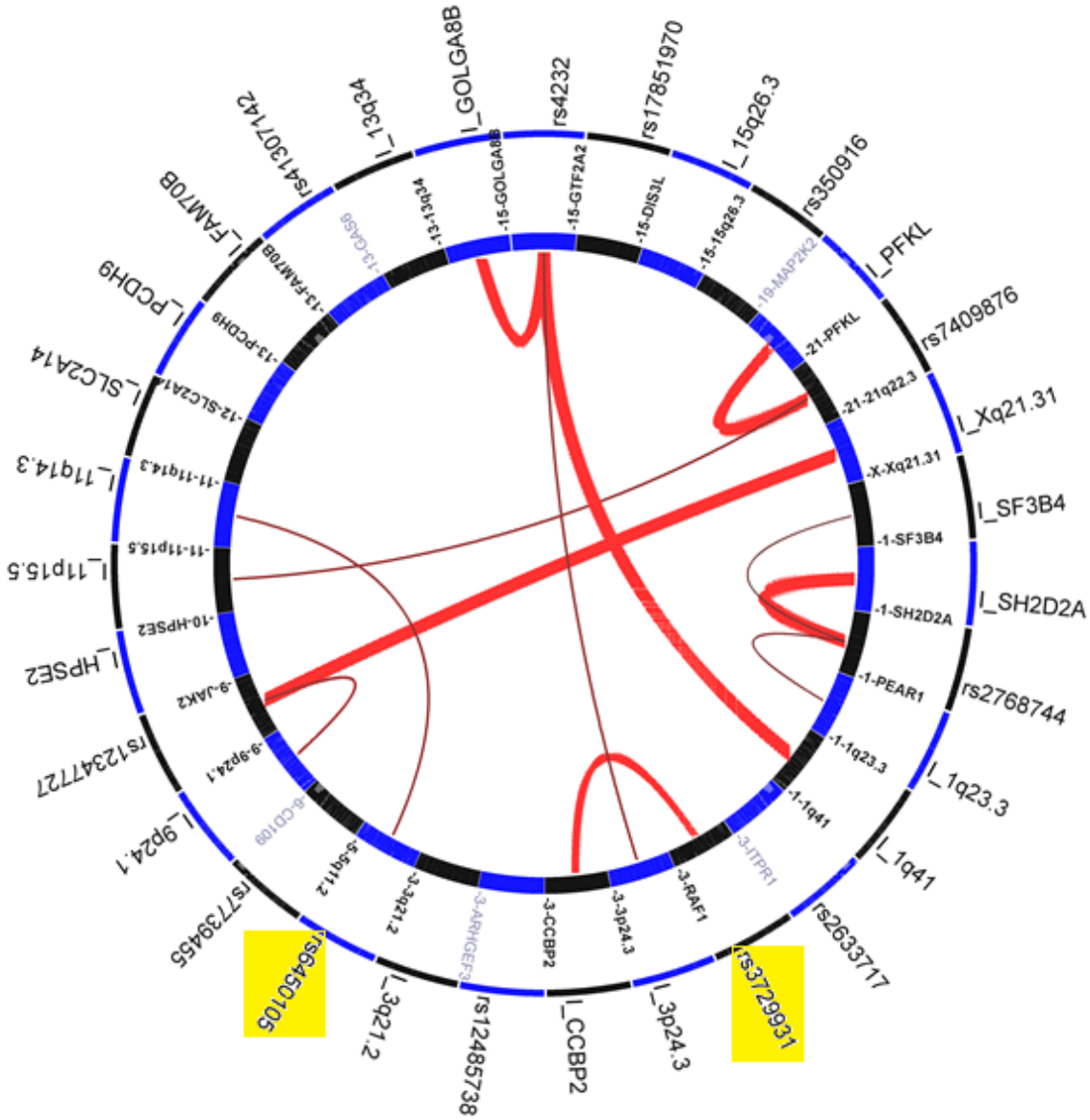


Figure 4.8 The representation of the long range interactions involving the predicted rSNPs associated with the FA response from dataset 2 or regulatory loci based on GWAS related HapMap LD SNPs. Two significant SNPs from RAPIDSNPs, which are likely to be involved with distal long range interactions with other genomic regions are yellow highlighted.

CHRPOS	SNPID	GENOTYPE	LOCUS	FINALP	LeadSNP	LEADSNP_P	RSQUARE	STATUS
3:12626516	rs3729931	G|A	RAF1	2.60E-05	rs3729931	1.00E+00	1	td, bda, enhancer
15:59931131	rs4232	A|G	GTF2A2	2.96E-05	rs7180408	1.00E+00	0.808	td, bda
9:5000811	rs12347727	A|G	JAK2	7.23E-04	rs10429491	1.00E+00	0.861	td, bda, enhancer
1:156872149	rs2768744	G|A	PEAR1	1.45E-03	rs11264579	1.00E+00	1	td, bda, enhancer
15:66612965	rs17851970	T|C	DIS3L	1.65E-03	rs11637556	1.00E+00	1	bda, enhancer
5:52276309	rs6450105	T|C	5q11.2	1.53E-02	rs6450105	1.00E+00	1	td, bda, enhancer
21:45763960	rs7409876	T|G	21q22.3	3.39E-02	rs7409876	1.00E+00	1	td, bda, enhancer
3:12641425	rs5746223	T|C	RAF1	4.15E-02	rs5746223	1.00E+00	1	td, bda, enhancer
5:52275718	rs41305896	T|C	5q11.2	4.33E-02	rs41305896	1.00E+00	1	td, bda, enhancer
6:74420628	rs7739455	A|T	CD109	5.00E-02	rs7739455	1.00E+00	1	td, bda, enhancer
19:4094775	rs350916	G|A	MAP2K2	7.73E-02	rs350916	1.00E+00	1	td, bda
3:56865776	rs12485738	A|G	ARHGEF3	8.02E-02	rs12485738	1.00E+00	1	td, bda, enhancer
13:114524843	rs41307142	C|G	GAS6	8.43E-02	rs41307142	1.00E+00	1	td, bda, enhancer
3:4828674	rs2633717	T|C	ITPR1	8.53E-02	rs17786144	1.00E+00	0.933	bda, enhancer

Table 4.10 The identified rSNPs from dataset 2 based on the HapMap GWAS related SNPs, which are used to compare with those submitted from the RAPIDSNPs key SNPs and are associated with FA in the dataset 2. The significant rSNPs are highlighted in green.

From Figure 4.8 and Table 4.10, two rSNPs were predicted to be significantly involved with the long range interactions. These SNPs are rs3729931 in RAF1, and rs6450105 in 5q11.2 in chromosomes 3 and 5, which are predicted to have long range interactions with *CCBP2* and *11q14.3* respectively.

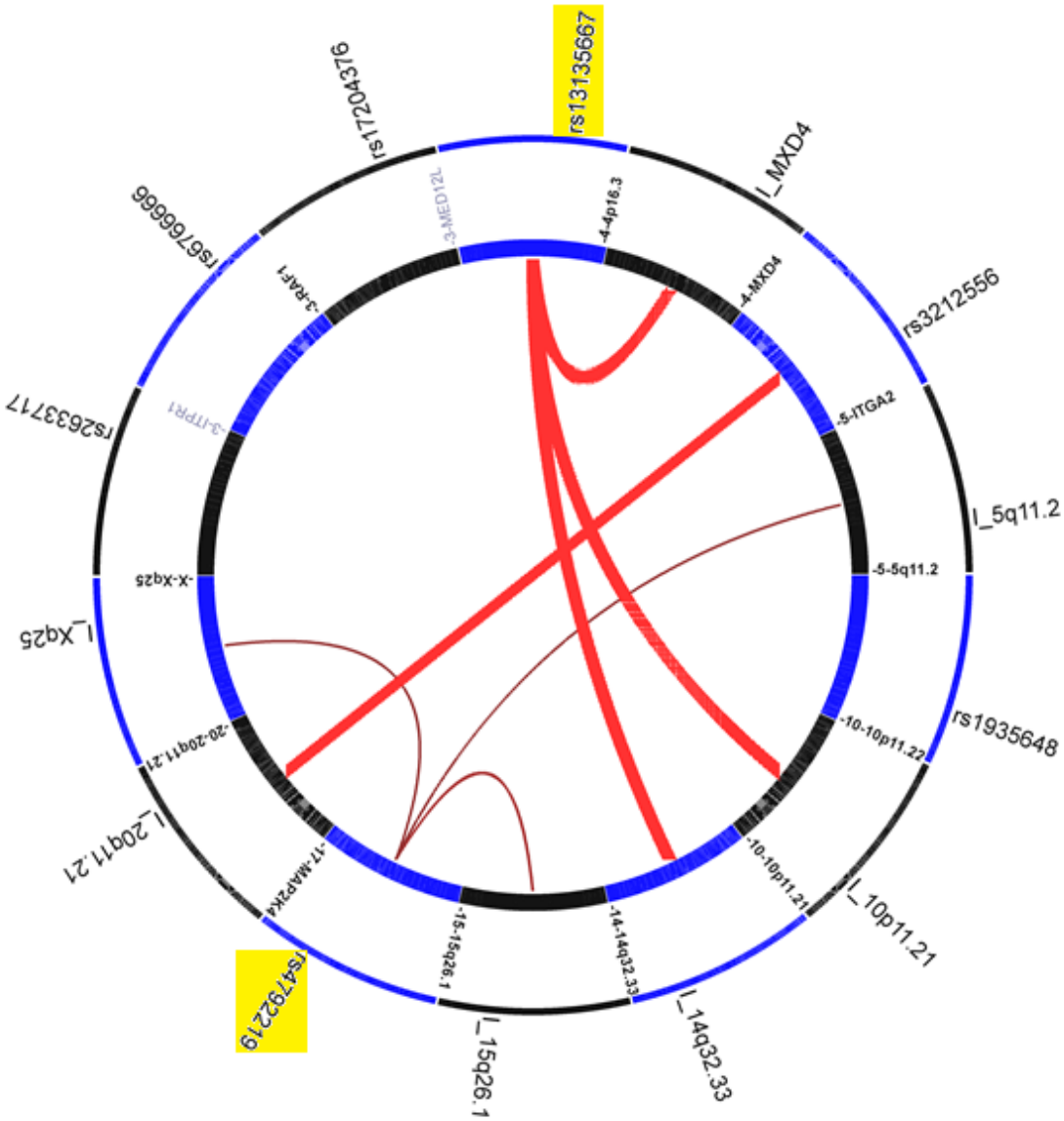


Figure 4.9 The representation of the long range interactions involving the predicted rSNPs associated with the FA response from dataset 3 or regulatory loci based on GWAS related HapMap LD SNPs. Two significant rSNPs from RAPIDSNPs, which are likely to be involved with distal long range interactions with other genomic regions are highlighted in yellow.

From Figure 4.9, the intergenic SNP rs13135667 is the most significant rSNP and is likely to be involved with the distal regulation and long range interactions with three loci in the chromosomal regions of *MXD4*, *10p11.21*, and *14q32.33* in the chromosome 4, 10, and 14 respectively.

CHRPOS	SNPID	GENOTYPE	LOCUS	FINALP	LeadSNP	LEADSNP_P	RSQUARE	STATUS
10:33252263	rs1935648	C|G	10p11.22	1.20E-04	rs17296289	1.00E+00	0.89	bda, enhancer
3:12690855	rs6766666	G|T	RAF1	2.59E-04	rs2290159	1.00E+00	0.879	bda, enhancer
10:33260699	rs17296289	G|A	10p11.22	8.07E-04	rs722432	1.00E+00	0.89	td, bda, enhancer
5:52362765	rs3212556	A|T	ITGA2	4.90E-03	rs3212603	1.00E+00	1	td, bda
17:12045905	rs4792219	G|A	MAP2K4	6.60E-03	rs4792219	1.00E+00	1	td, bda
4:1767067	rs13135667	C|G	4p16.3	3.28E-02	rs13135667	1.00E+00	1	td, bda, enhancer
3:151035961	rs17204376	G|T	MED12L	5.13E-02	rs17204376	1.00E+00	1	bda, enhancer
3:150811294	rs906766	C|T	MED12L	8.11E-02	rs906766	1.00E+00	1	bda, enhancer
3:4828674	rs2633717	T|C	ITPR1	8.53E-02	rs17786144	1.00E+00	0.933	bda, enhancer

Table 4.11 The identified rSNP from dataset 3 based on the HapMap GWAS related SNPs, which are used to compare with those submitted from the RAPIDSNPs key SNPs and are associated with FA in the dataset 3.

Furthermore, based on Table 4.11, the predicted rSNP rs4792219 in *MAP2K4*, is also significantly associated with the regulatory roles such as binding affinity with some transcription factors as predicted elsewhere above (Tables 4.3 and 4.5).

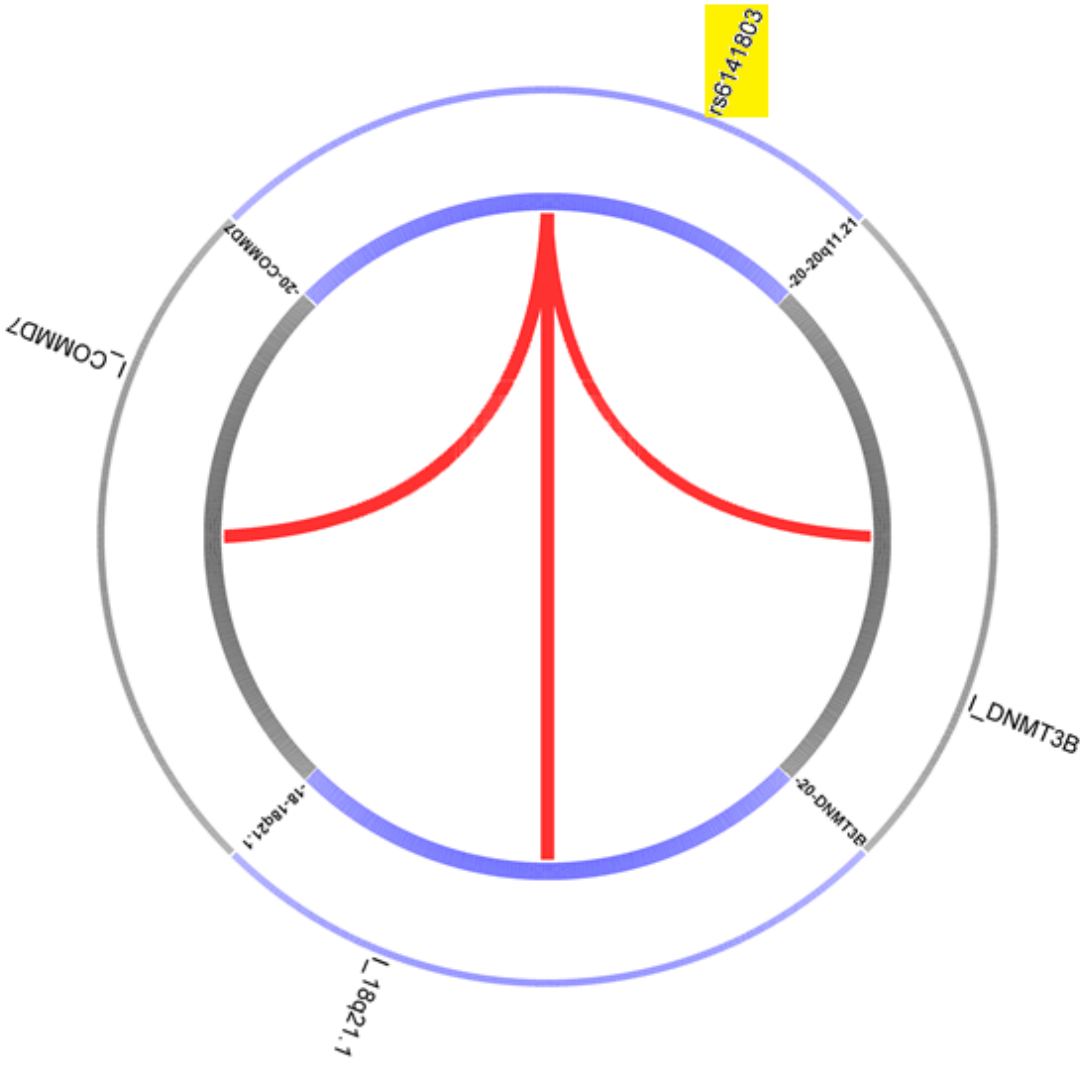


Figure 4.10 The representation of the long range interactions involving the predicted rSNPs associated with the PA response from dataset 1 or regulatory loci based on GWAS related 1000 Genomes LD SNPs. A significant SNP from RAPIDSNPs, which is likely to be involved with distal long range interactions with other genomic regions is highlighted in yellow.

CHRPOS	SNPID	GENOTYPE	*LOCUS*	FINALP	LeadSNP	LEADSNP_P	RSQUARE	STATUS
20:31340356	rs6141803	T|C	*20q11.21*	3.84E-02	rs6141803	1.00E+00	1	td, bda, enhancer,

Table 4.12 The identified rSNPs from dataset 1 based on the 1000 Genomes GWAS related SNPs, which are used to compare with those submitted from the RAPIDSNPs key SNPs and are associated with PA in the dataset 1. The SNPID column shows those from the similar LD within 1000 Genomes. The LeadSNP column shows the submitted RAPIDSNPs.

From Figure 4.10 and 4.12, only rs6141803 in the *ch20:32752550* locus was found to have distal and long range interactions with *COMMD7* in the chromosome 20, *18q21.1*, and *DNMT3B* loci. The selection of this rSNP reflects similar results when using HapMap GWAS SNPs.

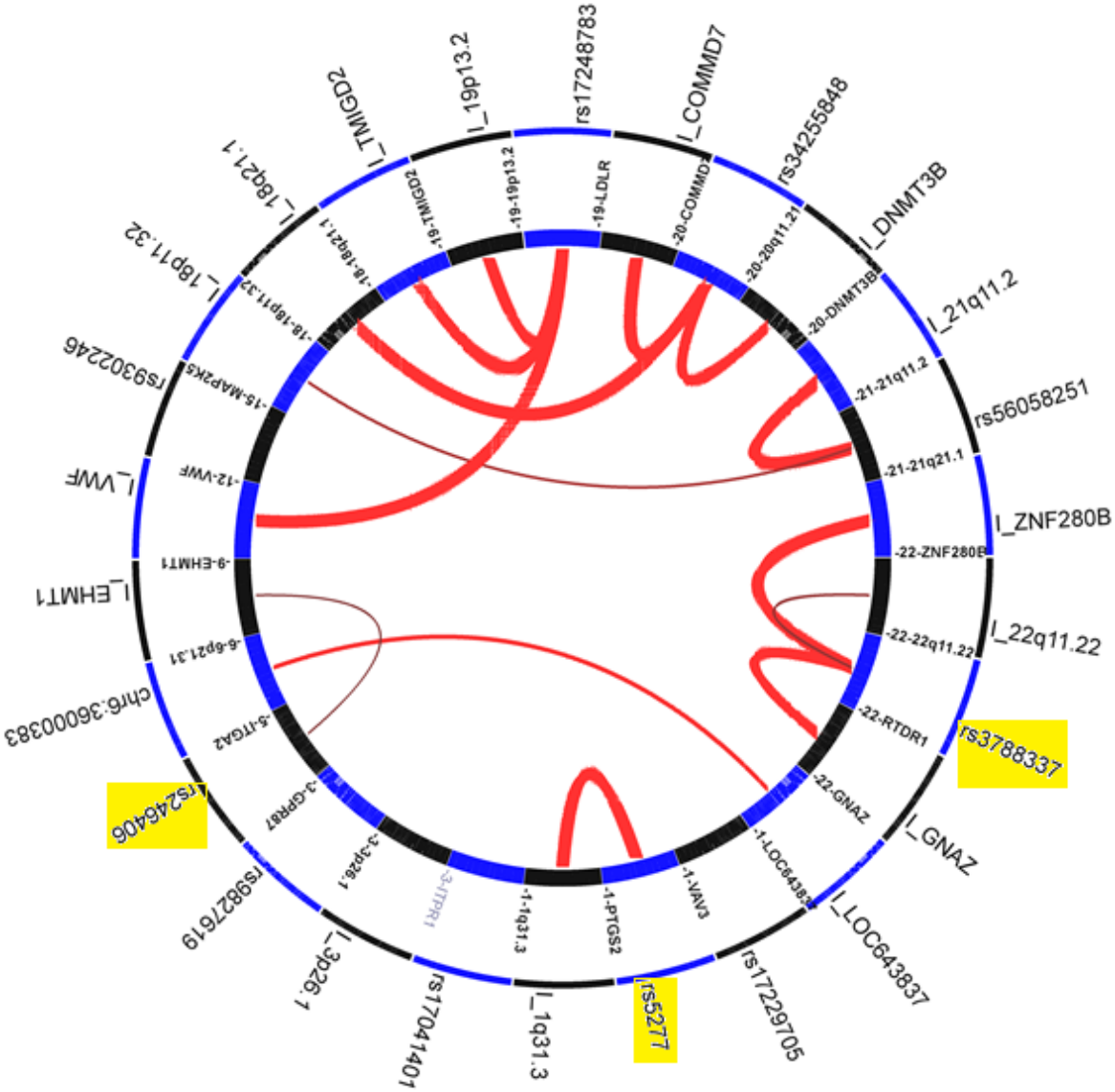


Figure 4.11 The representation of the long range interactions involving the predicted rSNPs associated with the PA response from dataset 2 or regulatory loci based on GWAS related 1000 Genomes LD. Three significant rSNPs from RAPIDSNPs, which are likely to be involved with distal long range interactions with with other genomic regions are highlighted in yellow.

From Figure 4.11, there are three predicted significant rSNPs from the RAPIDSNPs that are likely to be involved with distal regulation or interactions, and are associated with PA. These rSNPs are rs5277, rs246406 and rs3788337 in the *PTGS2*, *ITGA2*, *RTDR1* in the chromosomes loci 1, 5 and 22.

CHRPOS	SNPID	GENOTYPE	*LOCUS*	FINALP	LeadSNP	LEADSNP_P	RSQUARE	STATUS
1:108138961	rs17229705	T|C	*VAV3*	1.48E-05	rs17229705	1.00E+00	1	bda, enhancer,
15:68021280	rs9302246	A|G	*MAP2K5*	3.42E-04	rs11631474	1.00E+00	0.933	bda, enhancer,
19:11210314	rs17248783	G|A	*LDLR*	5.36E-04	rs2228671	1.00E+00	0.9839	td, bda, enhancer
5:52319076	rs246406	C|T	*ITGA2*	6.47E-04	rs246406	1.00E+00	1	td, bda, enhancer,
6:36000383	chr6:36000383	A|G	*6p21.31*	7.21E-04	rs2815805	1.00E+00	1	td, bda, enhancer
22:23412017	rs3788337	G|A	*RTDR1*	7.23E-04	rs3788337	1.00E+00	1	td, bda, enhancer,
3:151015872	rs9827619	T|C	*GPR87*	8.07E-04	rs1472122	1.00E+00	0.9222	bda, enhancer
1:186648197	rs5277	C|G	*PTGS2*	1.37E-03	rs5277	1.00E+00	1	td, bda, enhancer,
21:20216800	rs56058251	A|G	*21q21.1*	3.38E-03	rs950365	1.00E+00	0.9945	td, bda
20:31341267	rs34255848	T|A	*20q11.21*	1.43E-02	rs6057638	1.00E+00	0.9902	td, bda, enhancer
3:4809969	rs17041401	T|C	*ITPR1*	9.09E-02	rs17041401	1.00E+00	1	td, bda, enhancer,

Table 4.13 The identified rSNPs from dataset 2 based on the 1000 Genomes GWAS related SNPs, which are used to compare with those submitted from the RAPIDSNPs key SNPs and are associated with PA in the dataset 2.

From Table 4.13, the results appear to be similar to those obtained when using HapMap data in which the same rSNPs appear to significant.

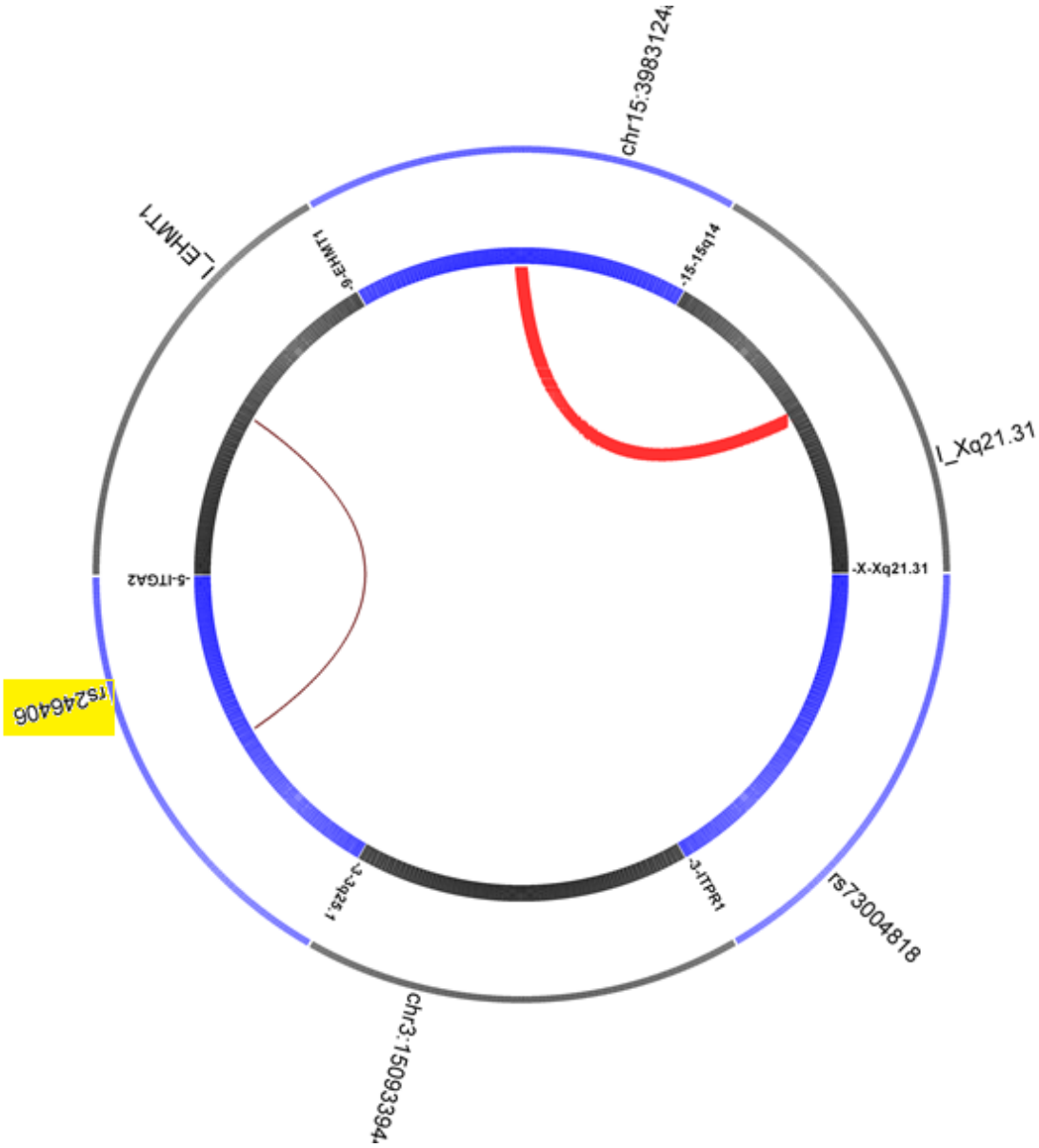


Figure 4.12 The representation of the long range interactions involving the predicted rSNPs associated with the PA response from dataset 3 or regulatory loci based on GWAS related 1000 Genomes LD SNPs. One significant rSNP from RAPIDSNPs, which is likely to be involved with distal long range interactions with with other genomic regions is highlighted in yellow.

From Figure 4.12, there is only one distal regulation, which involves rs246406 rSNP in the *ITGA2* locus in the chromosome 5 that interact with *EHMT1* in the chromosome 9.

CHRPOS	SNPID	GENOTYPE	LOCUS	FINALP	LeadSNP	LEADSNP_P	RSQUARE	STATUS
3:150933944	chr3:150933944	A|T	3q25.1	2.59E-04	rs906766	1.00E+00	0.8738	bda, enhancer
5:52319076	rs246406	C|T	ITGA2	6.47E-04	rs246406	1.00E+00	1	td, bda, enhancer,
15:39831248	chr15:39831248	C|T	15q14	9.66E-04	rs41305276	1.00E+00	0.8558	td, bda, enhancer
3:4831904	rs73004818	G|A	ITPR1	1.21E-03	rs2633717	1.00E+00	0.8697	bda, enhancer

Table 4.14 The identified rSNPs from dataset 3 based on the 1000 Genomes GWAS related SNPs, which are used to compare with those submitted from the RAPIDSNPs key SNPs and are associated with PA response in the dataset 3.

From Table 4.14, the similar rSNP rs246406, which was selected based on HapMap for the dataset 3 (Table 4.8) is also picked by using 1000 Genomes GWAS related SNPs.

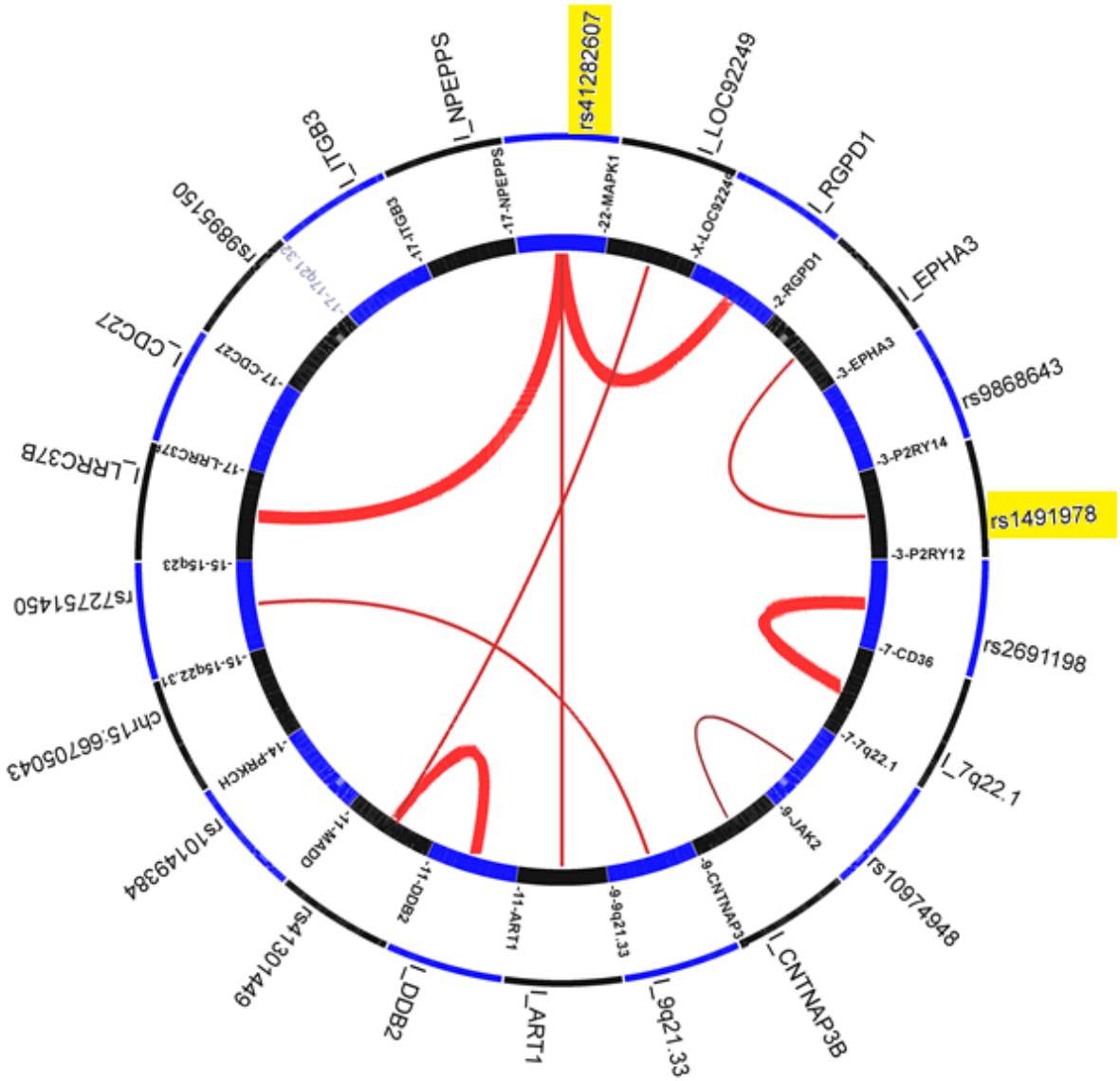


Figure 4.13 The representation of the long range interactions involving the predicted rSNPs associated with FA response from dataset 1 or regulatory loci based on GWAS related 1000 Genomes LD SNPs. Two significant rSNPs from RAPIDSNPs, which are likely to be involved with distal long range interactions with other genomic regions are highlighted in yellow.

From Figure 4.13, there are two predicted rSNPs, which are rs1491978 in *P2RY12* interacting with *EPHA3* in chromosome 3 and rs41282607 in *MAPK1*, which may involve *RGDP* in the chromosome 2, *ART* in the chromosome 11, and *LRRC37B* in the chromosomes 2, 11, and 17 respectively.

CHRPOS	SNPID	GENOTYPE	LOCUS	FINALP	LeadSNP	LEADSNP_P	RSQUARE	STATUS
7:80236603	rs2691198	A|C	CD36	4.06E-05	rs1537593	1.00E+00	0.8936	td, bda, enhancer
3:150950288	rs9868643	A|C	P2RY14	7.21E-04	rs17204437	1.00E+00	0.8696	bda, enhancer
15:66705043	chr15:66705043	G|A	15q22.31	8.07E-04	rs11637556	1.00E+00	0.8173	bda, enhancer
3:151080070	rs1491978	C|T	P2RY12	9.66E-04	rs1491978	1.00E+00	1	td, bda, enhancer
9:5075255	rs10974948	C|G,T	JAK2	9.66E-04	rs7034539	1.00E+00	0.9641	td, bda
7:80254792	rs9641866	T|A	CD36	1.64E-03	rs10499858	1.00E+00	0.9079	bda, enhancer
22:22118229	rs41282607	C|T	MAPK1	2.06E-03	rs41282607	1.00E+00	1	td, bda, enhancer
14:61846130	rs10149384	C|A	PRKCH	2.13E-03	rs3742633	1.00E+00	0.8492	bda, enhancer
11:47303821	rs41301449	C|T	MADD	2.48E-03	rs3736101	1.00E+00	0.9728	td, bda, enhancer
15:68102425	rs72751450	C|T	15q23	2.82E-02	rs41305272	1.00E+00	1	td, bda, enhancer
17:45307928	rs9895150	A|G	17q21.32	1.27E-01	rs9895150	1.00E+00	1	td, bda, enhancer

Table 4.15 The identified rSNPs from dataset 1 based on the 1000 Genomes GWAS related SNPs, which are used to compare with those submitted from the RAPIDSNPs key SNPs and are associated with FA in the dataset 1.

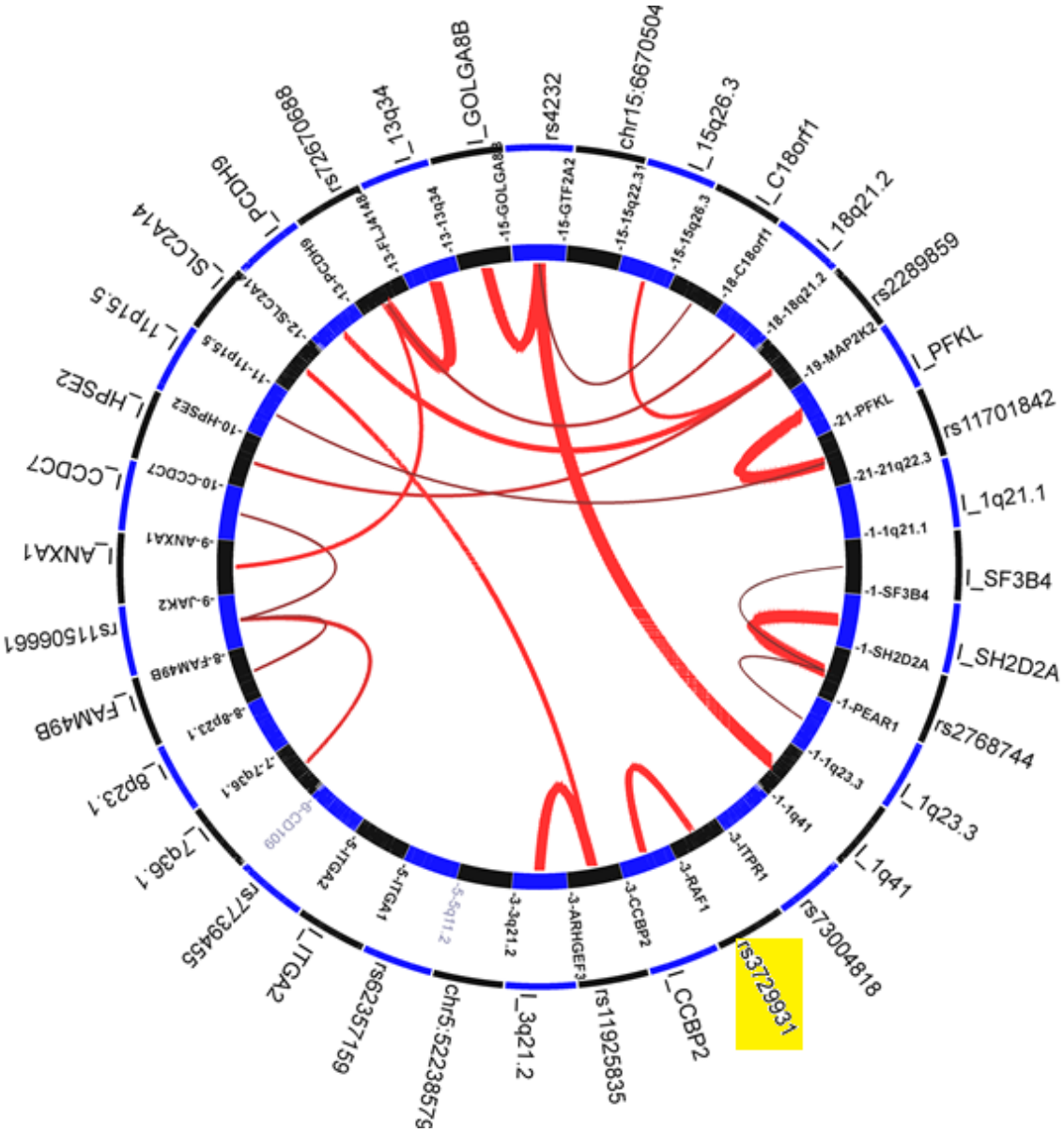


Figure 4.14 The representation of the long range interactions involving the predicted rSNPs from dataset 2 or regulatory loci based on GWAS related 1000 Genomes LD SNPs A yellow highlighted rSNP is from RAPIDSNPs, which is significant and likely to be involved with distal long range interactions with other genomic regions.

From Figure 4.14, it appears that there is one predicted rSNP with the distal interaction, which is rs3729931 in *RAF1* locus interacting with the *CCBP2* in the chromosome 3. The associated interacting genomic regions are more likely to be involved with the regulation of the FA response.

CHRPOS	SNPID	GENOTYPE	LOCUS	FINALP	LeadSNP	LEADSNP_P	RSQUARE	STATUS
3:12626516	rs3729931	G|A	RAF1	2.60E-05	rs3729931	1.00E+00	1	td, bda, enhancer,
15:59931131	rs4232	A|G	GTF2A2	2.96E-05	rs7180408	1.00E+00	0.863	td, bda,
9:5035586	rs11506661	A|T	JAK2	2.59E-04	rs10429491	1.00E+00	0.8341	td, bda
13:114519681	rs72670688	C|T	FLJ41484	2.59E-04	rs41307142	1.00E+00	0.8866	td, bda, enhancer
15:66705043	chr15:66705043	G|A	15q22.31	8.07E-04	rs11637556	1.00E+00	0.8173	bda, enhancer
3:12662426	rs61147639	G|A	RAF1	8.94E-04	rs5746223	1.00E+00	0.9497	td, bda, enhancer
21:45765331	rs11701842	T|C	21q22.3	9.66E-04	rs7409876	1.00E+00	0.9563	td, bda, enhancer
5:52239009	rs62357159	G|T	ITGA1	9.66E-04	rs6450105	1.00E+00	0.8558	td, bda
3:4831904	rs73004818	G|A	ITPR1	1.21E-03	rs17786144	1.00E+00	0.8193	bda, enhancer
1:156872149	rs2768744	G|A	PEAR1	1.45E-03	rs11264579	1.00E+00	0.9829	td, bda, enhancer
19:4094110	rs2289859	G|T	MAP2K2	4.09E-02	rs350916	1.00E+00	0.9895	td, bda
3:56865445	rs11925835	T|C	ARHGEF3	4.28E-02	rs12485738	1.00E+00	0.8008	td, bda, enhancer
6:74420628	rs7739455	A|T	CD109	5.00E-02	rs7739455	1.00E+00	1	td, bda, enhancer
5:52238579	chr5:52238579	G|A	5q11.2	8.55E-02	rs41305896	1.00E+00	0.8601	td, bda

Table 4.16 The identified rSNPs from dataset 2 based on the 1000 Genomes GWAS related SNPs, which are used to compare with those submitted from the RAPIDSNPs key SNPs and are associated with FA in the dataset 2.

From Table 4.16, there is one predicted rSNP with the distal interaction, which is rs3729931 in *RAF1* locus interacting with the *CCBP2* in the chromosome 3.

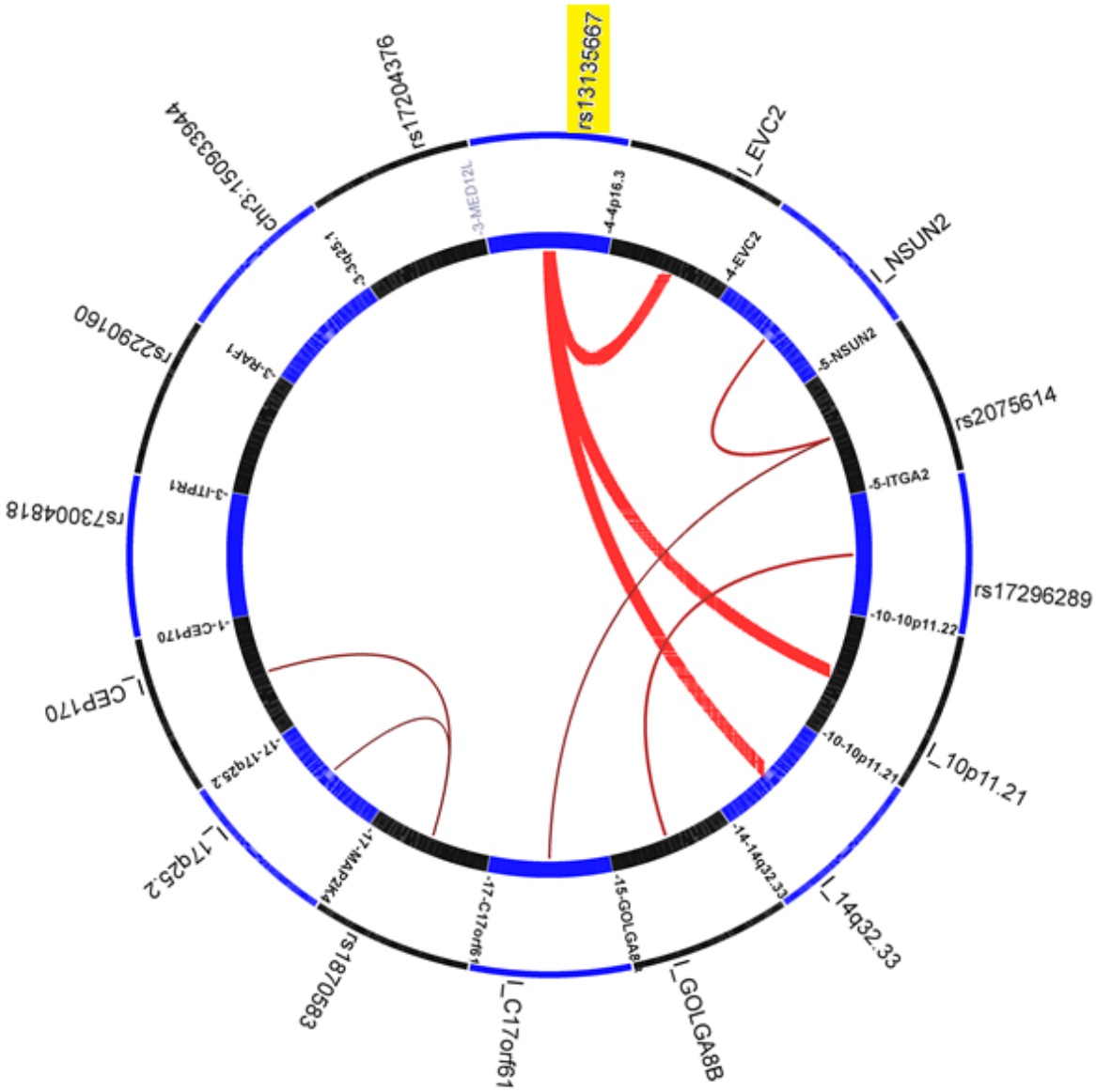


Figure 4.15 The representation of the long range interactions involving the predicted rSNPs from dataset 3 or regulatory loci based on GWAS related 1000 Genomes LD SNPs. A yellow highlighted rSNP is from RAPIDSNPs, which is significant and likely to be involved with distal long range interactions with other genomic regions.

From Figure 4.15, it can be observed that once again the intergenic SNP rs13135667 in the ch4:1765340 locus appear to have likely distal regulations or interactions with *14q32.33*, *10p11.21*, and *EVC2* in the chromosomes 14, 10, and 4 respectively.

CHRPOS	SNPID	GENOTYPE	LOCUS	FINALP	LeadSNP	LEADSNP_P	RSQUARE	STATUS
3:150933944	chr3:150933944	A|T	3q25.1	2.59E-04	rs906766	1.00E+00	0.8738	bda, enhancer
10:33260699	rs17296289	G|A	10p11.22	8.07E-04	rs722432	1.00E+00	0.8726	td, bda, enhancer
3:12633083	rs2290160	A|G	RAF1	9.25E-04	rs2290159	1.00E+00	0.9637	bda, enhancer
3:4831904	rs73004818	G|A	ITPR1	1.21E-03	rs17786144	1.00E+00	0.8193	bda, enhancer
5:52345035	rs2075614	T|G	ITGA2	2.02E-03	rs3212603	1.00E+00	0.928	td, bda, enhancer
17:12032944	rs1870583	A|G	MAP2K4	4.36E-03	rs4792219	1.00E+00	0.9816	td, bda,
4:1767067	rs13135667	C|G	4p16.3	3.28E-02	rs13135667	1.00E+00	1	td, bda, enhancer,
3:151035961	rs17204376	G|T	MED12L	5.13E-02	rs17204376	1.00E+00	1	bda, enhancer,

Table 4.17 The identified rSNPs from dataset 3 based on the 1000 Genomes GWAS related SNPs, which are used to compare with those submitted from the RAPIDSNPs key SNPs and are associated with FA in the dataset 3.

From Table 4.17, only the rs13135667 key SNP was identified to be significant

4.4.2.3.1 Summary of predicted interesting rSNPs likely to be involved in long range interactions

In examining all of the predicted rSNPs, which are likely to be involved in long range and distal interactions from GWAS3D, the significant and seems most interesting regulatory variant is rs2228671. This rSNP is a synonymous codon in the *LDLR* and newly identified by the RAPIDSNPs. The *LDLR* is known for its involvement in the cholesterol metabolism and hence is one of the determinants of the individuals' CVD risk. The associated rSNP has been predicted to be interacting with *VWF*, which is ADP platelet response associated gene.

4.4.3 The rSNPs that are predicted to be in the RNA binding sites and likely affecting the bound proteins

Several of the SNPs across the three datasets were identified to be likely to localise and chnage the RNA binding sites. Thus, they are likely to be involved in regulating the binding sites affinity of the bound proteins (or their complexes) and affecting post-transcription regulation.

Tables 4.18 and 4.19 show the predicted RNA binding proteins (RBPs) and the associated rSNP(s) for PA and FA respectively based on the rSNPBase method.

Data set	Identified rSNP	SNP Type	Locus/genome location	Possible Associated RBP
Dataset 1	**rs2300065**	intron	*chr:5,133510384*	SKP1
	rs8033381	Intron, downstream 500B	*chr:15,75080685*	CSK
Dataset 2	**rs17229705**	nc transcript, synonymous codon	*chr:1,108138961*	VAV3
	rs1472122	intron, utr 5'	*chr:3,151034602*	P2RY12, GPR87
	rs2815805	intron, stop lost, synonymous codon	*chr:6,36068041*	MAPK14
	rs5277	synonymous codon, upstream 2KB	*chr:1,186648197*	PTGS2
	rs41307147	intron	*chr:6,74527043*	CD109
	rs2769668	nc transcript variant, utr 3'	*chr:1,108115145*	VAV3
	rs3788337	intron, upstream 2KB, utr 5'	*chr:22,23412017*	RTDR1
Dataset 3	**rs906766**	intron	*chr:3,150811294*	MED12L
	rs41305276	intron	*chr:15,39878578*	THBS1

Table 4.18 The predicted rSNPs, which are likely to influence the binding affinity of the RNA binding proteins (RBPs) that are potentially associate with PA. These RBP are common in different cell-types using the ENCODE data. The previously identified SNPs by Jones et al. (2009) are italicised.

Data set	Identified rSNP	SNP Type	Locus/genome location	Possible Associated RBP
Dataset 1	**rs1491978**	intron	*chr:3,151080070*	MED12L, P2RY12,
	rs17204437	intron, upstream 2KB	*chr:3,151067007*	MED12L, P2RY12
Dataset 2	**rs12485738**	intron	*chr:3,56865776*	ARHGEF3
	rs41307142	intron	*chr:13,114524843*	GAS6AS1, GAS6
	rs11637556	intron	*chr:15,66728951*	MAP2K1, ELAVL1
	rs3729931	intron	*chr:3,12626516*	RAF1
Dataset 3	**rs17204376**	intron, upstream 2KB	*chr:3,151035961*	MED12L
	rs4792219	utr 3 prime	*chr:17,12045905*	MAP2K4
	rs2290159	intron	*chr:3,12628920*	RAF1
	rs906766	intron	*chr:3,150811294*	MED12L

Table 4.19 The predicted rSNPs, which are likely to influence the binding affinity of the RNA binding proteins (RBPs) that potentially may associate with FA. The major observable difference with PA platelet response is that for the dataset 2, all of the predicted rSNPs are in the intronic regions. In bold are the newly discovered significant SNPs, which were not previously identified in the study by Jones et al. (2009).

Furthermore, using the RBPmap method, several SNPs were identified to be significantly involved with the changing the binding affinity of the proteins. Each identified rSNP is mapped

with the significant possible motifs in which it may localise and the predicted binding site(s) or genomic position with its bound protein(s). Tables 4.20 and 4.21 show the selected identified rSNPs, which are associated with the PA and FA responses respectively, and are likely to be occurring to the binding positions and regulating or have affinity with respective RBPs. Most of these rSNPs have been also predicted to be involved with other regulatory functions above. In addition, several of these predicted rSNPs are likely to be occurring in the same position where the RNA proteins bind.

Dataset	**SNP id**	**Genomic coordinate**	**Motif sequence (Binding site)**	**Possible binding position**	**Bound protein(s) (RBPs)**	**Significance (p-value)**
Dataset 1	**rs3730051**	*chr19:40238790*	ugrwgvh	chr19:40238791	SRSF1	3.14e-03
	rs6141803	*chr20:32752550*	ugrwgvh	chr20:32752551	SRSF1	1.73e04
Dataset 2	rs5277	*chr1:186679065*	kgugukk	chr1:186679068	BRUNOL4	1.18e03
			ugugukk	chr1:186679070	BRUNOL5	3.85e04
			ugugug	chr1:186679068	TARDP	2.88e03
			aaguguu	chr1:186679069	TRA2B	1.66e04
	rs1472122	*chr3:151316814*	guaguagu	chr3:151316819	HNRNPA1	2.40e04
	rs17041401	*chr3:4768285*	yywcwsg	chr3:4768286	SRSF5	6.30e03
	rs41307147	*chr6:73817320*	chuuuuu	chr6:73817320	CPEB2	3.39e04
			uuuuuu	chr6:73817320	CPEB4	2.29e04
			huuuuuk	chr6:73817320	HNRNPC	1.02e04
			uukruuu	chr6:73817320	HuR	7.75e04
Dataset3	**rs2289171**	*chr2:208339990*	uguanaua	chr2:208339990	PUM2	1.16e04

Table 4.20 The identified significant rSNPs that are associated with the PA and binding affinity of the different RBP(s). Each SNP is shown with its genomic position or locus, the motif(s) and binding site (or genomic position) in which the RNA protein(s) are likely to bind, the predicted RNA binding protein(s) and the statistical significance of the binding site position. In bold are the newly discovered significant SNPs, which were not previously identified in the study by Jones et al. (2009).

From Table 4.20, it can be seen clearly that most of these rSNPs are predicted to be in the vicinity or within the binding sites of RBPs, which may bind to different motif(s). The most significant rSNP is rs41307147 in the *chr6:73817320* or intron of the *CD109* locus. This SNP might be localising in many different motifs, which are likely to be bound with different RBPs with the same binding position.

Dataset	SNP id	Genomic coordinate	Motif sequence (Binding site)	Possible binding position	Bound protein(s) (RBPs)	Significance (p-value)
Dataset 1	**rs7034539**	*chr9:5081585*	wuaauur	chr9:5081585	A1CF	3.11e04
	rs1491978	*chr3:151362282*	ugrwgvh	chr3:151362284	SRSF1	4.23e03
Dataset 2	rs6450105	*chr5:52980479*	aaguguu	chr5:52980476	TRA2B	5.77e03
	rs11637556	*chr15:66436613*	chuuuuu	chr15:66436613	CPEB2	4.77e03
			gcuugc	chr15:66436612	MBNL1	1.74e03
	rs12485738	*chr3:56831748*	rwucaag	chr3:56831748	SNRNP70	1.03e03
			wcwwc	chr3:56831747	SRSF3	8.82e03
Dataset 3	**rs2276829**	*chr3:42527208*	kgugukk	chr3:42527207	BRUNOL4	4.93e03
			ugugukk	chr3:42527207	BRUNOL5	4.02e03
			ugugug	chr3:42527207	TARDBP	1.52e03
	rs17296289	*chr10:32971771*	auaaaav	chr10:32971771	KHDRBS1	5.42e04
			rauaaam	chr10:32971770	KHDRBS2	9.87e04

Table 4.21 The potential identified significant rSNPs that are associated with the FA platelet responses and binding affinity of different RBP(s). In bold are the newly discovered significant SNPs, which were not previously identified in the study by Jones et al. (2009).

Based on the Tables 4.18, 4.19, 4.20 and 4.21, several rSNPs have been predicted to occur in the same position as the RBPS. This indicates that these rSNPs are likely to be regulating or affecting the binding affinity with corresponding RBPs. In addition, some of these RBPs such as HuR or ELAVL1 are not cell-type specific and have been implicated in several biological processes and diseases. Thus, they might be interesting to examine their role in platelet activation and cardiovascular diseases.

4.4.4 Predicted rSNP(s) that are involved with the chromatin state and histone modifications

Only one intron key SNP was found to be likely involved with the underlying chromatin state and histone modifications in the bone marrow derived cells (megakaryocytes), which are platelets factory. Table 4.22 represents the chromatin state and histone modifications and their associated rSNP(s) for PA.

Methods	Identified rSNPs	Gene locus	Predicted Chromatin and Histone modifications	Cell type Chromatin structure
rSNPBase	rs1472122	*P2YR12*	H3K4me1	bone_marrow_HS5

Table 4.22 The identified rSNP associated with PA that is likely to be involved with the chromatin state and histone modifications in bone-marrow related tissue type. The predicted chromatin and histone modification is histone H3 lysine 4 mono-methylation (H3K4me1) in distal enhancer.

The identified SNP is rs1472122 in *P2YR12*, which is likely to be involved with H3K4me1 histone modification in the distal enhancer of the gene potentially *P2RY12*. This is in line with the predictions given by GWAS3D related to the same SNP, which also shows to be likely to be involved with enhancer. This means the presence of the H3K4me1 mark may likely to indicate the underexpression of the related gene (Calo and Wysocka, 2013), which in this case is *P2RY12*. This result might tie with those from the RAPIDSNPs in chapter 2, which showed that the SNP is significantly associated with low response of PA. Moreover, it could be suggested that the potential presence of H3K4me1 in the enhancer of *P2Y12* may likely to lead or contribute to the individuals decrease of the expression of *P2RY12* that may likely to decrease the PA response and platelet aggregation. Intriguingly and in contrast, Jones et al. found that this SNP is associated with the high PA response (Jones et al., 2009).

4.4.5 Summary of results for the most significantly predicted rSNPs out of those identified by the RAPIDSNPs pipeline

Several of the key SNPs identified using the RAPIDSNPs pipeline across the three datasets are predicted to be significantly associated with several regulatory roles. The potential involvement of the SNPs in different regulatory roles indicates their underlying significance at the molecular level that is likely to contribute to the differential ADP platelet responses among individuals. Table 4.23 summarises the predicted rSNPs and their different regulatory roles.

Dataset	FA/PA	Identified rSNPs	Locus/ Genomic Position	Likely regulatory roles
2	FA	**rs12485738**	*ARHGEF3*	eQTL, TFBS regulation, regulation of RNA binding proteins
2	PA	**rs6057638**	*ch20:32751526*	eQTL, TFBS regulation, distal regulation,
		rs3788337	*GNAZ*	TFBS regulation, proximal and distal regulation, regulation of RNA binding proteins, long range interactions.
		rs2815805	*MAPK14*	eQTL, TFBS regulation, proximal and distal regulation, regulation of RNA binding proteins
		rs1472122	*P2RY12*	TFBS regulation, proximal regulation, chromatin and histone modifications, regulation of RNA binding proteins
3		**rs906766**	*MED12L*	TFBS regulation, proximal regulation, regulation of RNA binding proteins
		rs246406	*ITGA2*	eQTL, TFBS regulation, distal regulation, long range interactions
1	FA	**rs1491978**	*P2RY12*	eQTL, TFBS regulation, regulation, regulation of RNA binding proteins, long range interactions
		rs3736101	*MADD*	TFBS regulation, proximal regulation.
		rs7034539	*JAK2*	eQTL, TFBS regulation, regulation of RNA binding proteins
		rs3729931	*RAF1*	eQTL, TFBS regulation, distal regulation, regulation of RNA binding proteins.
		rs7180408	*GTF2A2*	TFBS regulation, distal regulation
		rs17296289	*LOC101929475*	TFBS regulation, regulation of RNA binding proteins
		rs17204376	*GPR87*	eQTL, TFBS regulation, proximal regulation, regulation of RNA binding proteins
	PA	**rs8033381**	*CSK*	TFBS regulation, proximal and distal regulation, regulation of RNA binding proteins
		rs2300065	*SKP1*	TFBS regulation, proximal and distal regulation, regulation of RNA binding proteins
		rs3730051	*AKT2*	TFBS regulation, proximal and distal regulation, regulation of RNA binding proteins

Table 4.23 The key SNPs from the RAPIDSNPs, which have been predicted to be involved with many regulatory roles. These SNPs were selected based on their likely involvement in at least two regulatory mechanisms for PA and FA platelet responses. In bold are the newly discovered significant SNPs, which were not previously identified in the study by Jones et al. (2009).

4.4.5.1 ***Non-deleterious, synonymous key SNPs with highly regulatory roles***

From Figure 1.14 in Chapter 1, we identified SNPs, which are missense (non-deleterious), or synonymous and thus, were not further investigated for their effect on the encoded proteins' structural/function in the previous chapter (Chapter 3). However, in this chapter, these SNPs were found to be involved with many regulatory roles (Table 4.24), and hence are likely to be contributing to the individuals' ADP platelet variability. These include the non-deleterious missense SNP rs3736101 in *MADD*, stop gained synonymous SNP rs2228671 in *LDLR* and stop lost synonymous SNP rs2815805 in *MAPK14*.

Furthermore, the identified *deleterious* nsSNPs (missense SNPs) in Chapter 3 are all unlikely to be involved with any of the regulatory functions, as none of them were predicted to be rSNP by the designed pipeline. This molecular feature may suggest that the damaging missense SNPs are unlikely to be perturbing regulatory mechanisms underpinning the complex trait (ADP platelet responses), though further investigation is needed.

4.5 Discussion

4.5.1 Predicted rSNPs that are involved with eQTL, and their likely association with ADP platelet responses and CVD

From Table 4.23, it is clear that the identified intronic, non-coding, non-deleterious missense, and synonymous key SNPs from the RAPIDSNPs are likely to be involved with many regulatory mechanisms. These mechanisms are likely to be contributing to the underlying ADP platelet activation variability and potentially contribute to the genetic basis of CVD risk. For instance, in the eQTL, the most interesting rSNP is rs12485738 in the *ARHGEF3* locus, which is significantly associated with the FA response and an increase of platelet aggregation. This rSNP has been found to be associated with many eQTL genes. Several of the related eQTL

genes are likely to be associated with the ADP platelet responses and CVD. Expression levels of one of the identified eQTL genes, *MYL9* (myosin light chain 9), are predicted to be regulated by rs12485738. The proteomic analyses show that the protein encoded by *MYL9* plays a critical role in the signalling of integrin αIIbβ3, which is the required receptor for fibrinogen binding during the ADP activated platelet response aggregation (Fröbel et al., 2013). Thus, the likely regulation of the expression level of MYL9 by rs12485738 may contribute to the upregulation of the platelet aggregation. Moreover, eQTL genes associated with rs12485738 may include *ADCY3*, *CD226*, *ADCY6*, *ARHGEF12*, *ARHGAP21*, *TLN1*, *VWF*, *CD9*, and *CETP*, which are discussed below.

CD226

CD226 encodes an adhesion molecule CD226, which mediates the binding of thrombin-activated platelets to endothelial cells and induces platelet adhesion (Kojima et al., 2003). The involvement of CD226 is likely to occur during the secondary platelet aggregation where ADP is involved (Jiang et al., 2013).

***ADCY3* and *ADCY6* isoforms**

Two isoforms of adenylate cyclase, *ADCY3* and *ADCY6*, which encode ADCY3 and ADCY6 respectively, are known to be associated with the formation of cyclic adenosine monophosphate (cAMP). Both isoforms are known to be associated with the de-aggregation of the platelet and are inhibited by the G(i) protein coupled receptors (Katsel et al., 2003). The cAMP is used to inhibit ADP activated platelet aggregation under the absence of G protein-coupled receptors and hence downregulate aggregation (Woulfe et al., 2001). In addition, the inhibition of ADCY6 was previously reported to be associated with the decrease of cAMP levels under thrombin stimulation (Werthmann et al., 2009). Therefore, the likely regulation of the expression levels of the ADCY3 and ADCY6 by the rs12485738

SNP might result in the possible downregulation of the cAMP and hence increased platelet aggregation. In addition, the ADCY6 isoform has been also associated with CVD (Hodges et al., 2010).

ARHGEF12

The proteins encoded by this gene are exchange proteins, which have been reported to form a complex with G-proteins and thus, it has been suggested that ARHGEF12 may be participating in the platelet activation and aggregation (Zou et al., 2014). The roles of G-proteins with respect to the ADP platelet responses are well characterised in several papers, where together with the G-protein coupled receptors, they are reported to mediate the ADP induced platelet responses and aggregation (Klages et al., 1999; Offermanns et al., 1997, 1994; Ohlmann et al., 1995; Woulfe et al., 2001). Additionally, rs12485738 is in *ARHGEF3*, which encodes another isoform of ARHGEF and has been elsewhere reported to have association with the MI and ischemic stroke (Meisinger et al., 2009). Therefore, the likely regulation of the expression levels of the ARHGEF12 by the rs12485738 SNP might have an effect on the underlying differential regulation of the ADP platelet aggregation and hence, could also be a marker for CVD risk.

ARHGAP21

The likely regulation of the expression levels of this gene by the rs12485738 SNP might be vital to the mechanisms underpinning the ADP platelet responses and CVD. This is because ARHGAP21 is a GTPase activating protein for Cdc42, which has been reported to be associated with the increase of ADP induced platelet activation and aggregation, and CVD (Pleines et al., 2010; Sinha and Yang, 2008).

TLN1

TLN1 (Talin 1) is another eQTL gene, the expression levels of which are likely to be regulated by the rs12485738 rSNP. The major function of *TLN1* is to activate integrin αIIbβ3 for binding of key ligands, such as the fibrinogen required for the platelet aggregation in the inside-out signalling fashion (Petrich et al., 2007). The activation is achieved when the two subdomains, F2/F3, of the expressed *TLN1* form a complex with αIIbβ3 (Provasi et al., 2014). Moreover, the activated integrin and binding to its ligand (fibrinogen) brings about outside-in signalling, which is crucial for further platelet responses, recruitment, spreading, adhesion and aggregation (Li et al., 2010). Thus, the likely regulation of the expression levels of *TLN1* by the rs12485738 rSNP might be critical for both inside-out and outside-in signalling. In addition, the mediation of the latter two processes was reported to be carried out by the G-proteins (Offermanns, 2006; Shen et al., 2012). It is well known that G-proteins are central to the ADP platelet responses and aggregation. Moreover, the deficiency or decrease in *TLN1* expression is thought to be associated with decrease of platelet aggregation and blood related disease (Fröbel et al., 2013). Overall, the rSNP rs12485738 regulated *TLN1* expression levels might be critical to the FA platelet response and CVD, and hence, warrants further investigation.

VWF (von Willebrand Factor)

VWF is another eQTL gene, which is predicted to be regulated by the rSNP rs12485738. *VWF* encodes an adhesive ligand, which forms large multimeric complexes and plays vital role in the increase of platelet aggregation mediated by the binding of fibrinogen to integrin αIIbβ3 under flow condition (Ikeda et al., 1991; Schmugge et al., 2003). Elsewhere, it was reported that VWF mediates the ADP-enhanced platelet activation, aggregation and persistent thrombus formation through P2Y1 and P2Y12 under shear flow (Mazzucato et al., 2004). Thus, it is likely that upregulation of the expression level of *VWF* by rs12485738 may increase the FA and

platelet aggregation. Additionally, v*WF*, is a target for antithrombotic treatments (Ruggeri, 1992), which might be potentially optimised for personalised CVD therapeutics based on its predicted regulation by rs12485738.

CD9

This is another eQTL gene, which is expressed in the platelet surface and its expression plays critical role in the platelet activation and adhesion likely in complex with other surface molecules (Lanza et al., 1991; Worthington et al., 1990). Furthermore, it was previously reported that the expressed *CD9* would co-localise with glycoprotein GPIIb-IIIa (integrin αIIbβ3) and mediate the fibrinogen binding during platelets adhesion and aggregation (Brisson et al., 1997). As rs12485738 associates with the FA response, then this suggests that it might be playing a role in increasing the expression levels of *CD9* and therefore platelet aggregation.

CETP (cholesteryl ester transfer protein)

The expression of *CETP* is associated with the regulation of cholesterol metabolisms and platelet counts (Hildebrand et al., 2010). Moreover, it has been linked with the risks of several CVD events such as MI, atherosclerosis, and coronary artery disease (CAD) and their related therapy (de Keyser et al., 2011; El Bouhassani et al., 2011; Kimura et al., 2011; Ridker et al., 2009). Thus, the predicted regulation of its expression levels by rs12485738 might be crucial in enhancing the CVD individualised treatment.

Moreover, the rs6141803 SNP in *COMMD7* was also identified to be likely influencing the expression levels of many platelet-related genes including *COMMD7*. Thus, the expression level of this gene is likely to be associated with risk of Myocardial Infarction (MI) (since this SNP was also reported to be associated with the risk of MI (Goodall *et al.*, 2010). In addition,

this SNP is associated with eQTL genes *CDK5RAP1* and *CPNE1* (or *CPN1*), which are likely to be physiologically associated with platelet functioning and increased CVD risk (Matthews et al., 2004; Zimman et al., 2014).

Taken together, the above discussed genes with their associated significant and likely biological interesting rSNPs such as rs12485738 are worth for experimental investigation, which may further illuminate our understanding of the molecular mechanism underpinning ADP platelet responses.

4.5.2 Predicted rSNPs occurring in the TFBS and their likely association with ADP platelet responses and aggregation

For transcription factor binding regulation, several rSNPs have been predicted to be occur within the TF binding sites or have binding affinity with several transcription factors. Thus, they are likely to be differentially regulating transcription of the involved genes and hence, potentially contribute to the ADP platelet response variability and CVD risks. Table 4.24 shows these rSNPs with their likely TFs and target genes.

rSNP	PA/ FA	Gene/Locus	Predicted TF	Selected potential target ADP platelet related genes	Decrease/increase of platelet aggr.
rs6057638	*PA*	*ch20:32751526*	GATA2	*GNB1* (Offermanns et al., 1994) *GBP1*(Lubeseder-Martellato et al., 2002)	Increase
rs3212391		*ITGA2*	STAT1, PARP	*APOE* (Riddell et al., 1997), *GNAZ* (Brass et al., 1993), *PARP* (Alexy et al., 2004)	Decrease
rs6141803		*COMMD7*	CEBPB	*PTGS1* (Catella-Lawson et al., 2001; Vane et al., 1994)	Decrease
			GATA1/2	*LRRFIP1* (Goodall *et al.*, 2010)	Increase
rs3788337		*GNAZ*	AR	*TLN1* (Nieswandt et al., 2007), *MAPK14* (Navarro-Núñez et al., 2010)	Decrease
			STAT5A	*ESR2* (Jayachandran et al., 2010)	Increase
rs2289171		*PIKFYVE*	MAFK	*PTK2B* (Cipolla et al., 2013),	Increase
rs3730051		*AKT2*	PAX5	*ESR1* (Jin et al., 2006)	*Decrease*
rs1472122		*P2YR12/ GPR87*	STAT3, CEBPB	*JAK3*(Tibbles et al., 2001), *PTGS1* (Catella-Lawson et al., 2001; Vane et al., 1994), (Siess and Tigyi, 2004)	*Decrease*
rs5277		*PTGS2*	POL2	*SP3*(Meinders et al., 2015)	*Increase (Potential)*
rs3729931	*FA*	*RAF1*	STAT1, PPARG	PPARG (Ray et al., 2008)	*Decrease*
rs12485738		*ARHGEF3*	FOXA1	*SP3*(Meinders et al., 2015),	*Increase*

Table 4.24 The most often predicted rSNPs, which are associated with several TFBS. These rSNPs are likely to be in the binding sites of the target gene(s). The predicted TFs is followed by selected potential gene(s) that they regulate. In bold are previously unidentified SNPs in the Jones et al. (2009) study.

4.5.3 Summary of rSNPs that are likely to be associated with different CVD risks

There are numerous key SNPs identified using RAPIDSNPs that are predicted to be regulatory and involved with many of the molecular functions, as described above. Additionally, these SNPs are likely to be associated with different CVD risks, based on whether the involved genes or SNPs are predicted to increase or decrease platelet aggregation (Table 4.24). The association with CVD can be directly or indirectly inferred through the SNP itself or related eQTL genes/TFs from the literature. Thus, there are new SNPs in addition to those mentioned in the Chapter 2 that are likely to be associated with CVD, or have potential to be associated with CVD. Table 4.25 provides the summary of the predicted rSNPs, which are likely to be associated with CVD.

rSNP	Gene/Locus	ADP Platelet response	Regulatory roles	Involved gene(s) / TF(s)	CVD risk type	reference
rs12485738	*ARHGEF3*	FA	eQTL, TFB	*ADCY6*, *CETP*, FOXA1	Blood pressure, Potential for MI & ischemic stroke	(Hodges et al., 2010), (Meisinger et al., 2009)
rs6141803	*COMMD7*	PA	eQTL, TFB	*CDK5RAP1*, *CPN1*, GATA1/2, CEBPB	MI	(Goodall *et al.*, 2010)
rs1491978	*P2RY12*	FA	eQTL, TFB	*P2RY12*, NFAT	Potential for MI & ischemic stroke	(Simon et al., 2009; Zee et al., 2008)
rs3729931	*RAF1*	FA	eQTL, TFB	PPARG, STAT1	Heart Failure/ Hypertension	(Ray et al., 2006)/ (Parsa et al., 2011)
rs3212391	*ITGA2*	PA	TFB, distal l. interactions	PARP, PPARG	Atherosclerosis (Stroke), Hypertension	(Deng and Shen, 2007; Gardener et al., 2011), (Ray et al., 2006)
rs1472122	*P2YR12*	PA	eQTL, TFB	CEBPB,	Potential for Ischemic Stroke	(Siess and Tigyi, 2004)
rs5277	*PTGS2*	PA	TFB	POL2/SP3	Potential for Ischemic stroke	(Kraus et al., 2013; Maguire et al., 2011)

Table 4.25 The predicted rSNPs, which are likely to be involved with CVD. In bold are previously unidentified SNPs in the Jones et al. (2009) study.

4.6 Conclusion

Based on Table 4.25, half of the identified SNPs, are new and were not previously identified in the Jones et al. study (Jones et al., 2009) implying that the RAPIDSNPs pipeline is capable of identifying, potentially crucial, novel rSNPs, which are likely to be associated with CVD. Moreover, not all of the SNPs that have been identified by RAPIDSNPs (Table 2.20 in Chapter 2) and shown to have a likely association with CVD, have been identified to be regulatory.

Taken together, the results may suggest that most of the identified key SNPs from the RAPIDSNPs are indeed likely to associate to several alternative mechanisms that may underlie ADP platelet responses variability. Additionally, the molecular mechanisms of the several

previously identified SNPs (Jones et al., 2009), have been further elucidated. These rSNPs also include rare variants, such as rs41305896 in the *ITGA2* gene, which is likely to be associated with transcription factor binding regulation and increased FA platelet responses.

Furthermore, most of the regulatory mechanisms that are likely to be involved with the predicted rSNPs are differential gene expression levels (eQTL), and gene transcription. This may suggest that the identified key SNPs are likely to down- or up-regulate expression levels depending on the individuals' genotypes. Moreover, they are likely to occur in the transcription factor binding sites of several related genes. The likely presence of these rSNPs in the TFBS may have an effect on the binding affinity of the related TFs and hence lead to differential transcript variants of the targeted genes. Moreover, numerous rSNPs were predicted to be likely involved with many other regulatory mechanisms such as, RNA protein binding and proximal and distal regulations of the related TFBS. Each of these predicted regulatory roles may further underpin the genetic basis of the variability of ADP platelet responses and aggregation, which may contribute to differential CVD risk levels.

The next chapter describes the development of a gene based prediction method for identifying individuals with extreme ADP platelet responses levels, based on the identified SNPs. The presumption is that some of the rSNPs identified in this chapter (along with the coding SNPs characterised in Chapter 3) may help to increase the predictive accuracy of the genetic models. Thus, the analysed SNPs in this and Chapter 3 are investigated their involvement in predicting high or low ADP platelet response levels in allelic specific manner, which may further explain genetic and molecular risk of CVD.

Chapter 5 - Predicting ADP platelet response levels using SNPs

5.0 Abstract

A major challenge in medicine is to understand and predict disease risk. Predicting disease risk could be achieved via elucidation of the intermediate phenotypes of the complex trait (biological system), such as its molecular mechanisms. The mechanistic combination of the complex trait prediction using genetics (or genotypic information) and prediction of intermediate phenotypes, may have profound impact on treatment outcomes. The last two chapters endeavoured in predicting the molecular mechanisms (intermediate phenotypes) due to the genetic variants identified in the Chapter 2. In this chapter, the genotypic information related to the SNPs, which were identified in the Chapter 2, are used to predict the ADP platelet responses levels of the individuals. It is postulated that combining this type of prediction with the predictive results from the last two chapters, may further illuminate the understanding of ADP activated platelet responses, CVD risk. In the long term, the results of these studies may add value to the current clinical strategies for CVD.

We have designed a predictive approach based on a supervised machine learning method (using an artificial neural network or ANN), which predicts the individuals' low or high ADP platelet response levels. Two underlying hypotheses are: 1) the specific SNPs genotypes among individuals are likely to contribute to varying individuals' extreme high or low expressed ADP platelet response levels, 2) key SNPs from the RAPIDSNPs significantly contribute to increasing the accuracy of predicting the individuals' low or high ADP platelet response levels. The main focus of the chapter is to find the specific individuals SNPs' alleles, which predictably contribute to the extreme high or low ADP platelet response levels. This information will be important for inferring the extent of the involvement of the genetic factor (SNPs) and their predicted molecular mechanisms (intermediate phenotypes) in complex trait variability. Moreover, accurate predictions of the ADP platelet response levels might also allow for

inferences to be made regarding the individuals' CVD risk levels, due to the combined genetic factors and molecular intermediate phenotypes. This could have potential to influence future CVD personalised healthcare settings and 'omic driven health policy development. The major assumption in this work is that environmental factors such as non-genetic, epigenetics and other covariates are controlled in the input data set.

Based on our approach, the key SNPs from the RAPIDSNPs have been found to be significant factors for predicting individuals with high or low ADP platelet response levels. The key SNPs were benchmarked against the SNPs identified from Jones et al. study (Jones et al., 2009). Additionally, we have identified the specific SNPs alleles/genotypes, which are likely to significantly contribute to the extreme individuals' low or high ADP platelet response levels. These results may help to further our understanding of individualised CVD risks for application in future personalised medicine (PM) strategies.

5.1 Introduction

In understanding the genetic basis for human disease or complex traits, biomedical experts have two long standing goals: 1) the identification of the disease or trait associated SNPs (variants), and 2) the ability to predict the phenotype variations from the SNPs genotypes (Burga and Lehner, 2013). The former has been largely achieved due to advances in GASs including genome-wide association studies (Burton et al., 2007; Hirschhorn and Daly, 2005), and the latter has now become a mainstream interest (Makowsky et al., 2011; Wray et al., 2013). Predicting phenotypic variations from the genotype is now performed in plant and animal breeding, model systems, and the human population (Campos et al., 2013; Meuwissen et al., 2001; Ober et al., 2012; Wray et al., 2013). For predicting disease or complex traits phenotype from genotype in the human population, the main interest lies in what and how this type of prediction may bring benefit to preventive and personalised medicine (Makowsky et al., 2011).

Most of the disease phenotype predictions are based on statistical models that attempt to relate the genetic variants such as SNPs with a particular phenotype (Burga and Lehner, 2013). In this case, early efforts involved building predictive models for the quantitative traits (Lynch et al., 1998). Nevertheless, there are several challenges/limitations that are to be addressed for the aim of achieving higher accuracy in these predictions (Burga and Lehner, 2013; Wray et al., 2013), even though predicting 100 percent accuracy for PM is improbable (Burga and Lehner, 2012). One of the fundamental challenges is that much of the complex trait or disease phenotype variations are still unexplained, or in other words, there are yet unknown SNPs or variants that might be influencing traits or diseases (Manolio et al., 2009). The RAPIDSNPs approach, which identified the novel (previously unidentified) significant SNPs, helps us further in bridging this gap. In addition, the identified key SNPs from RAPIDSNPs are postulated to help increase the accuracy of the genetic models, for predicting high or low ADP platelet response levels. Hence, the output from RAPIDSNPs may help us to address a major limitation of predicting complex traits with high accuracy - the inability to identify the key SNPs with high genetic effect (Janssens et al., 2006; Wray et al., 2013).

Furthermore, other limitations to designing useful methods may include the consideration of all SNP's with small effects contributing to the complex trait, which are likely to improve the prediction accuracy (de los Campos et al., 2010; Wray et al., 2013). With this regard, there are numerous designed statistical methods that are applied to the prediction of the complex trait phenotypes from genotypes (Campos et al., 2013). These methods could be classified into different types such as multivariate mixed models, Bayesian and penalised methods and other multivariate parametric based approaches (Campos et al., 2013; Stephan et al., 2015). However, most of these statistical methods have been reported elsewhere to be sub-optimal for genotype-phenotype association or prediction (Moore et al., 2010).

Methods that incorporate machine learning (ML) for optimising the selection of useful genetic variants and improving phenotypic prediction have recently gained attention, and have been reported to complement these standard statistical approaches (Okser et al., 2014). The underlying mechanism of supervised ML approaches is to construct a learning genotype-phenotype model using the training examples with useful genetic features and then testing the model with new examples. In doing genetic prediction, these methods might be focusing on predicting the genetic risks of both the continuous and non-continuous (binary or multi-class) complex trait phenotypes (Campos et al., 2013; Makowsky et al., 2011). One of the advantages of the ML methods is that they can maximise the individual model's predictive accuracy and hence have greater potential to be successfully applied in personalised medicine (Ashley et al., 2010; Kraft et al., 2009). Moreover, they can more efficiently deal with the genetic interactions (Lehner, 2007; Moore and Williams, 2009).

In applying these methods for standard genetic prediction, the prediction procedure might follow two steps (Shi et al., 2011). Firstly, selecting the most significant SNPs from the large set of the SNPs, and secondly, use the selected significant SNPs for predicting the SNPs effect on the phenotypic trait (Powell and Zietsch, 2011). However, this strategy should be performed with care as it may result in overestimation of the prediction accuracy, particularly when using the same SNPs in the same sample data during the validation stage. The use of *k*-fold cross-validation during the validation stage is always recommended to mitigate this problem (Powell and Zietsch, 2011; Purcell et al., 2009; Wray et al., 2013). Moreover, the overestimation might be further due to missing the predictive interactions across SNPs, which can lead to overfitting or reduced predictive accuracy (Abraham et al., 2013). The key SNPs from the RAPIDSNPs are postulated to be useful for predictive purposes, as described in the above stated two-way standard procedure for predicting the complex trait. In addition, the k-fold cross-validation (see

the Methods section) was applied in the underlying predictive approach described in this chapter to address overestimation.

Therefore, the study hypothesises that most of the key significant SNPs identified using RAPIDSNPs are likely to contribute to enhanced accuracy in predicting the individuals' low and high ADP platelet responses levels. Moreover, the study investigates combining information about the SNPs' associated molecular mechanisms to further increase predictive accuracy. Such combination appears to be useful despite the fact that it is difficult to understand the underlying detail of the molecular mechanisms (intermediate phenotypes) and how they are involved in increasing the trait predictive accuracy (Burga and Lehner, 2013).

In predicting the ADP platelet response levels, the underlying designed method employs a binary classifier for complex trait/disease risk prediction. This is because the high or low ADP platelet response levels phenotypes are categorical.

The underlying predictive approach employs an artificial neural network (ANN). ANNs can deal with non-linearity and complex interactions of the data, which make it very suitable for genetic predictions (González-Camacho et al., 2012). To date, ANNs have not been widely used for genotype-phenotype complex trait predictions in human. Notable attempts have been made in the prediction of the phenotypic drug resistance (Pasomsub et al., 2010; Wang and Larder, 2003). However, for complex trait phenotypes in other model systems, particularly in plants and livestock, genetic prediction using ANNs has been extensively applied (González-Camacho et al., 2016, 2012; Ornella et al., 2014).

Thus, we evaluate the predictive accuracy of the generated ANN genetic models when they are alternatively run using the key SNPs identified by RAPIDSNPs and those identified in the original study (non-RAPIDSNPs) (Jones et al. SNPs). Based on the model evaluation, a

significant improvement of the model accuracy is observed when the key SNPs identified by RAPIDSNPs are used in the predictive models compared with those from Jones et al. (non-RAPIDSNPs), in predicting individuals' high or low ADP platelet response levels.

5.1.1 Why it is important to predict ADP platelet response levels?

One of the advantages of predicting ADP platelet response levels using SNPs genotypes is that it allows us to gauge the extent to which the key SNPs are involved in the individuals' extreme low or high ADP platelet response levels. Furthermore it allows us to infer the likely involvement of the SNPs' allele in the underlying particular molecular mechanisms contributing to the individuals' low or high ADP levels and hence their CVD genetic risks (Burga and Lehner, 2013). For instance, the missense SNP rs2071676 in *CA9* was found to be likely associated with structural/functional changes of the CA IX. The minor allele of this SNP (see Results & Discussion section) was found to be significant associated with low FA response levels. Then, based on Burga and Lehner, this allele could be attributed or inferred to the structural/functional changes of the CA IX protein.

Moreover, predicting individuals' high or low ADP platelet response levels could be of biomedical importance. It was observed in the Chapter 1 that there is an inter-individual variability in the FA/PA platelet responses with those who showed normal, very low or high platelet responses. These variations are likely to be due to the individual's genetics in addition to non-genetic factors, as previously described. The RAPIDSNPs approach was developed in order to identify the most significant SNPs (key SNPs), which are associated with individuals' variability of ADP platelet responses. Genetically, it could be hypothesised that this variability might be due to variation of the individuals' SNPs alleles/genotypes, which may contribute to the individuals' differently expressed extreme high or low ADP platelet response levels. And thus, the individuals' SNPs genotypic information, which are involved with high or low ADP

platelet response levels could have potential of predicting the CVD risk levels. This type of predictive tool is likely to be informative in future personalised healthcare settings and in policy development.

Thus, the ANN was trained to learn predicting which individuals are likely to be associated with the extreme ADP platelet response levels, based on their SNPs genotypes. The next section describes the details of the underlying predictive approach.

5.2 Methods

The predictive approach is divided into three sub-methods based on the underlying SNPs groups, which are used as inputs. It should be stated that the major aim here is to genetically predict individuals with either high or low ADP platelet response levels from their SNPs genotypes/alleles (i.e. major, minor, or heterozygous). For details about the description of SNPs' alleles, refer to the Chapter 1 (section 1.4.1.1).

There are three types of SNPs groups that are going to be used: - 1) SNPs from Jones et al. findings that are associated with ADP platelet responses (Jones et al., 2009), 2) key SNPs from the RAPIDSNPs, and 3) the randomly selected SNPs from the entire SNPs dataset, including those from 1 & 2.

Then, each of the above SNPs groups are used as inputs to predict high or low PA and FA platelet responses levels using an ANN. The diagram below illustrates a general predictive approach, which involves these SNPs sets.

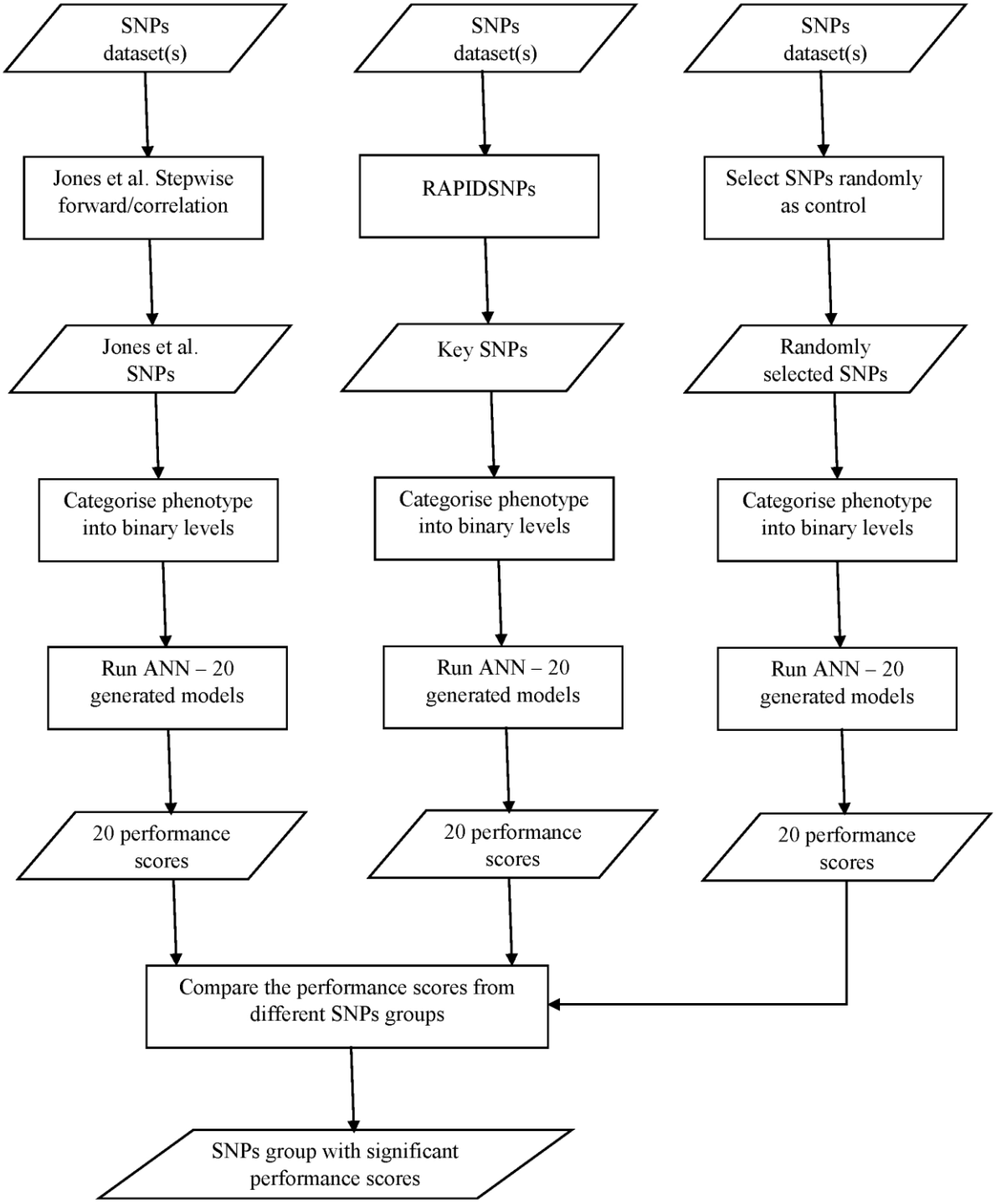


Figure 5.1 The underlying approach for genetically predicting the individuals with high or low ADP platelet response levels. Three different SNPs groups (Jones et al. SNPs or SJ, key SNPs from RAPIDSNPs or SR, & randomly selected SNPs or SD) are used as inputs to the artificial neural network models. These SNPs groups are due to the subsets or datasets 1, or 2 (i.e. the first inputs layer of SNPs dataset(s)). The performance scores of each method based on these groups, are compared to examine the significant and best performing group. Under null hypothesis the key SNPs are not significantly able to significantly predict with high accuracy the individual's high or low ADP levels over Jones at al SNPs.

From the flowchart diagram in Figure 5.1, the randomly selected SNPs group are randomly selected SNPs from either of the two datasets (dataset 1, or 2), which were derived as a result

of treating the missing values contained in the original cohort as explained in the Chapter 2. Moreover, the datasets were used for the SNPs selection based on the two major involved processes, which were either the RAPIDSNPs or Jones et al. stepwise forward approaches. The Jones et al SNPs were essentially taken from the previous study and alternatively used for predictions related to dataset 1 and 2. The binary levels are based on two stage prediction in which the first stage is predicting whether the individual is an extreme or normal ADP platelet responder. The second stage is related with only extreme individuals where the focus is to predict whether the individual is high or low ADP platelet responder. The major reason for this two stage binary prediction is to ensure the models are smoothly able to handle the normal individuals' ADP platelet response levels.

Furthermore, in comparing the performance scores from each SNPs group, 20 different ANN models were generated. The number of models for comparison could be higher than twenty, but twenty were presumed to be a reasonable size for statistically testing the significance of the performance scores.

5.2.1 Transforming ADP platelet responses into two classes – Categorising phenotype into high or low FA/PA levels

Each of the ADP platelet responses (PA and FA) cases were initially categorised into either of the three classes or levels (low, normal, high) based on the empirical distributions as described in the section 1.8 of Chapter 1. The levels were based on the 15 – 85% percentile of the PA or FA traits. In this case, the quantiles $Q_{0.15}$ and $Q_{0.85}$ were used to split each trait phenotype (y_i) into three levels: $y_i \in$ high level, if $y_i > Q_{0.85}$; $y_i \in$ normal level if $Q_{0.15} < y_i < Q_{0.85}$; $y_i \in$ low level if $y_i \leq Q_{0.15}$. Since the largest number of cases are 'normal' individuals, then it was assumed that the model accuracy or performance using a 3 – class problem would be relatively skewed toward the individuals with "normal" ADP levels, due to the cross-validation effect.

Besides, attempting a two-class prediction, by focusing only on high or low ADP levels, is more appropriate for the underlying biomedical purpose of the study, as the individuals of high or low ADP levels are more important for potentially determining CVD risk levels. Therefore, two class predictions were performed in two phases. The first phase involved predicting individuals whether they are likely to be of extreme or normal ADP platelet response levels. The second phase involved predicting extreme individuals whether are high or low ADP platelet response levels. Breaking the prediction into these two sub-tasks enabled to accommodate in the models the individuals with 'normal' cases.

5.2.2 Model induction or fitting (Multilayer Perceptron Neural Network – MLP)

In this study, the MLP was used to classify the individual ADP levels into 2 different disjoint classes S = (C1, C2). Each target class C_i was transformed into target vector of 0 or 1. Example C1 = (1,1) or C2 (0,0), which might imply high or low ADP levels respectively. The input vector of cases and SNPs were in the X matrix of n x p dimensions where n and p represented cases and SNPs respectively. When applying MLP, the method was trained without prior assumption about inputs and class outcomes distribution, which provided the parallel or extra advantage in capturing complex interactions (González-Camacho et al., 2016). The input features for the ANN are SNPs genotypes transformed into binary vectors where each neuron will be receiving one of the three genotype signals, i.e. (1,1), (0,0) or (1,0) representing major, minor, or heterozygous alleles from each SNP respectively.

In this study, the underlying basis of the MLP was represented by the topology, which is shown in the diagram below, Figure 5.2:

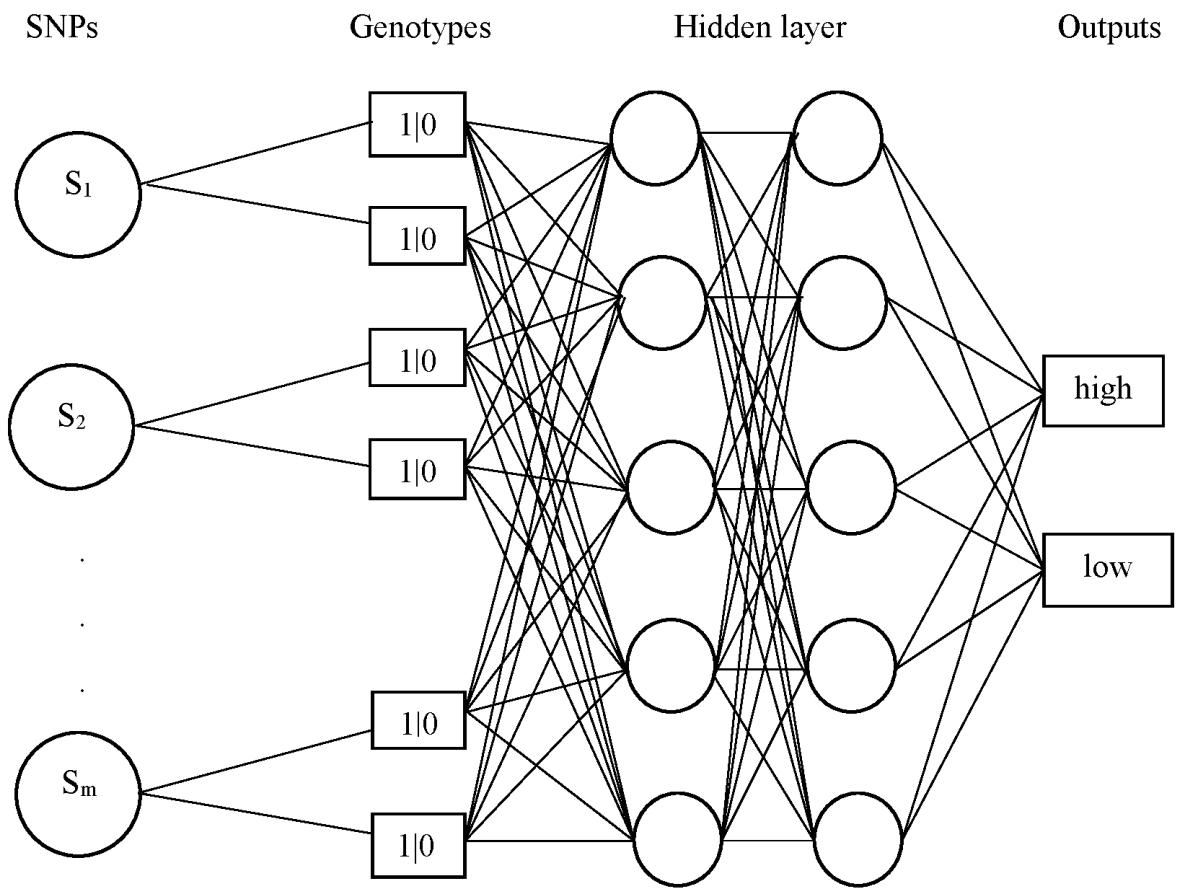


Figure 5.2 The underlying MLP ANN topology used for predicting the high or low ADP platelet response levels. Two Hidden Layers were used for activating the weights when fitting the model using back-propagation algorithm. The input features are each SNP (S_i) genotypes, which might take any of the binary values vectorised as {1,1}, {1,0} or {0,0} for the SNP's major, heterozygous or minor alleles respectively. i.e. each SNP's genotype is represented together by two rectangular shapes or boxes where each may contain and gives a signal of bit (1 or 0) to a neuron Thus, each neuron might take one of the above binary vectors from a single SNP (S_i), which represents a particular genotype.

For this study, the MLP is initially used for classifying binary class disjoint sets with levels, which are: 1) extreme, and normal, 2) low, and high individuals ADP platelet response levels. The MLP ANN topology contained two hidden layers. For all three SNPs group dataset (s), i.e. the randomly selected SNPs (SD), key SNPs from RAPIDSNPs (SR) and Jones et al. SNPs (SD), each hidden layer contains 5 neurons. The activation function for each hidden and output layers are governed by the equations below:

$$Z_m = \sigma(\alpha_{0m} + \alpha_m^T X),\ m = 1, \ldots, M,$$
$$T_k = \beta_{0k} + \beta_k^T Z,\ k = 1, \ldots, K,$$
$$f_k(X) = g_k(T),\ k = 1, \ldots, K,$$

where:

m is an individual input case for the total cases M, $Z = (Z_1, Z_2, \ldots, Z_M)$, and $T = (T_1, T_2, \ldots, T_K)$, Z_m are activated weights (Wij) in the hidden layers derived from the linear combinations of input SNPs genotypes features, which are represented by dimensional matrix X. The sigma $\sigma(v)$ is a sigmoid nonlinear function $1/1 + e^{-v}$, the vectorised class K represents the classes (ADP levels) to be modelled where kth unit is an output probability of class k. There are 2 target measurements Y_k, k =1,2, each is coded as 0 – 1 (one hot encoding) variable for the kth class. Y_k is an output function ($f_k(X)$), which is modelled as linear combinations of the Z_m. There is additional bias unit, which captures the intercepts α_{0m} and β_{0k} in the model above. The output function $g_k(T)$ is *softmax* function for transforming the vector of outputs T into probabilities K classes in which each kth class contains value in the interval [0,1]. The *softmax* function is given as follows:

$$g_k(T) = \frac{e^{T_k}}{\sum_{\ell=1}^{K} e^{T_\ell}}.$$

In fitting the ANN model with above architecture for genetic prediction of the ADP levels, the standard gradient descent *back-propagation* algorithm is used. Briefly, the theory underlying the algorithm is based on the error function that finds the summation of nodes' weights, which will minimise error in predicting the classes. Hastie et al., (2005) has provided further details of the algorithm and theoretical underpinnings.

Furthermore, in fitting the ANN model, the seed was used for random initialisation of the weights and thresholds. In this case, the new random seed was being auto generated and changed in each time when the model was fitted. Each of these random seed numbers were being recorded so as to be reused in fitting the models using the Jones et al. SNPs group. Moreover, as the models were run or implemented using KNIME tool, in most cases, each of these random seed auto generated numbers was observed to be positive or negative large integer (See section 5.3.3).

5.2.3 Model evaluation

In predicting the individuals' extreme and normal ADP response levels, the predictive accuracy of the models was assessed by dividing each subset into training and test cases based on the stratified samples by classes for each of the applied SNPs group (SR, SJ, or SD). This is because there was a sufficient number of individual cases, which were 462 for dataset 1. In this case, absolute partitioning was used in which the training cases contained 300 cases while the test cases contained 162 cases. However, for predicting the extreme cases alone (second stage prediction), i.e. predicting individuals' high or low ADP platelet response levels, the predictive accuracy of the models was assessed through cross-validation procedure. In this case, 5- fold cross-validation based on the stratified samples by classes (ADP platelet response levels) for each SNPs feature group was applied. This is because there were few individual extreme cases with high or low ADP platelet response levels based on the set up quantiles (approximately 140 individuals with high and low ADP platelet response levels). In this case, each cross-validation set for each SNP group contained 28 cases split evenly into high or low individuals. The number of SNPs for each of the SNP groups related to dataset 1 was 7, 10, & 10 for the SR, SJ, & SD groups respectively, which are associated with PA platelet response. For FA platelet response the number of SNPs was 7, 11, & 10 for the SR, SJ, & SD groups respectively.

Furthermore, the models' performance scores were evaluated using the accuracy measure and Cohen's kappa (Ornella et al., 2014). However, based on the assessment, using kappa score provides reasonable performance due its high sensitivity in predicting the true number of high or low ADP platelet response levels comparing to the accuracy measure (see section 5.3.2). In addition, in stratifying the data sets, random seed was initially set to be equal to 10 for reproducing the same results in the initial models. A confusion matrix was also used for evaluating the prediction performance of the models. The confusion matrix is given by the following Table 5.1.

	Classifier predicted value		**Sum**
Observed	1	0	
1	TP	FN	TP + FN
0	FP	TN	FP +TN
Sum	TP + FP	FN + TN	n

Table 5.1 Confusion matrix for binary or two-class classification with observed values and classifier predicted values. In the context of this study, 1 and 0 represent high and low ADP platelet response levels respectively. TP true positive, FP false positive, FN false negative, TN true negative, n is total number of cases.

All computations were performed using R language and KNIME 3.2.1 Analytics platform (Berthold et al., 2008). Figure 5.3 below shows the example of KNIME nodes workflow for ANN modelling. MLP neural network KNIME implementation illustration can be found here (https://www.knime.org/nodeguide/analytics/classification-and-predictive-modelling/example-for-learning-a-neural-network),

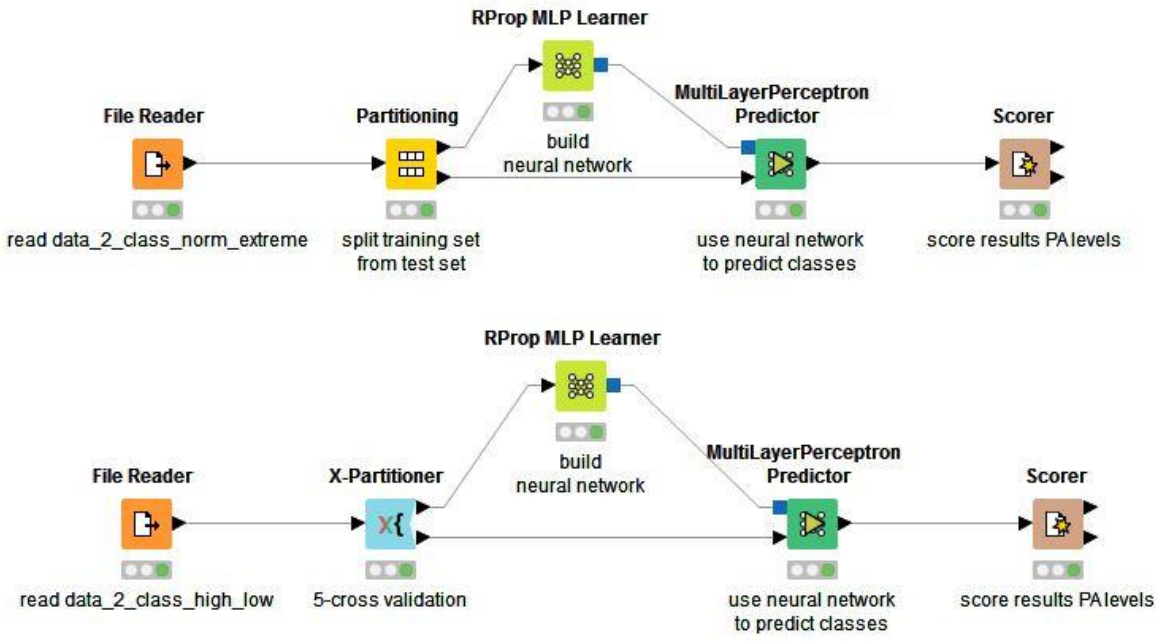


Figure 5.3 The sample KNIME workflow for modelling individuals' extreme or normal, and high or low ADP platelet response levels. The connected nodes show the flow of procedure in modelling the ADP levels. The SNPs data were loaded through the File Reader nodes. Then, the partitioning node was used to split the data into training and test cases. The partitioning of the data with 5-cross-validation was performed using X-Partitioner node, which was then applied to evaluate the trained model. The RProp MLP Learner node was used to train the data and build the ANN model. The MultiLayerPerceptron Predictor node tested or validated the trained model by using 5 cross-validated sets from the X-Partitioner node. The Scorer node evaluated the performance of the model using the Cohen's kappa and Accuracy.

5.2.4 KNIME implementation of ANN

In implementing ANN with back propagation algorithm, two key nodes in KNIME were used, as shown in Figure 5.3. These are RProp MLP Learner and MultiLayerPerceptron Predictor nodes. The configurations of the two nodes, in which the first node is used to train the model while the second is used to test the models were as follows; for the RProp MLP Learner, the total number of iterations for the entire epoch was 100, two hidden layers were used and each contained a maximum of five neurons, the auto generated random seed number was used to initialise the weights. For the MultiLayerPerceptron node, the configuration for this node is optional, which involves changing the attribute name of the new column of the predicted values.

Additionally, in running the ANN models in KNIME, for each generated model, a new RProp MLP Learner node was reset with new auto generated random seed, thus, 20 auto generated random seeds were used for a total of 20 models (see section 5.3.4 in the Results).

5.2.5 Statistical test of the model scores

The driving hypothesis of the chapter is that the key SNPs from RAPIDSNPs contribute most to the increased performance accuracy of the genetic models in predicting high or low ADP platelet response levels. In this case, prediction performance scores from the key SNPs from RAPIDSNPs were benchmarked against those from the Jones et al. SNPs. To test the hypothesis and hence, the statistical significance of the models, a paired t-test was conducted to compare the model scores, which were generated by the key SNPs from RAPIDSNPs (SR) with those from Jones et al. SNPs (SJ) and randomly selected SNPs (SD), which were used as controls. In conducting a paired t-test, 20 models scores were generated for each of the key SNPs and Jones et al. SNPs based on the 20 different random seeds (1, 2, 3, …, 20), which were alternatively set during stratifying the SNPs data in the partitioning and cross-validation stages. Thus, each run of the model was based on these random seeds, which were set before building and running the ANN model (See sub section 5.3.4 of the Results section for further clarification). In addition, for performing this statistical comparison, 20 models with their scores were presumed to be reasonably and statistically sufficient as further indicated with the scores pattern (see section 5.3.4 of the Results). The cut off for models' significance was set to be p-value of <= 0.05.

The kappa scores for 20 models were then noted for the three involved SNPs groups (key SNPs from RAPIDSNPs, randomly selected SNPs, and Jones et al. SNPs). Under null hypothesis, the key SNPs from RAPIDSNPs do not predict the individual's high or low ADP levels with significantly higher accuracy than those from Jones at al. SNPs. Thus, if the null hypothesis is

accepted, then any observed increases in the predictive accuracy of the model using the key SNPs from RAPIDSNPs, might be due to chance.

5.2.6 General prediction procedure

The general genetic prediction procedure (pseudocode) is outlined beneath:

Input: - A random selected SNPs with their genotypes and ADP platelet responses (phenotype)(SD). - Key SNPs with their genotypes and ADP platelet responses (phenotype) from the RAPIDSNPs (SR). - Previously obtained SNPs from Jones et al. with their genotypes and ADP platelet responses (phenotype) (SJ).

1. Transform ADP platelet response phenotype into three classes (low, high, and normal individual levels) based on 15% - 85% percentiles.
2. For each group (i) (SD, SR, and SJ), compute 20 different kappa scores
  a. Initialise total number of random seeds to 1
  b. While total number of random seeds are equal to 20
  i. Select a random seed.
  ii. Perform cross-validation with 5-fold using stratified sampling based on the individuals' high and low ADP levels.
  iii. Predict using ANN and get the kappa score.
  iv. Increment number of random seeds.

3. Perform a paired t-test on the scores for SR and SJ.
4. If the p-value is small (<= 0.05), then it may suggest that the model based on SR significantly outperforms that based on SJ for high and low ADP platelet response levels.
5. Use the Chi-square to test the significance of each SNP(s) and their genotypes in the model related to SR, which significantly predict individuals' high or low ADP levels.

5.2.7 Testing if the SNPs' alleles (genotypes) significantly contribute to high or low ADP platelet response levels.

For the SNPs sets, which significantly predicted, with high accuracy, the individuals with high or low ADP platelet response levels, a Chi-square test was applied to test which of the SNPs genotypes are likely to significantly to contribute to high or low ADP platelet response levels. In addition, if each individual SNP was found to have a significant association with extreme ADP platelet response levels, then its genotypes were further examined to determine which allele (genotype), is significantly related to either a high or low ADP platelet response levels. Hence, this enabled us to determine the allele, which is more likely to be involved with extreme ADP cases. The cut off for significance of the SNP's allele was a p-value of <= 0.01.

5.3 Results and Discussion

5.3.1 Prediction performance based on the dataset 1 – randomly selected SNPs, key SNPs from the RAPIDSNPs and Jones et al (2009). SNPs

5.3.1.1 ***Predicting extreme/normal PA response levels***

In predicting individuals' extreme or normal PA response levels, the following scores in Table 5.2 that are related to the initial models were obtained.

Selected SNPs sets for ANN	Model accuracy		Predicted true number of cases with extreme PA levels	Predicted true number of cases with normal PA levels
	Accuracy (%)	Kappa		
Random selected SNPs (SD)	64.815	-0.002	6/49	91/113
Jones et al. SNPs (SJ)	64.815	**0.098**	14/49	91/113
Key SNPs from RAPIDSNPs (SR)	**66.049**	-0.009	4/49	101/113

Table 5.2 The model scores for predicting extreme/normal individuals' PA response levels using three SNP sets stratified based on random seed equal to 10. The best models' scores are in bold.

From Table 5.2, the SR has high model score in accuracy measure while the SJ set has high model score in kappa indicating that it has an edge or is more sensitive in predicting extreme individuals. In addition, the SR set predicts well the normal individuals over other SNPs sets. However, SJ set outperform both SR and SD in predicting the extreme cases. Table 5.3 shows sample results of the predicted individuals' extreme and normal PA response levels based on the above model from SR SNPs group.

rs8033381	PselectinADP_l	P (PselectinADP_levels=Extreme)	P (PselectinADP_levels=Normal)	Prediction
0	Normal	0.126353965	0.890868561	Normal
1	Normal	0.261744804	0.722601814	Normal
1	Extreme	0.373296344	0.580593067	Normal
0	Normal	0.99994434	5.38E-05	Extreme
1	Normal	0.378374509	0.569413398	Normal
1	Extreme	0.178877646	0.836124727	Normal
1	Extreme	0.382557878	0.564009465	Normal
0	Normal	0.210278386	0.789467889	Normal
0	Normal	0.105305403	0.913963774	Normal

Table 5.3 The sample results of predicted individuals' extreme and normal PA response levels for the initial model fitted using key SNPs set related to dataset 1.

Based on the Tables 5.2 and 5.3, it appears that all of the models do not perform well in predicting the individuals with extreme PA response levels. As there are differences in models' scores, the statistical test is applied to examine the significance the models' scores (section 5.3.4).

5.3.1.2 ***Predicting high/low PA levels***

After applying the initial ANN models on the above SNPs sets (section 5.2.3), different prediction performance scores were examined. The results for predicting high/low PA platelet response levels that are related to dataset 1 are shown in the Table 5.4.

Selected SNPs sets for ANN	**Model accuracy**		**Predicted true number of cases with high ADP levels**	**Predicted true number of cases with low ADP levels**
	Accuracy (%)	**Kappa**		
Random selected SNPs (SD)	35.714	-0.286	5/14	5/14
Jones et al. SNPs (SJ)	46.429	-0.071	4/14	9/14
Key SNPs from RAPIDSNPs (SR)	**82.143**	**0.643**	11/14	12/14

Table 5.4 The model scores for predicting low/high individuals' PA levels using three SNP sets stratified based on random seed equal to 10. The best model score is in bold and from the SR.

Based on the Table 5.4, the model with the highest score appears to be from the SR. However, one score cannot judge the entirety or the significance of the SR in generalising other cases. Hence, section 5.3.4 determines the statistical significance of the differences in the performance of the models. Moreover, Table 5.5 shows sample results of the predicted individuals' high and low PA response levels based on the above model from SR group.

rs8033381	PselectinA	P (PselectinADP_levels=Low)	P (PselectinADP_levels=High	Prediction (Psel
1	Low	0.991690386	0.004773324	Low
0	High	0.02773625	0.976900756	High
1	High	0.116541829	0.902427363	High
1	High	0.448345006	0.558916499	High
0	Low	0.667258715	0.275381419	Low
1	High	0.005604971	0.996510906	High
0	Low	0.069228772	0.945529057	High
0	Low	0.996472933	0.001743898	Low
1	High	0.192823008	0.830757008	High

Table 5.5 The sample results of predicted individuals' extreme and normal PA response levels for the initial model fitted using key SNPs set related to dataset 1.

5.3.1.3 ***Predicting extreme/normal FA levels***

In determining the individuals with extreme or normal FA platelet response levels, Table 5.6 shows scores for the initial models for the involved three SNPs groups.

Selected SNPs sets for ANN	**Model accuracy when the SNP(s) in the model**		**True predicted number of cases with extreme ADP levels**	**True predicted number of cases with normal ADP levels**
	Accuracy (%)	**Kappa**		
Randomly selected SNPs (SD)	69.136	0.127	10/49	102/113
Jones et al. SNPs (SJ)	53.086	-0.155	7/49	79/113
Key SNPs from RAPIDSNPs (SR)	64.815	-0.092	8/49	92/113

Table 5.6 The model scores for predicting individuals' extreme/normal FA platelet response levels based on random seed equal to 10.

From Table 5.6, the control SD SNPs group has higher score than other groups in predicting extreme or normal individuals' FA response levels.

5.3.1.4 ***Predicting high/low FA levels***

Table 5.7 shows the prediction scores of the initial ANN models for predicting high and low individual FA levels based on the above three SNPs sets related to dataset 1.

Selected SNPs sets for ANN	**Model accuracy when the SNP(s) in the model**		**True predicted number of cases with high ADP levels**	**True predicted number of cases with low ADP levels**
	Accuracy (%)	**Kappa**		
Randomly selected SNPs (SD)	42.857	-0.143	4/14	8/14
Jones et al. SNPs (SJ)	53.571	0.071	6/14	9/14
Key SNPs from RAPIDSNPs (SR)	**67.857**	**0.357**	11/14	8/14

Table 5.7. The model scores for predicting individuals' high/low FA platelet response levels based on random seed equal to 10. In this case, RAPIDSNPs has produced a high scoring model comparing to and Jones et al. and randomly SNPs.

From the Table 5.7, the SR group appears to have higher score than other SNPs groups (SJ & SD) in predicting individuals with high or low FA response levels. Therefore, to determine

which of the two is more likely to significantly predict individual with high or low FA platelet response levels, see the statistical test results in section 5.3.6.

5.3.2 Prediction performance based on the dataset 2 – randomly selected SNPs, key SNPs from the RAPIDSNPs and Jones et al (2009). SNPs

5.3.2.1 ***Predicting extreme/normal PA levels***

In predicting the individuals with extreme or normal PA levels related to dataset 2, Table 5.8 shows the initial models scores generated by three SNPs groups (SR, SJ, & SD).

Selected SNPs sets for ANN	**Model accuracy when the SNP(s) in the model**		**True predicted number of cases with extreme PA levels**	**True predicted number of cases with normal PA levels**
	Accuracy (%)	**Kappa**		
Random SNPs (SD)	60.914	-0.06	8/59	112/138
Jones et al. SNPs (SJ)	56.853	-0.004	19/59	93/138
9 Key SNPs from RAPIDSNPs (SR)	**62.437**	**0.04**	15/59	108/138

Table 5.8 The initial model scores for predicting extreme/normal individual PA platelet response levels for dataset 2 based on random seed equal to 10.

From Table 5.8, the SR group has a slightly higher score in predicting the individuals with extreme or normal PA levels related to dataset 2 than other SJ and SD groups.

5.3.2.2 ***Predicting high/low PA levels***

Based on the three above SNPs groups, three initial models with their scores were generated for predicting high/low PA levels, Table 5.9.

Selected SNPs sets for ANN	Model accuracy when the SNP(s) in the model		True predicted number of cases with high PA levels	True predicted number of cases with low PA levels
	Accuracy (%)	Kappa		
Randomly selected SNPs (SD)	60	0.2	6/15	6/15
Jones et al. SNPs (SJ)	56.667	0.133	9/15	8/15
9 Key SNPs from RAPIDSNPs (SR)	**73**	**0.467**	**14/15**	**8/15**

Table 5.9 The model scores for predicting high/low individual PA platelet response levels for dataset 2 based on random seed equal to 10. The SR group appear to have an edge in predicting PA levels among other SNPs groups.

From Table 5.9, the SR group are predicting with relatively higher accuracy, whether the individuals have high or low PA levels.

5.3.2.3 ***Predicting extreme/normal FA levels***

Table 5.10 shows the initial models from three SR, SJ, and SD groups used in predicting the individuals' extreme or normal FA response levels.

Selected SNPs sets for ANN	Model accuracy		True predicted number of cases with extreme FA levels	True predicted number of cases with normal FA levels
	Accuracy	Kappa		
Randomly selected SNPs (SD)	63.452	0.025	12/59	113/138
Jones et al. SNPs (SJ)	64.467	0.009	9/59	118/138
11 Key SNPs from RAPIDSNPs (SR)	**63.452**	0.025	**12/59**	**113/1138**

Table 5.10 The model scores for predicting extreme/normal individual FA platelet response levels for dataset 2 based on random seed equal to 10.

From Table 5.10, the SR and SD groups have similar performance in predicting the individuals' extreme or normal FA response levels though outperformed SJ. In addition, the SJ group has better performance in predicting normal individuals than SR and SD. Since, there is no difference in the performance accuracy between SR and control group (SD), and SJ has higher

score in predicting normal individuals than SR and SD, then the SR are not outperforming other SNPs groups in predicting particularly normal individuals' FA response levels.

5.3.2.4 ***Predicting high/low FA levels***

The initial model scores based on the above SNPs sets for predicting high or low FA levels related to dataset 2 are shown in the Table 5.11.

Selected SNPs sets for ANN	**Model accuracy**		**True predicted number of cases with high FA levels**	**True predicted number of cases with low FA levels**
	Accuracy	Kappa		
Random selected SNPs (SD)	50	-0.067	10/15	4/15
Jones et al. SNPs (SJ)	**70**	**0.4**	8/15	12/15
11 Key SNPs from RAPIDSNPs (SR)	66.667	0.333	10/15	10/15

Table 5.11 The model scores for predicting of the individuals' FA platelet response levels for dataset 3 based on random seed equal to 10. The Jones et al., 2009 SNPs set (SJ) has the highest score.

Based on the Table 5.11, the model with the highest score is achieved by the Jones et al. SNPs set but, as it has been found in other results, the accuracy score alone does not tell us how significant the differences in performance are. Hence, further statistical tests (section 5.3.5) was performed on other 20 scores to determine the significance of the differences in performance.

5.3.3 Statistical significance of the models from the RAPIDSNPs (SR), randomly selected SNPs (SD), Jones et al (2009). SNPs (SJ)

5.3.3.1 ***For predicting individuals with high or low PA platelet response levels related to dataset 1***

For determining the significance of the models in predicting individuals with high or low PA platelet response levels related to dataset 1, the involved three SNPs sets from SJ, SD and SR were statistically tested as described in the section 5.2.5. The performance scores for 20 generated model scores from all three sets are shown in the Table 5.12.

Random seed	Auto-generated random seed initialisation number	key SNPs from the RAPIDSNPs (SR)	random selected SNPs	Jones et al. (10 SNPs) (SJ)
		Model scores (kappa)	Model scores (kappa)	Model scores (kappa)
1	1,556,544,133	0.357	0.143	0.143
2	789,320,481	0.429	0.071	-0.071
3	1,725,422,828	0.429	0.143	0.071
4	-2,139,474,743	0.643	0.357	0.286
5	-1,656,849,783	0.429	0.286	0
6	-1,279,092,758	0.286	0.071	-0.143
7	-882,689,422	0.357	-0.143	0.071
8	42,587,044	0.429	0.071	-0.429
9	-704,996,954	0.571	-0.071	-0.143
10	690,128,337	0.643	-0.286	-0.071
11	-1,889,184,159	0.429	0.071	0.286
12	1,449,888,884	0.143	0	0.143
13	-747,333,527	0.214	-0.071	0.071
14	-1,358,497,622	0.286	0	0.071
15	-1,681,722,665	0.357	-0.071	-0.071
16	-379,921,503	0.571	-0.143	-0.143
17	1,920,482,382	0.429	-0.071	-0.143
18	-1,683,843,630	0.143	0	0.143
19	-2,028,404,516	0.071	0.286	0
20	-1,840,315,556	0.571	0.143	0.214

Table 5.12. The scores of 20 generated models involving three SNPs groups for predicting high and low individuals' PA platelet response levels. The kappa score shows how confident is the model. The higher the kappa score the confident the model.

The above scores from Table 5.12, are further visualised using a boxplot below in Figure 5.5.

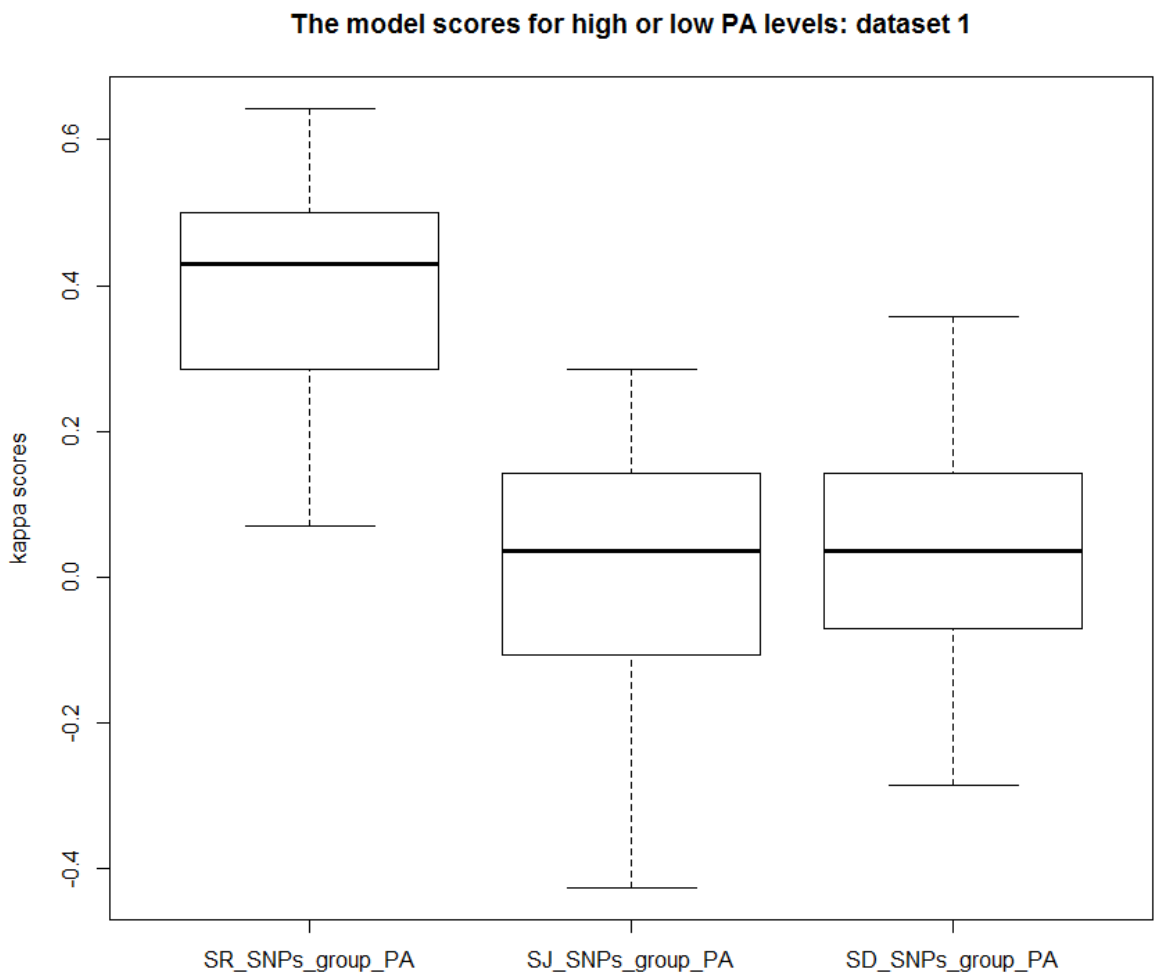


Figure 5.4 The distribution of the kappa scores showing the performance between the key SNPs from RAPIDSNPs (SR_SNPs_group_PA), randomly selected SNPs (SD_SNPs_group_PA) and Jones et al. (2009) SNPs (SJ_SNPs_group_PA) in predicting individuals with high or low PA levels.

In examining the above plot in Figure 5.4, there is an observable mean difference in scores with the highest scores appear to be from the models generated by the SR set. In addition, the range of scores from the SR is higher than other SNPs groups. Thus, the SNPs from SR are likely to have higher scores than the SD and SJ groups. Based on one tailed paired t-test, the models based on the SR are significantly outperforming those based on the SJ for predicting individuals with high or low PA platelet response levels related to dataset 1 at the 95% confidence level (***p-value = 9.891e-07***). In addition, the models from the SR group is significantly outperforming

those based on the SD controls group in predicting high or low PA response levels, (***p-value = 1.669e-06***).

5.3.3.2 ***For predicting individuals with extreme or normal FA levels related to dataset 1***

Table 5.13 shows the scores of 20 generated models predicting individuals' extreme or normal FA response levels.

Random seed	**Auto-generated random seed initialisation number**	**key SNPs from the RAPIDSNPs (SR)**	**Random selected 10 SNPs (SD)**	**Jones et al. (10 SNPs) (SJ)**
		Model scores (kappa)	**Model scores (kappa)**	**Model scores (kappa)**
1	1,938,763,782	-0.039	0.018	-0.022
2	-95,873,046	0.008	0.099	-0.006
3	468,293,175	-0.111	0.044	0.02
4	-883,905,004	-0.029	-0.054	-0.081
5	740,787,778	0.119	0.01	0.027
6	2,037,902,520	-0.063	0.036	-0.106
7	1,160,817,536	0.106	0.003	-0.024
8	589,853,289	0.075	0.09	-0.074
9	-119,798,776	-0.002	-0.026	0.049
10	-924,605,920	-0.026	0.127	-0.155
11	889,651,198	-0.009	-0.025	0.007
12	1,418,251,353	-0.054	0.053	-0.117
13	-237,240,713	0.028	0.015	-0.062
14	823,731,052	-0.026	0.072	-0.091
15	235,662,289	-0.002	0.011	-0.008
16	-2,003,489,478	-0.102	0.02	-0.161
17	-1,441,996,732	0.047	-0.005	0.041
18	-475,782,606	-0.087	0.125	-0.072
19	1,170,682,892	0.025	0.104	-0.099
20	276,789,895	0.049	-0.009	0.055

Table 5.13 The scores of 20 generated models involving three SNPs groups for predicting extreme and normal individuals' FA platelet response levels.

From Table 5.13, the SD SNPs group appear to have higher scores that other groups in predicting the individuals' FA platelet response levels. Figure 5.5 further examines the above scores in Table 5.13.

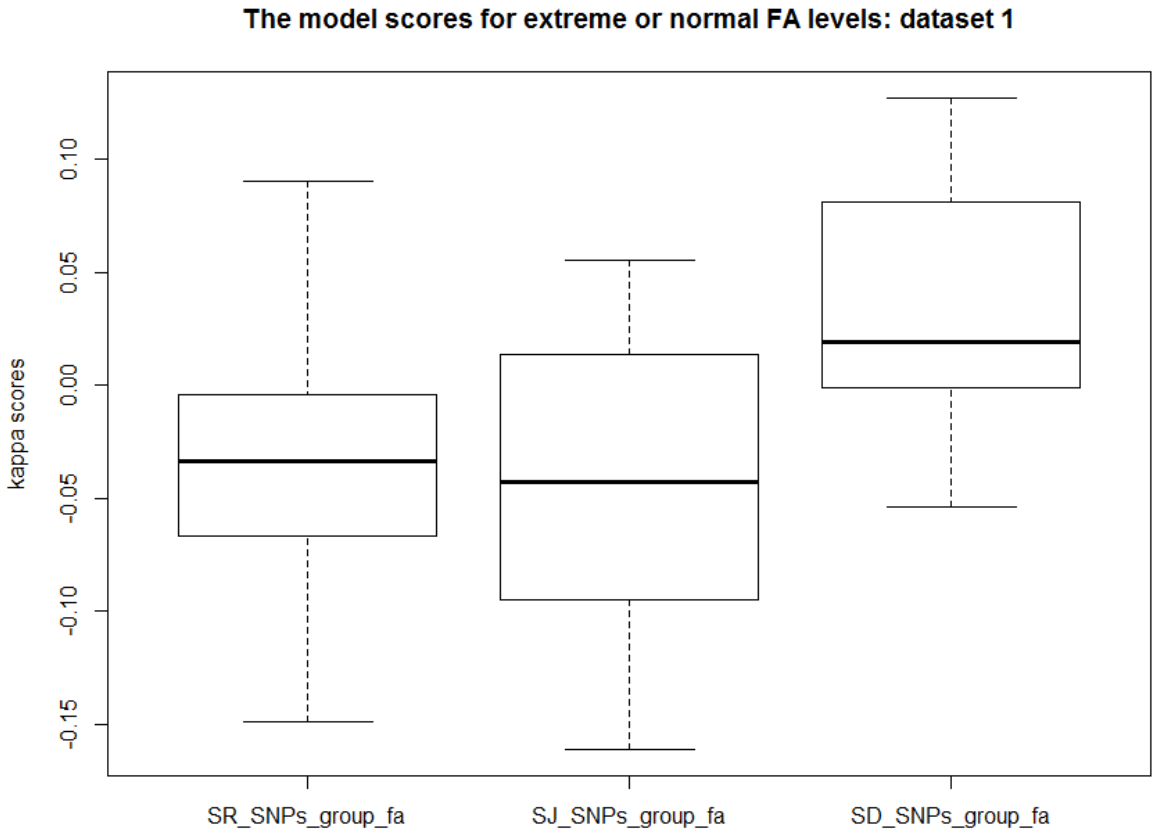


Figure 5.5 The distribution of the kappa scores showing the performance between the key SNPs from RAPIDSNPs (SR_SNPs_group_fa), randomly selected SNPs (SD_SNPs_group_fa) and Jones et al. (2009) SNPs (SJ_SNPs_group_fa) in predicting individuals with extreme or normal FA levels.

From Figure 5.5, the SD group appears to have higher mean score than other groups. However, the SR group appears to have a slight increase in the mean score than the SJ group in predicting individuals' extreme or normal FA levels. Based on the one tailed paired t-test, the SR group is not significantly outperforming the SJ group in predicting individuals' extreme or normal

FA levels (*p-value = 0.234*). Interestingly, the SD is significantly outperforming both SR (***p-value = 0.000147***) and SJ (***p-value = 0.001394***) respectively.

5.3.3.3 ***For predicting individuals with high or low FA levels related to dataset 1***

Furthermore, in predicting individuals' high or low FA levels, the scores from three SNPs group i.e. SD, SJ and SR are shown in Table 5.14.

Random seed	**Auto-generated random seed initialisation number**	**Key SNPs from the RAPIDSNPs (SR)**	**Random selected SNPs (SD)**	**Jones et al. SNPs (SJ)**
		Model scores (kappa)	Model scores (kappa)	Model scores (kappa)
1	-421,726,686	0.286	0.071	0
2	1,398,968,928	0.214	0	-0.429
3	-46,480,488	0.214	0.286	0.071
4	856,319,280	-0.143	0.143	-0.357
5	897,979,989	0.643	0	0.286
6	-1,945,224,330	0.357	-0.143	0.214
7	-797,115,398	0.429	0.286	0.214
8	-67,878,017	0.071	0.071	-0.143
9	-1,263,812,197	0.143	-0.071	0.071
10	1,263,672,307	0.357	-143	0.071
11	1,560,729,502	0.143	0.071	0
12	996,268,964	0.429	-0.071	0.429
13	-1,094,218,337	0.143	-0.143	-0.071
14	-940,900,883	0.5	0	-0.214
15	-93,667,453	0.357	-0.214	0
16	-857,543,894	-0.214	-0.429	-0.071
17	167,713,791	0.143	0.143	-0.214
18	1,116,500,401	-0.143	0	-0.214
19	-1,333,030,801	0.214	-0.429	-0.071
20	188,915,709	0.429	0.071	-0.071

Table 5.14. The scores of 20 generated models involving three SNP sets for predicting individuals with high or low FA levels related dataset 1.

The boxplot below further illustrates the performance differences between the SR, SD and SJ in predicting individual high or low FA levels.

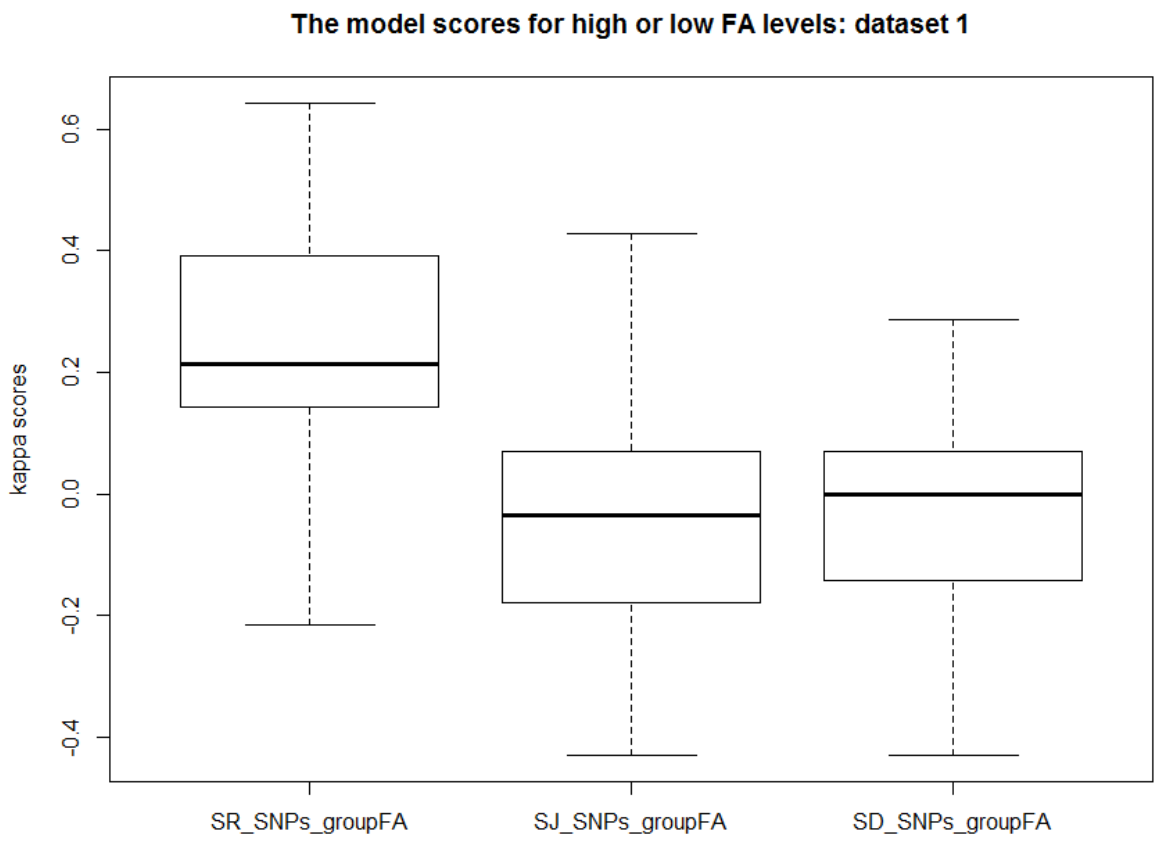


Figure 5.6 The distribution of the kappa scores showing the performance between the key SNPs from RAPIDSNPs (SR_SNPs_groupFA), Jones et al., (2009) SNPs (SJ_SNPs_ groupFA) and randomly selected SNPs (SD_SNPs_ groupFA) in predicting individuals with high or low FA platelet response levels related to dataset 1. Most of the high scores appear to have been achieved by the SR set.

From the Figure 5.6, SR set appear to be performing well in predicting the individuals with high or low FA levels. Based on one tailed a paired t-test, the ANN models using the SR set significantly outperform those using the SJ and SD sets at the 99% confidence level (***p-value = 1.156e-05***) and (***p-value = 0.0002488***) respectively.

5.3.3.4 ***For predicting individuals with extreme or normal PA levels related to dataset 2***

In determining the significance of the model's score from SR comparably to other groups, which were used in predicting the individuals with extreme or normal PA response levels that are related to dataset 2, 20 generated models' scores are shown in the Table 5.15.

Random seed	**Auto-generated random seed initialisation number**	**key SNPs from the RAPIDSNPs (SR)**	**Randomly selected SNPs (SD)**	**Jones et al. SNPs (SJ)**
		Model scores (kappa)	Model scores (kappa)	Model scores (kappa)
1	2,069,119,248	-0.05	-0.006	-0.133
2	623,544,118	-0.005	-0.096	0.133
3	159,738,235	-0.098	-0.051	0
4	1,436,454,075	0.076	-0.075	0.133
5	179,624,322	-0.14	0.007	0.333
6	-15,041,337	0.075	-0.026	0.2
7	-1,069,357,275	0.048	0.037	0
8	-707,736,224	0.128	-0.017	0.133
9	-442,226,650	0.16	-0.029	0.133
10	-2,080,497,923	0.04	-0.06	0
11	1,729,448,743	0.075	0.039	0
12	314,586,795	0.069	-0.044	0.2
13	181,773,638	0.06	-0.055	0.133
14	709,931,303	-0.002	-0.104	-0.2
15	1,872,043,444	0.018	-0.009	0.2
16	-949,821,970	-0.02	0.108	-0.2
17	-2,028,132,021	0.065	0.066	-0.067
18	1,941,732,180	-0.026	-0.083	0.133
19	700,486,429	0.075	-0.076	0.067
20	1,713,533,898	0.178	-0.037	0.067

Table 5.15 The scores of 20 generated models involving three SNP sets for predicting individuals with extreme or normal FA levels related dataset 2.

From Table 5.15, many ANN models from the SJ SNPs group have relatively higher scores than other groups (SR & SD). However, based on paired t-test (one sided & two sided), which compared SJ and SR scores, there is no significance high accuracy in predicting individuals with extreme or normal PA response levels (***p-value = 0.3148*** for SJ > SR, and ***p-value = 0.6295***

for SR = SJ or two sided). Moreover, in comparing with the SD group using one tailed paired t-test, both SR and SJ groups are significantly outperforming SD group, ***p-values: 0.001934*** and ***0.01576*** respectively.

5.3.3.5 ***For predicting individuals with high or low PA levels related to dataset 2***

Table 5.16 shows 20 generated model scores based on all three SNPs sets (SR, SJ, & SD), which were used to predict the individuals with high or low PA levels related to dataset 2.

Random seed	**Auto-generated random seed initialisation number**	**key SNPs from the RAPIDSNPs (SR)**	**Random selected SNPs(SD)**	**Jones et al. SNPs (SJ)**
		Model scores (kappa)	Model scores (kappa)	Model scores (kappa)
1	139,824,644	0.4	-0.2	0.067
2	-1,252,124,782	0.4	-0.133	0.267
3	-621,316,962	0.53	-0.267	0.133
4	1,004,258,364	0.333	-0.467	0.133
5	1,131,094,984	0.333	0.2	0.53
6	833,896,622	0.2	0.333	0.133
7	1,953,698,203	0.067	0	0
8	1,250,020,917	0.133	-0.067	0
9	-749,088,494	0.133	0.333	0.067
10	-1,245,268,760	0.467	0.2	0.133
11	1,151,016,445	0.067	0	0.067
12	-1,310,363,528	0.467	0.267	0.267
13	-77,163,649	0.333	-0.067	0.4
14	625,866,709	0.6	-0.2	0.267
15	-1,848,196,768	0.267	0.133	0.2
16	-2,123,498,527	0.267	-0.133	0.067
17	158,803,932	0.467	-0.267	0.133
18	1,842,726,089	0.467	0	0.4
19	710,805,618	0.4	0	0.267
20	1,571,695,510	0.4	0.067	0.2

Table 5.16 The scores of 20 generated models involving three SNP sets for predicting individuals with high or low PA levels related dataset 2.

In examining Table 5.16, it also appears that the ANN models based on the SR set are predicting the individuals with low or high PA platelet response levels with higher performance than those based on the SJ and SD sets. A one tailed paired t-test shows that the ANN models based on the SR significantly outperform those based on the Jones et al. SNPs and SD at the 95% confidence interval (***p-values: 0.003838 and 0.0001095 respectively***).

5.3.3.6 ***For predicting individuals with high or low FA levels related to dataset 2***

Table 5.17 shows the generated 20 model scores from the three SNPs sets (SR, SJ, & SD).

Random seed	**Auto-generated random seed initialisation number**	**11 key SNPs from the RAPIDSNPs (SR)**	**Random selected 10 SNPs (SD)**	**Jones et al. (10 SNPs) (SJ)**
		Model scores (kappa)	**Model scores (kappa)**	**Model scores (kappa)**
1	-2,121,331,986	0.4	-0.067	0
2	-344,902,931	0.6	-0.067	0
3	2,125,676,383	0.533	-0.267	0.267
4	607,475,327	0.533	-0.333	0.267
5	825,393,109	0.267	0	0.2
6	-202,623,379	0.133	0	0
7	-2,121,868,076	0.4	-0.133	0.267
8	947,198,504	0.4	-0.133	0.667
9	-1,459,234,751	0.467	-0.067	0.4
10	1,771,975,301	0.333	-0.067	0.4
11	-1,588,285,366	0.667	0.067	0.2
12	-1,101,656,008	0.667	-0.4	0.533
13	-1,446,700,401	0.2	-0.2	0.133
14	1,841,546,502	0.4	0	0.067
15	968,230,868	0.467	-0.2	-0.067
16	-543,957,020	0.667	0	0
17	1,648,053,797	0.333	-0.4	0.133
18	348,280,606	0.867	-0.2	0.333
19	-892,795,992	0.4	-0.2	0.267
20	-1,714,724,683	0.533	-0.333	0.2

Table 5.17 The scores of 20 generated models involving three SNP sets for predicting individuals with high or low FA platelet response levels related dataset 2.

The above scores in Table 5.17 can be further illustrated using a plot in Figure 5.7, which explores the performance scores between the key SNPs and Jones et al. SNPs.

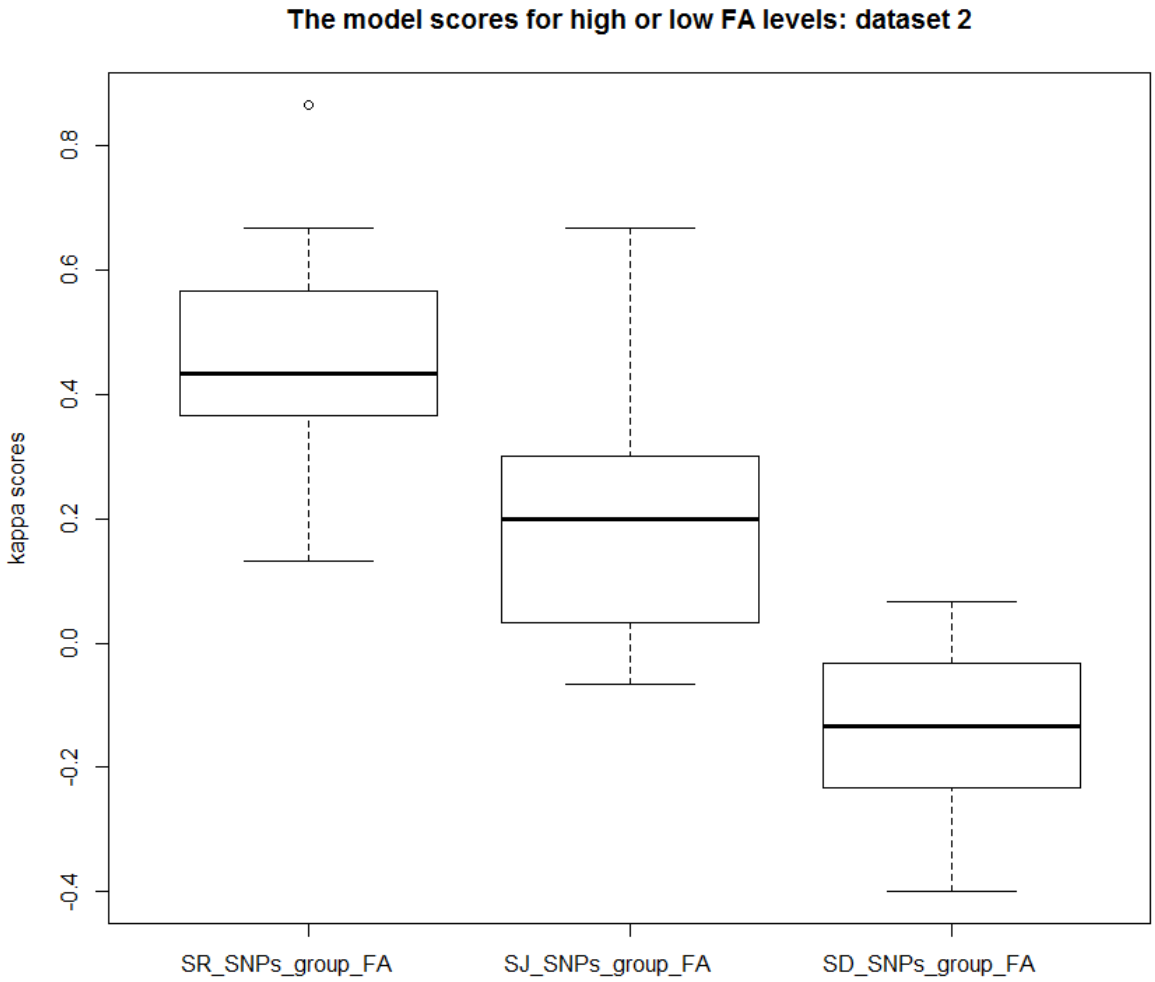


Figure 5.7 The distribution of the kappa scores showing the performance between the key SNPs from RAPIDSNPs (SR_SNPs_group_FA), Jones et al., (2009). SNPs (SJ_SNPs_group_FA) and SD_SNPs_group_FA in predicting individuals with high or low FA levels related to dataset 2.

From Figure 5.7, the SR appears to have the highest significant scores comparing to the other two groups (SJ & SD) in predicting individuals with high or low FA response level. A one tailed paired t-test shows that the ANN models based on the SR significantly outperform those based on the SJ and SD at the 95% confidence interval (***p-value = 7.482e-05 and 3.705e-10***) respectively.

The results summary for testing of the significance of predictions, which involved models from both SR, SJ and SD groups in two datasets are shown in the Tables 5.18 and 5.19 for extreme/**normal** and high/low predictions respectively.

Datasets	**ADP platelet response & p-values**	**Which SNPs group is significant?**	**ADP platelet response & p-values**	**Which SNPs group is significant?**
	FA	*RAPIDSNPs' key SNPs (SR) / Jones et al. SNPs (SJ) / Randomly selected SNPs (SD)*	*PA*	*SR / SJ / SD*
Dataset 1	**0.000147**	SD	0.05032	*SR*
Dataset 2	**N/A**	No difference in prediction scores	**0.001934 and 0.01576**	*SR & SJ respectively*

Table 5.18 The summary of the significance of the models in predicting the individuals normal ADP platelet response levels. Bolded are p-values, which show the models' significance in predicting either normal PA/FA platelet response levels.

From Table 5.18, the SR have overall significant predictions over SJ and the control SD in predicting the normal PA platelet response levels. In addition, both SR & SJ are less significant than SD, which was used as control in predicting normal individuals FA response levels related to dataset 1.

Datasets	**ADP platelet response & p-values**	**Which SNPs group is significant?**	**ADP platelet response & p-values**	**Which SNPs group is significant?**
	FA	*RAPIDSNPs' key SNPs (SR) / Jones et al. SNPs (SJ) / Randomly selected SNPs (SD)*	*PA*	*SR / SJ / SD*
Dataset 1	**1.156e-05**	SR	**9.891e-07**	*SR*
Dataset 2	**7.482e-05**	SR	**0.00383**	*SR*

Table 5.19 The summary of the significance of the models in predicting the individuals low or high ADP platelet response levels. Bolded are p-values, which show the models' significance in predicting either low or high PA/FA platelet response levels.

Moreover, based on the summary in Table 5.19, the SR have overall significant predictions over SJ and SD groups in predicting the high/low ADP platelet response levels. Hence, these results may suggest that the SNPs, which are selected from the RAPIDSNPs pipeline are likely to have higher performance in predicting the individuals with high or low ADP platelet response levels.

5.3.4 The significance of the SNPs genotypes in predicting individuals with high or low responses

In the previous sections, it was observed that in most cases, the key SNPs from RAPIDSNPs appeared to be performing significantly better at predicting the individuals with high or low PA and FA platelet response levels. Thus, the test for significance mainly involved alleles from the key selected SNPs.

5.3.4.1 ***Dataset 1 – SR set involved with high or low PA levels***

In testing the significance of the SNPs' alleles of the individual SR set from the dataset 1 (rs6141803, rs6442896, rs3730051, rs1527480, rs8033381, rs6442895, rs3212391), which significantly predicted individuals' high or low PA levels over the Jones et al. SNPs, Table 5.20 was created. The Table examines the significance of each SNP from the model and its genotypes used in predicting individuals high or low PA levels.

SNP ID	Allele type	Frequency		Overall SNP's significance (p-value)
		High	Low	
rs6141803	Major	55	40	0.02512
	Minor	1	2	
	Hetero	14	28	
rs6442896	Major	55	40	0.0123
	Minor	0	3	
	Hetero	15	27	
rs3730051	Major	30	48	0.009099
	Minor	6	3	
	Hetero	34	19	
rs1527480	Major	59	49	0.1034
	Minor	0	1	
	Hetero	11	20	
rs8033381	Major	43	39	0.05439
	Minor	1	8	
	Hetero	26	23	
rs6442895	Major	55	40	0.0123
	Minor	0	3	
	Hetero	15	27	
rs3212391	Major	30	15	0.0008449
	Minor	9	27	
	Hetero	31	28	

Table 5.20 The relationship between each SNP's genotypes and an individual's high or low PA platelet response levels related to dataset 1.

From Table 5.20, the Chi-square test, shows that only four SNPs' alleles from the model, which significantly predicted individuals with high or low PA levels, have significant relationship with high or low PA levels. These SNPs are rs6442896 in *ITPR1*, rs3730051 in *AKT2*, rs6442895 in *ITPR1*, and rs3212391 in *ITGA2*. The latter SNP rs3212391 appears to be the most significant among the four with the *minor allele* being significantly associated with low PA levels (p-value = 0.0027). Moreover, the closest significant allele, which is likely to be associated with high PA levels is rs3730051 heterozygous (p-value = 0.03936). These results may suggest that for high PA level the most involved SNP is the rs3730051 *heterozygous allele* in *AKT2*, and for low PA level the most involved SNP is the rs3212391 *minor allele*.

5.3.4.2 ***Dataset 1 – SR set involved with high or low FA levels***

The Chi-square test was performed to test the individual SR alleles, which may significantly contribute to the individuals with high or low FA platelet response levels, using dataset 1. The involved key SNPs in the model used in predicting individuals' high or low FA platelet response levels are: - rs11637556, rs1491978, rs1388622, rs1038639, rs2071676, rs10499858, rs10974955. Table 5.21 shows the relationship between each SNPs' genotypes from the model and their frequencies related to the individuals high or low FA platelet response levels.

SNP ID	**Allele type**	**Frequency**		**Overall SNP's significance (p-value)**
		High	**Low**	
rs11637556	Major	34	46	0.09633
	Minor	7	3	
	Hetero	29	21	
rs1491978	Major	19	36	0.01202
	Minor	11	6	
	Hetero	40	28	
rs1388622	Major	20	36	0.01104
	Minor	11	4	
	Hetero	39	40	
rs1038639	Major	16	32	0.01208
	Minor	14	13	
	Hetero	40	25	
rs2071676	Major	37	29	0.009777
	Minor	3	15	
	Hetero	30	26	
rs10499858	Major	67	57	0.01681
	Minor	0	0	
	Hetero	3	13	
rs10974955	Major	37	48	0.1428
	Minor	3	3	
	Hetero	30	19	

Table 5.21 The significance of each SNP with its genotypes' occurrences for the individuals' high or low FA platelet response levels related to dataset 1,

From Table 5.21, a Chi-square test shows that only four SNPs (rs1388622, rs1491978, rs1038639, and rs2071676) have significant relationship with the high or low FA levels. SNPs rs1388622 and rs1491978 are both in *P2YR12*, while SNPs rs1038639 and rs2071676 are in *ITPR1* and *CA9* respectively. The rs2071676 *minor allele* has a significant relationship with low FA levels (p-value = 0.004678). For rs1388622 and rs1491978, no observable significance for both alleles, however, the *major allele* is more likely to be associated with high FA levels, since it is more significant than other alleles (p-value = 0.03251). The rs1038639 *major allele* is more likely to have an effect on the low FA levels (p-value = 0.02092). Taken together, these results suggest that the most significant SNPs associated with high FA level are the rs1388622 and rs1491978 *major alleles* in *P2RY12* and the most significant SNP, which is involved with low FA level, is the rs2071676 *minor allele* in *CA9*.

5.3.4.3 ***Dataset 2 – SR set involved with high or low PA levels***

According to the dataset 2 results, the ANN models based on key SNPs from RAPIDSNPs significantly outperform those based on the Jones et al. SNPs, for predicting individual cases with high or low PA platelet response levels. The key SNPs were individually tested to examine which of the alleles significantly associate with individuals with low or high PA levels. The involved SNPs, which significantly predict the PA levels related to dataset 2 are: - rs17229705, rs246406, rs11631474, rs5277, rs1472122, rs2633717, rs2815805, rs17041401, and rs2825207. Table 5.22 shows the significance of each SNP and its genotypes' frequencies of the individuals' high or low PA platelet response levels related to dataset 2.

SNP ID	Allele type	Frequency		Overall SNP's significance (p-value)
		High	Low	
rs17229705	Major	60	70	0.03064
	Minor	0	0	
	Hetero	15	5	
rs246406	Major	30	24	0.003646
	Minor	2	15	
	Hetero	43	36	
rs11631474	Major	19	28	0.01102
	Minor	32	15	
	Hetero	24	32	
rs5277	Major	46	58	0.04642
	Minor	0	1	
	Hetero	29	16	
rs1472122	Major	23	11	0.003769
	Minor	24	9	
	Hetero	43	40	
rs2633717	Major	47	65	0.003309
	Minor	3	1	
	Hetero	25	9	
rs2815805	Major	66	74	0.02195
	Minor	0	0	
	Hetero	9	1	
rs17041401	Major	18	36	0.008767
	Minor	13	10	
	Hetero	44	29	
rs2825207	Major	19	28	0.05836
	Minor	21	10	
	Hetero	35	37	

Table 5.22 The significance of each SNP with its genotypes' occurrences for the individuals' high or low PA platelet response levels related to dataset 2.

The Chi-square test conducted on each individual SNP revealed that only five SNPs are significantly associated with either high or low PA platelet response levels, Table 5.22. These SNPs are rs246406, rs11631474, rs1472122, rs2633717, and rs17041401, which are associated low, high, low, high, and low PA platelet response respectively with their p-values of 0.003646, 0.01102, 0.003769, 0.003309, and 0.008767 respectively. Moreover, in examining the significance of their individual alleles, it was found that: the rs246406, *minor allele* is

significantly involved with low PA level (p=0.001616), the rs11631474 *minor allele* is significantly involved with high PA levels (p=0.01315), the rs1472122 *minor allele* is significantly involved with low PA levels (p=0.009023), the rs2633717 *heterozygous allele* is significantly involved with high PA levels (p=0.00607), and the rs17041401 *major allele* is significantly involved with low PA levels (p=0.01431) respectively.

5.3.4.4 ***Dataset 2 – SR set involved with high or low FA levels***

Based on the genetic models, the SR group, which are related to dataset 2, are: - rs10429491, rs11264579, rs11637556, rs17786144, rs350916, rs41307142, rs6450105, rs7180408, rs7409876, and rs822442. Table 5.23 shows the significance of each of these SNPs that are associated with high or low FA response levels.

SNP ID	Allele type	Frequency		Overall SNP's significance (p-value)
		High	Low	
rs10429491	Major	42	29	0.005526
	Minor	3	15	
	Hetero	30	31	
rs11264579	Major	54	43	0.01968
	Minor	1	9	
	Hetero	20	23	
rs11637556	Major	38	45	0.122
	Minor	8	2	
	Hetero	29	28	
rs17786144	Major	47	56	0.2744
	Minor	1	1	
	Hetero	27	18	
rs350916	Major	12	25	0.01823
	Minor	20	10	
	Hetero	43	40	
rs41307142	Major	73	64	0.02025
	Minor	0	0	
	Hetero	2	11	
rs6450105	Major	66	74	0.02195
	Minor	0	0	
	Hetero	9	1	
rs7180408	Major	31	46	0.01838
	Minor	5	7	
	Hetero	39	22	
rs7409876	Major	16	31	0.02273
	Minor	16	9	
	Hetero	43	35	
rs822442	Major	54	66	0.03385
	Minor	2	0	
	Hetero	19	9	

Table 5.23 The significance of each SNP with its genotypes' occurrences for the individuals' high or low FA platelet response levels related to dataset 2.

A Chi-square test indicated that only two SNPs, rs10429491 and rs7180408, are significantly associated with the individuals' high and low FA levels respectively, with p-values of 0.005526 and 0.01838 respectively, Table 5.23. Moreover, the rs10429491 *minor allele* and rs7180408

heterozygous allele are significantly associated with low and high FA levels respectively, with p-values of 0.0047, and 0.02951 respectively.

5.3.5 Summary of the most significant SNPs and their genotypes for predicting low or high ADP platelet response levels.

Taken together, from the above results, the following SNPs were found to be significantly involved in predicting individuals with high or low ADP platelet response levels, Table 5.24.

Datasets	SNP	*Gene/loci*	Allele/genotype	p-value	PA/FA	High/low
1	**rs3212391**	*ITGA2*	minor	0.0027	PA	Low
	rs2071676	*CA9*	minor	0. 004678	FA	Low
2	rs246406	*ITGA2*	minor	0.00162	PA	Low
	rs11631474	*MAP2K5*	minor	0.01315	PA	High
	rs1472122	*P2YR12*	minor	0.00902	PA	Low
	rs2633717	*ITPR1*	heterozygous	0.00607	PA	High
	rs17041401	*ITPR1*	major	0.01431	PA	Low
	rs10429491	*JAK2*	minor	0.0047	FA	Low

Table 5.24 The most significant SNPs associated with individuals' high or low ADP levels and their related genotypes/alleles. In bold are previously unidentified SNPs.

5.3.6 Discussion

Based on the prediction results, it appears that the key SNPs from RAPIDSNPs (SR group) confer an improvement in performance on the ANN models than those from the Jones et al. study (SJ group) and the randomly selected SNPs (SD group) used as controls. The scores from the models, which were generated by the SR group show significant performance increase over those from models using the SJ and SD groups. This has been shown to be the case for predicting individuals with both high or low ADP platelet response levels for both FA and PA platelet responses. Nevertheless, in few occasions, it was observed that the SD group outperformed the SR group indicating that there could be other uncaptured SNPs, which could further improve the accuracy in predicting high or low ADP levels. In particular, the SD group

outperforms both SR and SJ groups in predicting individuals with 'normal' FA levels related to dataset 1 and that might be interesting from the methodological and biological aspects. Moreover, in predicting the individuals with normal PA response level, the performances of the SR & SJ are significantly close particularly when using dataset 2. The exception is during applying the models to the dataset 1 in which the SR slightly outperforms the SJ set.

It might be noted that the dataset 3 was not entirely included in the modelling and analyses. The major reason for excluding this subset from the analyses is that it has relative very small number of cases, which would be relatively insufficient to train NN (Lawrence et al., 1998).

Therefore, based on the results, the ANN models can be used to predict human phenotype, and that this requires good SNP detection for which RAPIDSNPs is better.

5.3.6.1 ***The most significant SNPs allele(s) contributed to high/low ADP levels and their associated CVDs***

We have identified the most significant SNPs that are likely to be involved with an individual's high or low ADP levels. And therefore, we have answered one of the major underlying questions of the study: determining the genetic variants that are more likely to underpin individuals with extreme high or low ADP platelet response levels (Figures 1.8, 1.9, and 1.10 in Chapter 1). The rs246406 minor allele in *ITGA2* appears to be the most significant SNP for low PA platelet response levels. Moreover, the rs2071676 SNP in *CA9* and rs10429491 in *JAK2*, are the most significant minor alleles associated with low FA platelet response levels. Additionally, for high FA level, both SNPs rs1388622 and rs1491978 in *P2YR12* appear to be the most significant major alleles. For other alleles, the rs2633717 in *ITPR1*, which is a newly identified, is the most significant heterozygous allele for high PA platelet response levels.

Thus, these results further support our initial and underlying argument that bioinformatics based approaches may complement and add value to the mainstream standard biostatistical approaches, such as stepwise forward, which are used in genetic association studies. It is clear that some of these SNPs, such as the common SNP rs246406, and rs10429491 were indeed previously identified to be associated with PA (Jones et al., 2009). However, our approaches have also identified which of these SNPs alleles are significantly associated with extreme responders i.e. individuals with high or low ADP platelet responses levels.

Moreover, our methods have predicted additional SNPs, which significantly associate with extreme low or high ADP platelet response levels that were *previously unidentified* using the purely biostatical approach (forward stepwise).

Additionally, the SNPs alleles related to the individuals with normal ADP response levels were not further investigated as the focus of the study is to find the alleles that are likely to contribute to the individuals' extreme (high/low) ADP levels. Nevertheless, the individuals with normal ADP levels were accounted in the genetic models as they could be of the biomedical interest to predict whether an individual may respond 'normally' to ADP if is not extreme.

Finally, the SNPs identified in this chapter that significantly associate with extreme levels of platelet response activation were also predicted to be involved with molecular mechanisms such as structural or functional changes in the related proteins, and regulatory roles. The involved regulatory roles include transcription binding sites (TFBS), expression quantitative loci (eQTL), and others (see Chapter 3 and 4). Thus, these newly identified SNPs such as rs2071676 in *CA9* and rs3212391 in *ITGA2* warrant further experimental investigation, for a deeper understanding of the associated molecular mechanisms that may be underpinning ADP platelet responses. In addition, these SNPs might be used as markers to guide future CVD personalised medicine decisions.

5.4 Conclusion

Taken together, the above results overwhelmingly suggest that the accuracy of the predictive genetic models, which are based on supervised learning approaches (such as ANNs) are more likely to be enhanced when the key SNPs obtained from the RAPIDSNPs are used as input parameters. The majority of the generated scores from models based on the key SNPs from RAPIDSNPs were statistically significant, and predicted high or low ADP platelet response levels within individuals with the highest accuracy. In addition, a handful of these SNPs' alleles were found to be significantly associated with individuals extreme high or low ADP platelet response levels, which could have clinical or therapeutic implications in future CVD personalised medicine applications (see Chapter 6).

The results in this Chapter further validate the RAPIDSNPs approach, which appears to be robust in selecting the SNPs that are good predictors of the disease/trait states. The RAPIDSNPs approach might be generically utilised for optimising the selection of SNPs to improve predictive genetic models involving binary phenotypic traits, in data where non-genetic factors are controlled.

Chapter 6 - Synthesis and next direction

6.0 Aims of the Project

The initial aim of this project was to develop computational predictive methods for rapidly identifying and analysing key SNPs, which are associated with ADP platelet responses variability and may likely to be missed when using standard biostatistical methods. The ADP platelet response was chosen as a suitable case study, due to its wide used as one of the major treatment targets for common CVD problems. CVD continues to be one of the leading causes of death in both developed and developing countries (Stanner, 2008; "WHO | Cardiovascular diseases (CVDs)," 2016). Therefore, it is envisaged that the newly identified SNPs will have potential uses in the future implementation of CVD personalised medicine applications or for informing genomic based medicine / health decisions. In Chapter 2, the main hypothesis was that the designed computational method (RAPIDSNPs) was more likely to rapidly identify the most significant SNPs, which are associated with ADP platelet responses, than standard biostatistical methods.

Furthermore, the newly identified SNPs, using RAPIDSNPs, were hypothesised to be involved with the various molecular mechanisms that underpin the ADP platelet responses variability, and hence, we developed molecular bioinformatics predictive analytics protocols to explore these hypotheses further in Chapters 3 & 4. In addition, the identified key SNPs from the RAPIDSNPs method were investigated to examine their usefulness in predictive genetic models for identifying individuals with high/low ADP platelet responses. Thus, in Chapter 5, the underlying hypothesis was that these SNPs could be used to significantly improve predictions of individuals with high or low ADP platelet response levels. And hence, they are more likely to be potential candidates and have implications for future CVD genomic-based or personalised medicine applications.

This chapter provides a synthesis of the key findings, which focuses on a few of the key SNPs that were identified, predictively analysed and are thought to be more likely to have potential implications for future CVD research. The chapter further discusses how the results of the study may potentially impact on personalised CVD clinical decisions and therapeutics.

6.1 An integrated pipeline for omic data analyses associated with ADP platelet responses for CVD research

To investigate the variability of individuals ADP platelet responses, an integrated pipeline was designed to provide an analytical framework for investigating different omics data (section 1.10). The fundamental objective was to design a computational method for rapidly identifying novel (previously unidentified) key genetic variants. Also by using this framework, a further objective was to develop predictive approaches for examining the molecular effects, which are likely to contribute to ADP variability. In addition, for potential application of the identified key SNPs to the CVD personalised clinical/therapeutic settings, further predictive approaches were developed to determine whether the identified key genetic variants could be useful in predicting individuals with high or low ADP response levels.

6.2 Implementation of the integrated pipeline

The implementation of the integrated pipeline was performed in logical sequential manner and Chapters are presented in chronological order. The key components of the framework are summarised below.

6.2.1 RAPIDSNPS

In Chapter 2, the RAPIDSNPs was developed for improving identification of the key SNPs, which are associated with the ADP platelet responses. It has been reported elsewhere that the identification of the disease associated genetic variants could be improved if biostatistical

methods are supplemented/complemented with the computational/bioinformatics based approaches (Eichler et al., 2010; Moore et al., 2010). Hence, the RAPIDSNPs method was designed to help further bridge the gap due to the data generated from the projects such as HapMap, 1000 Genomes and GWAS, and to provide an increased ability for identifying true disease/trait associated SNPs, which potentially could be missed using standard methods. The performance of this method in identifying trait associated SNPs that are unlikely to be identified by the standard biostatistical methods was benchmarked against the stepwise forward approach, which was previously used to find the SNPs associated with ADP platelet responses.

The underlying functioning of the method is grounded on the random forests (RF) algorithm, which is a tree based methodology used in classification and regression for high dimensional and non-linear data. The RF is used as a base filtering for finding initial useful SNPs, which are further refined using an ensemble of other feature selection methods for eventual identification of the most significant SNPs associated with ADP platelet responses. The RAPIDSNPs method could also be generically applied to other continuous trait/disease phenotypes. Several new SNPs were identified by the RAPIDSNPs method, which are significantly associated with ADP platelet responses and which were previously unidentified. Furthermore, the majority of the genes were also found to have general associations with CVDs in other independent studies, indicating the robustness of the method in identifying crucial or potential disease associated variants.

6.2.2 A structural bioinformatics approach for investigating missense SNPs associated with ADP platelet response

The first key molecular investigation was into the effects of the identified missense SNP mutations on the related protein structures and functions. The aim was to determine how likely the missense SNPs were to damage or affect the structure and/or function of the proteins, which

may help to further explain the variability of ADP platelet responses. The missense SNPs lead to an amino acid substitution in the related proteins, which may cause changes to the associated proteins' folding and/or binding.

The structural bioinformatics approach was designed and applied to investigate the role of the missense SNPs that underpin extreme ADP platelet responses. The damaging level of the missense SNPs were initially investigated and subsequently several 3D models were then generated for the proteins associated with the identified harmful missense SNPs. The likely structural and/or functional effects on the proteins due to the identified damaging missense SNPs were then examined.

6.2.3 A bioinformatics approach for investigating the regulatory SNPs associated with ADP platelet responses

Another key molecular investigation involved the regulatory functional SNPs that are likely to be associated with the ADP platelet responses. For instance, the SNPs regulatory roles may lead to changes in the expression levels of the related gene transcripts, i.e. transcriptomic levels variation. This variation in the expressed genes may be caused by the presence of the SNPs in allelic specific manner, which may lead to the variation of the produced transcript levels among individuals. Thus, the variability of the ADP platelet responses was further hypothesised to be due these regulatory key SNPs, which were identified by the RAPIDSNPs.

Therefore, an analytical pipeline, which incorporated state-of-art bioinformatics methods, was designed to investigate the likelihood of the involvement of the key SNPs in regulatory roles. Several alternative regulatory roles were investigated.

6.2.4 A genetic predictive approach for ADP platelet response levels

The gene-based predictive approach was designed, using the SNPs from the RAPIDSNPs method and the SNPs identified from Jones et al study (Jones et al., 2009), in order to predict high or low ADP platelet response levels. The underlying hypothesis was that the SNPs from RAPIDSNPs key SNPs can be used to make significantly better predictions than those from the Jones et al study. This kind of prediction is potentially very useful for CVD personalised medicine. For instance, it would be of interest to identify the SNP's allele, which could correctly guide the individualised dosage/treatment strategies. In addition, based on the prediction outcomes, an individual's future CVD prognosis could be determined.

The developed predictive approach employed supervised learning method using an artificial neural network (ANN). Several models were generated using the SNPs from RAPIDSNPs and Jones et al., which were then evaluated to compare their predictive performance. The individual SNPs alleles (genotypes), which were significantly linked with high or low ADP platelet response levels, were then examined.

6.3 Summary of the key findings and conclusions

6.3.1 RAPIDSNPs

6.3.1.1 ***The key SNPs identified by RAPIDSNPs***

Using the RAPIDSNPs method, numerous (key) SNPs were identified, which are significantly associated with the ADP platelet responses. Among the key SNPs identified by RAPIDSNPs, many were previously unidentified, when the same data was analysed using the stepwise forward method. Furthermore, several of these were found in other studies to be generally associated with CVDs, suggesting the robustness of the approach in identifying crucial SNPs.

Moreover, the ability to identify key SNPs was further validated using simulated data with artificial SNPs, which also showed the same performance pattern as the real SNPs.

The notable key SNPs, which were previously unidentified and found to be associated with the ADP platelet responses, as well as having independently identified associations with CVDs, are outlined in the Table 6.1 below.

S/N	Identified SNPs	Gene/Locus	Associated ADP platelet responses	High response/ Increase Low response /Decrease	Likely associated CVD type based on the discussion
1	**rs6141803**	*COMMD7*	PA	low	Myocardial Infarction (MI) (Goodall et al.,2010)
2	**rs1491978**	*P2Y12*	FA	high	Has a potential for atherothrombosis, stroke, or MI (Zee et al., 2008)
3	**rs1388622**	*P2Y12*	FA	high	Has a potential for atherothrombosis, stroke, or MI(Zee et al., 2008; Ziegler et al., 2005)
4	**rs2071676**	*CA9*	FA	low	Has a potential for hypertension (Reardon et al., 2009; Woodman et al., 2010)
5	**rs12953**	*PECAM1*	PA	low	Likely MI (Listì et al., 2004; Sahebkar et al., 2013)
8	**rs12485738**	*ARHGEF3*	FA	high	MI & ischemic stroke (Meisinger et al., 2009)

Table 6.1 The key SNPs identified by the RAPIDSNPs pipeline that are associated with different ADP platelet response and were previously unidentified. Each of the SNPs corresponds to a particular ADP platelet response and whether is likely to increase or decrease to the platelet response and associate with a particular CVD. In bold are the newly discovered significant SNPs, which were not previously identified in the study by Jones et al. (2009).

The identified SNP rs12485738 in the Table 6.1, has been associated with CVD in different independent unrelated studies. Other SNPs such as rs12953 and rs1491978 have potential

association with CVD, as their related genes or proximal SNPs have been reported to be associated with CVD and thus might be worth for further investigation. The most interesting among these SNPs is rs2071676 in the *CA9* gene which was found to be significantly associated with FA platelet responses. To the author's knowledge this SNP is not mentioned in the platelet literature, however, the related gene isoform *CA2* was elsewhere reported to be associated with a decrease in the platelet aggregation and hypertension (Woodman et al., 2010). Thus, it might be of interest to investigate *CA9*, as the SNP's minor allele is also found to significantly associated with the decrease in the ADP platelet response (FA).

Furthermore, in our analyses using RAPIDSNPs we have found that two *P2YR12* SNPs (rs1388622 and rs1491978), which were previously found to have an insignificant association with ADP platelet responses, are actually significantly associated with FA platelet responses. It is understood that the P2YR12 plays pivotal role in the ADP mediated platelet responses and aggregation, and is a subject of several anti-platelet drugs for CVD treatment, thus, these SNPs are also worthy of further investigation.

6.3.2 Predicted structural/functional effects on the proteins related to ADP platelet response due to damage missense SNPs

Two missense SNPs, rs2071676 and rs12953 (which correspond to the Val33Met and Ser563Asn mutations respectively) were likely to be deleterious to their related proteins (CA IX, and PECAM1 respectively). These mutations are worthy of future experimental work to further investigate their impact on the molecular mechanisms of ADP platelet responses.

The Val33Met mutation was predicted to be in the HLH or EF-hand structural motifs within N-terminus region in the characterised PG domain of CA IX. To the best of our knowledge, to date, the characterisation of this mutation has not been mentioned in the literature. Further

examination of the potential functional impact revealed that the mutation may affect the binding affinity of the respective metal ion ligand (Mg2+ or Ca2+), which has a potential to affect the overall signal transduction of the protein. Future experiments could further focus on characterising and understanding the individuals' differences in the structural and/or function of this putative HLH/EF hand motif and the effect on the FA platelet response levels, in an allele specific manner. The focus might be on individuals with minor allele, which appears to significantly associate with decrease in FA platelet response or low FA level (Chapter 5). This might be of interest in a clinical setting, for instance in predicting the individuals' responses to treatment and prognosis, which may also influence dosage and monitoring.

Furthermore, for the Ser563Asn mutation in PECAM1, it was found that the mutation is likely to cause a crucial change in the structure, where the Ig domain 6 of the protein appears to be changing from the C-type to V-type. This structural change is likely to have an effect on the binding activities of PECAM1 partners, leading to a decrease in PA platelet responses and aggregation. Thus, this mutation also has potential for future investigations in understanding its therapeutic and clinical implications in the context of personalised medicine.

6.3.3 Predicted ADP platelet response associated regulatory SNPs

Furthermore, several key SNPs were identified to be significantly associated with regulatory roles, i.e. they are more likely to be rSNPs. These regulatory roles appear to be affecting the individuals' PA or FA platelet responses. Further investigation of their related genes or transcription factors suggested the possibility of the overall rSNPs involvement in CVD. The most significant rSNPs are shown in the Table 6.2, each of which are associated with high or low ADP levels and have potential association with CVD, thus worthy of further investigation.

Datasets	SNP	*Gene/loci*	Predicted molecular regulatory roles	PA/ FA	High /low	Likely CVD type associated
1	**rs3212391**	*ITGA2*	TFB, distal l. interactions	PA	Low	Atherosclerosis (Stroke), Hypertension(Deng and Shen, 2007)
	rs12485738	*ARHGEF3*	eQTL, TFB, regulation of RNA binding proteins	FA	High	Blood pressure, Potential for MI & ischemic stroke (Meisinger et al., 2009)
	rs1388622/ rs1491978	*P2RY12*	eQTL, TFB	FA	High	Potential for MI & ischemic stroke (Zee et al., 2008; Ziegler et al., 2005)
2	rs246406	*ITGA2*	eQTL, TFB	PA	Low	Potential for Ischemic stroke (Wu et al., 2014)
	rs1472122	*P2YR12*	eQTL, TFB	PA	Low	Potential for Ischemic Stroke (Zee et al., 2008; Ziegler et al., 2005)

Table 6.2 The identified rSNPs found to be more likely to be related with regulatory roles. These rSNPs were also found in Chapter 5 (excluding rs12485738) to be significantly associated with high or low ADP levels and are likely to be associated with CVD. In bold are the newly discovered significant SNPs, which were not previously identified in the study by Jones et al. (2009).

6.3.4 Genetic prediction of individuals' high or low ADP platelet response levels

Based on the designed predictive approach, it is clear that key SNPs from the RAPIDSNPs approach can be used to predict individuals with high or low ADP levels with significantly higher accuracy than those from the Jones et al. study. Moreover, we identified the most significant SNPs that are more likely to be involved with the high or low extreme ADP levels. Examples of these SNPs are rs1388622 and rs1491978 in *P2RY12*, which were identified to be associate with high FA platelet response levels. It was also found that the SNP rs2071676 in CA9 is most significantly associated with low FA levels. This reflects the previous findings, which reported that the related cytosolic CA II isoform is associated with ADP platelet responses and decreased platelet aggregation (Woodman et al., 2010). Moreover, it appears that

individuals with this mutation (minor alleles) are likely to have low platelet aggregation and are less-prone to hypertension, while those with other alleles are likely to have high platelet aggregation, and are prone to thrombosis and hypertension (Reardon et al., 2009; Woodman et al., 2010). The literature is silent on which individuals' alleles are involved with the related structural changes of the SNP that was mentioned above. Hence, our findings may guide future investigations aimed at this mutation and protein for platelet aggregation and hypertension.

Therefore, the identified SNPs involved with high or low extreme ADP platelet response levels are more likely to be of the higher interest for clinical and targeted therapeutic CVD applications for personalised medicine (PM). The next section is a further discussion of how these SNPs, and in general our computational predictive approaches, may potentially impact on future PM.

6.4 Implications of the approaches and findings to personalised medicine (PM) for CVD and other diseases

Based on the discussion of PM in Chapter 1, our computational predictive approaches have potential to be exploited in PM settings. They could be used to identify an individuals' likely CVD risk and predict their specific ADP platelet responses, while pinpointing the key associated SNPs.

The RAPIDSNPs method could be used to create *the overall or initial genetic profile* of individuals, based on the threshold SNPs' confidence level, which was used in RAPIDSNPs to identify the key SNPs (section 2.2.2.1). The specific molecular effects of the SNPs such as whether a SNP is damaging to structure/function of the related protein, transcription regulation could then be predicted and used to extend upon the initial genetic profile accordingly. Such predictive aspects could also include the SNPs alleles, which are significantly associated with the individuals' high or low ADP platelet response levels. Moreover, the extended profile may

be useful for molecular diagnostics and providing an informed and targeted treatment strategy for specific individuals.

For instance, for the individuals' related to the FA platelet response, the overall view of genetic and molecular results for potential CVD application could be as follows, Table 6.3.

RAPIDSNPS							
SNP ID	***Loci (gene)***	**Confidence level**	**Phenotypic state (+ve/-ve)**	**Molecular state**		**The significance in the phenotypic levels (Low/High FA levels): p-value**	**Specific allele involved & significance (p-value), Low/High FA levels)**
				Structural / functional	**Regulatory role**		
rs1388622	*P2YR12*	0.563	+ve			0.011 High	Major, 0.032, high
rs1491978	*P2YR12*	0.438	+ve		eQTL, TFBS	0.012 High	Major, 0.022, low
rs11637556	*MAP2K1*	1.000	+ve		TFBS, RBPS	0.09633 High	
rs1038639	*ITPR1*	0.375	-ve			0.012 Low	Major, 0.021, low
rs2071676	*CA9*	0.375	-ve	Damaging		0.01 Low	minor, 0.0047, low
rs10499858	*CD36*	0.688	-ve		RBPS	0.017 Low	Minor, 0.01242, low
rs10974955	*JAK2*	0.563	+ve			0.1428 High	

Table 6.3 The genetic and molecular overview of individuals associated with FA platelet response extreme level based on the predictive information. The blue, green, yellow, and red indicate the SNPs alleles with higher, standard, medium, and poor significance respectively that are associated with extreme levels.

Hence, from Table 6.3, for potential CVD PM applications, the generated genetic profile could be used to predict the disease state, or it could be used for targeted clinical investigation or treatment depending on the specific SNPs alleles and their molecular status. For example, it has been shown that the SNP rs1491978 is likely to be associated with myocardial infarction (MI). MI is triggered mainly due to the platelet hyperactivity, which is linked with increase in fibrinogen binding under ADP activation. Thus, under the PM settings, the minor/heterozygous alleles individuals would be the main focus of clinical investigation in determining the disease state, or response to drugs or dose monitoring.

Using a hypothetical example, suppose an individual whose two SNPs (rs1388622 and rs1491978 in *P2YR12*) with homozygous major genotypes, has been diagnosed with FA platelet hyperactivity. And we want to know CVD risk level for this particular individual. Based on the Table 6.3, the homozygous major alleles for the SNPs rs1388622 and rs1491978 in *P2YR12* are related to FA platelet hyperactivity (High FA response level). Based on Table 6.1, these SNPs are more likely to be associated with increased risk of myocardial infarction (MI) and/or ischemic stroke. Thus, further individual diagnosis for MI or/and ischemic stroke which would integrate omic (molecular) information of these SNPs (eQTL and TFBS) would be performed for better guiding clinical or dosage treatment focusing on individuals with major allele.

Moreover, such genetic profiles are said to be promising in facilitating the enhancement of new diagnostic tests based on genes or proteins, and therapies, which target the consequences of specific genetic alterations or aberrations (President's Council of Advisors on Science Technology, 2008).

6.4.1 Could the methods be applicable to the developing countries?

The main motive of the question is driven by the fact that the author is from Tanzania, which is one of the developing countries in the Sub-Saharan region. Based on the discussion in the Chapter 1, CVD and other non-communicable diseases are emerging as silent killers in the developing countries particularly in the Sub-Sahara. The stats showing the increased risks for the CVD are sharply rising (see Chapter 1, section 1.2). However, the omic data generation and analytical tools are not as advanced as in the developed countries. Based on the survey that the researcher conducted, most of the generated data from the omic based researches, including genetic association studies are partly analysed in-house but, largely are analysed either in Europe or South Africa where the tools are more advanced. Thus, the designed approaches have

potential of being implemented and applied for investigating CVD risks and for potential future personalised medicine applications even in the developing countries.

6.5 Major contributions and new insights gained

There are several major contributions from this study. The contributions might be divided into two major categories, which are computational and biological aspects.

6.5.1 Contributions to the computational aspects of genetic association analyses

The first major contribution is the RAPIDSNPs pipeline, which is a novel computational tool for rapidly analysing individuals' genotyped SNPs, which are associated with a continuous phenotype. The tool is able to identify key and significant SNPs, which were previously unidentified or found to be insignificant associated with ADP platelet responses signifying the robustness of the tool in identifying the most significant genetic variants. Moreover, the pipeline was tested for simultaneously analysing covariates or non-genetic factors. Based on the tests, it was shown to be effective at identifying key SNPs in addition to the significance of covariates in further explaining the continuous phenotypic variation. Thus, the RAPIDSNPs has further bridged the gap of identifying other unknown genetic variants that may account for the remaining unexplained variation ('missing heritability' problem) The paper describing the pipeline has been published in the journal PLOS ONE (Salehe et al., 2017).

The further computational contribution is the development of the computational predictive protocols for analysing the molecular aspects of the genetic variants associated with ADP platelet responses and CVD. In most cases, genetic association studies do not provide any details concerning the molecular effects of the causal SNPs associated with the trait/disease under investigation.

Finally, a supervised learning method based on ANN was developed for predicting the individual's high or low ADP levels for potential prediction of CVD risk levels and PM applications. Based on the literature, ANNs are rarely used for this type of genetic prediction for human disease/traits. Each of these developed protocols and predictive supervised learning methods also have potential for being generically applied in other disease/trait cases.

6.5.2 Contributions to the biological knowledge of ADP platelet responses and CVD.

The major contribution in this area is on the molecular genetics of the ADP platelet responses and the link with CVD. Several SNPs, which were newly identified in this study and those previously identified, were studied to elucidate their possible molecular effects through computational predictive methods. Each of the SNPs have been identified to be significantly associated with ADP platelet responses (FA/PA). Therefore, further experimental investigation to verify their underlying molecular mechanisms would be justified. In particular, to highlight a few, SNPs such as **rs12485738** in *ARHGEF3* and **rs2071676** in *CA9*, rs12953 in *PECAM1*, which are significantly associated with FA, rs1472122 in *P2YR12*, **rs3212391** and rs246406 in *ITGA2*, which are significantly associated with PA. The bolded SNPs were previously unidentified and are likely to be associated with several molecular mechanisms. For instance, the SNP rs12485738 has been predicted to be associated with different regulatory functions such as eQTL where it has been associated with several ADP platelet responses related eQTL genes. Moreover, this SNP is likely to be involved with the regulation of TFBS in which the rs3212391 SNP has been also associated with. With regard to the structural and/or functional aspect, both of the rs2071676 and rs12953 have been predicted to be affecting the structure or binding activity of their related proteins.

Each of these bolded SNPs are worthy of future investigation as they have been independently associated with CVD, such as stroke and MI (Tables 6.1 and 6.2) and in this study, for the first

time, we have identified them to be significantly involved with specific individuals' ADP platelet responses variation. In addition, the predicted molecular aspects of these SNPs could further be utilised for targeted pharmacogenomics or therapeutics purposes that might potentially benefit future CVD PM applications.

6.6 Next directions

6.6.1 ADP platelet responses (FA/PA) and CVD

There are notable SNPs, which were found to be significantly associated with FA/PA platelet responses, and their underlying molecular mechanisms are worthy of further experimental investigation. For instance, rs2071676 in CA9, which lead to V33M substitution in the related CA IX protein would be interesting for experimental studies aimed to validate the link between this mutation, low FA responses and related CVD state (likely hypertension). The experimental work could focus on minor allele individuals, as they have been found to be significantly associated with low FA responses (Chapter 5). In this regard, a guided molecular dynamics simulation, for example, followed by other experimental work could be vital in understanding and characterising the effects of mutation on the protein's binding and folding in the predicted putative HLH/EF hand motif, which is underpinning the low FA platelet response.

In addition, the missense SNP rs12953, which causes Ser563Asn mutation in the extracellular domain 6 of PECAM1 might be also worthy of further investigation. Based on the analyses of the predicted structural 3D models and existing experimental data, this mutation is likely to reduce the binding affinity and ligand specificity of PECAM1 (Baldwin et al., 1994; Wollscheid et al., 2009). Thus, investigating this mutation to determine its potential for targeted CVD therapeutics for future CVD personalised treatment would be vital.

6.6.2 RAPIDSNPs improvement

Furthermore, other future work could focus on improving the efficiency of the RAPIDSNPs method. Several highlighted areas have potential for improvement, many of which were also suggested by the reviewers of the paper.

6.6.2.1 ***Short-term objectives***

6.6.2.1.1 Handling genome-wide genetic variants (SNPs)

The most appealing area for RAPIDSNPs development is in the ability to handle genome-wide SNPs. This would address the exponential increases in the number of both curated (validated/RefSNP) and non-curated SNPs in the dbSNP databases (Sherry et al., 2001). For instance, the number of validated (rs#) SNP in the current dbSNP build 149 is estimated to be 89,404,961. In addition, millions of SNPs are normally being genotyped in the genome-wide association studies. It is clear that further optimised computational methods for analysing this voluminous data would be indispensable. In its current version the RAPIDSNPs is able to handle small numbers of genotyped SNPs (~10,000), but in relatively high dimensional spaces (~500 cases and large number of SNPs). Since, the RAPIDSNPs method is an optimised tool for finding the most significant SNPs, then its potential improvement for handling the whole genome-wide data for association studies would likely benefit the scientific community. Thus, in the near future, the focus would be to improve the RAPIDSNPs for efficient handling of whole genome SNPs.

6.6.2.1.2 Implementation of the complete pipeline as a web server/an R-package

For wider adoption by the scientific community, a potential next endeavour would be to implement the RAPIDSNPs as a web server or an official R-package. The web-server approach may be the preferred implementation with potential to reach a wider audience, as it would offer

a graphical user interface (GUI) accessible to anyone with a web browser. Nevertheless, server maintenance would be more challenging comparing to an R-package, which would be deposited on the CRAN (Comprehensive R Archive Network) sites. As the R language has gained much more attention particularly amongst bioinformatics and biomedical science community, then an R-package implementation would also be a priority.

6.6.2.2 ***Long-term objectives***

6.6.2.2.1 Inclusion of epigenetic and other variants

The epigenetic or methylomic data have gained much more attention, as these type of variants occur frequently in the genome and are considered to be associated with complex diseases including CVD (Feinberg, 2010; Gerasimova et al., 2013; Keating et al., 2016). Therefore, in terms of the RAPIDSNPs, it would be interesting to consider how it could handle the epigenetic variants. Several new aspects would need to be considered, such as data structure (representation) issues, as epigenetic data might involve histone modification or DNA methylation, which may add on the inter-individuals' phenotypic variability.

Furthermore, a related consideration would be the extension of the ADP platelet responses and CVD research by incorporating these type of variations. For instance, it would be interesting to know the effects of any identified epigenetic variations and how they might explain the inter-individual variability in the ADP platelet responses and CVD risks. This type of information could also be useful for potential future targeted clinical and therapeutic CVD applications.

References

Abraham, G., Kowalczyk, A., Zobel, J., & Inouye, M. (2013). Performance and robustness of penalized and unpenalized methods for genetic prediction of complex human disease. *Genetic Epidemiology*, *37*(2), 184–195. https://doi.org/10.1002/gepi.21698

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., … Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, *7*(4), 248–249. https://doi.org/10.1038/nmeth0410-248

Agüero, F., Correa-Oliveira, G., Roos, D. S., & Kissinger, J. C. (2007). Bioinformatics in tropical disease research: A practical approach. del Portillo H, Gruber A, Durham A, Hyung C, editors. *NCBI Electronic Book*.

Akerman, M., David-Eden, H., Pinter, R. Y., & Mandel-Gutfreund, Y. (2009). A computational approach for genome-wide mapping of splicing factor binding sites. *Genome Biology*, *10*, R30. https://doi.org/10.1186/gb-2009-10-3-r30

Albers, G. W., Amarenco, P., Easton, J. D., Sacco, R. L., & Teal, P. (2001). ANtithrombotic and thrombolytic therapy for ischemic stroke. *Chest*, *119*(1_suppl), 300S–320S. https://doi.org/10.1378/chest.119.1_suppl.300S

Albert, F. W., & Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics*, *16*(4), 197–212. https://doi.org/10.1038/nrg3891

Alevriadou, B. R., Moake, J. L., Turner, N. A., Ruggeri, Z. M., Folie, B. J., Phillips, M. D., … McIntire, L. V. (1993). Real-time analysis of shear-dependent thrombus formation and its blockade by inhibitors of von Willebrand factor binding to platelets. *Blood*, *81*(5), 1263–1276.

Alexy, T., Ambrus, T., Zsolt, M., Beata, H., Katalin, K., Gergely, F., … Kalman, T. (2004). Inhibition of ADP-evoked platelet aggregation by selected poly(ADP-ribose) polymerase inhibitors. *Journal of Cardiovascular Pharmacology*, *43*(3), 423–431.

Alj, Y., Georgiakaki, M., Savouret, J.-F., Mal, F., Attali, P., Pelletier, G., … Perlemuter, G. (2004). Hereditary persistence of α-fetoprotein is due to both proximal and distal hepatocyte nuclear factor-1 site mutations1. *Gastroenterology*, *126*(1), 308–317. https://doi.org/10.1053/j.gastro.2003.10.073

Alterio, V., Di Fiore, A., D'Ambrosio, K., Supuran, C. T., & De Simone, G. (2012). Multiple Binding Modes of Inhibitors to Carbonic Anhydrases: How to Design Specific Drugs Targeting 15 Different Isoforms? *Chemical Reviews*, *112*(8), 4421–4468. https://doi.org/10.1021/cr200176r

Alterio, V., Hilvo, M., Di Fiore, A., Supuran, C. T., Pan, P., Parkkila, S., … De Simone, G. (2009). Crystal structure of the catalytic domain of the tumor-associated human carbonic anhydrase IX. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(38), 16233–16238. https://doi.org/10.1073/pnas.0908301106

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402.

Andersen, M. C., Engström, P. G., Lithwick, S., Arenillas, D., Eriksson, P., Lenhard, B., … Odeberg, J. (2008). In Silico Detection of Sequence Variations Modifying Transcriptional Regulation. *PLOS Comput Biol*, *4*(1), e5. https://doi.org/10.1371/journal.pcbi.0040005

Anfinsen, C. B., Haber, E., Sela, M., & White, F. H. (1961). THE KINETICS OF FORMATION OF NATIVE RIBONUCLEASE DURING OXIDATION OF THE REDUCED POLYPEPTIDE CHAIN. *Proceedings of the National Academy of Sciences of the United States of America*, *47*(9), 1309–1314.

Ao, S.-I. (2008). *Data mining and applications in genomics* (Vol. 25). Springer. Retrieved from http://books.google.co.uk/books?hl=en&lr=&id=pZWAaN_y5qYC&oi=fnd&pg=PR2&dq=Data+Mining+and+Applications+in+Genomics&ots=Xe6Z9HsZck&sig=NXtMjf_x5pinJWsOddq_clibr2g

Ashley, E. A., Butte, A. J., Wheeler, M. T., Chen, R., Klein, T. E., Dewey, F. E., … Altman, R. B. (2010). Clinical assessment incorporating a personal genome. *Lancet (London, England)*, *375*(9725), 1525–1535. https://doi.org/10.1016/S0140-6736(10)60452-7

Atchley, W. R., & Fitch, W. M. (1997). A natural classification of the basic helix–loop–helix class of transcription factors. *Proceedings of the National Academy of Sciences*, *94*(10), 5172–5176.

Ayers, K. L., & Cordell, H. J. (2010). SNP selection in genome-wide and candidate gene studies via penalized logistic regression. *Genetic Epidemiology*, *34*(8), 879–891. https://doi.org/10.1002/gepi.20543

Balaur, I., Saqi, M., Barat, A., Lysenko, A., Mazein, A., Rawlings, C. J., … Auffray, C. (2016). EpiGeNet: A Graph Database of Interdependencies Between Genetic and Epigenetic Events in Colorectal Cancer. *Journal of Computational Biology*. https://doi.org/10.1089/cmb.2016.0095

Balding, D. J. (2006). A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, *7*(10), 781–791. https://doi.org/10.1038/nrg1916

Baldwin, H. S., Shen, H. M., Yan, H. C., DeLisser, H. M., Chung, A., Mickanin, C., … Albelda, S. M. (1994). Platelet endothelial cell adhesion molecule-1 (PECAM-1/CD31): alternatively spliced, functionally distinct isoforms expressed during mammalian cardiovascular development. *Development (Cambridge, England)*, *120*(9), 2539–2553.

Banerjee, A. (2012). A review of family history of cardiovascular disease: risk factor and research tool. *International Journal of Clinical Practice*, *66*(6), 536–543. https://doi.org/10.1111/j.1742-1241.2012.02908.x

Barclay, A. N. (1999). Ig-like domains: Evolution from simple interaction molecules to sophisticated antigen recognition. *Proceedings of the National Academy of Sciences*, *96*(26), 14672–14674. https://doi.org/10.1073/pnas.96.26.14672

Bates, S. (2010). Progress towards personalized medicine. *Drug Discovery Today*, *15*(3–4), 115–120. https://doi.org/10.1016/j.drudis.2009.11.001

Bayat, B., Werth, S., Sachs, U. J. H., Newman, D. K., Newman, P. J., & Santoso, S. (2010). Neutrophil transmigration mediated by the neutrophil-specific antigen CD177 is influenced by the endothelial S536N dimorphism of platelet endothelial cell adhesion molecule-1. *Journal of Immunology (Baltimore, Md.: 1950)*, *184*(7), 3889–3896. https://doi.org/10.4049/jimmunol.0903136

Belton, O., Byrne, D., Kearney, D., Leahy, A., & Fitzgerald, D. J. (2000). Cyclooxygenase-1 and -2-dependent prostacyclin formation in patients with atherosclerosis. *Circulation*, *102*(8), 840–845.

Bendtsen, J. D., Nielsen, H., von Heijne, G., & Brunak, S. (2004). Improved prediction of signal peptides: SignalP 3.0. *Journal of Molecular Biology*, *340*(4), 783–795. https://doi.org/10.1016/j.jmb.2004.05.028

Bennett, J. S. (2001). Platelet-Fibrinogen Interactions. *Annals of the New York Academy of Sciences*, *936*(1), 340–354. https://doi.org/10.1111/j.1749-6632.2001.tb03521.x

Berchner-Pfannschmidt, U., Petrat, F., Doege, K., Trinidad, B., Freitag, P., Metzen, E., … Fandrey, J. (2004). Chelation of cellular calcium modulates hypoxia-inducible gene expression through activation of hypoxia-inducible factor-1alpha. *The Journal of Biological Chemistry*, *279*(43), 44976–44986. https://doi.org/10.1074/jbc.M313995200

Berry, J. D., Dyer, A., Cai, X., Garside, D. B., Ning, H., Thomas, A., … Lloyd-Jones, D. M. (2012). Lifetime risks of cardiovascular disease. *The New England Journal of Medicine*, *366*(4), 321–329. https://doi.org/10.1056/NEJMoa1012848

Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., … Wiswedel, B. (2008). KNIME: The Konstanz Information Miner. In C. Preisach, P. D. H. Burkhardt, P. D. L. Schmidt-Thieme, & P. D. R. Decker (Eds.), *Data Analysis, Machine Learning and Applications* (pp. 319–326). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/978-3-540-78246-9_38

Bevers, E. M., Comfurius, P., van Rijn, J. L., & Hemker, H. C. (1982). Generation of Prothrombin-Converting Activity and the Exposure of Phosphatidylserine at the Outer Surface of Platelets. *European Journal of Biochemistry*, *122*(2), 429–436.

Bhartiya, D., & Scaria, V. (2016). Genomic variations in non-coding RNAs: Structure, function and regulation. *Genomics*, *107*(2–3), 59–68. https://doi.org/10.1016/j.ygeno.2016.01.005

Bianchi, M., Crinelli, R., Giacomini, E., Carloni, E., & Magnani, M. (2009). A potent enhancer element in the 5'-UTR intron is crucial for transcriptional regulation of the human ubiquitin C gene. *Gene*, *448*(1), 88–101. https://doi.org/10.1016/j.gene.2009.08.013

Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., … Schwede, T. (2014). SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Research*, gku340. https://doi.org/10.1093/nar/gku340

Bourne, P. E. (2003). CASP and CAFASP experiments and their findings. *Methods of Biochemical Analysis*, *44*, 501–508.

Boyanova, D., Nilla, S., Birschmann, I., Dandekar, T., & Dittrich, M. (2012). PlateletWeb: a systems biologic analysis of signaling networks in human platelets. *Blood*, *119*(3), e22-34. https://doi.org/10.1182/blood-2011-10-387308

Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. A., Kasowski, M., … Snyder, M. (2012). Annotation of functional variation in personal genomes using RegulomeDB. *Genome Research*, *22*(9), 1790–1797. https://doi.org/10.1101/gr.137323.112

Braga-Neto, U., Hashimoto, R., Dougherty, E. R., Nguyen, D. V., & Carroll, R. J. (2004). Is cross-validation better than resubstitution for ranking genes? *Bioinformatics (Oxford, England)*, *20*(2), 253–258.

Brass, L. F., Hoxie, J. A., Kieber-Emmons, T., Manning, D. R., Poncz, M., & Woolkalis, M. (1993). Agonist receptors and G proteins as mediators of platelet activation. *Advances in Experimental Medicine and Biology*, *344*, 17–36.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Breiman, L., & Cutler, A. (2004). RFtools-for predicting and understanding data. In *Interface Workshop, April*.

Briggs, A., Clark, T., Wolstenholme, J., & Clarke, P. (2003). Missing.... presumed at random: cost-analysis of incomplete data. *Health Economics*, *12*(5), 377–392. https://doi.org/10.1002/hec.766

Brisson, C., Azorsa, D. O., Jennings, L. K., Moog, S., Cazenave, J. P., & Lanza, F. (1997). Co-localization of CD9 and GPIIb-IIIa (·IIb ·3 integrin) on activated platelet pseudopods and ·-granule membranes. *The Histochemical Journal*, *29*(2), 153–165. https://doi.org/10.1023/A:1026437522882

British Heart Foundation. (2015). Cardiovascular Disease Statistics 2015. Retrieved October 13, 2016, from https://www.bhf.org.uk/publications/statistics/cvd-stats-2015

Bromberg, Y. (2013). Building a Genome Analysis Pipeline to Predict Disease Risk and Prevent Disease. *Journal of Molecular Biology*, *425*(21), 3993–4005. https://doi.org/10.1016/j.jmb.2013.07.038

Brookes, A. J. (1999). The essence of SNPs. *Gene*, *234*(2), 177–186. https://doi.org/10.1016/S0378-1119(99)00219-X

Brown, R. P., & Feder, M. E. (2005). Reverse transcriptional profiling: non-correspondence of transcript level variation and proximal promoter polymorphism. *BMC Genomics*, *6*, 110. https://doi.org/10.1186/1471-2164-6-110

Brown, S., Heinisch, I., Ross, E., Shaw, K., Buckley, C. D., & Savill, J. (2002). Apoptosis disables CD31-mediated cell detachment from phagocytes promoting binding and engulfment. *Nature*, *418*(6894), 200–203. https://doi.org/10.1038/nature00811

Browning, S. R., & Browning, B. L. (2007). Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized

Haplotype Clustering. *The American Journal of Human Genetics*, *81*(5), 1084–1097. https://doi.org/10.1086/521987

Bryzgalov, L. O., Antontseva, E. V., Matveeva, M. Y., Shilov, A. G., Kashina, E. V., Mordvinov, V. A., & Merkulova, T. I. (2013). Detection of Regulatory SNPs in Human Genome Using ChIP-seq ENCODE Data. *PLOS ONE*, *8*(10), e78833. https://doi.org/10.1371/journal.pone.0078833

Buanne, P., Renzone, G., Monteleone, F., Vitale, M., Monti, S. M., Sandomenico, A., … Zambrano, N. (2013). Characterization of Carbonic Anhydrase IX Interactome Reveals Proteins Assisting Its Nuclear Localization in Hypoxic Cells. *Journal of Proteome Research*, *12*(1), 282–292. https://doi.org/10.1021/pr300565w

Buckland, P. R. (2006). The importance and identification of regulatory polymorphisms and their mechanisms of action. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, *1762*(1), 17–28. https://doi.org/10.1016/j.bbadis.2005.10.004

Bulger, M., & Groudine, M. (2010). Enhancers: the abundance and function of regulatory sequences beyond promoters. *Developmental Biology*, *339*(2), 250–257. https://doi.org/10.1016/j.ydbio.2009.11.035

Burga, A., & Lehner, B. (2012). Beyond genotype to phenotype: why the phenotype of an individual cannot always be predicted from their genome sequence and the environment that they experience. *FEBS Journal*, *279*(20), 3765–3775. https://doi.org/10.1111/j.1742-4658.2012.08810.x

Burga, A., & Lehner, B. (2013). Predicting phenotypic variation from genotypes, phenotypes and a combination of the two. *Current Opinion in Biotechnology*, *24*(4), 803–809. https://doi.org/10.1016/j.copbio.2013.03.004

Burke, D. F., Worth, C. L., Priego, E.-M., Cheng, T., Smink, L. J., Todd, J. A., & Blundell, T. L. (2007). Genome bioinformatic analysis of nonsynonymous SNPs. *BMC Bioinformatics*, *8*, 301. https://doi.org/10.1186/1471-2105-8-301

Burkhart, J. M., Gambaryan, S., Watson, S. P., Jurk, K., Walter, U., Sickmann, A., … Zahedi, R. P. (2014). What Can Proteomics Tell Us About Platelets? *Circulation Research*, *114*(7), 1204–1219. https://doi.org/10.1161/CIRCRESAHA.114.301598

Burnette, R., Simmons, L. A., & Snyderman, R. (2012). Personalized Health Care as a Pathway for the Adoption of Genomic Medicine. *Journal of Personalized Medicine*, *2*(4), 232–240. https://doi.org/10.3390/jpm2040232

Burton, P. R., Clayton, D. G., Cardon, L. R., Craddock, N., Deloukas, P., Duncanson, A., … others. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, *447*(7145), 661–678.

Calo, E., & Wysocka, J. (2013). Modification of enhancer chromatin: what, how and why? *Molecular Cell*, *49*(5), 825–37. https://doi.org/10.1016/j.molcel.2013.01.038

Campos, G. de los, Hickey, J. M., Pong-Wong, R., Daetwyler, H. D., & Calus, M. P. L. (2013). Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. *Genetics*, *193*(2), 327–345. https://doi.org/10.1534/genetics.112.143313

Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., … Lander, E. S. (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genetics*, *22*(3), 231–238. https://doi.org/10.1038/10290

Cariaso, M., & Lennon, G. (2012). SNPedia: a wiki supporting personal genome annotation, interpretation and analysis. *Nucleic Acids Research*, *40*(Database issue), D1308-1312. https://doi.org/10.1093/nar/gkr798

Cartharius, K., Frech, K., Grote, K., Klocke, B., Haltmeier, M., Klingenhoff, A., … Werner, T. (2005). MatInspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics*, *21*(13), 2933–2942. https://doi.org/10.1093/bioinformatics/bti473

Casonato, A., Sartorello, F., Pontara, E., Gallinaro, L., Bertomoro, A., Grazia Cattini, M., … Pagnan, A. (2007). A novel von Willebrand factor mutation (I1372S) associated with type 2B-like von Willebrand disease: an elusive phenotype and a difficult diagnosis. *Thrombosis and Haemostasis*, *98*(6), 1182–1187.

Catella-Lawson, F., Reilly, M. P., Kapoor, S. C., Cucchiara, A. J., DeMarco, S., Tournier, B., … FitzGerald, G. A. (2001). Cyclooxygenase inhibitors and the antiplatelet effects of aspirin. *The New England Journal of Medicine*, *345*(25), 1809–1817. https://doi.org/10.1056/NEJMoa003199

Cavallari, U., Trabetti, E., Malerba, G., Biscuola, M., Girelli, D., Olivieri, O., … Pignatti, P. F. (2007). Gene sequence variations of the platelet P2Y12 receptor are associated with coronary artery disease. *BMC Medical Genetics*, *8*, 59. https://doi.org/10.1186/1471-2350-8-59

Cavallo, A., & Martin, A. C. R. (2005). Mapping SNPs to protein sequence and structure data. *Bioinformatics*, *21*(8), 1443–1450. https://doi.org/10.1093/bioinformatics/bti220

Chen, R., Davydov, E. V., Sirota, M., & Butte, A. J. (2010). Non-Synonymous and Synonymous Coding SNPs Show Similar Likelihood and Effect Size of Human Disease Association. *PLOS ONE*, *5*(10), e13574. https://doi.org/10.1371/journal.pone.0013574

Chen, X., & Ishwaran, H. (2012). Random forests for genomic data analysis. *Genomics*, *99*(6), 323–329. https://doi.org/10.1016/j.ygeno.2012.04.003

Chhibber, A., French, C. E., Yee, S. W., Gamazon, E. R., Theusch, E., Qin, X., … Brenner, S. E. (2016). Transcriptomic variation of pharmacogenes in multiple human tissues and lymphoblastoid cell lines. *The Pharmacogenomics Journal*, *17*(2), 137–145. https://doi.org/10.1038/tpj.2015.93

Chien, M.-H., Yang, J.-S., Chu, Y.-H., Lin, C.-H., Wei, L.-H., Yang, S.-F., & Lin, C.-W. (2012). Impacts of CA9 Gene Polymorphisms and Environmental Factors on Oral-Cancer Susceptibility and Clinicopathologic Characteristics in Taiwan. *PLoS ONE*, *7*(12), e51051. https://doi.org/10.1371/journal.pone.0051051

Chivian, D., Robertson, T., Bonneau, R., & Baker, D. (2003). Ab Initio Methods. In P. E. Bourne & H. Weissig (Eds.), *Structural Bioinformatics* (pp. 547–557). John Wiley & Sons, Inc. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/0471721204.ch27/summary

Chorley, B. N., Wang, X., Campbell, M. R., Pittman, G. S., Noureddine, M. A., & Bell, D. A. (2008). Discovery and verification of functional single nucleotide polymorphisms in regulatory genomic regions: Current and developing technologies. *Mutation Research/Reviews in Mutation Research*, *659*(1–2), 147–157. https://doi.org/10.1016/j.mrrev.2008.05.001

Chothia, C., & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *The EMBO Journal*, *5*(4), 823–826.

Cipolla, L., Consonni, A., Guidetti, G., Canobbio, I., Okigaki, M., Falasca, M., … Torti, M. (2013). The proline-rich tyrosine kinase Pyk2 regulates platelet integrin αIIbβ3 outside-in signaling. *Journal of Thrombosis and Haemostasis: JTH*, *11*(2), 345–356. https://doi.org/10.1111/jth.12099

Coles, C. H., Mitakidis, N., Zhang, P., Elegheert, J., Lu, W., Stoker, A. W., … Aricescu, A. R. (2014). Structural basis for extracellular cis and trans RPTPσ signal competition in synaptogenesis. *Nature Communications*, *5*. https://doi.org/10.1038/ncomms6209

Collins, F. S., Brooks, L. D., & Chakravarti, A. (1998). A DNA polymorphism discovery resource for research on human genetic variation. *Genome Research*, *8*(12), 1229–1231.

Collins, F. S., & Varmus, H. (2015). A New Initiative on Precision Medicine. *New England Journal of Medicine*, *372*(9), 793–795. https://doi.org/10.1056/NEJMp1500523

Consortium, I. H. (2005). A haplotype map of the human genome. *Nature*, *437*(7063), 1299–1320.

Consortium, T. 1000 G. P. (2010). A map of human genome variation from population-scale sequencing. *Nature*, *467*(7319), 1061–1073. https://doi.org/10.1038/nature09534

Consortium, T. 1000 G. P. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, *491*(7422), 56–65. https://doi.org/10.1038/nature11632

Consortium, T. E. P. (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, *306*(5696), 636–640. https://doi.org/10.1126/science.1105136

Consortium, T. U. (2008). The Universal Protein Resource (UniProt). *Nucleic Acids Research*, *36*(suppl 1), D190–D195. https://doi.org/10.1093/nar/gkm895

Coppinger, J. A., Cagney, G., Toomey, S., Kislinger, T., Belton, O., McRedmond, J. P., … Maguire, P. B. (2004). Characterization of the proteins released from activated platelets leads to localization of novel platelet proteins in human atherosclerotic lesions. *Blood*, *103*(6), 2096–2104.

Cordell, H. J., & Clayton, D. G. (2002). A Unified Stepwise Regression Procedure for Evaluating the Relative Effects of Polymorphisms within a Gene Using Case/Control or Family Data: Application to HLA in Type 1 Diabetes. *The American Journal of Human Genetics*, *70*(1), 124–141. https://doi.org/10.1086/338007

Cordell, H. J., & Clayton, D. G. (2005). Genetic association studies. *The Lancet*, *366*(9491), 1121–1131. https://doi.org/10.1016/S0140-6736(05)67424-7

Cosemans, J. M. E. M., Munnix, I. C. A., Wetzker, R., Heller, R., Jackson, S. P., & Heemskerk, J. W. M. (2006). Continuous signaling via PI3K isoforms β and γ is required for platelet ADP receptor function in dynamic thrombus stabilization. *Blood*, *108*(9), 3045–3052. https://doi.org/10.1182/blood-2006-03-006338

Coughlin, S. R. (2000). Thrombin signalling and protease-activated receptors. *Nature*, *407*(6801), 258–264.

Cox, D., Smith, R., Quinn, M., Theroux, P., Crean, P., & Fitzgerald, D. J. (2000). Evidence of platelet activation during treatment with a GPIIb/IIIa antagonist in patients presenting with acute coronary syndromes. *Journal of the American College of Cardiology*, *36*(5), 1514–1519. https://doi.org/10.1016/S0735-1097(00)00919-0

Craddock, N., Hurles, M. E., Cardin, N., Pearson, R. D., Plagnol, V., Robson, S., … Donnelly, P. (2010). Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature*, *464*(7289), 713–720. https://doi.org/10.1038/nature08979

Cule, E. (2015). Package "ridge." Retrieved from https://cran.r-project.org/web/packages/ridge/ridge.pdf

Cule, E., Vineis, P., & Iorio, M. D. (2011). Significance testing in ridge regression for genetic data. *BMC Bioinformatics*, *12*(1), 372. https://doi.org/10.1186/1471-2105-12-372

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

de Bono, J. S., & Ashworth, A. (2010). Translating cancer research into targeted therapeutics. *Nature*, *467*(7315), 543–549. https://doi.org/10.1038/nature09339

de Keyser, C. E., Eijgelsheim, M., Hofman, A., Sijbrands, E. J. G., Maitland-van der Zee, A.-H., van Duijn, C. M., … Ch Stricker, B. H. (2011). Single nucleotide polymorphisms in genes that are associated with a modified response to statin therapy: the Rotterdam Study. *The Pharmacogenomics Journal*, *11*(1), 72–80. https://doi.org/10.1038/tpj.2010.11

de los Campos, G., Gianola, D., & Allison, D. B. (2010). Predicting genetic predisposition in humans: the promise of whole-genome markers. *Nature Reviews Genetics*, *11*(12), 880–886. https://doi.org/10.1038/nrg2898

de Martino, M., Klatte, T., Seligson, D. B., LaRochelle, J., Shuch, B., Caliliw, R., … Belldegrun, A. S. (2009). CA9 Gene: Single Nucleotide Polymorphism Predicts Metastatic Renal Cell Carcinoma Prognosis. *The Journal of Urology*, *182*(2), 728–734. https://doi.org/10.1016/j.juro.2009.03.077

DeLano, W. L. (2002). The PyMOL user's manual. *DeLano Scientific, San Carlos, CA, 452*.

Deng, H., & Shen, H. (2007). *Current Topics in Human Genetics: Studies in Complex Diseases*. Singapore: World Scientific.

Díaz-Uriarte, R., & Andrés, S. A. de. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, *7*(1), 3. https://doi.org/10.1186/1471-2105-7-3

Dionisio, N., Redondo, P. C., Jardin, I., & Rosado, J. A. (2012). Transient receptor potential channels in human platelets: expression and functional role. *Current Molecular Medicine*, *12*(10), 1319–1328.

Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., … Gingeras, T. R. (2012). Landscape of transcription in human cells. *Nature*, *489*(7414), 101–108. https://doi.org/10.1038/nature11233

Drachkova, I. A., Ponomarenko, P. M., Arshinova, T. V., Ponomarenko, M. P., Suslov, V. V., Savinkova, L. K., … others. (2011). In vitro examining the existing prognoses how TBP binds to TATA with SNP associated with human diseases. *Health*, *3*(9), 577.

Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H., & Nadeau, J. H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews. Genetics*, *11*(6), 446–450. https://doi.org/10.1038/nrg2809

El Bouhassani, M., Gilibert, S., Moreau, M., Saint-Charles, F., Tréguier, M., Poti, F., … Huby, T. (2011). Cholesteryl ester transfer protein expression partially attenuates the adverse effects of SR-BI receptor deficiency on cholesterol metabolism and atherosclerosis. *The Journal of Biological Chemistry*, *286*(19), 17227–17238. https://doi.org/10.1074/jbc.M111.220483

Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., … Bernstein, B. E. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, *473*(7345), 43–49. https://doi.org/10.1038/nature09906

Faraday, N., Yanek, L. R., Yang, X. P., Mathias, R., Herrera-Galeano, J. E., Suktitipat, B., … Becker, L. C. (2011). Identification of a specific intronic PEAR1 gene variant associated with greater platelet aggregability and protein expression. *Blood*, *118*(12), 3367–3375. https://doi.org/10.1182/blood-2010-11-320788

Fazel-Najafabadi, E., Vahdat Ahar, E., Fattahpour, S., & Sedghi, M. (2015). Structural and functional impact of missense mutations in TPMT: An integrated computational approach. *Computational Biology and Chemistry*, *59, Part A*, 48–55. https://doi.org/10.1016/j.compbiolchem.2015.09.004

Feinberg, A. P. (2010). Genome-scale approaches to the epigenetics of common human disease. *Virchows Archiv: An International Journal of Pathology*, *456*(1), 13–21. https://doi.org/10.1007/s00428-009-0847-2

Fernald, G. H., Capriotti, E., Daneshjou, R., Karczewski, K. J., & Altman, R. B. (2011). Bioinformatics challenges for personalized medicine. *Bioinformatics*, *27*(13), 1741–1748.

Flaherty, D. K. (2007). Single nucleotide polymorphisms, drug metabolism and untoward health effects. *J Med Biol Sci*, *1*(2). Retrieved from https://www.yumpu.com/en/document/view/26129147/single-nucleotide-polymorphisms-drug-metabolism-and-untoward-

Fontana, P., Dupont, A., Gandrille, S., Bachelot-Loza, C., Reny, J.-L., Aiach, M., & Gaussem, P. (2003). Adenosine Diphosphate–Induced Platelet Aggregation Is Associated With P2Y12 Gene Sequence Variations in Healthy Subjects. *Circulation*, *108*(8), 989–995. https://doi.org/10.1161/01.CIR.0000085073.69189.88

Foulkes, A. S. (2009). *Applied Statistical Genetics with R: For Population-based Association Studies*. New York, USA: Springer Science+Business Media.

Franceschini, N., Muallem, H., Rose, K. M., Boerwinkle, E., & Maeda, N. (2009). LDL Receptor Polymorphisms and the Risk of Coronary Heart Disease: the Atherosclerosis Risk in Communities Study. *Journal of Thrombosis and Haemostasis : JTH*, *7*(3), 496–498. https://doi.org/10.1111/j.1538-7836.2008.03262.x

Franke, L., & Jansen, R. C. (2009). eQTL analysis in humans. *Methods in Molecular Biology (Clifton, N.J.)*, *573*, 311–328. https://doi.org/10.1007/978-1-60761-247-6_17

Fraser, H. B., & Xie, X. (2009). Common polymorphic transcript variation in human disease. *Genome Research*, *19*(4), 567–575. https://doi.org/10.1101/gr.083477.108

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1.

Fröbel, J., Cadeddu, R.-P., Hartwig, S., Bruns, I., Wilk, C. M., Kündgen, A., … Czibere, A. (2013). Platelet proteome analysis reveals integrin-dependent aggregation defects in patients with myelodysplastic syndromes. *Molecular & Cellular Proteomics: MCP*, *12*(5), 1272–1280. https://doi.org/10.1074/mcp.M112.023168

Frost, S. C., & McKenna, R. (2013). *Carbonic Anhydrase: Mechanism, Regulation, Links to Disease, and Industrial Applications*. Springer Science & Business Media.

Gachet, C., Hechler, B., Leon, C., Vial, C., Leray, C., Ohlmann, P., & Cazenave, J. P. (1997). Activation of ADP receptors and platelet function. *Thrombosis and Haemostasis*, *78*(1), 271–275.

Gardener, H., Beecham, A., Cabral, D., Yanuck, D., Slifer, S., Wang, L., … Rundek, T. (2011). Carotid plaque and candidate genes related to inflammation and endothelial function in Hispanics from northern Manhattan. *Stroke; a Journal of Cerebral Circulation*, *42*(4), 889–896. https://doi.org/10.1161/STROKEAHA.110.591065

Genuer, R., Poggi, J.-M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, *31*(14), 2225–2236. https://doi.org/10.1016/j.patrec.2010.03.014

Gerasimova, A., Chavez, L., Li, B., Seumois, G., Greenbaum, J., Rao, A., … Peters, B. (2013). Predicting Cell Types and Genetic Variations Contributing to Disease by Combining GWAS and Epigenetic Data. *PLoS ONE*, *8*(1), e54359. https://doi.org/10.1371/journal.pone.0054359

Gerstein, M. B., Kundaje, A., Hariharan, M., Landt, S. G., Yan, K.-K., Cheng, C., … Snyder, M. (2012). Architecture of the human regulatory network derived from ENCODE data. *Nature*, *489*(7414), 91–100. https://doi.org/10.1038/nature11245

Gibbins, J. M. (2004). Platelet adhesion signalling and the regulation of thrombus formation. *Journal of Cell Science*, *117*(16), 3415–3425. https://doi.org/10.1242/jcs.01325

Gibson, A. W., Edberg, J. C., Wu, J., Westendorp, R. G. J., Huizinga, T. W. J., & Kimberly, R. P. (2001). Novel Single Nucleotide Polymorphisms in the Distal IL-10 Promoter Affect IL-10 Production and Enhance the Risk of Systemic Lupus Erythematosus. *The Journal of Immunology*, *166*(6), 3915–3922. https://doi.org/10.4049/jimmunol.166.6.3915

Gieger, C., Radhakrishnan, A., Cvejic, A., Tang, W., Porcu, E., Pistis, G., … Soranzo, N. (2011). New gene functions in megakaryopoiesis and platelet formation. *Nature*, *480*(7376), 201–208. https://doi.org/10.1038/nature10659

Gifford, J. L., Walsh, M. P., & Vogel, H. J. (2007). Structures and metal-ion-binding properties of the Ca2+-binding helix–loop–helix EF-hand motifs. *Biochemical Journal*, *405*(2), 199–221. https://doi.org/10.1042/BJ20070255

Ginsburg, G. S., & Willard, H. F. (2009). Genomic and personalized medicine: foundations and applications. *Translational Research*, *154*(6), 277–287.

Glembotsky, A. C., Bluteau, D., Espasandin, Y. R., Goette, N. P., Marta, R. F., Marin Oyarzun, C. P., … Heller, P. G. (2014). Mechanisms underlying platelet function defect in a pedigree with familial platelet disorder with a predisposition to acute myelogenous leukemia: potential role for candidate RUNX1 targets. *Journal of Thrombosis and Haemostasis: JTH*, *12*(5), 761–772. https://doi.org/10.1111/jth.12550

Glisovic, T., Bachorik, J. L., Yong, J., & Dreyfuss, G. (2008). RNA-binding proteins and post-transcriptional gene regulation. *FEBS Letters*, *582*(14), 1977–1986. https://doi.org/10.1016/j.febslet.2008.03.004

Gnad, F., Baucom, A., Mukhyala, K., Manning, G., & Zhang, Z. (2013). Assessment of computational methods for predicting the effects of missense mutations in human cancers. *BMC Genomics*, *14*(3), S7. https://doi.org/10.1186/1471-2164-14-S3-S7

Godzik, A. (2003). Fold Recognition Methods. In P. E. Bourne & H. Weissig (Eds.), *Structural Bioinformatics* (pp. 525–546). John Wiley & Sons, Inc. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/0471721204.ch26/summary

Goldstein, B. A., Polley, E. C., & Briggs, F. B. S. (2011). Random Forests for Genetic Association Studies. *Statistical Applications in Genetics and Molecular Biology*, *10*(1), 1–34. https://doi.org/10.2202/1544-6115.1691

Gong, N., & Chatterjee, S. (2003). Platelet endothelial cell adhesion molecule in cell signaling and thrombosis. *Molecular and Cellular Biochemistry*, *253*(1–2), 151–158.

González, S., Montserrat-Sentís, B., Sánchez, F., Puiggròs, M., Blanco, E., Ramirez, A., & Torrents, D. (2012). ReLA, a local alignment search tool for the identification of distal and proximal gene regulatory regions and their conserved transcription factor binding sites. *Bioinformatics*, *28*(6), 763–770. https://doi.org/10.1093/bioinformatics/bts024

González-Camacho, J. M., Campos, G. de los, Pérez, P., Gianola, D., Cairns, J. E., Mahuku, G., … Crossa, J. (2012). Genome-enabled prediction of genetic values using radial basis

function neural networks. *Theoretical and Applied Genetics*, *125*(4), 759–771. https://doi.org/10.1007/s00122-012-1868-9

González-Camacho, J. M., Crossa, J., Pérez-Rodríguez, P., Ornella, L., & Gianola, D. (2016). Genome-enabled prediction using probabilistic neural network classifiers. *BMC Genomics*, *17*, 208. https://doi.org/10.1186/s12864-016-2553-1

Goodall, A. H., Burns, P., Salles, I., Macaulay, I. C., Jones, C. I., Ardissino, D., … Dudbridge, F. (2010). Transcription profiling in human platelets reveals LRRFIP1 as a novel protein regulating platelet function. *Blood*, *116*(22), 4646–4656.

Grabarek, Z. (2006). Structural Basis for Diversity of the EF-hand Calcium-binding Proteins. *Journal of Molecular Biology*, *359*(3), 509–525. https://doi.org/10.1016/j.jmb.2006.03.066

GuhaThakurta, D., Xie, T., Anand, M., Edwards, S. W., Li, G., Wang, S. S., & Schadt, E. E. (2006). Cis-regulatory variations: a study of SNPs around genes showing cis-linkage in segregating mouse populations. *BMC Genomics*, *7*, 235. https://doi.org/10.1186/1471-2164-7-235

Guo, L., Du, Y., Chang, S., Zhang, K., & Wang, J. (2014). rSNPBase: a database for curated regulatory SNPs. *Nucleic Acids Research*, *42*(Database issue), D1033–D1039. https://doi.org/10.1093/nar/gkt1167

Gutmanas, A., Alhroub, Y., Battle, G. M., Berrisford, J. M., Bochet, E., Conroy, M. J., … Kleywegt, G. J. (2014). PDBe: Protein Data Bank in Europe. *Nucleic Acids Research*, *42*(D1), D285–D291. https://doi.org/10.1093/nar/gkt1180

Haas, J., Roth, S., Arnold, K., Kiefer, F., Schmidt, T., Bordoli, L., & Schwede, T. (2013). The Protein Model Portal—a comprehensive resource for protein structure and model information. *Database*, *2013*, bat031. https://doi.org/10.1093/database/bat031

Haeseleer, F., Imanishi, Y., Sokal, I., Filipek, S., & Palczewski, K. (2002). Calcium-binding proteins: intracellular sensors from the calmodulin superfamily. *Biochemical and Biophysical Research Communications*, *290*(2), 615–623.

Hallikas, O., Palin, K., Sinjushina, N., Rautiainen, R., Partanen, J., Ukkonen, E., & Taipale, J. (2006). Genome-wide Prediction of Mammalian Enhancers Based on Analysis of Transcription-Factor Binding Affinity. *Cell*, *124*(1), 47–59. https://doi.org/10.1016/j.cell.2005.10.042

Hamid, J. S., Hu, P., Roslin, N. M., Ling, V., Greenwood, C. M. T., & Beyene, J. (2009). Data Integration in Genetics and Genomics: Methods and Challenges. *Human Genomics and Proteomics : HGP*, *2009*. https://doi.org/10.4061/2009/869093

Han, F., & Pan, W. (2010). A data-adaptive sum test for disease association with multiple common or rare variants. *Human Heredity*, *70*(1), 42–54. https://doi.org/10.1159/000288704

Hanson, C., Cairns, J., Wang, L., & Sinha, S. (2015). Computational discovery of transcription factors associated with drug response. *The Pharmacogenomics Journal*, *16*, 573–582. https://doi.org/10.1038/tpj.2015.74

Harrell, F. E. (2001). *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. New York, USA: Springer-Verlag.

Harrison, P. (2005). Platelet function analysis. *Blood Reviews*, *19*(2), 111–123.

Hartwig, J. H. (2006). The Platelet: Form and Function. *Seminars in Hematology*, *43*, *Supplement 1*, S94–S100. https://doi.org/10.1053/j.seminhematol.2005.11.004

Hastie, T., Tibshirani, R., Friedman, J., & Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, *27*(2), 83–85.

Hawkins, R. D., Hon, G. C., & Ren, B. (2011). Next-generation genomics: an integrative approach. *Nature Reviews Genetics*, *11*(7), 476–86. https://doi.org/10.1038/nrg2795

Heintzman, N. D., & Ren, B. (2009). Finding distal regulatory elements in the human genome. *Current Opinion in Genetics & Development*, *19*(6), 541–549. https://doi.org/10.1016/j.gde.2009.09.006

Helles, G. (2008). A comparative study of the reported performance of ab initio protein structure prediction algorithms. *Journal of The Royal Society Interface*, *5*(21), 387–396. https://doi.org/10.1098/rsif.2007.1278

Herrera-Galeano, J. E., Becker, D. M., Wilson, A. F., Yanek, L. R., Bray, P., Vaidya, D., … Becker, L. C. (2008). A novel variant in the platelet endothelial aggregation receptor-1 gene is associated with increased platelet aggregability. *Arteriosclerosis, Thrombosis, and Vascular Biology*, *28*(8), 1484–1490.

Hildebrand, R. B., Lammers, B., Meurs, I., Korporaal, S. J. A., De Haan, W., Zhao, Y., … Van Eck, M. (2010). Restoration of high-density lipoprotein levels by cholesteryl ester transfer protein expression in scavenger receptor class B type I (SR-BI) knockout mice does not normalize pathologies associated with SR-BI deficiency. *Arteriosclerosis, Thrombosis, and Vascular Biology*, *30*(7), 1439–1445. https://doi.org/10.1161/ATVBAHA.110.205153

Hirschhorn, J. N., & Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, *6*(2), 95–108.

Hodges, G. J., Gros, R., Hegele, R. A., Van Uum, S., Shoemaker, J. K., & Feldman, R. D. (2010). Increased blood pressure and hyperdynamic cardiovascular responses in carriers of a common hyperfunctional variant of adenylyl cyclase 6. *The Journal of Pharmacology and Experimental Therapeutics*, *335*(2), 451–457. https://doi.org/10.1124/jpet.110.172700

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67.

Hoffmann, T. J., Marini, N. J., & Witte, J. S. (2010). Comprehensive approach to analyzing rare genetic variants. *PloS One*, *5*(11), e13584. https://doi.org/10.1371/journal.pone.0013584

Hoggart, C. J., Whittaker, J. C., De Iorio, M., & Balding, D. J. (2008). Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genetics*, *4*(7), e1000130.

Hollopeter, G., Jantzen, H.-M., Vincent, D., Li, G., England, L., Ramakrishnan, V., … Conley, P. B. (2001). Identification of the platelet ADP receptor targeted by antithrombotic drugs. *Nature*, *409*(6817), 202–207. https://doi.org/10.1038/35051599

Holm, L., & Sander, C. (1996). Decision support system for the evolutionary classification of protein structures. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, *5*, 140–146.

Holzinger, E. R., & Ritchie, M. D. (2012). Integrating heterogeneous high-throughput data for meta-dimensional pharmacogenomics and disease-related studies. *Pharmacogenomics*, *13*(2), 213–222. https://doi.org/10.2217/pgs.11.145

Hood, L., & Flores, M. (2012). A personal view on systems medicine and the emergence of proactive P4 medicine: predictive, preventive, personalized and participatory. *New Biotechnology*, *29*(6), 613–624. https://doi.org/10.1016/j.nbt.2012.03.004

Hoogendoorn, B., Coleman, S. L., Guy, C. A., Smith, K., Bowen, T., Buckland, P. R., & O'Donovan, M. C. (2003). Functional analysis of human promoter polymorphisms. *Human Molecular Genetics*, *12*(18), 2249–2254. https://doi.org/10.1093/hmg/ddg246

Howie, B. N., Donnelly, P., & Marchini, J. (2009). A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLoS Genet*, *5*(6), e1000529. https://doi.org/10.1371/journal.pgen.1000529

Hudson, T. J. (2003). Wanted: regulatory SNPs. *Nature Genetics*, *33*(4), 439–440. https://doi.org/10.1038/ng0403-439

Hull, J., Campino, S., Rowlands, K., Chan, M.-S., Copley, R. R., Taylor, M. S., … Kwiatkowski, D. (2007). Identification of Common Genetic Variation That Modulates Alternative Splicing. *PLoS Genet*, *3*(6), e99. https://doi.org/10.1371/journal.pgen.0030099

Hutchins, J. R. A. (2014). What's that gene (or protein)? Online resources for exploring functions of genes, transcripts, and proteins. *Molecular Biology of the Cell*, *25*(8), 1187–1201. https://doi.org/10.1091/mbc.E13-10-0602

Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y., … Lee, C. (2004). Detection of large-scale variation in the human genome. *Nature Genetics*, *36*(9), 949–951. https://doi.org/10.1038/ng1416

Ihaka, R., & Gentleman, R. (1996). R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, *5*(3), 299–314. https://doi.org/10.1080/10618600.1996.10474713

Ihnatko, R., Kubes, M., Takacova, M., Sedlakova, O., Sedlak, J., Pastorek, J., … Pastorekova, S. (2006). Extracellular acidosis elevates carbonic anhydrase IX in human glioblastoma

cells via transcriptional modulation that does not depend on hypoxia. *International Journal of Oncology*, *29*(4), 1025–1033.

Ikeda, Y., Handa, M., Kawano, K., Kamata, T., Murata, M., Araki, Y., … Itagaki, I. (1991). The role of von Willebrand factor and fibrinogen in platelet aggregation under varying shear stress. *Journal of Clinical Investigation*, *87*(4), 1234–1240.

Italiano Jr, J. E., Richardson, J. L., Patel-Hett, S., Battinelli, E., Zaslavsky, A., Short, S., … Klement, G. L. (2008). Angiogenesis is regulated by a novel mechanism: pro-and antiangiogenic proteins are organized into separate platelet α granules and differentially released. *Blood*, *111*(3), 1227–1233.

Jackson, S. P., Nesbitt, W. S., & Kulkarni, S. (2003). Signaling events underlying thrombus formation. *Journal of Thrombosis and Haemostasis*, *1*(7), 1602–1612. https://doi.org/10.1046/j.1538-7836.2003.00267.x

Janssens, A. C. J. W., Aulchenko, Y. S., Elefante, S., Borsboom, G. J. J. M., Steyerberg, E. W., & van Duijn, C. M. (2006). Predictive testing for complex diseases using multiple genes: Fact or fiction? *Genetics in Medicine*, *8*(7), 395–400. https://doi.org/10.1097/01.gim.0000229689.18263.f4

Jardin, I., Albarrán, L., Bermejo, N., Salido, G. M., & Rosado, J. A. (2012). Homers regulate calcium entry and aggregation in human platelets: a role for Homers in the association between STIM1 and Orai1. *The Biochemical Journal*, *445*(1), 29–38. https://doi.org/10.1042/BJ20120471

Jayachandran, M., Preston, C. C., Hunter, L. W., Jahangir, A., Owen, W. G., Korach, K. S., & Miller, V. M. (2010). Loss of estrogen receptor β decreases mitochondrial energetic potential and increases thrombogenicity of platelets in aged female mice. *Age*, *32*(1), 109–121. https://doi.org/10.1007/s11357-009-9119-y

Jiang, C., Xuan, Z., Zhao, F., & Zhang, M. Q. (2007). TRED: a transcriptional regulatory element database, new entries and other development. *Nucleic Acids Research*, *35*(Database issue), D137–D140. https://doi.org/10.1093/nar/gkl1041

Jiang, L., Xu, C., Yu, S., Liu, P., Luo, D., Zhou, Q., … Hu, H. (2013). A critical role of thrombin/PAR-1 in ADP-induced platelet secretion and the second wave of aggregation. *Journal of Thrombosis and Haemostasis: JTH*, *11*(5), 930–940. https://doi.org/10.1111/jth.12168

Jiang, S., & Caffrey, M. (2007). Solution structure of the coxsackievirus and adenovirus receptor domain 2. *Protein Science: A Publication of the Protein Society*, *16*(3), 539–542. https://doi.org/10.1110/ps.062643507

Jin, J., Quinton, T. M., Zhang, J., Rittenhouse, S. E., & Kunapuli, S. P. (2002). Adenosine diphosphate (ADP)–induced thromboxane A2generation in human platelets requires coordinated signaling through integrin αIIbβ3 and ADP receptors. *Blood*, *99*(1), 193–198. https://doi.org/10.1182/blood.V99.1.193

Jin, Y., Marunde, R., Desta, Z., Nguyen, A., Storniolo, A. M., & Flockhart, D. A. (2006). Estrogen receptor genotype is associated with platelet inhibition after tamoxifen treatment. *<I>The FASEB Journal</I>*, *20*(4), A267.

Johnson, A. D., Yanek, L. R., Chen, M.-H., Faraday, N., Larson, M. G., Tofler, G., … Becker, L. C. (2010). Genome-wide meta-analyses identifies seven loci associated with platelet aggregation in response to agonists. *Nature Genetics*, *42*(7), 608–613. https://doi.org/10.1038/ng.604

Jones, C. I., Bray, S., Garner, S. F., Stephens, J., de Bono, B., Angenent, W. G. J., … Bloodomics Consortium. (2009). A functional genomics approach reveals novel quantitative trait loci associated with platelet signaling pathways. *Blood*, *114*(7), 1405–1416. https://doi.org/10.1182/blood-2009-02-202614

Jones, C. I., Garner, S. F., Angenent, W., Bernard, A., Berzuini, C., Burns, P., … On Behalf of the Bloodomics Consortium. (2007). Mapping the platelet profile for functional genomic studies and demonstration of the effect size of the GP6 locus. *Journal of Thrombosis and Haemostasis*, *5*(8), 1756–1765. https://doi.org/10.1111/j.1538-7836.2007.02632.x

Jones, D. T., Taylort, W. R., & Thornton, J. M. (1992). A new approach to protein fold recognition. *Nature*, *358*(6381), 86–89. https://doi.org/10.1038/358086a0

Jones, P. A., & Takai, D. (2001). The Role of DNA Methylation in Mammalian Epigenetics. *Science*, *293*(5532), 1068–1070. https://doi.org/10.1126/science.1063852

Källberg, M., Wang, H., Wang, S., Peng, J., Wang, Z., Lu, H., & Xu, J. (2012). Template-based protein structure modeling using the RaptorX web server. *Nature Protocols*, *7*(8), 1511–1522. https://doi.org/10.1038/nprot.2012.085

Kaluz, S., Kaluzová, M., & Stanbridge, E. J. (2003). Expression of the hypoxia marker carbonic anhydrase IX is critically dependent on SP1 activity. Identification of a novel type of hypoxia-responsive enhancer. *Cancer Research*, *63*(5), 917–922.

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S., Freimer, N. B., … Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, *42*(4), 348–354. https://doi.org/10.1038/ng.548

Kathiresan, S., Voight, B. F., Purcell, S., Musunuru, K., Ardissino, D., Mannucci, P. M., … Altshuler, D. (2009). Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nature Genetics*, *41*(3), 334–341. https://doi.org/10.1038/ng.327

Katsel, P. L., Tagliente, T. M., Schwarz, T. E., Craddock-Royal, B. D., Patel, N. D., & Maayani, S. (2003). Molecular and biochemical evidence for the presence of type III adenylyl cyclase in human platelets. *Platelets*, *14*(1), 21–33.

Kauskot, A., Michele, M. D., Loyen, S., Freson, K., Verhamme, P., & Hoylaerts, M. F. (2012). A novel mechanism of sustained platelet αIIbβ3 activation via PEAR1. *Blood*, *119*(17), 4056–4065. https://doi.org/10.1182/blood-2011-11-392787

Keating, S. T., Plutzky, J., & El-Osta, A. (2016). Epigenetic Changes in Diabetes and Cardiovascular Risk. *Circulation Research*, *118*(11), 1706–1722. https://doi.org/10.1161/CIRCRESAHA.116.306819

Kettlewell, S., Cabrero, P., Nicklin, S. A., Dow, J. A. T., Davies, S., & Smith, G. L. (2009). Changes of intra-mitochondrial Ca2+ in adult ventricular cardiomyocytes examined using a novel fluorescent Ca2+ indicator targeted to mitochondria. *Journal of Molecular and Cellular Cardiology*, *46*(6), 891–901. https://doi.org/10.1016/j.yjmcc.2009.02.016

Kim, D. E., Chivian, D., & Baker, D. (2004). Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Research*, *32*(suppl 2), W526–W531. https://doi.org/10.1093/nar/gkh468

Kimura, H., Miyazaki, R., Imura, T., Masunaga, S., Shimada, A., Mikami, D., … Yoshida, H. (2011). Smaller low-density lipoprotein size as a possible risk factor for the prevalence of coronary artery diseases in haemodialysis patients: associations of cholesteryl ester transfer protein and the hepatic lipase gene polymorphism with low-density lipoprotein size. *Nephrology (Carlton, Vic.)*, *16*(6), 558–566. https://doi.org/10.1111/j.1440-1797.2011.01454.x

Kirschbaum, N. E., Gumina, R. J., & Newman, P. J. (1994). Organization of the gene for human platelet/endothelial cell adhesion molecule-1 shows alternatively spliced isoforms and a functionally complex cytoplasmic domain. *Blood*, *84*(12), 4028–4037.

Klages, B., Brandt, U., Simon, M. I., Schultz, G., & Offermanns, S. (1999). Activation of G12/G13 Results in Shape Change and Rho/Rho-Kinase–mediated Myosin Light Chain Phosphorylation in Mouse Platelets. *The Journal of Cell Biology*, *144*(4), 745–754.

Koch, E., Ristroph, M., & Kirkpatrick, M. (2013). Long Range Linkage Disequilibrium across the Human Genome. *PLOS ONE*, *8*(12), e80754. https://doi.org/10.1371/journal.pone.0080754

Kohannim, O., Hibar, D. P., Stein, J. L., Jahanshad, N., Hua, X., Rajagopalan, P., … Thompson, P. M. (2012). Discovery and Replication of Gene Influences on Brain Structure Using LASSO Regression. *Frontiers in Neuroscience*, *6*, 115. https://doi.org/10.3389/fnins.2012.00115

Koivisto, U. M., Palvimo, J. J., Jänne, O. A., & Kontula, K. (1994). A single-base substitution in the proximal Sp1 site of the human low density lipoprotein receptor promoter as a cause of heterozygous familial hypercholesterolemia. *Proceedings of the National Academy of Sciences*, *91*(22), 10526–10530.

Kojima, H., Kanada, H., Shimizu, S., Kasama, E., Shibuya, K., Nakauchi, H., … Shibuya, A. (2003). CD226 Mediates Platelet and Megakaryocytic Cell Adhesion to Vascular Endothelial Cells. *Journal of Biological Chemistry*, *278*(38), 36748–36753. https://doi.org/10.1074/jbc.M300702200

Kopacek, J., Barathova, M., Dequiedt, F., Sepelakova, J., Kettmann, R., Pastorek, J., & Pastorekova, S. (2005). MAPK pathway contributes to density- and hypoxia-induced expression of the tumor-associated carbonic anhydrase IX. *Biochimica Et Biophysica Acta*, *1729*(1), 41–49. https://doi.org/10.1016/j.bbaexp.2005.03.003

Kouranov, A., Xie, L., Cruz, J. de la, Chen, L., Westbrook, J., Bourne, P. E., & Berman, H. M. (2006). The RCSB PDB information portal for structural genomics. *Nucleic Acids Research*, *34*(suppl 1), D302–D305. https://doi.org/10.1093/nar/gkj120

Kraft, P., Wacholder, S., Cornelis, M. C., Hu, F. B., Hayes, R. B., Thomas, G., … Chanock, S. (2009). Beyond odds ratios — communicating disease risk based on genetic profiles. *Nature Reviews Genetics*, *10*(4), 264–269. https://doi.org/10.1038/nrg2516

Kraus, S., Hummler, S., Toriola, A. T., Poole, E. M., Scherer, D., Kotzmann, J., … Ulrich, C. M. (2013). Impact of genetic polymorphisms on adenoma recurrence and toxicity in a COX2 inhibitor (celecoxib) trial: results from a pilot study. *Pharmacogenetics and Genomics*, *23*(8), 428–437. https://doi.org/10.1097/FPC.0b013e3283631784

Krieger, E., Nabuurs, S. B., & Vriend, G. (2003). Homology modeling. In *Structural Bioinformatics* (pp. 509–524). Wiley.

Kryshtafovych, A., & Fidelis, K. (2009). Protein structure prediction and model quality assessment. *Drug Discovery Today*, *14*(7–8), 386–393. https://doi.org/10.1016/j.drudis.2008.11.010

Kryshtafovych, A., Fidelis, K., & Moult, J. (2010). CASP: a driving force in protein structure modeling. In *Introduction to Protein Structure Prediction: Methods and Algorithms* (pp. 15–32). New Jersey: Wiley. Retrieved from https://books.google.co.uk/books?hl=en&lr=&id=LeRhAoz4NwEC&oi=fnd&pg=PA15&dq=casp:+a+driving+force+in+protein+structure+modelling&ots=yA5b6G2HQ1&sig=4dX0CK4MWxhJoZ2kmrF5XxWHEk0

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., … Ma'ayan, A. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*, *44*(W1), W90-97. https://doi.org/10.1093/nar/gkw377

Kursa, M. B. (2014). Robustness of Random Forest-based gene selection methods. *BMC Bioinformatics*, *15*, 8. https://doi.org/10.1186/1471-2105-15-8

Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the Boruta package. *Journal of Statistical Software*, *36*. Retrieved from http://www.jstatsoft.org/v36/i11/paper

Kvasnicka, T., Bobcikova, P., Malikova, I., Hajkova, J., Zima, T., Ulrych, J., … others. (2015). The Frequencies of Ten Platelet Polymorphisms Associated with Atherosclerotic Cardiovascular Disease in Patients with Venous Thromboembolism: A Population-Based Case-Control Study. *Hereditary Genetics: Current Research*, *4*, 153. https://doi.org/doi:10.4172/2161- 1041.1000153

Langreth, B. R., & Waldholz, M. (1999). New Era of Personalized Medicine Targeting Drugs For Each Unique Genetic Profile. *The Oncologist*, *4*(5), 426–427.

Lanke, E., Kristoffersson, A.-C., Isaksson, C., Holmberg, L., & Lethagen, S. (2008). N1421K mutation in the glycoprotein Ib binding domain impairs ristocetin- and botrocetin-mediated binding of von Willebrand factor to platelets. *European Journal of Haematology*, *81*(5), 384–390. https://doi.org/10.1111/j.1600-0609.2008.01123.x

Lanza, F., Wolf, D., Fox, C. F., Kieffer, N., Seyer, J. M., Fried, V. A., … Jennings, L. K. (1991). cDNA cloning and expression of platelet p24/CD9. Evidence for a new family of multiple membrane-spanning proteins. *The Journal of Biological Chemistry*, *266*(16), 10638–10645.

Lawrence, R., Day-Williams, A. G., Mott, R., Broxholme, J., Cardon, L. R., & Zeggini, E. (2009). GLIDERS - A web-based search engine for genome-wide linkage disequilibrium between HapMap SNPs. *BMC Bioinformatics*, *10*, 367. https://doi.org/10.1186/1471-2105-10-367

Lawrence, S., Giles, C. L., & Tsoi, A. C. (1998). *What size neural network gives optimal generalization? Convergence properties of backpropagation* (Technical report). University of Maryland. Retrieved from http://drum.lib.umd.edu/handle/1903/809

Lehner, B. (2007). Modelling genotype–phenotype relationships and human disease with genetic interaction networks. *Journal of Experimental Biology*, *210*(9), 1559–1566. https://doi.org/10.1242/jeb.002311

Lewis, J. P., Ryan, K., O'Connell, J. R., Horenstein, R. B., Damcott, C. M., Gibson, Q., … Shuldiner, A. R. (2013). Genetic Variation in PEAR1 Is Associated With Platelet Aggregation and Cardiovascular Outcomes. *Circulation: Cardiovascular Genetics*, *6*(2), 184–192. https://doi.org/10.1161/CIRCGENETICS.111.964627

Li, B., & Leal, S. M. (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. https://doi.org/10.1016/j.ajhg.2008.06.024

Li, M. J., Wang, L. Y., Xia, Z., Sham, P. C., & Wang, J. (2013). GWAS3D: detecting human regulatory variants by integrative analysis of genome-wide associations, chromosome interactions and histone modifications. *Nucleic Acids Research*, gkt456. https://doi.org/10.1093/nar/gkt456

Li, M. J., Yan, B., Sham, P. C., & Wang, J. (2015). Exploring the function of genetic variants in the non-coding genomic regions: approaches for identifying human regulatory variants affecting gene expression. *Briefings in Bioinformatics*, *16*(3), 393–412. https://doi.org/10.1093/bib/bbu018

Li, Z., Delaney, M. K., O'Brien, K. A., & Du, X. (2010). Signaling during platelet adhesion and activation. *Arteriosclerosis, Thrombosis, and Vascular Biology*, *30*(12), 2341–2349. https://doi.org/10.1161/ATVBAHA.110.207522

Liao, L., Ning, G., Liu, C., Zhang, W., & Bao, M. (2013). The intron from the 5′-UTR of the FBP11 gene in petunia displays promoter- and enhancer-like functions. *Scientia Horticulturae*, *154*, 96–101. https://doi.org/10.1016/j.scienta.2013.02.009

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*(3), 18–22.

Lino Cardenas, C. L., Renault, N., Farce, A., Cauffiez, C., Allorge, D., Lo-Guidice, J.-M., … Chevalier, D. (2011). Genetic polymorphism of CYP4A11 and CYP4A22 genes and in silico insights from comparative 3D modelling in a French population. *Gene*, *487*(1), 10–20. https://doi.org/10.1016/j.gene.2011.07.015

Linsel-Nitschke, P., Götz, A., Erdmann, J., Braenne, I., Braund, P., Hengstenberg, C., … Cardiogenics Consortium. (2008). Lifelong reduction of LDL-cholesterol related to a common variant in the LDL-receptor gene decreases the risk of coronary artery disease-

-a Mendelian Randomisation study. *PloS One*, *3*(8), e2986. https://doi.org/10.1371/journal.pone.0002986

Listì, F., Candore, G., Lio, D., Cavallone, L., Colonna-Romano, G., Caruso, M., … Caruso, C. (2004). Association between platelet endothelial cellular adhesion molecule 1 (PECAM-1/CD31) polymorphisms and acute myocardial infarction: a study in patients from Sicily. *European Journal of Immunogenetics*, *31*(4), 175–178. https://doi.org/10.1111/j.1365-2370.2004.00464.x

Liu, D. J., & Leal, S. M. (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. https://doi.org/10.1371/journal.pgen.1001156

Liu, Y., Beyer, A., & Aebersold, R. (2016). On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell*, *165*(3), 535–550. https://doi.org/10.1016/j.cell.2016.03.014

Lood, C., Amisten, S., Gullstrand, B., Jönsen, A., Allhorn, M., Truedsson, L., … Bengtsson, A. A. (2010). Platelet transcriptional profile and protein expression in patients with systemic lupus erythematosus: up-regulation of the type I interferon system is strongly associated with vascular disease. *Blood*, *116*(11), 1951–1957. https://doi.org/10.1182/blood-2010-03-274605

Lozano, M. L., Cook, A., Bastida, J. M., Paul, D. S., Iruin, G., Cid, A. R., … Rivera, J. (2016). Novel mutations in RASGRP2, which encodes CalDAG-GEFI, abrogate Rap1 activation, causing platelet dysfunction. *Blood*, *128*(9), 1282–1289. https://doi.org/10.1182/blood-2015-11-683102

Lubeseder-Martellato, C., Guenzi, E., Jörg, A., Töpolt, K., Naschberger, E., Kremmer, E., … Stürzl, M. (2002). Guanylate-Binding Protein-1 Expression Is Selectively Induced by Inflammatory Cytokines and Is an Activation Marker of Endothelial Cells during Inflammatory Diseases. *The American Journal of Pathology*, *161*(5), 1749–1759.

Lumley, T. (2015). Package "leaps." Retrieved from https://cran.r-project.org/web/packages/leaps/leaps.pdf

Lynch, M., Walsh, B., & others. (1998). *Genetics and analysis of quantitative traits* (Vol. 1). Sinauer Sunderland, MA. Retrieved from http://www.invemar.org.co/redcostera1/invemar/docs/RinconLiterario/2011/febrero/AG_8.pdf

Ma, Y.-Q., Qin, J., & Plow, E. F. (2007). Platelet integrin alpha(IIb)beta(3): activation mechanisms. *Journal of Thrombosis and Haemostasis: JTH*, *5*(7), 1345–1352. https://doi.org/10.1111/j.1538-7836.2007.02537.x

Maghrabi, A. H. A., & McGuffin, L. J. (2017). ModFOLD6: an accurate web server for the global and local quality estimation of 3D protein models. *Nucleic Acids Research*, *45*, W416–W421. https://doi.org/10.1093/nar/gkx332

Maguire, J., Thakkinstian, A., Levi, C., Lincz, L., Bisset, L., Sturm, J., … Attia, J. (2011). Impact of COX-2 rs5275 and rs20417 and GPIIIa rs5918 polymorphisms on 90-day ischemic stroke functional outcome: a novel finding. *Journal of Stroke and*

Cerebrovascular Diseases: The Official Journal of National Stroke Association, *20*(2), 134–144. https://doi.org/10.1016/j.jstrokecerebrovasdis.2009.10.011

Makowsky, R., Pajewski, N. M., Klimentidis, Y. C., Vazquez, A. I., Duarte, C. W., Allison, D. B., & Campos, G. de los. (2011). Beyond Missing Heritability: Prediction of Complex Traits. *PLOS Genetics*, *7*(4), e1002051. https://doi.org/10.1371/journal.pgen.1002051

Makrythanasis, P., & Antonarakis, S. (2013). Pathogenic variants in non-protein-coding sequences. *Clinical Genetics*, *84*(5), 422–428. https://doi.org/10.1111/cge.12272

Malo, N., Libiger, O., & Schork, N. J. (2008). Accommodating Linkage Disequilibrium in Genetic-Association Analyses via Ridge Regression. *The American Journal of Human Genetics*, *82*(2), 375–385. https://doi.org/10.1016/j.ajhg.2007.10.012

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., … Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, *461*(7265), 747–753. https://doi.org/10.1038/nature08494

Marbach, D., Lamparter, D., Quon, G., Kellis, M., Kutalik, Z., & Bergmann, S. (2016). Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nature Methods*, *13*(4), 366–370. https://doi.org/10.1038/nmeth.3799

Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, *11*(7), 499–511. https://doi.org/10.1038/nrg2796

Masood, E. (1999). As consortium plans free SNP map of human genome. *Nature*, *398*(6728), 545–546. https://doi.org/10.1038/19126

Matsubara, Y., Murata, M., Sugita, K., & Ikeda, Y. (2003). Identification of a novel point mutation in platelet glycoprotein Ibalpha, Gly to Ser at residue 233, in a Japanese family with platelet-type von Willebrand disease. *Journal of Thrombosis and Haemostasis: JTH*, *1*(10), 2198–2205.

Matthews, K. W., Mueller-Ortiz, S. L., & Wetsel, R. A. (2004). Carboxypeptidase N: a pleiotropic regulator of inflammation. *Molecular Immunology*, *40*(11), 785–793.

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., … Stamatoyannopoulos, J. A. (2012). Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science*, *337*(6099), 1190–1195. https://doi.org/10.1126/science.1222794

Mazzucato, M., Cozzi, M. R., Pradella, P., Ruggeri, Z. M., & De Marco, L. (2004). Distinct roles of ADP receptors in von Willebrand factor-mediated platelet signaling and activation under high flow. *Blood*, *104*(10), 3221–3227. https://doi.org/10.1182/blood-2004-03-1145

Mbewu, A., & Mbanya, J.-C. (2006). Cardiovascular Disease. In D. T. Jamison, R. G. Feachem, M. W. Makgoba, E. R. Bos, F. K. Baingana, K. J. Hofman, & K. O. Rogo (Eds.), *Disease and Mortality in Sub-Saharan Africa* (2nd ed.). Washington (DC): World Bank. Retrieved from http://www.ncbi.nlm.nih.gov/books/NBK2294/

McCarroll, S. A., Kuruvilla, F. G., Korn, J. M., Cawley, S., Nemesh, J., Wysoker, A., … Altshuler, D. (2008). Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature Genetics*, *40*(10), 1166–1174. https://doi.org/10.1038/ng.238

McGuffin, L. J. (2008a). Intrinsic disorder prediction from the analysis of multiple protein fold recognition models. *Bioinformatics*, *24*(16), 1798–1804. https://doi.org/10.1093/bioinformatics/btn326

McGuffin, L. J. (2008b). Protein fold recognition and threading. *Computational Structural Biology, World Scientific*, 37–60.

McGuffin, L. J., Atkins, J. D., Salehe, B. R., Shuid, A. N., & Roche, D. B. (2015). IntFOLD: an integrated server for modelling protein structures and functions from amino acid sequences. *Nucleic Acids Research*, *43*(W1), W169–W173. https://doi.org/10.1093/nar/gkv236

McGuffin, L. J., Buenavista, M. T., & Roche, D. B. (2013). The ModFOLD4 server for the quality assessment of 3D protein models. *Nucleic Acids Research*, *41*(W1), W368–W372. https://doi.org/10.1093/nar/gkt294

McGuffin, L. J., & Jones, D. T. (2003). Improvement of the GenTHREADER method for genomic fold recognition. *Bioinformatics*, *19*(7), 874–881.

McGuffin, L. J., & Roche, D. B. (2011). Automated tertiary structure prediction with accurate local model quality assessment using the IntFOLD-TS method. *Proteins*, *79 Suppl 10*, 137–146. https://doi.org/10.1002/prot.23120

McRedmond, J. P., Park, S. D., Reilly, D. F., Coppinger, J. A., Maguire, P. B., Shields, D. C., & Fitzgerald, D. J. (2004). Integration of Proteomics and Genomics in Platelets A PROFILE OF PLATELET PROTEINS AND PLATELET-SPECIFIC GENES. *Molecular & Cellular Proteomics*, *3*(2), 133–144. https://doi.org/10.1074/mcp.M300063-MCP200

Meinders, M., Kulu, D. I., Werken, H. J. G. van de, Hoogenboezem, M., Janssen, H., Brouwer, R. W. W., … Philipsen, S. (2015). Sp1/Sp3 transcription factors regulate hallmarks of megakaryocyte maturation and platelet formation and function. *Blood*, *125*(12), 1957–1967. https://doi.org/10.1182/blood-2014-08-593343

Meisinger, C., Prokisch, H., Gieger, C., Soranzo, N., Mehta, D., Rosskopf, D., … Döring, A. (2009). A Genome-wide Association Study Identifies Three Loci Associated with Mean Platelet Volume. *The American Journal of Human Genetics*, *84*(1), 66–71. https://doi.org/10.1016/j.ajhg.2008.11.015

Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2001). Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics*, *157*(4), 1819–1829.

Meyniel, J.-P., Cottu, P. H., Decraene, C., Stern, M.-H., Couturier, J., Lebigot, I., … Sastre-Garau, X. (2010). A genomic and transcriptomic approach for a differential diagnosis between primary and secondary ovarian carcinomas in patients with a previous history of breast cancer. *BMC Cancer*, *10*(1), 222. https://doi.org/10.1186/1471-2407-10-222

Michelson, A. D. (2004). Platelet function testing in cardiovascular diseases. *Circulation*, *110*(19), e489–e493.

Mikeska, T., & Craig, J. M. (2014). DNA Methylation Biomarkers: Cancer and Beyond. *Genes*, *5*(3), 821–864. https://doi.org/10.3390/genes5030821

Miles, C., Wayne, M., & others. (2008). Quantitative trait locus (QTL) analysis. *Nature Education*, *1*(1), 208.

Miyamoto, S., Ogawa, H., Soejima, H., Takazoe, K., Sakamoto, T., Yoshimura, M., … Yasue, H. (2000). Formation of platelet aggregates after attacks of coronary spastic angina pectoris. *The American Journal of Cardiology*, *85*(4), 494–497. https://doi.org/10.1016/S0002-9149(99)00779-1

Moghaddasian, M., Arab, H., Dadkhah, E., Boostani, H., Babak, A. R., & Abbaszadegan, M. R. (2014). Protein modeling of cathepsin C mutations found in Papillon–Lefèvre syndrome. *Gene*, *538*(1), 182–187. https://doi.org/10.1016/j.gene.2013.11.079

Monteagudo, L. V., Ferrer, L. M., Catalan-Insa, E., Savva, D., McGuffin, L. J., & Tejedor, M. T. (2015). In silico identification and three-dimensional modelling of the missense mutation in ADAMTS2 in a sheep flock with dermatosparaxis. *Veterinary Dermatology*, *26*(1), 49-e16. https://doi.org/10.1111/vde.12178

Moore, J. H., Asselbergs, F. W., & Williams, S. M. (2010). Bioinformatics challenges for genome-wide association studies. *Bioinformatics*, *26*(4), 445–455.

Moore, J. H., & Williams, S. M. (2009). Epistasis and Its Implications for Personal Genetics. *American Journal of Human Genetics*, *85*(3), 309–320. https://doi.org/10.1016/j.ajhg.2009.08.006

Moore, L. D., Le, T., & Fan, G. (2013). DNA Methylation and Its Basic Function. *Neuropsychopharmacology*, *38*(1), 23–38. https://doi.org/10.1038/npp.2012.112

Mora, A., Sandve, G. K., Gabrielsen, O. S., & Eskeland, R. (2015). In the loop: promoter–enhancer interactions and bioinformatics. *Briefings in Bioinformatics*, bbv097. https://doi.org/10.1093/bib/bbv097

Morgenthaler, S., & Thilly, W. G. (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. https://doi.org/10.1016/j.mrfmmm.2006.09.003

Morris, A. P., & Zeggini, E. (2010). An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genetic Epidemiology*, *34*(2), 188–193. https://doi.org/10.1002/gepi.20450

Motyer, A. J., McKendry, C., Galbraith, S., & Wilson, S. R. (2011). LASSO model selection with post-processing for a genome-wide association study data set. *BMC Proceedings*, *5*(Suppl 9), S24. https://doi.org/10.1186/1753-6561-5-S9-S24

Moult, J. (2005). A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Current Opinion in Structural Biology*, *15*(3), 285–289. https://doi.org/10.1016/j.sbi.2005.05.011

Moult, J., Pedersen, J. T., Judson, R., & Fidelis, K. (1995). A large-scale experiment to assess protein structure prediction methods. *Proteins: Structure, Function, and Bioinformatics*, *23*(3), ii–iv. https://doi.org/10.1002/prot.340230303

Moyra, S. (2011). *Phenotypic Variation: Exploration and Functional Genomics*. Oxford University Press.

Musunuru, K., Strong, A., Frank-Kamenetsky, M., Lee, N. E., Ahfeldt, T., Sachs, K. V., … Rader, D. J. (2010). From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature*, *466*(7307), 714–719. https://doi.org/10.1038/nature09266

Nakamura, S., Kugiyama, K., Sugiyama, S., Miyamoto, S., Koide, S., Fukushima, H., … Ogawa, H. (2002). Polymorphism in the 5′-Flanking Region of Human Glutamate-Cysteine Ligase Modifier Subunit Gene Is Associated With Myocardial Infarction. *Circulation*, *105*(25), 2968–2973. https://doi.org/10.1161/01.CIR.0000019739.66514.1E

Nanda, N., Bao, M., Lin, H., Clauser, K., Komuves, L., Quertermous, T., … Hart, M. J. (2005). Platelet endothelial aggregation receptor 1 (PEAR1), a novel epidermal growth factor repeat-containing transmembrane receptor, participates in platelet contact-induced activation. *The Journal of Biological Chemistry*, *280*(26), 24680–24689. https://doi.org/10.1074/jbc.M413411200

Navarro-Núñez, L., Lozano, M. L., Martínez, C., Vicente, V., & Rivera, J. (2010). Effect of quercetin on platelet spreading on collagen and fibrinogen and on multiple platelet kinases. *Fitoterapia*, *81*(2), 75–80. https://doi.org/10.1016/j.fitote.2009.08.006

Ng, P. C., & Henikoff, S. (2003). SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Research*, *31*(13), 3812–3814. https://doi.org/10.1093/nar/gkg509

Nguyen, T.-T., Huang, J., Wu, Q., Nguyen, T., & Li, M. (2015). Genome-wide association data classification and SNPs selection using two-stage quality-based Random Forests. *BMC Genomics*, *16 Suppl 2*, S5. https://doi.org/10.1186/1471-2164-16-S2-S5

NHS England. (2015, September 24). NHS England » Gearing up for one of the most fundamental changes in NHS history. Retrieved April 30, 2016, from https://www.england.nhs.uk/2015/09/personalised-medicine/

Nicodemus, K. K., & Malley, J. D. (2009). Predictor correlation impacts machine learning algorithms: implications for genomic studies. *Bioinformatics*, *25*(15), 1884–1890. https://doi.org/10.1093/bioinformatics/btp331

Nicodemus, K. K., Malley, J. D., Strobl, C., & Ziegler, A. (2010). The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC Bioinformatics*, *11*(1), 110. https://doi.org/10.1186/1471-2105-11-110

Nieswandt, B., Moser, M., Pleines, I., Varga-Szabo, D., Monkley, S., Critchley, D., & Fässler, R. (2007). Loss of talin1 in platelets abrogates integrin activation, platelet aggregation, and thrombus formation in vitro and in vivo. *The Journal of Experimental Medicine*, *204*(13), 3113–3118. https://doi.org/10.1084/jem.20071827

Noé, L., Peeters, K., Izzi, B., Van Geet, C., & Freson, K. (2010). Regulators of platelet cAMP levels: clinical and therapeutic implications. *Current Medicinal Chemistry*, *17*(26), 2897–2905.

Nonyane, B. A., & Foulkes, A. S. (2008). Application of two machine learning algorithms to genetic association studies in the presence of covariates. *BMC Genetics*, *9*, 71. https://doi.org/10.1186/1471-2156-9-71

Novinska, M. S., Pietz, B. C., Ellis, T. M., Newman, D. K., & Newman, P. J. (2006). The alleles of PECAM-1. *Gene*, *376*(1), 95–101. https://doi.org/10.1016/j.gene.2006.02.016

Nugent, T., & Jones, D. T. (2009). Transmembrane protein topology prediction using support vector machines. *BMC Bioinformatics*, *10*, 159. https://doi.org/10.1186/1471-2105-10-159

Ober, U., Ayroles, J. F., Stone, E. A., Richards, S., Zhu, D., Gibbs, R. A., … Simianer, H. (2012). Using Whole-Genome Sequence Data to Predict Quantitative Trait Phenotypes in Drosophila melanogaster. *PLOS Genet*, *8*(5), e1002685. https://doi.org/10.1371/journal.pgen.1002685

Offermanns, S. (2006). Activation of Platelet Function Through G Protein–Coupled Receptors. *Circulation Research*, *99*(12), 1293–1304. https://doi.org/10.1161/01.RES.0000251742.71301.16

Offermanns, S., Laugwitz, K. L., Spicher, K., & Schultz, G. (1994). G proteins of the G12 family are activated via thromboxane A2 and thrombin receptors in human platelets. *Proceedings of the National Academy of Sciences of the United States of America*, *91*(2), 504–508.

Offermanns, S., Toombs, C. F., Hu, Y.-H., & Simon, M. I. (1997). Defective platelet activation in Gαq-deficient mice. *Nature*, *389*(6647), 183–186. https://doi.org/10.1038/38284

Oganesyan, V., Mazor, Y., Yang, C., Cook, K. E., Woods, R. M., Ferguson, A., … Dall'Acqua, W. F. (2015). Structural insights into the interaction of human IgG1 with FcγRI: no direct role of glycans in binding. *Acta Crystallographica Section D Biological Crystallography*, *71*(11), 2354–2361. https://doi.org/10.1107/S1399004715018015

Ohlmann, P., Laugwitz, K. L., Nürnberg, B., Spicher, K., Schultz, G., Cazenave, J. P., & Gachet, C. (1995). The human platelet ADP receptor activates Gi2 proteins. *The Biochemical Journal*, *312 (Pt 3)*, 775–779.

Ohnishi, Y., Tanaka, T., Yamada, R., Suematsu, K., Minami, M., Fujii, K., … Nakamura, Y. (2000). Identification of 187 single nucleotide polymorphisms (SNPs) among 41 candidate genes for ischemic heart disease in the Japanese population. *Human Genetics*, *106*(3), 288–292.

Okser, S., Pahikkala, T., Airola, A., Salakoski, T., Ripatti, S., & Aittokallio, T. (2014). Regularized Machine Learning in the Genetic Prediction of Complex Traits. *PLOS Genetics*, *10*(11), e1004754. https://doi.org/10.1371/journal.pgen.1004754

Okuda, T., Fujioka, Y., Kamide, K., Kawano, Y., Goto, Y., Yoshimasa, Y., … Miyata, T. (2002). Verification of 525 coding SNPs in 179 hypertension candidate genes in the

Japanese population: identification of 159 SNPs in 93 genes. *Journal of Human Genetics*, *47*(8), 387–394. https://doi.org/10.1007/s100380200056

Orengo, C. A., Jones, D. T., Thornton, J. M., & others. (1994). Protein superfamilies and domain superfolds. *Nature*, *372*(6507), 631–634.

Orlowski, A., Giusti, V. C. D., Morgan, P. E., Aiello, E. A., & Álvarez, B. V. (2012). Binding of carbonic anhydrase IX to extracellular loop 4 of the NBCe1 Na+/HCO3− cotransporter enhances NBCe1-mediated HCO3− influx in the rat heart. *American Journal of Physiology - Cell Physiology*, *303*(1), C69–C80. https://doi.org/10.1152/ajpcell.00431.2011

Ornella, L., Pérez, P., Tapia, E., González-Camacho, J. M., Burgueño, J., Zhang, X., … Crossa, J. (2014). Genomic-enabled prediction with classification algorithms. *Heredity*, *112*(6), 616–626. https://doi.org/10.1038/hdy.2013.144

Ouyang, H. (2014, October 30). Africa's Top Health Challenge: Cardiovascular Disease. *The Atlantic*. Retrieved from http://www.theatlantic.com/health/archive/2014/10/africas-top-health-challenge-cardiovascular-disease/381699/

Overby, C. L., & Tarczy-Hornoch, P. (2013). Personalized medicine: challenges and opportunities for translational bioinformatics. *Personalized Medicine*, *10*(5), 453–462.

Pal, L. R., & Moult, J. (2015). Genetic Basis of Common Human Disease: Insight into the Role of Missense SNPs from Genome-Wide Association Studies. *Journal of Molecular Biology*, *427*(13), 2271–2289. https://doi.org/10.1016/j.jmb.2015.04.014

Pamuk, O. N., Tozkir, H., Uyanik, M. S., Gurkan, H., Saritas, F., Duymaz, J., … Pamuk, G. E. (2014). PECAM-1 gene polymorphisms and soluble PECAM-1 level in rheumatoid arthritis and systemic lupus erythematosus patients: any link with clinical atherosclerotic events? *Clinical Rheumatology*, *33*(12), 1737–1743. https://doi.org/10.1007/s10067-014-2771-3

Pancione, M., Remo, A., & Colantuoni, V. (2012). Genetic and epigenetic events generate multiple pathways in colorectal cancer progression. *Pathology Research International*, *2012*, 509348. https://doi.org/10.1155/2012/509348

Parker, S. C. J., Stitzel, M. L., Taylor, D. L., Orozco, J. M., Erdos, M. R., Akiyama, J. A., … Young, A. (2013). Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proceedings of the National Academy of Sciences*, *110*(44), 17921–17926. https://doi.org/10.1073/pnas.1317023110

Parsa, A., Chang, Y.-P. C., Kelly, R. J., Corretti, M. C., Ryan, K. A., Robinson, S. W., … Liggett, S. B. (2011). Hypertrophy-Associated Polymorphisms Ascertained in a Founder Cohort Applied to Heart Failure Risk and Mortality. *Clinical and Translational Science*, *4*(1), 17–23. https://doi.org/10.1111/j.1752-8062.2010.00251.x

Pasomsub, E., Sukasem, C., Sungkanuparph, S., Kijsirikul, B., & Chantratita, W. (2010). The application of artificial neural networks for phenotypic drug resistance prediction: evaluation and comparison with other interpretation systems. *Japanese Journal of Infectious Diseases*, *63*(2), 87–94.

Pastorek, J., Pastoreková, S., Callebaut, I., Mornon, J., Zelník, V., Opavský, R., … Stanbridge, E. (1994). Cloning and characterization of MN, a human tumor-associated protein with a domain homologous to carbonic anhydrase and a putative helix-loop-helix DNA binding segment. *Oncogene*, *9*(10), 2877–2888.

Paz, I., Akerman, M., Dror, I., Kosti, I., & Mandel-Gutfreund, Y. (2010). SFmap: a web server for motif analysis and prediction of splicing factor binding sites. *Nucleic Acids Research*, gkq444. https://doi.org/10.1093/nar/gkq444

Paz, I., Kosti, I., Ares, M., Cline, M., & Mandel-Gutfreund, Y. (2014). RBPmap: a web server for mapping binding sites of RNA-binding proteins. *Nucleic Acids Research*, gku406. https://doi.org/10.1093/nar/gku406

Petersen, T. N., Brunak, S., von Heijne, G., & Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods*, *8*(10), 785–786. https://doi.org/10.1038/nmeth.1701

Peterson, T. A., Doughty, E., & Kann, M. G. (2013). Towards Precision Medicine: Advances in Computational Approaches for the Analysis of Human Variants. *Journal of Molecular Biology*, *425*(21), 4047–4063. https://doi.org/10.1016/j.jmb.2013.08.008

Petrich, B. G., Marchese, P., Ruggeri, Z. M., Spiess, S., Weichert, R. A. M., Ye, F., … Ginsberg, M. H. (2007). Talin is required for integrin-mediated platelet function in hemostasis and thrombosis. *The Journal of Experimental Medicine*, *204*(13), 3103–3111. https://doi.org/10.1084/jem.20071800

Pevsner, J. (2009). *Bioinformatics and functional genomics*. John Wiley & Sons.

Pleines, I., Eckly, A., Elvers, M., Hagedorn, I., Eliautou, S., Bender, M., … Nieswandt, B. (2010). Multiple alterations of platelet functions dominated by increased secretion in mice lacking Cdc42 in platelets. *Blood*, *115*(16), 3364–3373. https://doi.org/10.1182/blood-2009-09-242271

Portales-Casamar, E., Arenillas, D., Lim, J., Swanson, M. I., Jiang, S., McCallum, A., … Wasserman, W. W. (2009). The PAZAR database of gene regulatory information coupled to the ORCA toolkit for the study of regulatory sequences. *Nucleic Acids Research*, *37*(suppl_1), D54–D60. https://doi.org/10.1093/nar/gkn783

Porto, W. F., Franco, O. L., & Alencar, S. A. (2015). Computational analyses and prediction of guanylin deleterious SNPs. *Peptides*, *69*, 92–102. https://doi.org/10.1016/j.peptides.2015.04.013

Powell, J. E., Henders, A. K., McRae, A. F., Wright, M. J., Martin, N. G., Dermitzakis, E. T., … Visscher, P. M. (2012). Genetic control of gene expression in whole blood and lymphoblastoid cell lines is largely independent. *Genome Research*, *22*(3), 456–466. https://doi.org/10.1101/gr.126540.111

Powell, J. E., & Zietsch, B. P. (2011). Predicting Sensation Seeking From Dopamine Genes Use and Misuse of Genetic Prediction. *Psychological Science*, *22*(3), 413–415. https://doi.org/10.1177/0956797610397669

Precision Medicine Initiative. (2015). Precision Medicine Initiative. Retrieved February 4, 2016, from https://www.nih.gov/precision-medicine-initiative-cohort-program

President's Council of Advisors on Science Technology. (2008). Priorities for personalized medicine : report of the President's Council of Advisors on Science and Technology. :: Georgetown Law Library. Retrieved April 30, 2016, from http://cdm16064.contentdm.oclc.org/cdm/ref/collection/p266901coll4/id/1735

Prokunina, L., & Alarcón-Riquelme, M. E. (2004). Regulatory SNPs in complex diseases: their identification and functional validation. *Expert Reviews in Molecular Medicine*, *6*(10), 1–15. https://doi.org/10.1017/S1462399404007690

Provasi, D., Negri, A., Coller, B. S., & Filizola, M. (2014). Talin-driven inside-out activation mechanism of platelet αIIbβ3 integrin probed by multimicrosecond, all-atom molecular dynamics simulations. *Proteins*, *82*(12), 3231–3240. https://doi.org/10.1002/prot.24540

Pruissen, D. M. O., Kappelle, L. J., Rosendaal, F. R., Algra, A., & SMART Study Group. (2009). Prothrombotic genetic variants and atherosclerosis in patients with cerebral ischemia of arterial origin. *Atherosclerosis*, *204*(1), 191–195. https://doi.org/10.1016/j.atherosclerosis.2008.08.033

Purcell, S. M., Wray, N. R., Stone, J. L., Visscher, P. M., O'Donovan, M. C., Sullivan, P. F., … Moran, J. L. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, *460*(7256), 748–752. https://doi.org/10.1038/nature08185

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., … Sham, P. C. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

Pytela, R., Pierschbacher, M. D., Ginsberg, M. H., Plow, E. F., & Ruoslahti, E. (1986). Platelet membrane glycoprotein IIb/IIIa: member of a family of Arg-Gly-Asp–specific adhesion receptors. *Science*, *231*(4745), 1559–1562.

Raghavachari, N., Xu, X., Harris, A., Villagra, J., Logun, C., Barb, J., … Kato, G. (2007). Amplified expression profiling of platelet transcriptome reveals changes in arginine metabolic pathways in patients with sickle cell disease. *Circulation*, *115*(12), 1551–1562.

Rauch, U., Osende, J. I., Fuster, V., Badimon, J. J., Fayad, Z., & Chesebro, J. H. (2001). Thrombus formation on atherosclerotic plaques: pathogenesis and clinical consequences. *Annals of Internal Medicine*, *134*(3), 224–238.

Ray, D. M., Spinelli, S. L., O'Brien, J. J., Blumberg, N., & Phipps, R. P. (2006). Platelets as a novel target for PPARgamma ligands : implications for inflammation, diabetes, and cardiovascular disease. *BioDrugs: Clinical Immunotherapeutics, Biopharmaceuticals and Gene Therapy*, *20*(4), 231–241.

Ray, D. M., Spinelli, S. L., Pollock, S. J., Murant, T. I., O'Brien, J. J., Blumberg, N., … Phipps, R. P. (2008). Peroxisome proliferator-activated receptor gamma and retinoid X receptor

transcription factors are released from activated human platelets and shed in microparticles. *Thrombosis and Haemostasis*, *99*(1), 86–95. https://doi.org/10.1160/TH07-05-0328

Reardon, D. A., Desjardins, A., Vredenburgh, J. J., Gururangan, S., Sampson, J. H., Sathornsumetee, S., … Friedman, H. S. (2009). Metronomic chemotherapy with daily, oral etoposide plus bevacizumab for recurrent malignant glioma: a phase II study. *British Journal of Cancer*, *101*(12), 1986–1994. https://doi.org/10.1038/sj.bjc.6605412

Reddy, K. S., & Yusuf, S. (1998). Emerging Epidemic of Cardiovascular Disease in Developing Countries. *Circulation*, *97*(6), 596–601. https://doi.org/10.1161/01.CIR.97.6.596

Reif, D. M., Motsinger, A. A., McKinney, B. A., Crowe, J. E., & Moore, J. H. (2006). Feature selection using a random forests classifier for the integrated analysis of multiple data types. In *Computational Intelligence and Bioinformatics and Computational Biology, 2006. CIBCB'06. 2006 IEEE Symposium on* (pp. 1–8). IEEE. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4133169

Reva, B., Antipin, Y., & Sander, C. (2011). Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Research*, gkr407. https://doi.org/10.1093/nar/gkr407

Riancho, J. A. (2012). Genome-Wide Association Studies (GWAS) in Complex Diseases: Advantages and Limitations. *Reumatología Clínica (English Edition)*, *8*(2), 56–57. https://doi.org/10.1016/j.reumae.2011.07.007

Riddell, D. R., Graham, A., & Owen, J. S. (1997). Apolipoprotein E Inhibits Platelet Aggregation through the L-Arginine:Nitric Oxide Pathway IMPLICATIONS FOR VASCULAR DISEASE. *Journal of Biological Chemistry*, *272*(1), 89–95. https://doi.org/10.1074/jbc.272.1.89

Ridker, P. M., Paré, G., Parker, A. N., Zee, R. Y. L., Miletich, J. P., & Chasman, D. I. (2009). Polymorphism in the CETP gene region, HDL cholesterol, and risk of future myocardial infarction: Genomewide analysis among 18 245 initially healthy women from the Women's Genome Health Study. *Circulation. Cardiovascular Genetics*, *2*(1), 26–33. https://doi.org/10.1161/CIRCGENETICS.108.817304

Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A., & Kim, D. (2015). Methods of integrating data to uncover genotype-phenotype interactions. *Nature Reviews Genetics*, *16*(2), 85–97. https://doi.org/10.1038/nrg3868

Riva, A. (2012). Large-scale computational identification of regulatory SNPs with rSNP-MAPPER. *BMC Genomics*, *13*(4), S7. https://doi.org/10.1186/1471-2164-13-S4-S7

Riva, A., & Kohane, I. S. (2004). A SNP-centric database for the investigation of the human genome. *BMC Bioinformatics*, *5*(1), 33. https://doi.org/10.1186/1471-2105-5-33

Rivera, J., Lozano, M. L., Navarro-Núñez, L., & Vicente, V. (2009). Platelet receptors and signaling in the dynamics of thrombus formation. *Haematologica*, *94*(5), 700–711.

Robertson, N., Potter, C., & Harris, A. L. (2004). Role of carbonic anhydrase IX in human tumor cell growth, survival, and invasion. *Cancer Research*, *64*(17), 6160–6165. https://doi.org/10.1158/0008-5472.CAN-03-2224

Robinson, M. R., Wray, N. R., & Visscher, P. M. (2014). Explaining additional genetic variation in complex traits. *Trends in Genetics*, *30*(4), 124–132. https://doi.org/10.1016/j.tig.2014.02.003

Roche, D. B., Buenavista, M. T., & McGuffin, L. J. (2013). The FunFOLD2 server for the prediction of protein–ligand interactions. *Nucleic Acids Research*, *41*(W1), W303–W307. https://doi.org/10.1093/nar/gkt498

Rose, P. W., Bi, C., Bluhm, W. F., Christie, C. H., Dimitropoulos, D., Dutta, S., … Bourne, P. E. (2013). The RCSB Protein Data Bank: new resources for research and education. *Nucleic Acids Research*, *41*(D1), D475–D482. https://doi.org/10.1093/nar/gks1200

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Engineering*, *12*(2), 85–94. https://doi.org/10.1093/protein/12.2.85

Roznovăţ, I. A., & Ruskin, H. J. (2013). A computational model for genetic and epigenetic signals in colon cancer. *Interdisciplinary Sciences, Computational Life Sciences*, *5*(3), 175–186. https://doi.org/10.1007/s12539-013-0172-y

Rudock, M. E., Liu, Y., Ziegler, J. T., Allen, S. G., Lehtinen, A. B., Freedman, B. I., … Bowden, D. W. (2009). Association of polymorphisms in cyclooxygenase (COX)-2 with coronary and carotid calcium in the Diabetes Heart Study. *Atherosclerosis*, *203*(2), 459–465. https://doi.org/10.1016/j.atherosclerosis.2008.07.018

Ruggeri, Z. M. (1992). von Willebrand factor as a target for antithrombotic intervention. *Circulation*, *86*(6 Suppl), III26-29.

Ruggeri, Z. M. (2003). Von Willebrand factor, platelets and endothelial cell interactions. *Journal of Thrombosis and Haemostasis*, *1*(7), 1335–1342.

Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, *23*(19), 2507–2517.

Sahebkar, A., Morris, D. R., Biros, E., & Golledge, J. (2013). Association of single nucleotide polymorphisms in the gene encoding platelet endothelial cell adhesion molecule-1 with the risk of myocardial infarction: a systematic review and meta-analysis. *Thrombosis Research*, *132*(2), 227–233. https://doi.org/10.1016/j.thromres.2013.07.007

Salehe, B. R., Jones, C. I., Di Fatta, G., & McGuffin, L. J. (2017). RAPIDSNPs: A new computational pipeline for rapidly identifying key genetic variants reveals previously unidentified SNPs that are significantly associated with individual platelet responses. *PloS One*, *12*(4), e0175957. https://doi.org/10.1371/journal.pone.0175957

Sander, C., & Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins: Structure, Function, and Bioinformatics*, *9*(1), 56–68. https://doi.org/10.1002/prot.340090107

Savinkova, L. K., Ponomarenko, M. P., Ponomarenko, P. M., Drachkova, I. A., Lysova, M. V., Arshinova, T. V., & Kolchanov, N. A. (2009). TATA box polymorphisms in human gene promoters and associated hereditary pathologies. *Biochemistry. Biokhimiia*, *74*(2), 117–129.

Scheibe, R. J., Gros, G., Parkkila, S., Waheed, A., Grubb, J. H., Shah, G. N., … Wetzel, P. (2006). Expression of Membrane-bound Carbonic Anhydrases IV, IX, and XIV in the Mouse Heart. *Journal of Histochemistry & Cytochemistry*, *54*(12), 1379–1391. https://doi.org/10.1369/jhc.6A7003.2006

Schierding, W. S., Cutfield, W. S., & O'Sullivan, J. M. (2014). The missing story behind Genome Wide Association Studies: single nucleotide polymorphisms in gene deserts have a story to tell. *Epigenomics and Epigenetics*, *5*, 39. https://doi.org/10.3389/fgene.2014.00039

Schmugge, M., Rand, M. L., & Freedman, J. (2003). Platelets and von Willebrand factor. *Transfusion and Apheresis Science: Official Journal of the World Apheresis Association: Official Journal of the European Society for Haemapheresis*, *28*(3), 269–277. https://doi.org/10.1016/S1473-0502(03)00046-6

Schwarz, D. F., Szymczak, S., Ziegler, A., & König, I. R. (2007). Picking single-nucleotide polymorphisms in forests. *BMC Proceedings*, *1*(Suppl 1), S59.

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., … Wigler, M. (2004). Large-Scale Copy Number Polymorphism in the Human Genome. *Science*, *305*(5683), 525–528. https://doi.org/10.1126/science.1098918

Shastry, B. S. (2002). SNP alleles in human disease and evolution. *Journal of Human Genetics*, *47*(11), 561–566. https://doi.org/10.1007/s100380200086

Shen, B., Delaney, M. K., & Du, X. (2012). Inside-out, outside-in, and inside-outside-in: G protein signaling in integrin-mediated cell adhesion, spreading, and retraction. *Current Opinion in Cell Biology*, *24*(5), 600–606. https://doi.org/10.1016/j.ceb.2012.08.011

Sherry, S. T., Ward, M.-H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, *29*(1), 308–311. https://doi.org/10.1093/nar/29.1.308

Shi, G., Boerwinkle, E., Morrison, A. C., Gu, C. C., Chakravarti, A., & Rao, D. C. (2011). Mining gold dust under the genome wide significance level: a two-stage approach to analysis of GWAS. *Genetic Epidemiology*, *35*(2), 111–118. https://doi.org/10.1002/gepi.20556

Shi, Z., & Moult, J. (2011). Structural and Functional Impact of Cancer-Related Missense Somatic Mutations. *Journal of Molecular Biology*, *413*(2), 495–512. https://doi.org/10.1016/j.jmb.2011.06.046

Shi, Z., Sellers, J., & Moult, J. (2012). Protein stability and in vivo concentration of missense mutations in phenylalanine hydroxylase. *Proteins: Structure, Function, and Bioinformatics*, *80*(1), 61–70. https://doi.org/10.1002/prot.23159

Shihab, H. A., Gough, J., Cooper, D. N., Stenson, P. D., Barker, G. L. A., Edwards, K. J., … Gaunt, T. R. (2013). Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models. *Human Mutation*, *34*(1), 57–65. https://doi.org/10.1002/humu.22225

Shukla, S., & Mishra, R. (2011). Functional analysis of missense mutations G36A and G51A in PAX6, and PAX6(5a) causing ocular anomalies. *Experimental Eye Research*, *93*(1), 40–49. https://doi.org/10.1016/j.exer.2011.04.001

Sieberts, S. K., & Schadt, E. E. (2007). Moving toward a system genetics view of disease. *Mammalian Genome*, *18*(6–7), 389–401. https://doi.org/10.1007/s00335-007-9040-6

Siess, W., Cuatrecasas, P., & Lapetina, E. G. (1983a). A role for cyclooxygenase products in the formation of phosphatidic acid in stimulated human platelets. Differential mechanisms of action of thrombin and collagen. *Journal of Biological Chemistry*, *258*(8), 4683–4686.

Siess, W., Siegel, F. L., & Lapetina, E. G. (1983b). Arachidonic acid stimulates the formation of 1, 2-diacylglycerol and phosphatidic acid in human platelets. Degree of phospholipase C activation correlates with protein phosphorylation, platelet shape change, serotonin release, and aggregation. *Journal of Biological Chemistry*, *258*(18), 11236–11242.

Siess, W., & Tigyi, G. (2004). Thrombogenic and atherogenic activities of lysophosphatidic acid. *Journal of Cellular Biochemistry*, *92*(6), 1086–1094. https://doi.org/10.1002/jcb.20108

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., … Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, *7*(1), 539. https://doi.org/10.1038/msb.2011.75

Siljander, P. R.-M., Munnix, I. C., Smethurst, P. A., Deckmyn, H., Lindhout, T., Ouwehand, W. H., … Heemskerk, J. W. (2004). Platelet receptor interplay regulates collagen-induced thrombus formation in flowing human blood. *Blood*, *103*(4), 1333–1341.

Simon, T., Verstuyft, C., Mary-Krause, M., Quteineh, L., Drouet, E., Méneveau, N., … French Registry of Acute ST-Elevation and Non-ST-Elevation Myocardial Infarction (FAST-MI) Investigators. (2009). Genetic determinants of response to clopidogrel and cardiovascular events. *The New England Journal of Medicine*, *360*(4), 363–375. https://doi.org/10.1056/NEJMoa0808227

Sinha, S., & Yang, W. (2008). Cellular signaling for activation of Rho GTPase Cdc42. *Cellular Signalling*, *20*(11), 1927–1934. https://doi.org/10.1016/j.cellsig.2008.05.002

Snyderman R, & Dinan MA. (2010). IMproving health by taking it personally. *JAMA*, *303*(4), 363–364. https://doi.org/10.1001/jama.2010.34

Söding, J. (2005). Protein homology detection by HMM–HMM comparison. *Bioinformatics*, *21*(7), 951–960. https://doi.org/10.1093/bioinformatics/bti125

Sondermann, P., Huber, R., & Jacob, U. (1999). Crystal structure of the soluble form of the human fcgamma-receptor IIb: a new member of the immunoglobulin superfamily at 1.7 A resolution. *The EMBO Journal*, *18*(5), 1095–1103. https://doi.org/10.1093/emboj/18.5.1095

Soranzo, N., Rendon, A., Gieger, C., Jones, C. I., Watkins, N. A., Menzel, S., … others. (2009). A novel variant on chromosome 7q22. 3 associated with mean platelet volume, counts, and function. *Blood*, *113*(16), 3831–3837.

Stankiewicz, P., & Lupski, J. R. (2010). Structural Variation in the Human Genome and its Role in Disease. *Annual Review of Medicine*, *61*(1), 437–455. https://doi.org/10.1146/annurev-med-100708-204735

Stanner, S. (2008). *Cardiovascular Disease: Diet, Nutrition and Emerging Risk Factors (The Report of the British Nutrition Foundation Task Force)*. Wiley. com.

Stepanova, M., Tiazhelova, T., Skoblov, M., & Baranova, A. (2006). Potential regulatory SNPs in promoters of human genes: A systematic approach. *Molecular and Cellular Probes*, *20*(6), 348–358. https://doi.org/10.1016/j.mcp.2006.03.007

Stephan, J., Stegle, O., & Beyer, A. (2015). A random forest approach to capture genetic effects in the presence of population structure. *Nature Communications*, *6*, 7432. https://doi.org/10.1038/ncomms8432

Stormo, G. D., & Fields, D. S. (1998). Specificity, free energy and information content in protein–DNA interactions. *Trends in Biochemical Sciences*, *23*(3), 109–113. https://doi.org/10.1016/S0968-0004(98)01187-6

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*(1), 307. https://doi.org/10.1186/1471-2105-9-307

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, *8*(1), 25. https://doi.org/10.1186/1471-2105-8-25

Strobl, C., & Zeileis, A. (2008). Danger: High power! Exploring the statistical properties of a test for random forest variable importance. In *18th International Conference on Computational Statistics*. Porto, Portugal.

Sun, J., Williams, J., Yan, H. C., Amin, K. M., Albelda, S. M., & DeLisser, H. M. (1996). Platelet endothelial cell adhesion molecule-1 (PECAM-1) homophilic adhesion is mediated by immunoglobulin-like domains 1 and 2 and depends on the cytoplasmic domain and the level of surface expression. *The Journal of Biological Chemistry*, *271*(31), 18561–18570.

Sun, L., Gorospe, J. R., Hoffman, E. P., & Rao, A. K. (2007). Decreased platelet expression of myosin regulatory light chain polypeptide (MYL9) and other genes with platelet dysfunction and CBFA2/RUNX1 mutation: insights from platelet expression profiling. *Journal of Thrombosis and Haemostasis: JTH*, *5*(1), 146–154. https://doi.org/10.1111/j.1538-7836.2006.02271.x

Supuran, C. T. (2008). Carbonic anhydrases: novel therapeutic applications for inhibitors and activators. *Nature Reviews Drug Discovery*, *7*(2), 168–181. https://doi.org/10.1038/nrd2467

Svastová, E., Zilka, N., Zat'ovicová, M., Gibadulinová, A., Ciampor, F., Pastorek, J., & Pastoreková, S. (2003). Carbonic anhydrase IX reduces E-cadherin-mediated adhesion of MDCK cells via interaction with beta-catenin. *Experimental Cell Research*, *290*(2), 332–345.

Swietach, P., Hulikova, A., Vaughan-Jones, R. D., & Harris, A. L. (2010). New insights into the physiological role of carbonic anhydrase IX in tumour pH regulation. *Oncogene*, *29*(50), 6509–6521. https://doi.org/10.1038/onc.2010.455

Szymczak, S., Biernacka, J. M., Cordell, H. J., González-Recio, O., König, I. R., Zhang, H., & Sun, Y. V. (2009). Machine learning in genome-wide association studies. *Genetic Epidemiology*, *33*(S1), S51–S57. https://doi.org/10.1002/gepi.20473

Taylor, W. R., Munro, R. E. J., Petersen, K., & Bywater, R. P. (2003). Ab initio modelling of the N-terminal domain of the secretin receptors. *Computational Biology and Chemistry*, *27*(2), 103–114. https://doi.org/10.1016/S1476-9271(03)00020-3

Teichmann, S. A., & Chothia, C. (2000). Immunoglobulin superfamily proteins in Caenorhabditis elegans1. *Journal of Molecular Biology*, *296*(5), 1367–1383. https://doi.org/10.1006/jmbi.1999.3497

Teng, S., Michonova-Alexova, E., & Alexov, E. (2008). Approaches and resources for prediction of the effects of non-synonymous single nucleotide polymorphism on protein function and interactions. *Current Pharmaceutical Biotechnology*, *9*(2), 123–133.

Thiry, A., Dogné, J.-M., Masereel, B., & Supuran, C. T. (2006). Targeting tumor-associated carbonic anhydrase IX in cancer therapy. *Trends in Pharmacological Sciences*, *27*(11), 566–573. https://doi.org/10.1016/j.tips.2006.09.002

Thomas-Chollier, M., Hufton, A., Heinig, M., O'Keeffe, S., Masri, N. E., Roider, H. G., … Vingron, M. (2011). Transcription factor binding predictions using TRAP for the analysis of ChIP-seq data and regulatory SNPs. *Nature Protocols*, *6*(12), 1860–1869. https://doi.org/10.1038/nprot.2011.409

Tibbles, H. E., Vassilev, A., Wendorf, H., Schonhoff, D., Zhu, D., Lorenz, D., … Uckun, F. M. (2001). Role of a JAK3-dependent Biochemical Signaling Pathway in Platelet Activation and Aggregation. *Journal of Biological Chemistry*, *276*(21), 17815–17822. https://doi.org/10.1074/jbc.M011405200

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Torrent, M., Rickert, K., Pan, B.-S., & Sepp-Lorenzino, L. (2004). Analysis of the activating mutations within the activation loop of leukemia targets Flt-3 and c-Kit based on protein homology modeling. *Journal of Molecular Graphics and Modelling*, *23*(2), 153–165. https://doi.org/10.1016/j.jmgm.2004.05.002

Touw, W. G., Bayjanov, J. R., Overmars, L., Backus, L., Boekhorst, J., Wels, M., & Hijum, S. A. F. T. van. (2013). Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? *Briefings in Bioinformatics*, *14*(3), 315–326. https://doi.org/10.1093/bib/bbs034

Türkmen, S., Guo, G., Garshasbi, M., Hoffmann, K., Alshalah, A. J., Mischung, C., … Robinson, P. N. (2009). CA8 Mutations Cause a Novel Syndrome Characterized by Ataxia and Mild Mental Retardation with Predisposition to Quadrupedal Gait. *PLoS Genet*, *5*(5), e1000487. https://doi.org/10.1371/journal.pgen.1000487

Ullah, A. Z. D., Lemoine, N. R., & Chelala, C. (2012). SNPnexus: a web server for functional annotation of novel and publicly known genetic variants (2012 update). *Nucleic Acids Research*, gks364. https://doi.org/10.1093/nar/gks364

Valencia, A., & Pazos, F. (2003). Prediction of protein-protein interactions from evolutionary information. In *Structural Bioinformatics* (pp. 411–426). Wiley.

van Steensel, B. (2011). Chromatin: constructing the big picture. *The EMBO Journal*, *30*(10), 1885–1895. https://doi.org/10.1038/emboj.2011.135

Vane, J. R., Mitchell, J. A., Appleton, I., Tomlinson, A., Bishop-Bailey, D., Croxtall, J., & Willoughby, D. A. (1994). Inducible isoforms of cyclooxygenase and nitric-oxide synthase in inflammation. *Proceedings of the National Academy of Sciences of the United States of America*, *91*(6), 2046–2050.

Varga-Szabo, D., Braun, A., & Nieswandt, B. (2009). Calcium signaling in platelets. *Journal of Thrombosis and Haemostasis*, *7*(7), 1057–1066. https://doi.org/10.1111/j.1538-7836.2009.03455.x

Vélez, P., & García, Á. (2015). Platelet proteomics in cardiovascular diseases. *Translational Proteomics*, *7*, 15–29. https://doi.org/10.1016/j.trprot.2014.09.002

Verikas, A., Gelzinis, A., & Bacauskiene, M. (2011). Mining data with random forests: A survey and results of new tests. *Pattern Recognition*, *44*(2), 330–349. https://doi.org/10.1016/j.patcog.2010.08.011

Viles-Gonzalez, J. F., Fuster, V., & Badimon, J. J. (2004). Atherothrombosis: A widespread disease with unpredictable and life-threatening consequences. *European Heart Journal*, *25*(14), 1197–1207. https://doi.org/10.1016/j.ehj.2004.03.011

Vizioli, L., Muscari, S., & Muscari, A. (2009). The relationship of mean platelet volume with the risk and prognosis of cardiovascular diseases. *International Journal of Clinical Practice*, *63*(10), 1509–1515.

Wan, Y., Qu, K., Zhang, Q. C., Flynn, R. A., Manor, O., Ouyang, Z., … Chang, H. Y. (2014). Landscape and variation of RNA secondary structure across the human transcriptome. *Nature*, *505*(7485), 706–709. https://doi.org/10.1038/nature12946

Wang, D., & Larder, B. (2003). Enhanced prediction of lopinavir resistance from genotype by use of artificial neural networks. *The Journal of Infectious Diseases*, *188*(5), 653–660. https://doi.org/10.1086/377453

Wang, X., Tomso, D. J., Liu, X., & Bell, D. A. (2005). Single nucleotide polymorphism in transcriptional regulatory regions and expression of environmentally responsive genes. *Toxicology and Applied Pharmacology*, *207*(2 Suppl), 84–90. https://doi.org/10.1016/j.taap.2004.09.024

Wang, Z., & Moult, J. (2001). SNPs, protein structure, and disease. *Human Mutation*, *17*(4), 263–270.

Ward, J. J., McGuffin, L. J., Bryson, K., Buxton, B. F., & Jones, D. T. (2004). The DISOPRED server for the prediction of protein disorder. *Bioinformatics*, *20*(13), 2138–2139.

Ward, L. D., & Kellis, M. (2012a). HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Research*, *40*(D1), D930–D934. https://doi.org/10.1093/nar/gkr917

Ward, L. D., & Kellis, M. (2012b). Interpreting noncoding genetic variation in complex traits and human disease. *Nature Biotechnology*, *30*(11), 1095–1106. https://doi.org/10.1038/nbt.2422

Weischenfeldt, J., Symmons, O., Spitz, F., & Korbel, J. O. (2013). Phenotypic impact of genomic structural variation: insights from and for human disease. *Nature Reviews Genetics*, *14*(2), 125–138. https://doi.org/10.1038/nrg3373

Werthmann, R. C., von Hayn, K., Nikolaev, V. O., Lohse, M. J., & Bünemann, M. (2009). Real-time monitoring of cAMP levels in living endothelial cells: thrombin transiently inhibits adenylyl cyclase 6. *The Journal of Physiology*, *587*(Pt 16), 4091–4104. https://doi.org/10.1113/jphysiol.2009.172957

Weyrich, A. S., Schwertz, H., Kraiss, L. W., & Zimmerman, G. A. (2009). Protein synthesis by platelets: historical and new perspectives. *Journal of Thrombosis and Haemostasis: JTH*, *7*(2), 241–246. https://doi.org/10.1111/j.1538-7836.2008.03211.x

WHO | Cardiovascular diseases (CVDs). (2016). Retrieved June 9, 2016, from http://www.who.int/mediacentre/factsheets/fs317/en/

WHO | Department of Measurement and Health Information. (2006). Projections of mortality and burden of disease by region. Retrieved from http://www.who.int/healthinfo/statistics/bod_deathbyregion.xls

WHO | Raised blood pressure. (2016). Retrieved June 9, 2016, from http://www.who.int/gho/ncd/risk_factors/blood_pressure_prevalence_text/en/

Wilhelm, M., Schlegl, J., Hahne, H., Moghaddas Gholami, A., Lieberenz, M., Savitski, M. M., … Kuster, B. (2014). Mass-spectrometry-based draft of the human proteome. *Nature*, *509*(7502), 582–587. https://doi.org/10.1038/nature13319

Williams, A. F., & Barclay, A. N. (1988). The Immunoglobulin Superfamily—Domains for Cell Surface Recognition. *Annual Review of Immunology*, *6*(1), 381–405. https://doi.org/10.1146/annurev.iy.06.040188.002121

Williams, M. S., Weiss, E. J., Sabatine, M. S., Simon, D. I., Bahou, W. F., Becker, L. C., … 2010 Platelet Colloquium Participants. (2010). Genetic regulation of platelet receptor

expression and function: application in clinical practice and drug development. *Arteriosclerosis, Thrombosis, and Vascular Biology*, *30*(12), 2372–2384. https://doi.org/10.1161/ATVBAHA.110.218131

Wilson, C. A., Kreychman, J., & Gerstein, M. (2000). Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores1. *Journal of Molecular Biology*, *297*(1), 233–249. https://doi.org/10.1006/jmbi.2000.3550

Wittkopp, P. J., & Kalay, G. (2012). Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nature Reviews Genetics*, *13*(1), 59–69. https://doi.org/10.1038/nrg3095

Wollscheid, B., Bausch-Fluck, D., Henderson, C., O'Brien, R., Bibel, M., Schiess, R., … Watts, J. D. (2009). Mass-spectrometric identification and relative quantification of N-linked cell surface glycoproteins. *Nature Biotechnology*, *27*(4), 378–386. https://doi.org/10.1038/nbt.1532

Wong, K. K., deLeeuw, R. J., Dosanjh, N. S., Kimm, L. R., Cheng, Z., Horsman, D. E., … Lam, W. L. (2007). A Comprehensive Analysis of Common Copy-Number Variations in the Human Genome. *The American Journal of Human Genetics*, *80*(1), 91–104. https://doi.org/10.1086/510560

Woodman, R., Brown, C., & Lockette, W. (2010). Chlorthalidone decreases platelet aggregation and vascular permeability and promotes angiogenesis. *Hypertension*, *56*(3), 463–470. https://doi.org/10.1161/HYPERTENSIONAHA.110.154476

Worthington, R. E., Carroll, R. C., & Boucheix, C. (1990). Platelet activation by CD9 monoclonal antibodies is mediated by the FCγII receptor. *British Journal of Haematology*, *74*(2), 216–222. https://doi.org/10.1111/j.1365-2141.1990.tb02568.x

Woulfe, D., Yang, J., & Brass, L. (2001). ADP and platelets: the end of the beginning. *Journal of Clinical Investigation*, *107*(12), 1503–1505. https://doi.org/10.1172/JCI13361

Wray, N. R., Yang, J., Hayes, B. J., Price, A. L., Goddard, M. E., & Visscher, P. M. (2013). Pitfalls of predicting complex traits from SNPs. *Nature Reviews. Genetics*, *14*(7), 507–515. https://doi.org/10.1038/nrg3457

Wu, G., Xi, Y., Yao, L., Su, L., Yan, Y., Li, M., & Gu, L. (2014). Genetic polymorphism of ITGA2 C807T can increase the risk of ischemic stroke. *The International Journal of Neuroscience*, *124*(11), 841–851. https://doi.org/10.3109/00207454.2013.879718

Wu, J., Devlin, B., Ringquist, S., Trucco, M., & Roeder, K. (2010). *Genetic Epidemiology*, *34*(3), 275–285. https://doi.org/10.1002/gepi.20459

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., & Lin, X. (2011). Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. *American Journal of Human Genetics*, *89*(1), 82–93. https://doi.org/10.1016/j.ajhg.2011.05.029

Wu, Q., Ye, Y., Liu, Y., & Ng, M. K. (2012). SNP selection and classification of genome-wide SNP data using stratified sampling random forests. *IEEE Transactions on Nanobioscience*, *11*(3), 216–227. https://doi.org/10.1109/TNB.2012.2214232

Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E., & Lange, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, *25*(6), 714–721. https://doi.org/10.1093/bioinformatics/btp041

Wu, Y.-P., Vink, T., Schiphorst, M., van Zanten, G. H., IJsseldijk, M. J., de Groot, P. G., & Sixma, J. J. (2000). Platelet thrombus formation on collagen at high shear rates is mediated by von Willebrand factor–glycoprotein Ib interaction and inhibited by von Willebrand factor–glycoprotein IIb/IIIa interaction. *Arteriosclerosis, Thrombosis, and Vascular Biology*, *20*(6), 1661–1667.

Würtz, M., Nissen, P. H., Grove, E. L., Kristensen, S. D., & Hvas, A.-M. (2014). Genetic Determinants of On-Aspirin Platelet Reactivity: Focus on the Influence of PEAR1. *PLoS ONE*, *9*(10). https://doi.org/10.1371/journal.pone.0111816

Xu, D., & Zhang, Y. (2012). Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins: Structure, Function, and Bioinformatics*, *80*(7), 1715–1735. https://doi.org/10.1002/prot.24065

Yáñez, M., Gil-Longo, J., & Campos-Toimil, M. (2012). Calcium binding proteins. *Advances in Experimental Medicine and Biology*, *740*, 461–482. https://doi.org/10.1007/978-94-007-2888-2_19

Yang, Y., Faraggi, E., Zhao, H., & Zhou, Y. (2011). Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. *Bioinformatics*, *27*(15), 2076–2082. https://doi.org/10.1093/bioinformatics/btr350

Yates, C. M., Filippis, I., Kelley, L. A., & Sternberg, M. J. E. (2014). SuSPect: Enhanced Prediction of Single Amino Acid Variant (SAV) Phenotype Using Network Features. *Journal of Molecular Biology*, *426*(14), 2692–2701. https://doi.org/10.1016/j.jmb.2014.04.026

Ye, H., Li, X., Wang, L., Liao, Q., Xu, L., Huang, Y., … Duan, S. (2013). Genetic associations with coronary heart disease: Meta-analyses of 12 candidate genetic variants. *Gene*, *531*(1), 71–77. https://doi.org/10.1016/j.gene.2013.07.029

Yu, S., Yoon, J., Lee, J., Myung, S., Jang, E., Kwak, M., … Lee, H. (2011). Inhibition of hypoxia-inducible carbonic anhydrase-IX enhances hexokinase II inhibitor-induced hepatocellular carcinoma cell apoptosis. *Acta Pharmacologica Sinica*, *32*(7), 912–920. https://doi.org/10.1038/aps.2011.24

Yue, P., & Moult, J. (2006). Identification and Analysis of Deleterious Human SNPs. *Journal of Molecular Biology*, *356*(5), 1263–1274. https://doi.org/10.1016/j.jmb.2005.12.025

Závada, J., Závadová, Z., Pastorek, J., Biesová, Z., Ježek, J., & Velek, J. (2000). Human tumour-associated cell adhesion protein MN/CA IX: identification of M75 epitope and of the region mediating cell adhesion. *British Journal of Cancer*, *82*(11), 1808–1813. https://doi.org/10.1054/bjoc.2000.1111

Zee, R. Y. L., Michaud, S. E., Diehl, K. A., Chasman, D. I., Emmerich, J., Gaussem, P., … Ridker, P. M. (2008). Purinergic receptor P2Y, G-protein coupled, 12 gene variants and

risk of incident ischemic stroke, myocardial infarction, and venous thromboembolism. *Atherosclerosis*, *197*(2), 694–699. https://doi.org/10.1016/j.atherosclerosis.2007.07.001

Zhang, F. L., Luo, L., Gustafson, E., Lachowicz, J., Smith, M., Qiao, X., … Monsma, F. J. (2001). ADP Is the Cognate Ligand for the Orphan G Protein-coupled Receptor SP1999. *Journal of Biological Chemistry*, *276*(11), 8608–8615. https://doi.org/10.1074/jbc.M009718200

Zhang, X.-S., Wang, R.-S., Wu, L.-Y., & Chen, L. (2006). Models and algorithms for haplotyping problem. *Current Bioinformatics*, *1*(1), 105–114.

Zhang, Y. (2008). I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics*, *9*, 40. https://doi.org/10.1186/1471-2105-9-40

Zhang, Y., & Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, *57*(4), 702–710. https://doi.org/10.1002/prot.20264

Zhang, Y., & Skolnick, J. (2005a). The protein structure prediction problem could be solved using the current PDB library. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(4), 1029–1034. https://doi.org/10.1073/pnas.0407152101

Zhang, Y., & Skolnick, J. (2005b). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, *33*(7), 2302–2309. https://doi.org/10.1093/nar/gki524

Zhao, K., Lu, Z., Park, J. W., Zhou, Q., & Xing, Y. (2013). GLiMMPS: robust statistical model for regulatory variation of alternative splicing using RNA-seq data. *Genome Biology*, *14*, R74. https://doi.org/10.1186/gb-2013-14-7-r74

Zhu, X. (2014). Comparison of four methods for handing missing data in longitudinal data analysis through a simulation study. *Open Journal of Statistics*, *4*(11), 933.

Ziegler, S., Schillinger, M., Funk, M., Felber, K., Exner, M., Mlekusch, W., … Mannhalter, C. (2005). Association of a functional polymorphism in the clopidogrel target receptor gene, P2Y12, and the risk for ischemic cerebrovascular events in patients with peripheral artery disease. *Stroke*, *36*(7), 1394–1399. https://doi.org/10.1161/01.STR.0000169922.79281.a5

Zimman, A., Titz, B., Komisopoulou, E., Biswas, S., Graeber, T. G., & Podrez, E. A. (2014). Phosphoproteomic Analysis of Platelets Activated by Pro-Thrombotic Oxidized Phospholipids and Thrombin. *PLOS ONE*, *9*(1), e84488. https://doi.org/10.1371/journal.pone.0084488

Zou, S., Teixeira, A. M., Sanada, C. D., Zhang, P., & Krause, D. (2014). ARHGEF12 Is Essential for Human Megakaryocyte Differentiation and Plays Critical Roles in Platelet Function. *Blood*, *124*(21), 341–341.

APPENDIX 1 – PUBLICATIONS ARISING OR RELATED TO THIS PROJECT

1. McGuffin, L.J., Atkins, J., **Salehe, B.R**., Shuid, A.N. & Roche, D.B. (2015) IntFOLD: an integrated server for modelling protein structures and functions from amino acid sequences. *Nucleic Acids Research*, 43, W169-73.
2. **Salehe BR**, Jones CI, Di Fatta G, McGuffin LJ (2017) RAPIDSNPs: A new computational pipeline for rapidly identifying key genetic variants reveals previously unidentified SNPs that are significantly associated with individual platelet responses. *PLoS ONE* 12(4): e0175957. https://doi.org/10.1371/journal.pone.0175957
3. McGuffin LJ, Shuid AN, Kempster R, Maghrabi AH, Nealon JO, **Salehe BR**, Atkins JD, Roche DB (2017) Accurate Template Based Modelling in CASP12 using the IntFOLD4-TS, ModFOLD6 and ReFOLD methods. *Proteins: Structure, Function, and Bioinformatics.* doi:10.1002/prot.25360

APPENDIX 2 – ACCEPTED ABSTRACTS FOR CONFERENCE POSTERS AND PROCEEDINGS

Salehe, B.R, Jones, C.I, Di Fatta, G, & McGuffin, L.J. (2014). Identification of novel genes associated with platelet activation signalling pathways in high dimensional data using an alternative regression approach. *Heart 2014;**100**:A2 doi:10.1136/heartjnl-2014-306916.5* **Proceeding: British Cardiovascular Research Society (BCRS) – Reading, UK**

Salehe, B.R, Jones, C.I, Di Fatta, G, & McGuffin, L.J. (2015). Genetic association analysis using an alternative approach involving random forest and shrinkage methods applied to the study of platelet activation pathways. **International Society of Computational Biology (ISCB) Africa ASBCB Conference on Bioinformatics, Dar es salaam, Tanzania.**

Salehe, B.R, Jones, C.I, Di Fatta, G, & McGuffin, L.J. (2015). An integrated computational approach for analysing genetic, molecular, and functional variations for understanding diseases and complex traits. **Proceeding: International Society of Computational Biology (ISCB) Reginal Student Groups (RSG) - UK – Norwich, UK**.

Salehe, B.R, Jones, C.I, Di Fatta, G, & McGuffin, L.J. (2015). An integrated computational pipeline for analysing genetic, molecular, and functional variations in complex diseases. **International Society of Computational Biology (ISCB/RECOMB) Conference, Philadelphia, USA.**