Vrije Universiteit Amsterdam

Universiteit van Amsterdam

MSc Bioinformatics and Systems Biology

Major Internship (36 ECTS)

Course Code: XM_405027

Identifying Biomarkers of Irinotecan-Induced Neutropenia with Machine Learning Methods

Olivier M. F. Martin, Pharm.D.

VU 2576272 | UvA 11110945

August 15^{th} 2016 – January 15^{th} 2017

Supervised by

Daniël Vis, Ph.D.
Sanne Abeln, Ph.D.

Nederlands Kanker Instituut

Computational Cancer Biology Group

Abstract

Irinotecan is a key drug in the treatment of cancer. Despite its efficacy, severe adverse-events such as neutropenia are frequent (up to 36% of patients). Unfortunately, our understanding of the variability of toxicity among patients is insufficient. Here, we set out to identify new Single Nucleotide Polymorphisms (SNPs) associated with irinotecan-induced neutropenia using whole-exome sequencing. To do so, we first apply standard univariate association testing. We then explore machine learning methods with elastic net and random forests classifiers. These were used as a complementary technique to identify new SNPs using feature importance measures, but also to estimate the predictive performance of identified SNPs. To reduce the number of SNPs fed to classifiers, we filter out SNPs based on functional consequence predictions. Finally, over-representation and induced network analysis are used to facilitate interpretation. Concerning results, no SNP reached significance in the univariate analysis and trained classifiers had no more predictive performance than expected by chance. However, certain SNPs had significantly higher feature importance values than expected by chance indicating that they contained information concerning irinotecan-induced neutropenia. Many of these identified SNPs concerned genes involved in processes relevant to neutrophils, such as immunological response and hematopoiesis. This is surprising because previously described associated variants were metabolizing enzymes and transporters of irinotecan. This study thus provides new hints on the physiopathology of irinotecan-induced neutropenia. Clinical studies are however necessary to confirm the validity of identified candidate SNPs.

Acknowledgements

I would like to thank all the member of the Computational Cancer Biology group for their hospitality, namely Lodewyk Wessels for welcoming me in the group and my supervisor Daniël Vis for the very insightful discussions. I would also like to thank Sander Bins for his valuable advice about handling the clinical data. Finally, I would like to thank Sanne Abeln for her supervision and constructive comments.

Contents

List of Figures

List of Tables

Acronyms

C/EBP CCAAT Enhancer Binding Protein. 23

CCNA1 Cyclin A1. 20, 22, 24

CDK Cyclin-Dependent Kinase. 22

CRLF3 Cytokine Receptor-like Factor 3. 19, 20, 22

CRTC1 CREB Regulated Transcription Coactivator 1. 20, 24

E2F1 Retinoblastoma-Associated Protein 1. 22, 23

FGF2 Fibroblast Growth Factor 2. 20, 23, 24

GOGB1 Golgin 1. 20, 22

HLA-DP Human Leukocyte Antigen DP. 18, 20–22, 24

IGH Heavy Immunoglobulin Locus. 18, 20–22, 24

LYN Tyrosine-protein kinase LYN. 22, 23

NUDT6 Nudix Hydrolase 6. 20, 23

OR8J1 Olfactory Receptor 8J1. 20, 23

PEAK1 Pseudopodium-Enriched Atypical Kinase 1. 20, 23

PKHH2 Pleckstrin Homology Domain-Containing Family H Member 2. 20

RBL2 RetinoBlastoma-Like 2 Protein. 22, 23

SUMO3 Small Ubiquitin-like MOdifier 3. 18, 21

VWA5B2 von Willebrand factor A domain containing 5B2. 20

ABC ATP-Binding Cassette. 1

AE Adverse-Event. 2, 4, 5, 7, 17–19, 23

APC 7-ethyl-10-[4-N-(5-aminopentanoic acid)-1-piperidino]-carbonyloxycamptothecin. 1

BH Benjamini–Hochberg. 15, 39, 40

CART Classification And Regression Trees. 12, 13

CCMO Central Committee on Research Involving Human Subjects. 7

CES Carboxylesterase. 1

CNV Copy-Number Variation. 3

CPCT Center for Personalized Cancer Treatment. 7

CRC ColoRectal Cancer. 1, 22

CREB cAMP Responsive Element Binding. 23

CTCAE Common Terminology Criteria for Adverse Events. 2, 7, 17

CYP Cytochrome P450. 1

ECOG Eastern Cooperative Oncology Group. 2

FATHMM Functional Analysis through Hidden Markov Models. 4, 9, 10, 18, 21, 35

Fc Fragment Crystallizable. 22

FN False Negative. 14

FP False Positive. 14

G-CSF Granulocyte-Colony Stimulating Factor. 2, 22

GO Gene Ontology. 9, 15, 20

GWAS Genome-Wide Association Study. 2, 4–6, 8, 10, 18

HW Hardy-Weinberg. 8, 9

IIN Irinotecan-Induced Neutropenia. 2, 6, 18, 19, 21–24, 27, 28, 35

KEGG Kyoto Encyclopedia of Genes and Genomes. 9, 15

LASSO Least Absolute Shrinkage and Selection Operator. 12

MAF Minor Allele Frequency. 7, 9, 18

METC Medical Ethics Review Committee. 7

NGS Next-Generation Sequencing. 5

NPC 7-ethyl-10-(4-amino-1-piperidino)-carbonyloxycamptothecin. 1

NPV Negative Predictive Value. 14, 19

PCA Principal Components Analysis. 5, 8, 17

PPV Positive Predictive Value. 14, 19

SLCO Solute Carrier Organic Anion Transporter. 1

SN-38 7-ethyl-10-hydroxycamptothecin. 1, 3

SNP Single Nucleotide Polymorphism. i, 3–7, 10, 11, 17–24, 26, 28, 35

SUMO Small Ubiquitin-like MOdifier. 21

TN True Negative. 14

TP True Positive. 14

UGT UDP-GlucuronosylTransferase. 1, 3, 4, 18, 21

VCF Variant Call Format. 7, 8

WES Whole Exome Sequencing. 5, 7, 18, 23

WGS Whole Genome Sequencing. 5, 23

WHO World Health Organization. 1

1 Introduction

1.1 Irinotecan

Irinotecan, also known as CPT-11 [72], is a widely used chemotherapeutic agent. It is a key drug in the treatment of metastatic ColoRectal Cancer (CRC): it is indicated as first line therapy [23] and is listed as an essential medicine by the World Health Organization (WHO) [60]. In addition, it is also used in the treatment of lung cancer and other cancers of the digestive tract, such as stomach and pancreatic cancer. Chemically, irinotecan is a semi-synthetic water-soluble analog of camptothecin, a compound extracted from the Chinese tree *Camptotheca acuminate* [12].

Concerning pharmacodynamics, that is how a drug affects the organism, irinotecan and all camptothecin derivatives are reversible DNA topoisomerase I inhibitors. Topoisomerase I is an indispensable enzyme in DNA replication; it transiently cleaves one strand of DNA, allowing the replication fork to proceed, before resealing the cleavage [70]. Its inhibition impedes the progression of the replication fork, thus leading to single strand breaks and irreversible inhibition of DNA synthesis. Finally, double strand breaks are observed which lead to S-phase arrest and apoptosis [23].

Concerning pharmacokinetics, that is how the organism affects the drug, irinotecan is a prodrug that must be converted into the active 7-ethyl-10-hydroxycamptothecin (SN-38) by endogenous Carboxylesterase (CES) enzymes, more specifically the CES2 isoform; CES1 only plays a minor role in activation [51]. SN-38 is at least 100 times more cytotoxic than its parent compound and is believed to be responsible for irinotecan's efficacy [12], but also for its dose-limiting toxicity [69]. Detoxification of SN-38 is dependant of glucuronidation catalyzed by UDP-GlucuronosylTransferases (UGTs), namely UGT1A1, UGT1A17 and UGT1A19 [35]. This reaction attaches a glucuronic acid residue to SN-38, thereby inactivating it, and making it more hydrophilic and more easily eliminated through urine. This reaction is reversible by microbial β-glucuronidase located in the intestine [47]. Irinotecan is also metabolized into inactive metabolites such as 7-ethyl-10-[4-N-(5-aminopentanoic acid)-1-piperidino]-carbonyloxycamptothecin (APC) and 7-ethyl-10-(4-amino-1-piperidino)-carbonyloxycamptothecin (NPC) by Cytochrome P450 (CYP), namely CYP3A4 and CYP3A5 [73]. Finally, transcellular transporters facilitate liver uptake or outtake. These are namely ATP-Binding Cassette (ABC) transported such as ABCB1, ABCC2 and ABCG2, and Solute Carrier Organic Anion Transporter (SLCO) such as SLCO1B1 [43]. The general pharmacokinetics are briefly summarized in Figure 1.

1.2 Irinotecan-Induced Neutropenia

Irinotecan can be responsible for serious Adverse-Events (AEs), which limit its use. Its two main AEs are severe diarrhea, and myelosuppression with neutropenia [12]. Neutropenia is defined as a low blood count of neutrophils, the most common blood leukocyte [7]. These phagocytes are an essential part of the innate immune system by being the first line of defense in acute inflammation. Accordingly, neutropenia is responsible for an increased risk of infections. It is a consequence of either a lack of granulopoiesis (*i.e.* production of granulocytes, namely neutrophils), or of a peripheric destruction, or both as in Irinotecan-Induced Neutropenia (IIN) [7]. Indeed, irinotecan is an aspecific chemotherapeutic agent and will also inhibit the DNA topoisomerase I enzyme of neutrophils and neutrophilic precursors, thus leading to their apoptosis.

The severity of neutropenia can be graded by the Common Terminology Criteria for Adverse Events (CTCAE) scale that defines five increasing grades on the basis of blood neutrophil counts [16]. Severe IINs are relatively common; the reported incidence of grade-3 and grade-4 AEs in clinical trials can be up to 36% of patients [44] [9]. Such toxicities can lead to reduction of dosage or withdrawal of treatment which is associated with lesser survival and tumor progression [54]. Granulocyte-Colony Stimulating Factor (G-CSF) can be used as a prophylactic or curative therapy of IIN [16].

1.3 Factor Affecting Irinotecan Response and Toxicity

It is now well established that not all cancer patients will experience the same response or toxicity when administered chemotherapy [18]. For illustration, for two patients undergoing an irinotecan treatment, one may simply experience manageable toxicity such as nausea, while the other may experience life-threatening AEs such as febrile neutropenia. Given the importance of irinotecan in the cancer therapeutic arsenal and the incidence of severe AEs (grade 3 or more), a clear understanding of variability between patient becomes indispensable. Factors influencing patient response or toxicity can be clinical or genetic; these will influence the drug's pharmacokinetics and pharmacodynamics and thus response and toxicity. Clinical factors associated with irinotecan toxicity are: age > 65 years, male sex, Eastern Cooperative Oncology Group (ECOG) performance status ≥ 1, elevated serum creatinine and bilirubin, smoking status and co-medication [44].

Pharmacogenomic biomarkers of irinotecan toxicity are still today lacking [51]. Most of the currently identified and confirmed biomarkers are issued from gene-candidate approaches; genome-wide approaches remain seldom. It is remarkable to note here that a search in the Genome-Wide Association Study (GWAS) catalog [83], a curated collection of published GWASs, only returns two studies related to irinotecan response [28] and toxicity [27]. A more extensive search of the literature allowed us identify two more GWASs concerning irinotecan response and toxicity [78] [29]. Results are summarized in

Table 1.

The most extensively studied pharmacogenomic factor influencing irinotecan pharmacokinetics is *UGT1A1*. Its expression and activity are inversely correlated with the number dinucleotide TA repeats found in the TATA box of its promoter region; higher number dinucleotide repeats result in decreased SN-38 detoxification. The wildtype genotype is considered to be variant *UGT1A1*1* with six repeats whereas *UGT1A1*28* has seven repeats [51]. Patients with homozygote *UGT1A1*28* genotypes were shown to have a higher risk of developing severe neutropenia when compared to *UGT1A1*1* genotypes [17]. Other variants such as *UGT1A1*6* have also shown to be predictive of irinotecan toxicity [59].

Although the mechanism of action of irinotecan is well understood, little is still known about factors influencing its pharmacodynamics [51]. *CDC45L*, *NFKB1*, *PARP1*, *TDP1* and *XRCC1* have been shown to influence drug response and toxicity [32]. Nonetheless, the effect of polymorphisms affecting these genes is thought to be modest or negligible [33].

1.4 Identifying Pharmacogenomic Biomarkers

To be able to measure and identify new polymorphisms in the human genome associated with irinotecan toxicity, that is pharmacogenomic biomarkers, one must first understand the origin and structure of these variations. The 1000 Genomes Project set out in 2008 to catalog common human genetic variation using sequencing technologies and individuals from diverse human populations [1]. We know today that, for a randomly chosen pair of individuals, the proportion of nucleotide differences is estimated to be between $\frac{1}{1000}$ and $\frac{1}{1500}$ per nucleotide; this amounts to a nucleotide diversity of about 0.1% [71]. Its origin is mutational events such as pointwise mutations and major chromosomal rearrangements, but also chromosomal crossover during meiosis. Around 80% of these variations are Single Nucleotide Polymorphisms (SNPs), that is differences in one single nucleotide concerning at least 1% of individuals in a population [42]. The NCBI database of genetic variation dbSNP [74] has catalogued more than 150,000,000 SNPs as of end-2016, each with its own rs identification number. Other variations are structural variations, that is variations in the structure of a chromosome caused by events such as insertions, deletion, inversions, and duplications. These include Copy-Number Variation (CNV) that concern the number of times a part of a chromosome is repeated. These polymorphisms may occur in genes within exomes or introns but can also take place in other parts of the genomes such as Alu insertions [14].

Mutations may be inherited or occur spontaneously during the patient's lifetime. Inheritable mutations occur in germ cells, that is sperm cells and oocytes, and are referred to as germline mutations. On the other hand, mutations concerning the *soma*, that is any cell of the body except germ cells, are not inheritable and are referred to as

somatic mutations. According to the neutral theory of molecular evolution, most of these polymorphisms are neutral, that is they are neither advantageous or deleterious to the organism that bears them [20]. In certain cases, mutations can affect protein function, namely by affecting its structure or expression. This change in protein function at the molecular level may have consequences at the phenotypic level [42]. For example, certain somatic mutations, such as those concerning oncogenes such as the retinoblastoma gene *RB*, may lead to cancer [48]. For this reason, in cancer research, there exist two relevant genomes: the tumor's genome referred to as the somatic genome, and the patient's genome referred to as the germline genome. It is thought that the somatic genome determines the tumor's characteristics and thus influences the prognosis of the patient and the drug response, whereas the germline genome determines the drug's metabolism and thus its toxicity [30].

Functional consequences of mutation, such as non-synonymous SNPs, can be computationally predicted [42]. Such methods typically classify mutations as neutral or deleterious, and are of two types: structural methods and evolutionary methods. Structural methods try and estimate the change in free energy of folding caused by a mutation ($\Delta\Delta G$). These methods require thus a protein structure as input. Given the sparsity of such data, usage of these methods is limited [42]. Evolutionary methods assume that over-represented mutations are neutral and under-represented mutations are deleterious [55]. Accordingly, they depend on multiple-sequence alignments of homologous sequences. To estimate conservation of sequence and thus predict functional consequences, early methods, such as SIFT, depended on simple substitution frequencies [55], whereas, more recent methods, such as Functional Analysis through Hidden Markov Models (FATHMM), make use of the Hidden Markov Model framework [75].

The identification of genetic polymorphisms has medical applications in association studies where we try to identify genotypes that are statistically associated with a disease or a phenotype such as AEs. Historically, genetic association studies have relied on candidate-gene association studies and gene linkage analysis. Candidate-gene studies focus on a pre-specified set of genes that are putatively linked to the phenotypic trait. The determination of the gene set depends on an *a priori* hypothesis based on the biology behind the phenotype or previous studies (*e.g.* linkage analysis or GWAS) [45] [63]. These studies rely upon a case-control design, although family-based designs are possible. The identification of *UGT1A* as a pharmacogenomic biomarker for irinotecan toxicity was discovered using this methodology. Indeed, patients with Gilbert's syndrome, a common genetic liver disorder resulting in hyperbilirubinemia due to reduced activity of UGT, were initially identified as being at risk for irinotecan toxicity [82] before the *UGT1A*11* variant was identified [37]. Although these candidate-gene studies are relatively quick and cheap, they are dependent on previous knowledge. Furthermore, the selection of the set of genes is somewhat subjective and depends on the researcher's understanding of biology.

Linkage analysis studies depend on linkage-disequilibrium: the fact that neighboring

genes tend to be inherited in blocks. This observation allows the identification of the chromosomal region of the responsible gene as a preliminary step [67]. These studies rely on genotyping different family members for different genetic markers and determining the segregation of these markers with the disease across different families [8]. However, these studies suffer from low power and fail to identify SNPs with a small effect size. This is problematic because it has been estimated that the median odds ratio of a SNP to be 1.33 [50]. Moreover, such studies require genotyping families, which is more laborious than simply genotyping independent individuals.

The advent of chip-based microarrays for genotyping millions of SNPs simultaneously made GWASs possible [8]. These SNP-chips are constructed so that the SNPs they probe are in maximal linkage disequilibrium with SNPs in neighboring regions and thus are representative of a region. For this reason, the genotyped SNPs are termed tag-SNPs. The remaining SNPs, can be inferred using computational methods [62]. This allows to minimize the number of SNPs to probe to a minimum of around 500,000 by not acquiring any redundant information. However, recently, with the plummeting costs of sequencing, Next-Generation Sequencing (NGS) allows us to obtain cohorts large enough for GWAS. This can refer to Whole Exome Sequencing (WES) or Whole Genome Sequencing (WGS). Although sequencing provides much more information than SNP-chips, the complexity of data, and of the observed variations, is still not well understood [4].

GWASs statistically test the association between a phenotypic trait (*e.g.* having had AEs) and each genotype. They are most often designed in a case-control manner, although quantitative and family-based designs are possible. They most often test for association for single loci by applying independence tests such Fisher's exact test or Armitage-Cochrane trend test to each genotyped position [89]. Poor design of the study can lead to spurious associations. Examples of this are population stratification (*i.e.* non-homogenous study group) and confounding (*i.e.* the association between two variables is an artifact due to a third variable being associated with both variables). Population stratification can be tested for by using Principal Components Analysis (PCA) [64]. Adjustment for confounding covariates, such as age or sex, can be taken care of using logistic regression. Finally, because of the colossal number of statistical testing involved, multiple-testing corrections or very stringent significance thresholds are required. Several software suites, such as PLINK [68], implement many of these operations.

Classical GWAS approaches tend to concentrate on univariate effects of SNPs. Nonetheless, their effects are moderate and can often only explain a subset of the heritable risk. Moreover, univariate approaches omit interaction between genetic loci. Because of this, recent approaches to identifying associated genetic variants have been influenced by machine learning [77]. This approach encounters different challenges: the number of samples (*i.e.* patients) is many orders of magnitude smaller than the number of features (*i.e.* polymorphisms). This is referred to as the small n big p problem in machine learning and is also more generally known as the curse of dimensionality. In this setting, the number of samples needed to cover the sample space becomes unrealistically high. For

illustration, a SNP-chip containing 500,000 SNP leads to $3^{500000} = 4 \times 10^{238562}$ unique possible genotype combinations. For comparison, the number of humans that ever lived is estimated to be around 10^{11}. Luckily, most of these polymorphisms are irrelevant to the problem and will only add noise to the classifier if retained. This justifies the use of feature selection methods or sparse methods that only retain certain polymorphisms. Another notable problem is that linkage disequilibrium leads to correlated features which reduce the predictive performance of the classifier [56].

Association studies typically result in a considerable amount of significantly associated variants, all requiring validation. This is classically undertaken by attempting to replicate results in independent and sufficiently large cohorts [8]. In machine learning methods, single or nested cross-validation may be utilized to limit over-fitting and consequently, reduce the number of false positives [77]. Furthermore, GWASs provide no functional information about the identified variants and identified associations often concern unsuspected genes. Examination of variants in linkage disequilibrium may reveal variants with clearer functional implications. Functional studies such as knockouts in cell or animal models may also provide valuable information [65]. Finally, bioinformatic analysis such as over-representation analysis and network analysis can be employed to facilitate interpretation [13].

1.5 Objective of the Study

As stated previously, irinotecan pharmacogenomics are still poorly understood. We thus set out to identify new pharmacogenomic biomarkers of IINs. The scope of this study is thus exploratory, that is to try to generate new hypothesis about their origin, and not to try to confirm previously identified biomarkers.

To do so we gathered data from an irinotecan clinical trial. We first applied standard univariate association analysis techniques. We then explored machine learning methods as a supplementary method to identify new biomarkers and to estimate the predictive performance of biomarkers. Finally, we used over-representation and induced network analysis to facilitate interpretation of the identified variants.

2 Patients and Methods

2.1 Patients

Data were collected retrospectively from a clinical trial of the Dutch Center for Personalized Cancer Treatment (CPCT). The study was entitled "Feasibility Study of Biomarker Development for Response Prediction by Large Scale DNA Mutational Analysis of Metastatic Lesion". It had the ClinicalTrials.gov identifier NCT01855061 and Bio2RDF identifier NL35198.041.11. Inclusion criteria were the following: 1. Patients with a metastatic solid tumor who have failed at least one line of palliative chemotherapy and are irinotecan naïve; 2. Patients who are, as per local protocol, eligible for palliative treatment with (standard of care) irinotecan; 3. Measurable metastatic lesion(s), according to RECIST 1.1 criteria [21]; 4. Radiological measurable metastatic lesion(s) of which a histological biopsy can safely be obtained; 5. Patients age 18 years or up, willing and able to comply with the protocol as judged by the investigator with a signed informed consent. The trial ran from 5^{th} of May 2011 to the 5^{th} of May 2016 in three locations: Antoni van Leeuwenhoek Hospital, Daniël den Hoed clinic and University Medical Center Utrecht. This study was conducted with approval from the following ethics committee: Medical Ethics Review Committee (METC) and Central Committee on Research Involving Human Subjects (CCMO). All subjects provided informed consent.

2.2 Definition of Cases and Controls

The severity of neutropenia was accessed using CTCAE v4.0. Cases were defined as having had at least one grade 3 or grade 4 leukopenia during the course of the irinotecan-based chemotherapy. The justification for why leukopenia and not neutropenia was used can be found in the discussion. It is also important to note here that AE reporting was not part of the clinical trial protocol and that these are likely to be under-reported.

2.3 Sequencing and Quality Control

Because we hypothesized that the germline DNA would have a higher implication that somatic DNA in determining irinotecan's pharmacokinetics and thus toxicity, we decided to simply analyze germline DNA. Peripheral blood was drawn at baseline and sequenced using WES. This resulted in individual BAM files. Variant calling on BAM files was performed using the freeBayes v0.9.14-17-g7696787 variant detection software [24] and the Ensembl *Homo sapiens* reference genome GRCh37 [88]. This resulted in individual Variant Call Format (VCF) files that were then merged using a Python script.

Quality control criteria were the following: 1. Only SNPs were retained; 2. Minor

Allele Frequency (MAF) > 10%; 3. Hardy-Weinberg (HW) equilibrium as tested by Haldane's exact test: p-value > 0.01; 4. Sequencing depth per position (DP column in VCF file): > 20; 5. Phred-scaled probability that a polymorphism exists at a given position (QUAL column in VCF file): > 20.

2.4 Detection of Population Stratification

Population stratification, that is the systematic variation in allele frequency between cases and controls due to ancestry, can be responsible for spurious associations between genotype and phenotype [64]. To detect any signs of population stratification, PCA was applied on the centered and scaled genotypic data. Patients that deviate from the main cluster were considered as outliers and were dropped from further analysis.

2.5 Univariate Analysis

Univariate association analysis of clinical variables was performed using multivariate logistic regression. Regression coefficients were bootstrapped 10,000 times in order to estimate their distribution and access their significance. Moreover, a randomly generated binomial variable was added to the model to compare distributions with a null distribution. The probability parameter p of the binomial variable was randomly sampled from a uniform distribution from 0 to 1.

Univariate association analysis of genomic variables was carried out in PLINK 1.90 (beta 3.44, 17 Nov) [68]. P-values were estimated using adaptive permutations with a minimum of 100 and a maximum of 1,000,000,000 permutations. Because of the exploratory nature of this study and the low number of samples, no multiple-testing correction was performed. Manhattan plots were generated with the `qqman` package [80] in R 3.3.1 [79].

2.6 Machine Learning

All machine learning related analyses were carried out in R 3.3.1 [79] using Bioconductor 3.4 [34].

2.6.1 MachineLearningGWAS R Package

Machine learning for GWASs is still in its infancy and no related software suite we know of currently exist. Because of this, we developed the `MachineLearningGWAS` package for

R. It is freely available at: github.com/olivmrtn/MachineLearningGWAS. The package vignette can be read by clicking here. This package is largely depended on the following R packages: `caret` [86] for machine learning, `ggplot2` [85] for plotting and `data.table` [19] for data manipulation. It also depends on various Bioconductor packages [34]. Briefly, this package provides an interface to 1. Import VCF file and clinical data into an R S4 object (`MDT`) based on the `data.table` package; 2. Annotate variants using Bioconductor annotations packages; 3. Estimate different statistics for variants such as p-values for HW equilibrium tests or association tests as well as MAF. 4. Filter out variants or samples; 5. Use this data to train classifiers using the `caret` package, and organize results (classifiers, predictions, errors, feature importance) in one R S4 object (`MLGWAS`) based on the `data.table` package; 6. Plot genomic and classification data using the `ggplot2` package.

2.6.2 Genotype Coding

Genotypes were coded using integers 0, 1 and 2 with the following meanings: 0: homozygote for the reference allele; 1: heterozygote; 2: homozygote for the alternative allele. The coding scheme was selected because it is the most memory efficient and was shown to be the one resulting in the highest prediction accuracy in Mittag *et al.* [53]. Only autosomes were considered; gonosomes and the mitochondrial chromosome were discarded.

2.6.3 Annotations

Most annotation were gathered from Bioconductor 3.4 annotation packages [34]. Annotations of gene Entrez identifiers [58] and coding changes was performed using `TxDb.Hsapiens.UCSC.hg19.knownGene` [11]. dbSNP rs identifiers [74] were annotated using `SNPlocs.Hsapiens.dbSNP144.GRCh37` [61]. Gene names, Gene Ontology (GO) terms [5] and associated Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways [41] were annotated using `org.Hs.eg.db` [10]. Functional predictions of variants were obtained from FATHMM v2.3 [75]. All these databases were up to date at time of analysis.

2.6.4 Resampling Schemes

In this study, two resampling schemes were used: nested cross-validation and bootstrapping. Nested cross-validation was used to obtain unbiased estimates of prediction performance measure distributions, whereas bootstrap was used to approximate the distribution of feature importance.

Our nested cross-validation was based on the protocol described in Wessels *et al.* [84]. More precisely, the inner loop was 10 iterations of a 10-fold cross-validation to identify model parameters with the highest predictive performance. This was handled using caret's `trControl` argument. The outer-loop was 100 repetitions of a 3-fold cross-validation to estimate performance measure distributions. This was handled using caret's `createMultiFolds` function. Our bootstrap procedure consisted of 1,000 iterations of random sampling with replacement as handled by caret's `createResample` function.

As a reminder, in k-fold cross-validation, the dataset is partitioned in k equally-sized data subsets. Out of these partitions, only one, the test set, is used for estimate performance; the other $k-1$ partitions can be used for classifier training. This protocol is then iterated k times so that each partition is the test set exactly once. This sampling scheme is mostly employed to obtain the unbiased estimates of performance measures (*i.e.* accuracy). In the case feature selection is applied to learning, nested cross-validation may be used. Indeed, feature selection is a component of learning and our estimate of performance must be based on data that was not employed in any part thereof [84]. On the other hand, partitioning in bootstrap is based on random sampling with replacement. It is mostly applied to obtain an approximation of the distribution of a statistical value [38].

2.6.5 Feature Selection

To considerably reduce the number of features fed to classifiers, functional consequence predictions were used. The FATHMM v2.3 predicts probabilities that a SNP will be deleterious [75]. Values under 0.5 were filtered out.

Different statistical feature selections were attempted to further reduce the number of features. These were three independence tests: R's Fisher's exact test `fisher.test`, R's correlation test `cor.test`, and R's package `bnlearn` mutual information test `ci.test`. Features were selected according to their p-value. Two thresholds were experimented: 50 and 500 lowest p-values. It is important to note here that estimation of p-values was done using only the training set of the outer cross-validation loop.

2.6.6 Classifiers

Classification of cases and controls was performed by using two classification methods: logistic regression with elastic net from the `glmnet` package (caret method `'glmnet'`) [22] and random forests from the `randomForest` pacakge (caret method `'rf'`) [46]. These classifiers were selected because they were shown to be suited for GWAS-type studies [77] [26].

Hereunder, we will briefly describe both of these classification models. Concerning notation, the number of samples/patients will be written as n and the number of features/SNPs as p. Our genotype matrix will be of dimension $n \times p$ and be written $\boldsymbol{X}$. It will take values 0, 1 and 2 as described in 2.6.2. To refer to a column of samples j, we will write X_j. Our binary response vector will have dimension $n \times 1$ and be written Y and takes values 1 for cases and 0 for controls.

Logistic Regression and Penalized Estimation Methods For logistic regression, it is the probability of observing a response that is modeled using the genotype matrix $\boldsymbol{X}$ through the logistic function and regression coefficients $\boldsymbol{\beta}$. The logistic function is used so that the probability P(Y) is bounded between 0 and 1.

$$P(Y) = \frac{e^{\beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p}}{1 + e^{\beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p}}$$

The textbook method of estimating coefficients βs is maximum likelihood defined hereunder.

$$L_{ML}(\boldsymbol{\beta}) = \prod_{i|Y_i=1}^{n} P(Y_i) \prod_{i|Y_i=0}^{n} (1 - P(Y_i))$$

$$\hat{\boldsymbol{\beta}} = \arg\max L_{ML}(\boldsymbol{\beta})$$

In the case where $p \gg n$, the number of parameters (*i.e.* $\hat{\beta}_1, \ldots, \hat{\beta}_p$) to be estimated largely outnumbers the number of samples n. In this setting, maximum-likelihood may not be the desired estimation method due to the bias – variance tradeoff. As a reminder, bias refers to the expected difference between predictions and the true outcomes; variance refers to the variability of predictions relative to the use of different training datasets. Maximum-likelihood tries to estimate all p regression coefficients resulting in a model with low bias but with high variance. Because of this, in the $p \gg n$ setting, maximum-likelihood estimation is highly unstable and is characterized by low predictive performance. Moreover, the large number of parameters to be analyzed makes the resulting model hard to interpret. One alternative is to use sparse regression, also referred to as penalized regression. These parameter estimation methods introduce a penalty term to disfavor models with a large number of parameters. They can be mathematically formalized as follows where λ is a penalty term, k some power and l refers to the norm of the $\boldsymbol{\beta}$ matrix.

$$L_P(\boldsymbol{\beta}) = L_{ML}(\boldsymbol{\beta}) + \lambda \left\| \boldsymbol{\beta} \right\|_l^k$$

In the case where $k = l = 2$ (*i.e.* squared sum of βs), we talk about ridge regression; in the case where $k = l = 1$ (*i.e.* sum of absolute βs), we talk about Least Absolute Shrinkage and Selection Operator (LASSO). The LASSO is a true parsimonious model in the sense that certain regression coefficients will actually be estimated as zero. On the other hand, ridge regression will shrink coefficients but nonetheless always keep all features. That being said, Zou and Hastie [91] pointed out that the LASSO has significant disadvantages in the $p \gg n$ setting: 1. it can at most select n features; 2. when two features are correlated, it will tend to only select one; 3. ridge regression outperforms LASSO in terms of prediction accuracy.

Because of this, Zou and Hastie [91] proposed a regularized regression method that linearly combines both penalties and named it the elastic net. In this case, the likelihood function to maximize is the following.

$$L_{EN}(\boldsymbol{\beta}) = L_{ML}(\boldsymbol{\beta}) + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|_2^2$$

In the `glmnet` package, this equation is written as follows where α is a mixing parameter that takes values from 0 to 1. When $\alpha = 0$, the likelihood function is the same as in the LASSO, whereas when $\alpha = 1$, the likelihood is the same as in ridge regression.

$$L_{EN}(\boldsymbol{\beta}) = L_{ML}(\boldsymbol{\beta}) + \lambda[\frac{1-\alpha}{2} \|\boldsymbol{\beta}\|_1 + \alpha \|\boldsymbol{\beta}\|_2^2]$$

Random Forests Random forests are a tree-based classification or regression ensemble method. At its core are decision trees, such as Classification And Regression Trees (CART), that recursively partition the input space into regions. The predicted value for Y is its mean value c_m in a region. For M nodes and corresponding partitions denoted R_m, a decision tree can be modelled by the following equation.

$$Y = \sum_{m=1}^{M} c_m \times 1_{X \in R_m}$$

$$c_m = mean(Y_i | X_i \in R_m)$$

The training procedure of a decision tree consists of identifying a sequence of variables and thresholds to partition the input space. Most algorithms, such as the CART algorithm, use a greedy method in order to identify a locally optimal solution. Mathematically, this can be formalized by the following equation that must be solved recursively.

$$(j,t) = \sum_{X_{ij} \in R_m} cost(X_{ij}, Y_j) + \sum_{X_{ij} \notin R_m} cost(X_{ij}, Y_j)$$

The most commonly used splitting metric is the Gini index G defined hereunder for categorical variables with K classes and where p_{mk} is the proportion of observations of the k^{th} class and the input space region m. The Gini index is referred to as a measure of node purity in the sense that the more observations are from an identical class, the lower the value it takes.

$$G = \sum_{i=1}^{K} = p_{mk}(1 - p_{mk})$$

The main problem with decision trees is that they suffer from a high variance, that is, different datasets will lead to significantly different tree models and thus significantly different predictions. To circumvent this, bootstrap aggregating (bagging) can be used. Bagging equates to using bootstrap samples of the datasets to train B trees and then average prediction results. This can be done by using a majority vote. The insight behind this is that averaging a large number of values reduces variance. Indeed, for B independent and identically distributed random variables $Z_1, Z_2, \ldots, Z_B$, its mean has a variance equal to $\frac{\sigma^2}{B}$ where σ^2 is the variance of individual random variables. The number of bootstrap samples B is often referred to as the `ntrees` arguments in the machine learning literature.

However, because the trees are constructed using the same dataset, they are not independent of each other. For non-independent variables, the variance of their mean is equal to the following where ρ is the correlation between features.

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$$

Although the second term does indeed tend to zero as B increases, the first term remains. Because of this, the reduction in variance of decision trees is limited by the correlation of their predictions. One way to bypass this is to use randomization, that is randomly select a subset of features to be selected from in the partitioning process. The number of randomly selected features is referred to as the `mtry` parameter. To summarize, random forests are bagged CART trees with randomization.

The `mtry` parameter is one of the most important parameters in the random forest algorithm. Its selection must be based on the bias – variance tradeoff. The lower it is, the less correlation between trees and thus the lower the variance. However, this may

result in an increase in bias. Conversely, the higher its value, the higher the variance but the lower the bias.

Selection of Classifier Parameters When using nested cross-validation, selection of classifier parameters was undertaken in its inner fold using grid search. The test Cohen's Kappa was estimated for every combination of predefined parameter values and recorded. The classifier using the combination of parameters that gave the lowest test error was declared to be the final model. For the elastic net possible parameters were: $\alpha = \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and $\lambda = \{0.001, 0.01, 0.1, 1, 10, 100\}$. For random forest possible parameters were: $mtry = \{5, 10, 25, 50, 100\}$. The combination most frequently chosen using nested cross-validation was then selected to be used for the bootstrap resampling scheme.

2.6.7 Performance Measures

In the case of the nested cross-validation scheme, the number of True Positives (TPs), True Negatives (TNs), False Positives (FPs) and False Negatives (FNs) was recorded for each fold. The following performance measures were computed to access predictive performance of trained classifiers: accuracy, sensitivity, specificity, Positive Predictive Value (PPV), Negative Predictive Value (NPV) and Cohen's Kappa. Cohen's Kappa was estimated using the `psych` package. The other measures were estimated using the following formulas:

$$
\begin{aligned}
Accuracy &= \frac{TP + TN}{TP + TN + FP + FN} \\
Sensitivity &= \frac{TP}{TP + FN} \\
Specificity &= \frac{TN}{TN + FP} \\
PPV &= \frac{TP}{TP + FP} \\
NPV &= \frac{TN}{TN + FN}
\end{aligned}
$$

2.6.8 Feature Importance

Features were extracted using the `varImp` function from `caret`. This function returns the mean $\boldsymbol{\beta}$ coefficients for the elastic net and the mean decrease in node impurity as measured by the mean decrease of Gini index per node across trees for random forests.

These measures were scaled so that the most important measure had a feature importance value of 100%.

For a tree t and a node m, the mean decrease of Gini index is equal to the following where l and r are the left and right node, n the total number of observation, and n_l and n_r the number of observations in the left and right nodes.

$$\Delta G(m,t) = G(m,t) - \frac{n_l}{n} G(l,t) + \frac{n_r}{n} G(r,t)$$

This local decrease of Gini index is then averaged across all trees in the forest to obtain the feature importance measure.

$$\Delta G(m) = \sum_{t=1}^{ntrees} \Delta G(m,t)$$

2.6.9 Over-representation Analysis

Over-representation analysis attempts to identify over-represented biological categories, such as GO terms or KEGG pathways, among a predefined set of genes. To do so, a hypergeometric test compares the number of genes pertaining to a certain category between the predefined set of genes and the genes in the experiment (*i.e.* background gene set). The null hypothesis H_0 here is that the probability of a certain category in the gene set is equal to the probability in the background gene set.

The construction of our set of genes was based on the value of feature importance measures. The analysis was carried out in the ConsensusPathDB web interface [40]. Categories that were selected for an over-representation test were: molecular function, biological process, cellular component GO terms levels 2, 3, 4, 5, and pathways from all databases in Kamburov *et al.* [40]. Benjamini–Hochberg (BH) multiple-testing correction was applied and a significance level of 0.05 was set.

2.6.10 Induced Network Analysis

Induced network analysis was first introduced by Berger *et al.* [6]. It aims at placing a seed set of genes identified in an experiment in the context of biological networks (*e.g.* regulatory, protein-protein interaction). The resultant network is made of the seed genes and their functional connections, but also intermediate genes that were not in the input. These intermediate genes are included because they are significantly connected to the

seed genes. To prune out low-quality interactions, a Z-score based on a binomial proportions test can be used. In summary, this analysis helps to understand the functional relationships between the initially identified genes and may help identify other genes or proteins associated with the phenotype.

The analysis was carried out on the ConsensusPathDB web interface [40]. Considered interactions were protein-protein interactions (low confidence excluded), genetic interactions, biochemical reactions, gene regulatory interactions and drug-target interactions. All databases in ConsensusPathDB were considered and a minimum Z-score of 10 was required.

3 Results

3.1 Patients' Characteristics and Population Stratification

Before displaying patients' characteristics, we will first apply methods to identify and correct for population stratification. Population stratification occurs when the allele frequency between cases and controls differs. Its identification is pivotal for it can be responsible for spurious associations between genotype and phenotype [64]. To detect it, PCA, also known as eigenanalysis, was performed on the 43 patients and 39,964 quality controlled SNPs. Figure 2 shows the two first principal components of patent's genotypes. Patients 15, 28 and 30 appeared as clear outliers from the main cluster of patients.

Methods for dealing with stratification were never shown to work for such small sample sizes [66]. Moreover, a closer analysis of outlying patients' genotypes revealed that they had a higher proportion of homozygous variants for the reference genome than others (results not shown). This stratification could thus potentially be attributed to sequencing artifacts and not necessarily ancestry. For these reasons, no correction for ancestry will be performed and samples 15, 28 and 30 have simply been removed from the analysis.

Characteristics of the remaining 40 patients are displayed in Table 2 for categorical variables and in Table 3 for continuous variables. Cases were defined as having had at least one grade 3 or grade 4 leukopenia during the course of the irinotecan-based chemotherapy as accessed by CTCAE v4.0.

3.2 Univariate Analysis

3.2.1 Clinical Variables

Clinical variables tend to be an important predictive factor for drug response and AEs [44]. Moreover, clinical variables such as age and sex can be used to prevent confounding in association studies [65]. For this reason, we studied the influence of clinical variables on irinotecan AEs using multivariate logistic regression. Because all cases were men, sex was not included in the regression model. Coefficients were bootstrapped in order to estimate their distribution and access their significance. Moreover, a randomly generate binary variable was added to the model to compare coefficients' distributions with a null distribution. Quantiles of bootstrapped coefficients are shown in Figure 3. All 95% confidence interval overlapped with $\beta = 0$ and the random variable β distribution. Accordingly, all clinical variables were deemed to be unassociated with AEs and hence not considered for further analysis.

3.2.2 Genotypic Variables

After quality control, 41,191 SNPs were available for testing for association with IIN. The increase in the number of SNPs, when compared to Section 3.1, is caused by the rise in estimates of MAFs. This is related to the fact that that samples that were removed had a higher proportion of homozygous variants for the reference genome.

Because no clinical variable was found to be significantly associated with AEs, no correction for confounding was attempted. Instead, a simple univariate independence test was used: the empirical p-value adaptive permutation calculation of PLINK [68]. Multiple-correction led to no significantly associated SNP; the lowest corrected p-value was 0.916. Because of this, and the exploratory nature of this study, no multiple-testing correction was applied.

Distribution of p-values is shown as a Manhattan plot in Figure 4. Two main regions seem to be associated with toxicity: 6p21 containing Human Leukocyte Antigen DP (*HLA-DP*) α and β genes and 14q32 containing the Heavy Immunoglobulin Locus (*IGH*). Table 4 shows SNPs with lowest p-values. Table 5 shows results for previously described associated SNP in our dataset; none of them were significantly associated with IIN. It is however notable that among those genes are Small Ubiquitin-like MOdifier 3 (*SUMO3*) and *UGT2A3* for which different isoforms have been described associated with irinotecan response and toxicity (Table 1).

3.3 Machine Learning

Machine learning was used to estimate the predictive performance of SNPs and as a complimentary technique to identify supplementary SNPs. Classifiers used were: logistic regression with elastic net and random forests. These classifiers were selected because they were shown to be suited for GWAS-type studies [77] [26].

Instead of training our classifiers on all the 41,191 quality controlled SNPs, we decided to use biological information to prune out irrelevant SNPs. First of all, since we are using WES, we decided to remove synonymous variants. Moreover, because we believe that most SNPs will not be associated to IIN, we decided to use a functional effect prediction method to prune out SNPs predicted to be without functional consequences. Values under 0.5, as predicted by FATHMM v2.3 [75], were filtered out. After applying these biological filters, a total of 3,918 SNPs remained in the dataset. Standard statistical feature selection methods were also applied. These were Fisher's exact test, mutual information and correlation test. Results did not significantly differ in terms of performance between these features selection methods and only results for Fisher's exact test are shown in this report. Different combinations of methods and their identifiers are shown in Table 6.

3.3.1 Performance

To estimate classifier performance, nested cross-validation was used. Measures used to access performance were: accuracy, Cohen's Kappa, sensitivity, specificity, PPV and NPV. Classifiers trained on permuted IINs vector were also constructed to estimate the null distribution of measures and thus access performance significance of classifiers trained on actual data.

Distribution of accuracy across cross-validation fold for all classifiers is shown in Figure 5. Results for other performance measures are shown in Table 7. Overall, no classification method performed better than expected by chance. Inspection of residuals indicate that lack of performance was due to lack of correct prediction of cases; in other words, classifiers tend to predict everyone as a control (results not shown). This is also reflected in the low specificity and negative predictive value of classification (Table 7).

3.3.2 Feature Importance

Feature importance measures can be used to identify which features are the most important for classification. In this study, we used absolute regression coefficients for the elastic net and mean decrease in Gini's coefficient for random forests. These values were scaled so that the most important feature had a value of 100%. Because all classifiers performed equally, and because we did not wish to use methods that concentrated on main effects of SNPs, no statistical feature selection method was applied. Classifier parameters were set to $mtry = 10$ for random forests, and $\alpha = 0.1$ and $\lambda = 1$ for the elastic net because there were the most selected parameters while using the nested cross-validation resampling scheme. Finally, to estimate the distribution of feature importance, 1,000 iterations of bootstrapping were used and results were compared with classifiers trained on the permuted AEs vector.

Results for random forests are shown in Figure 6 and for elastic net in Figure 7. Although the performance of these classifiers was shown to be not significantly different than random, the distribution of feature importance is clearly nonrandom. Indeed, certain SNPs had significantly higher feature importance values than in the permuted setting, indicating that these SNPs contain information concerning IIN. Many of these identified SNPs were related to genes involved in immunological responses and granulopoiesis. Although features selected by the elastic net and random forests largely overlapped, some SNPs highly important for random forests were never selected for the elastic net. This is the case of rs11867457 of Cytokine Receptor-like Factor 3 (*CRLF3*). Closer analysis showed that only heterozygous genotypes of rs11867457 developed IIN thus showing evidence of a non-linear relationship (results not shown). Finally, there is also a clear correlation between both feature importance values and univariate analysis p-values.

3.4 Gene Set Enrichment Analysis

Machine learning techniques provide us with an important amount of results for which the interpretation is not straightforward. To simplify this, we used over-representation and induced network analysis. Over-representation analysis attempts to identify biological categories that are over-represented in a gene set. On the other hand, induced network analysis attempts to place a gene set in the context of biological networks.

The 10 SNPs with the highest median feature importance from both classifiers were combined, mapped to genes, and resulted in a set of 14 genes. These genes were: *C9orf84*, CREB Regulated Transcription Coactivator 1 (*CRTC1*), Cyclin A1 (*CCNA1*), Cytokine Receptor-like Factor 3 (*CRLF3*), Fibroblast Growth Factor 2 (*FGF2*), Golgin 1 (*GOGB1*), Human Leukocyte Antigen DP (*HLA-DP*) α and β, Heavy Immunoglobulin Locus (*IGH*), Nudix Hydrolase 6 (*NUDT6*), Olfactory Receptor 8J1 (*OR8J1*), Pleckstrin Homology Domain-Containing Family H Member 2 (*PKHH2*), Pseudopodium-Enriched Atypical Kinase 1 (*PEAK1*), von Willebrand factor A domain containing 5B2 (*VWA5B2*). The analysis was also performed separately for classifiers but results were not significantly different (results not shown). Table 9 shows results for over-representation of pathways and Table 10 for GO terms. Figure 8 shows induced network analysis performed on the same gene set.

In genes identified by feature importance in machine learning, immune-related pathways and GO terms were overrepresented. Over-represented pathways were namely infectious, autoimmune diseases and immunological response. Over-represented GO terms related to antigen presentation, interferon γ and the Golgi compartment. In most of these over-represented categories contained Human Leukocyte Antigen DP (*HLA-DP*) chain α and β. Other involved genes were Cyclin A1 (*CCNA1*), CREB Regulated Transcription Coactivator 1 (*CRTC1*), Fibroblast Growth Factor 2 (*FGF2*), Cytokine Receptor-like Factor 3 (*CRLF3*) and Golgin 1 (*GOGB1*). Induced network analysis managed to connect 8 of the 14 genes in the input set. The following genes were not retained in the final network: *C9orf84*, CREB Regulated Transcription Coactivator 1 (*CRTC1*), Cytokine Receptor-like Factor 3 (*CRLF3*), Heavy Immunoglobulin Locus (*IGH*), Olfactory Receptor 8J1 (*OR8J1*), von Willebrand factor A domain containing 5B2 (*VWA5B2*).

4 Discussion

Irinotecan toxicity has mostly been studied through gene candidate studies by concentrating on genes involved in irinotecan pharmacodynamics and pharmacokinetics such as metabolizing enzymes and transporters. In this report, we provide a study of the association between SNPs found in exomes and Irinotecan-Induced Neutropenia (IIN). We first performed a standard univariate association analysis of clinical and genomic variables. None of the clinical variables (age, number of cycles, primary tumor location, previous radiotherapy and previous surgery) were associated with IIN. Concerning genomic variables, distribution of p-values along exomes identifies two main associated regions: 6p21 containing namely Human Leukocyte Antigen DP (*HLA-DP*) chains α and β, and 14q32 containing the Heavy Immunoglobulin Locus (*IGH*). However, after applying multiple-testing correction to p-values, no SNP achieved significance. Moreover, the association for previously described SNPs was not significant. This can be, at least partially, attributed to the small sample size of the study. This being said, the SNP with the lowest p-value was rs2838697 ($-log_{10}p = 4.44$) and is found in the *SUMO3* gene. Small Ubiquitin-like MOdifier (SUMO) proteins are small proteins attached to other proteins in a process called SUMOylation. The functions of this post-translational modification are diverse and range from protein stability to apoptosis. *SUMO1* has already been shown to link to topoisomerase I and consequently increase sensitivity to irinotecan [29]. Furthermore, a SNP found in *UGT2A3* (rs3749510) was also among the lowest p-values ($-log_{10}p = 3.09$). SNPs concerning other isoforms of UGT such as *UGT1A1* were already shown to correlate to IIN. Validation of these two variants may confirm new pharmacogenomics biomarkers for IIN.

In our second analysis, we applied machine learning methods to identify new biomarkers and to estimate their predictive performance. To do so, we used the elastic net and random forests in combination with different resampling schemes. To estimate predictive performance, we used nested cross-validation. To reduce the number of features fed to the classifier, we used functional consequences of variants as predicted by FATHMM. Statistical feature selection methods, such as Fisher's exact independence test were also used. Performance measure distributions between the real and a permuted data were then compared. Observing that both distributions overlapped, we concluded that our classifiers had no more predictive performance than expected by chance. This can again, at least partially, be attributed to the low statistical power of our study. Moreover, nested cross-validation considerably reduces the number of samples available for classifier training only accentuating already existing power problems.

To try and identify new candidate variants, we used the bootstrap distribution of feature importance measures. The bootstrap resampling strategy has the advantage of maintaining the same sample size as the full dataset while still allowing to estimate statistical distributions. Comparing feature importance distributions with those obtained from permuted models clearly showed that the feature selection process of classifiers was

non-random. Selected variants turned out to be very similar to those in the univariate analysis. For example, feature importance of *HLA-DP* and *IGH* was higher than expected by chance. Despite the similarity of results, variant rs11867457 of the Cytokine Receptor-like Factor 3 (*CRLF3*) was only identified by random forests. Linear methods, such as a trend test or logistic regression, could not identify this variant.

Over-representation analysis revealed that genes associated with IIN were often implicated in immunological processes. *HLA-DP* genes were found in almost all of the overrepresented categories. Variants concerning these genes appeared to be strongly associated with IIN throughout all methods. These SNPs were however not described in dbSNP and have no rs identification number. *HLA-DP* proteins are expressed on the membrane of antigen presenting cells such as B lymphocytes and dendrites, but also neutrophils [39]. These proteins are an essential component of adaptive immunity by participating in antigen presentation to T lymphocytes. Certain subpopulations of T lymphocytes have been shown to be implicated in granulopoiesis [49]. Activation of these T lymphocytes by antigen-presentation through *HLA-DP* may result in higher expression of granulopoietic growth factor for certain variants more than for others. Finally, two variants concerning Golgin 1 (*GOGB1*), also known as Giantin, were identified (rs3732410, rs3732407). This gene is expressed in the Golgi apparatus [87] making us think it may be involved in *HLA-DP* synthesis. Unfortunately, the precise function of *GOGB1* is not yet clearly elucidated.

Two variants concerning *IGH* were also associated with IIN (rs200504122, rs112170273). This locus codes for the heavy chain of immunoglobulins. Neutrophils express receptors the constant domain of immunoglobulins, the Fragment Crystallizable (Fc) receptor. Fc receptors have been described to affect neutrophils, namely leading to the release of chemoattractants and favoring phagocytosis [57]. We postulate that recognition of Fc may have consequences related to IIN, for example, by limiting neutrophil apoptosis or favoring granulopoiesis. Moreover, in Cox *et al.* [15], the authors show that G-CSF/Fc fusion proteins have a higher activity than Fc taken alone. They attribute this to the longer half-life of the fusion protein, but maybe the Fc could be playing an active role in granulopoiesis.

Another identified variant concerned Cyclin A1 (*CCNA1*) (rs17188012). Cyclins have an important role in cell proliferation by controlling, with Cyclin-Dependent Kinase (*CDK*), progression through the cell cycle. *CCNA1* binds to *CDK*1 and *CDK*2 giving it regulatory power in the S and the G2 phase of the cell cycle. In Abal *et al.* [2], the authors show that p53-deficient CRC cells are more sensitive to irinotecan than non-deficient cells because of an inhibition of *CDK*1. It is thus possible that certain variants of *CCNA1* lead to a higher sensitivity to irinotecan. Given the non-specificity of irinotecan, this could well be observed in CRC cells, but also neutrophils leading to a higher risk of IIN.

In induced network analysis, *CCNA1* was connected with Retinoblastoma-Associated Protein 1 (*E2F1*), *c-Myc*, the RetinoBlastoma-Like 2 Protein (*RBL2*), Tyrosine-protein

kinase LYN (*LYN*), and a member of the Fibroblast Growth Factor 2 (*FGF2*) signaling network. Surprisingly, all these proteins have a role in granulopoiesis. *E2F1*, *c-Myc* and *RBL2* have been shown to pivotal for granulocytic differentiation mediated by CCAAT Enhancer Binding Protein (*C/EBP*) [25]. *LYN* was previously shown to be a negative regulator of granulopoiesis [52]. Although absent from the induced network, cAMP Responsive Element Binding (CREB) proteins were shown to be involved in granulopoiesis by regulation of *C/EBP* proteins [31].

A SNP found in the *FGF2* gene was also associated with IIN (rs1048201). *FGF2*, of which Nudix Hydrolase 6 (*NUDT6*) lies on the opposite strand, is a known to be potent hematopoietic growth factor under stress conditions [3] [36]. This protein could thus have a considerable role in regenerating neutrophils during irinotecan-based chemotherapy. Moreover, induced network analysis identified a phosphorylation network of various tyrosine kinase, fibroblast growth factor receptors isoforms, and Pseudopodium-Enriched Atypical Kinase 1 (*PEAK1*) of which SNP rs1867780 was associated with IIN. Zhao *et al.* showed that FGF receptor 1 signaling facilitated post-injury recovery of the hematopoietic system by promoting proliferation and mobilization of hematopoietic stem and progenitor cell [90].

Our study suffers from four main limitations. First, our study is clearly underpowered due to its small sample size (40 patients), the large number of SNPs and their small effect size. Results concerning individual SNPs should be taken with caution and not naively considered as being causal of IIN. Indeed, many of the identified variants are probably false positives. This is most likely the case of the olfactory receptor *OR8J1* identified by the elastic net and random forests. Moreover, predictive models were no more performant than expected by chance. Nevertheless, the scope of this study was purely exploratory and its objective was to generate new hypothesis concerning variants associated with IIN. In this respect, our analysis is successful. Another limitation is that our definition of IIN cases and controls is not perfectly rigorous. Indeed, instead of defining cases based solely on blood neutrophil counts, we resorted to the more general leukocyte counts. This was done for two reasons. 1. irinotecan causes non-specific myelosuppression and thus irinotecan-induced leukopenias are most often accompanied with neutropenia; 2. annotations concerning AEs patients experienced were limited. Indeed, reporting of AEs was not part of the clinical trial protocol and it is thus possible that leukopenic and neutropenic patients were only annotated as being leukopenic. For these reasons, we assumed that leukopenic patients were also neutropenic. A third limitation is our use of WES and not WGS or SNP-chips. Doing so, we limited our investigation to SNPs found in exons. However, most of the variants implicated in irinotecan toxicity are introns (Table 1). These variants have been missed by our analysis. Because of this, only six previously described SNPs were in our dataset (Table 5). Finally, it can not be expected that genomics data will manage to explain all of the patient intervariability. For illustration, intestinal microbiota has already been shown to modulate irinotecan toxicity [47].

In this study, the identified associated genes were found to be hematopoietic growth factors or proteins expressed on the cell membrane of neutrophils. Future studies could thus attempt to confirm these associations using a candidate-gene approach and concentrating only on such genes. A list of genes we believe to be worthy of further investigation can be found in Table 11. Pooling data from previous studies could also be attempted in order to increase statistical power. Indeed, most whole exome or genome studies of irinotecan use independent datasets of at most a few hundred samples [51] [28] [78] [27]. Combining data from different sources may, however, not be trivial from a technical point-of-view due to the different datatypes and human populations [76]. That being said, analysis using greater sample size appears as a necessary step to go from pure hypothesis generation to more rigorous statistical analysis. In addition, non-genomic studies will also be necessary to better explain and predict IIN. Investigation of intestinal microbiota through techniques such as metagenomics could provide additional information on the influence of gut bacteria. Furthermore, transcriptomics and proteomics studies would also us to query a different level of biological variation [81]. Functional studies, such as gene knockouts in mice models, concerning newly identified and putative variants may also provide valuable insight into the mechanism of IIN.

5 Conclusion

In this report, we detail the results of an association analysis between exonic SNPs and Irinotecan-Induced Neutropenia (IIN). First, no SNP reached significance in the standard univariate association analysis after multiple-testing correction. Moreover, trained classifiers had no more predictive performance than expected by chance. These results reflect the low power of our study, namely due to small sample size and large feature space. Nonetheless, feature importance distributions of certain SNPs were significantly higher than expected by chance. Many of these identified SNPs were related to genes involved in immunological responses and involved in granulopoiesis, namely Human Leukocyte Antigen DP (*HLA-DP*), Heavy Immunoglobulin Locus (*IGH*), Cyclin A1 (*CCNA1*), Fibroblast Growth Factor 2 (*FGF2*) and CREB Regulated Transcription Coactivator 1 (*CRTC1*). These results are surprising because previously described variants related to irinotecan pharmacokinetics and pharmacodynamics, such as metabolizing enzymes and transporters. This would seem to indicate that, to better understand the physiopathology of IIN, we must also understand the sensitivity of neutrophils to irinotecan and their capacity to regenerate themselves through granulopoiesis in stress conditions. This being said, clinical studies are however necessary to confirm the validity of identified variants.

6 Figures

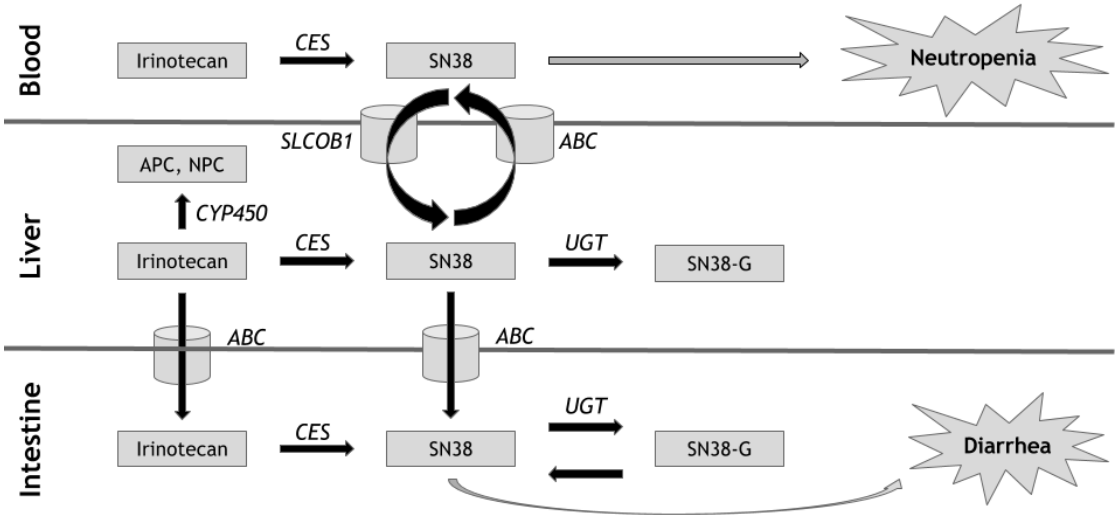

Figure 1: General Pharmacokinetics of Irinotecan

Abbreviations: SN38: 7-ethyl-10-hydroxycamptothecin; SN38-G: glucuronate 7-ethyl-10-hydroxycamptothecin; CES: carboxylesterase; UGT: UDP-glucuronosyltransferase; CYP450: cytochrome P450; ABC: ATP-binding cassette; SLCO: solute carrier organic anion transporter; APC: 7-ethyl-10-[4-N-(5-aminopentanoic acid)-1-piperidino]-carbonyloxycamptothecin; NPC: 7-ethyl-10-(4-amino-1-piperidino)-carbonyloxycamptothecin.

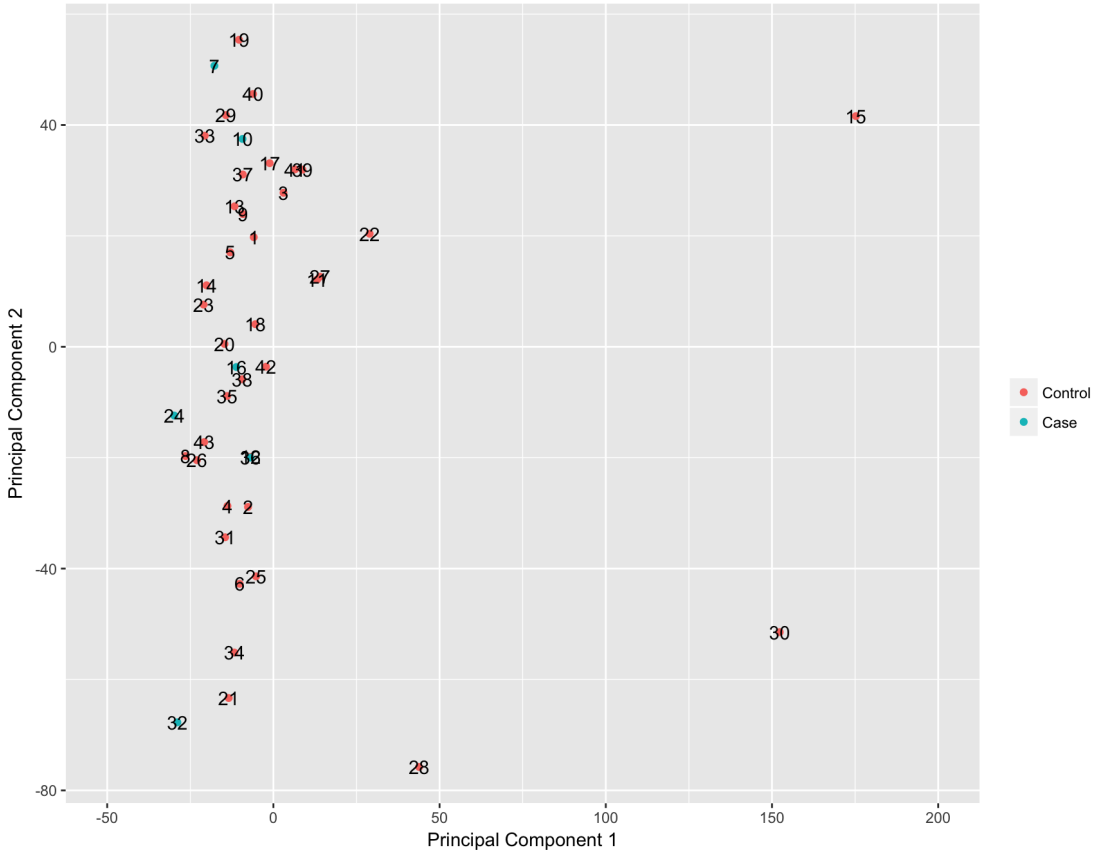

Figure 2: Eigenanalysis of Population Stratification

The first two principal components of 39964 quality controlled SNPs for 43 patients are shown. Samples 15, 28 and 30 were considered as outliers and were removed from further analysis.

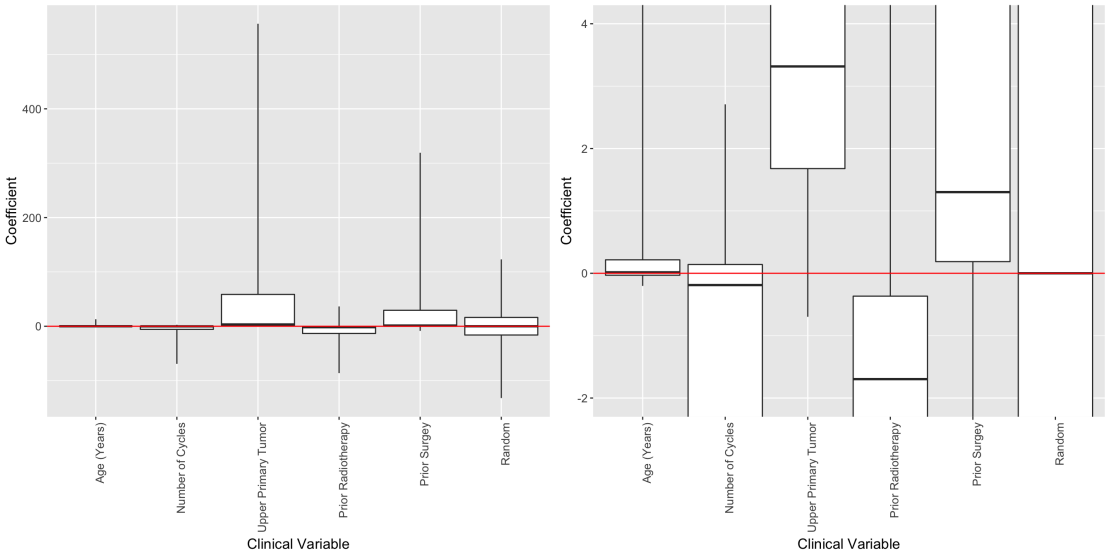

Figure 3: Univariate Association Analysis of Clinical Variables and Irinotecan-Induced Neutropenia

Bootstrap distribution of logistic regression coefficients of clinical variable is shown. Tails represent the 95% and boxes 75% bootstrapped confidence intervals. The left plot shows the full 95% confidence interval and the right plot is zoomed in the $\beta = 0$ axis. *Random* represents a randomly generate binary variable and can be thought of as the null distribution of coefficients in the model. All clinical variables were deemed to be unassociated with Irinotecan-Induced Neutropenia (IIN).

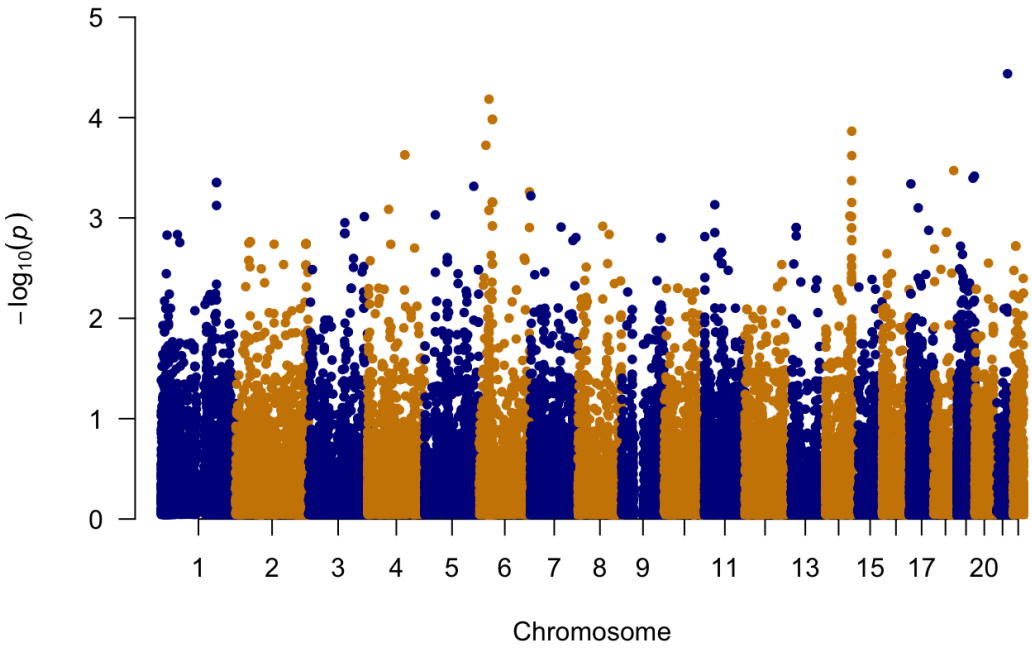

Figure 4: Manhattan Plot

Distribution of p-values from univariate association analysis of SNPs with IIN. P-values were computed using the empirical p-value adaptive permutation calculation of PLINK [68].

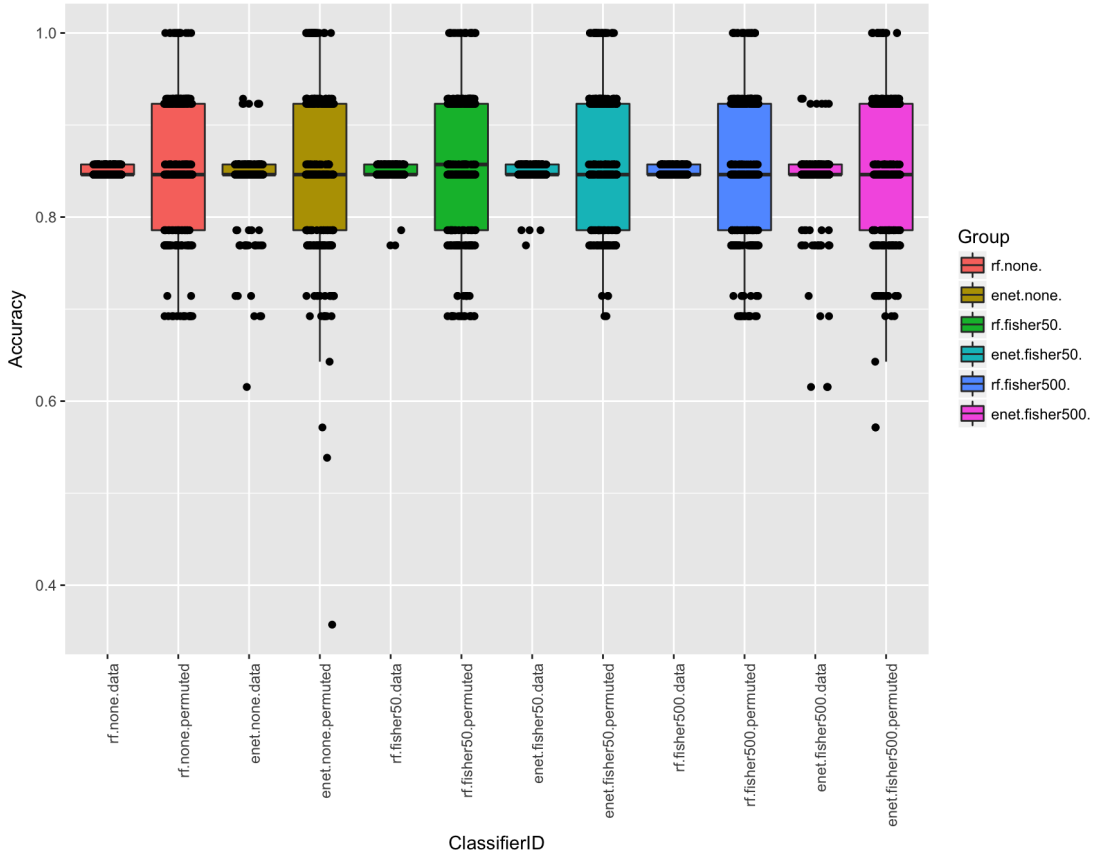

Figure 5: Classification Performance Accuracy

Boxplots represent the distribution of accuracy across cross-validation folds. Meaning of classier identifier can be found in Table 6. Overall, no classifier performed better than expected by chance.

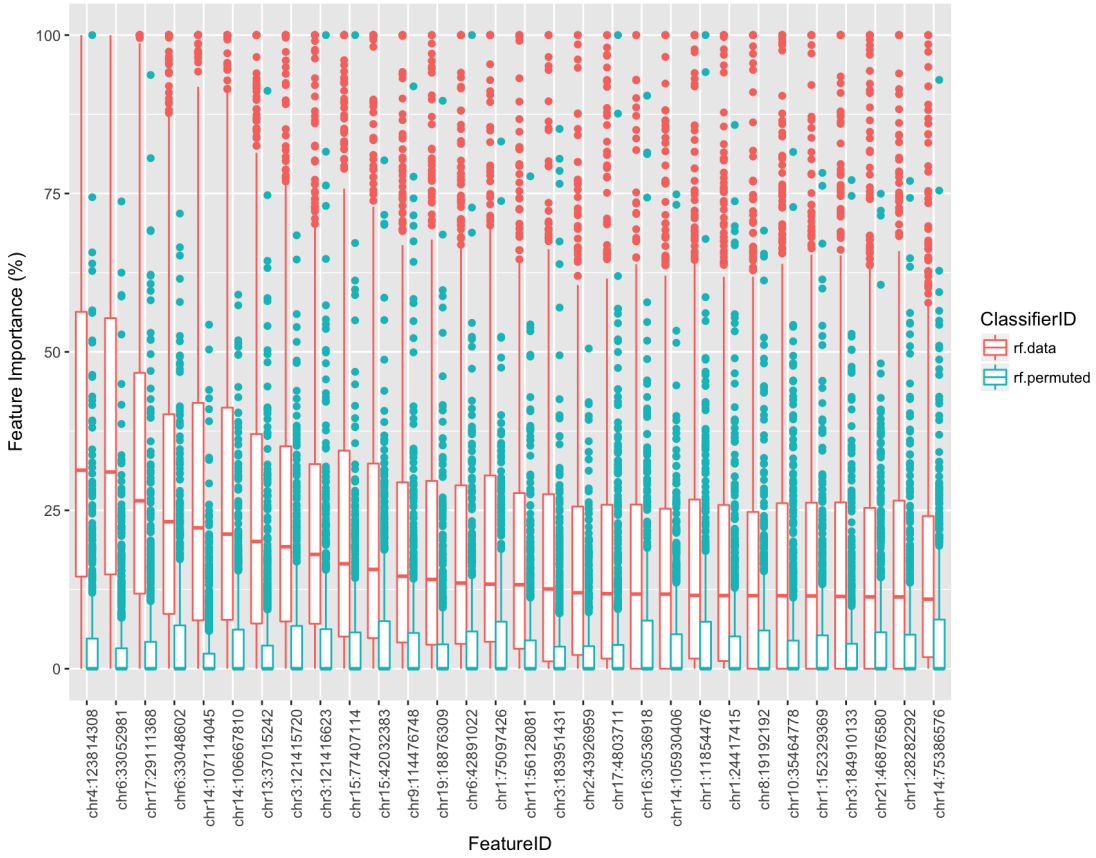

Figure 6: Feature Importance for Random Forests

Distribution of feature importance for random forest across bootstrap samples. Only the features with 30 highest median feature importance are shown. Feature importance is estimated using the mean decrease in Gini's index across individual trees. Values are scaled so that the maximum value is 100%. Genomic positions refer to the GRCh37 genome build.

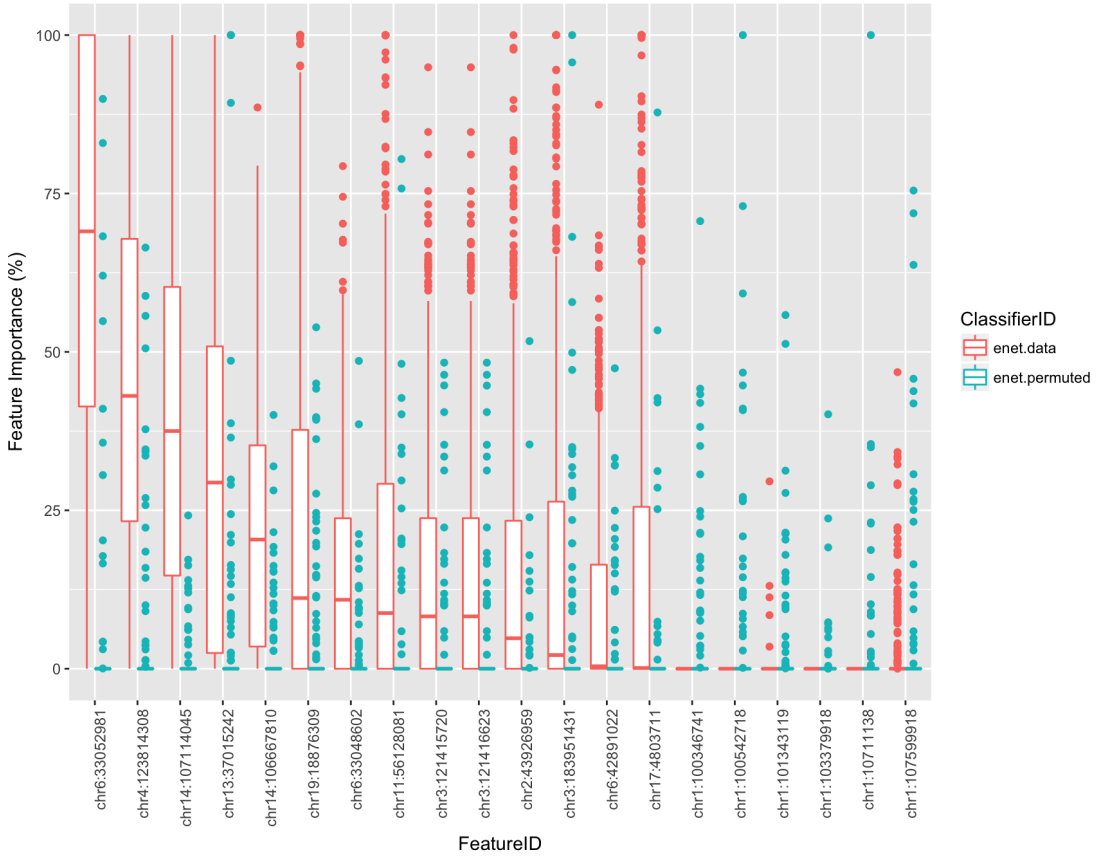

Figure 7: Feature Importance for the Elastic Net

Distribution of feature importance for logistic regression with the elastic net across bootstrap samples. Only the features with 20 highest median feature importance are shown. Feature importance is estimated using absolute regression coefficients. Values are scaled so that the maximum value is 100%. Genomic positions refer to the GRCh37 genome build.

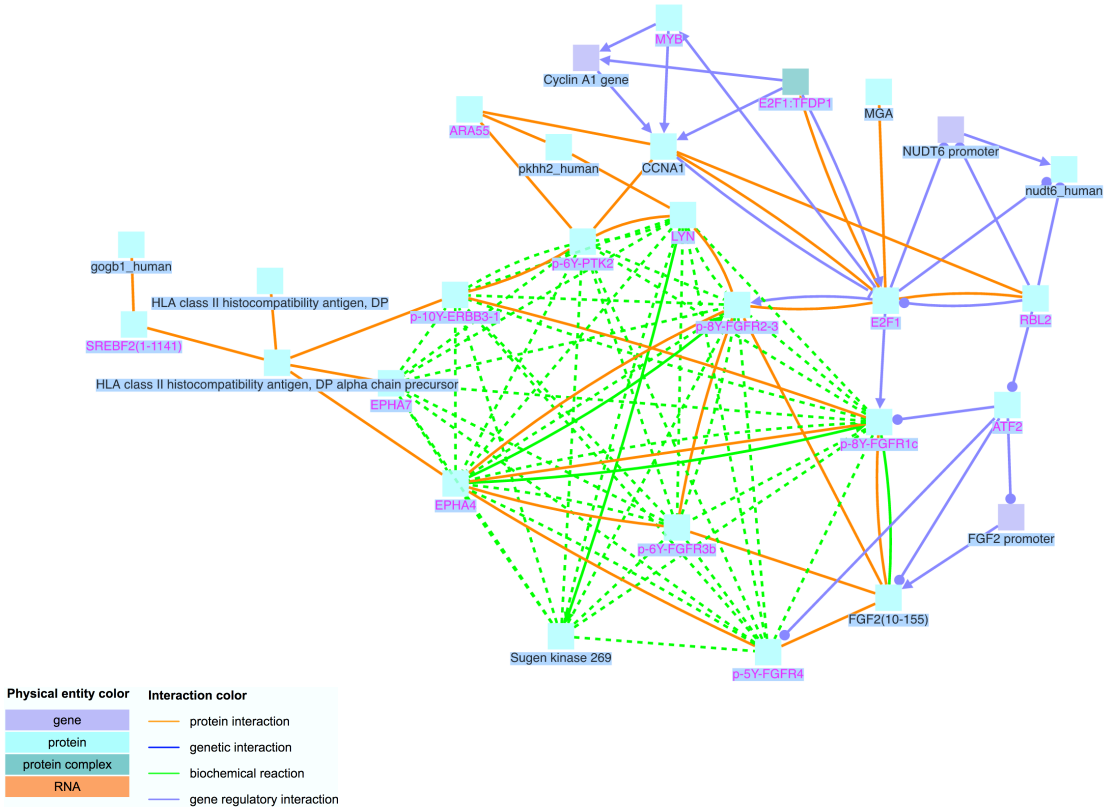

Figure 8: Induced Network Analysis

Induced network of features selected by machine learning methods. The gene set was composed of 14 genes and constructed using 10 genes with highest feature importance values from the elastic net and random forests classifiers. Abbreviations: ARA55: Androgen Receptor Coactivator/Transforming Growth Factor β1; ATF2: Activating Transcription Factor 2; CARD8: Caspase Recruitment Domain-Containing Protein 8; CCNA1: Cyclin A1; E2F1: Retinoblastoma-Associated Protein 1; EPHA4: Ephrin Type-A Receptor 4; EPHA7: Ephrin Type-A Receptor 7; FGF2: Fibroblast Growth Factor 2; GOGB1: Golgin 1; HLA: Human Leukocyte Antigen/Major Histocompatibility Complex; LYN: Tyrosine-protein kinase LYN; MGA: MGA, MAX Dimerization Protein; p-5Y-FGFR4: Phosphorylated Fibroblast Growth Factor Receptor 4; p-6Y-FGFR3b: Phosphorylated Fibroblast Growth Factor Receptor 3b; p-8Y-FGFR1c: Phosphorylated Fibroblast Growth Factor Receptor 2c; p-8Y-FGFR2-3: Phosphorylated Fibroblast Growth Factor Receptor 2b Long; p-10Y-ERBB3-1: Phosphorylated Erb-B2 Receptor Tyrosine Kinase 3 1; NUDT6: Nudix Hydrolase 6; PKHH2: Pleckstrin Homology Domain-Containing Family H Member 2 RBL2: Retinoblastoma-Like Protein 2; SREBF2: Sterol Regulatory Element-Binding Protein 2; Sugen kinase 269: Pseudopodium-Enriched Atypical Kinase 1; TFDP1: Transcription Factor Dp-1.

7 Tables

Gene	Entrez	Chromosome	Position	Variant	dbSNP	Functional Class	Variation Type	Role	Effect	Reference
ABCB1	5243	7	87138645	3435C>T	rs1045642	Exon	SNP	PK	Toxicity	[51]
ABCB1	5243	7	87160618	2677G>T	rs2032582	Exon	SNP	PK	Toxicity	[51]
ABCB1	5243	7	87179601	1236C>T	rs1128503	Exon	SNP	PK	Toxicity	[51]
ABCC2	1244	10	101604207	3972C>T	rs3740066	Exon	SNP	PK	Toxicity	[51]
ABCC4	10257	13	95775432		rs16950650	Exon	SNP		Sensitivity	[28]
ABCG2	9429	4	89061114	34G>A	rs2231137	Exon	SNP	PK	Toxicity	[51]
C8orf34	116328	8	70464006		rs1517114	Intron	SNP		Toxicity	[27]
CES2	8824	16	66969176	830C>G	rs11075646	UTR-5	SNP	PK	Insignificant	[51]
CYP3A4	1576	7	99361626	*18	rs28371759	Exon	SNP	PK	Insignificant	[51]
CYP3A4	1576	7	99768470	*16	rs12721627	Exon	SNP	PK	Insignificant	[51]
CYP3A5	1577	7	99270539	*3	rs776746	Intron	SNP	PK	Insignificant	[51]
DCBLD1	285761	6	117816128		rs17574269	Exon	SNP		Sensitivity	[28]
FLJ41856	388550	19	50247165		rs1661167	Intron	SNP		Toxicity	[27]
KCNQ5	56479	6	73749611		rs9351963	Intron	SNP		Toxicity	[78]
LOC101928196		2	19617302		rs2166219		SNP		Sensitivity	[28]
PDZRN3	23024	3	73751402		rs11128347	Intron	SNP		Toxicity	[27]
PLCB1	23236	20	8206590		rs2745761	Intron	SNP		Toxicity	[27]
SEMA3C	10512	7	79976508		rs11979430	Intron	SNP		Toxicity	[27]
SEMA3C	10512	7	79999364		rs7779029	Intron	SNP		Toxicity	[27]
SLCO1B1	10599	12	21222816		rs4149056	Exon	SNP	PK	Insignificant	[27]
SUMO1	7341	4	203079307	150G>A	rs3769817	Intron	SNP	PD	Sensitivity	[29]
SYNE3	161176	14	95943548		rs8020368	Promoter	SNP		Sensitivity	[28]
TDP1	55775	14	90456188	79T>G	rs2401863	Intron	SNP	PD	Sensitivity	[51]
UGT1A1	7361	2	233760235:233760236	*28	rs8175347	Intron	Microsatellite	PK	Toxicity	[51]
UGT1A1	7361	2	234669144	*6	rs4148323	Intron	SNP	PK	Toxicity	[51]
UGT1A7	54576	2	234590970	*2	rs17868323	Intron	SNP	PK	Toxicity	[51]
UGT1A7	54576	2	234590970	*3	rs11692021	Intron	SNP	PK	Toxicity	[51]
UGT1A9	54600	2	234580454	*22	rs3832043	Intron	Insertion	PK	Toxicity	[51]
XRCC1	7515	6	44055726	R399Q	rs25487	Exon	SNP	PD	Sensitivity	[51]
		1	68101640		rs4655567		SNP		Sensitivity	[28]
		7	29014195		rs2018683	Promoter	SNP		Sensitivity	[28]
		16	16957915		rs7186128		SNP		Sensitivity	[28]

Table 1: Previously Described Genetic Associations for Irinotecan Response and Toxicity
Genomic positions refer to the GRCh37 genome build. Abbreviations: SNP: Single Nucleotide Polymorphism; PK: Pharmacokinetics; PD: Pharmacodynamics.

Variable	**Levels**	**n**	%	∑ %
Response	Control	34	85.0	85.0
	Case	6	15.0	100.0
	all	40	100.0	
Sex	Female	14	35.0	35.0
	Male	26	65.0	100.0
	all	40	100.0	
Primary Tumor	Appendix	1	2.5	2.5
	Bile Duct	3	7.5	10.0
	Colo-Rectal	17	42.5	52.5
	Duodenum	1	2.5	55.0
	Esophagus	8	20.0	75.0
	Pancreas	2	5.0	80.0
	Papil Vater	1	2.5	82.5
	Stomach	4	10.0	92.5
	Unknown Primary	3	7.5	100.0
	all	40	100.0	
Prior Radiotherapy	No	28	70.0	70.0
	Yes	12	30.0	100.0
	all	40	100.0	
Prior Surgery	No	16	40.0	40.0
	Yes	24	60.0	100.0
	all	40	100.0	

Table 2: Patient Characteristics for Categorical Variables

Variable	n	Min	Q1	Median	Mean	Q3	Max	IQR	SD
Age (Years)	40	41	57.8	63.5	63.7	68.2	101	10.5	10.2
Total Number Of Cycles	40	1	2.0	2.0	3.7	5.2	12	3.2	2.6

Table 3: Patient Characteristics for Continuous Variables
Number of cycles refer to number of irinotecan cycles.

Chromosome	Position	dbSNP	Entrez	Name	FATHMM	-LOGP
21	46233836	rs2838697	6612	small ubiquitin-like modifier 3	0.06	4.44
6	33052981		3115	human leukocyte antigen DP beta 1	0.88	4.18
6	44269193	rs498512	221409	spermatogenesis associated serine rich 1	0.26	3.98
14	107114055	rs199709247	3492	immunoglobulin heavy locus	0.00	3.86
6	22570245	rs6900627	154150	hepatoma derived growth factor-like 1	0.74	3.72
4	123814308	rs1048201	2247	fibroblast growth factor 2	0.98	3.63
4	123814308	rs1048201	11162	nudix hydrolase 6	0.98	3.63
14	107114045	rs200504122	3492	immunoglobulin heavy locus	0.83	3.62
18	66504351	rs637051	79839	coiled-coil domain containing 102B	0.83	3.47
19	57649900	rs2370134	114026	zinc finger imprinted 3	0.01	3.42
19	52888522		400713	zinc finger protein 880	0.01	3.40
14	106667810	rs112170273	3492	immunoglobulin heavy locus	0.73	3.37
1	185143721	rs10489579	54823	SWT1, RNA endoribonuclease homolog	0.18	3.35
1	185171869	rs6698109	54823	SWT1, RNA endoribonuclease homolog	0.38	3.35
17	4722785	rs1052751	5338	phospholipase D2	0.75	3.34
5	162890953	rs380101	3161	hyaluronan mediated motility receptor	0.01	3.32
6	167789493	rs3010590	6953	t-complex 10	0.00	3.26
7	1590443	rs2304361	202915	transmembrane protein 184A	0.03	3.22
6	44250165	rs516582	202500	t-complex-associated-testis-expressed 1	0.02	3.16
6	44255459	rs324146	202500	t-complex-associated-testis-expressed 1	0.05	3.16
14	107049122				0.01	3.15
11	33631423	rs2076622	25758	KIAA1549 like	0.06	3.13
1	185135745	rs950327	54823	SWT1, RNA endoribonuclease homolog	0.06	3.12
17	29226630	rs28539246	79736	transcription elongation factor, mitochondrial	0.01	3.10
4	69817080	rs3749510	79799	UDP glucuronosyltransferase family 2 member A3	0.02	3.09
6	33048602		3113	human leukocyte antigen DP alpha 1	0.60	3.08
5	34811137	rs2271522	26064	retinoic acid induced 14	0.43	3.03
14	101200645	rs1802710	3492	immunoglobulin heavy locus	0.12	3.02
3	186435370	rs1050274	3827	kininogen 1	0.80	3.01
14	106641761	rs72695948	3492	immunoglobulin heavy locus	0.01	3.01
3	121400725	rs1463736	2804	golgin B1	0.32	2.95
6	44280812	rs167772	221409	spermatogenesis associated serine rich 1	0.09	2.92
8	82355621	rs6473276	5375	peripheral myelin protein 2	0.00	2.92
7	102412901	rs2539288	222234	family with sequence similarity 185 member A	0.33	2.91
6	167754661	rs909545	83887	tubulin tyrosine ligase like 2	0.02	2.90
13	36939758	rs3736920	100507135	SPG20 antisense RNA 1	0.00	2.90
14	106573352	rs201642285	3492	immunoglobulin heavy locus	0.02	2.90
14	106573354	rs200148307	3492	immunoglobulin heavy locus	0.01	2.90
17	64210757	rs4581	350	apolipoprotein H	0.05	2.88
18	42532693	rs1064204	26040	SET binding protein 1	0.86	2.86
18	42533315	rs3786177	26040	SET binding protein 1	0.12	2.86
11	33680371	rs12280103	25758	KIAA1549 like	0.12	2.85
3	121354583	rs3772126	3059	hematopoietic cell-specific Lyn substrate 1	0.19	2.85
3	121414061	rs1127412	2804	golgin B1	0.01	2.85
3	121415720	rs3732410	2804	golgin B1	0.94	2.85
3	121416623	rs3732407	2804	golgin B1	0.98	2.85
8	103851052	rs1062048	51582	antizyme inhibitor 1	0.90	2.84
1	54686374	rs612711	51253	mitochondrial ribosomal protein L37	0.01	2.83
1	20216860	rs10799593	23252	OTU deubiquitinase 3	0.01	2.83
13	37015242	rs17188012	8900	cyclin A1	0.70	2.82
11	244141	rs7128029	5719	proteasome 26S subunit, non-ATPase 13	0.02	2.81

Table 4: Univariate Association Analysis of Genomic Variables

The 50 lowest p-values from association analysis of SNPs of IIN are shown. $FATHMM$ refers to predictions of functional consequence obtained from FATHMM [75]. $-LOGP$ is the $-log_{10}$ transformed p-value computed using the empirical p-value adaptive permutation calculation of PLINK [68]. Genomic positions refer to the GRCh37 genome build.

Chromosome	Position	dbSNP	Entrez	Gene	FATHMM	-LOGP
2	234590970	rs17868323	54576	UGT1A7	0.00	0.74
7	87160618	rs2032582	5243	ABCB1	0.32	0.35
7	87138645	rs1045642	5243	ABCB1	0.03	0.33
7	87179601	rs1128503	5243	ABCB1	0.03	0.07
10	101604207	rs3740066	1244	ABCC2	0.08	0.04

Table 5: Univariate Association Analysis for Previously Described Associations

$FATHMM$ refers to predictions of functional consequence obtained from FATHMM [75]. $-LOGP$ is the $-log_{10}$ transformed p-value computed using the empirical p-value adaptive permutation calculation of PLINK [68]. Genomic positions refer to the GRCh37 genome build.

Identifier	Classifier	Feature Selection	Permuted
enet.fisher50.data	Elastic Net	50 Lowest P-value Fisher's Test	No
enet.fisher50.permuted	Elastic Net	50 Lowest P-value Fisher's Test	Yes
enet.fisher500.	Elastic Net	500 Lowest P-value Fisher's Test	No
enet.fisher500.permuted	Elastic Net	500 Lowest P-value Fisher's Test	Yes
enet.none.data	Elastic Net	None	No
enet.none.permuted	Elastic Net	None	Yes
rf.fisher50.data	Random Forests	50 Lowest P-value Fisher's Test	No
rf.fisher50.permuted	Random Forests	50 Lowest P-value Fisher's Test	Yes
rf.fisher500.data	Random Forests	500 Lowest P-value Fisher's Test	No
rf.fisher500.permuted	Random Forests	500 Lowest P-value Fisher's Test	Yes
rf.none.data	Random Forests	None	No
rf.none.permuted	Random Forests	None	Yes

Table 6: Classification Methods Identifiers
This tables aims at clarifying the meaning of classifier identifiers.

Variable	Levels	n	Min	q_1	$\bar{x}$	$\tilde{x}$	q_3	Max	s	IQR	#NA
Accuracy	enet.fisher50.data	60	0.77			0.85		0.86	0.01		
	enet.fisher50.permuted	58	0.69	0.79		0.85	0.92		0.07	0.14	
	enet.fisher500.data	60	0.69					0.92	0.03		
	enet.fisher500.permuted	55	0.50	0.77		0.83	0.92		0.10	0.15	
	enet.none.data	60	0.77			0.85		0.93	0.02		
	enet.none.permuted	57	0.62	0.79		0.83			0.08	0.07	
	rf.fisher50.data	60	0.79			0.85		0.86	0.01		
	rf.fisher50.permuted	60	0.69	0.77			0.92		0.08	0.15	
	rf.fisher500.data	60	0.85			0.85		0.86	0.01		
	rf.fisher500.permuted	60	0.69	0.77		0.83			0.07	0.09	
	rf.none.data	60	0.85			0.85		0.86	0.01		
	rf.none.permuted	60	0.69			0.86	0.92		0.07	0.08	
	all	710	0.50	0.85	0.85	0.84	0.86	1.00	0.06	0.01	0
Kappa	enet.fisher50.data	60	-0.11					0.00	0.01		0
	enet.fisher50.permuted	54	-0.08					0.00	0.01		4
	enet.fisher500.data	60	-0.18			0.03			0.16		0
	enet.fisher500.permuted	50				-0.01		0.00	0.05		5
	enet.none.data	60	-0.11			0.03			0.15		0
	enet.none.permuted	56	-0.14			-0.01		0.13	0.04		1
	rf.fisher50.data	60	-0.11					0.42	0.06		0
	rf.fisher50.permuted	58	-0.11						0.01		2
	rf.fisher500.data	60						0.00	0.00		0
	rf.fisher500.permuted	58						0.00	0.00		2
	rf.none.data	60						0.00	0.00		0
	rf.none.permuted	55						0.00	0.00		5
	all	691	-0.26	0.00	0.00	0.00	0.00	0.75	0.07	0.00	19
Sensitivity	enet.fisher50.data	60	0.91			1.00			0.01		
	enet.fisher50.permuted	58	0.85			1.00			0.02		
	enet.fisher500.data	60	0.82								
	enet.fisher500.permuted	55				0.97			0.10		
	enet.none.data	60	0.83						0.03		
	enet.none.permuted	57	0.69			0.98			0.05		
	rf.fisher50.data	60	0.92			1.00			0.02		
	rf.fisher50.permuted	60	0.92			1.00			0.01		
	rf.fisher500.data	60	1.00			1.00			0.00		
	rf.fisher500.permuted	60	1.00			1.00			0.00		
	rf.none.data	60	1.00			1.00			0.00		
	rf.none.permuted	60	1.00			1.00			0.00		
	all	710	0.57	1.00	1.00	0.99	1.00	1.00	0.04	0.00	0
Specificity	enet.fisher50.data	60				0.00		0.00	0.00		0
	enet.fisher50.permuted	53				0.00		0.00	0.00		5
	enet.fisher500.data	60				0.04			0.17		0
	enet.fisher500.permuted	47				0.00		0.00	0.00		8
	enet.none.data	60				0.04			0.19		0
	enet.none.permuted	56						0.33	0.04		1
	rf.fisher50.data	60						0.50	0.06		0
	rf.fisher50.permuted	58				0.00		0.00	0.00		2
	rf.fisher500.data	60				0.00		0.00	0.00		0
	rf.fisher500.permuted	58				0.00		0.00	0.00		2
	rf.none.data	60				0.00		0.00	0.00		0
	rf.none.permuted	55				0.00		0.00	0.00		5
	all	687	0.00	0.00	0.00	0.01	0.00	1.00	0.08	0.00	23
PPV	enet.fisher50.data	60	0.83			0.85		0.86	0.01		0
	enet.fisher50.permuted	53	0.69	0.79		0.85	0.92	0.93	0.06	0.13	5
	enet.fisher500.data	60	0.82			0.85			0.03		0
	enet.fisher500.permuted	47	0.69	0.77				0.93	0.06	0.08	8
	enet.none.data	60	0.83			0.86			0.03		0
	enet.none.permuted	56		0.79			0.92	0.93	0.07	0.13	1
	rf.fisher50.data	60	0.85			0.85		0.92	0.01		0
	rf.fisher50.permuted	58	0.69	0.77		0.83	0.91	0.93	0.07	0.14	2
	rf.fisher500.data	60	0.85			0.85		0.86	0.01		0
	rf.fisher500.permuted	58	0.69	0.77		0.82		0.93	0.07	0.09	2
	rf.none.data	60	0.85			0.85		0.86	0.01		0
	rf.none.permuted	55	0.69	0.82		0.85		0.93		0.04	5
	all	687	0.67	0.85	0.85	0.84	0.86	1.00	0.05	0.01	23
NPV	enet.fisher50.data	1				0.00	0.00	0.00	NA	0.00	59
	enet.fisher50.permuted	1				0.00	0.00	0.00	NA	0.00	57
	enet.fisher500.data	10				0.30	0.58		0.43	0.58	50
	enet.fisher500.permuted	5				0.00	0.00	0.00	0.00	0.00	50
	enet.none.data	5				0.43	0.67		0.43	0.67	55
	enet.none.permuted	6				0.06	0.00	0.33	0.14	0.00	51
	rf.fisher50.data	3				0.17	0.25	0.50	0.29	0.25	57
	rf.fisher50.permuted	1				0.00	0.00	0.00	NA	0.00	59
	rf.fisher500.data	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	rf.fisher500.permuted	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	rf.none.data	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	rf.none.permuted	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	all	32	0.00	0.00	0.00	0.19	0.33	1.00	0.34	0.33	678

Table 7: Predictive Performance of Classification Methods

Meaning of classifier identifiers is found in Table 6. *all* refers to the sum for n and $\#NA$ and the mean for other statistics. Unspecified values indicate that value was identical to mean of all methods. NA stands for Not Available and corresponds to values that were not computable (*e.g.* division by zero error).

Chromosome	Position	dbSNP	Entrez	Name	FATHMM	-LOGP	ENET	RF
6	33052981		3115	human leukocyte antigen DP beta 1	0.88019	4.183	69.029	31.038
4	123814308	rs1048201	11162	nudix hydrolase 6	0.98282	3.628	43.056	31.326
4	123814308	rs1048201	2247	fibroblast growth factor 2	0.98282	3.628	43.056	31.326
14	107114045	rs200504122	3492	immunoglobulin heavy locus	0.82741	3.621	37.512	22.238
13	37015242	rs17188012	8900	cyclin A1	0.69504	2.820	29.380	20.085
14	106667810	rs112170273	3492	immunoglobulin heavy locus	0.72884	3.371	20.386	21.242
19	18876309	rs3746266	23373	CREB regulated transcription coactivator 1	0.93362	2.359	11.149	14.085
6	33048602		3113	human leukocyte antigen alpha 1	0.60367	3.076	10.882	23.206
11	56128081	rs10896290	219477	olfactory receptor family 8 subfamily J member 1	0.97765	2.546	8.777	13.260
3	121415720	rs3732410	2804	golgin B1	0.94387	2.846	8.259	19.245
3	121416623	rs3732407	2804	golgin B1	0.97682	2.846	8.256	18.034
2	43926959		130271	pleckstrin homology, MyTH4 and FERM domain containing H2	0.9608	2.748	4.809	11.985
3	183951431	rs902417	90113	von Willebrand factor A domain containing 5B2	0.69968	2.262	2.157	12.588
6	42891022	rs9471966	171558	pre T-cell antigen receptor alpha	0.84672	2.189	0.389	13.525
17	4803711	rs35400274	1145	cholinergic receptor nicotinic epsilon subunit	0.81605	2.244	0.116	11.831
17	29111368	rs11867457	51379	cytokine receptor like factor 3	0.91744	2.400	0.000	26.507
15	77407114	rs1867780	79834	pseudopodium enriched atypical kinase 1	0.93206	1.166	0.000	16.563
15	42032383	rs17677991	23269	MGA, MAX dimerization protein	0.95177	0.718	0.000	15.675
9	114476748	rs11791445	158401	chromosome 9 open reading frame 84	0.81088	1.784	0.000	14.592
1	75097426	rs11210490	127254	glutamate rich 3	0.87429	1.560	0.000	13.341
16	30536918	rs10871453	79724	zinc finger protein 768	0.83551	2.445	0.000	11.783
14	105930406	rs4983413	9112	metastasis associated 1	0.8149	0.423	0.000	11.772
1	11854476	rs1801131	4524	methylenetetrahydrofolate reductase	0.9606	1.386	0.000	11.570
1	24417415	rs6700245	127294	myomesin 3	0.68077	2.022	0.000	11.546
8	19192192	rs17128221	63898	SH2 domain containing 4A	0.69264	1.017	0.000	11.533
10	35464778	rs1531550	1390	cAMP responsive element modulator	0.71793	1.704	0.000	11.521
1	152329369	rs2282302	388698	filaggrin family member 2	0.73011	1.865	0.000	11.486
3	184910133	rs11919970	1962	enoyl-CoA hydratase and 3-hydroxyacyl CoA dehydrogenase	0.83735	1.696	0.000	11.398
21	46876580	rs8133886	80781	collagen type XVIII alpha 1 chain	0.56493	2.063	0.000	11.339
1	28282292	rs3813803	27293	sphingomyelin phosphodiesterase acid like 3B	0.97513	1.088	0.000	11.332
14	75386576	rs2286913	83694	ribosomal protein S6 kinase like 1	0.94193	0.323	0.000	10.972
19	1079959	rs36084354	23526	Rho GTPase activating protein 45	0.86807	2.002	0.000	10.812
14	88407888	rs398607	2581	galactosylceramidase	0.98076	0.310	0.000	10.773
6	43184132	rs2273709	23113	cullin 9	0.56274	2.265	0.000	10.623
1	185240474	rs12041704	54823	SWT1, RNA endoribonuclease homolog	0.86341	2.183	0.000	10.289
3	126261207	rs1056523	166012	carbohydrate sulfotransferase 13	0.81958	1.628	0.000	10.235
19	48954421	rs2302951	83743	glutamate rich WD repeat containing 1	0.64103	2.007	0.000	10.186

Table 8: Feature Importance in Classification Methods

$FATHMM$ refers to predictions of functional consequence obtained from FATHMM [75]. $-LOGP$ is the $-log_{10}$ transformed p-value computed using the empirical p-value adaptive permutation calculation of PLINK [68]. $ENET$ and RF refers to median feature importance measures across 1000 bootstrap folds: $ENET$ is the scaled absolute regression coefficients for the elastic net classifier and RF is the scaled mean decrease in Gini Index for the random forest classifier. Genomic positions refer to the GRCh37 genome build.

Pathway	P-Value	Q-Value	Source
Epstein-Barr virus infection - Homo sapiens (human)	0.00013	0.00154	KEGG
HTLV-I infection - Homo sapiens (human)	0.00021	0.00154	KEGG
Translocation of ZAP-70 to Immunological synapse	0.00027	0.00154	Reactome
Phosphorylation of CD3 and TCR zeta chains	0.00027	0.00154	Reactome
PD-1 signaling	0.00027	0.00154	Reactome
Asthma - Homo sapiens (human)	0.00027	0.00154	KEGG
Generation of second messenger molecules	0.00035	0.00154	Reactome
Autoimmune thyroid disease - Homo sapiens (human)	0.00044	0.00154	KEGG
CD4 T cell receptor signaling-JNK cascade	0.00044	0.00154	INOH
Allograft rejection - Homo sapiens (human)	0.00053	0.00154	KEGG
Graft-versus-host disease - Homo sapiens (human)	0.00053	0.00154	KEGG
Intestinal immune network for IgA production - Homo sapiens (human)	0.00053	0.00154	KEGG
CD4 T cell receptor signaling-ERK cascade	0.00064	0.00154	INOH
Type I diabetes mellitus - Homo sapiens (human)	0.00064	0.00154	KEGG
Downstream TCR signaling	0.00064	0.00154	Reactome
MHC class II antigen presentation	0.00075	0.00154	Reactome
Costimulation by the CD28 family	0.00075	0.00154	Reactome
Viral myocarditis - Homo sapiens (human)	0.00075	0.00154	KEGG
TCR signaling	0.00075	0.00154	Reactome
Staphylococcus aureus infection - Homo sapiens (human)	0.00088	0.00155	KEGG
Inflammatory bowel disease (IBD) - Homo sapiens (human)	0.00088	0.00155	KEGG
Systemic lupus erythematosus - Homo sapiens (human)	0.00088	0.00155	KEGG
Leishmaniasis - Homo sapiens (human)	0.00101	0.00171	KEGG
Antigen processing and presentation - Homo sapiens (human)	0.00115	0.00180	KEGG
CD4 T cell receptor signaling-NFkB cascade	0.00115	0.00180	INOH
Allograft Rejection	0.00146	0.00204	Wikipathways
Rheumatoid arthritis - Homo sapiens (human)	0.00146	0.00204	KEGG
CD4 T cell receptor signaling	0.00146	0.00204	INOH
Toxoplasmosis - Homo sapiens (human)	0.00220	0.00286	KEGG
Adaptive Immune System	0.00220	0.00286	Reactome
Tuberculosis - Homo sapiens (human)	0.00262	0.00329	KEGG
Phagosome - Homo sapiens (human)	0.00382	0.00465	KEGG
Herpes simplex infection - Homo sapiens (human)	0.00408	0.00483	KEGG
Cell adhesion molecules (CAMs) - Homo sapiens (human)	0.00436	0.00500	KEGG
Influenza A - Homo sapiens (human)	0.00464	0.00517	KEGG

Table 9: Pathway Over-representation Analysis

The input gene set was composed of 14 genes and constructed using 10 genes with highest feature importance values from the elastic net and random forests. Q-values corresponds to BH multiple-testing corrected p-values.

GO Term	GO ID	P-Value	Q-Value
MHC class II protein complex	GO:0042613	0.00031	0.00251
trans-Golgi network membrane	GO:0032588	0.00097	0.00390
peptide antigen binding	GO:0042605	0.00081	0.00488
Golgi subcompartment	GO:0098791	0.00080	0.00596
MHC protein complex	GO:0042611	0.00081	0.00596
clathrin-coated endocytic vesicle membrane	GO:0030669	0.00081	0.00596
growth factor activity	GO:0008083	0.00304	0.00607
ER to Golgi transport vesicle membrane	GO:0012507	0.00198	0.00980
clathrin-coated vesicle membrane	GO:0030665	0.00223	0.00980
organelle subcompartment	GO:0031984	0.00085	0.01169
lumenal side of endoplasmic reticulum membrane	GO:0098553	0.00097	0.01169
trans-Golgi network	GO:0005802	0.00501	0.01335
transport vesicle	GO:0030133	0.00788	0.01575
integral component of endoplasmic reticulum membrane	GO:0030176	0.00501	0.01812
intrinsic component of endoplasmic reticulum membrane	GO:0031227	0.00576	0.01812
antigen binding	GO:0003823	0.00223	0.02004
lumenal side of membrane	GO:0098576	0.00097	0.02337
transport vesicle membrane	GO:0030658	0.00364	0.02378
endocytic vesicle membrane	GO:0030666	0.00396	0.02378
regulation of interferon-gamma production	GO:0032649	0.00067	0.02668
peptide binding	GO:0042277	0.00929	0.02786
coated vesicle membrane	GO:0030662	0.00616	0.02958
antigen processing and presentation of exogenous peptide antigen via MHC class II	GO:0019886	0.00248	0.03641
antigen receptor-mediated signaling pathway	GO:0050851	0.00364	0.03641
response to interferon-gamma	GO:0034341	0.00364	0.03641
lymphocyte costimulation	GO:0031294	0.00134	0.03876
regulation of leukocyte proliferation	GO:0070663	0.00657	0.04528
immune response-regulating signaling pathway	GO:0002764	0.00679	0.04528
interferon-gamma production	GO:0032609	0.00081	0.04961

Table 10: Gene Ontology Terms Over-representation Analysis

The input gene set was composed of 14 genes and constructed using 10 genes with highest feature importance values from the elastic net and random forests. Q-values corresponds to BH multiple-testing corrected p-values.

Entrez	Name
1869	retinoblastoma-associated protein 1
2247	fibroblast growth factor 2
3115	human leukocyte antigen DP beta 1
3492	immunoglobulin heavy locus
4067	LYN
4609	MYC
5934	Retinoblastoma-like protein 2
6612	small ubiquitin-like modifier 3
8900	cyclin A1
11162	nudix hydrolase 6
23373	CREB regulated transcription coactivator 1
51379	cytokine receptor like factor 3
79799	UDP glucuronosyltransferase family 2 member A3
79834	pseudopodium enriched atypical kinase 1

Table 11: List of Genes on which to Focus Further Analysis.

Sort order is simply based on Entrez identifier and not on importance.

8 References

[1] "A global reference for human genetic variation". In: *Nature* 526.7571 (Oct. 1, 2015), pp. 68–74. ISSN: 0028-0836. DOI: 10.1038/nature15393. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4750478/ (visited on 12/06/2016).

[2] Miguel Abal et al. "Enhanced sensitivity to irinotecan by Cdk1 inhibition in the p53-deficient HT29 human colon cancer cell line". In: *Oncogene* 23.9 (2004), pp. 1737–1744. ISSN: 0950-9232. DOI: 10.1038/sj.onc.1207299. URL: http://www.nature.com/onc/journal/v23/n9/full/1207299a.html (visited on 11/28/2016).

[3] M. Allouche and A. Bikfalvi. "The role of fibroblast growth factor-2 (FGF-2) in hematopoiesis". In: *Progress in Growth Factor Research* 6.1 (1995), pp. 35–48. ISSN: 0955-2235.

[4] Nerea Alonso, Gavin Lucas, and Pirro Hysi. "Big data challenges in bone research: genome-wide association studies and next-generation sequencing". In: *BoneKEy Reports* 4 (Feb. 11, 2015), p. 635. DOI: 10.1038/bonekey.2015.2. URL: http://www.nature.com/bonekeyreports/2015/150211/bonekey20152/full/bonekey20152.html (visited on 09/09/2016).

[5] Michael Ashburner et al. "Gene Ontology: tool for the unification of biology". In: *Nature Genetics* 25.1 (May 2000), pp. 25–29. ISSN: 1061-4036. DOI: 10.1038/75556. URL: http://www.nature.com/ng/journal/v25/n1/abs/ng0500_25.html (visited on 12/07/2016).

[6] Seth I. Berger, Jeremy M. Posner, and Avi Ma'ayan. "Genes2Networks: connecting lists of gene symbols using mammalian protein interactions databases". In: *BMC Bioinformatics* 8 (2007), p. 372. ISSN: 1471-2105. DOI: 10.1186/1471-2105-8-372. URL: http://dx.doi.org/10.1186/1471-2105-8-372 (visited on 12/13/2016).

[7] Varsha Bhatt and Abdus Saleem. "Drug-Induced Neutropenia – Pathophysiology, Clinical Features, and Management". In: *Annals of Clinical & Laboratory Science* 34.2 (Apr. 1, 2004), pp. 131–137. ISSN: 0091-7370, 1550-8080. URL: http://www.annclinlabsci.org/content/34/2/131 (visited on 12/18/2016).

[8] William S. Bush and Jason H. Moore. "Chapter 11: Genome-Wide Association Studies". In: *PLOS Comput Biol* 8.12 (Dec. 27, 2012), e1002822. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1002822. URL: http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002822 (visited on 08/16/2016).

[9] J. M. Campbell et al. "Irinotecan-induced toxicity pharmacogenetics: an umbrella review of systematic reviews and meta-analyses". In: *The Pharmacogenomics Journal* (Aug. 9, 2016). ISSN: 1470-269X. DOI: 10.1038/tpj.2016.58. URL: http://www.nature.com/tpj/journal/vaop/ncurrent/full/tpj201658a.html (visited on 09/06/2016).

[10] Marc Carlson. *org.Hs.eg.db: Genome wide annotation for Human.* R package version 3.4.0. 2016.

[11] Marc Carlson and Bioconductor Package Maintainer. *TxDb.Hsapiens.UCSC.hg19.knownGene: Annotation package for TxDb object(s)*. R package version 3.2.2. 2015.

[12] Guy G. Chabot. "Clinical Pharmacokinetics of Irinotecan". In: *Clinical Pharmacokinetics* 33.4 (Oct. 19, 2012), pp. 245–259. ISSN: 0312-5963, 1179-1926. DOI: 10.2165/00003088-199733040-00001. URL: http://link.springer.com/article/10.2165/00003088-199733040-00001 (visited on 09/05/2016).

[13] Gary K. Chen et al. "Genome-wide association analyses of expression phenotypes". In: *Genetic Epidemiology* 31 Suppl 1 (2007), S7–S11. ISSN: 0741-0395. DOI: 10.1002/gepi.20275.

[14] Donald F. Conrad et al. "Origins and functional impact of copy number variation in the human genome". In: *Nature* 464.7289 (Apr. 1, 2010), pp. 704–712. ISSN: 0028-0836. DOI: 10.1038/nature08516. URL: http://www.nature.com/nature/journal/v464/n7289/full/nature08516.html (visited on 12/09/2016).

[15] George N. Cox et al. "Hematopoietic Properties of Granulocyte Colony-Stimulating Factor/Immunoglobulin (G-CSF/IgG-Fc) Fusion Proteins in Normal and Neutropenic Rodents". In: *PLoS ONE* 9.3 (Mar. 17, 2014). ISSN: 1932-6203. DOI: 10.1371/journal.pone.0091990. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3956888/ (visited on 12/18/2016).

[16] Jeffrey Crawford, David C. Dale, and Gary H. Lyman. "Chemotherapy-induced neutropenia: risks, consequences, and new directions for its management". In: *Cancer* 100.2 (Jan. 15, 2004), pp. 228–237. ISSN: 0008-543X. DOI: 10.1002/cncr.11882.

[17] D. J. Crona et al. "Clinical validity of new genetic biomarkers of irinotecan neutropenia: an independent replication study". In: *The Pharmacogenomics Journal* 16.1 (Feb. 2016), pp. 54–59. ISSN: 1470-269X. DOI: 10.1038/tpj.2015.23. URL: http://www.nature.com/tpj/journal/v16/n1/full/tpj201523a.html (visited on 11/22/2016).

[18] Daniel Crona and Federico Innocenti. "Can knowledge of germline markers of toxicity optimize dosing and efficacy of cancer therapy?" In: *Biomarkers in medicine* 6.3 (June 2012), pp. 349–362. ISSN: 1752-0363. DOI: 10.2217/bmm.12.19. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3704209/ (visited on 12/06/2016).

[19] M. Dowle et al. *data.table: Extension of Data.frame*. R package version 1.9.6. 2015. URL: https://CRAN.R-project.org/package=data.table.

[20] Laurent Duret. "Neutral theory: The null hypothesis of molecular evolution." In: *Nature Education* 1.1 (2008), p. 218. URL: http://www.nature.com/scitable/topicpage/neutral-theory-the-null-hypothesis-of-molecular-839.

[21] E. A. Eisenhauer et al. "New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1)". In: *European Journal of Cancer*. Response assessment in solid tumours (RECIST): Version 1.1 and supporting papers 45.2 (Jan. 2009), pp. 228–247. ISSN: 0959-8049. DOI: 10.1016/j.ejca.2008.10.026. URL: http://www.sciencedirect.com/science/article/pii/S0959804908008733 (visited on 12/07/2016).

[22] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. "Regularization Paths for Generalized Linear Models via Coordinate Descent". In: *Journal of Statistical Software* 33.1 (2010), pp. 1–22. ISSN: 1548-7660. DOI: 10.18637/jss.v033.i01. URL: https://www.jstatsoft.org/index.php/jss/article/view/v033i01.

[23] Ken-ichi Fujita et al. "Irinotecan, a key chemotherapeutic drug for metastatic colorectal cancer". In: *World Journal of Gastroenterology* 21.43 (Nov. 21, 2015), pp. 12234–12248. ISSN: 1007-9327. DOI: 10.3748/wjg.v21.i43.12234. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4649109/ (visited on 12/01/2016).

[24] Erik Garrison and Gabor Marth. "Haplotype-based variant detection from short-read sequencing". In: *arXiv:1207.3907 [q-bio]* (July 17, 2012). arXiv: 1207.3907. URL: http://arxiv.org/abs/1207.3907 (visited on 12/07/2016).

[25] Sigal Gery et al. "C/EBP interacts with retinoblastoma and E2F1 during granulopoiesis". In: *Blood* 103.3 (Feb. 1, 2004), pp. 828–835. ISSN: 0006-4971, 1528-0020. DOI: 10.1182/blood-2003-01-0159. URL: http://www.bloodjournal.org/content/103/3/828 (visited on 11/25/2016).

[26] Benjamin A Goldstein, Eric C Polley, and Farren B. S. Briggs. "Random Forests for Genetic Association Studies". In: *Statistical Applications in Genetics and Molecular Biology* 10.1 (2011). ISSN: 1544-6115. DOI: 10.2202/1544-6115.1691. URL: https://www.degruyter.com/view/j/sagmb.2011.10.issue-1/sagmb.2011.10.1.1691/sagmb.2011.10.1.1691.xml (visited on 10/20/2016).

[27] J.-Y. Han et al. "A genome-wide association study for irinotecan-related severe toxicities in patients with advanced non-small-cell lung cancer". In: *The Pharmacogenomics Journal* 13.5 (Oct. 2013), pp. 417–422. ISSN: 1473-1150. DOI: 10.1038/tpj.2012.24.

[28] J.-Y. Han et al. "A genome-wide association study of survival in small-cell lung cancer patients treated with irinotecan plus cisplatin chemotherapy". In: *The Pharmacogenomics Journal* 14.1 (Feb. 2014), pp. 20–27. ISSN: 1473-1150. DOI: 10.1038/tpj.2013.7.

[29] Ji-Youn Han et al. "Association of SUMO1 and UBC9 genotypes with tumor response in non-small-cell lung cancer treated with irinotecan-based chemotherapy". In: *The Pharmacogenomics Journal* 10.2 (Apr. 2010), pp. 86–93. ISSN: 1470-269X. DOI: 10.1038/tpj.2009.46. URL: http://www.nature.com/tpj/journal/v10/n2/full/tpj200946a.html (visited on 12/18/2016).

[30] Daniel L. Hertz and Howard L. McLeod. "Use of pharmacogenetics for predicting cancer prognosis and treatment exposure, response and toxicity". In: *Journal of Human Genetics* 58.6 (June 2013), pp. 346–352. ISSN: 1434-5161. DOI: 10.1038/jhg.2013.42. URL: http://www.nature.com/jhg/journal/v58/n6/full/jhg201342a.html (visited on 12/06/2016).

[31] Hideyo Hirai et al. "Cyclic AMP Responsive Element Binding Proteins Are Involved in 'Emergency' Granulopoiesis through the Upregulation of CCAAT/Enhancer Binding Protein ". In: *PLoS ONE* 8.1 (Jan. 30, 2013). ISSN: 1932-6203. DOI: 10.1371/journal.pone.0054862. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3559830/ (visited on 12/20/2016).

[32] Janelle M. Hoskins et al. "Irinotecan Pharmacogenetics: Influence of Pharmacodynamic Genes". In: *Clinical Cancer Research* 14.6 (Mar. 15, 2008), pp. 1788–1796. ISSN: 1078-0432, 1557-3265. DOI: 10.1158/1078-0432.CCR-07-1472. URL: http://clincancerres.aacrjournals.org/content/14/6/1788 (visited on 12/06/2016).

[33] Janelle M Hoskins et al. "Pharmacodynamic genes do not influence risk of neutropenia in cancer patients treated with moderately high-dose irinotecan". In: *Pharmacogenomics* 10.7 (July 2009), pp. 1139–1146. ISSN: 1462-2416. DOI: 10.2217/pgs.09.35. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2748108/ (visited on 12/06/2016).

[34] Huber et al. "Orchestrating high-throughput genomic analysis with Bioconductor". In: *Nature Methods* 12.2 (2015), pp. 115–121. URL: http://www.nature.com/nmeth/journal/v12/n2/full/nmeth.3252.html.

[35] Kazuyuki Inoue et al. "Polymorphisms of the UDP-Glucuronosyl Transferase 1A Genes Are Associated with Adverse Events in Cancer Patients Receiving Irinotecan-Based Chemotherapy". In: *The Tohoku Journal of Experimental Medicine* 229.2 (2013), pp. 107–114. DOI: 10.1620/tjem.229.107.

[36] Tomer Itkin et al. "FGF-2 expands murine hematopoietic stem and progenitor cells via proliferation of stromal cells, c-Kit activation, and CXCL12 down-regulation". In: *Blood* 120.9 (Aug. 30, 2012), pp. 1843–1855. ISSN: 0006-4971, 1528-0020. DOI: 10.1182/blood-2011-11-394692. URL: http://www.bloodjournal.org/content/120/9/1843 (visited on 12/18/2016).

[37] Lalitha Iyer et al. "Phenotype-genotype correlation of in vitro SN-38 (active metabolite of irinotecan) and bilirubin glucuronidation in human liver tissue with UGT1A1 promoter polymorphism". In: *Clinical Pharmacology & Therapeutics* 65.5 (May 1, 1999), pp. 576–582. ISSN: 1532-6535. DOI: 10.1016/S0009-9236(99)70078-0. URL: http://onlinelibrary.wiley.com/doi/10.1016/S0009-9236(99)70078-0/abstract (visited on 12/05/2016).

[38] Gareth James et al. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated, 2014. ISBN: 978-1-4614-7137-0.

[39] Taku Kambayashi and Terri M. Laufer. "Atypical MHC class II-expressing antigen-presenting cells: can anything replace a dendritic cell?" In: *Nature Reviews Immunology* 14.11 (Nov. 2014), pp. 719–730. ISSN: 1474-1733. DOI: 10.1038/nri3754. URL: http://www.nature.com/nri/journal/v14/n11/abs/nri3754.html (visited on 12/18/2016).

[40] Atanas Kamburov et al. "ConsensusPathDB: toward a more complete picture of cell biology". In: *Nucleic Acids Research* 39 (Database issue Jan. 2011), pp. D712–D717. ISSN: 0305-1048. DOI: 10.1093/nar/gkq1156. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3013724/ (visited on 12/12/2016).

[41] Minoru Kanehisa et al. "KEGG as a reference resource for gene and protein annotation". In: *Nucleic Acids Research* 44 (D1 Jan. 4, 2016), pp. D457–D462. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkv1070. URL: http://nar.oxfordjournals.org/content/44/D1/D457 (visited on 12/07/2016).

[42] Panagiotis Katsonis et al. "Single nucleotide variations: Biological impact and theoretical interpretation". In: *Protein Science : A Publication of the Protein Society* 23.12 (Dec. 2014), pp. 1650–1666. ISSN: 0961-8368. DOI: 10.1002/pro.2552. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4253807/ (visited on 09/26/2016).

[43] Deanna L. Kroetz. "Role for Drug Transporters Beyond Tumor Resistance: Hepatic Functional Imaging and Genotyping of Multidrug Resistance Transporters for the Prediction of Irinotecan Toxicity". In: *Journal of Clinical Oncology* 24.26 (Sept. 10, 2006), pp. 4225–4227. ISSN: 0732-183X. DOI: 10.1200/JCO.2006.07.2355. URL: http://ascopubs.org/doi/abs/10.1200/JCO.2006.07.2355 (visited on 12/01/2016).

[44] Dinemarie Kweekel, Henk-Jan Guchelaar, and Hans Gelderblom. "Clinical and pharmacogenetic factors associated with irinotecan toxicity". In: *Cancer Treatment Reviews* 34.7 (Nov. 2008), pp. 656–669. ISSN: 0305-7372. DOI: 10.1016/j.ctrv.2008.05.002. URL: http://www.sciencedirect.com/science/article/pii/S0305737208001837 (visited on 09/06/2016).

[45] J. M. Kwon and A. M. Goate. "The candidate gene approach". In: *Alcohol Research & Health: The Journal of the National Institute on Alcohol Abuse and Alcoholism* 24.3 (2000), pp. 164–168. ISSN: 1535-7414.

[46] Andy Liaw and Matthew Wiener. "Classification and Regression by randomForest". In: *R News* 2.3 (2002), pp. 18–22. URL: http://CRAN.R-project.org/doc/Rnews/.

[47] Xiaoxi B. Lin et al. "The Role of Intestinal Microbiota in Development of Irinotecan Toxicity and in Toxicity Reduction through Dietary Fibres in Rats". In: *PLOS ONE* 9.1 (Jan. 14, 2014), e83644. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0083644. URL: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0083644 (visited on 09/06/2016).

[48] Harvey Lodish et al. *Proto-Oncogenes and Tumor-Suppressor Genes*. 2000. URL: https://www.ncbi.nlm.nih.gov/books/NBK21662/ (visited on 01/04/2017).

[49] Peter Lung et al. "An Apparent Lack of HLA Restriction in the Stimulation of Granulocyte-Macrophage Colony Formation from Normal Human Null Cells by Helper T Lymphocytes". In: *Scandinavian Journal of Haematology* 31.1 (July 1, 1983), pp. 23–30. ISSN: 1600-0609. DOI: 10.1111/j.1600-0609.1983.tb02132.x.

URL: http://onlinelibrary.wiley.com/doi/10.1111/j.1600-0609.1983.tb02132.x/abstract (visited on 12/18/2016).

[50] Teri A. Manolio. "Genomewide Association Studies and Assessment of the Risk of Disease". In: *New England Journal of Medicine* 363.2 (July 8, 2010), pp. 166–176. ISSN: 0028-4793. DOI: 10.1056/NEJMra0905980. URL: http://dx.doi.org/10.1056/NEJMra0905980 (visited on 12/06/2016).

[51] Sharon Marsh and Janelle M Hoskins. "Irinotecan pharmacogenomics". In: *Pharmacogenomics* 11.7 (July 2010), pp. 1003–1010. ISSN: 1462-2416. DOI: 10.2217/pgs.10.95. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2927346/ (visited on 09/05/2016).

[52] Craig H. Mermel et al. "Src family kinases are important negative regulators of G-CSF-dependent granulopoiesis". In: *Blood* 108.8 (Oct. 15, 2006), pp. 2562–2568. ISSN: 0006-4971. DOI: 10.1182/blood-2006-05-024307. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1895577/ (visited on 12/19/2016).

[53] Florian Mittag, Michael Römer, and Andreas Zell. "Influence of Feature Encoding and Choice of Classifier on Disease Risk Prediction in Genome-Wide Association Studies". In: *PLOS ONE* 10.8 (Aug. 18, 2015), e0135832. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0135832. URL: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0135832 (visited on 09/26/2016).

[54] Goro Nakayama et al. "The impact of dose/time modification in irinotecan- and oxaliplatin-based chemotherapies on outcomes in metastatic colorectal cancer". In: *Cancer Chemotherapy and Pharmacology* 73.4 (Apr. 1, 2014), pp. 847–855. ISSN: 0344-5704, 1432-0843. DOI: 10.1007/s00280-014-2416-x. URL: http://link.springer.com/article/10.1007/s00280-014-2416-x (visited on 12/19/2016).

[55] Pauline C. Ng and Steven Henikoff. "SIFT: predicting amino acid changes that affect protein function". In: *Nucleic Acids Research* 31.13 (July 1, 2003), pp. 3812–3814. ISSN: 0305-1048. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC168916/ (visited on 01/03/2017).

[56] Kristin K. Nicodemus and James D. Malley. "Predictor correlation impacts machine learning algorithms: implications for genomic studies". In: *Bioinformatics* 25.15 (Aug. 1, 2009), pp. 1884–1890. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btp331. URL: http://bioinformatics.oxfordjournals.org/content/25/15/1884 (visited on 09/26/2016).

[57] Falk Nimmerjahn and Jeffrey V. Ravetch. "Fc receptors as regulators of immune responses". In: *Nature Reviews Immunology* 8.1 (Jan. 2008), pp. 34–47. ISSN: 1474-1733. DOI: 10.1038/nri2206. URL: http://www.nature.com/nri/journal/v8/n1/full/nri2206.html (visited on 12/18/2016).

[58] Nuala A. O'Leary et al. "Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation". In: *Nucleic Acids Research* 44 (D1 Jan. 4, 2016), pp. D733–D745. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkv1189. URL: http://nar.oxfordjournals.org/content/44/D1/D733 (visited on 12/07/2016).

[59] Masahide Onoue et al. "UGT1A1*6 polymorphism is most predictive of severe neutropenia induced by irinotecan in Japanese cancer patients". In: *International Journal of Clinical Oncology* 14.2 (Apr. 2009), pp. 136–142. ISSN: 1341-9625. DOI: 10.1007/s10147-008-0821-z.

[60] World Health Organization. "19th WHO Model List of Essential Medicines". In: (May 8, 2015).

[61] Herve Pages. *SNPlocs.Hsapiens.dbSNP144.GRCh37: SNP locations for Homo sapiens (dbSNP Build 144)*. R package version 0.99.11. 2015.

[62] Peristera Paschou et al. "Intra- and interpopulation genotype reconstruction from tagging SNPs". In: *Genome Research* 17.1 (Jan. 2007), pp. 96–107. ISSN: 1088-9051. DOI: 10.1101/gr.5741407. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1716273/ (visited on 12/09/2016).

[63] Radhika Patnala, Judith Clements, and Jyotsna Batra. "Candidate gene association studies: a comprehensive guide to useful in silico tools". In: *BMC Genetics* 14 (2013), p. 39. ISSN: 1471-2156. DOI: 10.1186/1471-2156-14-39. URL: http://dx.doi.org/10.1186/1471-2156-14-39 (visited on 12/05/2016).

[64] Nick Patterson, Alkes L. Price, and David Reich. "Population Structure and Eigenanalysis". In: *PLOS Genet* 2.12 (Dec. 22, 2006), e190. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.0020190. URL: http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.0020190 (visited on 10/21/2016).

[65] Thomas A. Pearson. "How to Interpret a Genome-wide Association Study". In: *JAMA* 299.11 (Mar. 19, 2008), p. 1335. ISSN: 0098-7484. DOI: 10.1001/jama.299.11.1335. URL: http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.299.11.1335 (visited on 10/19/2016).

[66] Alkes L. Price et al. "Principal components analysis corrects for stratification in genome-wide association studies". In: *Nature Genetics* 38.8 (Aug. 2006), pp. 904–909. ISSN: 1061-4036. DOI: 10.1038/ng1847. URL: http://www.nature.com/ng/journal/v38/n8/full/ng1847.html (visited on 12/13/2016).

[67] Stefan M. Pulst. "Genetic Linkage Analysis". In: *Archives of Neurology* 56.6 (June 1, 1999), pp. 667–672. ISSN: 0003-9942. DOI: 10.1001/archneur.56.6.667. URL: http://jamanetwork.com/journals/jamaneurology/fullarticle/775035 (visited on 12/05/2016).

[68] Shaun Purcell et al. "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses". In: *American Journal of Human Genetics* 81.3 (Sept. 2007), pp. 559–575. ISSN: 0002-9297. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1950838/ (visited on 12/01/2016).

[69] Mullangi Ramesh, Preeti Ahlawat, and Nuggehally R. Srinivas. "Irinotecan and its active metabolite, SN-38: review of bioanalytical methods and recent update from clinical pharmacology perspectives". In: *Biomedical Chromatography* 24.1 (Jan. 1, 2010), pp. 104–123. ISSN: 1099-0801. DOI: 10.1002/bmc.1345. URL: http://onlinelibrary.wiley.com/doi/10.1002/bmc.1345/abstract (visited on 09/05/2016).

[70] M. L. Rothenberg. "Topoisomerase I inhibitors: Review and update". In: *Annals of Oncology* 8.9 (Aug. 1, 1997), pp. 837–855. ISSN: 0923-7534, 1569-8041. URL: http://annonc.oxfordjournals.org/content/8/9/837 (visited on 12/06/2016).

[71] R. Sachidanandam et al. "A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms". In: *Nature* 409.6822 (Feb. 15, 2001), pp. 928–933. ISSN: 0028-0836. DOI: 10.1038/35057149.

[72] Nagahiro Saijo et al. "7-Ethyl-10-[4-(1-piperidino)-1-piperidino] carbonyloxy camptothecin: mechanism of resistance and clinical trials". In: *Cancer Chemotherapy and Pharmacology* 34.1 (Jan. 1, 1994), S112–S117. ISSN: 0344-5704, 1432-0843. DOI: 10.1007/BF00684874. URL: http://link.springer.com/article/10.1007/BF00684874 (visited on 12/01/2016).

[73] Alexandre Santos et al. "Metabolism of Irinotecan (CPT-11) by CYP3A4 and CYP3A5 in Humans". In: *Clinical Cancer Research* 6.5 (May 1, 2000), pp. 2012–2020. ISSN: 1078-0432, 1557-3265. URL: http://clincancerres.aacrjournals.org/content/6/5/2012 (visited on 12/06/2016).

[74] S. T. Sherry et al. "dbSNP: the NCBI database of genetic variation". In: *Nucleic Acids Research* 29.1 (Jan. 1, 2001), pp. 308–311. ISSN: 0305-1048. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC29783/ (visited on 12/06/2016).

[75] Hashem A Shihab et al. "Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models". In: *Human Mutation* 34.1 (Jan. 2013), pp. 57–65. ISSN: 1059-7794. DOI: 10.1002/humu.22225. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3558800/ (visited on 11/29/2016).

[76] Lillian L. Siu et al. "Facilitating a culture of responsible and effective sharing of cancer genome data". In: *Nature Medicine* 22.5 (May 2016), pp. 464–471. ISSN: 1078-8956. DOI: 10.1038/nm.4089. URL: http://www.nature.com/nm/journal/v22/n5/abs/nm.4089.html (visited on 01/04/2017).

[77] Silke Szymczak et al. "Machine learning in genome-wide association studies". In: *Genetic Epidemiology* 33 Suppl 1 (2009), S51–57. ISSN: 1098-2272. DOI: 10.1002/gepi.20473.

[78] Hiro Takahashi et al. "Application of a Combination of a Knowledge-Based Algorithm and 2-Stage Screening to Hypothesis-Free Genomic Data on Irinotecan-Treated Patients for Identification of a Candidate Single Nucleotide Polymorphism Related to an Adverse Effect". In: *PLoS ONE* 9.8 (2014). DOI: 10.1371/

journal.pone.0105160. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4134257/ (visited on 12/18/2016).

[79] R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2016. URL: https://www.R-project.org/.

[80] Stephen D. Turner. "qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots". In: *bioRxiv* (May 14, 2014), p. 005165. DOI: 10.1101/005165. URL: http://biorxiv.org/content/early/2014/05/14/005165 (visited on 12/07/2016).

[81] Mike Tyers and Matthias Mann. "From genomics to proteomics". In: *Nature* 422.6928 (Mar. 13, 2003), pp. 193–197. ISSN: 0028-0836. DOI: 10.1038/nature01510. URL: http://www.nature.com/nature/journal/v422/n6928/full/nature01510.html (visited on 12/18/2016).

[82] E. Wasserman et al. "Severe CPT-11 toxicity in patients with Gilbert's syndrome: Two case reports". In: *Annals of Oncology* 8.10 (Oct. 1, 1997), pp. 1049–1051. ISSN: 0923-7534, 1569-8041. URL: http://annonc.oxfordjournals.org/content/8/10/1049 (visited on 12/05/2016).

[83] Danielle Welter et al. "The NHGRI GWAS Catalog, a curated resource of SNP-trait associations". In: *Nucleic Acids Research* 42 (Database issue Jan. 1, 2014), pp. D1001–D1006. ISSN: 0305-1048. DOI: 10.1093/nar/gkt1229. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3965119/ (visited on 12/06/2016).

[84] Lodewyk F. A. Wessels et al. "A protocol for building and evaluating predictors of disease state based on microarray data". In: *Bioinformatics* 21.19 (Jan. 1, 2005), pp. 3755–3762. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/bti429. URL: http://bioinformatics.oxfordjournals.org/content/21/19/3755 (visited on 12/09/2016).

[85] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009. ISBN: 978-0-387-98140-6. URL: http://ggplot2.org.

[86] Max Kuhn Contributions from Jed Wing et al. *caret: Classification and Regression Training*. R package version 6.0-73. 2016. URL: https://CRAN.R-project.org/package=caret.

[87] Tomasz M. Witkos and Martin Lowe. "The Golgin Family of Coiled-Coil Tethering Proteins". In: *Frontiers in Cell and Developmental Biology* 3 (Jan. 11, 2016). ISSN: 2296-634X. DOI: 10.3389/fcell.2015.00086. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4707255/ (visited on 12/20/2016).

[88] Andrew Yates et al. "Ensembl 2016". In: *Nucleic Acids Research* 44 (D1 Jan. 4, 2016), pp. D710–D716. ISSN: 0305-1048. DOI: 10.1093/nar/gkv1157. URL: https://academic.oup.com/nar/article/44/D1/D710/2502651/Ensembl-2016 (visited on 12/07/2016).

[89] Xiang Zhang et al. "Chapter 10: Mining Genome-Wide Genetic Markers". In: *PLOS Comput Biol* 8.12 (Dec. 27, 2012), e1002828. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1002828. URL: http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002828 (visited on 09/26/2016).

[90] Meng Zhao et al. "FGF signaling facilitates postinjury recovery of mouse hematopoietic system". In: *Blood* 120.9 (Aug. 30, 2012), pp. 1831–1842. ISSN: 0006-4971. DOI: 10.1182/blood-2011-11-393991. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3433089/ (visited on 12/19/2016).

[91] Hui Zou and Trevor Hastie. "Regularization and variable selection via the elastic net". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2 (Apr. 1, 2005), pp. 301–320. ISSN: 1467-9868. DOI: 10.1111/j.1467-9868.2005.00503.x. URL: http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9868.2005.00503.x/abstract (visited on 12/09/2016).