An Integrated Genomic Approach for the Identification and Analysis of Single Nucleotide Polymorphisms that Affect Cancer in Humans

Emmanouela Repapi

A thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

Trinity Term 2013

Queen's College

Ludwig Institute for Cancer Research

Nuffield Department of Clinical Medicine

University of Oxford

An Integrated Genomic Approach for the Identification and Analysis of Single Nucleotide Polymorphisms that Affect Cancer in Humans

Emmanouela Repapi, LICR, NDM, University of Oxford, Queen's College - D.Phil thesis, Trinity term 2013

Abstract

The identification of genetic variants such as single nucleotide polymorphisms (SNPs), which affect cancer progression, survival and response to treatments could help in the design of better prevention and treatment strategies. Genome-wide association studies (GWAS) have provided the first step of identifying SNPs associating with cancer risk. However, identifying the causal SNPs responsible for the associations has proven challenging, and GWAS have not been successful for time-to-event phenotypes such as cancer progression, due to the insurmountable obstacle of the large sample size needed. The aim of this thesis is to design and implement strategies that combine the identification of SNPs significantly associated with cancer, focusing on time-to-event phenotypes, with detailed bioinformatics analysis to allow for further experimental validation and modelling, to better understand cancer-associated genomic loci and accelerate their incorporation into the clinic.

First, a methodology that utilises the Random Survival Forest is developed and combined with a bioinformatics analysis that ranks SNPs according to their potential to result in differential protein levels or activity, in order to identify SNPs that affect the progression of B-cell chronic lymphocytic leukaemia. Next, an analysis that aims to extend our understanding of the role of SNPs in mediating the cellular responses to chemotherapeutic agents is applied. SNPs that could associate with differential cellular growth responses in cancer cell line panels are identified, and their association with the differential survival of cancer patients is explored. Finally, the potential roles of SNPs in affecting the transcriptional regulation of key cancer genes resulting in differential cancer risk are assessed. First, by focusing on SNPs in an important transcription factor binding motif that has been shown to be extremely sensitive to single base pair changes (the E-box) and next, by exploring the possibility that polymorphic transcription factor binding sites could underlie the significant associations noted in cancer GWAS.

Acknowledgements

I would first like to thank my supervisors Dr. Gareth Bond and Prof. Nicolai Meinshausen for their mentoring, guidance and support throughout my DPhil. It is needless to say that without their invaluable comments, this thesis would have never been possible. A big thank you also goes to Dr. Christopher Yau for all his help during my corrections. I also want to thank the Ludwig Institute for Cancer Research with all its members, for their funding and professional support. In addition, I would like to thank all my collaborators mentioned in the thesis for their contribution, Claire and Mark for helping out with editing this thesis and the other members of the group, Jorge, Anna, Alexander, Juliet and Elisabeth, for our insightful discussions.

Special thanks also to all my friends, both from the institute, as well as outside. Especially to Androniki, Melissa, Chrisanthi, Jorge, Anna and Alexander for their support in all matters, both professional and personal, and for making Oxford a fun place to be, and to Christoforos, the person that I can always count on to find me a solution for every problem. Finally, a huge thanks is owed to Olly and my family. Olly, thank you for being there for me in both the good and bad times in these four years. You are one of the main reasons I came back from Leicester to start this PhD and most importantly the one who made sure I finished it, by picking me up from every fall. Most of all, I am grateful to my family for their endless love and support and for making sure I always had everything I needed to make it this far. Mum and dad for believing in me and supporting me no matter what, my big brother Costis for always looking out for me and my godparents Kristy and George for all their love.

Declaration

The experimental validation described in Chapters 5.2.2, 6.2.2, 7.2.2 and 7.2.3 has been obtained through a collaboration with Jorge Zeron Medina Cuairan, and, as such, they have also been submitted as part of his theses for the degree of Doctor of Philosophy at the University of Oxford. Aspects of the work presented in Chapter 7 have been published in *Cell* as part of a collaborative project with Jorge Zeron Medina Cuairan (Zeron-Medina *et al.*, 2013), which I have co-authored. Permission from the co-authors of this publication has been granted to include the material in this thesis. In addition, the data on which the analyses have been performed have been provided by David Oscier and Jon Strefford (Chapters 4.2 and 5.2.2) and Mark Middleton and Anna Grawenda (Chapter 6.2.3).

All parts of the thesis that are not my own work, have been clearly indicated and referenced in the text.

Word count: approx. 47,600

Abbreviations

SNP Single Nucleotide Polymorphism

CNV Copy Number Variant

MAF Minor Allele Frequency

LD Linkage Disequilibrium

HWE Hardy-Weinberg Equilibrium

GWAS Genome-Wide Association Study

TFBS Transcription Factor Binding Sites

ChIP-Seq Chromatin immunoprecipitation sequencing

EMSA Electrophoretic Mobility Shift Assay

eQTL expression Quantitative Trait Loci

kb kilobase (1000 bases)

B-CLL B-cell chronic lymphocytic leukaemia

TFT time to first treatment

Populations

ASW African ancestry in Southwest USA

CEU Utah residents with Northern and Western European ancestry from the CEPH collection

CHB Han Chinese in Beijing, China

CHD Chinese in Metropolitan Denver, Colorado

GIH Gujarati Indians in Houston, Texas

JPT Japanese in Tokyo, Japan

LWK Luhya in Webuye, Kenya

MXL Mexican ancestry in Los Angeles, California

MKK Maasai in Kinyawa, Kenya

TSI Toscani in Italy

YRI Yoruba in Ibadan, Nigeria

DNA Nucleotide conventions

A Adenine

C Cytosine

G Guanine

T Thymine

N Any base

W A or T

K G or T

H Any base but G

Y Pyrimidine

R Purine

Contents

1 Introduction

1.1 SNPs that affect cancer in humans

1.1.1 A brief introduction to common genetic variants of the human genome

Human genetic variation encompasses all of the genetic differences within our species. It has long been believed that mapping the genetic variation will give us a better understanding of the phenotypic differences between humans. However, understanding the connection of genetic variants and disease (or any other trait) has been much more complex than originally anticipated in the majority of cases.

Genetic variants are classified in four broad groups: Single Nucleotide Polymorphisms (SNPs), Tandem Repeats (microsatellites and minisatellites), Indels (short insertions and deletions) and Structural Variants. SNPs are point mutations or single base pair changes in the DNA that occur in more than 1% of a population (Nakamura, 2009). Microsatellites, which are also called Short Tandem Repeats (STRs), are short nucleotide sequences of 1 to 5 base pairs (bp) repeated in tandem in the DNA (Cheng and Zhang, 2010). Minisatellites or Variable Number Tandem Repeats (VNTR), are also nucleotide sequences of tandem repeats but of longer size. Indels are short sequences (less than 50 bp) of DNA that have been inserted or deleted in a specific locus of the genome (Montgomery *et al.*, 2013). Finally, Structural Variants (SVs) are large segments of DNA (longer than 1kb) that have been inserted (in one or multiple copies), deleted, inverted or translocated.

Unbalanced Structural Variants are also called Copy Number Variants (CNVs) and comprise of insertions or deletions of large sequences of DNA (Conrad and Hurles, 2007; Mills *et al.*, 2011).

For all types of variation, a common observation is that some regions of the genome are highly variable, whereas others show very low levels of variation (1000 Genomes Project Consortium *et al.*, 2010). In order to understand and catalogue the evolution and function of the landscape of the human genome, many collaborative projects have been formed, providing the research community with an immense wealth of information. One of the first efforts to sequence and understand the human genome was the Human Genome Project (HGP), a major collaborative effort involving 20 groups in 6 countries, which resulted in the sequencing of more than 96% of the euchromatic part of the human genome (Lander *et al.*, 2001). One of HGP's most interesting findings was that the genomic landscape shows considerable variation in the distribution of genes, transposable elements, GC content, CpG islands and recombination rate. In addition, 30,000 - 40,000 protein-coding genes and more than 1.4 million SNPs were enumerated (Lander *et al.*, 2001).

Following the successes of HGP, a number of large consortium projects developed over the following decade that aimed to understand and catalogue the evolution and function of human genetic variation. One of the most important and pioneering projects of its time was the International HapMap Project, a multinational collaborative effort to catalogue genetic similarities and differences in humans, which was developed in three phases. In phase I, 1.1 million SNPs were discovered from DNA samples of 270 individuals from four populations (the Yoruba

people in Nigeria, Japanese in Tokyo, Han Chinese in Beijing, and the US Utah population with Northern and Western European ancestry-CEPH) (Gibbs *et al.*, 2003). In phase II, higher-resolution genotyping (of about one SNP per kb) was achieved, and an extra 3.1 million SNPs (of lower frequency) were discovered from the same populations as in phase I (Frazer *et al.*, 2007). Finally, in phase III, the number of DNA samples increased from 270 in phases I and II to 1,301 samples from 11 human populations (International HapMap 3 Consortium *et al.*, 2010). Together, the HGP, the International HapMap Project and The SNP Consortium (a collaborative effort between the two) collectively identified about 10 million common DNA variants, primarily SNPs (International HapMap 3 Consortium *et al.*, 2010).

In parallel with the HapMap project, the Human Diversity Genome Project (HGDP) was developed to study human genetic diversity and understand how and when it formed (Cavalli-Sforza, 2005). To this aim, they collected 1,063 lymphoblastoid cell lines (LCLs) from 1,050 individuals from 52 populations worldwide, and created a publicly available cell line panel. However, contrary to the HapMap project, the focus of the HDGP was the identification of microsatellites due to the fact that they are longer and thus have much higher mutation rates, which makes them more suitable for evolutionary studies (Cavalli-Sforza, 2005).

Finally, the advancement of sequencing technologies has given rise to the 1000 Genomes Project, which is the most recent major effort to characterise variation in the human genome (1000 Genomes Project Consortium *et al.*, 2010). Although the project is still ongoing, the initial discoveries have been published from the pilot

phase and phase I. The pilot phase of the project produced low coverage sequencing of 179 individuals (from 4 populations), high coverage sequencing of 6 individuals (from 2 populations) and exon-targeted sequencing of 697 individuals (from 7 populations) (1000 Genomes Project Consortium *et al.*, 2010). Phase I produced low coverage sequencing of 1,092 unrelated individuals from 14 populations from four continents. Phases II and III are expected to see this number to rise to 2,500.

In the pilot phase of the 1000 Genomes Project, more than 15 million SNPs were catalogued, of which approximately 60,000 were found to be synonymous SNPs (alleles in coding exons that do not change the amino acid sequence of proteins), more than 68,000 non-synonymous (alleles that change the amino acid sequence of proteins), approximately 1,000 introduce stop codons and approximately 500 disrupt splice sites (1000 Genomes Project Consortium *et al.*, 2010). In phase I, approximately 38 million SNPs were discovered, which tended to be rarer and reside in non-coding regions of the genome. In addition, more than a million short indels (1-50 bp) and 14,000 larger deletions were catalogued (1000 Genomes Project Consortium *et al.*, 2012).

Together, these projects have proven invaluable for the development of the techniques and methodologies used in a number of applications of genetic data. In addition, based on inference from closely linked polymorphisms, the reference genomes generated have provided the base on which genome-wide association studies (GWAS) have used these data to impute missing SNPs and identify thousands of associations of SNPs with human traits and disease.

1.1.2 SNPs in cancer research : From linkage analysis to genome-wide association studies (GWAS)

It is a common belief that genetic variants exist that affect individuals' susceptibility to cancer and cancer progression. Based on data from epidemiological studies, most common cancers show a 2 to 4-fold increase in familial risk compared to population studies (Easton and Eeles, 2008; Hosking *et al.*, 2011). These observations have led to the development and applications of linkage analysis methods, which use related individuals for the mapping of loci associated with disease (Botstein *et al.*, 1980).

These linkage studies were initially very successful in the identification of high-penetrance risk loci, such as the BRCA1 and BRCA2 genes associated with breast cancer (Hall *et al.*, 1990; Wooster *et al.*, 1994) and CDKN2A associated with melanoma (Cannon-Albright *et al.*, 1992). However, with the proposal of the 'common disease, common variant' hypothesis, by which alleles of high frequency ($> 5\%$) are hypothesised to be the cause of common disease (Lander, 1996), the focus changed to case control studies, as linkage studies were under-powered to detect the signals involved (Risch and Merikangas, 1996).

Case control studies of cancer initially focused on specific key genes that were hypothesised to carry SNPs that would affect cancer risk, or to specific polymorphisms of cancer genes. This became widely known as the candidate gene approach. Similar to linkage studies, the candidate gene approach identified some important associations, such as SNPs in the coding region of CASP8 associating with breast cancer (Cox *et al.*, 2007), but most of their findings were not replicated

in subsequent studies (Easton and Eeles, 2008; Hosking *et al.*, 2011).

With the advent of new technologies that allowed genome-wide genotyping at a relatively low cost for large cohorts, the era of the GWAS, an agnostic approach of scanning the whole genome for associations, began. To date, more than 1,000 SNPs have been associated with cancer in a GWAS[1]. However, in most cases the variants found are high-frequency, low effect size SNPs, preventing them from being used as screening tools in the clinic. In detail, the per allele estimated effect size has been found to range between 1.1 and 1.4, and the majority of new loci discovered have susceptibility alleles of minor allele frequency (MAF) that is larger than 10% (Chung and Chanock, 2011; Hosking *et al.*, 2011). Chung and Chanock (2011) note that this is not a surprising observation since GWAS is underpowered to detect associations of SNPs with low allele frequency. To overcome this lack of power, some studies have developed study designs that could bias their results. For example, they use publicly available data for the controls instead of matching populations. On the other hand, some of the designs, which appear to be lenient compared to the widely accepted quality controls used in GWAS, have been able to replicate their findings in more rigorous follow up studies (Chung and Chanock, 2011). This questions the necessity of such strict thresholds in cancer research, given that large sample sizes are hard to obtain.

An additional complication in cancer GWAS is the involvement of environmental factors, and their contribution to the complexity of associations of risk alleles to cancers with a strong environmental element. In most cases it is diffi-

[1]from the following databases: `http://www.genome.gov/gwastudies`, `http://www.ncbi.nlm.nih.gov/gap`, `http://www.hugenavigator.net/CancerGEMKB/`

cult to disentangle the potential confounding effects, and it has been suggested that such cases could benefit from more sophisticated statistical models and data-mining techniques to account for this complexity (Hosking *et al.*, 2011).

Apart from associating SNPs to cancer susceptibility, a vital step is better understanding their function in carcinogenesis. To date, the majority of SNPs identified in cancer GWAS have been located in regions not previously thought to be involved in carcinogenesis and, in most cases, in noncoding regions of the genome (Chung and Chanock, 2011; Hosking *et al.*, 2011). However, a number of loci have been associated with susceptibility to more than one type of cancer, so appear to have pleiotropic effects (Chung and Chanock, 2011; Hosking *et al.*, 2011). A notable example is the region spanning up to 800 kb upstream of the MYC oncogene (locus 8q24), which includes five independent loci that have been associated with breast, ovarian, colorectal and bladder cancer, as well as with chronic lymphocytic leukaemia (Amundadottir *et al.*, 2006; Easton *et al.*, 2007; Tomlinson *et al.*, 2007; Zanke *et al.*, 2007; Yeager *et al.*, 2007; Gudmundsson *et al.*, 2007; Kiemeney *et al.*, 2008; Al Olama *et al.*, 2009; Crowther-Swanepoel *et al.*, 2010). This supports the existence of central pathways that are important for carcinogenesis.

In addition, it has been noted that almost none of the susceptibility loci discovered to date have also been associated with the clinical outcomes of the same type of cancer, such as overall survival. This would suggest that the pathways involved in the two processes are distinct (Chung and Chanock, 2011). The application of GWAS on clinical outcomes such as survival has not been as well studied.

This is largely due to the large sample sizes required for a GWAS in combination with the low incidence rates of certain cancers. In addition, the cohort homogeneity (in terms of diagnosis and treatment) needed for such studies adds a further layer of complexity to the problem.

One of the most important databases for the collection and curation of the results of GWAS is the National Human Genome Research Institute (NHGRI) catalogue, a database of SNP-trait associations from published GWAS. To date, it includes more than 1,500 publications and 10,000 SNPs associated with human traits (Hindorff *et al.*, 2009). However, only 21 of these publications have identified SNPs associating with cancer survival or response to treatment (53 SNPs). Compared to the hundreds of SNPs identified to associate with cancer risk, this small number highlights the problems that are encountered in the design and implementation of survival analyses for cancer studies with low sample sizes and high heterogeneity.

1.2 High dimensionality in genome-wide studies

1.2.1 Multiple hypothesis testing

With the exponential increase in the volume of genomics data produced by high throughput technologies, data analysts have encountered the problem of having to obtain statistical estimates for many variables p when the sample size n is comparatively very low ($p \gg n$). This is a problem also known as 'large p, small n' that occurs in high-dimensional datasets. In the context of GWAS, this is commonly addressed by performing one test for each SNP, since testing marginally

avoids the 'curse of dimensionality'. However, this multiplicity introduces inflated type I errors (false positives) which need to be corrected for. The methods to correct for multiple hypothesis testing in genomic studies control either the familywise error rate (FWER) or the false discovery rate (FDR). The familywise error rate (FWER) is the probability of one or more false rejections, whilst the false discovery rate (FDR) is the expected proportion of incorrect rejections of null hypotheses among all rejections.

One of the most widely used methods of performing a multiple hypothesis correction is the Bonferroni correction, which provides strong control of the FWER. In this case, each test is controlled at α/p, with α being the desired type I error and p the number of tests. However, this type of correction is too conservative because many SNPs are highly correlated, violating the assumption of independence of tests. Correcting for the FWER in a less conservative and more exact way can be achieved via permutations where the empirical distributions of the p-values are used. Nonetheless, this is a very computationally intensive method, which is not easily employed in practice. More recently, a new methodology has been proposed, by which the tests are corrected based on the effective number of independent tests M_{eff}, which is estimated from the eigenvalue variance of the correlation matrix of the SNPs (Cheverud, 2001; Nyholt, 2004). However, these methods are still considered to be conservative, especially when there is high linkage disequilibrium (LD, a term which is explained in detail in section 1.4.1) between SNPs (Gao *et al.*, 2008).

Over the last two decades, a number of methodologies have emerged for mea-

suring significance based on the FDR. In a seminal paper, Benjamini and Hochberg (1995) presented a method for controlling FDR and showed that this method had greater power than the Bonferroni correction. This method also assumed independent test statistics, but Benjamini and Yekutieli (2001) proposed a modification for dependent test statistics. Based on these procedures, Storey (2003) introduced an extension, where an estimate of the proportion of features that are truly null was incorporated to the methodology. This estimate was based on the idea that these features would be uniformly distributed among $[0, 1]$, whereas the alternative features would be closer to 0. However, both FDR approaches are known to have a weaker control of FWER (Yang *et al.*, 2005).

In survival analysis, the most commonly used methods to obtain probability estimates for each SNP are the log-rank test and the Cox proportional hazards model (Mantel, 1966; Cox, 1972). In practice, however, both of these methods have important disadvantages in a genome-wide setting. The log-rank test has been found to be biased when the sample size is small and unbalanced (i.e. unequal between groups), or when the censoring is heavy in one group (Latta, 1981; Heinze *et al.*, 2003). As a consequence, when testing SNPs of low minor allele frequency (MAF), which produce very unbalanced groups, the statistics tend to be biased towards groups with smaller sample sizes, increasing false positive results. On the other hand, the results of the Cox proportional hazards models are dependent on assumptions in the model. This includes assumptions of proportionality between the groups as well as those on the underlying genetic effects of the SNPs, similarly to linear models. Specifically, the assumption of proportionality

cannot be tested for genome-wide studies and violations can result in loss of power (Schemper, 1992). In addition, it has been shown that assumptions on the genetic model and consequent coding/grouping of variables can lead to very different results from association studies (Lunetta, 2008; Zuk *et al.*, 2012). Even though the additive model is the most widely used model in GWAS, the choice of the most appropriate genetic model for association studies has been a subject of debate. Simulation studies done to assess the statistical power of different models have shown that the additive model has good performance for additive or dominant models but poor performance for recessive models and therefore a co-dominant model should be preferred (Lettre *et al.*, 2007). On the other hand, Cantor *et al.* (2010) have argued that the additive model is a good compromise between the genetic truth and parsimony, recommending it as the best model for the initial screening. Therefore, all models come with their advantages and disadvantages and the genetic model used could have an important impact on the results. A further discussion of this will be carried out throughout this thesis.

1.2.2 Data mining techniques and the discovery of biomarkers in survival analysis

Approaching the high-dimensionality problem from a different angle, many data mining techniques have been developed to deal with the increasing volume of genetic data available. Some of the most popular techniques in genetics and specifically in SNP analysis are discussed here, namely penalised regression methods and, more specifically, the lasso, Support Vector Machines (SMV), and Random

Forests (RF). This review is by no means exhaustive, as this would be beyond the scope of this thesis. It merely presents some of the most widely used data mining techniques, and discusses their use in genetic studies with a focus on their applications to survival data. Here, survival analysis is considered in a wide context, including any type of time-to-event phenotypes, such as overall survival and time to metastasis.

The lasso was the first of the penalised regression methods to be developed, and is a regression method by which the residual sum of squares is minimised subject to the sum of the absolute value of the coefficients being less than a constant (L_1 penalty) (Tibshirani, 1996). The penalised regression family of methods also includes ridge regression, which uses an L_2 penalty on the sum of the squares of the coefficients, and the elastic net, which uses a combination of the two penalties. Penalised regression models have been used in both candidate gene approaches and two-stage GWAS for prediction models and variable selection, as well as for rare SNP detection (Croiseau and Cordell, 2009; Wu *et al.*, 2009; Ayers and Cordell, 2010; Kooperberg *et al.*, 2010; Guo *et al.*, 2011). In a candidate gene context, Ayers and Cordell (2010) showed that penalised regression methods, in general, outperform single marker analysis providing sparse models, but Croiseau and Cordell (2009) did not observe any advantages over univariate analysis. However, the application of penalised regression methods in genome-wide analysis has been deemed computationally infeasible since the variable selection depends on a tuning parameter that needs to be determined by cross-validation, increasing the computational cost of this method (Szymczak *et al.*, 2009).

Support Vector Machines (SVMs) were first described by Cortes and Vapnik (1995) for binary classification, but can also be used for regression. It is a method that looks for the optimal separating hyperplane between two classes by maximising the distance between the closest points of the two classes. SVMs have been used in a variety of applications in genomic studies, including microarray analysis and predictions based on SNP data (Brown *et al.*, 2000; Ban *et al.*, 2010). However, Verikas *et al.* (2011) comment that predictors based on SVMs provide too little insight into what is the importance of the variables in the predictor derived and, as such, are not easily used for variable selection.

Random Forest (RF) is a tree ensemble technique developed by Breiman (2001), which has been found to be a stable and robust method for prediction. It has been extensively used in biological research for the analysis of a variety of data from high-throughput genomic technologies, because it yields low bias from the deep trees and reduced variance from the aggregation of the trees (Chen and Ishwaran, 2012). Specifically, it is most commonly used in microarray and SNP studies to construct predictive classifiers and rank SNPs based on their predictive importance, as well as to perform pathway analysis (Bureau *et al.*, 2005; Schwender *et al.*, 2004; Díaz-Uriarte and Alvarez de Andrés, 2006; Pang *et al.*, 2006; Meng *et al.*, 2007; Chang *et al.*, 2008; Goldstein *et al.*, 2010). In addition, many methodological papers have focused on fine-tuning Random Forest for implementation in SNP studies, aiming to test the variable importance measures and deal with the effect of SNP linkage on RF performance (Strobl *et al.*, 2007; Nicodemus and Malley, 2009; Meng *et al.*, 2009; Nicodemus *et al.*, 2010; Nicodemus, 2011; Walters *et al.*,

2012). Three recent reviews of RF in genomic studies provide an extensive list of their applications and advantages, as well as comparisons to other data mining techniques (Boulesteix *et al.*, 2012; Chen and Ishwaran, 2012; Touw *et al.*, 2013).

SVM and RF have been cited to be the "the most widely used classification techniques in the Life Sciences", where the performance of the former was noted to be slightly superior to the latter, when appropriately tuned (Touw *et al.*, 2013). On the other hand, Pers *et al.* (2009) compared the lasso, SVM and RF in a high-dimensional setting and showed a lower bootstrap cross-validation error for RF, followed by SVM and the lasso. In addition, Guo *et al.* (2010) compared RF to a number of other data mining techniques, including SVMs, and noted that it performed best when the class distribution was unbalanced, as it often is for SNP data. However, it is a common consensus that each method has specific applications in which it will perform better, and that there is no technique that is globally superior.

With regards to their extensions in survival analysis, the lasso can be used by minimising the log partial likelihood of the Cox proportional hazards models (Tibshirani, 1997). However, this method is not as widely utilised in genetic studies of survival analysis. To our knowledge, only Kim *et al.* (2013) have used the lasso for SNP data in a survival setting before. They propose a two stage approach using the gradient lasso method, but their main goal is prediction and not variable selection. A number of extensions of SVMs have been proposed for survival (censored) data (Shivaswamy *et al.*, 2007; Van Belle *et al.*, 2007; Khan and Zubek, 2008; Evers and Messow, 2008). However, only the variant of Evers and Messow (2008) has

been applied to SNP data for the identification of pathways associated with the overall survival of multiple myeloma patients (Pang *et al.*, 2011). This low number of applications for two of the most common data mining techniques demonstrates the lack of studies utilising high-dimensional techniques for survival analysis.

Based on the idea of tree ensembles and RF, over the last decade a number of methodologies have emerged that use survival trees. Two of these variants are bagging survival trees and survival ensembles (Hothorn *et al.*, 2003, 2006). The first one uses bootstrap samples to build the trees and compute the aggregated Kaplan-Meier estimates for prediction, whereas the second uses the estimated inverse probability of censoring weights for the bootstrap sampling of the RF, using the log-transformed survival time as the outcome. In addition, Ishwaran *et al.* (2008) presented an algorithm called Random Survival Forest (RSF) that utilised bootstrap samples for growing the trees, the Nelson-Aalen estimators for the terminal nodes and the averaged estimators for prediction (more details can be found in Materials and methods). RSF was also compared to survival ensembles in a variety of datasets and it was shown that the RSF had a much lower prediction error in high-dimensional settings. A more detailed introduction on the use of the RF and the RSF on SNP data is presented in Chapter 3.

The RSF has been previously compared to the Cox proportional hazards model and the Kaplan-Meier estimates. In a review of the applications of survival trees to censored data, Bou-Hamad *et al.* (2011) compared the performance of five methods on prediction: namely the Kaplan-Meier estimates, the Cox proportional hazards model, a single tree, the bagging survival trees and the Random Survival

Forest (RSF). They note that RSF presents the best results, followed by bagging and Cox, which have very similar results. However, the predictions were based on the results of the methods on a dataset with only 12 covariates and 300 patients, a scenario very different to the high-dimensional data met in genetic studies. Unfortunately, very few reviews comparing data mining techniques in survival analysis exist. The main reason for this is that there is no consensus on a measure for predictive performance for censored data (Bou-Hamad *et al.*, 2011; van Wieringen *et al.*, 2009). A more detailed analysis of the problems of the comparison of methodologies in survival analysis is given in van Wieringen *et al.* (2009).

Finally, there have been many papers comparing the RF to other data mining techniques for genomics data, but the RSF methodology has only been developed recently and, as such, has not been as extensively studied. Two exceptions are the studies by Datema *et al.* (2012) and Pang *et al.* (2011). Datema *et al.* (2012) compared the RSF to the Cox proportional hazards model and noted very similar error rates between the two approaches. In addition, Pang *et al.* (2011) compared the RSF to various other data mining techniques, including a variant of the SVMs, in a microarray study and showed improved performance of the RSF. However, no studies have compared the RSF to other methodologies in a simulated or real SNP analysis. This study aims to address this by studying the RSF under simulated datasets of SNP data based on real LD structure. The RSF methodology is fine-tuned for this data, and compared for its ranking abilities to the most commonly used methods of GWAS for survival analysis, namely the log-rank test and the Cox proportional hazards model, in a high dimensional setting.

1.3 SNPs and the chemotherapeutic response in humans

1.3.1 Cancer cell lines and the discovery of biomarkers for chemotherapies

GWAS have focused on the identification of SNPs associating with cancer risk and, as mentioned above, to a lesser extent with cancer survival. Another important area of genetic research is the potential role of SNPs in drug response. The most commonly used model systems in cancer drug research utilise tumour-derived cell lines. Cell lines have been extensively used for decades as model systems for the discovery of new experimental drugs and for the genomic characterisation of tumour types (Sharma *et al.*, 2010; Shoemaker, 2006; Gillet *et al.*, 2013). The first publicly organised effort to use a panel of cell lines for drug screening was the US National Cancer Institute (NCI) 60 panel. It consisted of 59 cell lines derived from 9 different types of cancer, namely leukaemia, colon, lung, renal, melanoma, ovarian, breast, prostate and central nervous system (CNS). This project constituted a shift from previous models in that it included cell lines from solid tumours and not just leukaemic ones (Shoemaker, 2006). Many technologies were developed as part of the challenges that had to be circumvented for the realisation of this project, many of which still represent the bases for cancer drug screening programmes today, such as the development of assays for the measurement of cytotoxicity (Sharma *et al.*, 2010; Shoemaker, 2006).

The initial aim of the NCI60 project was to screen more than 10,000 compounds, per year. To date, however, a number of different laboratories have expanded the screen to more than 100,000 compounds ranging from natural product

extracts to well-characterised drugs and chemotherapeutic agents with known and potential activity in cancer (Shoemaker, 2006; Weinstein, 2012).

Following the development of the NCI60 project, the Japanese Foundation for Cancer Research (JFCR-39) project, which utilised a cell line panel of 39 tumour cell lines, was established in 1999. This cell panel included a subset of the NCI60 as well as six stomach cancers to address the high prevalence of this cancer in Japan (Yamori, 2003). Similarly to the NCI60, the JFCR-39 included a screen of more than 300 standard compounds including anticancer drugs and inhibitors of known pathways.

At the time of the development of the NCI60 and the JFCR-39 anticancer screens, clinical response to known chemotherapeutics, such as cyclophosphamide and cisplatin, had achieved very high rates, reaching up to 65% of patients for certain types of cancer, justifying the use of only a few cell lines per cancer type for potentially high impact discoveries (Gillet *et al.*, 2013; Sharma *et al.*, 2010). Due to the high expectations of response rates and the labour intensive processes involved in developing such a grand scale experiment, these efforts were focused on what is nowadays considered a very limited number of cell lines per cancer type. However, with the development of high-throughput technologies, the expansion of human tumour cell line panels has been an achievable goal. Over the last decade, two major collaborative efforts, the Cancer Genome Project (CGP) and the Cancer Cell Line Encyclopedia (CCLE), have separately conducted a series of experiments aiming at characterising more than a thousand cell lines (Garnett *et al.*, 2012; Barretina *et al.*, 2012). Their databases provide access to many types of data, such as mutation

data for numerous genes implicated in cancer formation and progression, copy number variations and mRNA expression. Additionally, the response to more than a hundred compounds have been measured for many of these cell lines.

The advent of targeted therapies has changed the focus of the drug screens from cytotoxic agents to agents aimed at targeting specific tumour-associated proteins, such as oncogenes. In the Genomics of Drug Sensitivity in Cancer (GDSC) database, which is part of the CGP, 138 anticancer compounds have been screened but only 13 of them are conventional chemotherapeutic agents. The rest are targeted agents that are either already in clinical use, in development or are experimental compounds (Garnett *et al.*, 2012; Yang *et al.*, 2013). Similarly, 24 compounds have been selected for screening in the CCLE but only 3 are cytotoxic agents (Barretina *et al.*, 2012).

Although drugs that directly target proteins of oncogenes have been very successful in the clinic, most of them have low response rates (Sharma *et al.*, 2010; Caponigro and Sellers, 2011). An important and illustrative example is the tyrosine kinase inhibitors (TKI) gefitinib and erlotinib, which target epidermal growth factor (EGFR) in non-small cell lung cancer (NSCLC) patients. Even though more than 60% of all cases of NSCLC show an overexpression of EGFR, the response rates for both TKIs are as low as 10% (Sharma *et al.*, 2010; Laurie and Goss, 2013). Many potential explanations have been proposed, but at the core of all of them lies the immense genomic heterogeneity observed between tumours.

Contrary to the targeted agents that typically associate with somatic mutations and the overexpression of specific oncogenes, cytotoxic agents have not been

found to be as highly correlated with somatic mutations (Garnett *et al.*, 2012). Therefore, other predictive biomarkers are needed for the identification of patients that will respond well to the conventional and heavily utilised chemotherapeutic agents. Identifying these patients and understanding the biological processes that underlie the effects of these agents could also help explain the high rate of failure of many of chemotherapeutic agents in clinical trials (Caponigro and Sellers, 2011). Nonetheless, because the chemotherapeutic agents are acting on cell processes, predicting the subset of patients that would most benefit from a course of treatment is extremely challenging. Even though their role is not always curative, they have been found to improve progression-free survival and can be used as neoadjuvant therapies (DeVita and Chu, 2008). Therefore, the identification of biomarkers associating with chemotherapeutic response could be a great aid in classifying groups of patients that would most benefit from these agents.

1.3.2 SNPs in the p53 network of genes

p53 was discovered in 1979 as a protein that was expressed in cells transformed by the Simian virus 40 (SV40), but was then shown to be of cellular and not viral origin (Linzer and Levine, 1979). It is known to trigger the machinery for DNA repair, while also being able to induce cell cycle arrest. Most importantly, it controls cell growth by inducing senescence or apoptosis (Vousden and Lu, 2002; Vousden and Prives, 2009; Lane and Levine, 2010). Due to its pivotal role in essential cellular activities it is one of the most important tumour suppressor genes in the genome.

The TP53 gene is found in all vertebrates, whereas invertebrates have a p53-

like gene with a very similar DNA binding domain. It belongs to the p63/p73 family of proteins, which is also found in unicellular organisms. The conservation of this family of genes over more than a billion years of evolution underlines its importance for most organisms (Belyi *et al.*, 2010). The function of p53 varies greatly under different circumstances and it can act as a transcription factor or be part of complexes that trigger signalling cascades. Specifically, upon cellular stresses, such as hypoxia, UV radiation and DNA damage, p53 is post-translationally modified and its DNA binding domain is exposed, allowing it to bind to specific DNA sequences or response elements (REs) and activate the transcription of target genes (Vousden and Lu, 2002; Vousden and Prives, 2009; Lane and Levine, 2010). Without cellular stress, p53 is transcribed at constant rates but its protein is rapidly inactivated through its interaction with MDM2 (the murine double minute oncogene), and its subsequent ubiquitination and degradation. Cellular stresses activate a signalling cascade that results in the phosphorylation of p53 and its inability to interact with MDM2. The protein is, thus, stabilised and functions as a transcription factor for hundreds of genes, including MDM2. Attenuation of the cellular stresses results in the regained ability of the overexpressed MDM2 protein to interact with and inactivate p53, thereby restoring the equilibrium of this negative feedback loop (Lane and Levine, 2010; Weinberg, 2013).

The tumour suppressor TP53 gene, which encodes the p53 protein, is mutated in about 20 to 50% of most tumour cell genomes. Experiments performed in mice have shown that 75% of animals with an inactivated TP53 develop a tumour by the 6th month of their lives, and have multiple primary tumours of different

types (Donehower *et al.*, 1992). In humans, germline mutations in the TP53 gene are responsible for a familial autosomal dominant disorder called the Li-Fraumeni syndrome (LFS). LFS individuals have increased susceptibility to developing tumours at an early stage of their lives. It is a very rare syndrome and about half of the individuals affected develop cancer by the age of 40. It is associated with a variety of different cancer types, including soft tissue sarcoma, breast cancer, brain tumours, adrenocortical carcinomas and leukaemias. One important characteristic of the syndrome is that individuals with LFS have an increased risk of developing multiple and diverse primary tumours throughout their lives (Malkin *et al.*, 1990; Varley, 2003).

Contrary to the germline mutations seen in the LFS patients, more common genetic variants, such as SNPs, in the TP53 gene and its network of genes have been shown to have a moderate effect on cancer risk and survival (Whibley *et al.*, 2009; Lane, 2010). Nevertheless, the study of these SNPs can provide a biological insight into the function of these proteins, and aid in the targeted therapy of patients.

Over 120 other genes have been found to be regulated by p53 but this list is still expanding as our knowledge of this complex gene is even now still incomplete. Sequence analysis of p53's the DNA binding domain has identified, with high confidence, 542 sites that can interact with p53 (Wei *et al.*, 2006). Recently, two studies have focused on the identification of functional SNPs in the p53 pathway of genes and their effect on the cellular response to cytotoxic agents and ionising radiation (Vazquez *et al.*, 2008, 2010). This has been triggered by the indisputable role of p53 in the DNA damage and stress responses. These studies have resulted

in the identification of 2 SNPs that exhibit allelic differences in cellular responses to chemotherapeutic agents. In addition, they were found to associate with earlier onset and worse survival in a cohort of soft tissue sarcoma patients, highlighting the importance of functional SNPs in a pathway as important as p53. Following this paradigm, additional studies could further aid the identification of SNPs in pathways central to tumourigenesis and chemotherapeutic response.

1.4 Associations and causality in genetic studies: In search of functional SNPs

1.4.1 Association studies are only the first step in the process of discovering causal SNPs

Even though many SNPs have been associated with cancer-related phenotypes, the effect sizes observed are too low to be used as screening tools and the signal does not always correspond to the causal variant, providing no insight into the biological function of these variants. In detail, most SNPs are part of haplotype blocks, sequences of DNA that are inherited as a unit in chromosomes, and that are in linkage disequilibrium (LD) with the SNPs of the same haplotype block (Cheng and Zhang, 2010). Therefore, in association studies, the causal SNP cannot always be distinguished from the haplotype block that contains the trait-associated SNP because they have very similar p-values.

Furthermore, it has been shown that in most cases the SNP which is supported by experimental evidence it is not the reported SNP itself, but a linked one (Schaub

et al., 2012). It has been estimated that only about 12% of trait-associated SNPs or their proxies (other SNPs in strong LD with the associated variants) reside in protein coding regions, even though the protein coding regions of genes are over-represented in genotyping platforms (Manolio, 2010). In addition, about 80% reside in intergenic or intronic regions (Hindorff *et al.*, 2009; Manolio, 2010; Freedman *et al.*, 2011).

Although many studies, including the HapMap and 1000 Genomes projects, have provided the community with detailed LD maps of several populations, fine-mapping the associated regions in order to identify causal SNPs has proved to be a great challenge, and its success has been limited (Ioannidis *et al.*, 2009; Shea *et al.*, 2011). Fine mapping consists of deep sequencing the associated region in an independent cohort. One of the problematic points of fine mapping is that there is no consensus on the boundaries of the regions to be targeted (Freedman *et al.*, 2011). This is due to the fact that the boundaries depend on: a) the LD structure of the population that is being studied; b) the effect size; and c) the allele frequency of the causal variant. In addition, if the causal variant is a rare SNP or if the effect size is very small, the sample sizes required to identify it could further hinder the search (Ioannidis *et al.*, 2009).

In cancer GWAS, re-sequencing a replication cohort for the identification of the causative SNP requires very large sample sizes, and/or additional populations that have similar incidence rates of the cancer in question. Fine-mapping an associated region in order to find the causative SNP associating with survival or cancer progression is even more arduous, since it requires the validation cohort to have

similar clinical characteristics to the discovery cohort, and the phenotype to be measured in the same way.

For these reasons, GWAS and other types of association studies are being used as hypothesis-generating studies, which provide the candidate loci for further experimentation and investigation (Stranger *et al.*, 2011). This has been the topic of extensive research in the last few years, and some approaches have already been suggested regarding how to proceed with the causative/functional annotation of the associations observed, such as the one employed here, of proposing hypotheses of transcriptional regulation based on publicly available ENCODE data (Moore *et al.*, 2010; Freedman *et al.*, 2011; Manolio, 2010).

1.4.2 ENCODE as a tool for the functional annotation of associated SNPs

Many experimental techniques exist for the discovery of SNPs that affect gene regulatory regions (regulatory SNPs), such as electrophoretic mobility shift assays (EMSAs) for the *in vitro* detection of DNA-protein binding, and promoter reporter constructs for the detection of the differential expression of the reporter gene (e.g. luciferase activity). However, these standard techniques that aim to assess the effect of SNPs are bottlenecks in the new era of high-throughput technologies (Chorley *et al.*, 2008). With the development of the Encyclopedia of DNA Elements (ENCODE) project, the potential of using publicly available data to create functional annotation tools has changed dramatically. Even from the completion of the ENCODE pilot project, where only 1% of the human genome was analysed, it became

apparent that the human genome is extensively transcribed, suggesting that what we have so far considered as functional may not be precise.

Since the beginning of its design, the aim of the ENCODE project has been to "delineate all functional elements encoded in the human genome" (ENCODE Project Consortium *et al.*, 2012). The second phase of this huge feat, phase I, was completed in 2012. The data was published in 6 papers specifically highlighting the findings of this immense collaborative effort and many more accompanying papers. The ENCODE database now includes more than 1,640 genome-wide datasets from 147 cell types. Over 100 transcription factors, and an additional 13 histone or DNA modification markers, have been analysed to identify transcription factor occupied regions and patterns of DNA modification that represent regulatory regions across these cell types.

In the principal ENCODE paper, it was shown that 80% of the genome is comprised of elements linked to biochemical functions (ENCODE Project Consortium *et al.*, 2012) changing our view of what percentage of the genome is actually functional and what can still be referred to as 'junk DNA'. Moreover, it was estimated that 99% of the genome is less than 2 kb away from at least one of the biochemical events measured. In one of the accompanying papers, Djebali *et al.* (2012) noted that about 75% of the genome is transcribed at some point, even though no cell line transcribes more than 60% of the total observed transcriptome. This further supports the idea of abundant cell-type specific differences, and highlights the importance of identifying the appropriate systems for testing the allelic differences of potential regulatory SNPs.

Moreover, almost 400,000 regions with enhancer-like features, and more than 70,000 regions with promoter-like features were identified (ENCODE Project Consortium *et al.*, 2012). Seven classes of functional elements were defined, among which were transcription start sites (TSS), predicted enhancers (E) and predicted transcribed regions (T). It was noted that although TSS states had a similar number of cell invariant and cell specific occurrences, enhancer (E) and transcribed (T) states had an extensively cell-specific mode of function. Accounting and adjusting for these cell specific differences can be very laborious and expensive in the context of a standard experimental framework. However, the use of high-throughput ENCODE data can narrow-down the candidate cell types that a SNP could have an effect on.

Finally, an enrichment of SNPs in non-coding functional elements was noted (ENCODE Project Consortium *et al.*, 2012). Of about 4,500 GWAS SNPs examined, 12% of the SNPs were found to be in transcription factor occupied regions, but more than 30% were in DNAse I hypersensitivity sites (DHS), representing enriched subsets compared to the overall proportions of the SNPs detected in the 1000 Genomes project. These enrichments were present even after correcting for selection bias in the genotyping arrays. It should be noted that these percentages would be expected to be even higher, if the proxies of the trait-associated SNPs were also taken into account. Utilising this wealth of data for the identification of SNPs in cancer research could provide invaluable information about function of trait-associated SNPs. In this thesis, extensive use has been made of the ENCODE data, and how it can be incorporated into genetic studies is demonstrated.

1.4.3 Expression Quantitative Trait Loci (eQTL) studies and their role in understanding the effect of SNPs

Expression quantitative trait loci (eQTLs) are genomic regions which, in the context of this project, are SNPs that correlate with the mRNA expression levels of a gene (Jansen and Nap, 2001). The effects of eQTLs can be grouped into *cis*-eQTLS and *trans*-eQTLs, depending on whether they are acting on neighbouring (within 100 kb upstream and downstream of the surrounding genes) or distant genes (Cookson *et al.*, 2009). It has been estimated that most *cis*-eQTLs (about 95%) reside within 20 kb upstream of the transcription start site (Veyrieras *et al.*, 2008; Stranger *et al.*, 2012). Interestingly, in two recent studies by Nicolae *et al.* (2010) and Nica *et al.* (2010), it was shown that SNPs identified in GWAS are significantly more likely to be eQTLs. Furthermore, Nicolae *et al.* (2010) compared trait-associated SNPs to background SNPs of matching *MAF* that also belong to genotyping platforms, and showed that their results were robust to these effects. Therefore, they advocated the use of eQTLs for distinguishing true associations from noise, as well as characterising the biological mechanism underlying the associations.

With the era of high-throughput analyses, the availability of genome-wide eQTL studies has grown exponentially. One of the most widely used databases for the analysis and visualisation of eQTLs is the Genevar platform (Yang *et al.*, 2010). Genevar is based on three datasets; the Multiple Tissue Human Expression Resource (MuTHER) dataset of three tissue types (adipose, lymphoblastoid cell lines and skin) collected from female twins (Nica *et al.*, 2010), lymphoblastoid cell

lines (LCLs) from HapMap3 individuals of 8 populations (CEU, CHB, GIH, JPT, LWK, MEX, MKK and YRI) (Stranger *et al.*, 2012) and three cell types (fibroblasts, LCLs and T-cells) derived from the umbilical cords of the Geneva GenCord individuals (Dimas *et al.*, 2009). Therefore, a variety of tissue and cell types are available for analysis and the results can be incorporated into SNP analyses for the identification of functional genetic variants.

1.4.4 Conservation of functional regions and selective pressures

Even though evolutionary conservation is an important indication of functionality, it had been previously suggested that many functional elements of DNA do not show signs of conservation between species (Birney *et al.*, 2007; Odom *et al.*, 2007). This observation was confirmed with the data from the ENCODE project, where it was shown that the amount of non-coding but functional genome is greater than that of either known coding sequences or evolutionarily constrained bases (ENCODE Project Consortium *et al.*, 2012).

However, some classes of functional elements still preserve high rates of conservation. For example, for DNase I hypersensitive sites and loci where transcription factors were found to bind, there was an enrichment of mammalian constraint and a reduced population diversity within humans (ENCODE Project Consortium *et al.*, 2012). In addition, it was suggested that a large proportion of the elements that are not conserved are lineage specific, as they are required for the function of the organisms (ENCODE Project Consortium *et al.*, 2012). These elements could be distinguishable by evidence of natural selection within the human species.

Negative selection, also called 'background selection', is known to act as a purifying force, reducing the variability of the genome and causing the loss of deleterious alleles (Charlesworth, 1994; Pool *et al.*, 2010). Positive selection, on the other hand, causes an increase in allele frequency and longer haplotypes than expected by genetic drift alone, a phenomenon also known as 'genetic hitchhiking' (Castle, 2011; Casto and Feldman, 2011).

As mentioned above, ENCODE defines seven classes of functional elements. It had been previously asserted that more than 5% of the human genome has undergone purifying selection (Lindblad-Toh *et al.*, 2011). Based on these estimates, the ENCODE Project Consortium *et al.* (2012) suggested that all classes show signs of negative (purifying) selection on primate-specific elements, supporting the idea of functionality for these elements.

In the context of SNPs, it has recently been shown that there are less SNPs in conserved regions compared to non-conserved ones, and that functional regions also contain less SNPs compared to the genome (Castle, 2011). In contrast, SNPs that have been identified in GWAS have been shown to have higher integrated haplotype homozygosity scores (iHS, a measure of natural selection within a population) in Europe and East Asia (Casto and Feldman, 2011), suggesting that natural selection processes that affect SNP frequencies have phenotypic consequences. Therefore, conservation and natural selection tools could prove to be important aids in the identification of SNPs associating with cancer traits.

1.5 Summary and objectives of the study

To summarise, over the last decade there have been immense efforts to discover and catalogue human genetic variants, with the aim of understanding their functions and effects on human traits and disease. In cancer research, many GWAS have been successful in identifying new regions associated with cancer susceptibility, but associations with other clinical phenotypes of cancer have been understudied. Data mining techniques could aid in the analysis of thousands of SNPs in a single framework, providing a solution to the multiplicity problem of the traditional GWAS, which needs large sample sizes to deal with the strict thresholds of multiple hypothesis correction. However, their application in survival analyses has been limited so far.

In addition to survival phenotypes, response to therapy is a field of paramount importance. The use of cancer cell lines and publicly available drug screening databases could provide us with essential inherited biomarkers for response to both targeted therapies and chemotherapeutic agents. Specifically, for chemotherapeutic agents, somatic mutations have not been found to correlate with drug response, complicating the administration of these to patients who are most likely to respond to them.

Finally, association studies have contributed to the identification of many SNPs with a link to cancer, but the gap between genetic variation and clinical practice is long from being breached. Most of the effect sizes observed are too low to be used for clinical screening tools, and it has proved difficult to fine-map the associations in order to identify functional variants. Nonetheless, the wealth

of information that has become available with the completion of the ENCODE project, as well as with eQTL and evolution studies, could be the link needed for this crucial step.

The aim of this study was, therefore, to suggest alternative solutions to these problems by designing and implementing strategies which combine the identification of SNPs associating with cancer phenotypes with a detailed bioinformatics analysis to allow for further experimental validation of these candidates, to accelerate their incorporation into clinical strategies. The thesis can be divided into three themes: the first is focused on the identification of SNPs associated with the time-to-event phenotypes of cancer progression; the second on the identification of SNPs associated with chemotherapeutic response; and the third on the identification of SNPs affecting the transcriptional regulation of key cancer genes resulting in differential cancer risk.

For the first section (Chapters 3 and 4), a methodology named Variable Ranking was designed and tested, which is based on the Random Survival Forest for the ranking of candidate trait-associated SNPs. This methodology was then applied to a cohort of patients with B-cell chronic lymphocytic leukaemia for the identification of SNPs associating with cancer progression. For this purpose, the focus was on SNPs in genes whose protein products are clearly involved in cancer causation and progression, in order to reduce the number of SNPs tested in each patient cohort before applying the Variable Ranking method. Lastly, the list of candidate SNPs was filtered-out using publicly available ENCODE data and eQTL studies, in order to determine the SNPs for which validation would be sought in a replication cohort.

In the second section (Chapter 5), the NCI60 panel of cell lines was analysed for the identification of SNPs associating with chemotherapeutic response, and data from the Cancer Genome Project (CGP) and the Genomics of Drug Sensitivity in Cancer (GDSC) database are utilised as a replication cohort. Further characterisation of one of the candidate SNPs is then pursued, both by utilising expression data for the cell lines and experimental approaches.

Finally, in the third section (Chapters 6 and 7), efforts were made to assess the potential roles of cancer-associated SNPs in affecting the transcriptional regulation of key cancer genes. This was accomplished, firstly, by focusing on SNPs in E-box binding motifs, which have been shown to be very sensitive to single base pair changes. Evidence of functionality in the ENCODE data and eQTL studies was then used to further support the hypotheses derived and the results were followed up experimentally. Secondly, the possibility that SNPs in transcription factor binding sites could underlie the associations observed in cancer GWAS was explored.

2 Materials and methods

2.1 Patient cohorts and cell lines

2.1.1 B-cell Chronic Lymphocytic Leukaemia cohort

Discovery cohort The discovery cohort comprised 244 individuals with B-cell chronic lymphocytic leukemia (B-CLL) collected from two sources: 106 CLL cases that reported to the Royal Berkshire Hospital (RBH) in Reading between 1992 and 2010; and 138 cases from the UK multicentre CLL4 trial series, a description of which can be found in Catovsky *et al.* (2007). The selection of the 106 RBH cases was based on differing disease stability over time and/or genomic characteristics, whilst the 138 individuals from the CLL4 trial series were chosen as individuals with deletion of ATM (del11q) (40 cases) or randomised to the Fludarabine plus Cyclophosphamide arm of the trial (98 cases). Genotype data for all 244 cases from the discovery cohort was determined using the Affymetrix Genome-Wide Human SNP Array 6.0 by Dr Jonathan Strefford of the Cancer Genomics Group of the University of Southampton. Twenty-one samples (8.6%) with low quality control criteria (Contrast$QC < 0.4$) were excluded from further analysis, which may be reflecting the fact that some samples were historical samples with a potentially worse quality DNA.

Replication cohort The replication cohort comprised 601 individuals with B-CLL: 256 randomly selected from the Royal Berkshire Hospital (RBH); and 345 cases randomly selected from the UK CLL4 trial series. It should be noted that,

although the sources of the discovery and replication cohorts are the same, the individuals do not overlap between the two. The cases from the replication cohort were genotyped using Taqman genotyping from Applied Biosystems (by an experimentalist in our group). DNA samples were not available in the same quantities for all patients so it was not possible to genotype all SNPs (apart from rs3806318 that was first on the list because it had the lowest p-value in the discovery cohort) for all patients. Therefore, due to sample availability and genotyping quality control criteria, the SNPs were genotyped in different subsamples of patients. SNP rs3806318 was genotyped in 578 samples (247 from RBH and 331 from CLL4, 32 undetermined, 96.1% genotyping success rate), rs4693051 in 522 samples (246 RBH; 276 CLL4, 43 undetermined, 92.4% genotyping success rate), rs11740785 in 506 samples (247 RBH; 259 CLL4, 15 undetermined, 97.1% genotyping success rate), rs6690837 in 489 samples (227 RBH; 262 CLL4, 20 undetermined, 96.1% genotyping success rate) and rs1550871 in 478 samples (229 RBH; 249 CLL4, 29 undetermined, 94.3% genotyping success rate). Similarly to the discovery cohort, the relatively low genotyping success rates may be reflecting the fact that some samples were historical samples with a potentially worse quality DNA. However, no systematic genotyping failure was observed between the samples and therefore the results should not be biased by systematic differences.

2.1.2 Melanoma cohort

The cohort comprised 105 individuals with skin melanoma collected from the Churchill Hospital, Oxford, UK, between 1989 and 2008. There were 4 individ-

uals with missing data for metastatic outcome and these were excluded from the analysis for time-to-metastasis.

2.1.3 NCI60 panel

The Human Tumour Cell Line Assay (NCI60) The National Cancer Institute 60 (NCI60) human tumour cell line assay contains 59 cell lines from nine types of cancer. These include leukaemia, melanoma and cancers of the lung, colon, central nervous system, ovary, breast, prostate and kidney (Shoemaker, 2006). The mutational status of 21 important cancer-related genes (including TP53 and EGFR) and growth responses to 132 compounds (GI50) were retrieved from the National Cancer Institute/NIH Developmental Therapeutics Program (DTP) database (`http://dtp.nci.nih.gov/`).

Genotyping The genotyping for the NCI60 panel of cell lines was performed by another member of the lab, Philip Grochola, using a customised Illumina GoldenGate Genotyping Assay, a medium-throughput, 96-well plate assay. The assay comprised 671 SNPs with a *MAF* over 20%, and a haplotype coverage of 77.2% of 114 genes of the p53 network (Table 2.1). A high *MAF* was chosen due to the very low number of cell lines in this panel.

2.1.4 The Cancer Genome Project (CGP) panel

The genotype data for 974 samples was retrieved from the Cancer Genome Project Archive (`http://www.sanger.ac.uk/genetics/CGP/Archive/`) with the Affymetrix

AKT1	CSE1L	GSK3B	MAP2K4	PIK3R1	SRC
AKT2	CSNK2A1	GTF2H1	MAP3K5	PIK3R2	TCF7L1
AKT3	CTNNB1	GTF2H3	MAP4	PIK3R3	TCF7L2
APAF1	DUSP16	GTF2H4	MAPK1	PMAIP1	TDG
APC	DYRK2	hCG_1789827	MAPK10	PPP1R13B	TNFRSF10B
APEX1	E2F1	HDAC2	MAPK8	PTEN	TP53BP1
ATM	E4F1	HIPK2	MAPK9	PTK2	TP53BP2
ATR	EGF	HRAS	MDM2	PXN	TP63
BAD	EGFR	IGF1	MDM4	RAF1	TP73
BAX	EP300	IGF1R	MMP2	RB1	TSC1
BCAR1	ERCC2	ILK	MTA2	RBBP7	TSC2
CASP3	ERCC3	ITGB1	NFKB1	RCHY1	USP7
CASP9	ESR1	JMY	NUMB	REL	VEGFA
CBL	ESR2	KDR	P2RY5	RELA	XPC
CCNG1	FAS	LEF1	PCNA	RELB	
CDK2	FHL2	LOC643932	PDPK1	RFWD2	
CDKN1A	FLT1	LOC643983	PIAS1	RHEB	
CDKN2A	FOXO3	LRDD	PIK3CA	SHC1	
CHEK2	FRAP1	MAP2K1	PIK3CB	SOS1	
CREBBP	GADD45A	MAP2K2	PIK3CD	SOS2	

Table 2.1: Table of the 114 genes of the p53 network included in the analysis.

SNP6.0 array. The samples contained 729 cancer cell lines and 245 normal samples. The cancer cell lines consisted of samples from 31 primary tissue types and one that was not specified (Table 2.2).

Genomics for Drug Sensitivity in Cancer (GDSC) The IC50 drug response to 138 anticancer agents was retrieved from the Genomics for Drug Sensitivity in Cancer (GDSC) database (`http://www.cancerrxgene.org/`). Drug response was established using a fluorescence-based cell viability assay (Garnett *et al.*, 2012; Yang *et al.*, 2013). Of the 138 agents, 122 were targeted agents, 13 were cytotoxic chemotherapeutics (2 alkylating agents, 4 antimetabolites, 4 antimitotic agents, 1 topoisomerase I inhibitor and 2 topoisomerase II inhibitors) and 3 were 'other' (unknown target). A different number of cell lines was screened per drug (mean:

Primary Site	Number of Samples
adrenal gland	2
autonomic ganglia	37
biliary tract	6
bone	33
breast	45
central nervous system	59
cervix	12
endometrium	10
eye	1
gastrointestinal tract (site indeterminate)	1
haematopoietic and lymphoid tissue	127
kidney	21
large intestine	39
liver	10
lung	150
oesophagus	22
ovary	22
pancreas	17
placenta	2
pleura	6
prostate	5
salivary gland	1
skin	52
small intestine	1
soft tissue	19
stomach	21
testis	3
thyroid	12
upper aerodigestive tract	22
urinary tract	18
vulva	3
Non-specified (carcinoma)	1
Total number of samples	**729**

Table 2.2: Primary tissue sites for the cancer cell lines from the CGP archive.

525, ranging from 329 to 668).

2.1.5 Cancer Cell Line Encyclopedia (CCLE) data

The mRNA levels and CNV (processed) data for the relevant 15 genes (discovered from the analysis of the NCI60 data in Chapter 5) were retrieved from the Can-

cer Cell Line Encyclopedia database (`http://www.broadinstitute.org/ccle/`). Data was available for 1,041 cell lines (994 cell lines for the CNV data and 1,036 for the mRNA levels).

2.1.6 Melanoma cell line panel

A panel of 36 melanoma cell lines (Table 2.3) was assembled by another member of the group, Anna Grawenda. Measurement of the mRNA levels and the genotyping of the SNPs (analysed in Chapter 6) was performed by Anna Grawenda and Jorge Zeron.

501	IGR 39	Me235	Sk mel 14	Sk mel 26	T333A
1205 Lu	LND1	Me248.3	Sk mel 17	Sk mel 28	UACC 257
Colo 783	LOXIMVI	Me260	Sk mel 19	Sk mel 37	UACC 62
Colo 800	M14	Me300	Sk mel 2	Sk mel 5	WM115
Colo 853	MALME 3M	MM0117	Sk mel 21	Sk mel 7	WM1789
IGR 37	MDA MB 435	MM031	Sk mel 23	T135 3A	WM278

Table 2.3: Table of the melanoma cell lines included in the analysis.

2.2 Statistical methodology

2.2.1 Random Survival Forest

Random Survival Forest (RSF) is a tree ensemble method for data with right censoring, i.e. some patients' event times are not recorded either because they are lost to follow-up or because the events occur after the study follow-up time ends (Ishwaran *et al.*, 2008). It was based on the idea of the Random Forest (RF) of Breiman (2001), which was proved to be a very accurate classifier with the advantage of handling high-dimensional data under model-free assumptions.

Briefly, the RSF algorithm comprises the following steps:

- Draw a number of bootstrap samples from the original dataset, and for each sample grow a full size tree.
- On each node of the tree, a random subset of predictors is tested, and the variable that maximises the survival difference of the daughter nodes - based on a survival splitting criterion - is chosen.
- For each individual, an ensemble cumulative hazard estimate can be calculated and the out-of-bag (OOB) error rate evaluated.

The key points of the algorithm for handling survival data is the use of appropriate splitting rules and a way for computing the error rate of the RF in the survival setting. The RSF has four different options for survival rules, all of which aim to maximise the node separation in terms of survival differences. The 'logrank' rule and the 'logrankscore' rule were shown to be the ones with best predictive performance, followed by the 'conservation of events' splitting and the 'approximate logrank' splitting (Ishwaran and Kogalur, 2007).

The algorithm estimates a cumulative hazard function for each node h by Nelson-Aalen estimator:

$$\hat{H_h}(t) = \sum_{t_{(i,h)}<t} d_{(i,h)}/Y_{(i,h)}$$

where $d_{i,h}$ is the number of deaths at t_i and $Y_{i,h}$ the number of individuals at risk at t_i. The cumulative hazard function for each individual is then estimated

by averaging the cumulative hazard functions of all the terminal nodes that the individual belongs to.

The OOB estimate for an individual is computed by averaging the cumulative hazard functions of the trees for which the individual has not been used for their fitting. The OOB error for the forest is then given by comparing the hazard rates of all possible pairs of individuals, and scoring the predictive power of the forest for each pair. The OOB error is an important estimator of the predictive value of the RSF. An error rate of 0.5 indicates a model that does no better than random guessing, and an error rate of 0 suggests perfect accuracy by the RF.

For variable selection, the most commonly used feature of the RSF is the variable importance (VIMP) measure, for assessing the predictive value of each variable in the forest. Given a variable X, there are two options for the VIMP estimation. The 'randomsplit' option assigns a random daughter node to each OOB case every time X is the splitting variable, and estimates the average cumulative hazard, which is then used for the random OOB' estimation. The 'permute' option permutes the X variable in the OOB data and calculates the OOB'. In both cases, the VIMP is then the difference between the OOB' and the original OOB.

Ishwaran *et al.* (2010) also introduced the concept of the minimal depth for assigning an importance measure to the variables. The idea was based on the notion that the smaller the distance of a splitting variable to the root, the more influential/predictive the variable would be. However, this concept is only suitable for the settings that the p (number of variables) is large, but does not dominate n (number of samples). Otherwise the trees cannot be built deep enough for an

appropriate variable selection.

An extension of the RSF is a regularisation algorithm for variable selection (varSel function) (Ishwaran *et al.*, 2010). There are two variants of this algorithm, one based on the VIMP measure and one on the newly introduced concept of minimum depth. In more detail, the two options available for the implementation of the variable selection algorithm are as follows: using all the variables and all the data for building the forests, the minimum depth (md) method; or using a K-fold Monte Carlo validation, in which many forests are built (nrep times), each using a subsample of the data, the variable hunting (vh) method. For the first option, the variables are selected using an adaptive threshold on their minimal depth. For the second option, each forest is built using a subsample of the data (based on the K-fold size) and random sample of the variables (mtry), with weights estimated from an initial forest built on the training data. For each forest, the significant variables are selected using a forward selection based on the joint importance measure (VIMP) of the top variables, ranked either according to their minimal depth (vh method) or their VIMP (vhVIMP method). The final variables are selected based on their frequency (rel.freq) of importance in the nrep forests. The varSel function of the RSF is used throughout the analyses in this study, to take advantage of the improved stability of the forests provided by the nrep iterations.

2.2.2 Additional association analyses

Two additional methods for survival analysis were used throughout Chapters 3-6. The log-rank test (also known as the MantelCox test) was used extensively through-

out the results to identify associations with survival phenotypes (Mantel, 1966). It is a non-parametric test appropriate for right-censored data with non-informative censoring, such as the ones analysed here. The Cox proportional hazards models were used for adjusting for potential confounders and modelling the effects (Cox, 1972). Typically, two nested models were applied to the data, with and without the covariate of interest, and the models were compared using a likelihood ratio test. The Wald test and the likelihood ratio test give asymptotically equivalent results. However, the latter was preferred because of its property of being invariable under monotonic transformations of the data and because it does not use an approximation of the standard error of the effect (Klein and Moeschberger, 2003).

The t-test, Wilcoxon rank sum test and the Jonckheere trend test were also used extensively for non-censored, quantitative data. The t-test was used for testing for equality of means in normally distributed outcomes. The Wilcoxon rank sum test (also known as the Mann-Whitney U test) was applied to test for differences in populations as a non-parametric equivalent to the t-test (Wilcoxon, 1945). Finally, the Jonckheere trend test (or Jonckheere-Terpstra test) was used to test for ordered alternative hypotheses (Jonckheere, 1954; Terpstra, 1952).

2.2.3 Natural selection analysis

The integrated haplotype score (iHS) was used to test for a natural selection signal in the European populations (Voight *et al.*, 2006). The iHS statistic is an extension of the statistic of Sabeti *et al.* (2002), the extended haplotype homozygosity (EHH). The EHH denotes the probability that two random chromosomes with the haplo-

type of interest are identical by descent. The (unstandardised) iHS was estimated using:

$$iHS = log(iHH_A/iHH_D)$$

where iHH_A is the integrated EHH (area under the EHH curve) for the ancestral allele and iHH_D for the derived allele. When the rate of decay is similar between the two alleles the iHS is ≈ 0.

2.3 Statistical applications

2.3.1 Simulations analysis

Simulations of genetic data Genotype SNP data was simulated using HAPGEN v2 (Su *et al.*, 2011). The haplotypes of HapMap3 CEU were used as a reference population. Although HapGen is designed to create populations for cases-control studies, no disease SNP was used in this implementation so both samples of cases and controls were treated as one population. Two different populations were generated with 200 and 300 individuals, respectively, each with 100 simulations.

Selection of SNPs Only SNPs belonging to or in close proximity (10kb on either side) to the genes of the 15 cancer-related pathways, as defined by the Kyoto Encyclopaedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000), were retained for the simulation analysis. From these SNPs, the ones with low minor allele frequency for the CEU population ($MAF < 0.1$) were excluded, to test the performance of the methodology in SNPs of similar MAF as the ones to be used

in the B-CLL analysis of Chapter 4. A total of 35,199 SNPs were included in the simulation analysis.

Simulations of phenotypes Time-to-event phenotypes were simulated under several scenarios: different sample sizes (200 and 300); different effect sizes (1.5 and 2 multiplicative effect on the hazard ratio); different percentages of censoring (0 and 20%); and the addition of effects of two factor covariates acting as prognostic factors. For all scenarios, survival times and censoring times were simulated under the exponential model using the package prodlim (R package by Gerds). For all models the effect of the causative SNPs was additive on the log-hazard. For each scenario, 4 SNPs were simulated to have an effect on survival time (Figure 2.1). The SNPs were selected using SNAP (Johnson *et al.*, 2008) (`http://www.broadinstitute.org/mpg/snap/`), so that, in CEU, two SNPs would be in strong LD with exactly 4 other SNPs (with $r^2 > 0.8$ to 4 SNPs and $r^2 < 0.5$ to the remaining surrounding SNPs) and two would not be in strong LD with any other SNPs (with $r^2 < 0.5$ to all surrounding SNPs). For each of these groups, one SNP was of low minor allele frequency (0.1-0.2), and one of high minor allele frequency (0.4-0.5). The scenario of having a high D' but low r^2 values was not considered in these simulations since only common SNPs were assumed to be the causative SNPs.

Selection of tag SNPs The tag SNPs for the implementation of the RSF were selected using the algorithm of H-clust (Rinaldo *et al.*, 2005). No minor allele frequency filter was set and a cut-off value of 0.8 was used for finding clusters of

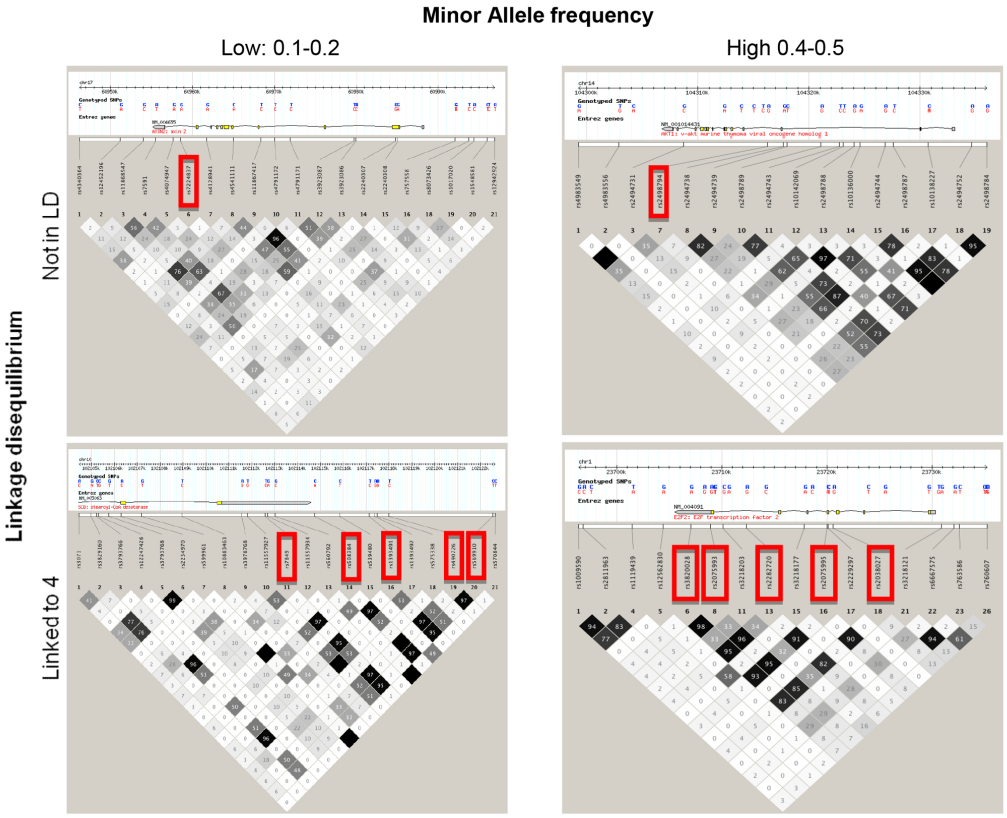


Figure 2.1: Four causative SNPs for the simulations. A combination of two LD structures (for the causative SNP with its surrounding SNPs) and two *MAF* windows was used for the selection of the SNPs. Two SNPs are in strong LD with exactly 4 other SNPs and 2 are not in LD with any other SNPs (in the CEU population). For each LD group, a SNP is of low *MAF* (0.1-0.2) and one of high *MAF* (0.4-0.5).

SNPs.

Application of the simple RSF The performance of the varSel function was also compared to the application of the simple RSF (rsf function), referred to as simple RSF in Chapter 3. The default values for all parameters apart from the number of trees were used for all applications of the rsf function. Taking into consideration the low sample size, the ntree parameter was increased to 10,000 trees.

2.3.2 B-CLL cohort analysis

Imputation IMPUTE v2 was used for imputation of the data (Howie *et al.*, 2009). The imputation was done using two reference panels, the 1000 genomes pilot 1 data (June 2010, Build 36 CEU haplotypes) and the HapMap data (HapMap3, release 2, Build 36, all minus 1000 genomes pilot 1 CEU haplotypes). Only SNPs that were originally genotyped and with imputation certainty higher than 0.8 were included in the analysis. Contrary to other GWAS approaches, imputation in this case was not done for inference of the un-genotyped SNPs. The aim was to minimise the missing genotypes in the data in order to increase the power.

SNP filtering Only SNPs residing in, or in close proximity to (10 kb on either side), the 1,208 genes of the 16 pathways associated with cancer and B-CLL, as defined by the KEGG database, were tested for associations (Table 2.4). After filtering for minor allele frequency ($MAF > 10\%$), analysis was performed on 19,122 SNPs. No exclusions were made for Hardy-Weinberg equilibrium because none of the SNPs were significantly in disequilibrium after Bonferroni correction. As discussed in the introduction, Bonferroni correction is the most commonly used approach for these types of analyses, but it is not always ideal due to the linkage disequilibrium between the SNPs. Therefore, association results of SNPs with weaker evidence of Hardy-Weinberg disequilibrium should be interpreted with caution.

Pathway	Number of genes	Number of SNPs
PPAR signalling pathway	68	787
MAPK signalling pathway	258	5123
ErbB signalling pathway	84	2499
Cytokine-cytokine receptor interaction	254	2353
Cell cycle	121	973
p53 signalling pathway	67	563
mTOR signalling pathway	49	691
Apoptosis	81	926
Wnt signalling pathway	146	2321
TGF-beta signalling pathway	84	968
VEGF signalling pathway	76	1068
Focal adhesion	193	4697
ECM-receptor interaction	83	2471
Adherens junction	72	2439
Jak-STAT signalling pathway	147	1433
B-cell receptor signalling	73	1319
Total	1,208	19,122

Table 2.4: KEGG cancer-associated pathways and the numbers of genes and SNPs included in this study.

2.3.3 NCI60 analysis

From the 673 SNPs included in the Illumina assay, 33 were excluded due to unsuccessful genotyping (or having been called monomorphic) and 56 due to low p-values for the HWE test (after Bonferroni correction). In addition, of the 132 chemotherapeutic agents, 6 were excluded because of extremely low variability between the cell lines (more than a third of the cell lines required the same amount of drug concentration). Therefore, the analysis was performed on 584 SNPs and 126 chemotherapeutic agents.

2.3.4 CGP analysis

Genotype calling Genotype quality control was performed using the Affymetrix Power Tools 1.15 (apt), and 66 cell lines were excluded from any further analysis

because they failed on quality control criteria ($contrastqc < 0.4$). The genotype calling for the cell lines was performed using the birdseed algorithm, and genotype calling was applied per tray of samples to minimise batch effects (apart from trays 11 and 13, which were merged because the number of samples per tray was too small to be analysed separately). The call rates for the cancer samples were all higher than 85% (minimum 88.49%) so no exclusions were made at this stage, with the aim to regenotype any potential hits with low call rate.

2.3.5 Software

Most of the above methodologies were performed in R software (`http://www.r-project.org/`) with the use of the randomsurvivalforest (Ishwaran *et al.*, 2008; Ishwaran and Kogalur, 2007), prodlim (Gerds), survival (Therneau) and survSNP () packages.

2.3.6 Visualisation software

All the haplotype plots in this thesis were produced using Haploview 4.2 (Barrett *et al.*, 2005; Barrett, 2009).

2.4 Bioinformatics analysis

2.4.1 Bioinformatics filter

The top SNPs from the RSF were investigated, in order to identify whether the SNPs or their proxies (with correlation coefficient $r^2 > 0.8$) lay in potential regula-

tory regions. This was ascertained in two respects: firstly by determining whether SNPs were in 5′UTR, 3′UTR or exonic regions; and secondly whether they resided in regions that were associated with enhancer or promoter activities. The former data, which was the 'consequence type' of the SNPs, was downloaded from biomaRt (`http://www.biomart.org/`).

For the latter, three criteria were utilised and the SNPs that fulfilled all of the criteria were further selected for replication. The criteria included evidence of DNase I hypersensitive areas, regions where multiple transcription factors have been found to bind and signs of enrichment of histone modification markers. The bed and broadPeak files were downloaded from the UCSC website (`http://genome.ucsc.edu/`, accessed 11/09/2011) and included: the DNase I Hypersensitivity Clusters of 75 cell lines (University of Washington ENCODE group, (Sabo *et al.*, 2006)); transcription factor ChIP-seq data (148 TF for 67 cell lines, from multiple ENCODE groups); and H3K4Me1, H3K4Me2, H3K4Me3, H4K20Me1, H3K27Ac, H3K9Ac and CTCF histone modification markers data (for 9 cell lines: Gm12878; H1hesc; Hmec; Hsmm; Hsmmt; Huvec; K562; Nhek; Nhlf, from the Broad Institute, (Ernst *et al.*, 2011; Bernstein *et al.*, 2005)) (build 37). The histone modifications selection was based on completeness for all cell lines and evidence of enhancer/promoter/repressor activity (ENCODE Project Consortium *et al.*, 2012). Markers signifying tight chromatin were not considered. Enrichment was assessed according to the presence of a signal in at least 2 cell lines (p-value< 0.05) in any of the histone markers. The requirement of having a signal in 2 cell lines served as a stringency criterion, since no q-values (FDR corrected

values) were provided. For the DNase I hypersensivity sites and transcription factor binding sites any signal was assessed as enrichment (data only included peaks within FDR 1%), independent of the number of cell lines in which the enrichment was observed. In addition, the conservation of the regions (PhastCons score) (Siepel *et al.*, 2005; Pollard *et al.*, 2010) was taken into account, and regions with a conservation score of above 0.5 were considered as conserved. The score (0-4) was computed as the sum of potential enrichment (0/1) of each of the above elements.

2.4.2 SNPs in E-boxes

Retrieval of the SNP and filtering The SNPs in the E-box transcription factor binding motif CANNTG, used the analysis of Chapter 6, were downloaded from the chromosome reports of the NCBI (`ftp://ftp.ncbi.nlm.nih.gov/snp/organisms/human_9606/chr_rpts/`, accessed 01/07/2010). Only SNPs residing in, or in close proximity to the 1,168 genes of the 15 pathways associated with cancer, as defined by the KEGG database, were included in the E-box pattern search (Table 2.5).

In addition, the SNPs' alleles and allelic frequencies were downloaded from NCBI (`http://www.ncbi.nlm.nih.gov/SNP/`). In order for the SNPs to be able to be followed up experimentally, only SNPs with MAF equal to or over 10% (in CEU) were selected. This was based on calculations on the estimated number of heterozygous cell lines in the NCI60 cell line panel in order to be able to perform allele specific ChIP experiments in multiple cell lines and tissues, taking into account the potential differences in the ethnicity of the cell lines. In more detail, it

Pathway	Number of genes
Adherens Junction	78
WNT Signalling	149
ECM-Receptor Interaction	84
Focal Adhesion	202
mTor Signalling	50
Cytokine-Cytokine Receptor Interaction	262
Jak-Stat Pathway	154
ErbB Signalling	86
MAPK Signalling	264
PPAR Signalling	70
Cell Cyle	119
TP53 Signalling	69
TGF-beta Signalling	90
Apoptosis	89
VEGF Signalling	72
Total number of unique genes	1168

Table 2.5: KEGG cancer-associated pathways, and the numbers of genes included in the analysis for the E-box SNPs.

was estimated that with a 10% *MAF* (in CEU) there would be approximately 10 cell lines available for a allele specific ChIP experiments to determine differential binding to various transcription factors.

Conservation scores of SNPs The conservation scores (PhastCons) for each sequence were retrieved via GVS (Genome Variation Server) (`http://gvs.gs.washington.edu/GVS137/`, accessed 27/08/2010) (Siepel *et al.*, 2005; Pollard *et al.*, 2010).

Retrieval of SNPs associated in a GWA study In order to prioritise our candidates, the SNPs of all published GWAS (`http://www.genome.gov/`, accessed 31/08/2010) were downloaded and compared to the candidate SNPs or their proxies, obtained from SNAP (Johnson *et al.*, 2008) (`http://www.broadinstitute.org/mpg/snap/`, accessed 31/08/2010). No threshold was applied for the GWAS

catalogue results so in effect the results included all published SNPs with a p-value $< 10^{-}5$.

2.4.3 SNPs in transcription factor binding sites

Identification of SNPs associated with cancer susceptibility traits A search in all published GWAS (`http://www.genome.gov/`, accessed 29/11/2010) was performed for SNPs associated with any type of cancer. The search terms were cancer, carcinoma, melanoma, tumor or tumour, neoplasm or neoplasia, leukemia, blastoma, lymphoma, sarcoma, glioma, as keywords identified from the National Cancer Institute List (`http://www.cancer.gov/cancertopics/types/alphalist`, accessed 29/11/2010), plus another 13 that were included to make the list more comprehensive (myeloma, malignant, mesothelioma, myeloproliferative or myelodisplastic, craniopharyngioma, histiocytosis, macroglobulinemia, mycosis fungoides, papillomatosis, pheochromocytoma, sezary). Only studies with a keyword included in the disease trait column were considered and 271 cancer-associated SNPs were retrieved from 74 publications.

Proxy SNPs All 5,056 SNPs in linkage disequilibrium ($r^2 > 0.8$ in the CEU population) with the trait-associated SNPs were considered as potentially causative. The proxies were obtained from SNAP (Johnson *et al.*, 2008) (`http://www.broadinstitute.org/mpg/snap/`, accessed 30/11/2010).

Retrieval of the alleles and flanking sequences The alleles of the SNPs were downloaded from biomaRt (`http://www.biomart.org/`, accessed 15/12/2010), ver-

sion Ensembl Variation 60. The flanking sequences of the SNPs (plus-minus 7 base pairs for the longest binding motif) were obtained from Galaxy (Goecks *et al.*, 2010) (`http://main.g2.bx.psu.edu/`, accessed 2/12/2010).

Extraction of minor allele frequencies and genes The minor allele frequencies for the SNPs and the genes in which they reside (or are in close proximity to - within 5kb) were downloaded from biomaRt (`http://www.biomart.org/`, accessed 15/12/2010). Only SNPs with minor allele frequencies of over 10% were studied further.

Natural Selection Analysis Integrated haplotype scores (iHS) were estimated using the package 'rehh' (Gautier and Vitalis, 2012), following the method described by Voight *et al.* (2006). Haplotype data for the CEU population was downloaded from the 1000 Genomes (phase 1) data (1000 Genomes Project Consortium *et al.*, 2012). For the calculation of the scores, only SNPs with a $MAF > 0.05$ were used. The normalisation of the scores was performed with a 0.05 allele frequency window. The fixation index statistic (Fst) values were downloaded from SPSmart (Amigo *et al.*, 2008) (`http://spsmart.cesga.es/`) using the 1000 Genomes phase I data (1000 Genomes Project Consortium *et al.*, 2012).

3 The Variable Ranking algorithm for the analysis of time-to-event phenotypes

3.1 Introduction

GWA studies have been the focus of extensive research over the last decade. Many articles and reviews have discussed the limitations of GWAS at length, but not much has been mentioned about the lack of application of GWAS on survival data. Specifically, time-to-event phenotypes such as survival and progression require a well-characterised, homogeneous cohort with similar patient diagnosis, treatments and measurements of prognostic factors. Therefore, the large sample sizes required for a GWAS are prohibitive for such applications. The statistical methods available for survival analysis in a GWAS setting also remain limited, and only some of the machine learning techniques that are appropriate for high dimensional datasets have been extended for the analysis of survival phenotypes. One of these techniques is the Random Survival Forest (RSF), which is an extension of the Random Forest (RF) developed by Breiman (2001).

RF has been applied extensively on many types of genomic data, including microarray data and GWAS (Díaz-Uriarte and Alvarez de Andrés, 2006; Statnikov *et al.*, 2008; Goldstein *et al.*, 2010; Maenner *et al.*, 2009; Ziegler *et al.*, 2007). In more detail, it has been used for the classification of cases and controls based on their genotypes, in order to select and rank SNPs with predictive importance and identify possible interactions between them (Bureau *et al.*, 2005; Schwender *et al.*, 2004; Lunetta *et al.*, 2004) and with environmental factors (Maenner *et al.*, 2009). It

has also been applied in two stage approaches, in combination with other statistical approaches/tests, such as chi-square tests and Bayesian networks (Roshan *et al.*, 2011; Meng *et al.*, 2007). Roshan *et al.* (2011) showed that applying the RF to the top SNPs from a chi-squared statistic could improve the ranks of causal SNPs, whereas Meng *et al.* (2007) used RF as a screening procedure to identify subsets of SNPs for further analysis with Bayesian networks. A number of simulations have also been performed to test the performance of the RF for various models, for example when a number of SNPs interact (Lunetta *et al.*, 2004) and for various options of the RF, such as the two available variable importance measures, the Gini index and permutation VIMPs (Strobl *et al.*, 2007; Nicodemus, 2011; Nicodemus *et al.*, 2010).

Nevertheless, RSF, an extension of the RF for survival data by Ishwaran *et al.* (2008), has not been extensively utilised. To our knowledge, no studies using the RSF for the association of SNPs with survival phenotypes have been published to date. However, one study has been reported that used the RSF with SNP data for the identification of pathways associating with survival of multiple myeloma patients (Pang *et al.*, 2011). More SNP association studies using the RSF are, therefore, warranted. Finally, most of the simulation studies performed with the RF algorithm only include up to a few thousand SNPs, which means that their results could only be applied to studies with a few candidate genes.

In this study, a methodology has been designed to rank candidate SNPs associated with survival phenotypes in disease-related pathways. The genotypes that have been generated are based on the LD patterns of a population of CEU,

and therefore reflect the true LD structure within the genes of interest. In addition, the newly introduced variable selection function of the RSF (Ishwaran *et al.*, 2010, 2011) is used, which includes another measure of importance for the variables, the minimum depth, and a K-fold subsampling variant that gives more stability to the results.

In this chapter, the four basic models under which the simulations were performed are presented, and a summary is provided of the steps of the algorithm proposed, referred to from here as Variable Ranking. Following on from this, the process under which the methodology was built is presented, and it is compared to the ranking of SNPs according to the log-rank test. Finally, the new the methodology is compared to the results provided by the Cox proportional hazards model, and the simulations are extended with two additional complex models.

3.2 Results

In this section, an algorithm is presented that aims to aid the ranking and identification of genetic variants that associate with survival phenotypes using the RSF. In order to achieve this, the genotypic data of 300 individuals was simulated based on LD patterns of the CEU population, thus creating SNP correlations based on real genetic structure. The aim of this study was to identify SNPs in cancer related genes, thus the simulations were also focused on the SNPs residing in cancer pathways as defined by the KEGG database (Kanehisa and Goto, 2000) (details in Materials and methods).

To assess the performance of the RSF in ranking the associated SNPs, 100

datasets were simulated under 4 models.

- *Model 1*: the simple model without censoring and only the 4 SNPs having an effect.
- *Model 2*: a model with 20% censoring.
- *Model 3*: a model with two added associating covariates.
- *Model 4*: a model with both 20% censoring and the two additional covariates.

To take into account varying effect sizes, two scenarios were simulated for each model.

- *Scenario 1*: of four SNPs having strong effects (additive effect of 0.7 on the log-hazard - multiplicative effect of 2 on the hazard function).
- *Scenario 2*: of four SNPS having weaker effects (additive effect of 0.4 on the log-hazard - multiplicative effect of 1.5 on the hazard function).

For each scenario, two of the effect SNPs were not linked to other SNPs and two were strongly linked to another four each (details in the Materials and methods). When analysing the results, all 12 SNPs (4 causative SNPs and 8 correlated SNPs) were considered to be true positives, in that the aim was to identify associated SNPs and not necessarily causative SNPs.

In order to examine in more detail the simulated datasets and to understand the effect of the correlation on the estimation of the coefficients, the bias and coverage of the estimators from the Cox proportional hazards models was examined

under the full model (using all 4 causal SNPs) and under 12 single predictor models for the 12 associated SNPs, using 1 predictor at a time and the additional covariates which are playing the role of the known prognostic factors for models 3 and 4. The full model using all 12 associated SNPs could not be fit for all simulations, due to the many highly correlated variables. The high level of singularity produced an error for the models. Tables 3.1 and 3.2 present for model 1 a summary of the bias and coverage of the simulations for strong (0.7 additive) and weaker (0.4 additive) effect sizes respectively. The same results for models 2-4 are found in the Appendix and are also summarised in Figure 3.1. For each covariate the bias was calculated as the mean difference of the estimated values to the true values, and the coverage as the number of the 95% confidence intervals that contain the true value. Unexpectedly, when the full model was used, the observed bias was centred around 0 and typically small, with its absolute values ranging from 0.0009 to 0.0173. Similarly, the coverage was centred around 95%, with values ranging from 93 to 97 for all covariates and both effect sizes. This verified the high accuracy of the Cox proportional hazards model when the 'true' model is fit, i.e. when all causal covariates are included in the model building.

However, when only one covariate is included in the model at a time, the coefficients are considerably biased. In more detail, the positive coefficients were consistently underestimated and the negative ones (for the SNPs where the homozygous for the major, and not the minor, allele is coded as 2) were overestimated, shrinking all estimators towards the zero. This bias is known to occur in non-linear models and in particular under the Cox proportional hazard models (Gail *et al.*,

1984; Struthers and Kalbfleisch, 1986; Bretagnolle and Huber-Carol, 1988). Moreover, this effect is highlighted by omitting covariates with large effects, as shown in Table 3.1 and Figure 3.1. These observations are also reflected in the poor coverage, with the causal SNPs having a coverage of about 65% for the strong effects and 93% for the weak effects. The coefficients for the associated SNPs were investigated under two hypotheses, one of the true scenario of them having no effect (true value equals to 0) and one of the scenario of them having an effect equivalent to their correlated counterpart (true value equals to 0.7 and 0.4 for Tables 3.1 and 3.2 respectively). Accepting their effect sizes as 0, produces a very high bias with an absolute value ranging from 0.513 to 0.525 for an effect size of 0.7 and ranging from 0.334 to 0.346 for an effect size of 0.4. In addition, their coverage is 0 for strong effects and ranging between 2 and 20 for weaker effects. Assuming that the 'true value' of the estimators for the associated SNPs are the same as their correlated counterparts, the coverage is only slightly worse than the causative SNPs for the models of weak effect, but considerably worse for the models of strong effect.

Moreover, the ranges of the p-values of the associated SNPs from the single predictor association tests (with the Cox proportional hazards model) are shown in order to give an estimate of how strongly associated the predictors are for this sample size (Figure 3.2). In this case, the minimum of the p-value is taken for each set of correlated predictors, aiming to identify whether the region was shown to be associated with the outcome. For the strong effect size (scenario 1), the high MAF SNPs passed the Bonferroni correction threshold (correcting for 35,199 SNPs) plotted in a red dashed line in most datasets (Figure 3.2a), whereas for the

Associated SNP	True Value	Causal SNP (r^2 in CEU)	FM: Bias	FM: 95% Cov.	SM: Bias	SM: 95% Cov.
Scenario 1 Effect=0.7						
rs7224837	0.7	rs7224837 (1)	-0.0042	93	-0.1916	65
rs2498794	0.7	rs2498794 (1)	0.0063	96	-0.1503	54
rs490726	0.7	rs490726 (1)	-0.0103	97	0.1712	65
rs508384	0 (0.7)	rs490726 (1)	NA	NA	-0.5250 (0.1750)	0 (65)
rs569910	0 (0.7)	rs490726 (1)	NA	NA	0.5229 (-0.1771)	0 (67)
rs1393491	0 (0.7)	rs490726 (0.97)	NA	NA	-0.5251 (0.1749)	0 (69)
rs7849	0 (0.7)	rs490726 (0.97)	NA	NA	-0.5156 (0.1844)	0 (66)
rs2075995	0.7	rs2075995 (1)	0.0009	94	-0.1539	61
rs2075993	0 (0.7)	rs2075995 (0.95)	NA	NA	-0.5226 (0.1774)	0 (47)
rs3820028	0 (0.7)	rs2075995 (0.93)	NA	NA	0.5148 (-0.1852)	0 (43)
rs2282720	0 (0.7)	rs2075995 (0.92)	NA	NA	-0.5131 (0.1869)	0 (44)
rs2038027	0 (0.7)	rs2075995 (0.90)	NA	NA	-0.5141 (0.1859)	0 (45)

Table 3.1: Bias and coverage of estimators for scenario 1 of strong effects (additive effect of 0.7 - multiplicative effect of 2). The single models were fit for all associated SNPs, whereas the full model was fit with the causative effects only. The correlation (r^2) of the associated SNPs to the causative SNPs is given in a parenthesis (all D' are 1). For the associated (but not causative) SNPs the true value in the strict sense is 0 but the value by association is also given in a parenthesis. Abbv: NA:not available, FM: Full Model, SM: Single Model

weak effects (scenario 2) only approximately 25% of the times the SNPs passed the correction (Figure 3.2b). As expected, the SNPs with the low *MAF* have high *p*-values for both scenarios and, strikingly, in the scenario of weak effects (multiplicative effect of 1.5 on the hazard function) their *p*-values almost never pass the Bonferroni correction threshold. In addition, the *p*-values for models 2 and 4 are slightly higher for both scenarios of weak and strong effects, as expected by the reduced power due to censoring. Interestingly, this is only apparent for the high *MAF* SNPs, perhaps due to the differences in power introduced when the very

Associated SNP	True Value	Causal SNP (r^2 in CEU)	FM: Bias	FM: 95% Cov.	SM: Bias	SM: 95% Cov.
Scenario 2 Effect=0.4						
rs7224837	0.4	rs7224837 (1)	0.0173	97	-0.0398	94
rs2498794	0.4	rs2498794 (1)	-0.0072	97	-0.0511	92
rs490726	0.4	rs490726 (1)	0.0060	94	0.0514	95
rs508384	0 (0.4)	rs490726 (1)	NA	NA	-0.3468 (0.0532)	19 (94)
rs569910	0 (0.4)	rs490726 (1)	NA	NA	0.3445 (-0.0555)	20 (94)
rs1393491	0 (0.4)	rs490726 (0.97)	NA	NA	-0.3461 (0.0539)	18 (91)
rs7849	0 (0.4)	rs490726 (0.97)	NA	NA	-0.3396 (0.0603)	18 (92)
rs2075995	0.4	rs2075995 (1)	-0.0013	94	-0.0456	92
rs2075993	0 (0.4)	rs2075995 (0.95)	NA	NA	-0.3426 (0.0574)	2 (92)
rs3820028	0 (0.4)	rs2075995 (0.93)	NA	NA	0.3359 (-0.0641)	2 (91)
rs2282720	0 (0.4)	rs2075995 (0.92)	NA	NA	-0.3391 (0.0609)	2 (89)
rs2038027	0 (0.4)	rs2075995 (0.90)	NA	NA	-0.3343 (0.0657)	2 (86)

Table 3.2: Bias and coverage of estimators for scenario 2 of weaker effects (additive effect of 0.4 - multiplicative effect of 1.5). Similarly to scenario 1, the single models were fit for all associated SNPs, whereas the full model was fit with the causative effects only. The correlation (r^2) of the associated SNPs to the causative SNPs is given in a parenthesis (all D' are 1). For the associated (but not causative) SNPs the true value in the strict sense is 0 but the value by association is also given in a parenthesis. Abbv: NA:not available, FM: Full Model, SM: Single Model

low numbers of the homozygous (less than 5 individuals) were grouped with the heterozygous, which occurs only for low MAF SNPs.

3.2.1 The Variable Ranking algorithm and its comparison to the log-rank test

The algorithm proposed here, Variable Ranking, can be summarised in 4 steps. In the first step, the log-rank test was applied to all SNPs, and only those with signs of association were selected for step 2. In step 2, tag-SNPs (with $r^2 > 0.8$)

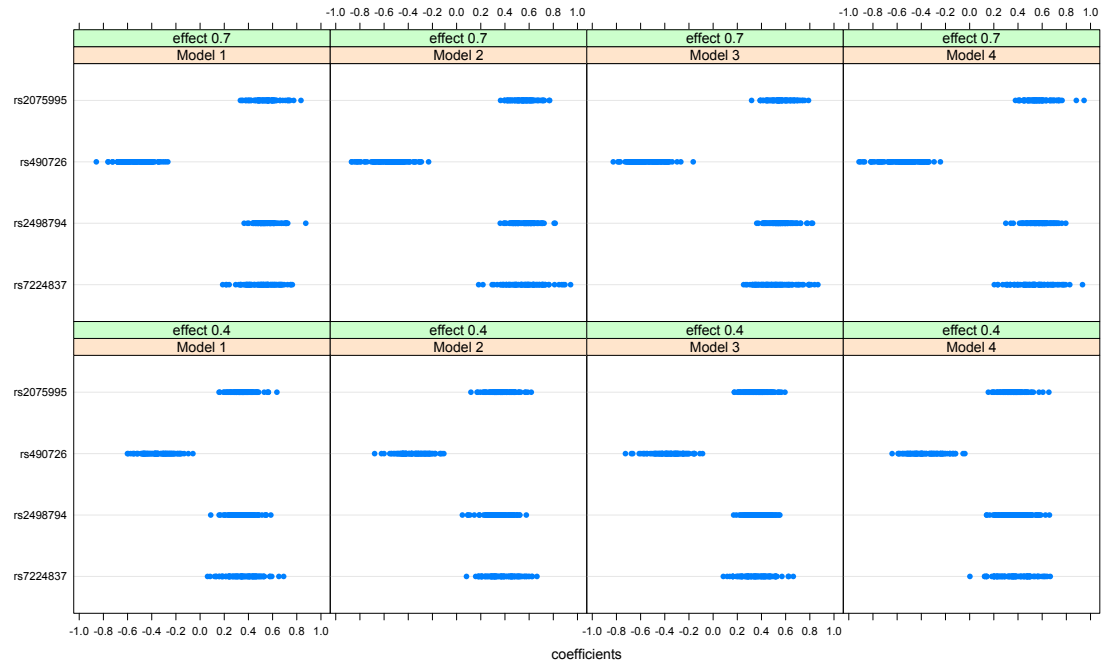
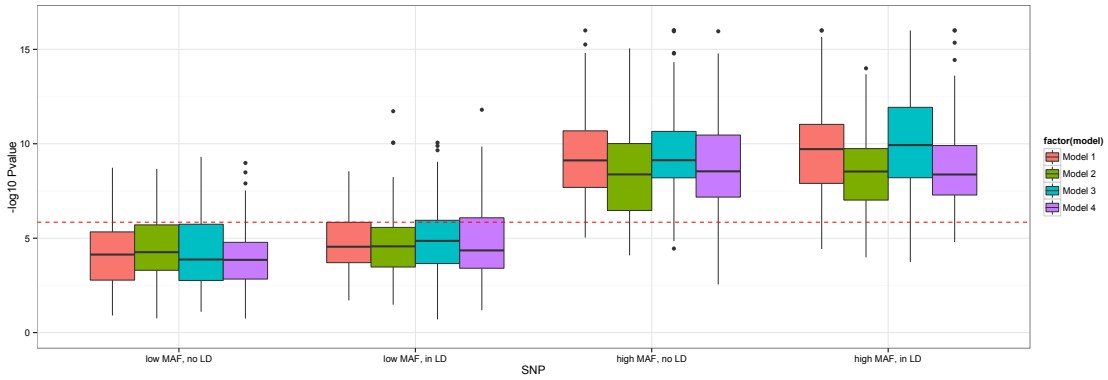


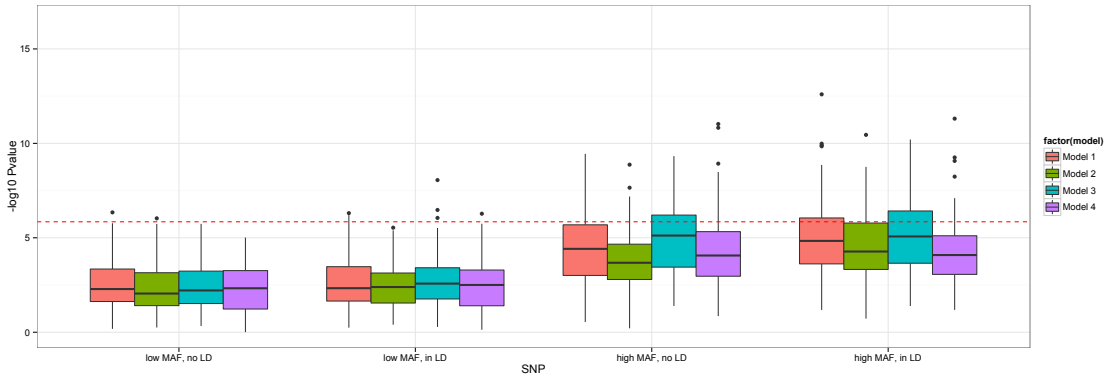
Figure 3.1: Estimated values of coefficients for both scenarios of effects under the single models. As also shown in the tables 3.1 and 3.2, the estimators for the coefficients of the strong effect are considerably biased. SNP rs490726 has negative coefficients because the homozygous for the major allele are coded as 2.

were retrieved from the subset of the potentially associated SNP. In step 3, the varSel of the RSF was applied using the *vhVIMP* method with parameters K=2 and *ntree*=1500. Finally, in step 4, the SNPs were ranked using two features of the varSel, the relative frequency (*rel.freq*) of the SNPs and their variable importance (*vimp*) measure (details of both measures are listed in the Materials and methods).

The following section outlines how the algorithm was constructed, supporting the choice of parameters for each step. For the choice of thresholds and parameters, the results are presented for the simplest model as a reference point (model 1) unless stated otherwise. Nonetheless, similar trends were observed for the remaining models.



(a)



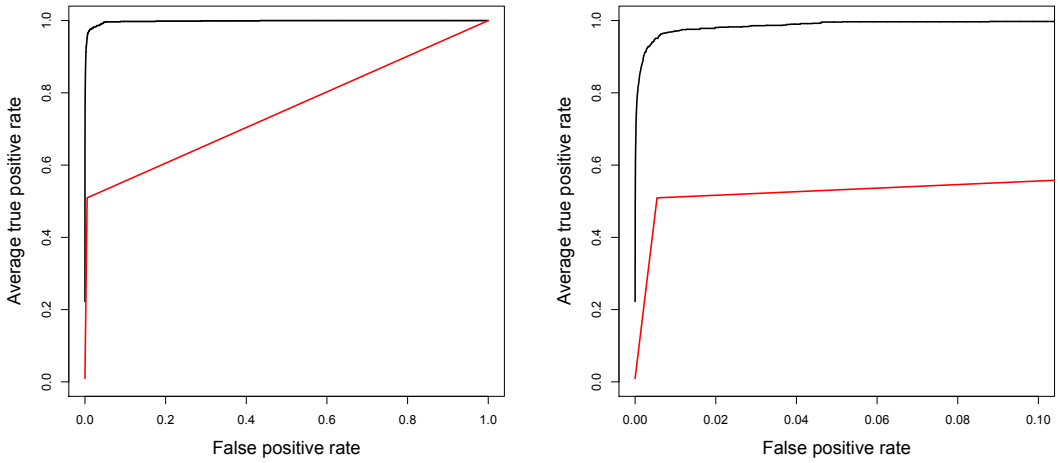
(b)

Figure 3.2: Boxplots of the p-values from the single predictor Cox proportional hazards models for each of the 4 sets of associated SNPs under the 4 models for scenario 1 (strong effects) (a) and scenario 2 (weak effects) (b). For the linked groups of SNPs, the SNP with the minimum p-value was selected as proxy. The red dashed line marks the threshold of the p-values if Bonferroni correction is used.

Step 1

As a first approach to identify the best performing method, the RSF (varSel function) was applied to all 35,199 SNPs, and it was compared with the p-values from the log-rank test using a sensitivity-specificity plot. However, the log-rank test outperformed the RSF for all models. An example is shown in Figure 3.3, for model 1 and effect size 0.7. The ranking that the p-values of the log-rank provided was superior to the rankings from the variable selection (varSel) algorithm of the RSF. Due to the large number of variables and low sample size, the trees were not deep enough for all the variables to be selected in the trees. Therefore, many variables could not be ranked by the algorithm, creating a straight line after the top (approximately) 300 variables had been ranked. Increasing the number of trees (*ntree*) to 1,500 (the default is 500) did not seem to have a significant effect in this case. It needs to be mentioned that due to the Monte Carlo validation method used for the varSel function (by which the forest building was repeated nrep times), the effective number of trees is larger than the reported one (by a factor of 10 in this application). If there was no computational cost, perhaps an appropriate increase of the *ntree* parameter (to an order of 50,000 trees) would have had an effect, but this was not computationally feasible with the current algorithm and resources.

In order to improve the performance of the RSF, most of the non-trait-associated/noise SNPs were removed via a pre-selection step, using the p-values of the log-rank test. Various thresholds were tried and the rankings of the SNPs for the different p-value thresholds compared using a sensitivity-specificity plot (Figure 3.4). Under certain thresholds, for some of the weak effects, some datasets



(a) (b)

Figure 3.3: Receiver operating characteristic (ROC) curves for the comparison of the rankings of the log-rank test (black line) to the variable selection algorithm (varSel) (red line), when all the SNPs are included in the varSel for model 1 and effect size 0.7 (a). The area of interest, below a false positive rate of 0.1, is enlarged (b). When all the SNPs are included for the building of the forests, the noise overwhelms the RSF and the true positive rate is much lower than the ranking that would be achieved according to the p-values of the log-rank test.

contained no associated SNPs under certain thresholds, and so were removed for the application of the ROC curves. Figure 3.4 illustrates that the performance of the Variable Ranking improved when applied to subsets of SNPs with a smaller proportion of noise, for three p-value thresholds (0.005, 0.01 and 0.05). Since the number of SNPs analysed differed for each group, the sensitivity-specificity plot of the log-rank test for each subset of SNPs was added to each plot, for comparison purposes. While the rankings of the log-rank test remained the same over all plots, the sensitivity-specificity plot of the Variable Ranking lost power by increasing the number of SNPs used in each run. This was in line with the results from Winham *et al.* (2012), where it was noted that the probability of detection of the causal SNPs declined as the number of predictors increased, and Roshan *et al.* (2011),

where the RF was applied to significant SNPs from a chi-squared test using different thresholds. From here on, the results are presented for effect 0.7, unless stated otherwise, for the comparison of the remaining parameters.

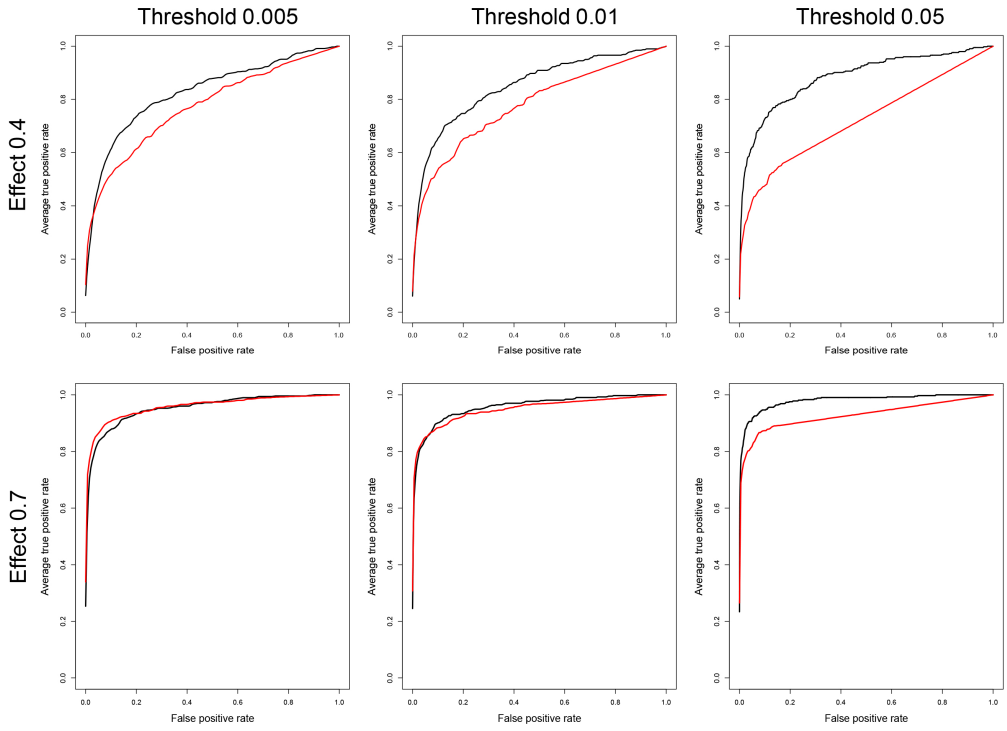


Figure 3.4: Comparison of different thresholds of filtering for step 1 of the methodology. Ranking of SNPs according to the p-values of the log-rank test (black lines) and after the Variable Ranking has been applied to the tag SNPs of the top 289 SNPs (red lines) (p-value = 0.005), 502 SNPs (p-value = 0.01) and 2,030 SNPs (p-value = 0.05) (values based on the mean of the 100 simulations).

Step 2

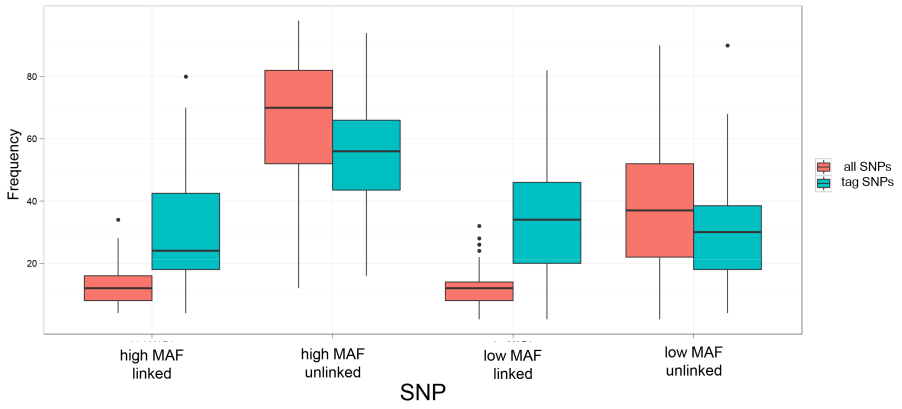
LD between SNPs has previously been noted to affect the performance of the RF (Walters *et al.*, 2012; Nicodemus and Malley, 2009; Meng *et al.*, 2009; Nicodemus *et al.*, 2010). In more detail, the RF has been shown to select uncorrelated predictors more frequently than correlated ones, when all the splits are considered (Nicodemus and Malley, 2009; Nicodemus *et al.*, 2010). This observation was repli-

cated in our simulations and it was also stronger for the models with effect size 0.7, as noted in previously reports (Meng *et al.*, 2009; Strobl *et al.*, 2008; Nicodemus and Malley, 2009). In Figure 3.5, for model 1, two cases are examined, where the relative frequencies of the SNPs (Figure 3.5a) and their ranking according to the *vhVIMP* algorithm (Figure 3.5b) were compared, before and after having removed the strongly linked SNPs ($r^2 > 0.8$) from the analysis. The relative frequency of the RSF denotes how frequently a variable is picked out of the 100 iterations (nrep) of the *vhVIMP* algorithm (details in the Materials and methods) and, therefore, a higher relative frequency suggests that the SNP is going to be among the top SNPs (will have a lower ranking). It is apparent that the relative frequencies of the linked SNPs increase significantly if the forests contain only tag SNPs. As a consequence, their ranking becomes lower, indicating that they are more likely to be selected as top variables. In addition, the effect of correlation between SNPs is not as strong when the effect size is moderate (Figure 3.5c).

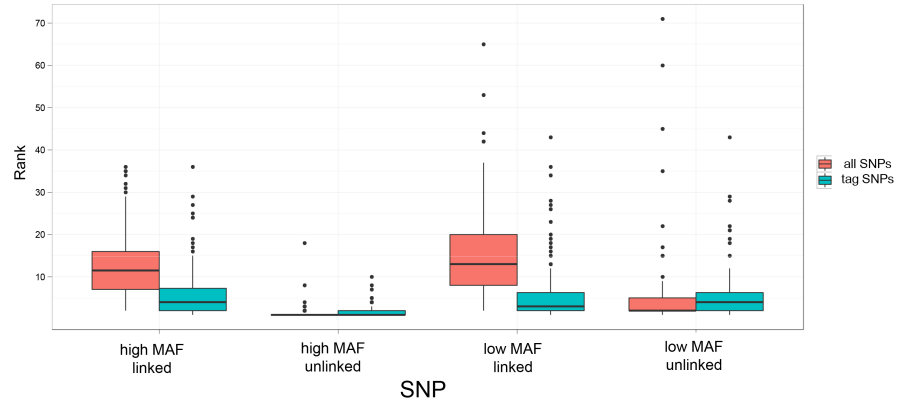
It is also notable that the frequencies of the non-linked SNPs decrease slightly. This is most likely because the signal from the other SNPs becomes stronger and, as a result, the RSF selects all four with more equal probabilities. However, the increases of the ranks of the unlinked SNPs are minimal compared to the decreases of the ranks of the linked SNPs.

Step 3

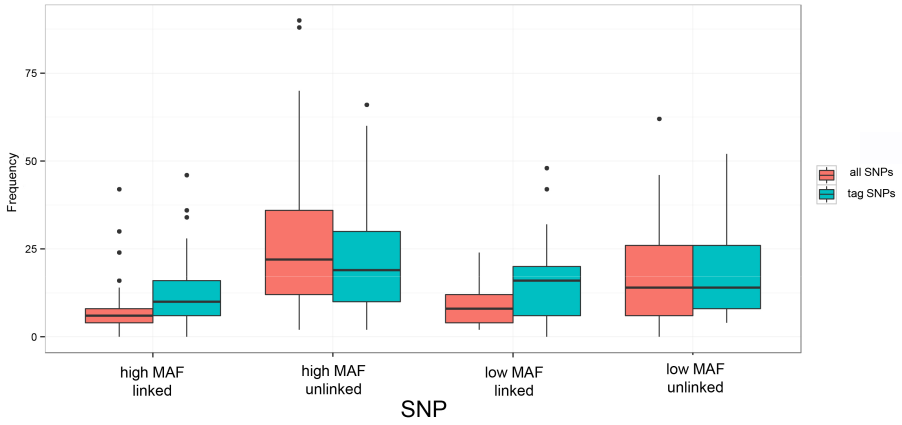
The algorithm for the RSF has an abundance of parameter values. Trying all combinations would be beyond the scope of this study, and many of the parameters



(a)



(b)



(c)

Figure 3.5: Boxplots of the relative frequency of each of the 4 causative or their linked SNPs (a) and of their ranking by the *vhVIMP* algorithm (b). The relative frequency signifies how often each variable (or its linked ones) is selected in the top variables by the varSel algorithm and the ranking how the SNPs are ordered according to the relative frequency. The relative frequency of the associated SNPs is not as severely affected when the effect sizes are weak (additive effect of 0.4) (c). For the linked groups of SNPs, the SNP with the highest relative frequency (or the minimum rank) was selected as proxy.

are known to have a minimal effect on the outcome. Nonetheless, three parameters were tested extensively for this analysis. The first was the number of trees for each forest. This parameter is of interest since the increase of the number of trees has been previously noted to improve the stability of the forest (Goldstein *et al.*, 2010). The second was the split of the data in training and testing samples for each repetition of the iterative process of the varSel function (K-fold size), which was expected to affect the ranking of the low MAF SNPs. The third was the number of variables randomly sampled at each split ($mtry$).

In Figure 3.6, the default parameters of the varSel are compared to different values for the $ntree$ and K-fold. Using sensitivity-specificity plots, it is shown that by increasing the number of trees to 1,500 (the default is 500) and using an equal split for the training and testing sample (K=2, the default being K=5), the method $vhVIMP$ performs better than the default values for either of the parameters for most models. The fact that the parameter $ntree$ does not seem to have an effect in some cases may mean that the trees have reached convergence, and so the addition of more trees does not improve their performance.

Moreover, the effect of the $mtry$ parameter was explored in Figure 3.7. However, increasing the $mtry$ did not appear to influence the algorithm, probably because the default value ($p/3$) is already large enough for this type of applications. Nonetheless, having a $mtry < p$ is important in order to increase the variability between the trees within the forest, and therefore the parameter was not tried with a larger value.

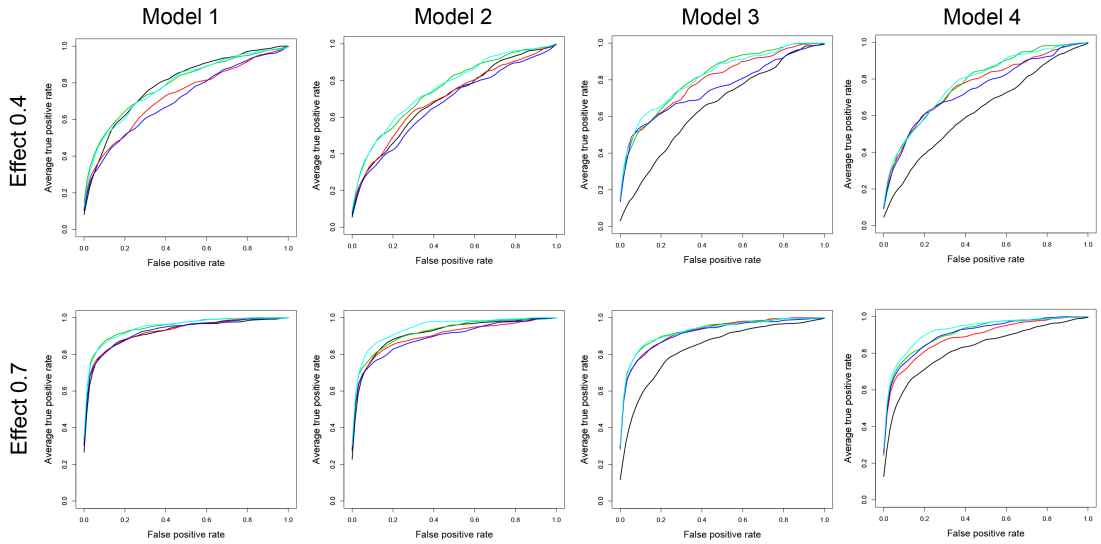


Figure 3.6: Comparison of different parameter settings for effect 0.4 and 0.7. The default values (red lines) are compared to *ntree*=1,500 (dark blue line), K=2 (green line) and *ntree*=1,500 with K=2 together (light blue line). The ranking with the log-rank p-values is also shown with a black line for comparative purposes.

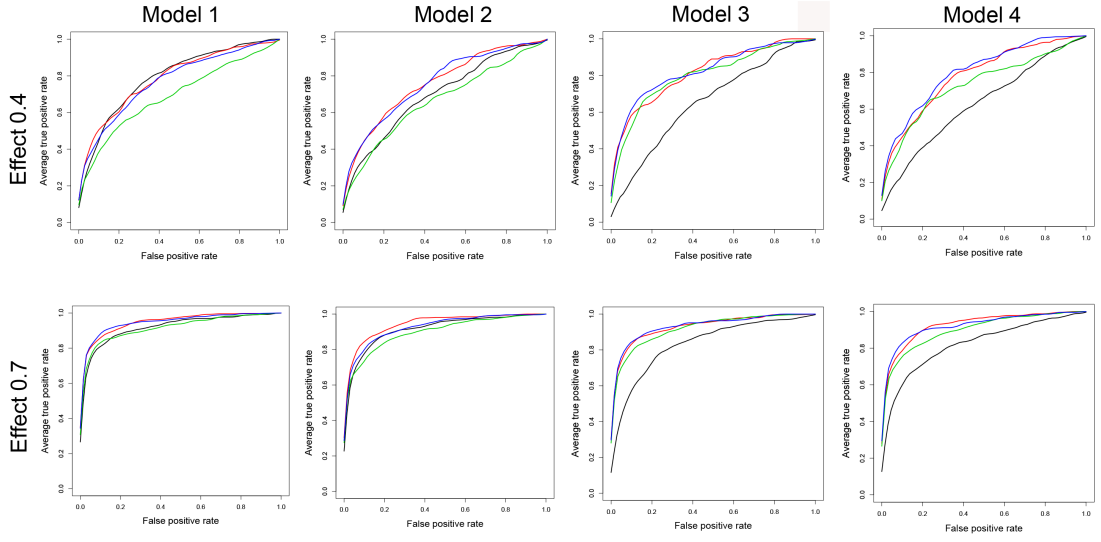


Figure 3.7: Comparison of the default settings for *mtry* and a larger *mtry* ($p/2$) for the two effect sizes, 0.4 and 0.7. The rankings from the RSF with K=2 and the default value for *mtry* (red line) are compared to K=2 and *mtry*=p/2 (blue line) and to K=5 (default) and *mtry*=p/2 (green line). For all the applications an increased *ntree* (1500) is used for additional stability. The ranking with the log-rank p-values is also shown with a black line for comparative purposes.

Step 4

In the final step of the Variable Ranking methodology, a new measure is defined for ranking the top SNPs according to the product of the ranks of the relative frequency (*rel.freq*) of the SNPs to the ranks of the variable importance (*vimp*) measure:

$$rank.vhVIMP = rank(rel.freq) * rank(vimp) \quad (1)$$

In Figure 3.8, the ranking of the SNPs as provided by the relative frequency of the SNPs (of the varSel function) is compared to the ranking gained using the measure described here, which takes into account both the variable importance and relative frequency. The new ranking measure outperforms the relative frequency alone, which underlines the fact that although the relative frequency is constructed from the ranking of the SNPs according to the variable importance measure, it does not take into account its size. Therefore, the combination of the two measures can better distinguish the associated SNPs (an example of which is shown in Figure 3.9).

Next, the Variable Ranking measure is compared to the default ranking measures under different parameter settings and methods. In the paper of Ishwaran *et al.* (2010), two variations of the variable hunting method are presented for performing the variable selection of the features of interest: the *vh* method and the *vhVIMP* method. The main difference is that the *vh* algorithm uses the minimum depth of the SNPs for the ranking, whereas the *vhVIMP* uses the variable

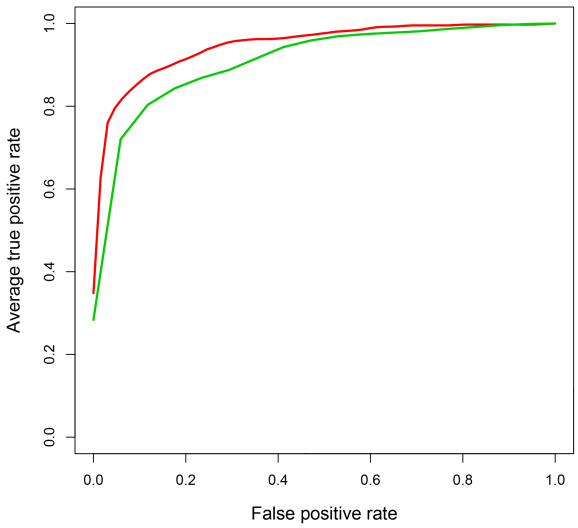


Figure 3.8: ROC curves comparing the relative frequency of the *vhVIMP* (green line) to the new measure of the products of the ranks of the relative frequency and the vimp (red line).

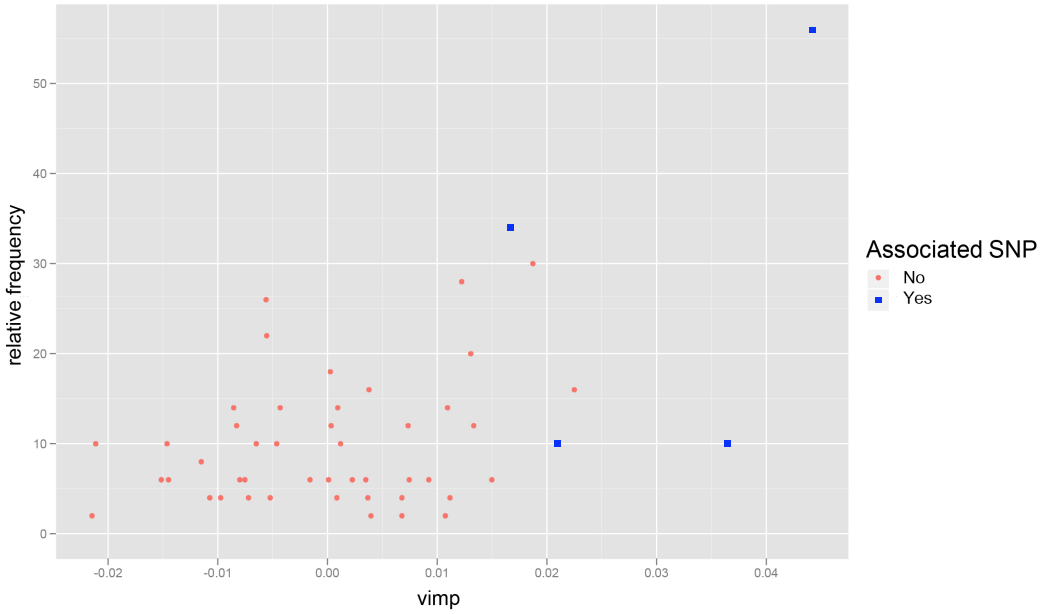


Figure 3.9: Example of an application of the *vhVIMP* algorithm. The blue squares are the associated SNPs and the red dots noise. The relative frequency of two of the associated SNPs (the ones of high *MAF*) is noticeably higher than the rest. However, for the SNPs of low *MAF* this is not the case and they would be mistakenly regarded as noise by the algorithm.

importance measure, as discussed in the Materials and methods.

Under all combinations of the parameter values, the *rank.vhVIMP* ranking measure was contrasted to the default ranking measures for the *vhVIMP* variant, as well as the default relative frequency for the *vh* method and an equivalent measure to *rank.vhVIMP* (1):

$$rank.vh = rank(rel.freq) * rank(min.depth) \tag{2}$$

This measure is the equivalent measure of *rank.vhVIMP* but for the *vh* method, which ranks the variables according to minimum depth (min.depth) and not to variable importance (vimp).

The *rank.vhVIMP* is shown to be significantly better than all the other ranking measures (a summary of which is given in Table 3.3), independently of the parameters of the RSF used (Figure 3.10). Interestingly, the equivalent of our proposed measure for the *vh* method is only better under the default K=5, whereas for K=2 the *rank.vh* under-performs.

Finally, the results of the Variable Ranking and the *rank.vh* measure were compared to the *md* method, which is recommended for settings where the number of variables does not dominate the sample size (Ishwaran *et al.*, 2010). Ranking according to the minimum depth is used for the *md* method, which is the default setting, and no K-fold is used in the settings, since it is not an option for this method. For the *rank.vh*, the same parameters were used as for the Variable Ranking. In the sensitivity-specificity plot of Figure 3.11, it is apparent that the

Ranking Measure	Description
vhVIMP	Default ranking measure of the *vhVIMP* method, according to the relative frequency.
rank.vhVIMP	New ranking measure where the ranks of the relative frequencies are multiplied to the ranks with the variable importance measure (Variable Ranking).
vh	Default ranking measure of the *vh* method, according to the relative frequency.
rank.vh	Equivalent ranking measure to *rank.vhVIMP*, where the ranks of the relative frequencies are multiplied with the ranks of the minimum depth measure.
md	Default ranking measure of the *md* method, according to the minimum depth.

Table 3.3: Brief descriptions of each ranking measure used in this Chapter.

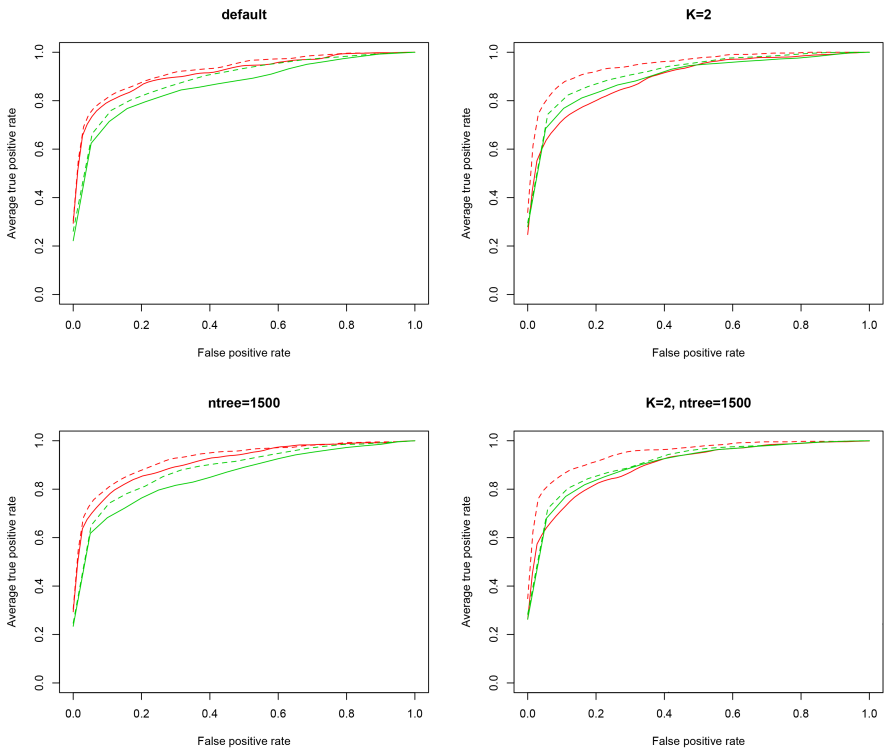


Figure 3.10: Comparison of the ranking measures and the two variable hunting methods for different parameter settings (default of the method is K=5 and *ntree*=500). The Variable Ranking (red broken line) is superior to the *vhVIMP* method with the ranking according to the default relative frequency (green broken line), *rank.vh* (red line) and the *vh* default ranking (green line) in all cases.

Variable Ranking also outperforms *md*. However, the *md* method is better than the *vh* method, showing that when the minimum depth is used for the ranking, the

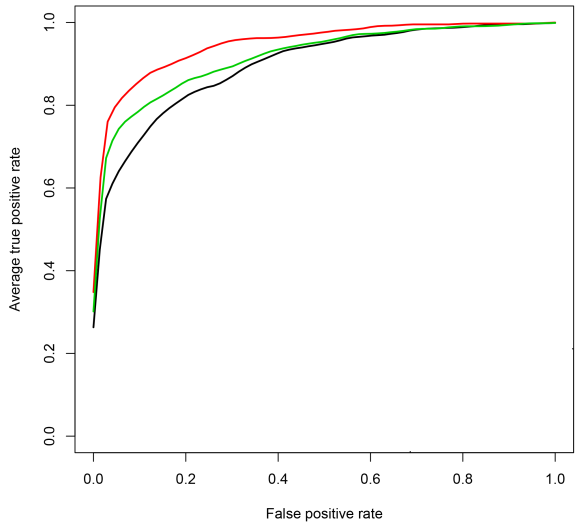


Figure 3.11: Sensitivity-specificity plot comparing Variable Ranking (red line) to the two alternative methods, the *md* method (green line) and the *rank.vh* (black line).

trees need to be deep enough for a precise minimal depth to be assigned to each variable. If the trees are too shallow, many variables are assigned to the maximum depth of the tree. Therefore, the use of a K-fold Monte Carlo validation, where the data is split between training and testing samples, is not appropriate for this setting, as it reduces the number of samples used for the building of the trees. This is in line with the previous observations (Figure 3.10), where it was shown that using a K=2 for the splitting of the data was not beneficial for the *vh* method, because it reduced the training sample.

Comparison to the ranking of the log-rank test

The Variable Ranking methodology was compared to the ranking of the SNPs by the p-values of the log-rank test, and to the ranking provided when applying the simple rsf function (see Materials and methods for details), for the four different

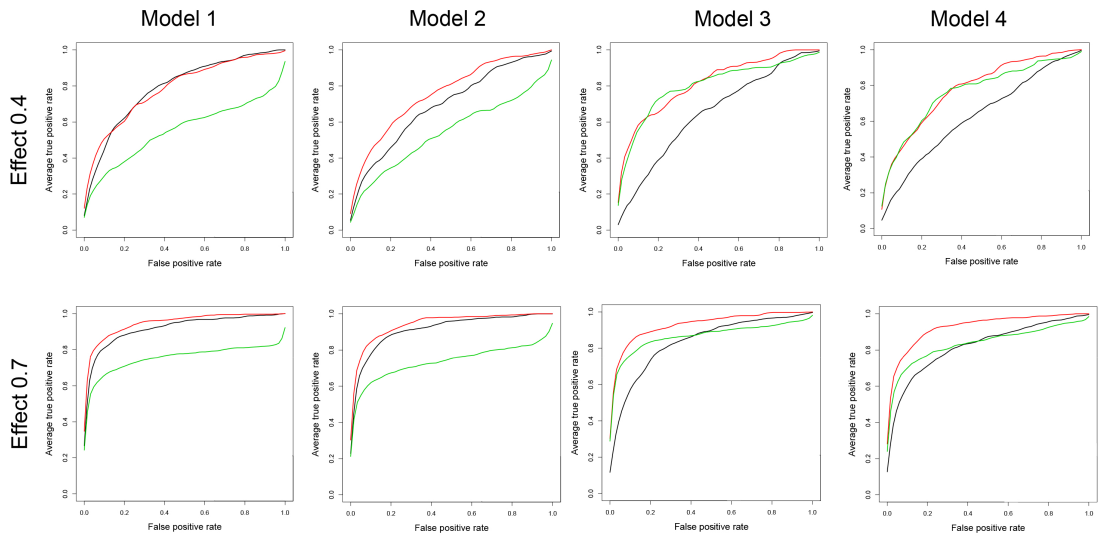


Figure 3.12: Sensitivity-specificity plots for the comparison of the ranking of the SNPs by the log-rank test alone (black line), and after the application of the simple RSF (green line) and the Variable Ranking (red line), on the tag SNPs with signs of association.

models. Figure 3.12 shows that the Variable Ranking was either as successful as, or outperformed, the other two methods for all cases. Although for the two models without covariates the Variable Ranking demonstrated a very similar performance to the log-rank for both effect sizes, for the more complex models the Variable Ranking's performance was significantly improved for both the strong and weaker effect sizes. Interestingly, the performance of the simple RSF was also superior to the log-rank test and comparable to the Variable Ranking for the models with the additional covariates, suggesting that prognostic factors can help better define appropriate trees and thus tune the variable importance measure. This observation, however, was only noticeable for the weak effects (scenario 2), whereas the Variable Ranking had an advantage for the strong effects. A possible explanation could be that the split of the data into training and testing subsamples (K parameter of the varSel function) is only beneficial when the effects are strong.

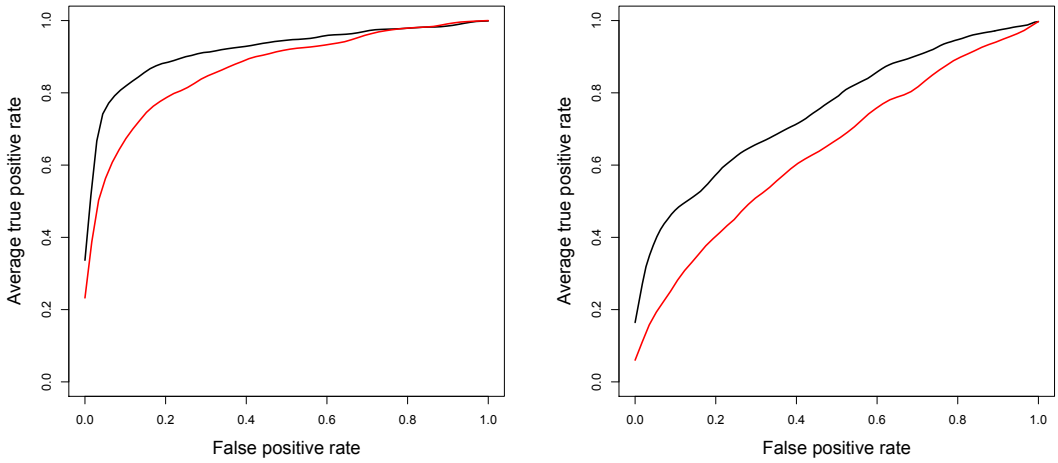
In the above comparisons of Figure 3.12, the p-value threshold chosen for the application of the Variable Ranking was 0.001. This threshold was shown to provide enhanced performance across all models. However, the choice of p-value threshold depends heavily on the sample size and the effect sizes of the covariates.

It is noteworthy that, although the RSF theoretically performs best in more complex scenarios of multiple interactions between covariates, it is also suited for models of no interactions, as in the above scenarios. The results show that the Variable Ranking methodology could aid the better ranking of candidate SNPs that associate with survival phenotypes, compared to their selection based on p-values from the log-rank test alone. This methodology was also tested for a sample size of 200 patients, the details of which can be found in the Appendix.

3.2.2 Comparison to the Cox proportional hazards model

Given the promising results described in the previous section, it was hypothesised that applying this algorithm to the top results from a Cox proportional hazards (Cox phz) model could further improve the rankings of the top SNPs. Similarly to Section 3.2.1, the varSel function was applied to SNPs with p-values under various thresholds. However, the Variable Ranking did not perform better for any threshold. An example of this comparison is shown in Figure 3.13, where the Variable Ranking of the SNPs is compared to the ranking according to the Cox phz p-values for the model with 20% censoring and the additional prognostic factors, model 4.

This is not surprising, since all the models simulated above were the ones



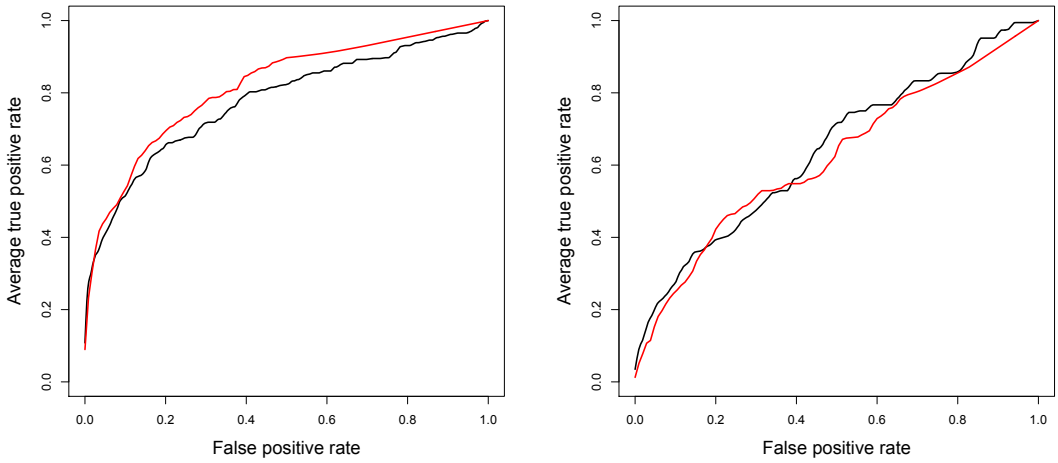
(a) (b)

Figure 3.13: The methodology is applied to SNPs with low p-values (< 0.001) from the Cox phz model for scenario 1 (strong effects) (a) and scenario 2 (weak effects) (b) for model 4. The rankings of the Cox phz (black line), however, are superior to the Variable Ranking ones (red line) for both scenarios.

for which the Cox phz model is known to have the optimal performance. The performance of the Variable Ranking was also examined in two alternative models: *model 5*, a model where the effects of the SNPs are covariant specific, i.e. are only observed in subsamples of the data according to a specific trait (e.g. gender); and *model 6*, a model where multiple SNPs have small size effects (additive effect of 0.262 on the log-hazard - multiplicative effect of 1.3 on the hazard function).

For model 5, a factor with two levels of equal probabilities was simulated for each dataset, and the 4 SNPs were set to have an effect for only one of the levels (the same for all 4 SNPs). Again, this was replicated for the same two scenarios as in Section 3.2.1 (an additive effect of 0.4 and 0.7 on the log-hazard).

Figure 3.14 shows that the Variable Ranking was on average of equal performance to the ranks of the SNPs by the p-values of the Cox phz model, when

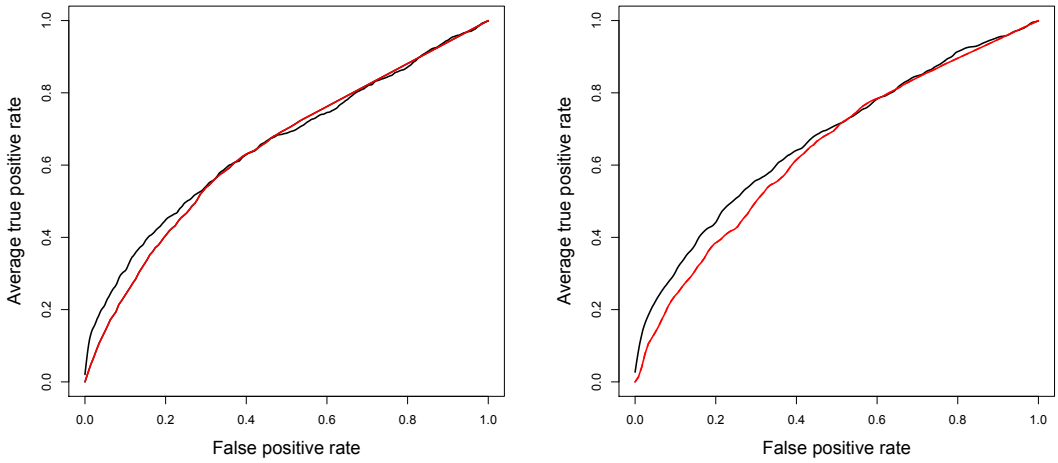


(a) (b)

Figure 3.14: Comparison of the rankings of the Cox phz model (black line) to the Variable Ranking (red line) for scenario 1 (strong effects) (a) and scenario 2 (weak effects) (b) for model 5 of covariate specific effects.

covariate-specific effects associate with the phenotype. Variable Ranking performed slightly better for scenario 1, where the effects were strong, but slightly worse for weaker effects (scenario 2).

Finally, model 6 was simulated to have 20% censoring, and two additional prognostic factors were included in the models (similarly to model 4), but the causative SNPs were 30 instead of 4. The SNPs were randomly picked from the set of all SNPs and, assuming that the LD is similar to the HapMap CEU population, the associated SNPs in this case were 185 SNPs in total (30 causative SNPs and 155 linked SNPs). The Variable Ranking algorithm was not better than the ranking of the SNPs by the Cox phz model alone for either p-value threshold of 0.01 (Figure 3.15a) or 0.005 (Figure 3.15b). A p-value threshold of 0.001 was not examined for this model because too many associated SNPs did not pass this threshold due to the size of the effects.



(a) (b)

Figure 3.15: Comparison of the rankings of the Cox phz model (black line) to the Variable Ranking (red line) for a p-value threshold of 0.01 (a) and 0.005 (b) for model 6.

A likely explanation for the inferior performance of the Variable Ranking is that, due to the large number of associated SNPs and the low sample size, the trees could not be deep enough to identify all candidates with high frequencies and the associated SNPs had to compete with one another. The shallow trees due to small sample size could also be the cause of high variability, which can be a disadvantage in a low-signal setting. In addition, even though the tightly linked SNPs were removed prior to the application of the varSel function, some SNPs that were on the boundary of linkage ($r^2 = 0.8$) are bound to have remained for the final analysis. That could further deteriorate the performance of the Variable Ranking, as shown in Section 3.2.1.

3.3 Discussion

Performing variable ranking in high dimensional survival settings of small sample sizes and weak effects is a challenge for any existing methodology. Furthermore, complex structures of genetic data can provide an extra layer of complication to the process of the identification of SNPs associating with disease. Often, some of these obstacles are ignored in theoretical studies of simulations of models used to assess the performance of methodologies. However, in this study attempts have been made to use realistic scenarios in order to make the results of these simulations directly applicable to our datasets. In addition, the performance of this methodology, the Variable Ranking, is compared to two of the most common tests, the log-rank test and the Cox proportional hazards model.

The Variable Ranking algorithm was shown to outperform the log-rank test in a number of models tried. It showed optimal performance for models of strong effects and had the greatest advantage in cases where additional prognostic factors had an effect on the outcome. It was better than the simple RSF in most cases and of equal ranking abilities for the scenarios of weak effects and additional covariates. However, this was not the case when the Variable Ranking was applied to the significant SNPs from the Cox model. The Cox phz was better than the Variable Ranking for most cases, apart from the case of the covariant-specific effects, for which they had equal performance. Nevertheless, all the scenarios in which the Cox phz outperformed the Variable Ranking were simulated using an exponential distribution, with no time-specific effects or interactions, models for which the Cox model is known to perform well. It would, therefore, be interesting to test the

sensitivity of the Cox phz model in a wider variety of applications. In such cases, a combination of the two methods would perhaps provide the best results, with the Cox model identifying strong linear effects and the Variable Ranking any possible non-linear effects.

In addition, it is encouraging that the Variable Ranking was equally as good as the Cox phz model for the covariate specific model. Model 5 was a special case of an interactions scenario, where the covariate had a strong main effect as well as an interaction effect with the SNPs. This would suggest that pure interactions models, without main effects, between the SNPs as well as with prognostic factors, would perhaps be more appropriate for this type of analysis. For interactions models however, the first step of the Variable Ranking where the log-rank test is applied, would need to be modified in order to allow for a more appropriate test to be used for the pre-selection of SNPs. Therefore, although identifying potential interactions is a very interesting and important issue in SNP studies, and specifically with the use of data mining techniques such as the Random Survival Forest, key aspects of the Variable Ranking methodology would need to be modified to accommodate pure interaction effects.

The identification of interaction effects between SNPs has been a notoriously difficult topic in genomic studies, due to the huge amount of hypotheses that need to be tested in a genome-wide setting. As a consequence, even studies of thousands of individuals can be underpowered to detect these effects. Moreover, the data mining techniques that deal with high dimensional settings have not been extensively tested in such scenarios, a topic that is discussed at length in the Discussion (Chapter 8).

Specifically for RF, interactions models have been the focus of research for both classification and variable selection purposes since the early applications of the RF (Schwender *et al.*, 2004; Bureau *et al.*, 2005; Lunetta *et al.*, 2004). It was shown that the RF outperformed the Fisher's exact test when interactions models were tested (Bureau *et al.*, 2005), and that the performance of the RF increased with the numbers of interacting SNPs (Lunetta *et al.*, 2004). However, more recent research has provided evidence that the detection probability of the interacting SNPs is dependent on the marginal effects of the SNPs and not their total effects (Winham *et al.*, 2012), and that the RF can identify the individual SNPs with high variable importance but cannot identify the pairs of SNPs that interact (Chen and Ishwaran, 2012). However, the effect of interactions on a survival setting and more specifically using the variable hunting algorithm has not been examined in previous studies. Additional simulations of interaction scenarios under varying effect and sample sizes would therefore be an interesting extension for the Variable Ranking. Nonetheless, this would be a very challenging task, as the current implementation of the varSel function of the RSF to identify interactions is too computationally intensive to apply on hundreds of variables. The many difficulties that are encountered by statisticians to detect interaction in genetic studies, along with a further discussion of the data-mining techniques for detecting interactions in genetic studies is given in the final Discussion Chapter.

An important question in data mining techniques is the stability of the results produced by each algorithm. For RF and the variable importance measures (VIMP), it has been previously noted that ranks from permutation based VIMPs

are not stable (Nicodemus, 2011). To adjust for this instability, the variable hunting algorithm was used as the main component of the VR. As mentioned in the Materials and methods (Chapter 2) in more detail, the variable hunting is a regularization algorithm that involves running multiple times (nrep) the RSF. The VIMPs are consequently averaged over the runs and the frequency by which the variables are selected produce the relative frequency of the variables (rel.freq) (Ishwaran *et al.*, 2010, 2011). Moreover, VR utilises both the averaged VIMPs and the relative frequency of the variables in the new measure, which has the added advantage of further increasing the stability of the rankings because it depends on two measures rather than just the VIMP.

The Variable Ranking methodology has the great advantage of being flexible for use under a number of possible models, because it neither requires model specification nor depends on the coding of the variables, unlike the application of the Cox model. However, the Variable Ranking is dependent on some user-defined parameters that can greatly affect its power. One of these parameters is the p-value threshold to be set after applying the log-rank test. As shown here, setting an appropriate threshold can have a large effect on the results. By limiting the p-value threshold, the low frequency or weak effect SNPs are not always included in the second stage of the analyses. On the other hand, allowing too many SNPs in the variable selection (varSel) function deteriorates its power. This is an important limitation of this methodology so various thresholds should be tried out to validate the stability of the results independently of the threshold chosen.

Roshan *et al.* (2011), showed that RF achieved higher power and stability

Effect	Causative SNP	Proportion passing Bonferonni correction
1.5	SNP1 - low freq, not linked	0.01
	SNP2 - high freq, not linked	0.03
	SNP3 - low freq, linked to 4	0.02
	SNP4 - high freq, linked to 4	0.03
2	SNP1 - low freq, not linked	0.08
	SNP2 - high freq, not linked	0.44
	SNP3 - low freq, linked to 4	0.13
	SNP4 - high freq, linked to 4	0.46

Table 3.4: The frequency at which each type of SNP is passing the Bonferroni correction with the log-rank test. In the case of the weak effect sizes, hardly any SNPs pass the Bonferroni correction, whereas for strong effects only the high frequency SNPs come close to a 50% success rate. The success rates here are estimated for model 4, from 100 simulations.

in its rankings according to the variable importance measure (vimp) when the RF was applied to the top SNPs passing the Bonferroni correction using a chi-squared test. The findings described in this chapter are in line with Roshan *et al.* (2011), in that the rankings of the RSF also improve with the removal of excessive noise. However, the thresholds here are much more flexible because using the Bonferroni correction as a threshold in this sample size would have been too exclusive, as shown in Table 3.4. Nonetheless, the trade-off of excluding potentially associated SNPs in order to achieve a better ranking for the remaining SNPs is also an important limitation of this methodology.

Another parameter to be set is the correlation coefficient (r^2), which is used when defining the tag SNPs. For all the analyses described here, an r^2 of 0.8 was used in the clustering algorithm that defined the tag SNPs. This is a standard threshold for defining SNPs in LD in association studies (de Bakker *et al.*, 2006; Pe'er *et al.*, 2006; Barrett and Cardon, 2006). Further simulations are needed to fully understand the impact of this parameter setting for the identification of the

associated SNPs, with varying LD structures. In more detail, one of the most important parameters in simulation studies of SNP datasets is the effect of linkage disequilibrium in the data. The simulations of the datasets of this chapter are based on effects of SNPs with two types of linkage disequilibrium (LD): two independent SNPs which do not belong to any strong haplotype blocks and two SNPs which belong to two small haplotype blocks and are strongly linked to 4 other SNPs each. Here, it was shown that removing the linked SNPs greatly improved the ranking of the associated SNPs and therefore the performance of the algorithm. Hence it would not be expected that the number of SNPs belonging to the haplotype would influence the ability of the algorithm to identify the associated SNPs, since the linked SNPs are removed by default from the dataset before the application of the RSF. It also needs to be mentioned that the aim of this algorithm was to identify the associated SNPs and not necessarily the causative SNPs since the project intended to identify causality and ultimately functionality by means of bioinformatics analysis and in experimental procedures.

However, the threshold of correlation coefficient (r^2) which was used to define the haplotype blocks during the simulations process was 0.5 while the threshold used to select the tag SNPs in the algorithm was 0.8. Therefore, in real life scenarios, SNPs residing in long haplotypes of extended homozygosity but with medium size correlation coefficients ($0.5 < r^2 < 0.8$) could also influence the performance of the algorithm, since the linked SNPs wouldnt be removed from the dataset before the RSF. Additional simulations are hence needed to identify the best r^2 threshold to be used in the process of selecting the tag SNPs under various correlation struc-

tures. Along the same lines, in future simulations, it would be interesting to test to what extent of LD the RSF is influenced in the selection process of the highly ranked SNPs as a function of the sample size to which is applied.

Finally, the potential effect of causative SNPs that have a low r^2 but a high D' with the genotyped SNPs would also be an interesting scenario to test. Such scenarios would depict the effect that rare SNPs would have on the phenotype of interest. However, the chance of identifying the right haplotype in that case, would depend on the strength of the effect of the rare causative SNP in relation to the sample size at hand. Therefore, many different scenarios of effect sizes and sample sizes would need to be tested to fully explore these types of effects.

Additional solutions have been recommended for the appropriate handling of correlated predictors in genetic studies (Meng *et al.*, 2009; Strobl *et al.*, 2008). In an extended simulation analysis, Nicodemus *et al.* (2010) evaluated the various variable importance measures under the null and under models of association of correlated predictors. It was shown that when models of correlated predictors were simulated, artefacts were introduced into the study and the correlated predictors were much more strongly associated when each was considered on its own. Moreover, it was shown that the unconditional RF VIM was unbiased under the null hypothesis and gave a higher importance to correlated predictors in the first split of the trees, but also a slight preference for uncorrelated predictors across all splits. Similar studies for the RSF are needed for the selection of the best-performing scheme for dealing with LD structure but, to our knowledge, alternative measures have not yet been implemented in a survival context.

Further adjusting the parameters employed for the building of the trees could enhance the power of the Variable Ranking. The number of variables tried per node (*mtry*) is known to affect the trade-off between bias and variance, and has been studied extensively in Goldstein *et al.* (2010). Here, an increased value for the *mtry* was explored but a more detailed analysis is needed to fully explore the most appropriate value as a function of the number of parameters. Moreover, the terminal node size of the RSF has been shown to affect the depth and the balancedness of the trees (Ishwaran *et al.*, 2011). Therefore, adjusting this parameter could provide more stability to the results, but additional analyses are needed to further test this hypothesis in the context of the Variable Ranking.

The varSel function does not allow for different splitting rules, except for the default which is the *logrank* rule, based on the maximisation of the log-rank test statistic. The *logrank* and *logrankscore* splitting rules were shown to have the lowest prediction errors (Ishwaran *et al.*, 2008), but no study has been done to compare the performance of the splitting rules for variable ranking purposes. In addition, the first step of Variable Ranking already uses the log-rank test for the filtering of SNPs, and so an alternative splitting rule could provide greater variability and further benefit the performance of the methodology.

An important limitation of the Variable Ranking methodology is that it does not provide a threshold for significance, which would be a valuable tool for studies where variable selection is the aim. Even though the threshold of significance in its strict sense is a measure tied to frequentist approaches and not data mining techniques, as the one used here, the varSel function includes in its features a

variable selection algorithm which provides a subset of selected variables based on either the minimum depth or the variable importance (Ishwaran *et al.*, 2010). The ranking on which the variable selection is based, however, was shown in this study to be slightly inferior to the ranking by *rank.VIMP* (Figure 3.10). This is the reason that variable selection was not used here.

Moreover, this algorithm cannot be used for the identification of causative SNPs, since only the tag SNPs are used for the ranking. Nevertheless, given the low sample sizes in such studies, a replication study, in which fine-mapping of the associated regions can be performed, is an essential step for the validation of the observed associations. This is not, therefore, considered to be a disadvantage of the specific algorithm, but of this type of analysis in general. In the next chapter, however, an alternative approach is proposed for the detection of the causative variant. Finally, no genotyping error was assumed in this analysis and therefore this scenario was not studied, although any future simulations would ideally take into account genotyping error as well.

Finally, simulations were performed to test the VR under different models and effect scenarios, based on the type of real data that was available to study in this project. Small cohorts of a few hundreds of patients are very common in cancer studies and as such this methodology aims to address the low power issue that comes with the small sample sizes. In order to make the simulations resemble real life cohorts, two sample sizes were tried out, with 200 and 300 patients under 4 models and 2 effect size scenarios. Undoubtedly, testing VR under other scenarios of larger cohorts of a few thousand patients would add to the understanding of its

performance but its results would not be transferable to our data.

In conclusion, this study presents a model-free algorithm by which SNPs can be ranked based on signs of association with survival phenotypes, and it is utilised in the following chapters to identify the most promising candidates for associations with cancer progression. Moreover, it highlights that many of the observations that have been published on the RF can be directly applicable to the RSF too, which has not been demonstrated before. This is important, because over the last decade great advances have been made on the applications of the RF in genetic epidemiology and bioinformatics, but the RSF is a new methodological approach that may hold great promise for the field of genetics.

4 Variable Ranking identifies SNPs in the LEPR and ITGA1 genes that associate with B-CLL progression

4.1 Introduction

Chronic lymphocytic leukaemia is the most common type of leukaemia in Western countries with a higher occurrence in elderly individuals (Rozman and Montserrat, 1995). It is a haematopoietic disorder that arises from the uncontrolled expansion of lymphocytes of B-cell lineage. The disease is characterised by a varied symptom course and a diverse pattern of survival. For this reason, it is important that genetic markers are identified that can help detect patients at high risk of faster progression and poorer survival so that their treatment options can be handled accordingly. Furthermore, the identification of such genetic variants could improve our understanding of the cellular processes involved with the hope of exposing potential nodes of therapeutic intervention.

B-cell chronic lymphocytic leukaemia (B-CLL) presents itself with an accumulation of mature but non-functional lymphocytes in the blood, bone marrow and lymphoid tissues. It is usually diagnosed from routine blood cells with B-lymphocytes that exceed the normal levels of 5,000 cells per microlitre (Hallek *et al.*, 2008). Most cases are first diagnosed by chance and are asymptomatic (DeVita *et al.*, 2011). Some of these patients will progress quickly and die within a few years of their diagnosis, whereas others will have a normal life span and die of

unrelated causes (Rozman and Montserrat, 1995).

Patients diagnosed with CLL are typically classified into 3 groups according to the level of disease progression (Binet *et al.*, 1981): group A includes patients with 2 or fewer lymphoid involved areas; patients with more than 3 areas are classified as group B; and when they also have either anaemia or thrombocytopenia, in group C. About 80% of the patients are diagnosed at an early stage (Byrd *et al.*, 2004), underlining the need for markers for survival prediction and progression. In general, CLL progresses slowly and even high-risk patients have a mean survival of several years. Treatment depends on the patients' staging and disease progression. Initiation of treatment is recommended for patients in stage B or C that also have evidence of progressive or symptomatic disease (Hallek *et al.*, 2008). However, it has been shown that earlier treatment of stage A patients does not improve survival rates (Dighiero *et al.*, 1998), suggesting that further stratification of these patients into high or low progression risk groups could highlight those that would benefit from earlier treatment.

Currently, the most important prognostic factor for the progression of the disease is the mutation status of the immunoglobulin variable-region heavy chain (IgVH) gene (Damle *et al.*, 1999; Hamblin *et al.*, 1999). The un-mutated gene is associated with earlier progression and worse overall survival. The stratification of the patients is due to B-cells being transformed at different stages of their differentiation and activation process. B-cells of patients with an un-mutated IgVH gene show characteristics of either naive B-cells, or cells that have been activated but have not entered a germinal centre (GC). Mutated IgVH is an indication of

hypermutated B-cells that are either post-GC or memory cells. Following these studies, patients are classified as high or low risk depending on their mutation status for IgVH. The screening of patients for IgVH mutations is not easily done using routine tests and surrogate markers have been identified for this purpose. Damle *et al.* (1999) presented evidence that patients with unmutated IgVH showed higher expression of the cell surface molecule cluster of differentiation 38 (CD38), a glycoprotein that is found on the surface of many immune cells and functions in cell adhesion, signal transduction and calcium signalling. Further analyses have also associated higher expression of Z-chain associated protein kinase-70 (ZAP-70) with the mutation status of IgVH (Crespo *et al.*, 2003). ZAP-70 is a protein of the protein-tyrosine kinase family which is usually expressed in T-cells. Crespo *et al.* (2003) have shown that higher levels of ZAP-70 expression are found in leukaemic cells in patients with earlier progression and worse survival, and were correlated with un-mutated IgVH status. Yet, the mechanism of any relationship between the two is unknown.

Another type of important prognostic indicator for overall survival and disease progression is the genomic aberrations in patients with CLL. Deletions in 17p and 11q, which target the TP53 and ATM tumour suppressor genes, respectively, associate with worst survival, whereas the 13q deletion (resulting in Rb loss) associates with better survival and later progression. A trisomy in chromosome 12 was also associated with lower survival rates compared to patients with normal karyotypes, but the data has been controversial regarding its significance (Dohner *et al.*, 2000). Furthermore, chromosomal translocations as well as elevated genomic

complexity, have been associated with more aggressive CLL (Ouillette *et al.*, 2010; Mayr *et al.*, 2006).

Over the last decade, a number of studies have focused on the genetic component of CLL (Goldin *et al.*; Caporaso *et al.*, 2007; Sellick *et al.*, 2006). However, the aetiology of CLL is still unknown. A number of GWA studies have been conducted and reported (and validated) up to 11 variants with significant associations with susceptibility to B-CLL (Di Bernardo *et al.*, 2008; Crowther-Swanepoel *et al.*, 2010; Slager *et al.*, 2010, 2011a,b; Di Bernardo *et al.*, 2013). In addition, in a study of Knight *et al.* (2012) a number of recurrent copy number alterations (CNAs) and copy neutral loss of heterozygosity regions (cnLOHs) were identified in a CLL cohort (Knight *et al.*, 2012). In contrast, GWAS for the associations of SNPs with B-CLL progression and survival have not been as fruitful. To our knowledge, only one GWAS has been carried out that looked for associations of SNPs with progression-free survival but the findings could not be validated in a replication cohort (Wade *et al.*, 2011). In addition, a few candidate gene approaches have attempted to find associations with progression and overall survival, but so far the evidence has been contradictory and none have been replicated consistently (Slager *et al.*, 2007). Furthermore, Sellick *et al.* (2008) performed an analysis in 755 genes, which included 977 non-synonymous SNPs, to look for associations with the progression-free survival and overall survival of a group of patients. They identified 78 SNPs with evidence of association but no replication was reported. A follow-up study was later carried out, but only two of the previously reported SNPs were replicated (Rasi *et al.*, 2010). Therefore, a more comprehensive un-

dertaking to identify SNPs that are associated with faster progression and poorer overall survival of B-CLL patients is needed.

In this chapter, the Variable Ranking algorithm, which was described in Chapter 3, is applied to conduct an association analysis to identify SNPs that associate with differences in the progression rate of patients with B-CLL (measured by time to first treatment) in 16 pathways of genes that have been shown to be associated with cancer susceptibility and progression. First, the Variable Ranking algorithm was applied and next, the top results were screened using bioinformatics filters. Five candidate SNPs were taken forward for validation in a replication cohort. It was demonstrated that a SNP upstream of the promoter region of the LEPR gene significantly associates with time to first treatment in both cohorts, and a potential interaction of another SNP in intron 1 of the ITGA1 gene with deletion 11q was explored. The data presented suggests that the SNPs in transcriptional regulatory regions of these genes, could lead to altered levels of the protein products, with a potential effect on disease progression.

4.2 Results

4.2.1 Analysis of SNPs for association with B-CLL time to first treatment

Clinical characteristics of the discovery cohort

The discovery cohort comprised 181 patients with B-CLL and complete data for time to first treatment (TFT). TFT was recorded as the time between the patients' diagnosis and the point in time at which they required treatment, so act as a

marker for progression. In this cohort, 74% of the patients were males (134 males; 47 females), in line with the observed gender-specific incidence rates, by which the prevalence of B-CLL is higher in males than in females (Siegel *et al.*, 2013). The age at the time of diagnosis ranged from 39 to 88 years (median 64). At the time of diagnosis 78 patients were at stage A, 62 at stage B and 37 at stage C (4 had no records of stage). The patients' samples came from two sources (details in the Materials and methods): 88 of them were collected as part of a clinical trial (CLL4) and presented with advanced and faster-progressing disease; 93 were patients of a local hospital in Bournemouth and had slower-progressing disease (locals). This heterogeneity between the subgroups could introduce bias in the analysis and for this reason the source of the samples has been used as a prognostic factor in the following analyses and discussed in more detail in the discussion of this chapter. In total, 142 patients received treatment before the end of the study (22% censoring). Table 4.1 lists the immunophenotypic data and genomic aberrations (deletions 11q and 17p) that were recorded for the patients. Deletion 13q was not recorded in a similar manner between the two subgroups (CLL4 and locals) so the data could not be summarised in the table.

Power calculations

In order to assess the power of this study to identify the causative (or associated with $r^2 = 1$) SNPs using the most common analysis with the Cox proportional hazards model the power curves were drawn for SNPs of various MAF and genotype hazard ratios (GHR). The genotype hazard ratio is used as the equivalent of

Clinical characteristics	Discovery cohort Number of Patients (%) N=181	Replication cohort Number of Patients (%) N=601
Gender		
Male	134 (74)	396 (66)
Female	47 (26)	195 (32)
Missing	-	10 (2)
Stage		
MBL	-	90 (15)
A	78 (43)	210 (35)
B	62 (34)	181 (30)
C	37 (21)	117 (20)
Missing	4 (2)	3 (0.005)
IgVH mutation		
Yes	72 (40)	282 (47)
No	98 (54)	246 (41)
Missing	11 (6)	73 (12)
cd38		
0	59 (33)	180 (30)
1	91 (50)	188 (31)
Missing	31 (17)	233 (39)
zap70		
0	79 (44)	183 (31)
1	83 (46)	146 (24)
Missing	19 (10)	272 (45)
deletion 11q		
0	123 (68)	505 (84)
1	48 (27)	65 (11)
Missing	10 (5)	31 (5)
deletion 17p		
0	157 (88)	542 (90)
1	12 (6)	20 (3)
Missing	12 (6)	39 (7)

Table 4.1: Clinical characteristics of discovery and replication cohorts of B-CLL patients.

the effect size in a survival setting and defined here as the ratio of two conditional hazard rates evaluated at time $t > 0$ (Owzar *et al.*, 2012). The additive model is assumed for these calculations and the results are based on a single marker analysis using the Cox proportional hazards model, that is considered the most powerful approach for these types of analyses. The calculations were based on 181 individ-

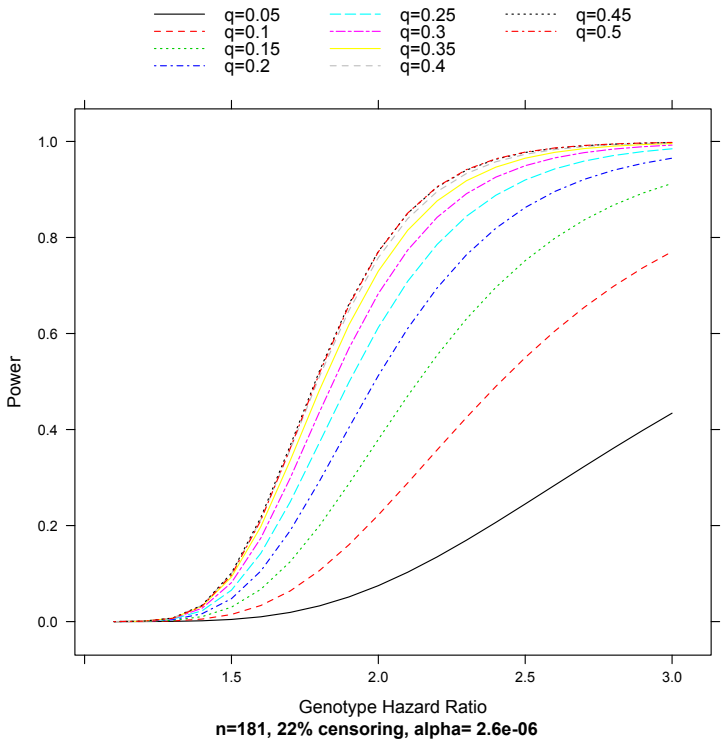


Figure 4.1: The power curves show the power to detect a SNP associated with time to first treatment at a level of significance of alpha=$2.6 * 10^(-6)$ (after Bonferroni correction for the 19,122 SNPs). The curves are shown for SNPs of *MAF* ranging between 0.05 and 0.5 and for genotype hazard ratios (GHR) ranging between 1.1 and 3.

uals and 22% censoring, similar to the characteristics of this cohort. As expected, this study was very underpowered to detect SNP associations with this sample size using the Cox proportional hazards model. In more detail, for 80% power, the identification of SNPs with $MAF > 0.35$ was only possible when the GHR was over 2. Furthermore, the GHR needed to be greater than 2.6 in order to achieve 80% power to identify SNPs with *MAF* of 0.15 or more (Figure 4.1 and Table 4.2). Moreover, a SNP with a *MAF* of 10% can only be identified with a power of 0.55 if it has a GHR greater than 2. These calculations highlighted the need to take into account information on regulatory regions, to aid in the discovery of SNPs which associate with time to first treatment.

	GHR			
MAF	1.5	2	2.5	3
0.1	0.015	0.222	0.550	0.771
0.2	0.048	0.512	0.862	0.965
0.3	0.081	0.683	0.950	0.992
0.4	0.100	0.758	0.973	0.997
0.5	0.097	0.770	0.978	0.998

Table 4.2: Power calculations for a sample size of 181 patients, under different effect sizes and minor allele frequencies (*MAF*).

Quality control and filtering

As described in the Materials and methods, the SNPs were filtered according to location and only SNPs of the 16 cancer-pathways, as defined by KEGG, including 10kb on either side of each gene, were included. Further filtering included a minor allele frequency threshold of 10% (to have a power of at least 0.50 for a GHR greater than 2) and an analysis was performed on 19,122 SNPs of 1,208 genes. Since B-CLL is a haematopoietic disorder, confirmation was needed that the blood samples collected for the genotyping did not have abnormal karyotypes that would bias the results. For this reason, the allele frequencies of the SNPs were compared to observed allele frequencies for the CEU population of HapMap (Figure 4.2). No extreme deviations from the expected *MAF* were observed so no further filtering was applied.

Selection of candidate SNPs associating with Time to First Treatment (TFT)

The methodology employed was designed to identify either coding or regulatory SNPs associating with TFT using the Variable Ranking (Figure 4.3). For the first step of the analysis, a log-rank test was performed and 451 SNPs showed evidence

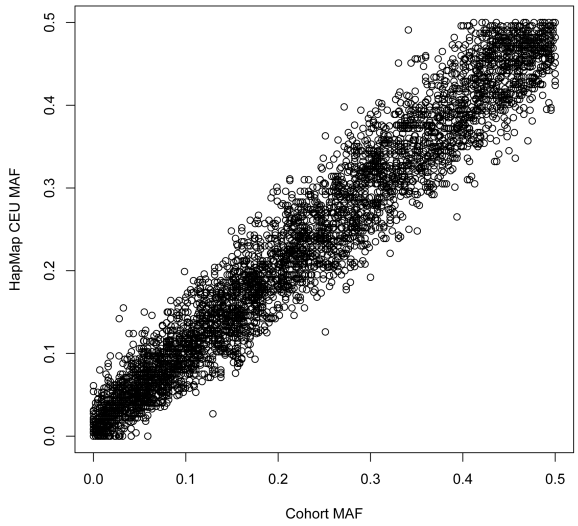


Figure 4.2: Comparison of the cohort minor allele frequencies (*MAF*) to the HapMap CEU *MAF*. No severe discrepancies were observed, confirming that the genotyping was not affected by abnormal karyotypes from the blood samples.

of association (p-value< 0.01) with TFT. After filtering to retrieve the tag SNPs, 294 SNPs were selected and further ranked using the RSF (varSel function with *ntree*=1500 and K=2, as shown in Chapter 3). The prognostic factors described in Table 4.1 and age at diagnosis were also included in the analysis with the RSF in order to account for any confounding effects. Any missing data for these confounders, as well as for the SNPs, was imputed by the varSel function using the new missing data algorithm for forests, as described in Ishwaran *et al.* (2008) by which only in-bag data are used for imputing, reducing the bias in prediction error. The ranking was performed by applying the measure *rank.vhVIMP*, and the top 15 SNPs selected for the final stage of the bioinformatics analysis. It needs to be mentioned, however, that the varSel function is using a Monte Carlo validation in which the algorithm is repeated *nrep* times to increase stability. For this application the *nrep* parameter was increased to 100 (default is 50) to ensure stability of

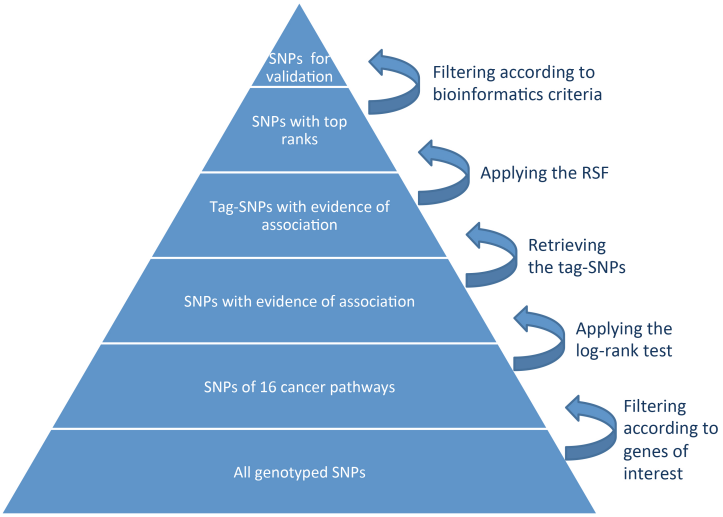


Figure 4.3: Pyramid chart depicting the methodology used to identify SNPs in regulatory regions associating with time to first treatment.

the random forest rankings, as it has been noted before that the vimp measure can have reduced stability (Nicodemus, 2011). Moreover, due to the low availability of DNA samples in the replication cohort, only a few SNPs could be followed up with additional genotyping. After various thresholds had been tested, the top 15 threshold followed by a bioinformatics analysis gave a realistic number of SNPs to be tested in the replication cohort, in terms of sample availability. Moreover, the top 15 filter and the bioinformatics analysis ensured that only the SNPs with the best rankings which were also supported by regulatory evidence would be followed up.

Using the data from the ENCODE project, it was determined that of the 15 SNPs, 4 resided in areas with evidence of DNase hypersensitivity, 6 where many transcription factors have been found to bind, and all but 2 were linked to SNPs in regions with signs of enrichment for histone modification markers indicating potential transcriptional regulatory regions (Table 4.3). Furthermore, 1 SNP was

linked to a SNP in an evolutionarily conserved area. In total, 3 candidates were linked to SNPs in regions with strong evidence of being associated with enhancer or promoter activities (rs11740785, rs3806318, rs4693051) and 2 were linked to synonymous SNPs located in coding regions (rs1550871 and rs6690837). More details on the bioinformatics analysis can be found in the Materials and methods. As these five SNPs could reside in potential regulatory regions and associate with differential CLL progression (Figure 4.4), they were taken forward for replication in an independent cohort of 601 B-CLL patients.

SNP	SNPs in exonic regions	SNPs in potential enhancer or promoter regions				Selected SNPs
		DNaseI	TFBS	Histone Markers	Conserv. Score	
rs11740785		✓	✓	✓	✓	✓
rs12432304				✓		
rs1432579				✓		
rs1550871	Synonymous, exon 14 in PTPN5 (via proxy)		✓	✓		✓
rs16959941				✓		
rs17234079						
rs17293443				✓		
rs17632458			✓	✓		
rs2282588		✓		✓		
rs290490						
rs3806318		✓	✓	✓		✓
rs4693051		✓	✓	✓		✓
rs6690837	Synonymous, exon 6 in VAV3 (via proxy)			✓		✓
rs6715129						
rs889294			✓	✓		

Table 4.3: Bioinformatics filter for the top 15 SNPs from the Variable Ranking. Aiming to find SNPs in functional regions of the DNA, two SNPs were selected for being linked to coding SNPs of the genes in which they reside, and three for being linked to SNPs in potential enhancer or promoter regions.

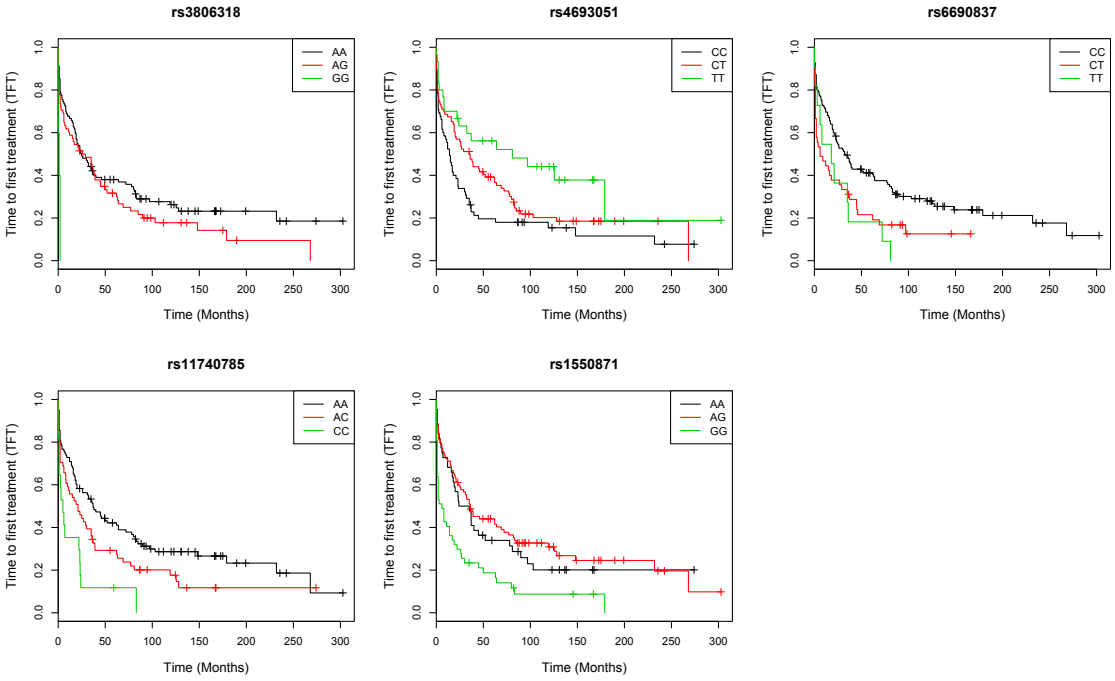


Figure 4.4: Kaplan-Meier curves for time to first treatment (TFT) for the 5 candidate SNPs in the discovery cohort.

4.2.2 Replication analysis of candidate SNPs

Clinical characteristics of the replication cohort

The replication cohort comprised of 601 patients with B-CLL (or its precursor monoclonal B-cell lymphocytosis, MBL) from the same two sources as the discovery cohort (345 patients were from the CLL4 clinical trial and 256 locals). 90 patients had been diagnosed with MBL (a group which was not present in the discovery cohort); 210 with stage A, 181 with stage B and 117 with stage C. The age of diagnosis ranged from 34 to 93 (median 64). Finally, 443 patients required treatment before the end of the study (26% censoring, a percentage similar to the discovery cohort). A summary of the clinical characteristics of the cohort is given in Table 4.1.

Analysis of candidate SNPs in the replication cohort

The log-rank test and the Cox proportional hazards model (adjusting for stage at diagnosis-`stage`, IgVH mutation-`IgVH`, deletion of 11q-`del11q`, and source of patients-`source`) were applied to validate the associations of the 5 SNPs with TFT. The results were compared to the p-values of the discovery cohort in Table 4.4. In the replication cohort the variable `IgVH` appeared to have a time-varying effect, violating the proportional hazards assumption with a p-value of 0.0095 (Figure 4.5), even though this was not the case in the discovery cohort. For this reason the p-values of both the simple model and the model with the stratified effect for `IgVH` are shown, to demonstrate that the conclusions remain the same for both analyses.

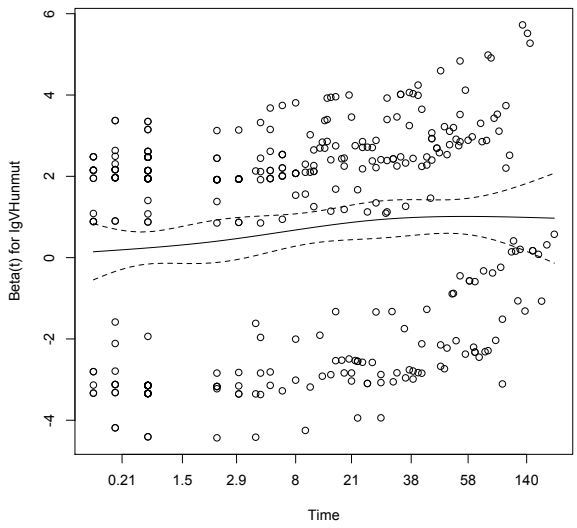


Figure 4.5: Plot of the scaled Schoenfeld residuals against transformed time for the variable IgVH. The trend of the smoothing line fit to the plot demonstrates that the effect of the variable IgVH increases with time, indicating non proportional hazards.

Only SNP rs3806318 in the LEPR gene showed strong evidence of association with TFT in the replication cohort, with a p-value of 0.0042 and with the same direction of effect (log-rank test). This was still significant after multiple hypothesis

SNP	Gene	Discovery Cohort		Replication Cohort		
		Log-rank	Cox ph	Log-rank	Cox ph	
rs3806318	LEPR	$2.73\text{x}10^{-5}$	0.22	0.0042	0.014	0.014
rs4693051	SCD5	$6.15\text{x}10^{-3}$	0.067	0.64	0.12	0.14
rs6690837	VAV3	$3.31\text{x}10^{-3}$	0.047	0.11	0.29	0.41
rs11740785	ITGA1	$4.91\text{x}10^{-4}$	0.010	0.55	0.089	0.064
rs1550871	PTPN5	$1.20\text{x}10^{-4}$	0.045	0.29	0.24	0.24

Table 4.4: *P*-values of the five selected SNPs in the discovery and replication cohorts. The effects of the SNPs are adjusted for `stage`, `IgVH`, `del11q`, `source`), under the Cox proportional hazards model (assuming an additive genetic model on the log-hazard). In the first column the Cox is applied similarly to the discovery cohort and in the second it is applied after stratification for the time-varying variable `IgVH`.

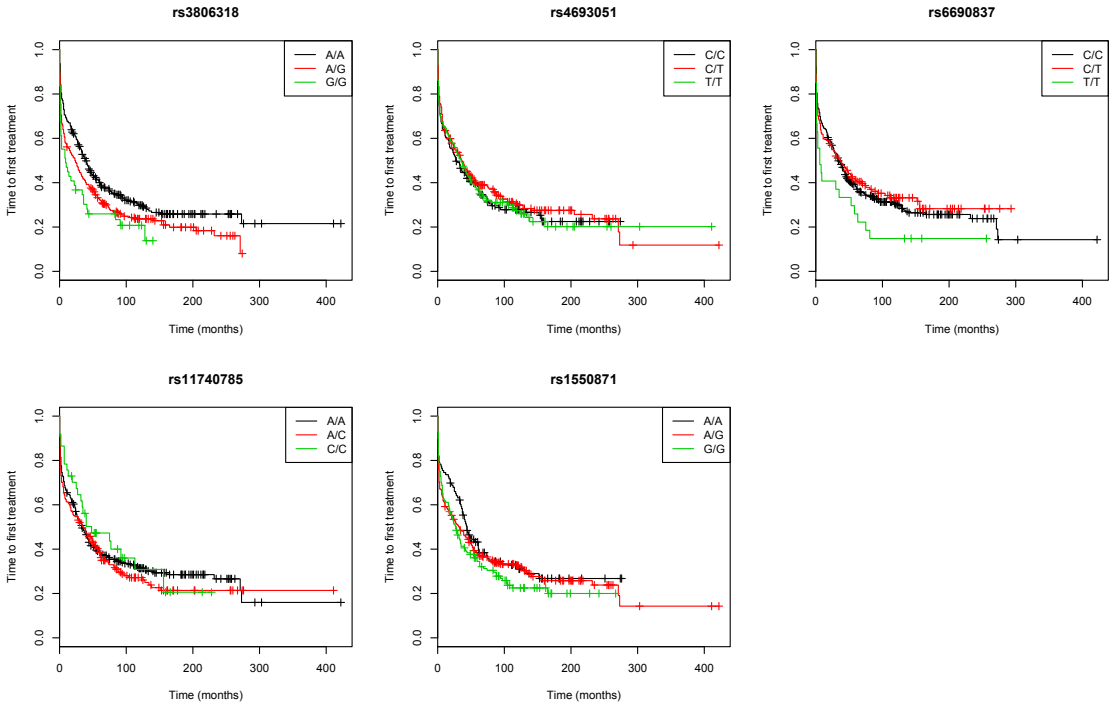


Figure 4.6: Kaplan-Meier curves for the 5 candidate SNPs in the replication cohort. SNPs rs3806318 and rs6690837 have similar trends to the discovery cohort event, though they give support to different genetic models.

correction (Bonferroni correction) with a p-value of 0.021. Specifically, the G allele of SNP rs3806318 was associated with faster progression, which manifests itself in the need for earlier treatment (Figures 4.4 and 4.6). This association remained significant after correction for other prognostic factors (stage at time of diagnosis,

IgVH mutation, 11q deletion and source of patients) with the Cox proportional hazards model with a p-value of 0.014, but was not significant with a Bonferroni correction (p-value 0.07). To further test the significance of this association, the prognostic factors were regressed out of the Cox proportional hazards model and then the (deviance) residuals were used in a linear regression, giving a highly significant a p-value of 0.00072, which was significant after Bonferroni correction (p-value 0.0036). These differences in p-values that can arise from the different applications of the Cox proportional hazards model highlight the fact that even a semi-parametric model can give very different levels of significance depending on the modelling of the variable and application of the model and lead to different conclusions. Therefore, these models should be interpreted with caution. To further underline these differences and to try and identify the genetic model under which the SNP could be associating with the outcome, the analyses under difference genetic models were performed. In the discovery cohort, the SNP was not significant under the Cox model assuming a genetic model of per allele multiplicative effect on the hazard ratio, but was significant when the model was not defined (ordered factor coding) with a p-value of 0.021, or when a recessive mode was assumed with a p-value of 0.0055. Although the direction of effects was identical across both cohorts (the G allele increases the hazard ratio), the model by which the SNP affected TFT was not consistent. In the discovery cohort, a recessive model would best fit the data, with the effect of the homozygous GG increasing the hazard ratio by a factor of 5.23 (95% CI: 1.97, 13.92). However, for the replication cohort, a per allele multiplicative model is more suitable with a p-value of 0.014 compared to a recessive model (p-value 0.15). In addition, the effect of having a G allele

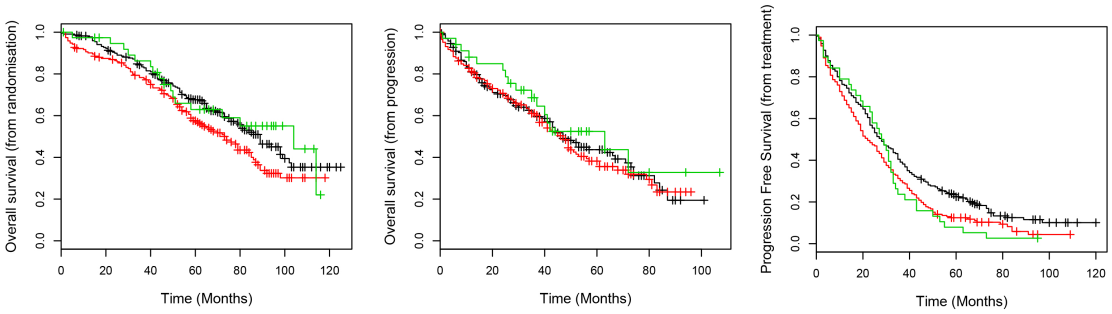


Figure 4.7: Kaplan-Meier curves for overall survival (from randomisation and from progression) and progression-free survival from treatment for SNP rs3806318. The SNP was significant for progression-free survival but not for overall survival. The black line corresponds to the AA patients, the red to AG, and the green to GG patients.

increases the hazard of faster progression by a factor of 1.24 (95% CI: 1.05, 1.47) in the replication cohort.

As both the replication and discovery cohorts included patients that were part of the CLL4 clinical trial, it was possible to investigate whether rs3806318 is also associated with progression-free and overall survival, as these phenotypes were recorded. When the CLL4 patients of both cohorts were grouped, SNP rs3806318 showed significant differences associated with progression-free survival (with treatment time the starting point), with p-values of 0.012 and 0.030 for the log-rank test and the Cox proportional hazards model (adjusting for stage, IgVH, del11 and treatment arm of the clinical trial), respectively (Figure 4.7). No association was observed with overall survival with starting points of either diagnosis or progression. From the Kaplan-Meier plot of the SNP's effect on progression-free survival, there was uncertainty on the proportionality of hazards between the groups for the Cox analysis. However, this was not supported by the model diagnostics for either the specific variable (p-value of 0.052), or with the global test (p-value of 0.22).

The SNP rs11740785 in the ITGA1 gene did not replicate in the second

cohort with a p-value of 0.55 for the log-rank test and 0.089 for the Cox model (adjusting for `source`, `stage`, `IgVH` and `del11q`). However, in an exploratory analysis, potential interactions of the SNP with the additional prognostic factors were examined in more detail performing a stepAIC. A significant interaction effect of the SNP with the deletion 11q was noted using a likelihood ratio test with a p-value of 0.0041. To examine the effect of the increased false positive rates due to the multiple hypothesis testing of these analyses, the stepAIC was run under the null of no interaction of the SNP with any of the four remaining prognostic factors. This was achieved by permuting the SNP 10,000 times while retaining the correlation structure between the prognostic factors and the outcome. Each time the stepAIC was performed and the smallest of the p-values of the interaction tests was recorded. If none of the interactions were retained by the model, no p-value was returned and a random value ranging from 0.1 to 1 was generated instead (assuming a uniform distribution). As expected, the distribution of the p-values was highly skewed towards the smaller p-values, with almost a quarter of the permutations having a p-value less than 0.05, demonstrating the increased false positive rate due to the multiple hypothesis testing (Figure 4.8). It needs to be mentioned, however, that only 3.6% of the permutations had a p-value less than 0.0041.

Treating this analysis as an exploration of the data, a hypothesis was formed that assuming that this interaction is truly significant, the same interaction should also be observed in the discovery cohort. Therefore, an analysis was performed to compare the models with and without the interaction in the discovery cohort using

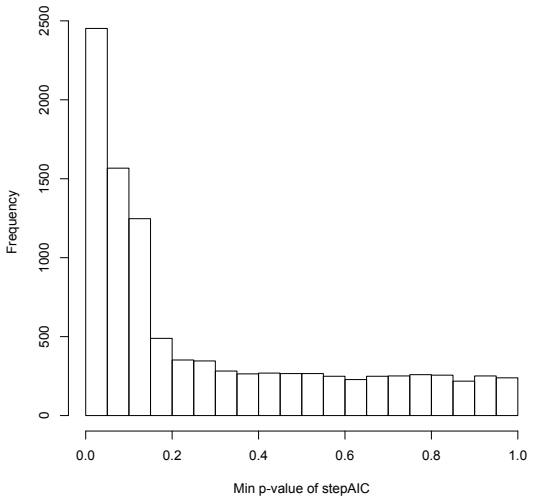


Figure 4.8: Testing the stepAIC function under the null demonstrates the increased false positive rate due to multiple hypothesis testing. Therefore, the distribution of the p-values under the null are highly skewed towards 0.

the likelihood ratio test, but it was non-significant (p-value of 0.082). Furthermore, the Kaplan Meier curves were plotted but the SNP appeared to have a strong effect on the subgroup of patients without deletion 11q and not in the subgroup of patients with deletion 11q (Figure 4.9 (a)). Moreover, the effect of the SNP was stronger and of opposite direction in the replication cohort for the group of patients with deletion 11q (Figure 4.9 (b)), suggesting a false positive result for the interaction effect. However, looking at the subgroup of patients that are classified as locals and have a less aggressive form of the disease, the trends were the same for the patients without deletion 11 in both groups (Figure 4.9 (c) and (d)). Given the nature of these analyses, it should be noted that the observations of these exploratory analyses are only generating interesting hypotheses, but they should be validated in an additional cohort to correct for the fact that exploratory analysis using a stepwise regression technique such as stepAIC produces inflated type I errors from the multiplicity of the tests and as such should not be treated as validated results.

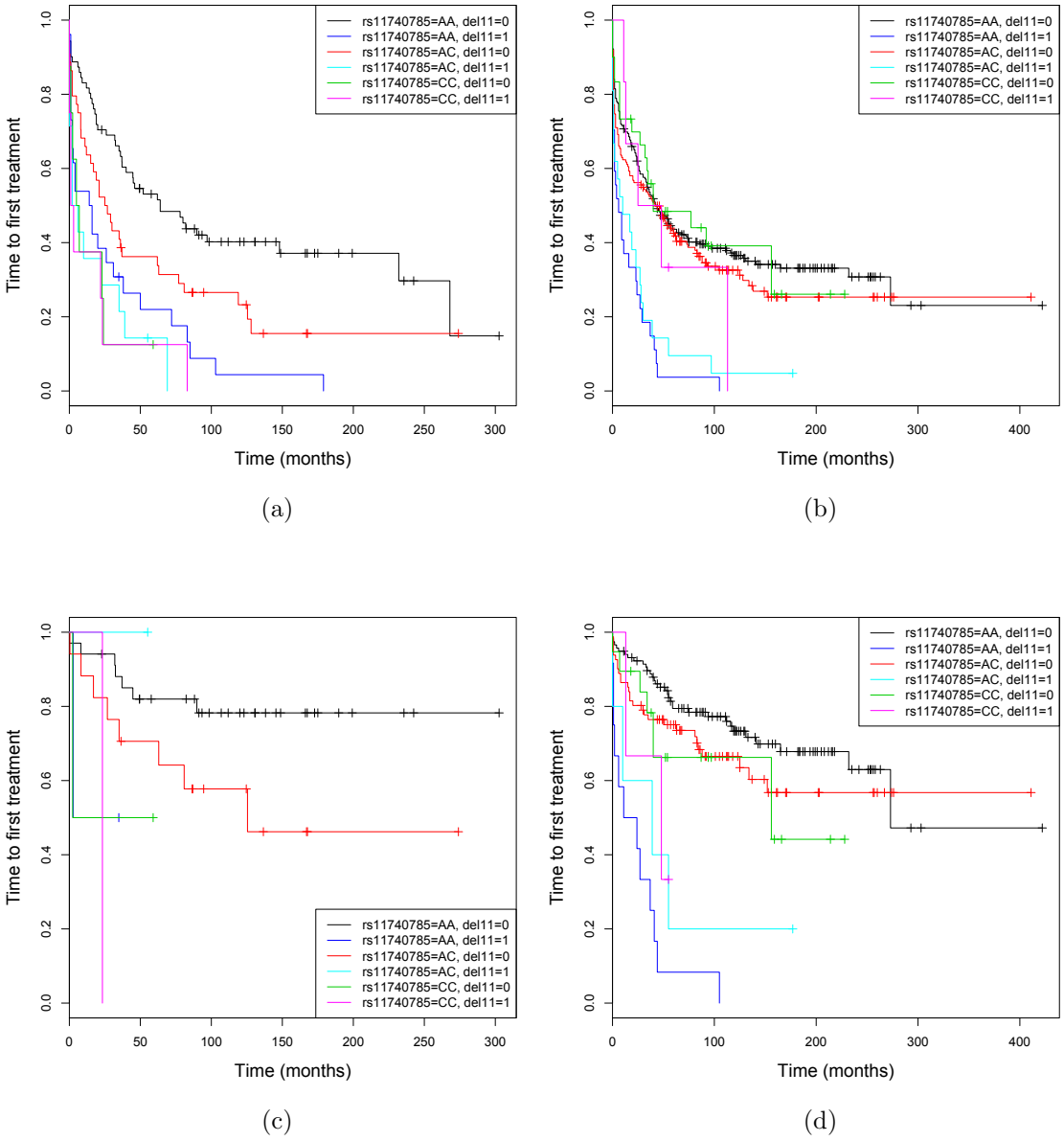


Figure 4.9: Kaplan-Meier curves of the interaction of SNP rs11740785 with deletion 11q for TFT in the discovery cohort (a) and in the replication cohort (b). The direction of effects is not consistent for both groups of patients with and without deletion 11q in these two cohorts. However, when only the subgroup of less advanced patients (locals) is considered, the effect of the SNP is equally prominent between the discovery (c) and replication (d) cohorts.

4.2.3 Further characterisation of LEPR SNP rs3806318

SNP rs3806318 resides in a region with strong signatures of regulatory potential, upstream of the leptin receptor (LEPR) gene (Figure 4.10). In addition, it is not linked to any other SNPs of the region in any of the populations of the 1000 genomes data ($r^2 < 0.3$). Therefore, it was hypothesised that the SNP identified could be the causative SNP of the observed associations.

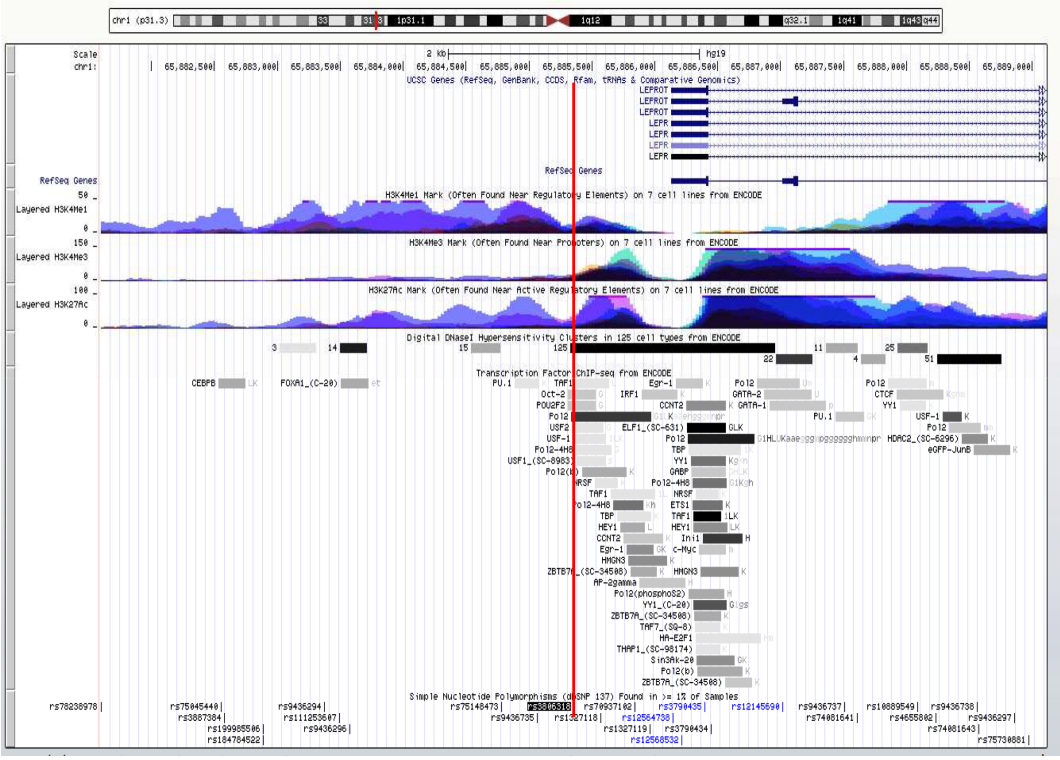
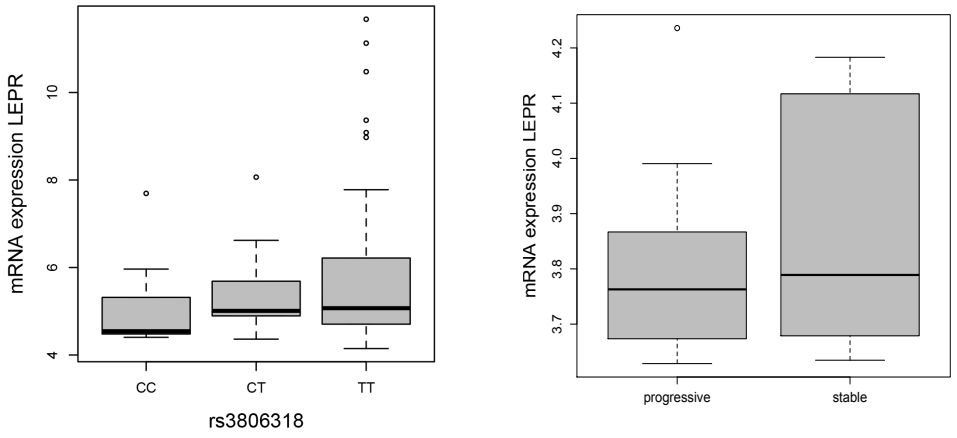


Figure 4.10: SNP rs3806318 lies upstream of the LEPR gene in a region of high regulatory potential, as signified by the enrichment of histone markers and DNase I hypersensitivity markers. In addition, in that region many transcription factors are found to bind, suggesting that a base pair change could alter the binding affinity of a number of transcription factors.

In order to explore the possibility of SNP rs3806318 being a functional SNP, its association with mRNA expression levels of LEPR (or LEPROT) was investigated. In a panel of 72 cell lines of haematopoietic and lymphoid origin (from the CCLE panel, details of which can be found in the Materials and methods), the



(a) (b)

Figure 4.11: In a cell panel of 72 cell lines of haematopoietic and lymphoid origin, the CC cell lines showed evidence of lower mRNA expression of the leptin receptor (LEPR) gene (a). A trend of the association of LEPR mRNA expression levels with status of CLL disease (stable or progressive) in a cohort of 22 patients is shown in (b). This trend suggests that lower mRNA levels could be associated with a progressive disease.

CC genotypes of the cell lines had a trend for lower expression of LEPR (Figure 4.11a). However, the trend was not significant with a p-value of 0.16 (Wilcoxon test). Furthermore, the role of LEPR in CLL progression was examined via publicly available microarray data (from the R package by Whalen). The expression levels of the LEPR gene were compared for samples of 22 CLL patients with two types of disease progression: 14 were progressive patients; and 8 had a stable disease. In this analysis, a trend of decreased expression of the LEPR gene was observed in progressive samples (Figure 4.11b). Again, however, this was not significant. Both datasets examined in this section had a very limited number of cell lines included and no associations could be provided, therefore, a more detailed functional analysis is warranted for the characterisation of rs3806318 and its association with B-CLL progression.

4.3 Discussion

Almost 20,000 SNPs which reside in, or are proximal to, 1,208 cancer-associated genes were interrogated in a cohort of B-CLL patients, using the Variable Ranking and bioinformatics criteria, and five candidate SNPs were identified for their association with TFT, as a marker for progression. One SNP, which lies in a regulatory region upstream of the LEPR gene, was validated in an independent cohort of 601 patients using the log-rank test, but the level of significance of the association was not consistent using the Cox proportional hazards model under different applications and genetic models. This underlines the need for further validation in an independent cohort. In addition, a potential interaction of another SNP of the ITGA1 gene with deletion of the ATM gene (del11q) was identified but an additional validation is needed to verify its association. Nonetheless, this study generates hypotheses on the function of these two genes and their involvement in the progression of B-CLL, which could be of value both as biomarkers and for a better understanding of this disease.

SNP rs3806318 (chr 1p31.3) is located about 1 kb upstream of the promoter of the leptin receptor (LEPR) and leptin receptor gene-related (LEPROT) proteins, which are encoded by the same gene. The LEPR gene encodes the protein LEPR, which is a member of a family of cytokine receptors and is involved in the regulation of fat metabolism. LEPROT is encoded by the LEPR gene, with which it only shares the first two 5' UTR exons, and has been found to negatively regulate LEPR. SNPs in both LEPR and LEPROT have previously been associated with obesity and type 2 diabetes (Loos *et al.*, 2006; Park *et al.*, 2006; Couturier *et al.*,

2007). Obesity has been suggested to promote many types of cancer and a link with CLL has also been shown (Dalamaga *et al.*, 2010; Larsson and Wolk, 2007). However, the CLL cohorts analysed in this study did not have measurements for body mass index (BMI) and so this hypothesis could not be tested here, but will be pursued in further studies.

Interestingly, the LEPR gene is also thought to be involved in haematopoiesis, and to act in the haematopoietic pathway by transducing a proliferative signal in haematopoietic cells (Bennett *et al.*, 1996). Specifically, leptin has been found to induce the expression of Bcl-2 and Cyclin D1 in B-cells, therefore protecting them from apoptosis, and even promoting cell cycle progression under co-stimulatory signals (Lam *et al.*, 2010). These results contradict the observations described here, where a trend of lower mRNA expression of the LEPR gene for the GG genotype is shown, suggesting a link of lower mRNA with a more a progressive disease. However, the sample sizes of the cell line analyses are too low, therefore, a more detailed functional analysis would be necessary for a link between SNP rs3806318 and the expression of LEPR, and ultimately with CLL progression, to be suggested.

SNP rs11740785 (chr 5q11.2) is located in intron 1 of the integrin alpha 1 gene (ITGA1), and belongs to a long haplotype that includes the 5' UTR region of the gene. The PELO gene overlaps the ITGA1 gene at the 5' end, but has not been extensively studied in humans. ITGA1 encodes the alpha 1 subunit of the heterodimeric receptor $\alpha 1 \beta 1$, which is a cell-surface receptor for laminin and collagen. Integrins are involved in processes such as vascular development, haematopoiesis,

immune regulation and homeostasis (Ye *et al.*, 2012). Interestingly, the SNPs of this haplotype have been found to associate with high fasting glucose and high mRNA expression of the ITGA1 gene in a candidate gene meta-analysis study of more than 40,000 individuals (Billings *et al.*, 2012). In more detail, the major allele of rs6867040 (r^2=0.963 with rs11740785) was associated with an additive effect of 0.0166 mmol/L increase in fasting glucose. In addition, the major allele was also associated with higher ITGA1 mRNA expression in liver tissue. Moreover, insulin receptor substrate proteins (IRS1 and IRS2) have been shown to promote carcinoma invasion in an integrin-dependent manner (Shaw, 2001). In relation to insulin's interaction with deletion of the ATM gene (del11q), Saiya-Cork *et al.* (2011) showed an over-expression of the insulin receptor gene (INSR) in patients with the 11q deletion, compared to those without it. They also suggested that higher levels of insulin inhibit apoptosis in CLL cells, and that CLL patients with higher INSR expression had a shorter TFT and overall survival. These differences in INSR expression could provide clues of the mechanism of function of the SNP and to why there could be an interaction of the SNP with del11q associating with B-CLL progression.

The biological model explaining the effect of the balance between insulin and glucose in CLL progression remains unclear and warrants further study. However, the observations reported here, if replicated, would imply that patients without deletion 11q, who are also carriers of the minor allele, would have lower fasting glucose, suggesting a deregulation of the insulin signalling pathway that could lead to faster B-CLL progression (Figure 4.12). A validation of these associations of

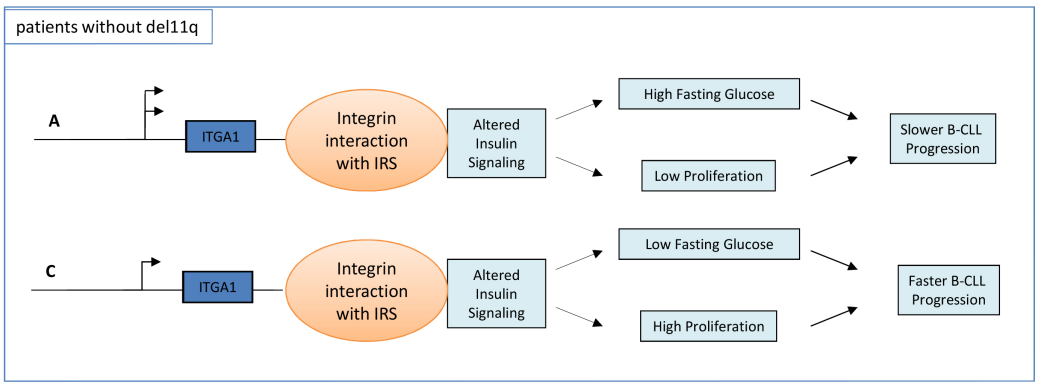


Figure 4.12: Model of the function of the ITGA1 SNP whereby the C allele associates with lower mRNA levels in patients without deletion 11q. High levels of ITGA1 interact with the insulin receptor signalling pathway, producing lower fasting glucose and faster B-CLL progression.

the SNPs of the LEPR and ITGA1 genes could expose a node for intervention, since there are many inexpensive and well-tolerated drugs for diabetes, such as metformin, which are already in the clinic.

One of the limitations of these analyses is that the discovery cohort of this study was slightly different to the replication cohort, in that the later included 90 MBL patients whereas the former had none. However, the thresholds of the diagnostic criteria used to distinguish MBL from CLL patients continue to evolve, with the latest update taking effect in 2008 (Hallek *et al.*, 2008). It has also been shown that up to half the patients now classified as MBL would, with the previous definitions, be classified as CLL and that the concentration of CLL-cells (used for the classification of patients in the MBL group) does not associate with TFT or overall survival (Rawstron *et al.*, 2010; Shanafelt *et al.*, 2008; Kern *et al.*, 2012). Therefore, the MBL patients were included in the replication cohort and were adjusted for in the Cox analysis.

Another limitation of this study is that both the discovery and replication

cohorts comprised two subgroups of patients, the CLL4 patients forming part of a clinical trial study and the locals representing a random sample of patients with CLL. The patients of the CLL4 cohort were selected because they were advanced patients in need of treatment, whereas the locals formed part of a prospective study of whether patients would progress. As such, the two subgroups are intrinsically different and would ideally be dealt with in separate analyses. However, due to the restricted sample size this could not be achieved and the subgroups were joined. This potential source of bias was adjusted for by including `source` as a prognostic factor in the analyses but further replication should be sought in a more homogeneous cohort.

One of the interesting observations of this analysis is the very different results obtained under different modelling for the candidate SNPs. Specifically, for the LEPR SNP rs3806318, assuming a multiplicative model on the hazard ratio, the p-value with the Cox model was 0.22 for the discovery cohort and 0.014 in the replication cohort, whereas the p-values shift to 0.0055 and 0.15 for the discovery and replication cohort, respectively, assuming a recessive model. These differences can be explained by the lack of power to detect the exact type of model in the discovery cohort, due to the low sample size. Therefore, a multiplicative effect on the hazard is the most likely scenario, but additional replication could further define an appropriate model. Moreover, the fact that the SNP doesn't pass Bonferroni correction after adjustment for the prognostic factors with the Cox proportional hazards model (p-value 0.07), but it does if a linear model is applied to the deviance residuals from the Cox model after adjustment for the prognostic factors (p-value

0.0036) complicates interpretation of these results. This strengthens the argument outlined here of using model-free techniques in a GWAS setting where sample size is small. This was also the aim of this analysis, to take advantage of the simplicity of the log-rank test and enhance its power using the random survival model.

Finally, the p-value threshold of the log-rank test selected for this application of the Variable Ranking was considerably higher than the one used in the simulations (p-values of 0.01 and 0.001 respectively). This reflected the decision to be more flexible for the filtering of the SNPs given the low sample size, and because the differences for the two thresholds in the simulations analyses were not substantial for this sample size (see the Appendix).

In this chapter, the Variable Ranking is used to identify two promising SNPs for their associations with TFT, a progression phenotype. It needs to be mentioned that further replication is needed to validate the interaction of the ITGA1 SNP with del11q and to suggest a genetic model for the LEPR SNP. As the next step, both these associations will be tested in an additional cohort of B-CLL patients from Barcelona. To our knowledge, no other study has successfully identified and replicated any associations of SNPs with B-CLL progression phenotypes to date and, as such, this could provide exciting opportunities for their use as biomarkers in future practice.

5 Identification of SNPs associating with chemosensitivity

5.1 Introduction

The second part of this thesis aims to identify SNPs of the p53 network of genes associating with chemotherapeutic response, utilising the NCI60 cell line panel of the Developmental Therapeutics Program (DTP). The NCI60 Human Tumour Cell Line Screen was launched in 1990, and represented a shift in the field of drug discovery from focusing on leukaemia to solid tumours (Shoemaker, 2006). The cell lines used were derived from many types of cancer, including colon, lung, renal, melanoma, ovarian, breast, prostate and central nervous system (CNS) tumours. Thousands of agents were screened for the development of this programme, and the potential of using cell lines for drug discovery was set as an example for the decades to come.

With the advent of high-throughput technologies, two new cell line panels were developed in the beginning of the 21st century, the cell line panels of the Cancer Genome Project (CGP) and the Cancer Cell Line Encyclopedia (CCLE), both of which included more than a thousand cell lines (Garnett *et al.*, 2012; Barretina *et al.*, 2012). As part of the CGP, the Genomics of Drug Sensitivity in Cancer (GDSC) database was developed as an accessible interface for the available drug responses of the cell lines. However, the drugs screened for these panels were focused on targeted agents and included only a handful of conventional chemotherapeutic agents.

There are five main groups of conventional chemotherapeutic agents used in both the NCI60 and GDSC panels and they are classified according to their mode of action. Briefly, alkylating agents are molecules that promote single or double strand breaks in the DNA, as well as DNA crosslinking, causing cells to go into apoptosis or growth arrest. Antimitotic agents are molecules that inhibit mitosis by interfering with microtubule function, and can be divided into two main groups: vinca alkaloids, which inhibit the formation of the mitotic spindle; and taxanes, which inhibit its disassembly. Antimetabolites are molecules that inhibit the metabolic processes within cells. They are S-phase-specific agents that resemble metabolites (purines and pyrimidines) and, as such, become incorporated in the DNA and RNA molecules. This causes incorrect code and strand breaks, halting cell division and growth. Finally, topoisomerases are enzymes that control the coiling and uncoiling of DNA, changing its topology. Topoisomerase I inhibitors obstruct the topoisomerase I and DNA bond, and topoisomerase II inhibitors prevent the re-ligation of the DNA (Kaye, 1998; Pelengaris and Khan, 2013; Nitiss, 2009).

The DTP human tumour cell line screen includes 132 anti-cancer agents that belong to the above families of chemotherapeutic agents, and for which the response of 59 cell lines have been recorded. Responses to most of these families of agents have been shown to correlate well with p53 status, with p53 mutant cell lines responding worse in terms of growth inhibition compared to p53 wild type lines (O'Connor *et al.*, 1997). In additional studies, it has also been shown that depletion of mutant p53 reduces resistance to chemotherapeutic agents (Bossi *et al.*, 2006;

Tsang *et al.*, 2005). These observations highlighted the pivotal role of the p53 gene in the chemotherapeutic response. They also suggest that polymorphisms in the p53 network of genes could also be crucial in the chemotherapeutic response.

Over the last decade, a number of studies have focused on SNPs in the p53 gene and its negative regulator MDM2 in the context of response to chemotherapeutic agents. SNP rs1042522 (codon 72 of the p53 protein) was found to be associated with differential induction of apoptosis (Thomas *et al.*, 1999; Sakamuro *et al.*, 1997), and later an allele-specific apoptotic response to chemotherapeutic agents has been shown in the case of wild type p53 (Sullivan *et al.*, 2004). Moreover, SNP rs2279744 (SNP309 of the promoter/enhancer region of the MDM2 gene) was shown to affect the p53 apoptotic response to many types of chemotherapeutics, including mitomycin C (an alkylating agent) and topoisomerase I and II inhibitors (Bond *et al.*, 2004; Arva, 2005; Nayak *et al.*, 2007).

More recently, the NCI60 panel was used to determine additional SNPs associating with chemotherapeutic response in candidate gene studies of the p53 stress response pathway. These studies resulted in the identification of two SNPs in the YWHAQ and CD44 genes, which associate with differences in chemosensitivity response and the incidence and survival of soft tissue sarcoma (Vazquez *et al.*, 2008, 2010).

The aim of this chapter is to expand these analyses using a custom-made genotyping platform and an alternative ranking method to explore potential associations of genetic variants from the p53 network of genes with chemotherapeutic response in the NCI60 cell line panel. 21 candidates were identified that were

followed up in the GDSC panel. Unfortunately no SNPs were validated in the GDSC panel, but SNP rs4966013 of the IGFR1 gene, which was not available for the GDSC panel, was shown to also associate with IGF1R mRNA and protein expression levels. In addition, a potential interaction with del11q was identified in a cohort of B-CLL patients.

5.2 Results

5.2.1 Identification of SNPs associating with chemosensitivity response in the NCI60 panel of cell lines

For this analysis, 584 SNPs belonging to 114 genes of the p53 network were tested for associations with 126 chemotherapeutic agents, as detailed in the Materials and methods. For each SNP-drug combination, four tests were performed to assess the association of each pair according to three potential genetic models: additive; dominant; and recessive models. The Jonckheere test was performed to test for a monotonic effect of the number of alleles on drug sensitivity. The Wilcoxon rank sum test tested for a dominant or recessive model for the minor allele (grouping the heterozygous with the homozygous of the minor allele or the major allele, respectively) and for significant differences between the homozygous for both alleles. The minimum of the p-values of these four tests was collected for each SNP-drug pair (p'-value). The multiplicity of the tests was corrected by permutations at the next step of the analysis. In cases of low homozygous counts for the minor allele (< 3), the homozygous lines were grouped with the heterozygous ones, and only the test of the dominant model (for the minor allele) was performed. The

response to treatment was not always normally distributed and the variance differed considerably between the agents (Figure 5.1). Therefore, only non-parametric tests were used, to avoid diverse transformations between the agents and for consistency between the analyses.

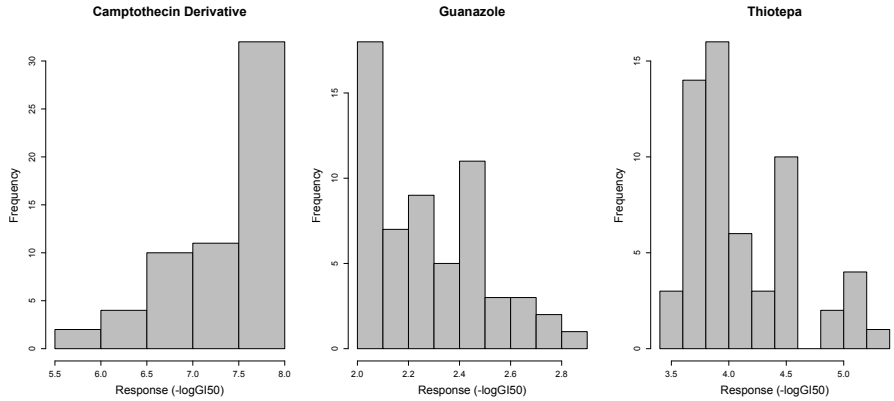


Figure 5.1: A camptothecin derivative, a taxol analogue and thiotepa provide examples of non-normally distributed responses of chemotherapeutic agents. The 3 drugs have different mechanisms of action (they are a topoisomerase I inhibitor, an antimitotic agent and an alkylating agent, respectively) demonstrating that this phenomenon was not restricted to a single group of agents.

For each SNP, the p'-values were used to devise two statistics: firstly, to single out the most significant SNPs for each chemotherapeutic agent; and secondly, to compare the most significant SNPs with a wide range of effects across most drug groups. For the first purpose, all the SNPs were ranked for each drug and the SNPs in the top 5% of each drug (according to their p'-values) were selected. Subsequently, the number of drugs for which each SNP was in the top 5% (top 5% statistic) was calculated. For the second purpose, the number of drugs for which the SNP had signs of association (p'-value< 0.05) was counted, creating a statistic called 'number of drugs' (D).

Interestingly, the results between the top results of the two analyses were very similar. All but one of the SNPs that had a value over the 97.5% quantile of the D

statistic, were also over the 95% quantile of the distribution of the top 5% statistic (total of 15 SNPs). Conversely, all of the 16 SNPs that were ranked as top SNPs for the top 5% statistic (with a value over the 97.5% quantile), also had very high estimates for the D statistic (higher than the 95% quantile of the distribution of D). These thresholds were used to compare the two statistics and not to analyse the chemosensitivity data. In view of the fact that the results between the two statistics were extremely similar, and focusing on SNPs that would have a wide range of effects as part of the p53 network, further analyses were performed using only the D statistic.

Estimation of statistical significance for the D statistic

One important aspect of the chemosensitivity panel is that the cell lines tend to be sensitive to drug groups, creating correlations between the GI50s of different drugs (Figure 5.2). Moreover, some of the drugs included, e.g. the topoisomerase I inhibitors, are very similar to each other structurally, and as such cell lines respond in the same manner to these subsets of drugs. As a result, the associations across SNPs and the chemotherapeutic agents are not independent. Therefore, a bias was observed on the D towards SNPs associating with the drugs of the largest correlated groups. For example, SNPs demonstrating allelic differences for uracil nitrogen mustard (an alkylating agent) were also associated with response to most of the remaining alkylating agents since they are highly correlated.

The minor allele frequency (MAF) of the SNPs was another potential confounder of the D estimates. SNPs with a relatively small MAF ($< 20\%$) would be

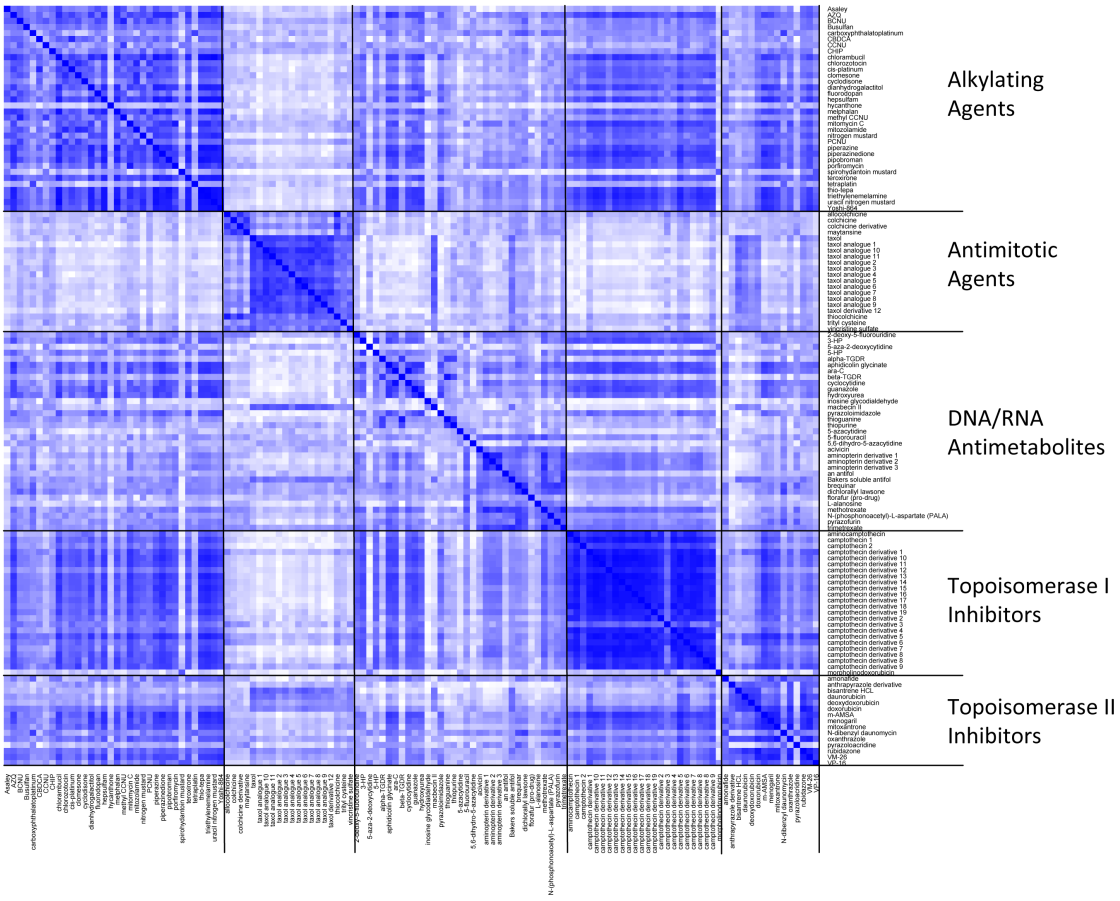


Figure 5.2: Heatmap of the absolute value of the correlation coefficients between all of the chemotherapeutic agents in the NCI60 panel. White signifies no correlation ($r^2 = 0$). The stronger the intensity of the blue colour, the higher the absolute value of the coefficient. The agents within groups show patterns of strong correlations, the most correlated group being the topoisomerase I inhibitors. Most of the alkylating agents are also strongly correlated to topoisomerase I inhibitors.

expected to associate with fewer drugs. This would be due to the lack of power of the study to detect such associations, because of the small numbers of minor alleles homozygous for this sample size (59 cell lines). Hence, the D statistic was not directly comparable between SNPs of different *MAF*.

In order to correct for these two confounders, permutations were performed and the significance of each SNP estimated. For each SNP the genotypes of the cell lines were permutated 1,000 times and the D recalculated. Subsequently, a p-value was estimated per SNP using the empirical distribution of D, under the

null hypothesis. Finally, SNPs were selected that had an estimate more extreme than the 95th percentile of the distribution (p-value< 0.05). These p-values where primarily used for ranking the SNPs according to the supporting evidence, but no multiple hypothesis correction was performed at this stage to adjust for the 584 SNPs.

In Figure 5.3, the distributions of the number of drugs are shown for the SNPs with a nominal p-value. It is notable that, for most SNPs, the 95% quantile of the distribution was approximately 40 drugs. This suggests that a SNP would need to be associated with chemotherapeutic response in at least two groups of agents in order to have been selected, given that the largest group (alkylating agents) included 34 drugs. Exceptions to this observation were 5 SNPs (rs3738948, rs2075677, rs6019621, rs6019618, rs3213150), which were amongst the 6 SNPs with the smallest minor allele frequencies ($MAF < 0.25$), as seen in Table 5.1. As discussed above, this was expected and the distribution of D for small MAF SNPs was less skewed towards large values. In addition, there were 7 SNPs (out of the 23 SNPs) with a HWE p-value smaller than 0.05, a number higher than expected by chance. This could either reflect some systematic genotyping error or differences in ethnicity between cell lines. Unfortunately, no ethnicity information was available for most of the cell lines so no conclusions could be drawn as to the cause of this observation. However, the results should be interpreted with caution due to this observation.

In order to visualise the groups of chemotherapeutic agents that each SNP has significant effects on, a heatmap was drawn (Figure 5.4). Depending on the

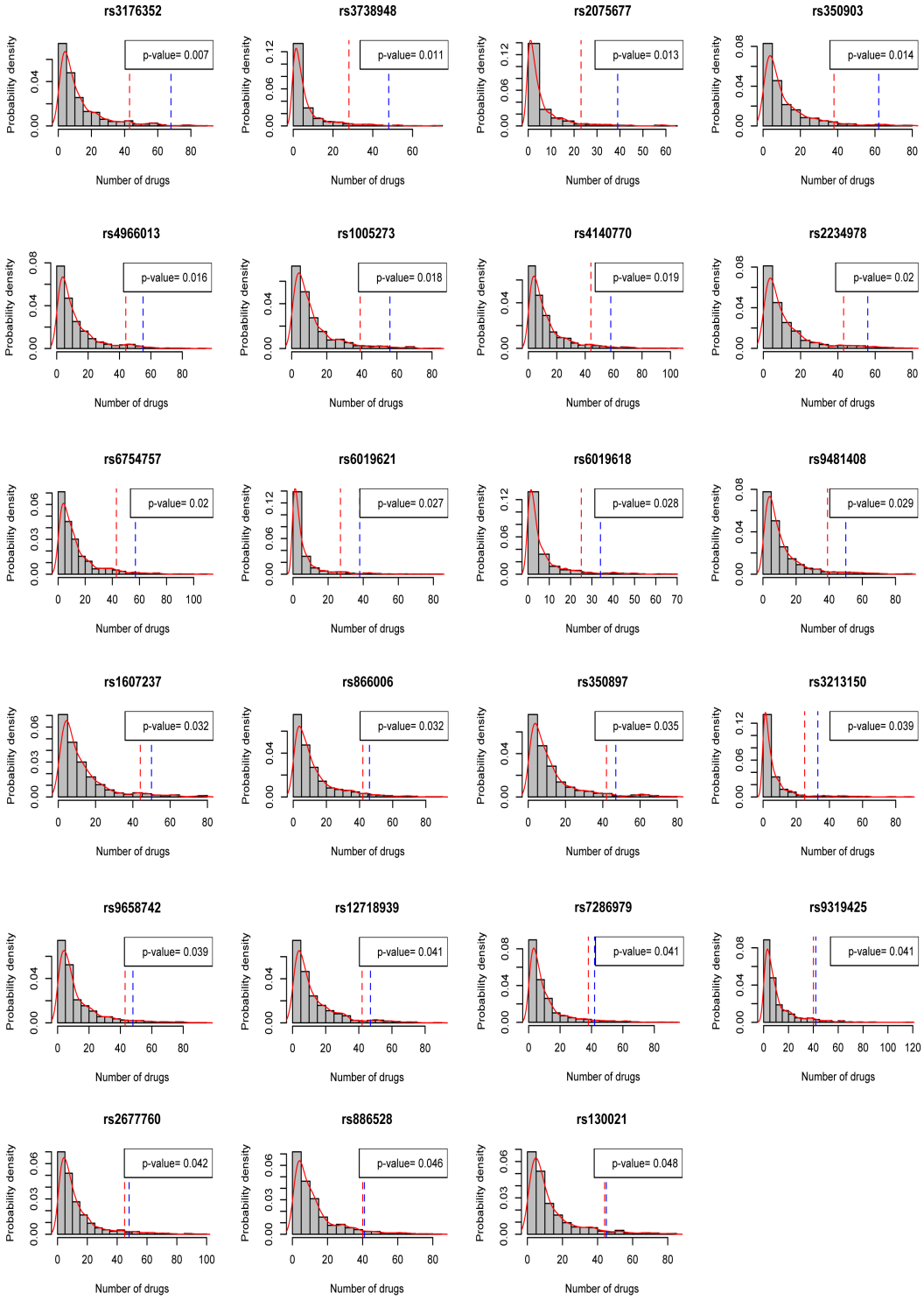


Figure 5.3: Empirical distributions of the D (number of drugs) estimates for the 23 SNPs with a p-value smaller than 0.05 under the null hypothesis. For each SNP, the red line depicts the cut-off value and the blue line the observed d value of the SNP.

SNP	Gene name	Major allele	Minor allele	*MAF*	HWE	Fisher test *p*-value	D *p*-value
rs1005273	PDPK1	G	A	0.44	0.1806	0.2994	0.018
rs12718939	EGFR	G	A	0.36	0.0527	0.0084	0.041
rs130021	CREBBP	A	G	0.34	0.77	0.71	0.048
rs1607237	PIK3CA	A	G	0.35	0.3826	0.1539	0.032
rs2075677	CSE1L	A	G	0.11	0.5372	0.7554	0.013
rs2234978	FAS	G	A	0.20	0.0039	0.1041	0.020
rs2677760	PIK3CA	A	G	0.44	0.1012	0.2047	0.042
rs3176352	CDKN1A	C	G	0.33	0.2416	0.4796	0.007
rs3213150	E2F1	G	A	0.21	1	0.1592	0.039
rs350897	MAP2K2	G	A	0.42	0.0057	0.0303	0.035
rs350903	MAP2K2	G	A	0.48	3e-04	0.0443	0.014
rs3738948	ERCC3	A	G	0.20	1	0.7508	0.011
rs4140770	EGFR	A	G	0.31	0.1191	0.7513	0.019
rs4966013	IGF1R	A	G	0.26	0.1487	0.8241	0.016
rs6019618	CSE1L	A	G	0.20	1	0.4985	0.028
rs6019621	CSE1L	G	A	0.18	0.67	0.65	0.027
rs6754757	TCF7L1	A	C	0.45	1	0.2927	0.020
rs7286979	EP300	G	A	0.38	1e-04	2e-04	0.041
rs866006	APC	C	A	0.32	0.2111	0.3242	0.032
rs886528	CREBBP	G	A	0.48	0.4223	0.4579	0.046
rs9319425	FLT1	G	A	0.39	1e-04	0.0517	0.041
rs9481408	HDAC2	G	A	0.28	0.0483	0.4839	0.029
rs9658742	FAS	C	G	0.31	0.0042	0.1169	0.039

Table 5.1: Table of top 23 SNPs, according to their p-values for the D statistic. 7 SNPs show evidence of Hardy-Weinberg disequilibrium, but none of them had a significant p-value after Bonferroni correction for the 584 SNPs. A Fisher's test was also performed to assess differences in genotype distributions across the tissue of origin.

direction of effects of each SNP, the significant associations are depicted with red and blue bars. A SNP with the same colour bars across all agents demonstrates allelic differences of the same direction for all agents, whereas the direction of effect is not consistent between drugs for those with both blue and red lines. The intensity of colour is proportional to the size of the p-value (the stronger the intensity, the lower the p-value). Two SNPs (rs6019621 and rs130021) had an inconsistent direction of effect for more than 10% of the drugs for which they were significant. In addition, the different directions of effects were not specific to a group of agents, which would have been biologically plausible, but were scattered across the panel

of agents. Therefore, these two SNPs were excluded from any further analysis.

The 21 top SNPs belonged to 17 different haplotypes of 16 genes. For each SNP, a Fisher's test was performed to assess significant differences of genotype distribution between the tissues of origin. However, at this stage, no exclusions were made because it would not be possible to perform a subgroup analysis for specific tissues, due to the small sample size.

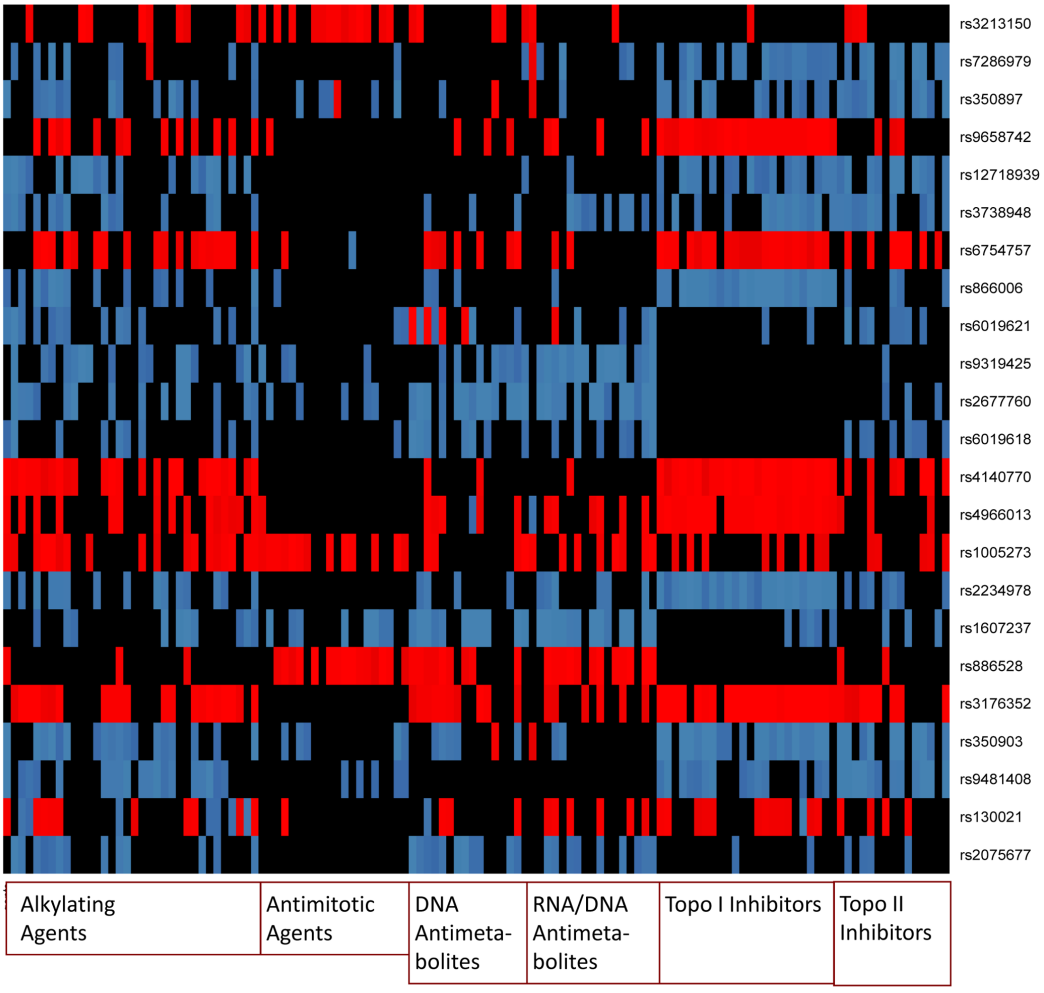


Figure 5.4: Heatmap of the significance of the top SNPs per agent group. The red and blue bars represent the direction of effects for each SNP, and the intensity of colour the strength of association (p-value).

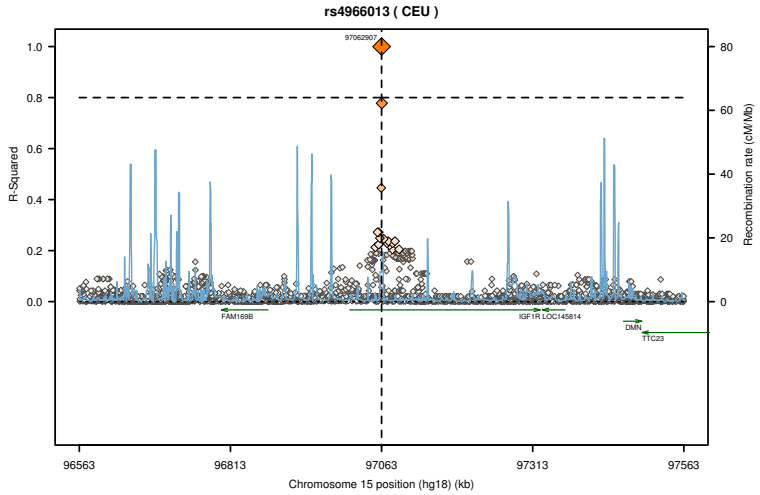
Replication analysis in the GDSC project

The top 21 SNPs from the analysis of the NCI60 panel were selected for further replication of the associations in the Genomics of Drug Sensitivity in Cancer (GDSC) project. However, only 15 SNPs were included (themselves or their proxies) in the Affymetrix 6.0 array and were able to be followed up. An analysis was performed to examine their potential associations with the 13 chemotherapeutic agents of the GDSC project and 3 SNPs showed significant associations with subsets of these agents (a detailed analysis can be found in the Appendix). However, after multiple hypothesis correction, none of the 3 candidate SNPs remained significant.

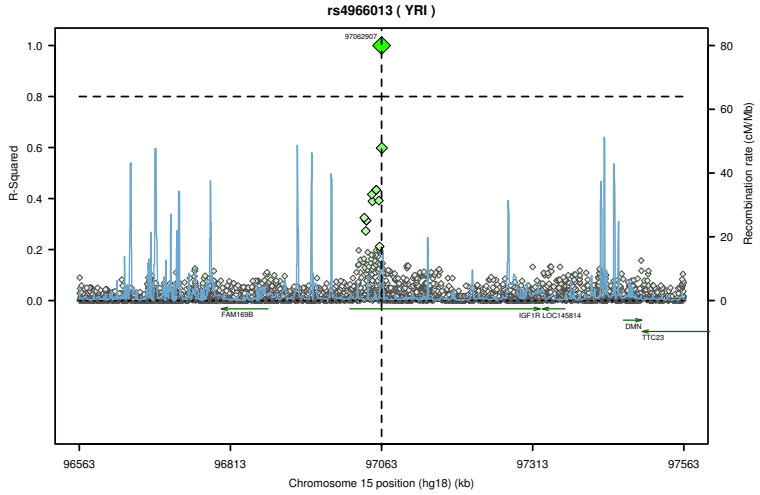
5.2.2 Further analysis of SNP rs4966013

SNP rs4966013 was in the top 5 SNPs of the 21 highlighted by the NCI60 analysis, and also resides in intron 1 of type 1 insulin-like growth factor receptor (IGF1R), a gene that has been previously associated with the cell's response to DNA damaging agents (Turner *et al.*, 1997; Macaulay *et al.*, 2001; Turney *et al.*, 2012). In addition, the SNP is not strongly linked to any other SNPs of the region (Figure 5.5), so it was hypothesised that it was the causative SNP for the allelic differences in drug response. For these reasons, it was followed-up experimentally.

SNP rs4966013 was associated with responses to 55 chemotherapeutic agents, most of them belonging to the groups of alkylating agents and topoisomerase I inhibitors (Figure 5.6). For 52 of the 55 agents, the cell lines that were GG homozygous (6 cell lines) required significantly higher drug concentrations (lower



(a)



(b)

Figure 5.5: Region plots of SNP rs4966013 for CEU and YRI populations, taken from SNAP (Johnson *et al.*, 2008). In neither of the two populations in the SNP is strongly linked ($r^2 > 0.8$) to any other SNP in the region.

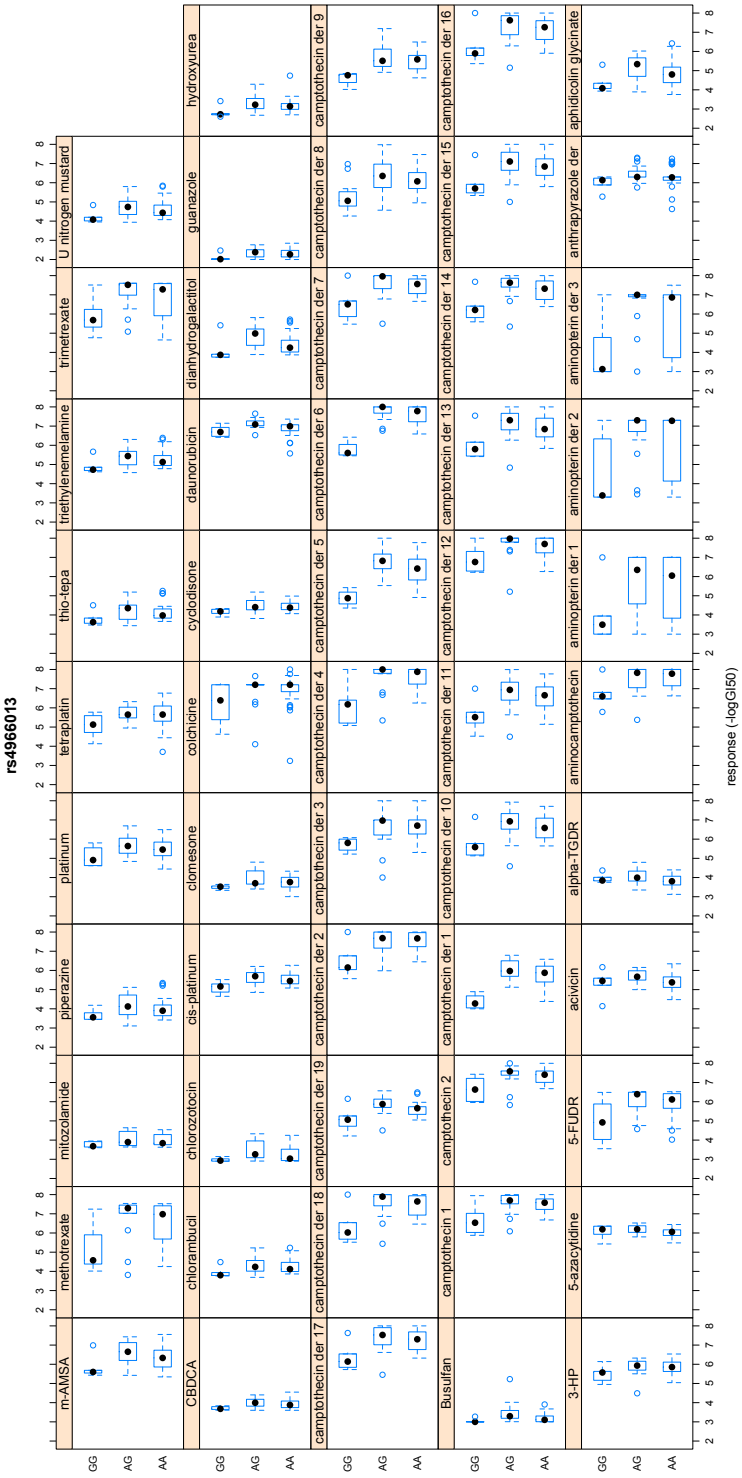


Figure 5.6: Boxplots for SNP rs4966013 and all the chemotherapeutic agents that it associates with. The range of responses vary extensively between chemotherapeutic agents (e.g. for cyclodisone the -logGI50 ranges from 3.8 to 5.2 whereas for colchicine from 3.24 to 8). However, they were all plotted on the same scale, in order for the differences to be directly comparable.

-logGI50), compared to the AA and AG lines (32 and 16, respectively). The mean fold differences between the AA and GG homozygous lines for the 52 drugs was 6.28, ranging from 1.03 to 44.90-fold. These associations did not follow an additive effect, but instead a dominant model for the major allele was prominent for most of the drug associations.

SNP rs4966013 associates with IGF1R inhibitors and chemotherapeutic agents in GDSC

SNP rs4966013 could not be followed up in the GDSC panel because it was not genotyped by the Affymetrix 6.0 array (or had any proxies). However, 53 of the 59 NCI60 cell lines were also part of the cell line panel for the GDSC project. Two IGF1R inhibitors were tested in 50 of these cell lines (BMS-754807 and OSI-906) and another one in 22 out of the 53 cell lines (BMS-536924). As shown in Figure 5.7, for two inhibitors the SNP had the same trend of direction of effects, with GG requiring less drug for an inhibition of 50% (measured with a fluorescence-based cell viability assay). rs4966013 was significantly associated with response to BMS-754807, with a p-value of 0.012 for the A allele carriers compared to those homozygous for the G allele (Wilcoxon test). The median concentration required for 50% inhibition of the GG cell lines was $0.42\mu M$ whereas for the AA or AG cell lines it was $3.01\mu M$. For OSI-906, the SNP showed a similar trend but it was not significant (p-value 0.11). Only 3 cell lines that were homozygous for GG had a recorded drug response for BMS-536924, so no trend was apparent for that inhibitor.

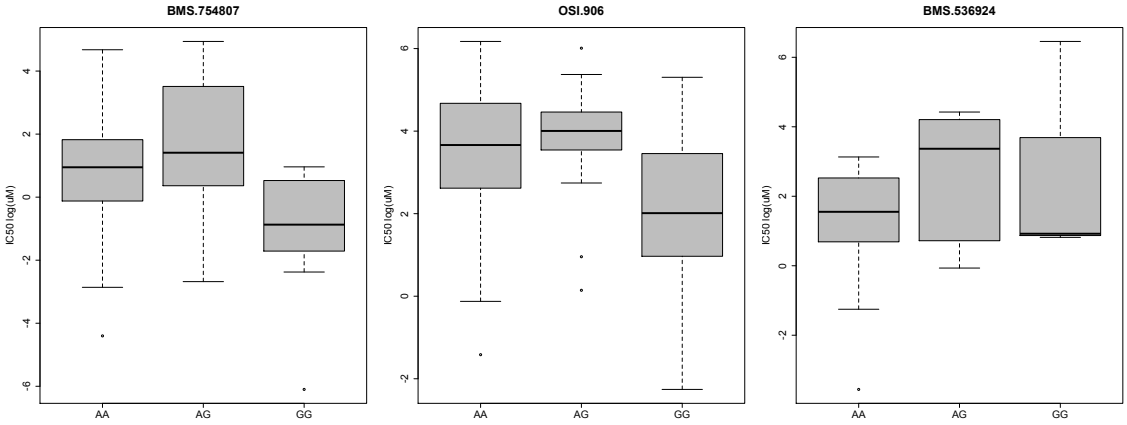


Figure 5.7: Allelic differences for responses to three IGF1R inhibitors; BMS-754807, OSI-906 and BMS-536924. For two of the inhibitors, the SNP shows the same direction of effects, with GG requiring less drug compared to AA and AG. For BMS-539624 no trend was observed, which could be due to the low numbers of GG cell lines (3 cell lines).

SNP rs4966013 associates with expression levels of IGF1R

The mRNA transcript levels were measured for the IFG1R gene in the NCI60 panel by qPCR (by Jorge Zeron, an experimentalist in the lab). Significant associations were noted for rs4699013, where GG was associated with higher mRNA levels (lower ΔCt values) compared to AA and AG (t-test p-value 0.0020, Wilcoxon test p-value 0.0078) or to AA alone (t-test p-value 0.0027, Wilcoxon test p-value 0.0037) (Figure 5.8). This amounted to a mean difference of 2.64-fold ($1.32\Delta Ct$) between the IGF1R transcript levels of the GG cell lines compared to the A allele carriers.

Finally, rs4966013 also associated with differences in relative protein levels for the IGF1R gene (Figure 5.9). The protein levels for IGF1R were downloaded from the Developmental Therapeutics Program (dtp) webpage (`http://dtp.nci.nih.gov/mtweb/`). Following the same trends as for the mRNA levels, the cell lines

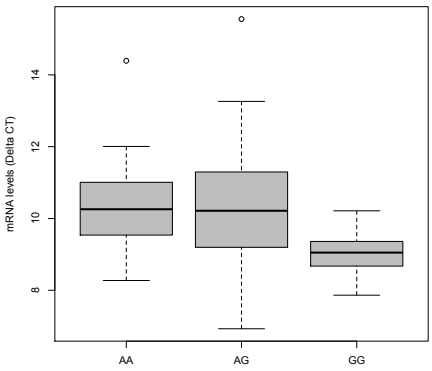


Figure 5.8: Boxplot of mRNA expression levels for rs4966013 for the NCI60 panel. A 2.64-fold difference in the mRNA levels was noted between the GG homozygous cells compared to AA and AG. The values represent ΔCt readings which are inversely proportional to the mRNA levels.

homozygous for the G allele associated with higher IGF1R expression levels and had a 1.44-fold difference of relative expression compared to the AA or AG cell lines, with a p-value of 0.0081 using a t-test (Wilcoxon p-value 0.0062).

Clinical association of rs4966013 with B-CLL progression

The IGF system is an essential regulator of energy metabolism, development and growth, and has been studied extensively for its role in cancer formation and progression (Bartella *et al.*, 2012; Pollak, 2012). IGF1R overexpression has been associated with tumourigenesis, metastatic progression and resistance to therapy (Chitnis *et al.*, 2008; Turner *et al.*, 1997; Spentzos *et al.*, 2007; Parker *et al.*, 2002). Given the associations of the SNP alleles with differing cell line chemosensitivities and mRNA expression levels of IGF1R, associations with aspects of cancer biology linked to IGF1R, such as progression, might be expected. For this reason, the SNP was genotyped in a cohort of 495 patients with B-CLL, to test the hypothesis that the SNP plays a role in cancer progression in the most common type of adult

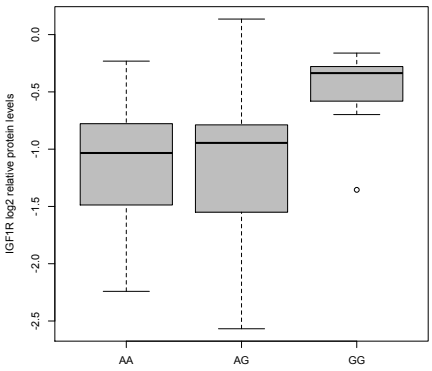
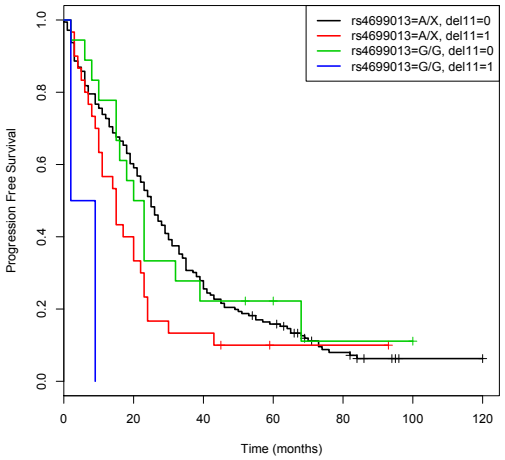


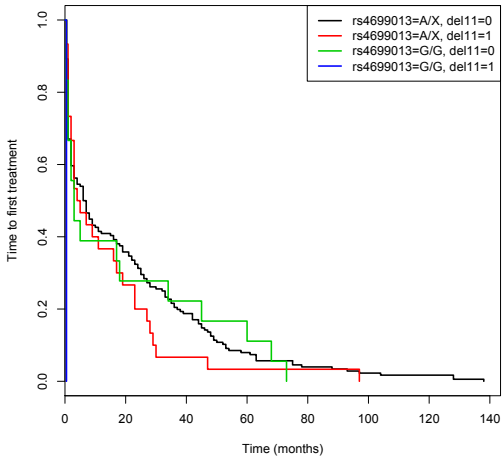
Figure 5.9: Relative IGF1R expression levels associate with rs4966013. The cells homozygous for the G allele expressed 1.44-fold higher levels of IGF1R protein compared to cells carrying the A allele.

leukaemia.

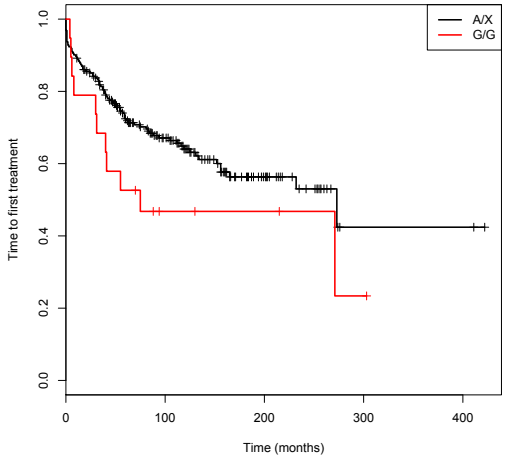
The cohort was comprised of patients with B-CLL from two sources (as detailed in the Materials and methods and Chapter 4). Of the 495 patients, 246 belonged to the CLL4 trial for which progression-free survival (PFS) was measured as a phenotype, time from treatment to relapse, defined either as a need for further treatment / progression or death from any cause. Of the 246 patients, 145 were treated with chlorambucil, 79 with fludarabine and 22 with fludarabine and cyclophosphamide together. The patients homozygous for the major allele were grouped with those heterozygous for the SNP, to look for effects of a similar type as those seen in NCI60. The association of PFS to rs4966013 was not significant when examined by the log-rank test or the Cox proportional hazards model. An exploratory analysis was performed using stepAIC to assess possible interactions in which a potential interaction of the SNP with deletion of the ATM region (del11) was shown (Figure 5.10a) but the p-value of the interaction effect was not significant using a Cox model adjusting for stage, IgVH mutation and treatment arm,



(a)



(b)



(c)

Figure 5.10: Kaplan-Meier survival curves of SNP rs4966013 for two progression phenotypes, progression-free survival and time to first treatment, in a cohort of 495 B-CLL patients. Homozygous patients for the A allele were grouped with heterozygous patients for rs4966013, following the observations of the association of the SNP with chemotherapeutic response. A potential interaction effect was observed for rs4966013 with deletion of the ATM gene (del11) for progression-free survival (p-value 0.055) (a) and time to first treatment (p-value 0.0079) (b) in the subcohort of 246 CLL4 patients. This interaction effect was not significant for time to first treatment in the subcohort of the slowly progressing patients (locals), but the main effect of the association of the SNP with time to first treatment was weakly significant with a p-value of 0.025 (c).

suggesting a possible false positive, due to an inflated type I error of the stepAIC. In order to further explore this observation, the interaction of the SNP with del11q was tested for TFT (time to first treatment from time of diagnosis) and it was shown to be significant with a p-value of 0.0079 (Cox proportional hazards model adjusting for stage and IgVH mutation) (Figure 5.10b). This significant association, however, seemed to be dependent on the two patients that had both the deletion for the ATM gene and the risk G/G genotype. Thus, the results should be interpreted with caution as they can not be conclusive when the p-values depend on only two influential patients. Further replication would be needed to properly assess the validity of this association.

249 patients with a more slowly progressing disease (locals), were not part of the CLL4 clinical trial. Only TFT (time to first treatment) was available as a progression phenotype in this subcohort. The interaction of the SNP with del11 was not significant (p-value 0.21, Cox model adjusted for stage and IgVH mutation). However, the main effect of the SNP was significant under the Cox model adjusted for stage, IgVH mutation and del11, with a p-value of 0.025 (Figure 5.10c).

Overall, in all 495 B-CLL patients, SNP rs4966013 was non-significant for TFT under a Cox model, adjusting for the subcohort (CLL4 or locals), stage, IgVH mutation and del11 (p-value 0.094). However, the interaction of the SNP with deletion of the gene ATM was borderline significant with a p-value of 0.035 (Cox model adjusting for the subcohort (CLL4 or locals), stage, IgVH mutation and del11). Consistent with the chemosensitivity data, GG homozygous individuals progressed faster and required earlier treatment.

Unfortunately, the data is not conclusive regarding the association of the SNP with time to first treatment (TFT) for neither the main effect of the SNP nor for the interaction with del11. The main effect only has a weak association signal for one of the subcohorts, but not with the other, not providing enough evidence of association. Moreover, the potential interaction of the SNP was observed during an exploratory analysis, using a methodology which is known to have an inflated type I error, and was not consistently significant for the rest of the analyses. In more detail, the interaction was only highly significant (p-value 0.0079) for one of the subcohorts where the results seemed to reflect the influence of two patients on the association analysis. Finally, due to the multiple tests conducted in this section for the subcohort analyses, the type I error is expected to be inflated, further weakening the observed associations. Therefore, these results should only be considered as the first step of an association analysis and can not be treated as conclusive evidence without a replication in an independent cohort.

SNP rs4966013 alters a FOXC2 binding site

To characterise the SNP's function, a transcription factor binding site search for the region around the SNP identified a potential forkhead binding motif (Figure 5.11, performed by Jorge Zeron). This predicted that the G allele would have stronger binding with several of the forkhead family members (particularly FOXA, FOXO, FOXC and FOXJ3). To verify this, an electromobility shift assay (EMSA) demonstrated allele-specific binding for FOXC2 (Figure 5.12, Jorge Zeron). When the binding of the *in vitro* translated protein of FOXC2 (lanes 3 and 4) and the oligonucleotide probes with either A or G was compared with the control (lanes

1 and 2), it was apparent that FOXC2 bound the G allele with a much higher affinity than the A allele. In addition, cold competition of the A allele was unable to completely abolish this binding. Together, this data suggests that the G allele binds a member of the forkhead family of transcription factors with higher affinity, enhancing the transcription of the IGF1R gene.

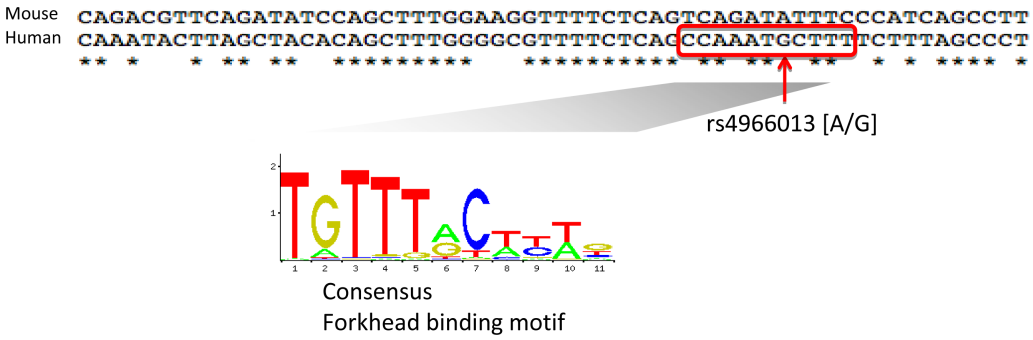
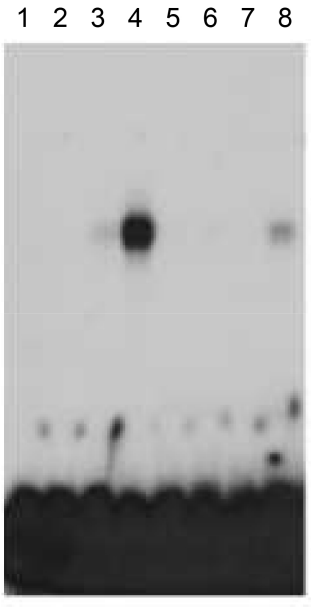


Figure 5.11: The region where SNP rs4966013 resides, is where transcription factors of the forkhead family were predicted to bind. The stars signify the conserved base pairs between the human DNA sequence and the mouse sequence. The G allele (C counterpart) is the one predicted to bind to the forkhead family members.

5.3 Discussion

In this chapter, more than 500 SNPs of the p53 stress response network of genes were analysed for differences in sensitivity to 126 well-characterised chemotherapeutic agents in the NCI60 cell line panel. Fifteen candidate SNPs were followed up in the CGP-GDSC panel, and three showed signs of association with chemotherapy response but were not significant after Bonferroni correction. However, the potential significance of rs4966013, which lies in intron 1 of the IGF1R gene, was further supported by associations of the SNP with response to an additional IGF1R inhibitor, as well as with mRNA and relative protein levels of the IGF1R gene. The association of the SNP with two progression phenotypes (progression free survival



Lane	Protein & Competitor	Labeled Oligo
1	Ribosome lysate	A allele
2	Ribosome lysate	G allele
3	FOXC2 DBD	A allele
4	FOXC2 DBD	G allele
5	FOXC2 DBD + cold A allele	A allele
6	FOXC2 DBD + cold G allele	G allele
7	FOXC2 DBD + cold G allele	A allele
8	FOXC2 DBD + cold A allele	G allele

Figure 5.12: Electromobility shift assay (EMSA) with the *in vitro*-translated protein FOXC2. Lines 1 and 2 represent the controls of the ribosome lysate with the A and G oligonucleotide probes and lines 3 and 4 include the FOXC2 *in vitro*-translated protein. The G allele has a higher affinity which is competed away with the cold G oligonucleotide probe (line 6), but not with the A (line8).

and time to first treatment) was tested in a cohort of B-CLL patients but the results were not conclusive. Finally, allele-specific binding to the transcription factor FOXC2 was shown by a protein DNA interaction assay.

IGF1R's role in cancer progression and survival has been the focus of research for more than two decades (Gualberto and Pollak, 2009). IGF1R is a cell surface receptor tyrosine kinase, which becomes activated by the ligands IGF-I and IGF-II. It signals via two main pathways, the PI3K-AKT and Ras-Raf-MAPK pathways, two of the most frequently dysregulated pathways in cancer (Liu *et al.*, 2009; Montagut and Settleman, 2009). Under normal conditions, IGF1R has been shown to regulate cell cycle progression at several points, most importantly by increasing cyclin D1 synthesis. However, in malignancies it has been shown to lead to proliferation and apoptosis protection (Chitnis *et al.*, 2008; Samani *et al.*, 2007).

In epidemiological studies, elevated levels of IGF-I have been associated with an increased risk of developing solid tumours and adverse prognosis in many cancers, including lung, colon and prostate cancer (Samani *et al.*, 2007; Chitnis *et al.*, 2008; Pollak, 2008). The role of IGF1R in cancer has also been tested in *in vivo* models, where it was shown that overexpression of IGF can confer an increase in tumour incidence and progression, and that mice with genetic mutations resulting in IGF1 underexpression have slower tumour growth (Samani *et al.*, 2007; Pollak, 2008).

In addition, there is increasing evidence supporting the role of IGF1R expression in response to UVB and ionising radiation, and DNA damaging chemotherapeutic agents. IGF1R overexpression was associated with a reduced response to radiotherapy in a cohort of breast cancer patients (Turner *et al.*, 1997). Furthermore, whilst IGFI administration reduced the drug sensitivity of breast cancer cells lines, IGF1R depletion sensitised prostate cancer cells to radiotherapy and DNA damaging chemotherapeutic agents (Dunn *et al.*, 1997; Rochester *et al.*, 2005; Turney *et al.*, 2012). To date, insulin-like growth factor I (IGFI) and the receptor family (IIRF) have been the targets of more than 100 clinical trials, many of which are still ongoing. However, the results have been disappointing and it has been suggested that the lack of predictive biomarkers is one potential reason for this failure (Pollak, 2012).

SNP rs4966013 demonstrated allelic differences in response to many chemotherapeutic agents and an IGF1R inhibitor, BMS-754807. BMS-754807 is a small molecule Receptor Tyrosine Kinase Inhibitor (RTKI) that competes for the ATP

binding site of IGF1R, obstructing its ability to activate the PI3K-AKT and ERK (part of RAF-MAPK) pathways (Dinchuk *et al.*, 2010; Heidegger *et al.*, 2011). One interesting observation was the opposite direction of associations of rs4966013 with the chemotherapeutic agents of NCI60, compared to the IGF1R inhibitors for the GG homozygous cell lines. These cell lines demonstrated increased IGF1R protein expression. Cells that overexpress IGF1R are often dependent on it for proliferation, thus, they may be more sensitive to its inhibition than cells that do not express it. On the other hand, IGF1R is important for proliferation and cell survival, and therefore higher expression of this receptor may result in resistance to the cytotoxic effects of chemotherapy.

The SNP showed weak signs association with time to first treatment for the subgroup of 'locals' patients, but not for progression after treatment in the CLL4 group. All three cytotoxic agents administered to the CLL4 group were alkylating agents, and rs4966013 had shown significant results for many of the alkylating agents of the NCI60 panel, including chlorambucil, which was one of the treatments of the three groups. However, the variety of treatments adds heterogeneity to the group, which could explain why looking at all the patients together did not give any significant results. On the other hand, there was only a small group of minor allele homozygous patients for each arm of treatment, notably 13 patients for chlorambucil, 10 for fluradabine and 2 for the combination of fluradabine with cyclophosphamide, so there was not enough power to detect any associations within either group.

In addition, in an exploratory analysis, an interaction of rs4966013 with dele-

tion 11q of the region of ATM was observed for time to first treatment but there was not enough evidence to prove this association. Nonetheless, a link between downregulation of the IGF1R gene and a defective ATM function after irradiation has been previously reported (Macaulay *et al.*, 2001), and so this interaction would be interesting to follow up in an additional cohort.

Although cell line panels have been the topic of extended debate since the first panel (the NCI60) was developed, they have played an essential role in cancer research. However, acquiring genotypic data in cancer cell lines can be susceptible to larger genotyping errors compared to patient cohorts, because of their genomic instability. Duplications, deletions and loss of heterozygosity can introduce bias to the results. Moreover, the ethnic origin of the cell lines is not always available and, as such, population stratification can also be a confounder in the analysis, a topic which has not been discussed widely in the context of the NCI60 panel. Finally, cell line panels inherently come with many limitations such as heterogeneity between tumour types, which can add an additional layer of confounding effects, and low samples sizes. In this study, the potential hits of the results were regenotyped to verify the validity of the genotypes and CNV data was used, where available, to correct for events of genomic instability. Unfortunately, CNV data was not available for NCI60, but only for CGP, and so the results with rs4966013 were not corrected in this manner. Moreover, the results were not corrected for potential population stratification. In most common scenarios, the first principal components would have been added to the models to adjust for population stratification. However, in this analysis no parametric modelling was possible to be done because of the huge

differences in distributions between the chemosensitivities of the cell lines, which meant that different transformations would have to be performed for the various phenotypes, complicating the interpretation of the results. However, high rates of departures from HWE have highlighted the potential dangers of this problem and in hindsight, this should have been addressed. Finally, a Fisher's test was performed to assess potential differences in genotype distributions due to the differences in the tissue of origin and replication was sought in an additional cell line panel to address the problem of the high false positive rates due to the low sample size. Additionally, an effort was done to validate one of the candidates in an independent system, namely a cohort of B-CLL patients.

To conclude, a potential biomarker for response to treatment is discovered from the NCI60 cell line panel analysis. Its association with time to first treatment as a marker of progression is also explored in a patient cohort but further replication is needed for conclusive evidence. A hypothetical model, however, is proposed by which one of the forkhead transcription factors binds preferentially to the G allele and activates transcription of the IGF1R gene, increasing its protein levels (Figure 5.13). This could manifest itself as resistance to chemotherapeutic agents, suggesting a potential therapeutic node of intervention with targeted therapies of IGF1R inhibitors.

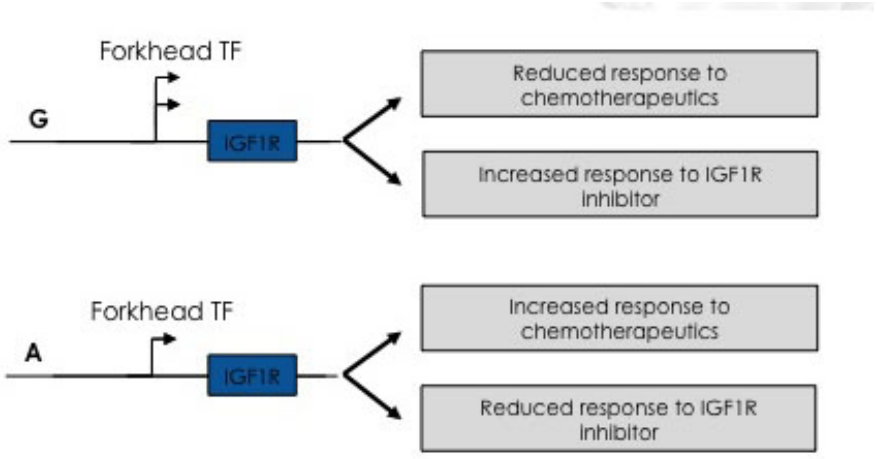


Figure 5.13: A member of the forkhead family of transcription factors binds preferentially to the G allele of SNP rs4966013, activating transcription of the IGF1R gene which leads to a reduced response to chemotherapeutics and an increased response to IGF1R inhibitors.

6 SNPs in E-box binding motifs and the transcriptional regulation of cancer genes

6.1 Introduction

As previously mentioned, the aim of this thesis is to design and implement strategies, which combine the identification of SNPs associated with cancer with a bioinformatics analysis, to better understand cancer-associated genomic loci and accelerate their incorporation into the clinic. In the previous chapters, three different cohorts of patients were analysed to find polymorphisms that could be contributing to cancer progression and metastasis: B-CLL; melanoma patients; and pancreatic cancer patients. In the following two chapters, the potential roles of SNPs in affecting the transcriptional regulation of key cancer genes resulting in differential cancer risk and progression is explored. This chapter specifically focuses on SNPs in the E-box transcription factor binding motif, which has been shown to be extremely sensitive to single base pair changes.

E-boxes are transcription factors binding motifs that consist of a core hexanucleotide sequence motif, CANNTG (Massari and Murre, 2000). The proteins that recognise the E-box motif belong to the helix-loop-helix (HLH) family of transcriptional regulatory proteins, and were first discovered as intronic enhancers of the immunoglobulin heavy-chain (IgH) gene, which shares the E-box signature motif (Ephrussi *et al.*, 1985). Transcription factors that recognise the E-box motif have been found to act as both transcriptional enhancers and repressors of genes that are involved in a number of crucial developmental processes, such as heart and pan-

creatic development, haematopoiesis and neurogenesis. The E-box DNA binding domain is bound by a family of transcription factors called E-box binding proteins. The family includes over 240 transcription factors, which are divided into classes based on tissue distribution and DNA-binding specificity, and usually contain a basic helix-loop-helix structural motif (bHLH). The bHLH-containing proteins are highly conserved and are found in many organisms from yeast to humans (Massari *et al.*, 1996; Massari and Murre, 2000; Jones, 2004).

Apart from being highly conserved, the E-box binding motif is important because a single base pair change can dramatically alter the binding affinity and transcriptional activation of genes that are directly regulated by it. In addition, recent studies have discovered SNPs in E-boxes that affect the immune response to inflammation, skeletal muscle and the response to an anticoagulant called warfarin (Szalai *et al.*, 2005; Teng *et al.*, 2009; Wang *et al.*, 2008). Therefore, it was hypothesised that those SNPs that affect the E-box binding motifs in genes known to be involved in cancer may alter their transcriptional activation, and play a role in tumourigenesis and cancer progression.

Over the last decade, the role of bHLH proteins has been studied extensively in cancer research and it has been shown that some members of the HLH family act as tumour suppressors, while others act as oncogenes. For example, E2A is a transcription factor that is crucial for lymphocyte development, and its loss has been shown to enhance proliferation and cell cycle progression in T-cell lymphoma (Steininger *et al.*, 2011). The inhibitor of DNA binding (Id) family of HLH proteins, on the other hand, are members of the family that lack the basic domain and so

cannot bind to DNA. Instead, they form protein-protein dimers with other members of the family, preventing them from binding to DNA. For example, ID2's binding to the retinoblastoma (RB) protein induces cell cycle progression (Lasorella *et al.*, 2000; Perk *et al.*, 2005).

In order to identify SNPs in E-boxes that could be affecting cancer risk and progression, the search was focused on genes that are part of the pathways that are involved in cancer, as defined by the KEGG database (Kanehisa and Goto, 2000). These pathways include the cell cycle, apoptosis and p53 signalling. Therefore, a SNP affecting the E-box-dependent regulation of these genes could have a direct role in tumourigenesis. Subsequently, and similarly to the methods utilised in previous chapters, ENCODE data was used to identify the E-box SNPs that actually reside in genomic regions with regulatory potential, in order to distinguish the true signal from the noise. This was an important step, since the E-box binding motif relies on four base pairs only, and so one would expect that many SNPs would lie in a site that resembles an E-box binding motif but which has no functional value. Seven candidate SNPs were identified that were then taken forward for experimental validation using electromobility shift assays (EMSAs). The four most promising SNPs were genotyped in a melanoma patient cohort and were examined for associations as eQTLs. Finally, the association of SNP in PPP3R1 with overall survival was explored in melanoma and was shown to have allelic differences for mRNA expression levels of PPP3R1 in a melanoma cell line panel, suggesting a biological role for this SNP in the metastatic progression of melanoma.

6.2 Results

6.2.1 Identification of SNPs residing in cancer genes that can create E-box elements

In order to identify SNPs in cancer genes that could be creating or abolishing an E-box element, the following methodology was developed: (i) the SNPs residing in the cancer genes of interest were retrieved; (ii) a pattern search was performed to identify SNPs that could reside in potential E-boxes; (iii) mono-allelic and multi-allelic SNPs and SNPs without reported MAF were excluded; (iv) SNPs for which one of the two alleles was creating an E-box were selected; (v) SNPs with $MAF>$ 10% (in CEU) were selected; and (vi) SNPs in highly evolutionarily conserved regions were selected.

In more detail, all reported SNPs were first retrieved from the chromosome reports (NCBI dbSNP), following which the SNPs were filtered according to their positions. Only SNPs belonging to pathways that are involved in cancer causation and progression, such as the cell cycle and apoptosis, were of interest. Therefore, only the SNPs that lay in the 1,168 genes of the 15 cancer associated pathways were retained (as defined by the KEGG pathway database, see Materials and methods). This resulted to 387,966 SNPs residing in the genes of interest according to the NCBI annotation, which includes the gene body as well as 2kb upstream of the transcriptional start site and 0.5kb downstream from the transcriptional end site (Bet, 2005).

In order to compare the known binding sequence of an E-box (CANNTG) to

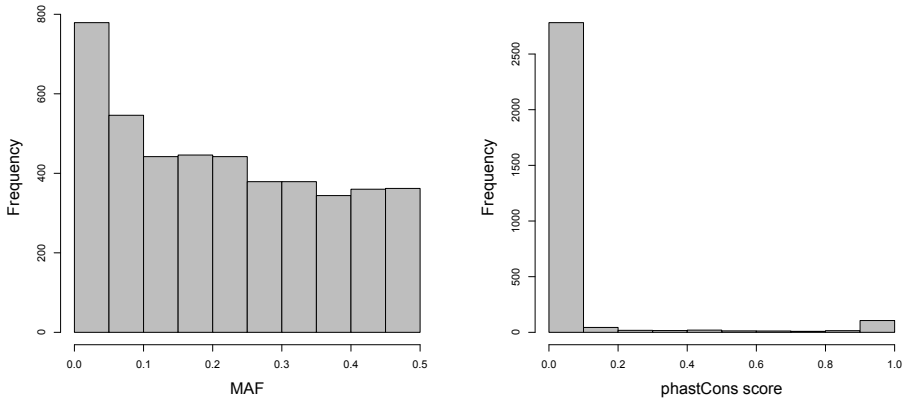
the DNA sequences of the region of each SNP, flanking sequences were extracted (5 bp on either side) for each variant. Performing a pattern search for each of the 4 possible positions that the SNP could occupy (for each strand), 39,885 candidate SNPs were extracted that belonged to one of the 8 sequence combinations (Table 6.1). From these, 28,219 had been validated using a non-computational method or by having frequency data (Sherry *et al.*, 2001), and were kept for further analysis.

E-box on positive strand	E-box on negative strand
xANNTG	xTNNAC
CxNNTG	GxNNAC
CANNxG	GTNNxC
CANNTx	GTNNAx

Table 6.1: Table of 8 possible sequences where a predicted E-box can be found. x symbolises the position of the SNP and N is for any base.

Next, the alleles and the frequencies of the SNPs were retrieved from the NCBI database and, after further filtering for bi-allelic SNPs with a reported allele frequency for the HapMap CEU population, 7,205 SNPs were retained. From these, 4,241 SNPs contained an allele creating an E-box element. However, a large proportion of these SNPs had a very low *MAF* (Figure 6.1a). In order for the SNPs to be able to be followed up experimentally or in patient cohorts, only SNPs with *MAF* equal to or over 10% (in CEU) were selected. After filtering for low *MAF*, 3,036 SNPs remained in the study.

Finally, SNPs were screened for those residing in evolutionarily conserved regions using the *phastCons* score, a cross-species conservation tool that uses multiple alignments of five vertebrate species: human, mouse, rat, chicken and *Fugu rubripes* (Siepel *et al.*, 2005). Functional regulatory elements, and specifically E-boxes, are often highly conserved, and so the aim of filtering for SNPs in conserved



(a) (b)

Figure 6.1: Histogram of the minor allele frequencies (*MAF*) and conservation scores of the candidate SNPs. 4,241 SNPs were deemed as potential E-box binding sites, but only 3,040 had a *MAF* higher than 10% in CEU (a). Distribution of the conservation *phastCons* score for the 3,036 SNPs with high *MAF* (b). 58% of the candidate SNPs had a conservation score of 0. Only the top 5% had a score above 0.5 and were selected to be taken forward.

regions was to aid in the identification of functional transcription factor binding sites (Whitfield *et al.*, 2012; Massari and Murre, 2000). The histogram of the conservation scores for the 3,036 SNPs is shown in Figure 6.1b. More than 50% of the SNPs had a conservation score of 0 (1792 SNPs), and only 5% had a score of over 0.5, indicating cross-species conservation. Therefore, the 156 SNPs having a *phastCons* score above 0.5 were selected for further analysis.

6.2.2 Bioinformatics and functional analysis of candidate E-box SNPs

To further narrow our SNP pool, the SNPs were screened to exclude exonic SNPs, since they would be less likely to have evolved to be transcriptional regulatory elements. Additionally, another member of the lab and a close collaborator, Jorge Zeron, filtered the SNPs based on their regulatory potential. Briefly, he scored the SNPs based on ENCODE evidence of functional regions. Moreover, the SNPs were

prioritised if they, or their linked SNPs, had been found to associate with a trait as recorded in the NHGRI catalogue of GWAS. Finally, the candidate E-box SNPs were followed up experimentally by Jorge Zeron, to determine whether differential binding would be observed in electrophoretic mobility shift assays (EMSAs). Consequently, 52 of the 156 SNPs were excluded from further analysis because they were exonic. This enrichment in exonic SNPs could be expected due to the filtering for conserved regions and the pressures on protein coding sequences.

Two criteria were used for shortlisting the remaining 104 SNPs, using publicly available data from the ENCODE project on markers of regulatory regions and from the GWAS catalogue on associations with human traits. One of the markers of regulatory regions is their presence in chromatin-precipitated DNA using antibodies of known transcription factors, and the ENCODE project has defined these regions for many transcription factors. Taking advantage of the transcription factor ChIP-seq data, SNPs were screened for those that resided in (or are close to) regions where transcription factors with the bHLH motif (Myc, Max, USF1, Hey1) have been found to bind (Hudson and Snyder, 2006; Euskirchen *et al.*, 2007). Four SNPs were selected on this basis (Table 6.2).

The second criterion used was whether the SNPs (or any strongly linked SNPs) had been found to be associated with GWAS traits and reside in DNA regions with high regulatory potential as defined by: a) being enriched for histone modification marker H3K4Me3 as a sign for promoter regions; b) being enriched for histone modification markers H3K4Me1 and H3K27Ac for enhancer regions; c) having evidence of DNase I hypersensitivity; and d) having high regulatory

potential scores (as defined by the regulatory potential (RP) scores from alignments of human, chimpanzee, macaque, mouse, rat, dog, and cow (Kolbe *et al.*, 2004; King *et al.*, 2005)). Five SNPs had been associated with a trait from a GWAS (see Materials and methods) supporting the functionality of these SNPs. Even though the GWAS were not necessarily related to cancer phenotypes, the fact that the genes are known to be involved in tumourigenesis and cancer progression would suggest that a functional SNP of these genes could also be affecting these processes. Three of the five SNPs also reside in regions with signs of regulatory potential (Table 6.2). Specifically, SNP rs3807989 was associated with heart rate and PR interval (Holm *et al.*, 2010; Pfeufer *et al.*, 2010), whilst SNPs rs11635424 and rs191777 were linked with an $r^2 = 1$ to SNPs (rs12593813 and rs2040494, respectively) associating with restless leg syndrome (p-value 2.5E-10, Winkelmann *et al.* (2007)) and height (p-value 4E-07 Lango Allen *et al.* (2010)). All three of these SNPs had some enrichment for the histone modification markers H3K4Me1 and H3K27Ac, two of them were in a DNase I hypersensitive region and two had a high score for ESPERR. Using these two criteria, 7 SNPs were identified that were deemed worthy of further experimental validation and exploration.

Furthermore, the ability for the predicted allele to bind to transcription factors with greater affinity was explored. To do this, DNA-protein binding assays (Electromobility shift assays, EMSAs) using cellular nuclear extracts which contain many transcription factors were conducted with a breast cancer cell line (MDA-MB-231) and a human embryonic kidney cell line (HEK293). Interestingly, 5 of the 7 SNPs demonstrated significant allelic differences in the protein-DNA complexes

SNP ID	Location	bHLH ChIP-seq	GWAS	Promoter ENCODE	Enhancer ENCODE	DNaseI ENCODE	ESPERR
rs10415219	Intron 1	USF1	No	0.1	1	1	0.5
rs681271	Intron 3	Max	No	0	1	1	1
rs1875455	Intron 1	c-Myc, Max	No	0.5	0.2	1	1
rs2029091	Intron 1	c-Myc, Max, Hey1	No	1	1	0	0
rs11635424	Intron 14	NA	Yes	0	0.3	0	1
rs191777	Intron 4	NA	Yes	0	0.3	1	0
rs3807989	Intron 2	NA	Yes	0	0.3	1	1

Table 6.2: Summary of the bioinformatics selection criteria for the screening of the 104 SNPs. Four SNPs resided in regions where bHLH transcription factors have been found to bind, and three SNPs were selected as residing in haplotypes identified from a GWAS and having regulatory potential.

that were noted in either of the nuclear extracts used. Specifically, in Figure 6.2, the allele-specific binding of 5 of the 7 SNPs is clearly seen. Importantly, for the 5 SNPs in ACTN4, FGF12, PPP3R1, CDK6 and CAV1, the allele with greater amounts of protein DNA complexes was the predicted E-box (Table 6.3). In order to explore the potential presence of a bHLH protein in these complexes, four commercial antibodies to well-described bHLH proteins (MYC, MAX and USF1 and USF2) were utilised in an attempt to super-shift the noted bands. This was performed for four of the five SNPs, which presented the strongest binding. Interestingly, protein-DNA complexes formed in the gel-shifts of the E-box alleles for two of the four SNPs shifted with USF antibodies (Figure 6.3). Specifically, a shift of the band that the A allele of the SNP in CAV1 was forming was shifted with the antibody of USF1, and that of the T allele of the SNP in FGF12 with the antibody of USF2. Together, these data lend strong support for our methodology of identifying polymorphic E-boxes that can affect the binding of bHLH proteins.

SNP	Gene	Alleles	Sequence	Binding allele	Minor allele	*MAF*	PhastCons score
rs10415219	ACTN4	A/G	cagctNcctta	G	G	0.45	0.914
rs11635424	MAP2K5	A/G	atttcNtatgt	A	A	0.32	0.813
rs1875455	FGF12	C/T	tcataNgcctg	T	C	0.35	1
rs191777	CDK6	A/T	tggacNgctga	A	A	0.49	0.966
rs2029091	PPP3R1	C/T	tgactNacctg	C	C	0.36	0.932
rs3807989	CAV1	A/G	tcaacNtgtgc	A	A	0.44	0.969
rs681271	CACNA1E	C/T	tccagNacctg	C	C	0.44	0.863

Table 6.3: Summary of predicted binding information for the seven candidate SNPs. All the selected SNPs were of high *MAF* and with a high PhastCons score.

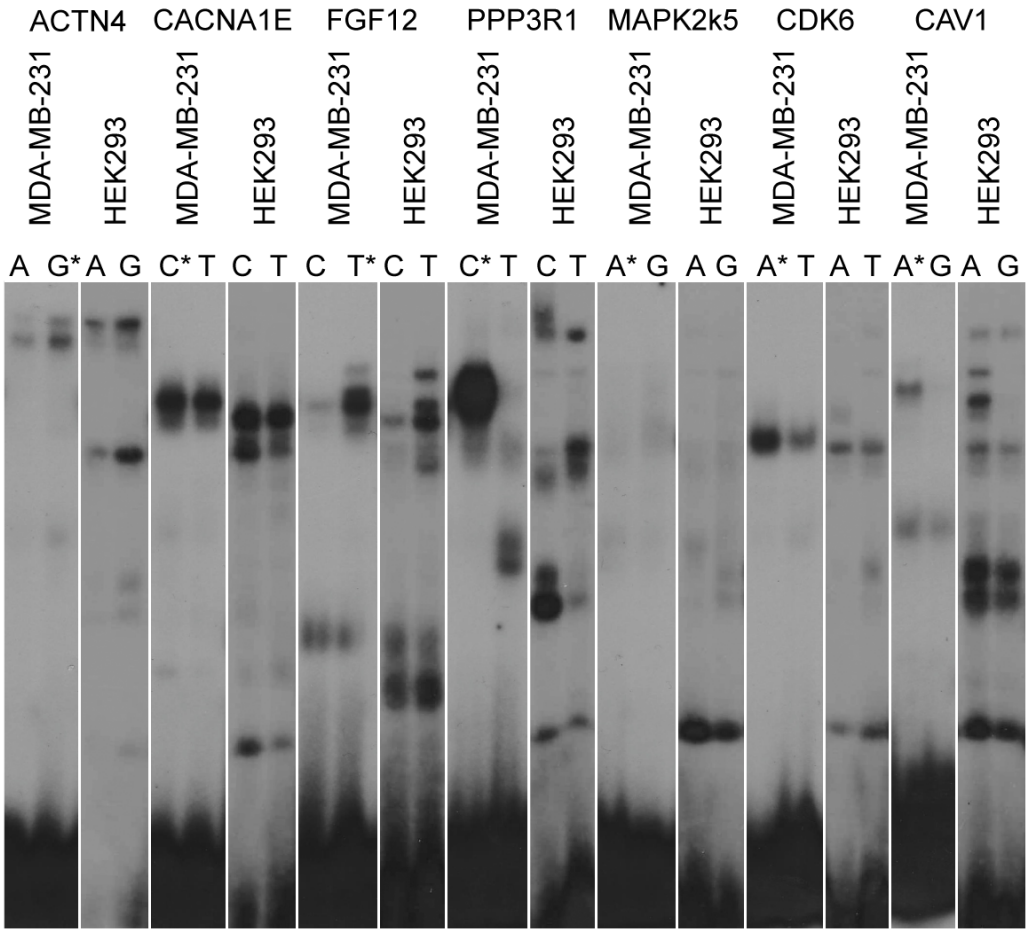


Figure 6.2: Electromobility shift assays (EMSAs) with the nuclear extracts from cell lines MDA-MB-231 (breast cancer cell line) and HEK293 (human embryonic kidney cell line). For all the SNPs, the allele with the higher affinity corresponds to the allele with the predicted binding. The $*$ symbol denotes the allele with predicted binding.

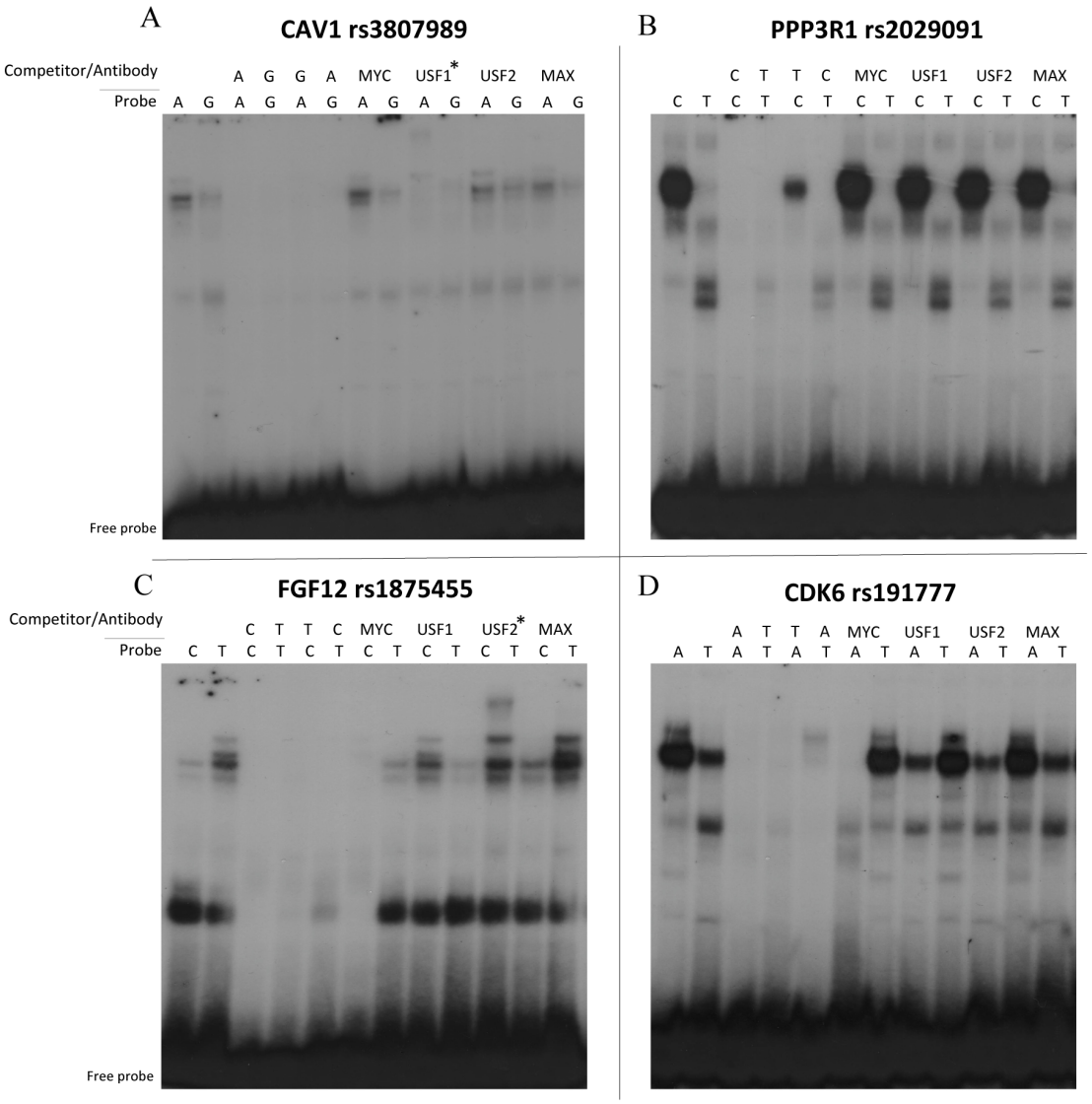


Figure 6.3: Electromobility shift assays (EMSAs) with the nuclear extracts from cell line MDA-MB-231 and antibodies against bHLH transcription factors Myc, Max, USF1 and USF2. Two of the allele-specific bands for SNP rs3807989 in CAV1 and SNP rs1875455 in FGF12 are successfully super-shifted with antibodies against USF1 and USF2, respectively. The $*$ symbol denotes the protein for which bands super-shifted.

6.2.3 E-box SNPs in a melanoma patient cohort

The four predicted E-box SNPs with the strongest allelic differences in protein binding (Figure 6.2) were taken forward to be tested for potential clinical associations. All four genes in which the E-box SNPs reside are part of signalling pathways that are crucial for melanoma risk and progression. Specifically, PPP3R1 is a regulatory subunit of calcineurin which, when inhibited, has been shown to have apoptotic effects in human melanoma cell lines HT168 and WM35 (Juhász *et al.*, 2009). CAV1 has been found to affect melanoma tumour growth and metastasis in mice (Trimmer *et al.*, 2010; Capozza *et al.*, 2012). FGF12 is part of the FGF pathway, which signals to the MAPK signalling pathway, a pathway that is essential for melanoma development and progression (Katoh and Katoh, 2006; Meier *et al.*, 2007). Finally, CDK4 and CDK6 have been associated with inherited predisposition to melanoma (Lin *et al.*, 2008).

The four candidate SNPs were consequently genotyped in a melanoma cohort of 105 patients to explore the potential associations of these loci with altered melanoma progression (for a cohort description see Materials and methods and Chapter 5). The p-values of the associations of the SNPs with overall survival, progression-free survival and time to metastasis are listed in Table 6.4. The log-rank test and the Cox proportional hazards model (adjusted for `AJCC stage` and `anatomic location`, for progression and metastasis-free analyses, and also including `age at diagnosis` for overall survival) were used to test for the associations. Three of the SNPs showed no sign of association, but SNP rs2029091 of PPP3R1 was strongly associated with all three outcomes with the Cox model, and showed

SNP (Gene)	Overall survival		Progression free survival		Metastasis free survival	
	Log-rank	Cox	Log-rank	Cox	Log-rank	Cox
rs2029091 (PPP3R1)	0.038	0.00059	0.078	0.016	0.30	0.0044
rs3807989 (CAV1)	0.69	0.96	0.94	0.82	0.99	0.26
rs1875455 (FGF12)	0.66	0.66	0.55	0.89	0.90	0.49
rs191777 (CDK6)	0.32	0.16	0.22	0.20	0.97	0.87

Table 6.4: p-values of the four SNPs that were genotyped in a melanoma cohort. SNP rs2029091 associated with all three progression phenotypes

clear signs of association with the log-rank test.

Interestingly, the Cox model with SNP rs2029091 showed much stronger signs of association for all three outcomes. Diagnostic tests for proportionality were performed for these three models, and signs of non-proportionality were found for the models for overall survival and time to metastasis for the covariate `stage`. However, when performing a stratified model for `stage`, the SNP was still very significant, with p-values of 0.0017 for overall survival and 0.0069 for time to metastasis.

Therefore, the smaller p-values compared to the log-rank test could be due to a better fitting model when the remaining prognostic factors were included, highlighting the evidence of association with the SNP. An additional reason would be that the log-rank test assigns no weights to the events, therefore underestimating early effects.

The allelic differences in the three progression-related phenotypes are shown in Figure 6.4. As expected from the nature of the outcomes, a strong correlation is noticeable between them. In more detail, patients with a homozygous T genotype have a longer survival and progression-free survival than C allele carriers. The CC

patients are only clearly distinguishable from the CT group for progression-free survival, but this could be due to the higher number of events for progression-free survival compared to the other two phenotypes (13 events compared to 6 events for metastasis and 8 for overall survival).

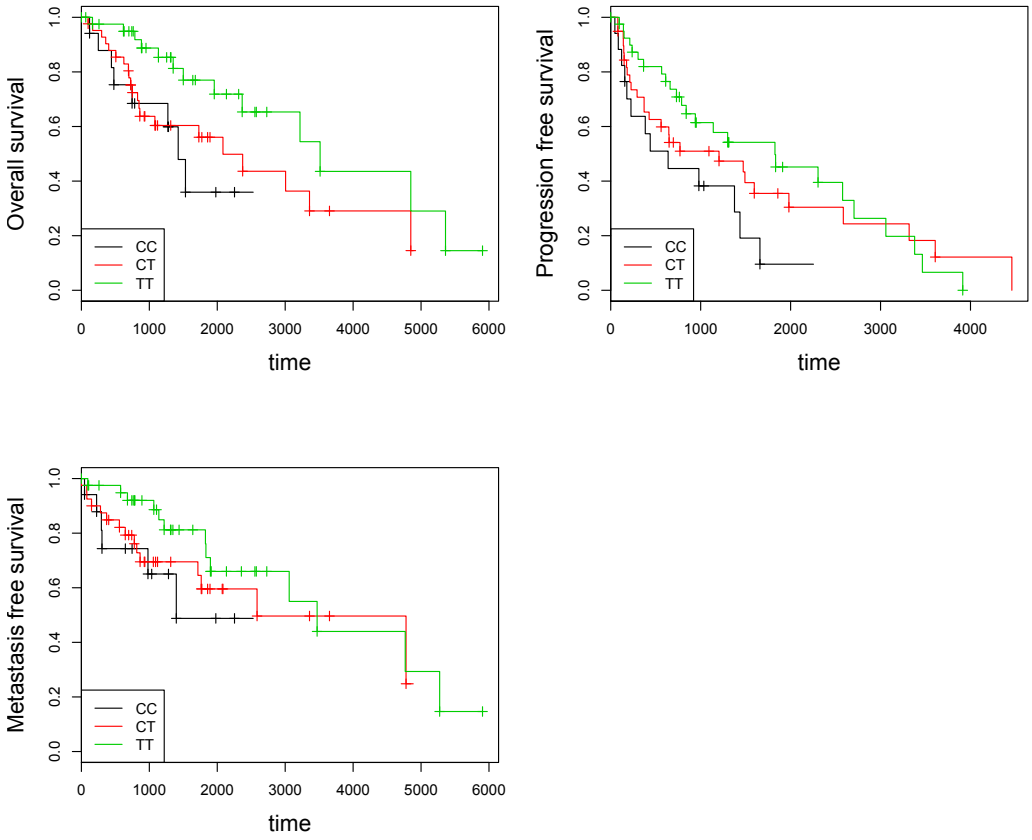


Figure 6.4: Kaplan-Meier survival curves for SNP rs2029091 for overall survival, progression-free survival and time to metastasis. For all three phenotypes the patients with the TT genotype have a much better prognosis compared to the individuals with either CT or CC.

The observed per C allele difference in hazard ratio was 2.25-fold (95% CI: 1.35, 3.74) for overall survival (for the stratified model), 1.57 (95% CI: 1.09, 2.27) for progression-free survival and 2.19 (95% CI: 1.24, 3.87) for metastasis-free survival (again for the stratified model). These data suggest that the identified polymorphic E-box in PPP3R1 could affect melanoma progression and survival.

In order to adjust for multiple hypothesis correction taking into account the correlation of the phenotypes, 1000 permutations of the data were performed, to compare the above p-values to their empirical distribution. This was achieved by permuting the genotypes of the data by resampling without replacement from the indexes of the patients, while keeping the correlation structure of the phenotypes and performing the same tests as above. Then the minimum of each simulation was retrieved and the min$p$-values were compared to the observed p-values. After correction for multiple hypothesis testing no p-value remained significant, with the adjusted p-value for overall survival and SNP rs2029091 being 0.064. Therefore, an exploration of this potential association in a larger cohort is necessary to draw any definitive conclusions of its association with melanoma progression and overall survival.

6.2.4 E-box SNPs as expression Quantitative Trait Loci (eQTLs)

SNP associations in GENEVAR

In order to test our hypothesis that the candidate SNPs would alter the E-box binding sites and affect transcription levels, the SNPs were explored for eQTL associations in the data derived from three studies of the GENEVAR (GENe Expression VARiation) project (Yang *et al.*, 2010). Only one SNP (rs2029091) was genotyped in all three studies of HapMap3, MuTHER and GeneCord (Stranger *et al.*, 2012; Nica *et al.*, 2010; Dimas *et al.*, 2009). Two of the SNPs (rs1875455 and rs3807989) were genotyped only in the studies of Stranger *et al.* (2012) (on

SNP	GENE	MuTHER Study	HapMap3	GeneCord
rs3807989	CAV1	Significant for Adipose tissue, Skin	Significant for CEU, CHB, GIH, JPT, LWK, MKK, YRI	NA
rs2029091	PPP3R1	Significant for LCL	Significant for CEU, JPT, MEX	Significant for T-cells
rs1875455	FGF12	NS	NS	NA
rs191777	CDK6	NS	NS	NS

Table 6.5: Table of eQTL results for the 4 top SNPs. Two SNPs associated with mRNA expression for the MuTHER and HapMap3 studies. Abbreviations: NS - non-significant (for either the main SNP or the proxies), NA - not available (for neither the main SNP nor the proxies), LCL - lymphoblastoid cell lines. Population abbreviations can be found in the Abbreviations section.

HapMap3 data) and Nica *et al.* (2010) (MuTHER study), but had no available proxies for the GeneCord study (Dimas *et al.*, 2009). For rs191777 a proxy (from the 1000 Genomes project) was used for the associations, the studies of Dimas *et al.* (2009) and Nica *et al.* (2010), and in the HapMap3 populations 2 linked SNPs in CEU, YRI and CHB/JPT were found and were both tested. Only when a SNP was significant in both twins of the MuTHER study was it recorded as significant. The results of this analysis are summarised in Table 6.5, where significant associations were noted for both SNPs rs2029091 and rs3807989, at the 0.05 p-value threshold for the permutations analysis.

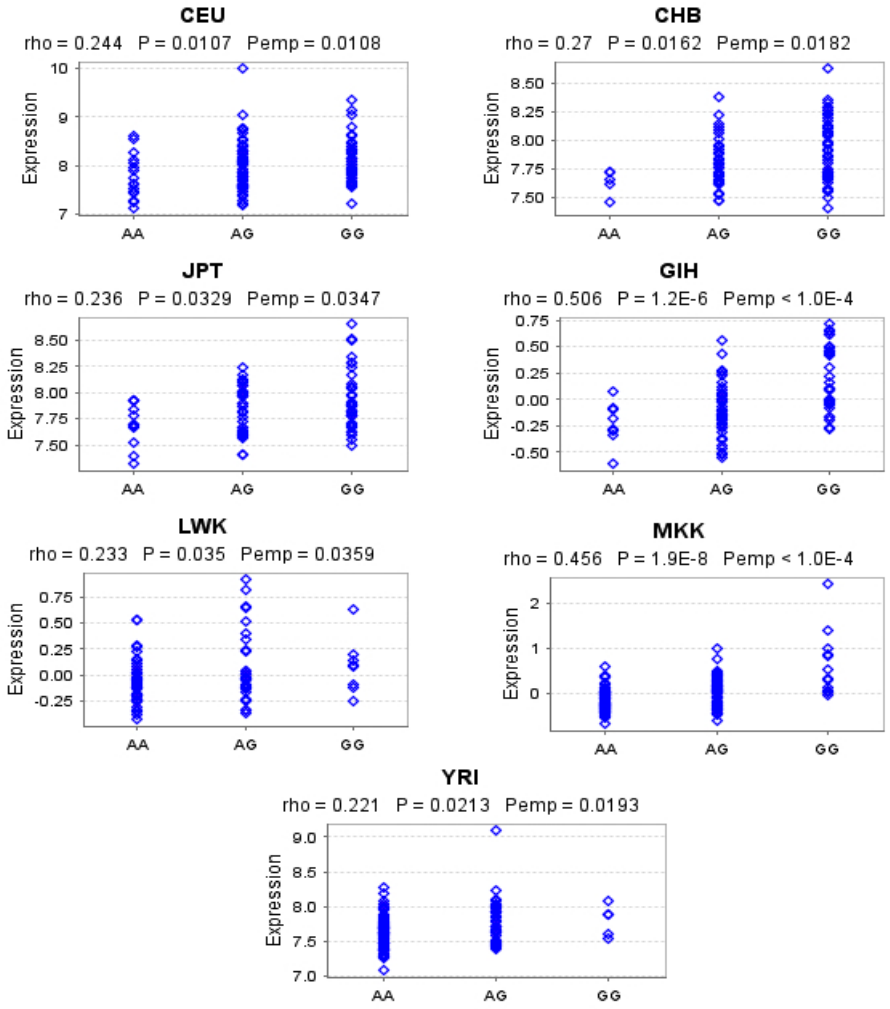
More specifically, SNP rs3807989 had a significant correlation with mRNA expression of CAV1 in lymphocytes of 7 out of 8 HapMap populations (Figure 6.5a). In addition, associations were also observed in the adipose tissue and skin of both twins in the MuTHER study (Figures 6.5b). The A allele, which was the allele predicted to bind to an E-box element, was therefore associated with lower expression levels for CAV1, suggesting that the protein binding to the DNA

sequence would be acting as a repressor. SNP rs3807989 unfortunately had no proxies genotyped in the GeneCord study so no conclusions could be drawn for this dataset.

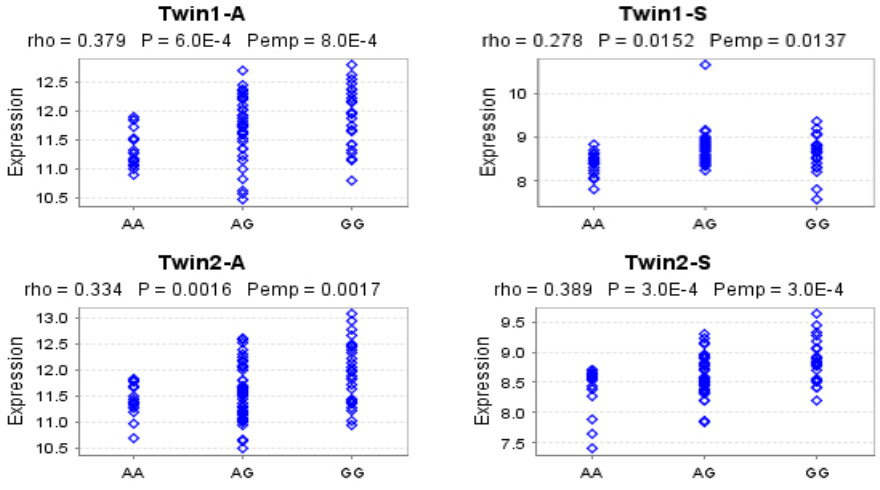
SNP rs2029091 also associated with differential mRNA expression levels for PPP3R1 in the CEU, JPT and MEX HapMap populations (Figure 6.6a) and in lymphoblastoid cell lines of the MuTHER study (Figure 6.6). Finally, the proxy of SNP rs2029091, rs7560138 with an $r^2 = 1$ in CEU, also showed an association with PPP3R1 mRNA expression for T cells of the study of the Geneva GeneCord individuals. Interestingly, the C allele was associated with lower mRNA levels in the HapMap populations and the T cells from GeneCord, but with higher expression in the lymphoblastoid cell lines of the MuTHER study. This paradox could be explained by the varying roles of bHLH transcription factors as activators and repressors in different scenarios, as shown in Massari and Murre (2000). Together, potential associations of 4 SNPs were explored in the 3 eQTL studies, and it was noted that 2 SNPs did indeed associate with differential levels of gene expression. These data, therefore, lend support to the hypothesis that the expression of genes CAV1 and PPP3R1 could be altered by differential binding to the E-box SNPs rs3807989 and rs2029091, respectively.

SNP rs2029091 in a melanoma cell line panel

The rs2029091 SNP in the PPP3R1 gene showed significant differences in PPP3R1 levels in two different cell types in the GENEVAR project and although the association with progression and survival of melanoma was not significant after



(a)



(b)

Figure 6.5: Allele-specific differences in mRNA expression levels of CAV1 were observed for SNP rs3807989. The A allele of rs3807989 associated with lower mRNA expression in (a) 7 HapMap populations and (b) adipose tissue and skin in twins. Note: No comparison can be made on the expression levels between populations because the measurements come from different experiments and as such they are not comparable (they are sensitive to systematic differences between arrays, such as batch effects).

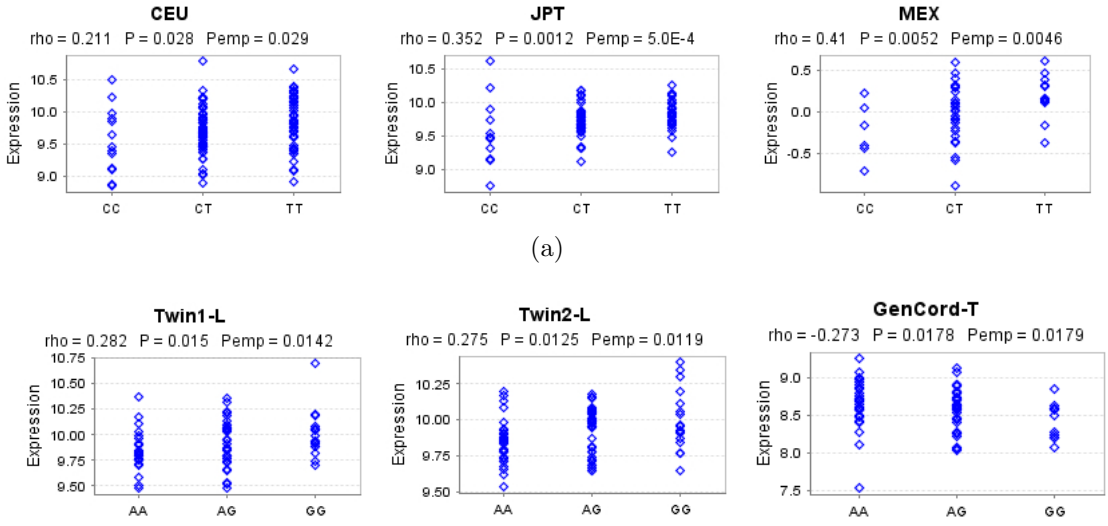


Figure 6.6: SNP rs2029091 and its proxy rs7560138 associated with allelic differences in mRNA expression for the gene PPP3R1. The C allele of SNP rs2029091 associated with lower mRNA expression levels in LCL of 3 HapMap populations (a), but with higher levels in lymphoblastoid cell lines from the MuTHER study (b). The G allele of SNP rs7560138 lies in the same haplotype as the C allele of SNP rs2029091 and also associated with lower mRNA levels (c).

multiple hypothesis correction, the trend suggested a possible association. Therefore, the association of rs2029091 with mRNA expression levels was explored in a melanoma panel of 39 cell lines. The mRNA transcript levels of PPP3R1 were measured in the panel, and the SNP genotyped in all cell lines by Jorge Zeron. The C allele was associated with higher mRNA expression levels (which corresponds to lower ΔCt values) in the melanoma cell lines (Figure 6.7) with a p-value of 0.032 (Jonckheere test). Cell line Me248.3 appeared to be an outlier with an extremely low ΔCt value of 0.037. The values of the rest of the cell lines ranged between 4.37 and 9.49, with a standard deviation of 1.05. A non-parametric test was used to avoid having to exclude the cell line. Therefore, these data suggest that this E-box SNP is associated with the mRNA expression of PPP3R1 in melanoma, which could affect melanoma progression and survival.

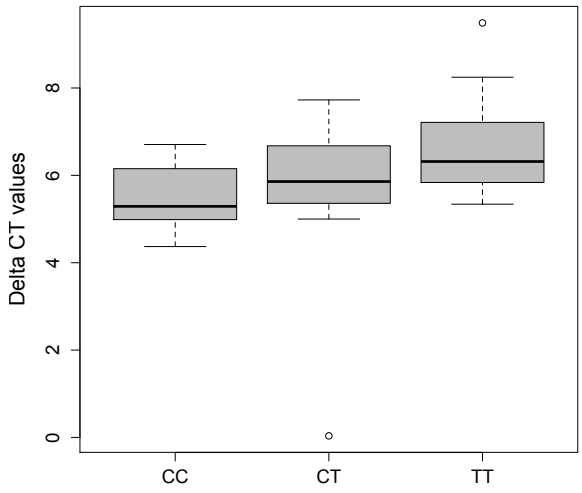


Figure 6.7: SNP rs2029091 exhibits deferential allele mRNA expression for the PPP3R1 gene. The C allele associates with higher expression in a panel of 39 melanoma cell lines.

6.3 Discussion

In this chapter, more than 1,100 cancer genes were screened for all common SNPs, to identify the SNPs that lie in potential functional E-box binding sites. From 104 candidate SNPs, which resided in regions with the E-box binding motif and which were evolutionarily conserved, 7 were selected based on supporting evidence regarding their regulatory potential from publicly available data (from the ENCODE project): firstly, based on whether transcription factors of the HLH protein family have been shown to bind to the relevant regions; and secondly, whether the SNPs associated with a GWAS phenotype and lay in regions of regulatory potential. Of the 7 selected SNPs, 4 showed strong allelic binding differences in protein-DNA assays (EMSAs), and the allele-specific bands of 2 SNPs were super-shifted with bHLH proteins USF1 and USF2. Moreover, SNPs rs2029091 and rs3807989, of the PPP3R1 and CAV1 genes, also functioned as eQTLs in GENEVAR. Finally, SNP

rs2029091 in the PPP3R1 gene showed significant allelic differences with mRNA expression of PPP3R1 in a panel of melanoma cell lines, although its association with overall and metastasis-free survival in a melanoma cohort of patient did not remain significant after multiple hypothesis correction. This suggests that the methodology used here for the identification of SNPs in E-boxes of known cancer genes could have phenotypic effects in cancer patient cohorts.

SNP rs2029091 is located in intron 1 of Protein Phosphatase 3, Regulatory Subunit B, alpha (PPP3R1). PPP3R1 is also known as calcineurin B type 1 gene, since it is a subunit of calcineurin (CN). Calcineurin is a calcium-dependent phosphatase that activates the T-cells of the immune system. Interestingly, in the eQTL studies, the SNP showed evidence of association in T-cells (GenCord) and lymphoblastoid cell lines (MuTHER study and HapMap populations), but not in fibroblasts, skin or adipose tissue (MuTHER and GenCord studies). This supports the hypothesis that the SNP affects the transcription of PPP3R1 in cell types that are involved in the immune system.

Calcineurin is known to regulate the NFAT (nuclear factor of activated T-cells) proteins by dephosphorylating them, after which they are translocated to the nucleus and become transcriptionally active (Crabtree and Olson, 2002; Hogan *et al.*, 2003). Members of the NFAT family have been shown to control tumour cell proliferation and regulate apoptosis. In addition, overexpression of NFAT proteins has been noted in tumour progression and metastasis but, to date, no mutations have been found in human cancers (Müller and Rao, 2010). Moreover, the disruption of NFAT signalling has been observed in many types of cancer,

such as colon cancer, haematological malignancies, breast cancer and pancreatic adenocarcinomas (Müller and Rao, 2010). For example, it has been suggested that NFATc2 has an essential role in the control of inflammation-derived colorectal cancer (Gerlach *et al.*, 2012). Interestingly, in this study, SNP rs2029091 associated with mRNA expression in a melanoma cell line panel. Calcineurin and NFAT signalling has been shown to have an antiapoptotic role and activate tumour growth in melanoma (Fedida-Metula *et al.*, 2012; Perotti *et al.*, 2012). Moreover, disrupting the calcineurin cascade was shown to extend survival *in vivo* (Fedida-Metula *et al.*, 2012). Even though the association with melanoma progression and overall survival did not withstand multiple hypothesis correction and needs further exploration, a model could be hypothesised in which an E-box protein is bound differentially to the two alleles and, when bound to the C allele, increases the protein levels of PPP3R1 promoting tumour progression and metastasis (Figure 6.8). Interestingly, the pharmacological implications of this system have also been explored. Hernández *et al.* (2001) have previously provided evidence that blocking NFAT translocation with cyclosporin A (CsA) prevents tumour angiogenesis in mice. Importantly, this agent is already used in the clinic in transplant therapy (Crabtree and Olson, 2002) and, thus, rs2029091 could aid as a biomarker to target patients likely to respond to treatment.

The results described in Section 6.2.4, in which the C allele of rs2029091 which forms the E-box is associated with higher mRNA levels in some cell lines but with lower mRNA levels in others, can appear contradictory at first. However, the bHLH transcription factors have been previously shown to act as both transcriptional

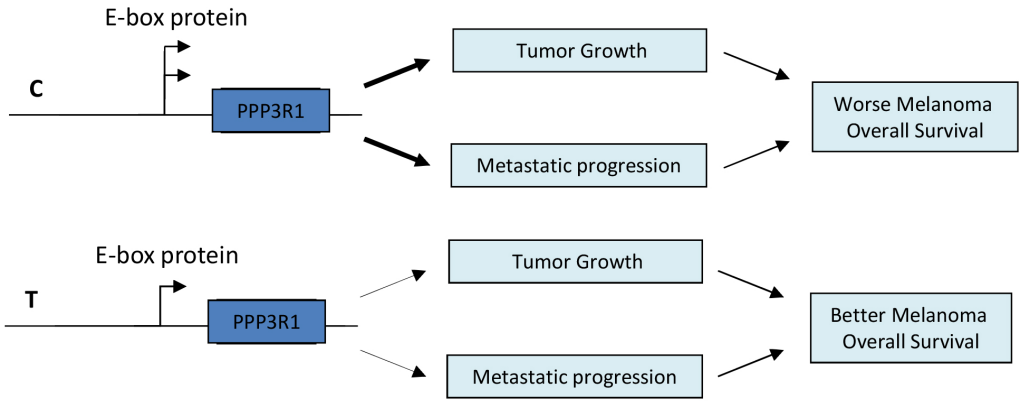


Figure 6.8: SNP rs2029091 is shown to reside in an E-box binding motif, affecting the transcription of the PPP3R1 gene, the overexression of which could affect the tumour growth and metastatic progression of melanoma patients.

activators and repressors (Massari and Murre, 2000). Therefore, these associations further support this study's hypothesis that the SNP acts as a bHLH transcription factor and that it is located in a functional region, independently of the direction of the effect. Nonetheless, an extended eQTL analysis of more tissue types would be needed to elucidate the function of this SNP and further identify the types of cancer that this SNP could be a biomarker for.

Evidence is provided here of SNP rs3807989 associating with mRNA expression levels of CAV1, but no clinical associations were found. SNP rs3807989 is located in the last intron of caveolin 1 (CAV1), a tumour suppressor and a negative regulator of the RAS signalling cascade. It has been shown that downregulation of CAV1 can play an important role in cancer metastasis (Lu *et al.*, 2003). However, further research regarding the potential clinical associations of rs3807989 is essential for a better understanding of the role of this SNP in cancer.

In this study, two SNPs with potential implications in cancer research have been identified. Even more importantly, however, the promising results of this

chapter suggest that the novel methodology by which SNPs in E-box binding motifs have been identified could be used to screen for many more interesting transcription factors in genes involved in diseases of interest, an idea which is pursued in the following chapter. In addition, the importance of incorporating publicly available data into genomic studies has been demonstrated. Given the plethora of data becoming available every day, these studies have the potential to yield ground-breaking discoveries in the field of biological research.

7 A polymorphic p53 response element in the KITLG gene influences cancer risk and has undergone natural selection

7.1 Introduction

Following the promising results outlined in the previous chapter, the search for SNPs in E-box binding sites was extended to additional transcription factors binding sites (TFBS) in which a single base pair change could dramatically alter binding affinities. The aim of this analysis was to further characterise SNPs identified in cancer-related GWAS and provide insight into the mechanisms underlying the observed associations. To date, there have been over 70 cancer GWAS, with more than 250 novel loci associated with an increased propensity to develop cancer (Hindorff *et al.*, 2009). However, determining the causal SNP from the haplotype of associated SNPs has been challenging, and only very few of these associations have been followed through systematically. In a GWAS context, this can be achieved by a number of methods, like targeted sequencing, eQTL analyses, and DNA methylation profiling (Freedman *et al.*, 2011).

Performing a pattern search for SNPs that alter TFBS is an attractive additional method for attributing functionality to trait-associated SNPs. SNPs associated with cancer risk have, in fact, been previously shown to reside in TFBS and impair or activate transcriptional control (Bond *et al.*, 2004; Post *et al.*, 2010; Schödel *et al.*, 2012). In addition, SNPs identified in GWAS have been shown to be

enriched in non-coding functional DNA elements (Ernst *et al.*, 2011), especially in transcription factor binding regions (ENCODE Project Consortium *et al.*, 2012).

An important feature of this method is that it proposes hypotheses that can be supported by publicly available data. This data includes experiments that have been performed by the Encyclopedia of DNA Elements (ENCODE) project and other international collaborative projects, such as genome-wide scans for TFBS and regions with differential methylation status, experiments on regulatory elements of genes, etc. In addition, the filters applied are highly customisable to highlight the regions of interest.

The binding motifs used for the identification of sequences that contain SNPs that could affect transcription share a common trait, in which single base substitutions can dramatically affect transcription factor binding. This has been shown before to be an important feature. One such example is SNP 1867227 of the gene FOXE1, which has been found to alter the binding affinities of the transcription factors USF1 / USF2, resulting in the allele-dependent regulation of the transcription of FOXE1, and has also been associated with thyroid cancer susceptibility (Landa *et al.*, 2009).

In this chapter, a screen is presented through which a SNP residing in a TFBS where p53 was shown to bind and affect transcription of the KITLG gene, was identified. Evidence is also presented that this SNP, which has been previously associated with an increased risk of developing testicular cancer in 3 GWAS, has been subjected to positive selection in Europeans. Aspects of the work presented in this chapter have been published in *Cell* as part of a collaborative project with

Jorge Zeron (Zeron-Medina *et al.*, 2013).

7.2 Results

7.2.1 Screen to identify SNPs residing in potential transcription factor binding sites (TFBS)

To identify the cancer-associated SNPs that could be residing in transcription factor binding sites (TFBS), all the SNPs that have been found by a GWAS for any cancer-associated trait were examined, with a focus on those with the potential to create or break a TFBS. Based on publicly available resources (details to be found in the Materials and methods), 271 SNPs associated with any type of cancer (using search terms listed in the Materials and methods) were identified in 74 publications. The 271 GWAS SNPs, as well as all common SNPs linked to them with an $r^2 > 0.8$, were examined, which gave a total of 5,053 SNPs. SNPs with high D' values (and low r^2 values) to the associated SNP were not considered in this analysis because the focus was on common SNPs and not rare SNPs which could provide a high association signal without having a high r^2 value with the associated SNP.

In addition, 9 transcription factors were selected based on the hypothesis that a base pair change could destroy the binding, i.e. the stringency of their binding motif, and on the availability of agents that our group would be able to test for. A pattern search was performed for the 5,053 SNPs, filtering for ones for which one of the two alleles was part of a sequence containing the binding motifs of any of the 9 transcription factors (Table 7.1) and the other allele was abolishing the binding, and 537 SNPs were identified in potential TFBS.

Transcription Factor	Binding motif	Sequences identified
MEF2C	TAWWWWTA	62
Klf4	GGGHGKGG	32
FOX:ETS	AACAGGAW	10
Rbp-j	GTGRGAA	30
Runx1	TGTGGT	49
E box	CANNTG	274
USF	CACGTG	15
Hif1a	RCGTG	161
Mbox	CATGTG	53

Table 7.1: A search was conducted for SNPs lying in regions of potential transcription factor binding sites. The binding motif is shown for each transcription factor, with W=A or T, H=any base but G, K=G or T, R=purine, and N=any base.

The SNPs were further filtered according to frequency and GWAS characteristics (Table 7.2). In order to also be able to characterise the candidate SNPs with a functional approach, we needed to ensure that cell lines both heterozygous and homozygous for the minor allele would be available. Therefore, only SNPs with a MAF higher than 10% in the CEU were selected (368 SNPs). In addition, only SNPs identified in populations of European ancestry were considered, since most of the cell lines that were to be utilised for the functional analysis were of European ancestry (298 SNPs). Furthermore, in order to focus on SNPs with a strong signal of association, 60 SNPs that did not replicate in an independent cohort were excluded, along with 72 SNPs for which the p-values had not reached the GWAS threshold (p-value$> 10^{-8}$). Finally, 14 SNPs were excluded because both the reported gene of the trait-associated SNP and the proxy SNP were intergenic, and as such their functional characterisation would have been challenging, giving a total of 171 SNPs fulfilling our criteria.

SNPs associated with cancer susceptibility in a GWAS	271
All SNPs and their proxies	5053
Filtering	
SNPs in sequences of TF binding motifs	537
SNPs with *MAF* (in CEU) > 10%	368
Exclusions	
SNPs from studies in Chinese, Japanese or African populations	51
SNPs from studies that didn't include a replication cohort	60
SNPs with high *p*-values in GWAS (*p*-value > 10^{-8})	72
Cases when both proxy SNP and trait-associated SNP were intergenic	14
Remaining SNPs	171

Table 7.2: Criteria for the selection procedure for the identification of SNPs on potential binding sites of transcription factors.

7.2.2 Bioinformatics and functional analysis of candidate SNPs

Functional SNPs have been shown to be enriched in regulatory regions (Ernst *et al.*, 2011). For this reason, all 171 SNPs underwent a bioinformatics analysis by another member of the lab, Jorge Zeron, and were ranked according to further criteria, suggesting that the selected SNPs had strong signatures of regulatory potential. Groups of SNPs belonging to the same haplotype block that were all altering TFBS were favoured, and so were SNPs in binding sites that were linked to more than one trait-associated SNP. Nine SNPs in 3 haplotype blocks resided in regions of transcriptional enhancer potential (Table 7.3), and were followed up by Jorge Zeron via experimental analyses for differential binding with electrophoretic mobility shift assays (EMSAs).

Four SNPs showed signs of differential binding; one in the FGFR2 gene and three in the KITLG gene, as listed in Table 7.4. Figure 7.1, supports this prediction of SNPs that alter binding with a single base pair change. The assay was performed using nuclear extracts from the breast cancer cell line MDA MB 231 for the SNPs

SNP ID	Gene (Location)	bHLH ChIP-seq	Promoter ENC.	Enhancer ENC.	DNase I ENC.	ESPERR
rs35444713	CDH1	HEY1	0.8	0.7	0	0
rs4076177	(Intron 2)	USF1	0.6	0.2	0	0
rs9928796		HEY1, USF1	0.6	0	1	1
rs1078806	FGFR2	No	1	0	1	1
rs4752570	(Intron 1/2)	No	1	0	1	0
rs1219648		No	0.1	0	0	1
rs2046971	KITLG	No	0.75	0	1	1
rs3907470	(Intron 1)	No	0.75	0	0	0
rs4590952		No	0.75	0	1	0

Table 7.3: Summary of the bioinformatics selection criteria for the screening of the 171 SNPs. Three SNPs resided in regions where bHLH transcription factors had been found to bind, and all resided in regions of regulatory potential. These were defined as regions with signs of enrichment of the H3K4Me3 histone marker (Promoter ENC.), the H3K4Me1, H3K27Ac histone markers (Enhancer ENC.), the DNase I hypersensitivity marker (DNaseI ENC.) and with a high score of regulatory potential (ESPERR).

in FGFR2, and the testicular cell line TERA 2 for the SNPs in KITLG, since they have been associated with breast and testicular cancer, respectively (Rapley *et al.*, 2009; Turnbull *et al.*, 2010; Kanetsky *et al.*, 2009).

The four SNPs were then followed up with experiments using luciferase reporter vectors. SNP rs450952 of the KITLG gene, and SNP rs4752570 of FGFR2, demonstrated clear allele-specific differences in overexpression of luciferase activity (Figure 7.2). SNPs rs2046971 and rs3907470 also showed decreased activity, suggesting that they could be acting as repressors.

SNP rs4590952 is in high LD with three other SNPs, rs4474514, rs3782191 and rs995030 (with r^2 of 1, 1 and 0.89, respectively), which have been found to be associated with testicular cancer risk in three GWAS studies, two of which were performed in independent samples (Thomas *et al.*, 2009; Easton *et al.*, 2007; Hunter *et al.*, 2007; Rapley *et al.*, 2009; Turnbull *et al.*, 2010; Kanetsky *et al.*,

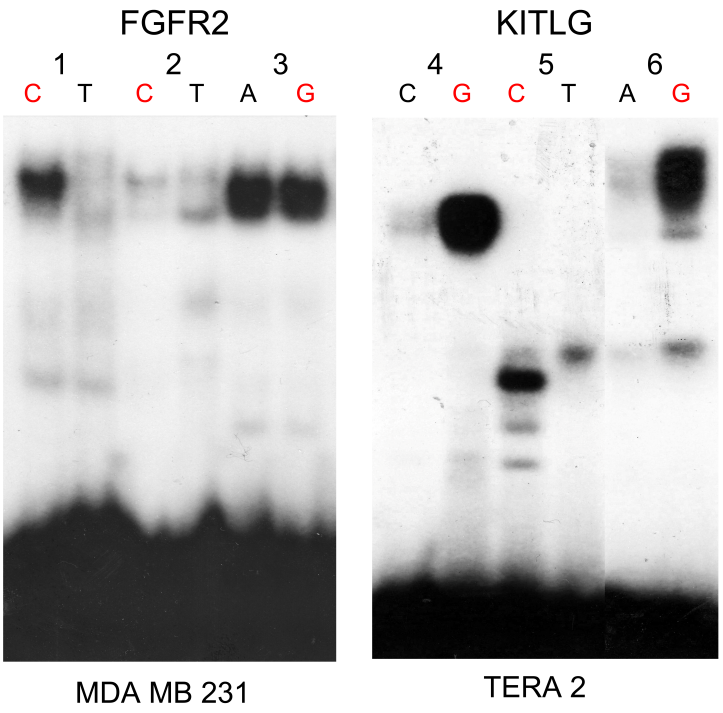


Figure 7.1: Electrophoretic mobility shift assays (EMSAs) using nuclear extracts from the breast cancer cell line MDA MB 231 for rs1078806 (1), rs4752570 (2) and rs1219648 (3), and the testicular cell line TERA 2 for rs2046971 (4), rs3907470 (5) and rs4590952 (6). The EMSAs show signs of differential binding in-line with the predicted binding (red alleles).

2009). Interestingly, the G allele was linked to the risk allele of the three studies, which was reported to have an odds ratio (OR) of up to 3.07, which is one of the highest ORs noted to date in a GWAS (Chanock, 2009). Consequently, this SNP was further followed-up and tested via *in silica* prediction of TFBS (Figure 7.3a). All the predictions that were differential between the two alleles were tested with EMSAs, by Jorge Zeron. Interestingly, only the p53 TF was found to bind in an allele-specific manner in the region, with the band of the G allele being super-shifted using a p53 antibody (Figure 7.3b).

Many transcription factors have similar binding motifs, such as hif1a and USF1 (Table 7.1). Likewise, the hif1a binding motif (RCGTG, with R=purine)

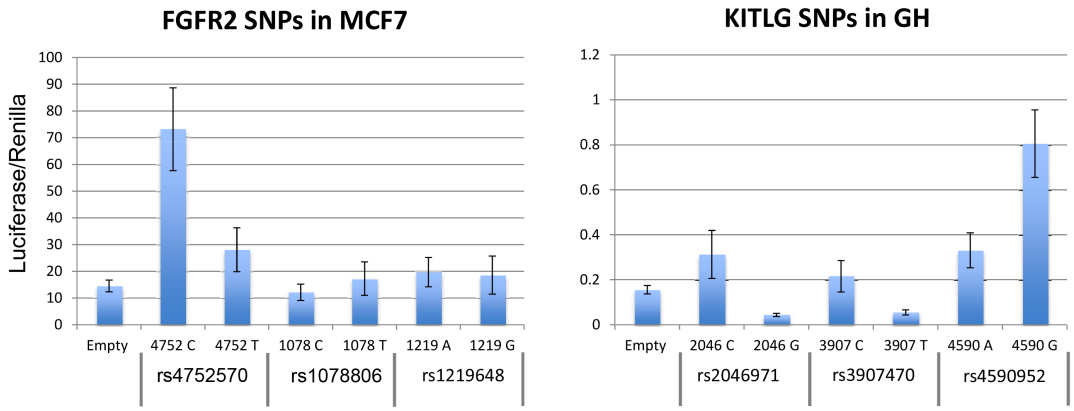
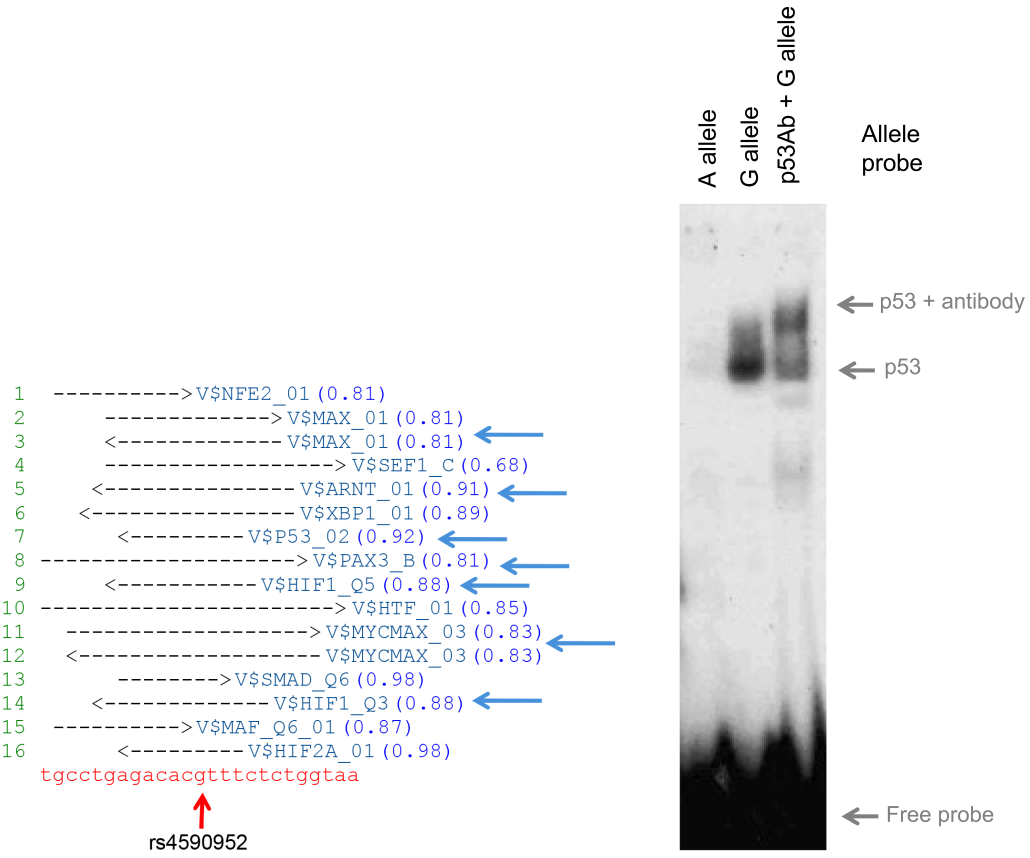


Figure 7.2: Experiments in luciferase reporter vectors showed allelic differences in luciferase activity for SNPs rs4752570 and rs4590952, in line with the differential binding.



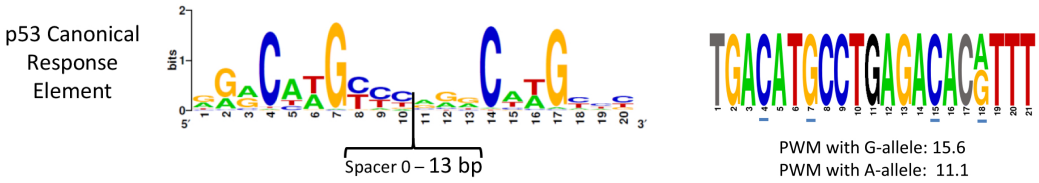
(a) (b)

Figure 7.3: (a) The transcription factors (TF) predicted to bind to the sequence when the G allele is present. The arrows point to the TF that were differential between the two alelles. p53 was one of the TF predicted to bind and the only one that was super-shifted with a p53 antibody (DO1).

TFBS	Proxy SNP	GWAS SNP	r^2	Gene	GWAS trait
E-box	rs1078806	rs2981579 rs1219648 rs2981582	1 0.967 0.935	intron 1, FGFR2	breast cancer
E-box E-box hif1a	rs2046971 rs3907470 rs4590952	rs3782181 rs4474514 rs995030	1 1 0.861	intron 1, KITLG	testicular cancer

Table 7.4: SNP rs1078806 in intron 1 of FGFR2 is part of an E-box binding motif and is linked to 3 SNPs (rs2981579, rs2981582 and rs1219648, with a r^2 of 1, 0.935 and 0.967, respectively) associated with breast cancer susceptibility in 3 independent GWAS (Thomas *et al.*, 2009; Easton *et al.*, 2007; Hunter *et al.*, 2007). Three SNPs in intron 1 of KITLG showed differential binding; SNPs rs2046971 and rs3907470 were located in 2 E-box binding sites, and rs4590952 in a hif1a binding site. All of them were linked to 3 SNPs (rs3782181, rs4474514 and rs995030, with a r^2 of 1, 1 and 0.861, respectively) that are associated with testicular cancer (Rapley *et al.*, 2009; Turnbull *et al.*, 2010; Kanetsky *et al.*, 2009) .

has high similarity to the half-site of the p53 consensus motif (RRRCWWGYYY, with R=purine, Y=pyrimidine and W=A or T), and 4 of the 5 hif1a binding motif base pairs overlap with it. The p53 consensus motif, which will be referred to here as the p53-response element (p53-RE), is created by two decameric half sites with a spacer of 0-13 nucleotides between them (Figure 7.4a). The PWM scores were calculated based on 228 published p53-REs, and the scores of the two alleles compared to the average of 13.8 (Zeron-Medina *et al.*, 2013). SNP rs4590952 resides in a key nucleotide of a p53-RE with a position weight matrix (PMW) score of 15.6 when the G allele is present, and 11.1 when the A allele is present (Figure 7.4b), supporting the above evidence that the G allele could differentially bind to the p53 transcription factor.



(a) (b)

Figure 7.4: The p53 consensus binding motif consists of two decamers with a spacer of 0 to 13 nucleotides (a). The C and G bases in the 4th and 7th position respectively, of each decamer constitute the core nucleotides of the motif. SNP rs4590952 resides on the 7th base of the 2nd decamer of a p53-RE (b). The PWM score increases from 11.1 to 15.6, with the substitution of A with a G allele in the position of our SNP. The spacer nucleotide is shown in black and the core bases are underlined in blue.

7.2.3 SNP rs4590952 shows allele-specific differences in p53-dependent transactivation of KITLG

To further explore whether the p53-RE SNP rs4590952, which lays in the first intron of KITLG, could affect the transcriptional activation of KITLG in a p-53 dependent manner, Jorge Zeron undertook two experiments, which will be described briefly here.

In the first experiment, a 60-bp region around the SNP was cloned into a luciferase reporter vector, and then inserted into a wild type p53 cell line (HCT116 p53+/+) and its isogenic p53-null form (HCT116 p53-/-). In the HCT116 p53+/+ cells, a 7-fold increase in activity was measured in the reporter containing the G allele (the allele with the stronger predicted p53-RE) compared to the A allele (p-value=0.003, Mann-Whitney test), whereas no significant difference was observed in the HCT116 p53-/- cells.

In the second experiment, the reporter vectors were transfected into four p53 wild type cell lines (three of which were of testicular cancer origin), and three p53

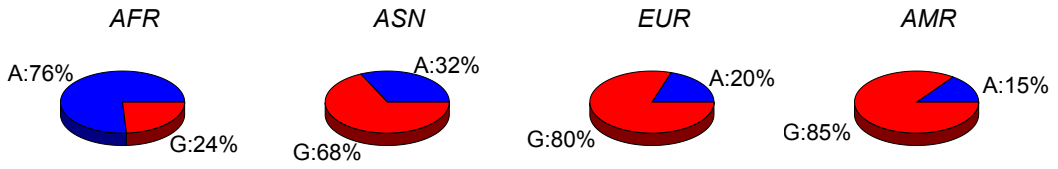


Figure 7.5: Frequencies of SNP rs4590952 for the 4 super populations of the 1000 Genomes project. Abbr: AFR=African, ASN = East Asian, EUR = European, AMS=Ad Mixed American

mutant or null cell lines. Similarly to the first experiment, no significant difference in luciferase activity was measured between the two alleles in the p53 mutant or null cell lines, but a 188-fold difference (on average, ranging from 93 to 373-fold) was measured for the wild type cell lines. A more detailed description and figures for these experiments can be found in Zeron-Medina *et al.* (2013).

7.2.4 The KITLG p53-RE SNP displays signatures of natural selection

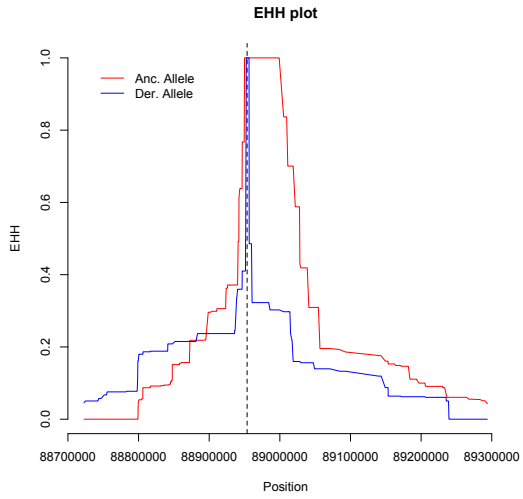
The role of p53 has been extensively studied over the last three decades and has been demonstrated to be crucial in many cellular processes, one of the most important being the stress response, via which it induces DNA repair, cell cycle arrest, senescence or apoptosis. Because of its essential role, many functional SNPs in the p53 pathway have undergone natural selection (Atwal *et al.*, 2007, 2009; Shi *et al.*, 2009). In order to assess whether this was the case with SNP rs4590952, two tests were performed to compare the allele frequencies between populations and the haplotype structures between the two alleles within the populations.

The differences between the allele frequencies in the 4 super populations of the 1000 Genomes project can be seen in Figure 7.5. To test for genetic differentiation between the super populations the fixation index statistic (Fst) was used,

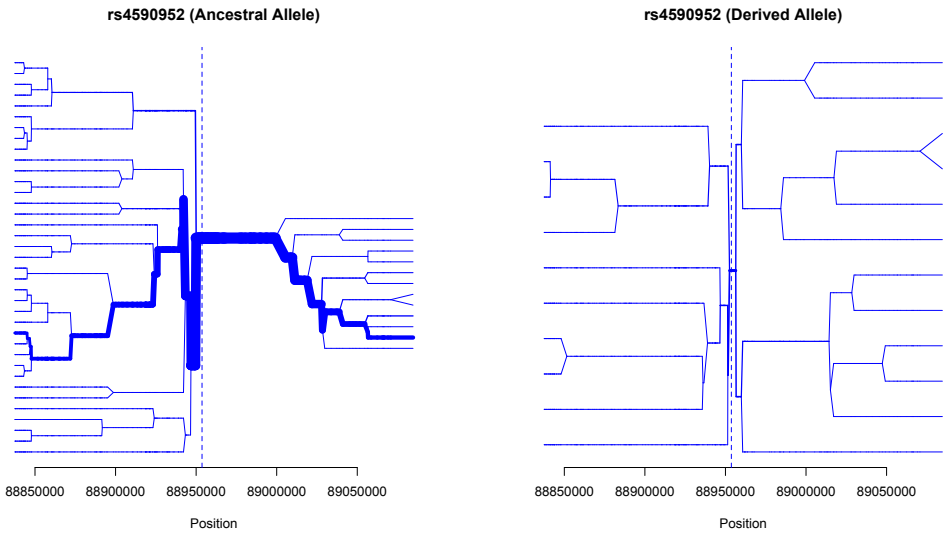
which compares the expected heterozygosities in the subpopulations to the overall population. The allelic differences noted were significantly larger than those that would have been expected by genetic drift alone, with an Fst of 0.228 between all super populations. In more detail, the frequency of the G allele (the allele with the stronger predicted p53-RE) in Europeans (which included CEU, FIN, GBR, IBS and TSI) was 80%, in contrast to the African populations (which included ASW, LWK and YRI) for which it was on average only 24%. These differences alone gave an even higher estimate for the fixation index of 0.301 between the EUR and AFR. This suggested that the estimated subdivision of populations accounted for approximately 30.1% of the total genetic variation between the EUR and AFR populations.

In addition, the haplotype homozygosities of the two alleles were compared in the CEU population, to explore whether natural selection could have driven the G allele to such a high frequency in the European populations. Theoretically, strong selection can increase the frequency of one allele unusually fast, leading to a long haplotype of low diversity. The haplotype structure between the two alleles showed an unusually long haplotype for the G allele (ancestral) compared to the A allele (derived) (Figure 7.6).

In order to formally assess the strength of the natural selection driving these changes in frequency, the integrated haplotype homozygosity score (iHS) was used (details in the Materials and methods). The iHS for rs4590952 was 2.3, one of the highest in the KITLG gene (Figure 7.7), and in the top 5% of all signals of positive selection in the genome (Voight *et al.*, 2006). The fact that the score was positive



(a)



(b)

Figure 7.6: The plot of the extended haplotype homozygosity (EHH) from the core SNP (rs4590952) shows the decay of the two haplotypes according to position, in CEU (a). The G allele (ancestral allele) appears to have a longer haplotype compared to the A allele (derived allele). The bifurcation diagram shows the extent of the haplotypes in distance, in CEU (b). The root of the diagram is the core SNP and a split occurs when more than one haplotype is found in the population for that region (between the position of interest and the core SNP). The thicker the line, the more frequent the haplotype. The thickness of the line for the G allele shows that there is one prominent haplotype with long range homozygosity. In contrast, the haplotypes of the A allele are more evenly distributed in frequency.

indicated longer haplotypes of the ancestral allele, in line with the EHH plot of Figure 7.6.

It was intriguing that the G allele was the ancestral allele (as defined by sequences found in the last common ancestor of humans and chimpanzees), whilst the major allele for the African populations was the A allele (the derived allele). It was, therefore, hypothesised that different selective pressures could be acting in the African populations, but no such evidence was found when the iHS was calculated for the YRI (iHS=1.2). This was also apparent from the LD analysis of the YRI, where the very long haplotype observed in CEU was non-existent in the YRI (Figure 7.7) and the bifurcation diagram of the SNP in YRI, which presents no pattern of long range homozygosity (Figure 7.8). The latter suggests that the derived allele was neutral in Africa, but that after migration out of Africa there was strong selection for the ancestral allele. Therefore, the G allele (which had stronger p53 predicted binding) was assessed to have been subjected to selective pressure throughout human evolution, probably after migration out of Africa.

7.3 Discussion

The aim of this chapter was to identify and characterise SNPs in TFBS from the pool of thousands of SNPs that have been associated with cancer risk. Over 5,000 SNPs were screened that have either themselves been associated with cancer risk or are in linkage with trait-associated SNPs. Nine transcription factors (TF) were chosen, and 171 SNPs selected as candidates residing in regions where these TF could be found to bind. Bioinformatics and functional analyses showed that two

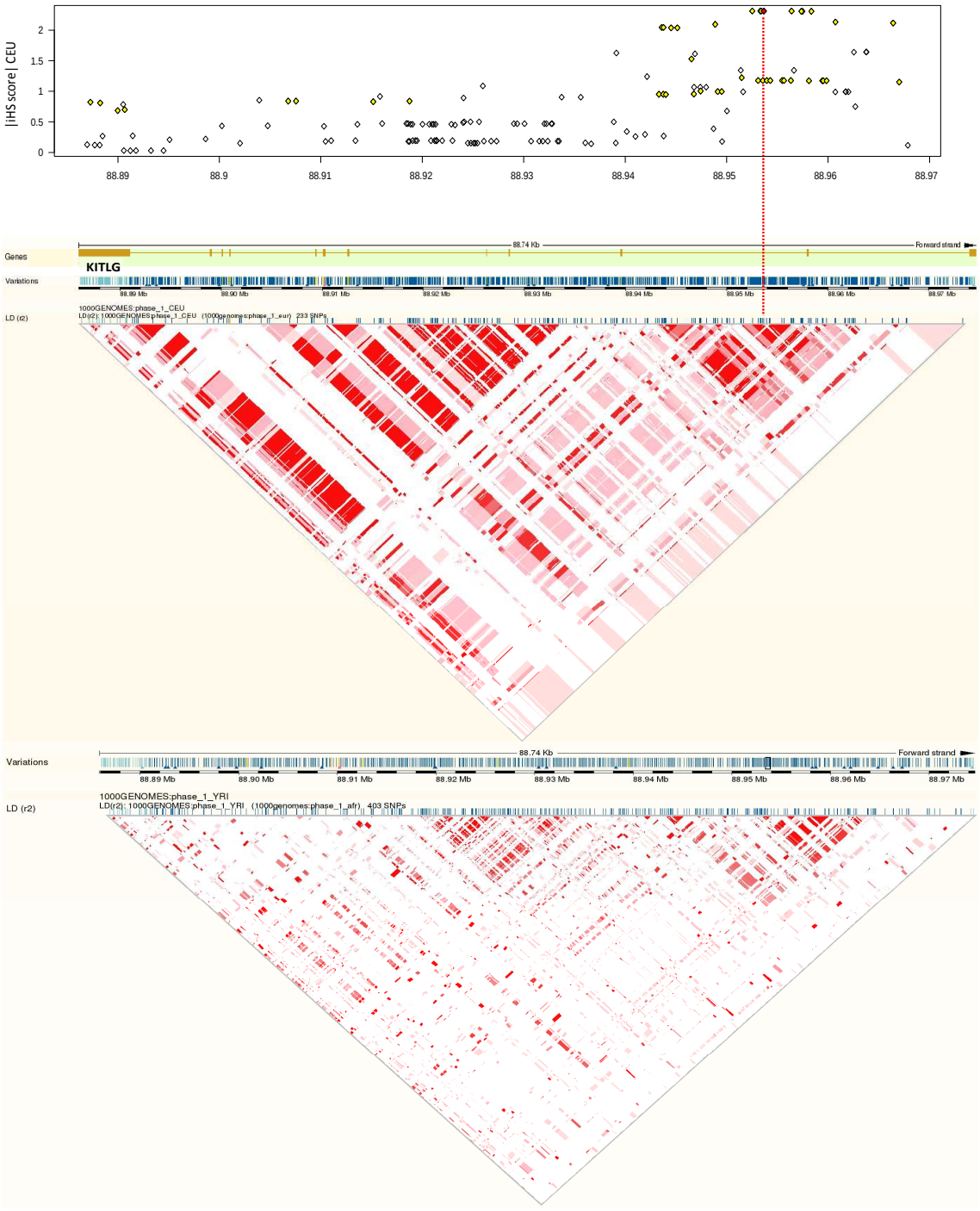


Figure 7.7: A plot of the absolute values of the standardised integrated haplotype homozygosity scores (iHS) for SNPs spanning the KITLG gene, demonstrating significant allelic differences in haplotype structure indicative of positive selection. The red line denotes the location of rs4590952 (marked in red), the yellow colouring denotes SNPs known to be in strong LD with rs4590952. Below is a schematic depiction of the KITLG gene with exons and introns noted, as well as the genetic variations in the gene, together with an LD analysis (r^2) for the CEU and YRI populations. The red colour denotes variations with high r^2 values.

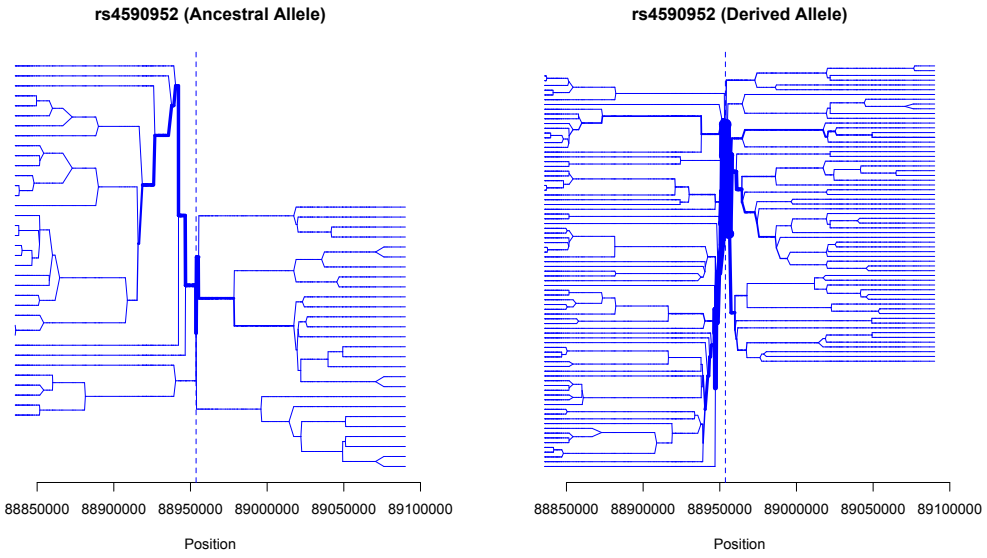


Figure 7.8: The patterns of long range homozygosity are not apparent in the YRI population, where neither the G allele (ancestral) nor the A allele (derived) belong to unusually long haplotypes.

SNPs had allelic differences in both binding and luciferase activity (Figures 7.1 and 7.2), and SNP rs4590952 was taken forward to further define the TF that binds in that region. The p53 TF was predicted to bind differentially *in silica*, and this binding was confirmed with a super-shift using a p53 antibody (Figure 7.3). Furthermore, rs4590952 showed allele-specific differences in the transactivation of the KITLG gene depending on p53 status. Finally, the SNP also showed evidence of signatures of natural selection throughout human evolution (Figures 7.6 and 7.7).

Based on linkage disequilibrium data from the 1000 Genomes project (1000 Genomes Project Consortium *et al.*, 2012), the G allele predisposes to a higher testicular cancer risk (Rapley *et al.*, 2009; Turnbull *et al.*, 2010; Kanetsky *et al.*, 2009). Testicular cancer is the most frequent cause of death from solid tumours in 20-40 year old men, and accounts for 60% of all malignancies that are diagnosed

in this age group (di Pietro *et al.*, 2005).

In this work, the high-risk G allele is shown to be one of two key nucleotides comprising a half-site of a p53-RE. p53 is the most commonly mutated gene in cancer and as many as 50% of all cancers have mutations in the p53 gene. However, almost all testicular germ cell tumours (TGCT) are characterised by the presence of wild type p53 (Lutzker and Barnard, 1998; di Pietro *et al.*, 2005). Even more interestingly, the p53 pathway is not attenuated in this type of cancer, but it is expressed in TGCT at higher than normal levels (Peng *et al.*, 1993; Lutzker and Barnard, 1998). In addition, more than 90% of testicular tumors can be cured with DNA-damaging agents in patients with good prognostic criteria. These very high response rates have been attributed to the function of p53 (di Pietro *et al.*, 2005; Lutzker and Levine, 1996; Gutekunst *et al.*, 2011). Together with the fact that p53 is a known tumour suppressor, these observations suggest that p53 fails to respond effectively against malignancies in TGCT and that its functionality is altered. The work described here inspired other experiments to identify SNPs in functional p53 response elements using genome-wide datasets of p53 occupancy and binding sites. SNP rs4590952 was shown to be the only one affecting a key nucleotide in a region with a strong binding site and frequently occupied by p53 in four independent ChIP-seq studies that were analysed, the rarity of which provides evidence of its direct influence on cancer susceptibility and of negative selection of polymorphic p53-REs (Zeron-Medina *et al.*, 2013).

SNP rs4590952 is located in the first intron of the KITLG gene, and evidence is provided here that it activates KITLG in a p53-dependent manner. Furthermore,

Zeron-Medina *et al.* (2013) performed allele-specific p53 ChIP analyses in four different lymphoblastoid cell lines and allele-specific KITLG transcript measurements in seven different cell lines, to show that the KITLG p53-RE is occupied by p53 in an allele-specific manner, and demonstrate an allelic imbalance in the transcript levels of the KITLG gene in a p53-dependent manner. KITLG is the ligand of a receptor tyrosine-kinase and is also known as Stem Cell Factor. KIT signalling has been found to regulate cell survival, migration, and proliferation, and it has been noted that gain-of-function mutations in c-Kit can promote tumour formation and progression (Lennartsson and Rönnstrand, 2012). As such, the KIT signalling pathway has been the target of a number of anti-cancer therapies (Lennartsson and Rönnstrand, 2006, 2012). Interestingly, SNPs in two more genes of the KIT signalling pathway (BAK1 and SPRY4) have also been associated with cancer risk in testicular cancer GWAS (Rapley *et al.*, 2009; Turnbull *et al.*, 2010; Kanetsky *et al.*, 2009), which highlights the importance of this pathway in TGCT. In this study, it is suggested that the G allele allows the p53 TF to bind to this particular DNA region and up-regulate KITLG expression, driving tumourigenesis and proliferation, and explaining the low rates of p53 mutations in testicular cancer, as depicted in the graphical abstract from Zeron-Medina *et al.* (2013) (Figure 7.9).

Under normal conditions, KIT's function is crucial in the haematopoietic system, nervous system, and intestinal motility, as well as pigmentation and fertility (Lennartsson and Rönnstrand, 2012). In more detail, in the skin, KIT signalling has been linked with increased melanin production (especially in response to UV radiation) and melanocyte migration (April and Barsh, 2007; McGowan *et al.*,

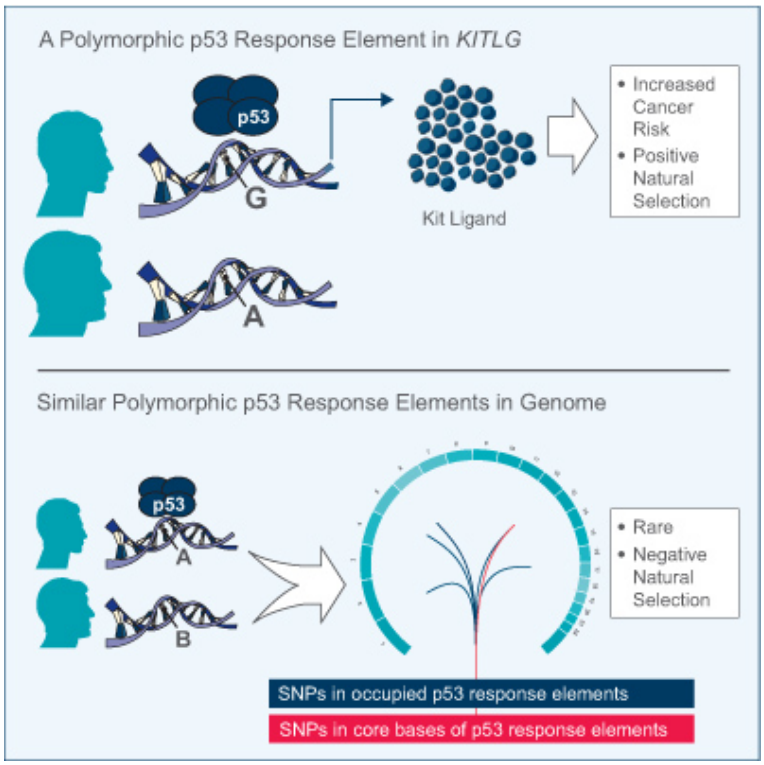


Figure 7.9: p53 binds to the KITLG SNP in an allele-specific manner, regulating the transcription of the KITLG gene, which drives proliferation and increases the risk of testicular cancer. This type of polymorphic p53-RE has been shown to be rare in the genome, as a result of negative selection.

2008; Terzian *et al.*, 2010). Interestingly, the G allele, which was linked to higher expression of the KITLG gene, has a very high frequency in Europeans compared to African populations. In addition, it has been previously noted that SNPs associated with pigmentation, among other traits, exhibit elevated iHS scores in certain geographical locations (Casto and Feldman, 2011). An intriguing hypothesis could be that the G allele offers a protective UV response, and that is why it is subject to natural selection forces in the European populations and not in African ones.

Finally, testicular cancer incidence varies by race, with non-white races having extremely low rates in comparison to white populations (Garner *et al.*, 2005; Holmes *et al.*, 2008). African-Americans do not have an increased risk of testicular cancer but are instead at higher risk for other types of cancer (Ross *et al.*, 1979;

Chia *et al.*, 2010). Together with the fact that testicular cancer's incidence rate has been increasing in Western countries since the middle of the 20th Century (Garner *et al.*, 2005), understanding the aetiology of the disease is crucial. The racial differences reported for testicular cancer incidence are in-line with the data, where it was shown that the G-risk allele of testicular cancer (Rapley *et al.*, 2009; Turnbull *et al.*, 2010; Kanetsky *et al.*, 2009) is of much lower frequency in the African populations compared to the Europeans and, therefore, could partly explain the observed differences in testicular cancer risk between the populations. However, the identification of an association in African (origin) populations is necessary to provide further support for these findings.

8 Discussion

The objective of this thesis was to design and implement strategies that combine the identification of SNPs associating with cancer phenotypes with the systematic follow-up and functional analysis of these candidates, to facilitate a better understanding of the cancer-associated genomic loci and accelerate their incorporation into the clinic. The focus was on the associations of SNPs with three types of cancer-related phenotypes: a) survival phenotypes that included time to treatment, time to metastasis/progression and overall survival; b) chemotherapeutic response; and c) cancer risk. The functional analysis of the candidate SNPs encompassed their functional annotation based on data from the ENCODE project, the inclusion of results from eQTL studies, as well as evolutionary conservation, recent natural selection and transcription factor binding site analyses.

In Chapter 3, a methodology based on Random Survival Forest was presented, namely the Variable Ranking method. Simulated data were produced based on realistic scenarios of LD structure between the SNPs, and the method tested in a number of scenarios and models, of varying effect sizes and censoring percentages, with the addition of effects of simulated prognostic factors. The performance of the Variable Ranking, was compared to two of the most common methods for survival analysis, the log-rank test and the Cox proportional hazards model. It was shown to outperform the log-rank test for most of the models tried, with an obvious advantage over the models where additional prognostic factors had an effect on the outcome. In contrast, the Variable Ranking was of inferior performance to the Cox model for these same models, but was of equal performance for the models of

covariate-specific effects.

The identification of associations of genetic markers with time-to-event phenotypes has been a challenging task for genome-wide association studies. One of the most important reasons is the insurmountable obstacle of acquiring large and homogeneous cohorts, in terms of diagnosis and treatment, for genome-wide analysis, which are required to deal with the strict thresholds incurred by multiple hypothesis testing. This is reflected in the very low numbers of successful GWAS that have been conducted to identify SNPs associating with cancer progression, survival or response to treatment. Up to the summer of 2013, over 10,000 SNPs had been associated with human traits, but only about 50 of them were associated with cancer survival and response to treatment [2]. Furthermore, the log-rank test and the Cox proportional hazards model, the two most commonly used methods in survival analysis can have important disadvantages in a genome-wide setting. The log-rank test is known to be biased when the sample size is small and unbalanced between the groups or when the censoring is heavy in one group (Latta, 1981; Heinze *et al.*, 2003) and the power of the Cox proportional hazards models is dependent on assumptions of proportionality between the groups and on the underlying genetic effects of the SNPs (Schemper, 1992; Lunetta, 2008; Zuk *et al.*, 2012).

In this work, an attempt was made to find alternative methodologies for the identification of SNPs associated with time-to-event phenotypes of cancer progression and survival, using a novel algorithm, the Variable Ranking. Variable Ranking uses a variant of the Random Forest, which has not been widely used in SNP stud-

[2] from the database: `http://www.genome.gov/gwastudies`

ies before, the Random Survival Forest (RSF). The RSF is a model-free technique that does not require any a priori information on the potential genetic models underlying the genetic effects, so it was hypothesised that it could potentially be more powerful under certain genetic models. Moreover, even though VR uses the log-rank test to make a preliminary selection on the SNPs and reduce the amount of noise for the RSF, the final ranking depends on the RSF ranking based on the variable importance and the relative frequency on which the SNPs are selected, aiming to reduce the effect of any potential bias due to unbalanced groups on the selection process. Furthermore, a new ranking measure is introduced, which utilises both the averaged VIMPs and the relative frequency of the variables, suggesting an improved ranking performance to the VIMP alone that has so far been the most commonly used ranking measure for Random Forests. Finally, even though some of the filtering steps demonstrated in Chapter 3 have been previously shown to be having an improved performance on RF, the framework for ranking SNPs based on their associations with time-to-event data on small datasets which is introduced in this work is novel. Nonetheless, additional simulations are needed to adjust its parameters and examine its performance on other common scenarios, such as effects of rare SNPs.

The Variable Ranking methodology was subsequently applied to a cohort of patients with B-cell Chronic Lymphocytic Leukaemia (B-CLL), for the ranking of the SNPs associating with time to first treatment (TFT) as a marker for disease progression. The top SNPs were then filtered based on evidence of regulatory potential from the ENCODE data, and five SNPs were selected to be taken forward

for validation. The association of SNP rs3806318 in the LEPR gene with TFT was validated in the replication cohort, and an association with progression-free survival was also noted in the clinical trial cohort CLL4. The role of LEPR in B-CLL is not known, but there are hypotheses involving its role in haematopoiesis, suggesting that LEPR promotes cell cycle progression under co-stimulatory signals in B-cells (Lam *et al.*, 2010). A functional characterisation of this SNP could further elucidate the role of LEPR in CLL and aid in the identification of patients who are at greater risk of faster progression, offering a therapeutic opportunity.

In addition, an interesting interaction effect was detected for SNP rs11740785 with deletion 11q of the region where the ATM gene resides, in an exploratory analysis using stepwise regression (stepAIC). Stepwise multiple regression is a well-known model building technique that identifies the most parsimonious model in agreement with the observed data. It involves a stepwise process of adding (or removing) covariates into a model with the aim to identify the optimal subset that best describes the data, using specific criteria (Hocking, 1976). However, it comes with a number of assumptions and limitations the most important being the inconsistencies among the model selection algorithms, which also depend on the order by which the variables are added, and an inflated type I error due to the multiple hypothesis testing (Whittingham *et al.*, 2006). The inconsistencies are partly explained by the fact that competing models can explain the data equally well and therefore, there is no best model. Furthermore, stepwise methods involve a consecutive testing of multiple models, which is known to fall under the limitations of multiple hypothesis testing and as such have an inflated type I error (Wilkinson,

1979). The inflation of type I errors was also observed in the permutation analysis in Chapter 4, where the function `stepAIC` was run under the null of no significant interactions, to obtain empirical p-values.

The aim of this work is to generate testable hypotheses that could aid in the discovery of biomarkers in cancer research. Stepwise multiple regression is used, therefore, as a hypothesis generating technique based on the observed data of the cohorts that are analysed and not as a validation technique. Moreover, identifying the best model for prediction was never the aim of this project, as the data is too preliminary for this kind of analysis. Therefore, the fact that there may be additional models that can equally well explain the data of interest does not lessen the importance of the identified associations, if they validate in an independent cohort. Furthermore, although an inflated type I error from the multiple hypothesis testing can lead to false positives and reduce the reproducibility of the identified associations, this is an inherent problem of all methods that aim to identify novel associations from genetic studies with large sets of markers. As a first approach to this problem, permutations of the data were performed and the results from the stepwise regression were collected to obtain empirical p-values for comparison. Moreover, this project aims to overcome this important issue by testing the putative associations in independent systems. This is accomplished by, firstly, testing these specific hypotheses for validation in replication cohorts and, consequently, testing them experimentally, with the aim not only to identify and catalogue these associations but also to construct potential biological models describing their function. Therefore, as the next step, the interaction of SNP rs11740785 with deletion

11q will be tested in an additional cohort of B-CLL patients from Barcelona.

In this project, two interaction effects of deletion 11q with SNPs rs11740785 and rs4966013 were identified by stepwise regression but only the first was shown to be significant. However, interestingly, both potential interactions involve the insulin pathway in connection with the ATM gene for differences in B-CLL progression. SNP rs11740785 resides in intron 1 of the integrin alpha 1 (ITGA1) gene and has been associated with levels of fasting glucose (Billings *et al.*, 2012), whilst SNP rs4966013 resides in intron 1 of the insulin growth factor 1 receptor (IGF1R), a cell surface receptor that becomes activated by the ligands insulin-like growth factors 1 and 2 (IGF-I, IGF-II). The interaction of this pathway with deletion 11q has been noted before by Saiya-Cork *et al.* (2011), who noted an over-expression of the insulin receptor gene (INSR) in patients with deletion 11q, and in the work of Yaktapour *et al.* (2013) where patients with del11q showed the best responsiveness toward an IGF1R inhibitor, AG1024. It is tempting to hypothesise that the RSF identified ITGA1 SNP rs11740785 in the discovery cohort because of its combined marginal and interaction effect on the phenotype. However, this could not be estimated from the initial analysis, since the rankings of the SNPs were not provided from a single forest but from the iterative function of the varSel function. In addition, the function provided by the RSF to identify interactions is too computationally intensive to apply to hundreds of variables. Therefore, an interesting extension of the Variable Ranking methodology would be to rank interactions of pairs of variables in a computationally non-intensive manner. However, the VR algorithm is used here as a ranking algorithm to take forward for replication the SNPs with

evidence of association mainly based on their main effects. Stepwise regression, on the other hand, was used in the replication cohort as an exploratory technique since the number of interaction effects to be tested is tractable. Alternative approaches, such as the RSF or shrinkage methods (e.g. lasso) could have been used to identify pairs of variables in a similar fashion, but the results would have been equally exploratory. Therefore, any hypotheses generated by these methods would also need to be tested in a third cohort for replication.

Interaction effects between SNPs, meant here in the context of statistical epistasis, have not been as intensively studied as marginal effects in genomic studies. The most important reason for this is the lack of power to detect those effects, due to the multiplicity incurred by the genome-wide setting, and the fact that data mining techniques aimed at dealing with high dimensional settings are not yet extensively tested in such scenarios. The variety of definitions and types of interactions that might occur within the studied systems further complicates matters. In addition, the successes of traditional statistical tests available for the identification of interacting SNPs, based on specific modelling of the interactions effects, have so far been limited. In more detail, one of the major problems in dealing with epistasis is the multiple definitions that exist and how they translate into statistical/mathematical modelling. Phillips (2008) and Cordell (2002) have both reviewed the possible definitions of epistasis and the confusion in the field created by using the term in a variety of ways, usually without further explanations of their exact meaning. They both separate the term 'statistical epistasis' from the term 'epistasis' as it is currently used by population geneticists or biologists. They

note that the term 'epistatic' as used by Fisher (1919) is meant in the context of statistical epistasis, as the deviation from the additive effect of two loci on a phenotype. On the other hand, biologists tend to use the term in a broader sense by which the mechanism of action of a factor depends on the presence or absence of another factor (Cordell, 2002).

Furthermore, an epistatic effect as defined by biologists does not always correspond to a statistical epistasis, further confusing the use of the term (Cordell, 2002; Phillips, 2008). For example, according to the 'heterogeneity model', an individual becomes affected if he/she carries a predisposing genotype at either locus A or locus B. Even though this model does not correspond to a 'statistical epistasis', it does correspond to a broader meaning of epistasis, by which the effects of recessive gene A are affected (masked) by the effects of recessive gene B.

The term 'epistasis' in the broader sense is too difficult to define mathematically. In addition, the use of common tools in genetic epidemiology for the quantification of statistical interactions, such as linear or logistic regression, comes with the inherent problem of defining a very specific type of interactions that you would expect to be affecting a phenotype. The models are, therefore, underpowered to detect any deviation from these hypotheses, which leads to the general statement that no interactions have been detected within the system. This is one of the reasons that Cordell (2002) notes that "the degree to which statistical tests of epistasis can elucidate underlying biological interactions may be more limited than previously assumed". In Chapter 4, evidence is provided of how the genetic model assumed can affect the statistical significance observed in a cohort. This

has been substantiated before by Lunetta (2008). In addition, it has been noted that the additivity usually assumed is not supported by biological evidence, and there are many cases where this has not been the case in nature (Zuk *et al.*, 2012). In the case of interactions, this problem is magnified by the potential combinations of genetic models involved, as well as the functional effects of the biological elements. For example, if a recessive gene B operates downstream of a dominant gene A, the observed effects are going to be different to the opposite scenario (of a dominant gene A operating downstream of a recessive gene B) and as such, the two scenarios would need to be modelled in a different way. Therefore, the employment of methods that are model free will provide more flexible, as well as more stable results across the board at the cost of having reduced power under specific scenarios. Flexibility will be achieved by not needing to define precise models, and stability by not running the risk of observing different results under different model assumptions. However, under well defined assumptions testing for specific models can have a considerable gain in power.

Nonetheless, the use of data mining techniques in identifying interacting SNPs is also limited, due to the fact that the power of these methodologies is yet undefined and the only way of comparing them is in the context of very specific scenarios, usually via simulations. This leads back to the initial problem of using specific models to simulate the effects that are then to be detected, creating a vicious circle for the problem of detecting statistical epistasis. An immediate consequence of these observations is that, for a given set of data mining techniques, the superior methodology is going to depend on the type of interaction effects. An illustration

of this problem can be found in the studies of García-Magariños *et al.* (2009) and of Chen *et al.* (2011), where the Random Forest is compared to other data mining techniques to detect statistical epistasis. García-Magariños *et al.* (2009) simulated models of varying sample sizes, causal SNP *MAF* and missing data, and compared Random Forests to logistic regression, classification and regression trees (CART) and Multifactor Dimensionality Reduction (MDR) for the detection of interaction effects. They noted that, in the case of interacting SNPs which also exert marginal effects, the RF and the CART perform equally well and better than MDR. However, in the case of pure interactions for weak effects, the RF had higher detection rates than CART, but was worse than MDR. Chen *et al.* (2011), performed a review of six statistical methods, namely four variants of Logic regression (LR), Random Forest and Bayesian logistic regression, and their performance for the identification of SNP interactions in case-control studies. They noted that Genetic Programming for Association Studies and Logic Feature Selection, both variants of LR, performed the best and that Random Forest was surprisingly unsuccessful. However, they suggested that this was due to the fact that the data was simulated using Boolean expressions, thus favouring the methods based on LR, and due to their measurements of interaction importance for the Random Forest.

These observations support the idea that the conventions needed for the appropriate comparison of methods are not yet in place. This stems from the fact that the nature of interactions/epistatic effects likely to be affecting complex diseases is still largely unknown. We believe, however, that in the decades to come, advances in statistical methodologies and data mining techniques will play an important role

in elucidating these effects.

The methodology proposed in Chapter 3, which was applied in the following chapter, aims to rank SNPs based on evidence of association. With the application of the Random Survival Forest, no significance values are assigned to the SNPs and, for this reason, there is no multiple hypothesis correction step. In settings where the sample size is very low and the linkage between the SNPs is high, the multiple hypothesis correction methods are too conservative and usually fail to separate the true signal from noise. Therefore, Variable Ranking is an alternative approach to select SNPs to take forward for validation, but cannot provide genome-wide significance and as such is not a standalone technique for association analysis. Further biological support is needed in such cases and here this was provided by bioinformatics analysis, eQTL studies and ultimately by experimental evidence. These multiple layers are meant to compliment the statistical analysis and support the hypotheses in a biological way.

The use of biological knowledge in genetic association studies has been previously employed in a number of investigations, but there has been no consensus on how to best implement this idea. Pathway information has been widely used by many groups to integrate biological knowledge in a meaningful way in GWAS, methods that have been reviewed in detail in Moore *et al.* (2010) and Baranzini *et al.* (2009). More recently, groups have focussed on the idea of identifying regulatory SNPs for the prioritisation of GWAS hits. Macintyre *et al.* (2010) presented a software named is-rSNP which is used to detect SNPs that are likely to affect transcription factor binding motifs based on position weight matrix (PWM) scores,

but it does not take into account the wealth of experimental data already published by ENCODE or other sources. However, it has been shown that the use of ENCODE data and eQTL studies has been a successful way of providing putative functional annotations to GWAS SNPs, and that the trait-associated SNPs of GWAS are enriched for functional SNPs compared to background SNPs (Schaub *et al.*, 2012; Boyle *et al.*, 2012). Therefore, the approach used here of filtering based on regulatory annotations can be valuable in supporting the observed associations.

In the second part of this project, SNP rs4966013 was shown to associate with response to chemotherapeutic agents in the NCI60 cell line panel. rs4966013 resides in intron 1 of the insulin-like growth factor I receptor (IGFIR), a gene shown to play a role in cancer by activating proliferation and apoptosis protection (Samani *et al.*, 2007; Chitnis *et al.*, 2008). It was shown to bind to the FOXC2 transcription factor in an allele-specific manner and associate with IGF1R mRNA expression and relative protein levels. Furthermore, using data from the CCLE, an association with response to an IGF1R inhibitor, BMS-754807, was also noted. Finally for rs4966013, a differential time to first treatment was explored in a cohort of B-CLL patients. The proposed model hypothesised that a member of the forkhead family of transcription factors binds to rs4966013 in an allele-specific manner, activating the transcription of the IGF1R gene which leads to a reduced response to chemotherapeutics, an increased response to IGF1R inhibitors and a potential worse prognosis for B-CLL patients. However, replication of these associations and a functional characterisation of this SNP is necessary to understand its role in proliferation and cancer progression.

Throughout this thesis, cell lines were used extensively in a number of analyses and experimental procedures. However, using cell lines as model systems comes with many limitations, one of the most important being their debated similarity to their original tissue type counterparts. Since the beginning of their use, it has been argued that, after a large number of replications, some cancer cell lines have undergone many modifications and are therefore no longer representative of the tissue type that they were originally derived from. To explore this further, the transcriptomic landscape of cell lines has been extensively investigated by a number of large and small scale studies (Lukk *et al.*, 2010; Barretina *et al.*, 2012; Gillet *et al.*, 2013), but the conclusions to date have been inconsistent.

Moreover, with regards to cell line response to cytotoxic agents, cell cultures are not representative environments in which drug response can be measured consistently. In more detail, immortalised cell lines have been removed from their natural environment, and so their interactions with other cell types and the signalling molecules that form their microenvironment cannot be studied. Therefore, the drug sensitivity in culture may not be indicative of the respective *in vivo* counterpart (Weinstein, 2012). Nonetheless, cancer cell lines have been instrumental in the discovery of many findings that have been important mile stones in cancer biology. Their value as tools in cancer research has, therefore, been indubitable.

In the third part of this thesis, attempts were made to detect SNPs acting as enhancers to genes associated with cancer. In Chapter 6, a methodology was presented for the detection and characterisation of SNPs in E-box binding motifs. Four SNPs were identified that presented strong allelic binding differences in EMSAs,

and two of these (rs2029091 and rs3807989) were also eQTLs in GENEVAR. Additionally, rs2029091 was associated with mRNA expression levels of PPP3R1 in a melanoma cell line panel, supporting its functionality. Interestingly, the calcineurin cascade, of which PPP3R1 is part, has been shown to have a anti-apoptotic effect and to activate tumour growth in melanoma (Fedida-Metula *et al.*, 2012; Perotti *et al.*, 2012). Therefore, a model is proposed by which an E-box protein is bound with a higher affinity to the C allele of rs2029091, increasing the protein levels of PPP3R1 and promoting tumour progression and metastasis in melanoma, but a characterisation of the function of the SNP is crucial to support this model.

In this project, eQTL studies have been widely used to support the functionality of the findings reported. In Chapter 6, evidence was provided for 2 SNPs that associated with mRNA expression levels of their target genes. Although identifying SNPs that act as eQTLs can be important for the functional characterisation of the variants, the study of eQTLs does not provide us with conclusive evidence of whether or not SNPs alter the transcription levels of the relevant genes. The reason behind this is that this approach also has many limitations, which are briefly discussed below.

Firstly, eQTL studies use the same SNP subsets that belong to the genotyping platforms used in GWAS. Therefore, similarly to the GWAS, they cannot be used to infer causality but only associations with transcript levels. High LD structures prevent us from using eQTLs to fine-map the associations, thus eQTL studies are of more use for identifying functional haplotypes. Secondly, the genetic regulation of gene expression levels has been shown to be tissue-specific for a large proportion

of genes (Dimas *et al.*, 2009; Gerrits *et al.*, 2009; Nica *et al.*, 2011; Price *et al.*, 2011). In more detail, Nica *et al.* (2011) showed that about 30% of the identified eQTLs appeared to be tissue specific, and that only another 30% was common between all three tissue types studied (adipose tissue, skin and lymphoblastoid cell lines). These percentages were similar to those of a study by Price *et al.* (2011), where the proportion of the heritable cis-regulatory gene expression was estimated to be 37% in blood and 24% in adipose tissue.

In addition, the methodology described here is successful in identifying target *cis*-eQTLs. However, the definition of *cis*-eQTLS is not very precise in regions of extended linkage disequilibrium (Cookson *et al.*, 2009). In this study, the eQTL databases were only used to test specific hypotheses (trait-associated SNPs with mRNA expression levels of the target gene), and not whether the SNPs could be acting as eQTLs for any of the neighbouring genes. However, our approach, by which the SNPs included in the studies were selected based on their proximity to the cancer genes, supports the use of eQTLs in this way.

Finally, most human studies of eQTLs have been performed in lymphoblastoid cell lines (LCLs) from peripheral blood (Cookson *et al.*, 2009), limiting the applicability of eQTLs for other cell types. Therefore, the eQTL associations observed are bound to be biased for these types of cell lines. For example, Nica *et al.* (2010) showed that the regulatory effects that they detected were enriched for associations with immunity-related conditions, as would be expected from the use of LCLs. Nonetheless, specifically for SNPs associating with certain types of cancer, tissue specificity is of great importance for cancer progression.

Similarly to the eQTL studies, the bioinformatics filters applied using the ENCODE data were restricted to a limited number of cell lines. In more detail, the DNase I hypersensitivity clusters were based on 75 cell lines, the transcription factor ChIP-seq data on 67 cell lines and the histone modification markers data on only 9 cell lines. This choice was not based on relevance to the system studied but on the availability of data. However, with the expansion of the ENCODE databases, it will soon be possible to filter the cell lines based on tissue type to obtain the most relevant information.

In the second part of the third section, in Chapter 7, the above methodology was extended to identify GWAS SNPs in other transcription factor binding sites. A SNP associated with testicular cancer in intron 1 of the KIT ligand (KITLG) gene was identified, and the high-risk G allele was shown to be one of two key nucleotides comprising a half-site of a p53-RE. Furthermore, p53-binding was confirmed with a super-shift using a p53 antibody, and the transactivation of the KITLG gene was noted to be allele-specific in a p53-dependent manner. Based on these observations and the known proliferative effect of KIT signalling, p53 is predicted to bind to the G allele of the KITLG SNP, regulating its transcription and activating a signalling cascade that leads to a greater testicular cancer risk. Finally, the SNP was shown to be selected for throughout human evolution, promoting a potential protective UV response in European populations.

To conclude, this work proposes novel methodologies for an integrated approach for the identification and analysis of SNPs associating with cancer phenotypes. Undoubtedly there are still many limitations in the use of this approach,

such as the extensive parameter setting required for achieving an increased performance with the use of the Variable Ranking methodology and the tissue specific limitations imposed by the availability of publicly available data. Future work involving more extensive simulations on larger sample sizes and of interaction scenarios between SNPs could elucidate further the advantages of this methodology over single-marker analyses. Nonetheless, data mining techniques are shown to be invaluable in the search of biomarkers for cancer progression and response to treatment, and publicly available data of regulatory loci has been integrated for the identification of functional polymorphisms. This project highlights the importance of the functional characterisation of trait-associated SNPs, and suggests ways of bridging the gap between clinical associations and biological knowledge. This brings us a step closer to the ultimate goal of understanding the peculiarities of our genome and utilising this knowledge to personalise medicine.

A APPENDIX

A.1 Chapter 3 - Bias and coverage for models 2-4

A.2 Chapter 3 - Simulations of model 4 with sample size of 200 patients

To assess the performance of the Variable Ranking for a cohort of an equivalent size of the B-CLL cohort, the sample size was reduced to 200 patients and the analysis repeated for model 4, a model with 20% censoring and 2 additional prognostic factors associated with survival time. The filtering, according to the log-rank p-values, was performed under three different thresholds: 0.05; 0.01; and 0.005. For the sample size of 200 individuals, the threshold of 0.001 was considered too low since for many simulations no associated SNPs had a p-value under this threshold.

For the three thresholds tried, 0.01 and 0.005 did not have any significant differences between them for either scenario of effects, and both outperformed the ranking of the p-values from the log-rank test alone (Figure A.1). For this reason, the less strict threshold of 0.01 was selected for the analysis of the cohort of 181 B-CLL patients.

Associated SNP	True Value	Causal SNP (r^2 in CEU)	FM: Bias	FM: 95% Cov.	SM: Bias	SM: 95% Cov.
Scenario 1 Effect=0.7						
rs7224837	0.7	rs7224837 (1)	0.0587	92	-0.1314	79
rs2498794	0.7	rs2498794 (1)	0.0178	99	-0.1364	67
rs490726	0.7	rs490726 (1)	-0.03349	92	0.1453	81
rs508384	0 (0.7)	rs490726 (1)	NA	NA	-0.5530 (0.1470)	1 (81)
rs569910	0 (0.7)	rs490726 (1)	NA	NA	0.5526 (-0.1474)	1 (78)
rs1393491	0 (0.7)	rs490726 (0.97)	NA	NA	-0.5534 (0.1466)	0 (80)
rs7849	0 (0.7)	rs490726 (0.97)	NA	NA	-0.5406 (0.1594)	3 (77)
rs2075995	0.7	rs2075995 (1)	0.0131	97	-0.1466	67
rs2075993	0 (0.7)	rs2075995 (0.95)	NA	NA	-0.5305 (0.1695)	0 (61)
rs3820028	0 (0.7)	rs2075995 (0.93)	NA	NA	0.5250 (-0.1795)	0 (53)
rs2282720	0 (0.7)	rs2075995 (0.92)	NA	NA	-0.5250 (0.1750)	0 (57)
rs2038027	0 (0.7)	rs2075995 (0.90)	NA	NA	-0.5241 (0.1759)	0 (50)
Scenario 2 Effect=0.4						
rs7224837	0.4	rs7224837 (1)	0.0236	98	-0.0252	98
rs2498794	0.4	rs2498794 (1)	-0.0124	95	-0.0568	93
rs490726	0.4	rs490726 (1)	-0.0069	97	0.0419	97
rs508384	0 (0.4)	rs490726 (1)	NA	NA	-0.3586 (0.0414)	21 (97)
rs569910	0 (0.4)	rs490726 (1)	NA	NA	0.3578 (-0.0422)	19 (97)
rs1393491	0 (0.4)	rs490726 (0.97)	NA	NA	-0.3605 (0.0395)	22 (97)
rs7849	0 (0.4)	rs490726 (0.97)	NA	NA	-0.3514 (0.0486)	24 (97)
rs2075995	0.4	rs2075995 (1)	0.0141	98	-0.0277	94
rs2075993	0 (0.4)	rs2075995 (0.95)	NA	NA	-0.3571 (0.0429)	2 (97)
rs3820028	0 (0.4)	rs2075995 (0.93)	NA	NA	0.3571 (-0.0483)	2 (97)
rs2282720	0 (0.4)	rs2075995 (0.92)	NA	NA	-0.3532 (0.0468)	2 (94)
rs2038027	0 (0.4)	rs2075995 (0.90)	NA	NA	-0.3532 (0.0468)	4 (90)

Table A.1: Bias and coverage of estimators for model 2. The single models were fit for all associated SNPs, whereas the full model was fit with the causative effects only. Abbv: NA:not available, FM: Full Model, SM: Single Model

Associated SNP	True Value	Causal SNP (r^2 in CEU)	FM: Bias	FM: 95% Cov.	SM: Bias	SM: 95% Cov.
Scenario 1 Effect=0.7						
rs7224837	0.7	rs7224837 (1)	-0.0015	94	-0.1879	61
rs2498794	0.7	rs2498794 (1)	0.0056	93	-0.1445	63
rs490726	0.7	rs490726 (1)	-0.0254	94	0.1571	73
rs508384	0 (0.7)	rs490726 (1)	NA	NA	-0.5416 (0.1584)	1 (72)
rs569910	0 (0.7)	rs490726 (1)	NA	NA	0.5404 (-0.1596)	1 (71)
rs1393491	0 (0.7)	rs490726 (0.97)	NA	NA	-0.5402 (0.1598)	1 (69)
rs7849	0 (0.7)	rs490726 (0.97)	NA	NA	-0.5328 (0.1672)	1 (71)
rs2075995	0.7	rs2075995 (1)	0.0099	94	-0.1386	68
rs2075993	0 (0.7)	rs2075995 (0.95)	NA	NA	-0.5377 (0.1623)	0 (49)
rs3820028	0 (0.7)	rs2075995 (0.93)	NA	NA	0.5295 (-0.1705)	0 (42)
rs2282720	0 (0.7)	rs2075995 (0.92)	NA	NA	-0.5285 (0.1715)	0 (47)
rs2038027	0 (0.7)	rs2075995 (0.90)	NA	NA	-0.5335 (0.1665)	0 (52)
Scenario 2 Effect=0.4						
rs7224837	0.4	rs7224837 (1)	0.0133	96	-0.0412	96
rs2498794	0.4	rs2498794 (1)	0.0151	97	-0.0274	96
rs490726	0.4	rs490726 (1)	-0.0199	93	0.0338	91
rs508384	0 (0.4)	rs490726 (1)	NA	NA	-0.3639 (0.0361)	15 (91)
rs569910	0 (0.4)	rs490726 (1)	NA	NA	0.3636 (-0.0369)	15 (91)
rs1393491	0 (0.4)	rs490726 (0.97)	NA	NA	-0.3640 (0.0360)	16 (90)
rs7849	0 (0.4)	rs490726 (0.97)	NA	NA	-0.3596 (0.0404)	17 (92)
rs2075995	0.4	rs2075995 (1)	0.0096	93	-0.0331	89
rs2075993	0 (0.4)	rs2075995 (0.95)	NA	NA	-0.3507 (0.0493)	1 (88)
rs3820028	0 (0.4)	rs2075995 (0.93)	NA	NA	0.3444 (-0.0556)	1 (86)
rs2282720	0 (0.4)	rs2075995 (0.92)	NA	NA	-0.3463 (0.0537)	1 (85)
rs2038027	0 (0.4)	rs2075995 (0.90)	NA	NA	-0.3468 (0.0532)	0 (85)

Table A.2: Bias and coverage of estimators for model 3. The single models were fit for all associated SNPs, whereas the full model was fit with the causative effects only. Abbv: NA:not available, FM: Full Model, SM: Single Model

Associated SNP	True Value	Causal SNP (r^2 in CEU)	FM: Bias	FM: 95% Cov.	SM: Bias	SM: 95% Cov.
Scenario 1 Effect=0.7						
rs7224837	0.7	rs7224837 (1)	-0.0090	96	-0.1716	74
rs2498794	0.7	rs2498794 (1)	0.0052	96	-0.1263	77
rs490726	0.7	rs490726 (1)	-0.0001	93	0.1505	71
rs508384	0 (0.7)	rs490726 (1)	NA	NA	-0.5471 (0.1529)	1 (70)
rs569910	0 (0.7)	rs490726 (1)	NA	NA	0.5454 (-0.1546)	1 (67)
rs1393491	0 (0.7)	rs490726 (0.97)	NA	NA	-0.5498 (0.1502)	1 (71)
rs7849	0 (0.7)	rs490726 (0.97)	NA	NA	-0.5429 (0.1571)	1 (72)
rs2075995	0.7	rs2075995 (1)	0.0105	93	-0.1293	71
rs2075993	0 (0.7)	rs2075995 (0.95)	NA	NA	-0.5520 (0.1480)	0 (59)
rs3820028	0 (0.7)	rs2075995 (0.93)	NA	NA	0.5441 (-0.1559)	0 (57)
rs2282720	0 (0.7)	rs2075995 (0.92)	NA	NA	-0.5435 (0.1565)	0 (59)
rs2038027	0 (0.7)	rs2075995 (0.90)	NA	NA	-0.5483 (0.1517)	0 (64)
Scenario 2 Effect=0.4						
rs7224837	0.4	rs7224837 (1)	0.0150	98	-0.0279	97
rs2498794	0.4	rs2498794 (1)	0.0098	92	-0.0243	91
rs490726	0.4	rs490726 (1)	-0.0063	93	0.0353	94
rs508384	0 (0.4)	rs490726 (1)	NA	NA	-0.3643 (0.0357)	24 (94)
rs569910	0 (0.4)	rs490726 (1)	NA	NA	0.3637 (-0.0363)	25 (94)
rs1393491	0 (0.4)	rs490726 (0.97)	NA	NA	-0.3690 (0.0310)	23 (93)
rs7849	0 (0.4)	rs490726 (0.97)	NA	NA	-0.3588 (0.0412)	27 (93)
rs2075995	0.4	rs2075995 (1)	-0.0032	95	-0.0368	92
rs2075993	0 (0.4)	rs2075995 (0.95)	NA	NA	-0.3440 (0.0560)	4 (91)
rs3820028	0 (0.4)	rs2075995 (0.93)	NA	NA	0.3397 (-0.0603)	5 (92)
rs2282720	0 (0.4)	rs2075995 (0.92)	NA	NA	-0.3399 (0.0601)	5 (90)
rs2038027	0 (0.4)	rs2075995 (0.90)	NA	NA	-0.3446 (0.0554)	5 (90)

Table A.3: Bias and coverage of estimators for model 4. The single models were fit for all associated SNPs, whereas the full model was fit with the causative effects only. Abbv: NA:not available, FM: Full Model, SM: Single Model

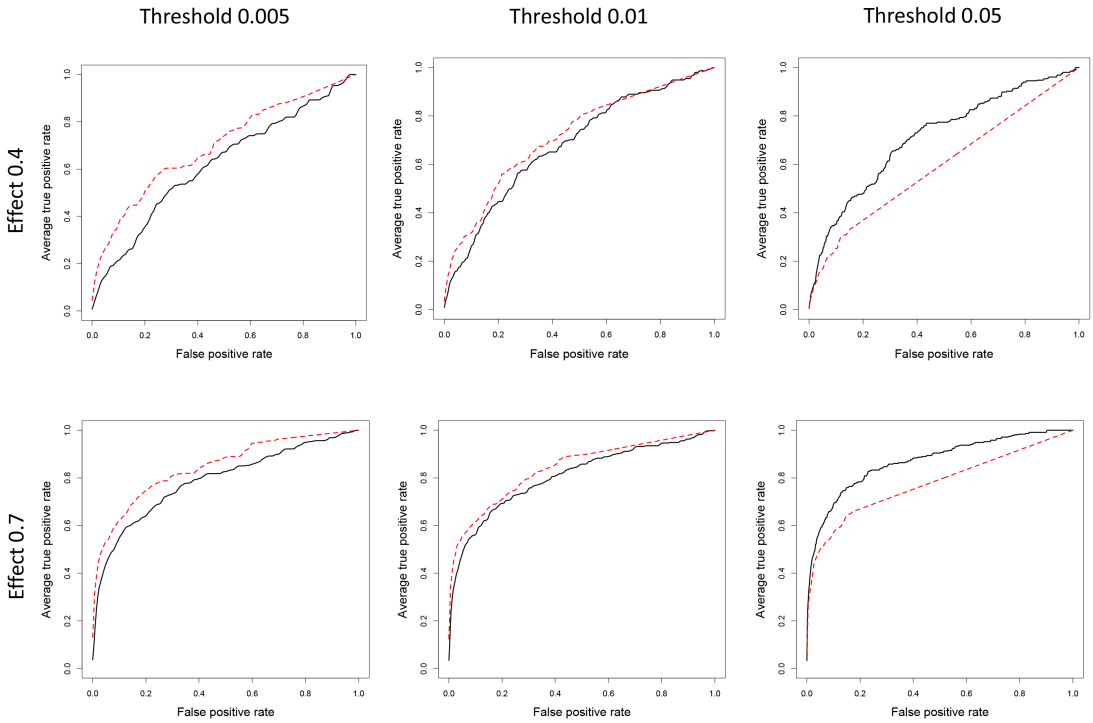


Figure A.1: Comparison of different thresholds of filtering for step 1 of the Variable Ranking for a sample size of 200 individuals. Ranking of SNPs according to the p-values of the log-rank test (black lines) and after the Variable Ranking has been applied to the tag SNPs of the top SNPs based on a p-value threshold of 0.005, 0.01 and 0.05 (red broken lines).

A.3 Chapter 5 - Replication analysis of the top hits associating with chemotherapeutic response from the NCI60 panel on the GDSC panel

21 SNPs from the analysis of the NCI60 panel had a significant p-value after the permutation tests (Chapter 7), and were selected for further replication in the Genomics of Drug Sensitivity in Cancer (GDSC) project. However, not all of the 21 SNPs were genotyped with the Affymetrix 6.0 array, and proxies were needed for the SNPs not in the platform. Because the cell lines are not all of European origin, the proxies of the SNPs that belonged to both CEU and YRI populations

were selected with the following procedure. The SNPs that were in the Affymetrix platform themselves were taken forward for validation (7 SNPs). 6 SNPs were not in the platform, nor were strongly linked to any other SNPs ($r^2 > 0.8$), so could not be further pursued. SNP rs3176352 was not in the platform and was not linked to anything in the YRI population, but was linked to another SNP ($r^2 = 0.874$) in the CEU, so that proxy SNP was selected for replication. Similarly, SNP rs9658742 was not strongly linked to any other SNPs in the CEU population, but was linked to another SNP in the YRI ($r^2 = 0.819$), which was selected for the replication. For the 6 remaining SNPs, only the proxies that were strongly linked ($r^2 > 0.8$) to the tag SNP in both populations (CEU and YRI) were taken for replication (16 SNPs in total). This amounted to 25 SNPs linked to 15 SNPs that were able to be followed up in total (Table A.4).

The same four univariate tests as for the analysis of the NCI60 panel were applied for each of the thirteen chemotherapeutic agents included in the GDSC panel. In order for the results to be comparable to the initial analysis, only cell lines of the same cancer types as in the NCI60 panel were utilised (398 cell lines). Three SNPs were identified as associating with response to chemotherapeutic agents, in the same direction of effects as in the discovery cell line panel.

SNP rs12718939 was associated with response to cisplatin and camptothecin, with p-values of 0.022 and 0.0049, respectively (Jonckheere test). The A allele was associated with higher drug concentration following an additive model (Figure A.2). More specifically, for cisplatin the median for the AA cell lines was 116 μM versus 28 μM for the GG cell lines, and for camptothecin was 0.059 μM for the

SNP	Genes	Chromosome	Position	In Affy 6.0	SNPs in GDSC
rs1005273	PDPK1	chr16	2585966	Y	rs1005273
rs12718939	EGFR	chr7	55072814	Y	rs12718939
rs1607237	PIK3CA	chr3	180432991	Y	rs1607237
rs2075677	CSE1L	chr20	47134431	Proxy	rs1983528 rs7274612
rs2234978	FAS	chr10	90761809	Y	rs2234978
rs2677760	PIK3CA	chr3	180385958	N	NA
rs3176352	CDKN1A	chr6	36760317	Proxy	rs12191972
rs3213150	E2F1	chr20	31735862	Proxy	rs3213183
rs350897	MAP2K3	chr19	4063895	N	NA
rs350903	MAP2K2	chr19	4058971	N	NA
rs3738948	ERCC3	chr2	127734533	Proxy	rs2134794 rs4300780 rs6430937
rs4140770	EGFR	chr7	55108970	Y	rs4140770
rs4966013	IGF1R	chr15	97062907	N	NA
rs6019618	CSE1L	chr20	47104945	Proxy	rs1885163 rs1983639 rs1997854 rs2295579 rs2295580 rs3818224 rs6019582 rs6019601
rs6754757	TCF7L1	chr2	85220482	N	NA
rs7286979	EP300	chr22	39828573	Proxy	rs2076578
rs866006	APC	chr5	112204458	Proxy	rs501250
rs886528	CREBBP	chr16	3751557	Y	rs886528
rs9319425	FLT1	chr13	27790985	N	NA
rs9481408	HDAC2	chr6	114371980	Y	rs9481408
rs9658742	FAS	chr10	90765609	Proxy	rs2862833

Table A.4: Table of the top 21 SNPs to be taken forward for replication in the GDSC panel. 7 SNPs were in the Affymetrix 6.0 platform, whilst 8 SNPs had proxies. Therefore, 15 SNPs could be followed up.

AA compared to 0.015 μM for the GG. This follows the same direction of effects as in the NCI60 panel. In the NCI60, rs12718939 was most prominently associated with response to alkylating agents and topoisomerase I and II inhibitors (Figure 5.4). Cisplatin and camptothecin are an alkylating agent and a topoisomerase I inhibitor, respectively. As such, the SNP was found to associate with two out of

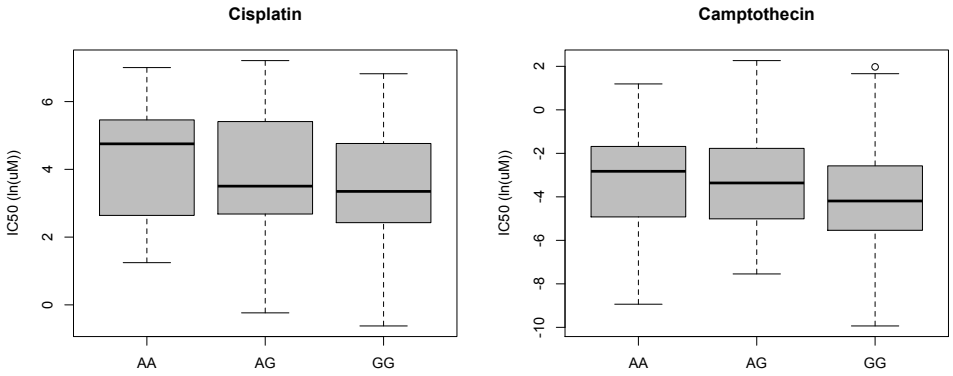


Figure A.2: Boxplots of differences in the IC50 of Cisplatin and Camptothecin between the genotypes of rs12718939. Following the same trend as in the NCI60, the AA genotype required more drug than the AG and GG, respectively.

the five chemotherapeutic agents belonging to the same groups as in the NCI60 panel of drugs.

SNP rs12718939 lays in intron 1 of the epidermal growth factor receptor (EGFR), a gene frequently deregulated in a number of different types of cancer. As such, the gene undergoes a number of genomic aberrations, including CNVs, which could be a confounding factor for this analysis. In fact, the mRNA level of EGFR is significantly correlated with the CNV within the gene, with a correlation coefficient of 0.46 (p-value$< 2.2e-16$). However, the association with camptothecin was also shown to be significant, even after adjusting for CNV in the EGFR gene, with a p-value of 0.017. The association with cisplatin was borderline not significant (p-value 0.059), probably due to the loss of power compared to the original analysis. The SNP, however, was not shown to be significantly associated with mRNA expression after adjusting for CNV in EGFR. In the NCI60 analysis, rs12718939 was found to have significant differences in the distribution of the genotypes according to cancer type (Table 5.1). However, these differences did not replicate in the CGP dataset

(Fishers test p-value 0.64), suggesting that the differences in the NCI60 panel could be solely due to chance and the small number of cell lines.

Two more SNPs were shown to associate with chemotherapeutic response in GDSC in the same direction of the effects in the NCI60 panel. SNP rs1607237 was associated with response to vinblastine, camptothecin and docetaxel, with p-values of 0.014 (Jonckheere test), 0.047 and 0.021, respectively (Wilcoxon tests). SNP rs886528 demonstrated signs of association with docetaxel, with a p-value of 0.029 (Wilcoxon tests). However, none of the two SNPs remained significant after adjusting for CNV in PIK3CA and CREBBP, respectively. Furthermore, they also showed no association with mRNA levels.

After multiple hypothesis correction (Benjamini-Hochberg), none of the 3 candidate SNPs remained significant, indicating the need for further validation for the replicability of these results.

References

SNP FAQ Archive. Bethesda (MD): National Center for Biotechnology Information (US).

1000 Genomes Project Consortium, Abecasis, G. R., Altshuler, D. *et al.* (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.

1000 Genomes Project Consortium, Abecasis, G. R., Auton, A. *et al.* (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.

Al Olama, A. A., Kote-Jarai, Z., Giles, G. G. *et al.* (2009) Multiple loci on 8q24 associated with prostate cancer susceptibility. *Nature genetics*, **41**, 1058–1060.

Amigo, J., Salas, A., Phillips, C. and Carracedo, A. (2008) SPSmart: adapting population based SNP genotype databases for fast and comprehensive web access. *Bmc Bioinformatics*, **9**, 428.

Amundadottir, L. T., Sulem, P., Gudmundsson, J. *et al.* (2006) A common variant associated with prostate cancer in European and African populations. *Nature genetics*, **38**, 652–658.

April, C. S. and Barsh, G. S. (2007) Distinct pigmentary and melanocortin 1 receptor-dependent components of cutaneous defense against ultraviolet radiation. *PLoS Genet*, **3**, e9.

Arva, N. C. (2005) A Chromatin-associated and Transcriptionally Inactive p53-Mdm2 Complex Occurs in mdm2 SNP309 Homozygous Cells. *Journal of Biological Chemistry*, **280**, 26776–26787.

Atwal, G. S., Bond, G. L., Metsuyanim, S. *et al.* (2007) Haplotype structure and selection of the MDM2 oncogene in humans. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 4524–4529.

Atwal, G. S., Kirchhoff, T., Bond, E. E. *et al.* (2009) Altered tumor formation and evolutionary selection of genetic variants in the human MDM4 oncogene. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 10236–10241.

Ayers, K. L. and Cordell, H. J. (2010) SNP Selection in genome-wide and candidate gene studies via penalized logistic regression. *Genet Epidemiol*, **34**, 879–891.

de Bakker, P. I. W., Burtt, N. P., Graham, R. R. *et al.* (2006) Transferability of tag SNPs in genetic association studies in multiple populations. *Nature genetics*, **38**, 1298–1303.

Ban, H.-J., Heo, J. Y., Oh, K.-S. and Park, K.-J. (2010) Identification of type 2 diabetes-associated combination of SNPs using support vector machine. *Bmc Genetics*, **11**, 26.

Baranzini, S. E., Galwey, N. W., Wang, J. *et al.* (2009) Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. *Human Molecular Genetics*, **18**, 2078–2090.

Barretina, J., Caponigro, G., Stransky, N. *et al.* (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, **483**, 603–607.

Barrett, J. C. (2009) Haploview: Visualization and Analysis of SNP Genotype Data. *Cold Spring Harbor Protocols*.

Barrett, J. C. and Cardon, L. R. (2006) Evaluating coverage of genome-wide association studies. *Nature genetics*, **38**, 659–662.

Barrett, J. C., Fry, B., Maller, J. and Daly, M. J. (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, **21**, 263–265.

Bartella, V., De Marco, P., Malaguarnera, R., Belfiore, A. and Maggiolini, M. (2012) New advances on the functional cross-talk between insulin-like growth factor-I and estrogen signaling in cancer. *Cellular signalling*, **24**, 1515–1521.

Belyi, V. A., Ak, P., Markert, E. *et al.* (2010) The origins and evolution of the p53 family of genes. *Cold Spring Harbor perspectives in biology*, **2**, a001198.

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

Benjamini, Y. and Yekutieli, D. (2001) The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, **29**, 1165–1188.

Bennett, B. D., Solar, G. P., Yuan, J. Q. *et al.* (1996) A role for leptin and its cognate receptor in hematopoiesis. *Curr Biol*, **6**, 1170–1180.

Bernstein, B. E., Kamal, M., Lindblad-Toh, K. *et al.* (2005) Genomic Maps and Comparative Analysis of Histone Modifications in Human and Mouse. *Cell*, **120**, 169–181.

Billings, L. K., Hsu, Y.-H., Ackerman, R. J. *et al.* (2012) Impact of common variation in bone-related genes on type 2 diabetes and related traits. *Diabetes*, **61**, 2176–2186.

Binet, J. L., Auquier, A., Dighiero, G. *et al.* (1981) A new prognostic classification of chronic lymphocytic leukemia derived from a multivariate survival analysis. *Cancer*, **48**, 198–206.

Birney, E., Stamatoyannopoulos, J. A., Dutta, A. *et al.* (2007) Identification and analysis of functional elements in 1by the ENCODE pilot project. *Nature*, **447**, 799–816.

Bond, G. L., Hu, W., Bond, E. E. *et al.* (2004) A single nucleotide polymorphism in the MDM2 promoter attenuates the p53 tumor suppressor pathway and accelerates tumor formation in humans. *Cell*, **119**, 591–602.

Bossi, G., Lapi, E., Strano, S. *et al.* (2006) Mutant p53 gain of function: reduction of tumor malignancy of human cancer cell lines through abrogation of mutant p53 expression. *Oncogene*, **25**, 304–309.

Botstein, D., White, R. L., Skolnick, M. and Davis, R. W. (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American journal of human genetics*, **32**, 314–331.

Bou-Hamad, I., Larocque, D. and Ben-Ameur, H. (2011) A review of survival trees. *Statistics Surveys*, **5**, 44–71.

Boulesteix, A.-L., Janitza, S., Kruppa, J. and König, I. R. (2012) Overview of Random Forest Methodology and Practical Guidance with Emphasis on Computational Biology and Bioinformatics.

Boyle, A. P., Hong, E. L., Hariharan, M. *et al.* (2012) Annotation of functional variation in personal genomes using RegulomeDB. *Genome research*, **22**, 1790–1797.

Breiman, L. (2001) Random forests. *Machine Learning*, **45**, 5–32.

Bretagnolle, J. and Huber-Carol, C. (1988) Effects of Omitting Covariates in Cox's Model for Survival Data. *Scandinavian Journal of Statistics*, **15**, 125–138.

Brown, M. P., Grundy, W. N., Lin, D. *et al.* (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 262–267.

Bureau, A., Dupuis, J., Falls, K. *et al.* (2005) Identifying SNPs predictive of phenotype using random forests. *Genetic Epidemiology*, **28**, 171–182.

Byrd, J. C., Stilgenbauer, S. and Flinn, I. W. (2004) Chronic lymphocytic leukemia. *Hematology Am Soc Hematol Educ Program*, 163–183.

Cannon-Albright, L. A., Goldgar, D. E., Meyer, L. J. *et al.* (1992) Assignment of a locus for familial melanoma, MLM, to chromosome 9p13-p22. *Science*, **258**, 1148–1152.

Cantor, R. M., Lange, K. and Sinsheimer, J. S. (2010) Prioritizing GWAS results: A review of statistical methods and recommendations for their application. *American journal of human genetics*, **86**, 6–22.

Caponigro, G. and Sellers, W. R. (2011) Advances in the preclinical testing of cancer therapeutic hypotheses. *Nature reviews. Drug discovery*, **10**, 179–187.

Caporaso, N., Goldin, L., Plass, C. *et al.* (2007) Chronic lymphocytic leukaemia genetics overview. *Br J Haematol*, **139**, 630–634.

Capozza, F., Trimmer, C., Castello-Cros, R. *et al.* (2012) Genetic ablation of Cav1 differentially affects melanoma tumor growth and metastasis in mice: role of Cav1 in Shh heterotypic signaling and transendothelial migration. *Cancer research*, **72**, 2262–2274.

Castle, J. C. (2011) SNPs occur in regions with less genomic sequence conservation. *PloS one*, **6**, e20660.

Casto, A. M. and Feldman, M. W. (2011) Genome-wide association study SNPs in the human genome diversity project populations: does selection affect unlinked SNPs with shared trait associations? *PLoS Genet*, **7**, e1001266–e1001266.

Catovsky, D., Richards, S., Matutes, E. *et al.* (2007) Assessment of fludarabine plus cyclophosphamide for patients with chronic lymphocytic leukaemia (the LRF CLL4 Trial): a randomised controlled trial. *Lancet*, **370**, 230–239.

Cavalli-Sforza, L. L. (2005) Opinion: The Human Genome Diversity Project: past, present and future. *Nat Rev Genet*, **6**, 333–340.

Chang, J. S., Yeh, R.-F., Wiencke, J. K. *et al.* (2008) Pathway analysis of single-nucleotide polymorphisms potentially associated with glioblastoma multiforme susceptibility using random forests. *Cancer Epidemiol Biomarkers Prev*, **17**, 1368–1373.

Chanock, S. (2009) High marks for GWAS. *Nature genetics*, **41**, 765–766.

Charlesworth, B. (1994) The effect of background selection against deleterious mutations on weakly selected, linked variants. *Audio and Electroacoustics Newsletter, IEEE*, **63**, 213–227.

Chen, C. C. M., Schwender, H., Keith, J. *et al.* (2011) Methods for Identifying SNP Interactions: A Review on Variations of Logic Regression, Random Forest and Bayesian Logistic Regression. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, **8**, 1580–1591.

Chen, X. and Ishwaran, H. (2012) Random forests for genomic data analysis. *Genomics*, **99**, 323–329.

Cheng, L. and Zhang, D. Y. (2010) *Molecular Genetic Pathology*. Springer.

Cheverud, J. M. (2001) A simple correction for multiple comparisons in interval mapping genome scans. *Heredity*, **87**, 52–58.

Chia, V. M., Quraishi, S. M., Devesa, S. S. *et al.* (2010) International trends in the incidence of testicular cancer, 1973-2002. *Cancer Epidemiology Biomarkers & Prevention*, **19**, 1151–1159.

Chitnis, M. M., Yuen, J. S. P., Protheroe, A. S., Pollak, M. and Macaulay, V. M. (2008) The type 1 insulin-like growth factor receptor pathway. *Clin Cancer Res*, **14**, 6364–6370.

Chorley, B. N. B., Wang, X. X., Campbell, M. R. M. *et al.* (2008) Discovery and verification of functional single nucleotide polymorphisms in regulatory genomic regions: Current and developing technologies. *Mutation Research/Reviews in Mutation Research*, **659**, 11–11.

Chung, C. C. and Chanock, S. J. (2011) Current status of genome-wide association studies in cancer. *Hum Genet*, **130**, 59–78.

Conrad, D. F. and Hurles, M. E. (2007) The population genetics of structural variation. *Nature genetics*, **39**, S30–6.

Cookson, W., Liang, L., Abecasis, G., Moffatt, M. and Lathrop, M. (2009) Mapping complex disease traits with global gene expression. *Nat Rev Genet*, **10**, 184–194.

Cordell, H. J. (2002) Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum Mol Genet*, **11**, 2463–2468.

Cortes, C. and Vapnik, V. (1995) Support-vector networks. *Machine Learning*, **20**, 273–297.

Couturier, C., Sarkis, C., Seron, K. *et al.* (2007) Silencing of OB-RGRP in mouse hypothalamic arcuate nucleus increases leptin receptor signaling and prevents diet-induced obesity. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 19476–19481.

Cox, A., Dunning, A. M., Garcia-Closas, M. *et al.* (2007) A common coding variant in CASP8 is associated with breast cancer risk. *Nature genetics*, **39**, 352–358.

Cox, D. R. (1972) Regression models and life-tables. *Journal of the Royal Statistical Society Series B-Methodological*, **34**, 187–220.

Crabtree, G. R. and Olson, E. N. (2002) NFAT signaling: choreographing the social lives of cells. *Cell*, **109 Suppl**, S67–79.

Crespo, M., Bosch, F., Villamor, N. *et al.* (2003) ZAP-70 expression as a surrogate for immunoglobulin-variable-region mutations in chronic lymphocytic leukemia. *N Engl J Med*, **348**, 1764–1775.

Croiseau, P. and Cordell, H. J. (2009) Analysis of North American Rheumatoid Arthritis Consortium data using a penalized logistic regression approach. *BMC proceedings*, **3**, S61.

Crowther-Swanepoel, D., Broderick, P., Di Bernardo, M. C. *et al.* (2010) Common variants at 2q37.3, 8q24.21, 15q21.3 and 16q24.1 influence chronic lymphocytic leukemia risk. *Nature genetics*, **42**, 132–136.

Dalamaga, M., Crotty, B. H., Fargnoli, J. *et al.* (2010) B-cell chronic lymphocytic leukemia risk in association with serum leptin and adiponectin: a case-control study in Greece. *Cancer Causes Control*, **21**, 1451–1459.

Damle, R. N., Wasil, T., Fais, F. *et al.* (1999) Ig V gene mutation status and CD38 expression as novel prognostic indicators in chronic lymphocytic leukemia. *Blood*, **94**, 1840–1847.

Datema, F. R., Moya, A., Krause, P. *et al.* (2012) Novel head and neck cancer survival analysis approach: random survival forests versus Cox proportional hazards regression. *Head & neck*, **34**, 50–58.

DeVita, V. T. and Chu, E. (2008) A History of Cancer Chemotherapy. *Cancer research*, **68**, 8643–8653.

DeVita, V. T., Vincent T DeVita, J. M. D., Lawrence, T. S. and Rosenberg, S. A. (2011) *Cancer*. Principles & Practice of Oncology : Primer of the Molecular Biology of Cancer. Lippincott Williams & Wilkins.

Di Bernardo, M. C., Broderick, P., Catovsky, D. and Houlston, R. S. (2013) Common genetic variation contributes significantly to the risk of developing chronic lymphocytic leukemia. *Haematologica*, **98**, e23–4.

Di Bernardo, M. C., Crowther-Swanepoel, D., Broderick, P. *et al.* (2008) A genome-wide association study identifies six susceptibility loci for chronic lymphocytic leukemia. *Nature genetics*, **40**, 1204–1210.

Díaz-Uriarte, R. and Alvarez de Andrés, S. (2006) Gene selection and classification of microarray data using random forest. *Bmc Bioinformatics*, **7**, 3.

Dighiero, G., Maloum, K., Desablens, B. *et al.* (1998) Chlorambucil in indolent chronic lymphocytic leukemia. French Cooperative Group on Chronic Lymphocytic Leukemia. *N Engl J Med*, **338**, 1506–1514.

Dimas, A. S., Deutsch, S., Stranger, B. E. *et al.* (2009) Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science*, **325**, 1246–1250.

Dinchuk, J. E., Cao, C., Huang, F. *et al.* (2010) Insulin receptor (IR) pathway hyperactivity in IGF-IR null cells and suppression of downstream growth signaling using the dual IGF-IR/IR inhibitor, BMS-754807. *Endocrinology*, **151**, 4123–4132.

Djebali, S., Davis, C. A., Merkel, A. *et al.* (2012) Landscape of transcription in human cells. *Nature*, **489**, 101–108.

Dohner, H., Stilgenbauer, S., Benner, A. *et al.* (2000) Genomic aberrations and survival in chronic lymphocytic leukemia. *N Engl J Med*, **343**, 1910–1916.

Donehower, L. A., Harvey, M., Slagle, B. L. *et al.* (1992) Mice deficient for p53 are developmentally normal but susceptible to spontaneous tumours. *Nature*, **356**, 215–221.

Dunn, S. E., Hardman, R. A., Kari, F. W. and Barrett, J. C. (1997) Insulin-like growth factor 1 (IGF-1) alters drug sensitivity of HBL100 human breast cancer cells by inhibition of apoptosis induced by diverse anticancer drugs. *Cancer research*, **57**, 2687–2693.

Easton, D. F. and Eeles, R. A. (2008) Genome-wide association studies in cancer. *Hum Mol Genet*, **17**, R109–15.

Easton, D. F., Pooley, K. A., Dunning, A. M. *et al.* (2007) Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*, **447**, 1087–1093.

ENCODE Project Consortium, Dunham, I., Kundaje, A. *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

Ephrussi, A., Church, G. M., Tonegawa, S. and Gilbert, W. (1985) B lineage–specific interactions of an immunoglobulin enhancer with cellular factors in vivo. *Science*, **227**, 134–140.

Ernst, J., Kheradpour, P., Mikkelsen, T. S. *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43–49.

Euskirchen, G. M., Rozowsky, J. S., Wei, C.-L. *et al.* (2007) Mapping of transcription factor binding regions in mammalian cells by ChIP: comparison of array- and sequencing-based technologies. *Genome research*, **17**, 898–909.

Evers, L. and Messow, C. M. (2008) Sparse kernel methods for high-dimensional survival data. *Bioinformatics*, **24**, 1632–1638.

Fedida-Metula, S., Feldman, B., Koshelev, V. *et al.* (2012) Lipid rafts couple store-operated Ca2+ entry to constitutive activation of PKB/Akt in a Ca2+/calmodulin-, Src- and PP2A-mediated pathway and promote melanoma tumor growth. *Carcinogenesis*, **33**, 740–750.

Fisher, R. A. (1919) XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh*, **52**, 399–433.

Frazer, K. A., Ballinger, D. G., Cox, D. R. *et al.* (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–861.

Freedman, M. L., Monteiro, A. N., Gayther, S. A. *et al.* (2011) Principles for the post-GWAS functional characterization of cancer risk loci. *Nature genetics*, **43**, 513–518.

Gail, M. H., Wieand, S. and Piantadosi, S. (1984) Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika*, **71**, 431–444.

Gao, X., Starmer, J. and Martin, E. R. (2008) A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genetic Epidemiology*, **32**, 361–369.

García-Magariños, M., López-de Ullibarri, I., Cao, R. and Salas, A. (2009) Evaluating the Ability of Tree-Based Methods and Logistic Regression for the Detection of SNP-SNP Interaction. *Annals of Human Genetics*, **73**, 360–369.

Garner, M. J., Turner, M. C., Ghadirian, P. and Krewski, D. (2005) Epidemiology of testicular cancer: an overview. *International journal of cancer. Journal international du cancer*, **116**, 331–339.

Garnett, M. J., Edelman, E. J., Heidorn, S. J. *et al.* (2012) Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, **483**, 570–575.

Gautier, M. and Vitalis, R. (2012) rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinformatics*, **28**, 1176–1177.

Gerds, T. A., *prodlim: Product Limit Estimation for event history and survival analysis.*

Gerlach, K., Daniel, C., Lehr, H. A. *et al.* (2012) Transcription factor NFATc2 controls the emergence of colon cancer associated with IL-6-dependent colitis. *Cancer research*, **72**, 4340–4350.

Gerrits, A., Li, Y., Tesson, B. M. *et al.* (2009) Expression Quantitative Trait Loci Are Highly Sensitive to Cellular Differentiation State. *PLoS Genet*, **5**, e1000692–e1000692.

Gibbs, R. A., Belmont, J. W., Hardenbol, P. *et al.* (2003) The International HapMap Project. *Nature*, **426**, 789–796.

Gillet, J.-P., Varma, S. and Gottesman, M. M. (2013) The Clinical Relevance of Cancer Cell Lines. *J Natl Cancer Inst.*

Goecks, J., Nekrutenko, A. and Taylor, J. (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*, **11**, R86.

Goldin, L. R., Slager, S. L. and Caporaso, N. E. () Familial chronic lymphocytic leukemia. *Curr Opin Hematol*, **17**, 350–355.

Goldstein, B. A., Hubbard, A. E., Cutler, A. and Barcellos, L. F. (2010) An application of Random Forests to a genome-wide association dataset: Methodological considerations & new findings. *Bmc Genetics*, **11**.

Gualberto, A. and Pollak, M. (2009) Emerging role of insulin-like growth factor receptor inhibitors in oncology: early clinical trial results and future directions. *Oncogene*, **28**, 3009–3021.

Gudmundsson, J., Sulem, P., Manolescu, A. *et al.* (2007) Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nature genetics*, **39**, 631–637.

Guo, W., Elston, R. C. and Zhu, X. (2011) Evaluation of a LASSO regression approach on the unrelated samples of Genetic Analysis Workshop 17. *BMC proceedings*, **5 Suppl 9**, S12.

Guo, Y., Graber, A., McBurney, R. N. and Balasubramanian, R. (2010) Sample size and statistical power considerations in high-dimensionality data settings: a comparative study of classification algorithms. *Bmc Bioinformatics*, **11**, 447.

Gutekunst, M., Oren, M., Weilbacher, A. *et al.* (2011) p53 hypersensitivity is the predominant mechanism of the unique responsiveness of testicular germ cell tumor (TGCT) cells to cisplatin. *PloS one*, **6**, e19198.

Hall, J. M., Lee, M. K., Newman, B. *et al.* (1990) Linkage of early-onset familial breast cancer to chromosome 17q21. *Science*, **250**, 1684–1689.

Hallek, M., Cheson, B. D., Catovsky, D. *et al.* (2008) Guidelines for the diagnosis and treatment of chronic lymphocytic leukemia: a report from the International Workshop on Chronic Lymphocytic Leukemia updating the National Cancer Institute-Working Group 1996 guidelines. *Blood*, **111**, 5446–5456.

Hamblin, T. J., Davis, Z., Gardiner, A., Oscier, D. G. and Stevenson, F. K. (1999) Unmutated Ig V(H) genes are associated with a more aggressive form of chronic lymphocytic leukemia. *Blood*, **94**, 1848–1854.

Heidegger, I., Pircher, A., Klocker, H. and Massoner, P. (2011) Targeting the insulin-like growth factor network in cancer therapy. *Cancer biology & therapy*, **11**, 701–707.

Heinze, G., Gnant, M. and Schemper, M. (2003) Exact LogRank Tests for Unequal FollowUp. *Biometrics*.

Hernández, G. L., Volpert, O. V., Iñiguez, M. A. *et al.* (2001) Selective inhibition of vascular endothelial growth factor-mediated angiogenesis by cyclosporin A: roles of the nuclear factor of activated T cells and cyclooxygenase 2. *The Journal of experimental medicine*, **193**, 607–620.

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

Hocking, R. R. (1976) A Biometrics invited paper. The analysis and selection of variables in linear regression. *Biometrics*.

Hogan, P. G., Chen, L., Nardone, J. and Rao, A. (2003) Transcriptional regulation by calcium, calcineurin, and NFAT. *Genes & Development*, **17**, 2205–2232.

Holm, H., Gudbjartsson, D. F., Arnar, D. O. *et al.* (2010) Several common variants modulate heart rate, PR interval and QRS duration. *Nature genetics*, **42**, 117–122.

Holmes, L., Escalante, C., Garrison, O. *et al.* (2008) Testicular cancer incidence trends in the USA (1975-2004): plateau or shifting racial paradigm? *Public health*, **122**, 862–872.

Hosking, F. J., Dobbins, S. E. and Houlston, R. S. (2011) Genome-wide association studies for detecting cancer susceptibility. *Br Med Bull*, **97**, 27–46.

Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A. and van der Laan, M. J. (2006) Survival ensembles. *Biostatistics*, **7**, 355–373.

Hothorn, T., Lausen, B., Benner, A. and Radespiel-Tröger, M. (2003) Bagging survival trees. *Statistics in medicine*, **23**, 77–91.

Howie, B. N., Donnelly, P. and Marchini, J. (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet*, **5**, e1000529.

Hudson, M. E. and Snyder, M. (2006) High-throughput methods of regulatory element discovery. *BioTechniques*, **41**, 673–675– 677 passim.

Hunter, D. J., Kraft, P., Jacobs, K. B. *et al.* (2007) A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nature genetics*, **39**, 870–874.

International HapMap 3 Consortium, Altshuler, D. M., Gibbs, R. A. *et al.* (2010) Integrating common and rare genetic variation in diverse human populations. *Nature*, **467**, 52–58.

Ioannidis, J. P. A., Thomas, G. and Daly, M. J. (2009) Validating, augmenting and refining genome-wide association signals. *Nat Rev Genet*, **10**, 318–329.

Ishwaran, H. and Kogalur, U. B., *Random Survival Forests*, r package version 3.6.4 edn.

— (2007) Random survival forests for R. *R News*, **7**, 25–31.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H. and Lauer, M. S. (2008) Random Survival Forests. *Annals of Applied Statistics*, **2**, 841–860.

Ishwaran, H., Kogalur, U. B., Chen, X. and Minn, A. J. (2011) Random survival forests for high-dimensional data. *Statistical Analysis and Data Mining*, **4**, 115–132.

Ishwaran, H., Kogalur, U. B., Gorodeski, E. Z., Minn, A. J. and Lauer, M. S. (2010) High-Dimensional Variable Selection for Survival Data. *Journal of the American Statistical Association*, **105**, 205–217.

Jansen, R. C. and Nap, J. P. (2001) Genetical genomics: the added value from segregation. *Trends in genetics : TIG*, **17**, 388–391.

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

Jonckheere, A. R. (1954) A distribution-free k-sample test against ordered alternatives. *Biometrika*, **41**, 133–145.

Jones, S. (2004) An overview of the basic helix-loop-helix proteins. *Genome Biol*, **5**, 226.

Juhász, T., Matta, C., Veress, G. *et al.* (2009) Inhibition of calcineurin by cyclosporine A exerts multiple effects on human melanoma cell lines HT168 and WM35. *International journal of oncology*, **34**, 995–1003.

Kanehisa, M. and Goto, S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*, **28**, 27–30.

Kanetsky, P. A., Mitra, N., Vardhanabhuti, S. *et al.* (2009) Common variation in KITLG and at 5q31.3 predisposes to testicular germ cell cancer. *Nature genetics*, **41**, 811–815.

Katoh, M. and Katoh, M. (2006) FGF signaling network in the gastrointestinal tract (review). *International journal of oncology*, **29**, 163–168.

Kaye, S. B. (1998) New antimetabolites in cancer chemotherapy and their clinical impact. *British journal of cancer*, **78 Suppl 3**, 1–7.

Kern, W., Bacher, U., Haferlach, C. *et al.* (2012) Monoclonal B-cell lymphocytosis is closely related to chronic lymphocytic leukaemia and may be better classified as early-stage CLL. *Br J Haematol*, **157**, 86–96.

Khan, F. M. and Zubek, V. B. (2008) Support Vector Regression for Censored Data (SVRc): A Novel Tool for Survival Analysis. In *2008 Eighth IEEE International Conference on Data Mining (ICDM)*, 863–868. IEEE.

Kiemeney, L. A., Thorlacius, S., Sulem, P. *et al.* (2008) Sequence variant on 8q24 confers susceptibility to urinary bladder cancer. *Nature genetics*, **40**, 1307–1312.

Kim, J., Sohn, I., Son, D.-S. *et al.* (2013) Prediction of a time-to-event trait using genome wide SNP data. *Bmc Bioinformatics*, **14**, 58.

King, D. C., Taylor, J., Elnitski, L. *et al.* (2005) Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. *Genome research*, **15**, 1051–1060.

Klein, J. P. and Moeschberger, M. L. (2003) *Survival Analysis*. Techniques for Censored and Truncated Data. Springer.

Knight, S. J. L., Yau, C., Clifford, R. *et al.* (2012) Quantification of subclonal distributions of recurrent genomic aberrations in paired pre-treatment and relapse samples from patients with B-cell chronic lymphocytic leukemia. *Leukemia*, **26**, 1564–1575.

Kolbe, D., Taylor, J., Elnitski, L. *et al.* (2004) Regulatory potential scores from genome-wide three-way alignments of human, mouse, and rat. *Genome research*, **14**, 700–707.

Kooperberg, C., LeBlanc, M. and Obenchain, V. (2010) Risk prediction using genome-wide association studies. *Genet Epidemiol*, **34**, 643–652.

Lam, Q. L. K., Wang, S., Ko, O. K. H., Kincade, P. W. and Lu, L. (2010) Leptin signaling maintains B-cell homeostasis via induction of Bcl-2 and Cyclin D1. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 13812–13817.

Landa, I., Ruiz-Llorente, S., Montero-Conde, C. *et al.* (2009) The variant rs1867277 in FOXE1 gene confers thyroid cancer susceptibility through the recruitment of USF1/USF2 transcription factors. *PLoS Genet*, **5**, e1000637.

Lander, E. S. (1996) The new genomics: global views of biology. *Science*, **274**, 536–539.

Lander, E. S., Linton, L. M., Birren, B. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Lane, D. and Levine, A. (2010) p53 Research: the past thirty years and the next thirty years. *Cold Spring Harbor perspectives in biology*, **2**, a000893.

Lane, D. P. (2010) The P53 Family. Cold Spring Harbor Laboratory Press.

Lango Allen, H., Estrada, K., Lettre, G. *et al.* (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, **467**, 832–838.

Larsson, S. C. and Wolk, A. (2007) Overweight and obesity and incidence of leukemia: A meta-analysis of cohort studies. *International journal of cancer. Journal international du cancer*, **122**, 1418–1421.

Lasorella, A., Noseda, M., Beyna, M., Yokota, Y. and Iavarone, A. (2000) Id2 is a retinoblastoma protein target and mediates signalling by Myc oncoproteins. *Nature*, **407**, 592–598.

Latta, R. B. (1981) A Monte Carlo Study of Some Two-Sample Rank Tests with Censored Data. *Journal of the American Statistical Association*, **76**, 713–719.

Laurie, S. A. and Goss, G. D. (2013) Role of epidermal growth factor receptor inhibitors in epidermal growth factor receptor wild-type non-small-cell lung cancer. *J Clin Oncol*, **31**, 1061–1069.

Lennartsson, J. and Rönnstrand, L. (2006) The stem cell factor receptor/c-Kit as a drug target in cancer. *Current cancer drug targets*, **6**, 65–75.

— (2012) Stem cell factor receptor/c-Kit: from basic science to clinical implications. *Physiological reviews*, **92**, 1619–1649.

Lettre, G., Lange, C. and Hirschhorn, J. N. (2007) Genetic model testing and statistical power in population-based association studies of quantitative traits. *Genet Epidemiol*, **31**, 358–362.

Lin, J., Hocker, T. L., Singh, M. and Tsao, H. (2008) Genetics of melanoma predisposition. *The British journal of dermatology*, **159**, 286–291.

Lindblad-Toh, K., Garber, M., Zuk, O. *et al.* (2011) A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, **478**, 476–482.

Linzer, D. I. and Levine, A. J. (1979) Characterization of a 54K dalton cellular SV40 tumor antigen present in SV40-transformed cells and uninfected embryonal carcinoma cells. *Cell*, **17**, 43–52.

Liu, P., Cheng, H., Roberts, T. M. and Zhao, J. J. (2009) Targeting the phosphoinositide 3-kinase pathway in cancer. *Nature reviews. Drug discovery*, **8**, 627–644.

Loos, R. J., Rankinen, T., Chagnon, Y. *et al.* (2006) Polymorphisms in the leptin and leptin receptor genes in relation to resting metabolic rate and respiratory quotient in the Quebec Family Study. *Int J Obes (Lond)*, **30**, 183–190.

Lu, Z., Ghosh, S., Wang, Z. and Hunter, T. (2003) Downregulation of caveolin-1 function by EGF leads to the loss of E-cadherin, increased transcriptional activity of beta-catenin, and enhanced tumor cell invasion. *Cancer cell*, **4**, 499–515.

Lukk, M., Kapushesky, M., Nikkilä, J. *et al.* (2010) A global map of human gene expression. *Nature biotechnology*, **28**, 322–324.

Lunetta, K. L. (2008) Genetic association studies. *Circulation*, **118**, 96–101.

Lunetta, K. L., Hayward, L. B., Segal, J. and Van Eerdewegh, P. (2004) Screening large-scale association study data: exploiting interactions using random forests. *Bmc Genetics*, **5**, 32.

Lutzker, S. G. and Barnard, N. J. (1998) Testicular germ cell tumors: molecular understanding and clinical implications. *Molecular medicine today*, **4**, 404–411.

Lutzker, S. G. and Levine, A. J. (1996) A functionally inactive p53 protein interatocarcinoma cells is activated by either DNA damage or cellular differentiation. *Nature medicine*, **2**, 804–810.

Macaulay, V. M., Salisbury, A. J., Bohula, E. A. *et al.* (2001) Downregulation of the type 1 insulin-like growth factor receptor in mouse melanoma cells is associated with enhanced radiosensitivity and impaired activation of Atm kinase. *Oncogene*, **20**, 4029–4040.

Macintyre, G., Bailey, J., Haviv, I. and Kowalczyk, A. (2010) is-rSNP: a novel technique for in silico regulatory SNP detection. *Bioinformatics*, **26**, i524–i530.

Maenner, M. J., Denlinger, L. C., Langton, A. *et al.* (2009) Detecting gene-by-smoking interactions in a genome-wide association study of early-onset coronary heart disease using random forests. *BMC proceedings*, **3 Suppl 7**, S88.

Malkin, D., Li, F. P., Strong, L. C. *et al.* (1990) Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. *Science*, **250**, 1233–1238.

Manolio, T. A. (2010) Genomewide association studies and assessment of the risk of disease. *N Engl J Med*, **363**, 166–176.

Mantel, N. (1966) Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep*, **50**, 163–170.

Massari, M. E., Jennings, P. A. and Murre, C. (1996) The AD1 transactivation domain of E2A contains a highly conserved helix which is required for its activity in both Saccharomyces cerevisiae and mammalian cells. *Molecular and cellular biology*, **16**, 121–129.

Massari, M. E. and Murre, C. (2000) Helix-loop-helix proteins: regulators of transcription in eucaryotic organisms. *Molecular and cellular biology*, **20**, 429–440.

Mayr, C., Speicher, M. R., Kofler, D. M. *et al.* (2006) Chromosomal translocations are associated with poor prognosis in chronic lymphocytic leukemia. *Blood*, **107**, 742–751.

McGowan, K. A., Li, J. Z., Park, C. Y. *et al.* (2008) Ribosomal mutations cause p53-mediated dark skin and pleiotropic effects. *Nature genetics*, **40**, 963–970.

Meier, F., Busch, S., Lasithiotakis, K. *et al.* (2007) Combined targeting of MAPK and AKT signalling pathways is a promising strategy for melanoma treatment. *The British journal of dermatology*, **156**, 1204–1213.

Meng, Y., Yang, Q., Cuenco, K. T. *et al.* (2007) Two-stage approach for identifying single-nucleotide polymorphisms associated with rheumatoid arthritis using random forests and Bayesian networks. *BMC proceedings*, **1 Suppl 1**, S56.

Meng, Y. A., Yu, Y., Cupples, L. A., Farrer, L. A. and Lunetta, K. L. (2009) Performance of random forest when SNPs are in linkage disequilibrium. *Bmc Bioinformatics*, **10**, 78.

Mills, R. E., Walter, K., Stewart, C. *et al.* (2011) Mapping copy number variation by population-scale genome sequencing. *Nature*, **470**, 59–65.

Montagut, C. and Settleman, J. (2009) Targeting the RAF–MEK–ERK pathway in cancer therapy. *Cancer Lett*, **283**, 125–134.

Montgomery, S. B., Goode, D. L., Kvikstad, E. *et al.* (2013) The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. *Genome research*, **23**, 749–761.

Moore, J. H., Asselbergs, F. W. and Williams, S. M. (2010) Bioinformatics challenges for genome-wide association studies. *Bioinformatics*, **26**, 445–455.

Müller, M. R. and Rao, A. (2010) NFAT, immunity and cancer: a transcription factor comes of age. *Nature reviews. Immunology*, **10**, 645–656.

Nakamura, Y. (2009) DNA variations in human and medical genetics: 25 years of my experience. *J Hum Genet*, **54**, 1–8.

Nayak, M. S., Yang, J.-M. and Hait, W. N. (2007) Effect of a single nucleotide polymorphism in the murine double minute 2 promoter (SNP309) on the sensitivity to topoisomerase II-targeting drugs. *Cancer research*, **67**, 5831–5839.

Nica, A. C., Montgomery, S. B., Dimas, A. S. *et al.* (2010) Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet*, **6**, e1000895.

Nica, A. C., Parts, L., Glass, D. *et al.* (2011) The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genet*, **7**, e1002003.

Nicodemus, K. K. (2011) Letter to the editor: on the stability and ranking of predictors from random forest variable importance measures. *Brief Bioinform*, **12**, 369–373.

Nicodemus, K. K. and Malley, J. D. (2009) Predictor correlation impacts machine learning algorithms: implications for genomic studies. *Bioinformatics*, **25**, 1884–1890.

Nicodemus, K. K., Malley, J. D., Strobl, C. and Ziegler, A. (2010) The behaviour of random forest permutation-based variable importance measures under predictor correlation. *Bmc Bioinformatics*, **11**, 110.

Nicolae, D. L. D., Gamazon, E. E., Zhang, W. W. *et al.* (2010) Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet*, **6**, e1000888–e1000888.

Nitiss, J. L. (2009) Targeting DNA topoisomerase II in cancer chemotherapy. *Nat Rev Cancer*, **9**, 338–350.

Nyholt, D. R. (2004) A Simple Correction for Multiple Testing for Single-Nucleotide Polymorphisms in Linkage Disequilibrium with Each Other. *American journal of human genetics*, **74**, 5–5.

O'Connor, P. M., Jackman, J., Bae, I. *et al.* (1997) Characterization of the p53 tumor suppressor pathway in cell lines of the National Cancer Institute anticancer drug screen and correlations with the growth-inhibitory potency of 123 anticancer agents. *Cancer research*, **57**, 4285–4300.

Odom, D. T., Dowell, R. D., Jacobsen, E. S. *et al.* (2007) Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nature genetics*, **39**, 730–732.

Ouillette, P., Fossum, S., Parkin, B. *et al.* (2010) Aggressive chronic lymphocytic leukemia with elevated genomic complexity is associated with multiple gene defects in the response to DNA double-strand breaks. *Clin Cancer Res*, **16**, 835–847.

Owzar, K., Li, Z., Cox, N. and Jung, S.-H. (2012) Power and sample size calculations for SNP association studies with censored time-to-event outcomes. *Genet Epidemiol*, **36**, 538–548.

Pang, H., Hauser, M. and Minvielle, S. (2011) Pathway-based identification of SNPs predictive of survival. *Eur J Hum Genet*, **19**, 704–709.

Pang, H., Lin, A., Holford, M. *et al.* (2006) Pathway analysis using random forests classification and regression. *Bioinformatics*, **22**, 2028–2036.

Park, K. S., Shin, H. D., Park, B. L. *et al.* (2006) Polymorphisms in the leptin receptor (LEPR)–putative association with obesity and T2DM. *J Hum Genet*, **51**, 85–91.

Parker, A. S., Cheville, J. C., Janney, C. A. and Cerhan, J. R. (2002) High expression levels of insulin-like growth factor-I receptor predict poor survival among women with clear-cell renal cell carcinomas. *Human pathology*, **33**, 801–805.

Pe'er, I., de Bakker, P. I. W., Maller, J. *et al.* (2006) Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nature genetics*, **38**, 663–667.

Pelengaris, S. and Khan, M. (2013) *The Molecular Biology of Cancer*. A Bridge from Bench to Bedside. John Wiley & Sons.

Peng, H. Q., Hogg, D., Malkin, D. *et al.* (1993) Mutations of the p53 gene do not occur in testis cancer. *Cancer research*, **53**, 3574–3578.

Perk, J., Iavarone, A. and Benezra, R. (2005) Id family of helix-loop-helix proteins in cancer. *Nat Rev Cancer*, **5**, 603–614.

Perotti, V., Baldassari, P., Bersani, I. *et al.* (2012) NFATc2 is a potential therapeutic target in human melanoma. *The Journal of investigative dermatology*, **132**, 2652–2660.

Pers, T. H., Albrechtsen, A., Holst, C., Sørensen, T. I. A. and Gerds, T. A. (2009) The validation and assessment of machine learning: a game of prediction from high-dimensional data. *PloS one*, **4**, e6287.

Pfeufer, A., van Noord, C., Marciante, K. D. *et al.* (2010) Genome-wide association study of PR interval. *Nature genetics*, **42**, 153–159.

Phillips, P. C. (2008) Epistasis–the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet*, **9**, 855–867.

di Pietro, A., Vries, E. G. E. d., Gietema, J. A., Spierings, D. C. J. and de Jong, S. (2005) Testicular germ cell tumours: the paradigm of chemo-sensitive solid tumours. *The international journal of biochemistry & cell biology*, **37**, 2437–2456.

Pollak, M. (2008) Insulin, insulin-like growth factors and neoplasia. *Best practice & research. Clinical endocrinology & metabolism*, **22**, 625–638.

— (2012) The insulin and insulin-like growth factor receptor family in neoplasia: an update. *Nature Reviews Cancer*, **12**, 159–169.

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. and Siepel, A. (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome research*, **20**, 110–121.

Pool, J. E., Hellmann, I., Jensen, J. D. and Nielsen, R. (2010) Population genetic inference from genomic sequence variation. *Genome research*, **20**, 291–300.

Post, S. M., Quintás-Cardama, A., Pant, V. *et al.* (2010) A high-frequency regulatory polymorphism in the p53 pathway accelerates tumor development. *Cancer cell*, **18**, 220–230.

Price, A. L., Helgason, A., Thorleifsson, G. *et al.* (2011) Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. *PLoS Genet*, **7**, e1001317–e1001317.

Rapley, E. A., Turnbull, C., Al Olama, A. A. *et al.* (2009) A genome-wide association study of testicular germ cell tumor. *Nature genetics*, **41**, 807–810.

Rasi, S., Forconi, F., Bruscaggin, A. *et al.* (2010) Impact of the host genetic background on prognosis of chronic lymphocytic leukemia. *Blood*, **115**, 1106–1107.

Rawstron, A. C., Shingles, J., de Tute, R. *et al.* (2010) Chronic lymphocytic leukaemia (CLL) and CLL-type monoclonal B-cell lymphocytosis (MBL) show differential expression of molecules involved in lymphoid tissue homing. *Cytometry. Part B, Clinical cytometry*, **78 Suppl 1**, S42–6.

Rinaldo, A., Bacanu, S.-A., Devlin, B. *et al.* (2005) Characterization of multilocus linkage disequilibrium. *Genetic Epidemiology*, **28**, 193–206.

Risch, N. and Merikangas, K. (1996) The future of genetic studies of complex human diseases. *Science-AAAS-Weekly Paper Edition*.

Rochester, M. A., Riedemann, J., Hellawell, G. O., Brewster, S. F. and Macaulay, V. M. (2005) Silencing of the IGF1R gene enhances sensitivity to DNA-damaging agents in both PTEN wild-type and mutant human prostate cancer. *Cancer gene therapy*, **12**, 90–100.

Roshan, U., Chikkagoudar, S., Wei, Z., Wang, K. and Hakonarson, H. (2011) Ranking causal variants and associated regions in genome-wide association studies by the support vector machine and random forest. *Nucleic Acids Res*, **39**, e62.

Ross, R. K., McCurtis, J. W., Henderson, B. E. *et al.* (1979) Descriptive epidemiology of testicular and prostatic cancer in Los Angeles. *British journal of cancer*, **39**, 284–292.

Rozman, C. and Montserrat, E. (1995) Chronic lymphocytic leukemia. *N Engl J Med*, **333**, 1052–1057.

Sabeti, P. C., Reich, D. E., Higgins, J. M. *et al.* (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature*, **419**, 832–837.

Sabo, P. J., Kuehn, M. S., Thurman, R. *et al.* (2006) Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. *Nat Methods*, **3**, 511–518.

Saiya-Cork, K., Collins, R., Parkin, B. *et al.* (2011) A pathobiological role of the insulin receptor in chronic lymphocytic leukemia. *Clin Cancer Res*, **17**, 2679–2692.

Sakamuro, D., Sabbatini, P., White, E. and Prendergast, G. C. (1997) The polyproline region of p53 is required to activate apoptosis but not growth arrest. *Oncogene*, **15**, 887–898.

Samani, A. A., Yakar, S., LeRoith, D. and Brodt, P. (2007) The role of the IGF system in cancer growth and metastasis: overview and recent insights. *Endocrine reviews*, **28**, 20–47.

Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S. and Snyder, M. (2012) Linking disease associations with regulatory information in the human genome. *Genome research*, **22**, 1748–1759.

Schemper, M. (1992) Cox analysis of survival data with non-proportional hazard functions. *The Statistician*, **41**, 455–465.

Schödel, J., Bardella, C., Sciesielski, L. K. *et al.* (2012) Common genetic variants at the 11q13.3 renal cancer susceptibility locus influence binding of HIF to an enhancer of cyclin D1 expression. *Nature genetics*, **44**, 420–5– S1–2.

Schwender, H., Zucknick, M., Ickstadt, K., Bolt, H. M. and GENICA network (2004) A pilot study on the application of statistical classification procedures to molecular epidemiological data. *Toxicology letters*, **151**, 291–299.

Sellick, G. S., Catovsky, D. and Houlston, R. S. (2006) Familial chronic lymphocytic leukemia. *Semin Oncol*, **33**, 195–201.

Sellick, G. S., Wade, R., Richards, S. *et al.* (2008) Scan of 977 nonsynonymous SNPs in CLL4 trial patients for the identification of genetic variants influencing prognosis. *Blood*, **111**, 1625–1633.

Shanafelt, T. D., Kay, N. E., Call, T. G. *et al.* (2008) MBL or CLL: which classification best categorizes the clinical course of patients with an absolute lymphocyte

count ¿or= 5 x 10(9) L(-1) but a B-cell lymphocyte count . *Leuk Res*, **32**, 1458–1461.

Sharma, S. V., Haber, D. A. and Settleman, J. (2010) Cell line-based platforms to evaluate the therapeutic efficacy of candidate anticancer agents. *Nat Rev Cancer*, **10**, 241–253.

Shaw, L. M. (2001) Identification of insulin receptor substrate 1 (IRS-1) and IRS-2 as signaling intermediates in the alpha6beta4 integrin-dependent activation of phosphoinositide 3-OH kinase and promotion of invasion. *Molecular and cellular biology*, **21**, 5082–5093.

Shea, J. J., Agarwala, V. V., Philippakis, A. A. A. *et al.* (2011) Comparing strategies to fine-map the association of common SNPs at chromosome 9p21 with type 2 diabetes and myocardial infarction. *Nature genetics*, **43**, 801–805.

Sherry, S. T., Ward, M. H., Kholodov, M. *et al.* (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*, **29**, 308–311.

Shi, H., Tan, S.-j., Zhong, H. *et al.* (2009) Winter temperature and UV are tightly linked to genetic changes in the p53 tumor suppressor pathway in Eastern Asia. *American journal of human genetics*, **84**, 534–541.

Shivaswamy, P. K., Chu, W. and Jansche, M. (2007) A Support Vector Approach to Censored Targets. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, 655–660. IEEE.

Shoemaker, R. H. (2006) The NCI60 human tumour cell line anticancer drug screen. *Nat Rev Cancer*, **6**, 813–823.

Siegel, R., Naishadham, D. and Jemal, A. (2013) Cancer statistics, 2013. *CA: a cancer journal for clinicians*, **63**, 11–30.

Siepel, A., Bejerano, G., Pedersen, J. S. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research*, **15**, 1034–1050.

Slager, S. L., Goldin, L. R., Strom, S. S. *et al.* (2010) Genetic susceptibility variants for chronic lymphocytic leukemia. *Cancer Epidemiology Biomarkers & Prevention*, **19**, 1098–1102.

Slager, S. L., Kay, N. E., Fredericksen, Z. S. *et al.* (2007) Susceptibility genes and B-chronic lymphocytic leukaemia. *Br J Haematol*, **139**, 762–771.

Slager, S. L., Rabe, K. G., Achenbach, S. J. *et al.* (2011a) Genome-wide association study identifies a novel susceptibility locus at 6p21.3 among familial CLL. *Blood*, **117**, 1911–1916.

— (2011b) Genome-wide association study identifies a novel susceptibility locus at 6p21.3 among familial CLL. *Blood*, **117**, 1911–1916.

Spentzos, D., Cannistra, S. A., Grall, F. *et al.* (2007) IGF axis gene expression patterns are prognostic of survival in epithelial ovarian cancer. *Endocrine-related cancer*, **14**, 781–790.

Statnikov, A., Wang, L. and Aliferis, C. F. (2008) A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *Bmc Bioinformatics*, **9**, 319.

Steininger, A., Möbs, M., Ullmann, R. *et al.* (2011) Genomic loss of the putative tumor suppressor gene E2A in human lymphoma. *The Journal of experimental medicine*, **208**, 1585–1593.

Storey, J. D. (2003) Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, **100**, 9440–9445.

Stranger, B. E., Montgomery, S. B., Dimas, A. S. *et al.* (2012) Patterns of cis regulatory variation in diverse human populations. *PLoS Genet*, **8**, e1002639.

Stranger, B. E., Stahl, E. A. and Raj, T. (2011) Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics*, **187**, 367–383.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T. and Zeileis, A. (2008) Conditional variable importance for random forests. *Bmc Bioinformatics*, **9**, 307.

Strobl, C., Boulesteix, A.-L., Zeileis, A. and Hothorn, T. (2007) Bias in random forest variable importance measures: illustrations, sources and a solution. *Bmc Bioinformatics*, **8**, 25.

Struthers, C. A. and Kalbfleisch, J. D. (1986) Misspecified Proportional Hazard Models. *Biometrika*, **73**, 363–369.

Su, Z., Marchini, J. and Donnelly, P. (2011) HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics*, **27**, 2304–2305.

Sullivan, A., Syed, N., Gasco, M. *et al.* (2004) Polymorphism in wild-type p53 modulates response to chemotherapy in vitro and in vivo. *Oncogene*, **23**, 3328–3337.

Szalai, A. J., Wu, J., Lange, E. M. *et al.* (2005) Single-nucleotide polymorphisms in the C-reactive protein (CRP) gene promoter that affect transcription factor binding, alter transcriptional activity, and associate with differences in baseline serum CRP level. *Journal of molecular medicine (Berlin, Germany)*, **83**, 440–447.

Szymczak, S., Biernacka, J. M., Cordell, H. J. *et al.* (2009) Machine learning in genome-wide association studies. *Genet Epidemiol*, **33**, S51–S57.

Teng, A. C. T., Adamo, K., Tesson, F. and Stewart, A. F. R. (2009) Functional characterization of a promoter polymorphism that drives ACSL5 gene expression in skeletal muscle and associates with diet-induced weight loss. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, **23**, 1705–1709.

Terpstra, T. J. (1952) The asymptotic normality and consistency of Kendall's test against trend, when ties are present in one ranking. *Indagationes Mathematicae*, **14**, 327–333.

Terzian, T., Torchia, E. C., Dai, D. *et al.* (2010) p53 prevents progression of nevi to melanoma predominantly through cell cycle regulation. *Pigment Cell Melanoma Res*, **23**, 781–794.

Therneau, T. M., *A package for Survival Analysis in S.*

Thomas, G., Jacobs, K. B., Kraft, P. *et al.* (2009) A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1). *Nature genetics*, **41**, 579–584.

Thomas, M., Kalita, A., Labrecque, S. *et al.* (1999) Two polymorphic variants of wild-type p53 differ biochemically and biologically. *Molecular and cellular biology*, **19**, 1092–1100.

Tibshirani, R. (1996) Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B-Methodological*, **58**, 267–288.

— (1997) The lasso method for variable selection in the cox model. *Statistics in medicine*, **16**, 385–395.

Tomlinson, I., Webb, E., Carvajal-Carmona, L. *et al.* (2007) A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nature genetics*, **39**, 984–988.

Touw, W. G., Bayjanov, J. R., Overmars, L. *et al.* (2013) Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? *Brief Bioinform*, **14**, 315–326.

Trimmer, C., Whitaker-Menezes, D., Bonuccelli, G. *et al.* (2010) CAV1 inhibits metastatic potential in melanomas through suppression of the integrin/Src/FAK signaling pathway. *Cancer research*, **70**, 7489–7499.

Tsang, W.-p., Ho, F. Y. F., Fung, K.-p., Kong, S.-k. and Kwok, T.-t. (2005) p53-R175H mutant gains new function in regulation of doxorubicin-induced apoptosis. *International journal of cancer. Journal international du cancer*, **114**, 331–336.

Turnbull, C., Rapley, E. A., Seal, S. *et al.* (2010) Variants near DMRT1, TERT and ATF7IP are associated with testicular germ cell cancer. *Nature genetics*, **42**, 604–607.

Turner, B. C., Haffty, B. G., Narayanan, L. *et al.* (1997) Insulin-like growth factor-I receptor overexpression mediates cellular radioresistance and local breast cancer recurrence after lumpectomy and radiation. *Cancer research*, **57**, 3079–3083.

Turney, B. W., Kerr, M., Chitnis, M. M. *et al.* (2012) Depletion of the type 1 IGF receptor delays repair of radiation-induced DNA double strand breaks. *Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology*, **103**, 402–409.

Van Belle, V., Pelckmans, K., Suykens, J. and Van Huffel, S. (2007) Support vector machines for survival analysis. 1–8.

Varley, J. M. (2003) Germline TP53 mutations and Li-Fraumeni syndrome. *Human mutation*, **21**, 313–320.

Vazquez, A., Bond, E. E., Levine, A. J. and Bond, G. L. (2008) The genetics of the p53 pathway, apoptosis and cancer therapy. *Nature reviews. Drug discovery*, **7**, 979–987.

Vazquez, A., Grochola, L. F., Bond, E. E. *et al.* (2010) Chemosensitivity profiles identify polymorphisms in the p53 network genes 14-3-3tau and CD44 that affect sarcoma incidence and survival. *Cancer research*, **70**, 172–180.

Verikas, A., Gelzinis, A. and Bacauskiene, M. (2011) Mining data with random forests: A survey and results of new tests. *Pattern Recognition*, **44**, 330–349.

Veyrieras, J.-B. J., Kudaravalli, S. S., Kim, S. Y. S. *et al.* (2008) High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet*, **4**, e1000214–e1000214.

Voight, B. F., Kudaravalli, S., Wen, X. and Pritchard, J. K. (2006) A map of recent positive selection in the human genome. *PLoS Biol*, **4**, e72.

Vousden, K. H. and Lu, X. (2002) Live or let die: the cell's response to p53. *Nat Rev Cancer*, **2**, 594–604.

Vousden, K. H. and Prives, C. (2009) Blinded by the Light: The Growing Complexity of p53. *Cell*, **137**, 413–431.

Wade, R., Di Bernardo, M. C., Richards, S. *et al.* (2011) Association between single nucleotide polymorphism-genotype and outcome of patients with chronic lymphocytic leukemia in a randomized chemotherapy trial. *Haematologica*, **96**, 1496–1503.

Walters, R., Laurin, C. and Lubke, G. H. (2012) An integrated approach to reduce the impact of minor allele frequency and linkage disequilibrium on variable importance measures for genome-wide data. *Bioinformatics*, **28**, 2615–2623.

Wang, D., Chen, H., Momary, K. M. *et al.* (2008) Regulatory polymorphism in vitamin K epoxide reductase complex subunit 1 (VKORC1) affects gene expression and warfarin dose requirement. *Blood*, **112**, 1013–1021.

Wei, C.-L., Wu, Q., Vega, V. B. *et al.* (2006) A global map of p53 transcription-factor binding sites in the human genome. *Cell*, **124**, 207–219.

Weinberg, R. A. (2013) *The Biology of Cancer*. Garland Science.

Weinstein, J. N. (2012) Drug discovery: Cell lines battle cancer. *Nature*, **483**, 544–545.

Whalen, E., *CLL: A Package for CLL Gene Expression Data.*

Whibley, C., Pharoah, P. D. P. and Hollstein, M. (2009) p53 polymorphisms: cancer implications. *Nature Reviews Cancer*, **9**, 95–107.

Whitfield, T. W., Wang, J., Collins, P. J. *et al.* (2012) Functional analysis of transcription factor binding sites in human promoters. *Genome Biol*, **13**, R50.

Whittingham, M. J., Stephens, P. A., Bradbury, R. B. and Freckleton, R. P. (2006) Why do we still use stepwise modelling in ecology and behaviour? *Journal of Animal Ecology*, **75**, 1182–1189.

van Wieringen, W. N., Kun, D., Hampel, R. and Boulesteix, A.-L. (2009) Survival prediction using gene expression data: A review and comparison. *Computational Statistics & Data Analysis*, **53**, 1590–1603.

Wilcoxon, F. (1945) Individual comparisons by ranking methods. *Biometrics bulletin*, **1**, 80–83.

Wilkinson, L. (1979) Tests of significance in stepwise regression. *Psychological Bulletin*.

Winham, S. J., Colby, C. L., Freimuth, R. R. *et al.* (2012) SNP interaction detection with Random Forests in high-dimensional genetic data. *Bmc Bioinformatics*, **13**, 164.

Winkelmann, J., Schormair, B., Lichtner, P. *et al.* (2007) Genome-wide association study of restless legs syndrome identifies common variants in three genomic regions. *Nature genetics*, **39**, 1000–1006.

Wooster, R. R., Neuhausen, S. L. S., Mangion, J. J. *et al.* (1994) Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13. *Science*, **265**, 2088–2090.

Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E. and Lange, K. (2009) Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, **25**, 714–721.

Yaktapour, N., Ubelhart, R., Schüler, J. *et al.* (2013) Insulin-like growth factor-1 receptor (IGF1R) as a novel target in chronic lymphocytic leukemia. *Blood*, **122**, 1621–1633.

Yamori, T. (2003) Panel of human cancer cell lines provides valuable database for drug discovery and bioinformatics. *Cancer chemotherapy and pharmacology*, **52 Suppl 1**, S74–9.

Yang, Q., Cui, J., Chazaro, I., Cupples, L. A. and Demissie, S. (2005) Power and type I error rate of false discovery rate approaches in genome-wide association studies. *Bmc Genetics*, **6 Suppl 1**, S134.

Yang, T.-P., Beazley, C., Montgomery, S. B. *et al.* (2010) Genevar: a database and Java application for the analysis and visualization of SNP-gene associations in eQTL studies. *Bioinformatics*, **26**, 2474–2476.

Yang, W., Soares, J., Greninger, P. *et al.* (2013) Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res*, **41**, D955–61.

Ye, F., Kim, C. and Ginsberg, M. H. (2012) Reconstruction of integrin activation. *Blood*, **119**, 26–33.

Yeager, M., Orr, N., Hayes, R. B. *et al.* (2007) Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nature genetics*, **39**, 645–649.

Zanke, B. W., Greenwood, C., Rangrej, J. and Kustra, R. (2007) Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nature*.

Zeron-Medina, J., Wang, X., Repapi, E. *et al.* (2013) A Polymorphic p53 Response Element in KIT Ligand Influences Cancer Risk and Has Undergone Natural Selection. *Forthcoming in Cell*.

Ziegler, A., DeStefano, A. L., Konig, I. R. *et al.* (2007) Data mining, neural nets, trees–problems 2 and 3 of Genetic Analysis Workshop 15. *Genetic Epidemiology*, **31 Suppl 1**, S51–60.

Zuk, O., Hechter, E., Sunyaev, S. R. and Lander, E. S. (2012) The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 1193–1198.