Cross-validation and cross-study validation of chronic lymphocytic leukaemia with exome sequences and machine learning

Abdulrhman Aljouie

Department of Computer Science,
New Jersey Institute of Technology,
Newark, NJ 07102, USA
Email: aa547@njit.edu

Nihir Patel and Bharati Jadhav

Department of Genetics and Genomics Sciences,
Icahn School of Medicine at Mount Sinai Hospital,
Hess Center for Science and Medicine,
New York City, NY 10029, USA
Email: nihir.patel@mssm.edu
Email: bharati.jadhav@mssm.edu

Usman Roshan*

Department of Computer Science,
New Jersey Institute of Technology,
Newark, NJ 07102, USA
Email: usman@njit.edu
*Corresponding author

Abstract: The era of genomics brings the potential of better DNA-based risk prediction and treatment. We explore this problem for chronic lymphocytic leukaemia that is one of the largest whole exome data set available from the NIH dbGaP database. We perform a standard next-generation sequence procedure to obtain Single-Nucleotide Polymorphism (SNP) variants and obtain a peak mean accuracy of 82% in our cross-validation study. We also cross-validate an Affymetrix 6.0 genome-wide association study of the same samples where we find a peak accuracy of 57%. We then perform a cross-study validation with exome samples from other studies in the NIH dbGaP database serving as the external data set. There we obtain an accuracy of 70% with top Pearson ranked SNPs obtained from the original exome data set. Our study shows that even with a small sample size we can obtain moderate to high accuracy with exome sequences, which is encouraging for future work.

Keywords: exome wide association study; chronic lymphocytic leukaemia; machine learning; disease risk prediction.

Reference to this paper should be made as follows: Aljouie, A., Patel, N., Jadhav, B. and Roshan, U. (2016) 'Cross-validation and cross-study validation of chronic lymphocytic leukaemia with exome sequences and machine learning', *Int. J. Data Mining and Bioinformatics*, Vol. 16, No. 1, pp.47–63.

Biographical notes: Abdulrhman Aljouie is currently a PhD student in the Computer Science Department at the New Jersey Institute of Technology. His research interests are in genomics, in particular machine learning applications for disease risk prediction and high performance GPU based genomics methods.

Nihir Patel is currently an Associate Bioinformatician in the Icahn School of Medicine at the Mount Sinai Hospital. He received his Masters in Bioinformatics from the New Jersey Institute of Technology.

Bharati Jadhav is currently a Bioinformatician in the Icahn School of Medicine at the Mount Sinai Hospital. She received her Masters in Bioinformatics from the New Jersey Institute of Technology.

Usman Roshan is an Associate Professor in the Department of Computer Science at the New Jersey Institute of Technology. He received his PhD in Computer Science from The University of Texas at Austin in 2004. His research interests are in risk prediction from genomic data, high performance genomic methods, and machine learning methods in particular methods for data representation and 0/1 loss optimisation.

1 Introduction

In the last few years there have been many studies exploring disease risk prediction with machine learning methods and genome-wide association studies (GWAS) (Abraham et al., 2013; Chatterjee et al., 2013; Kruppa et al., 2012; Roshan et al., 2011; Sandhu et al., 2010; Kooperberg et al., 2010; Evans et al., 2009; Janssens et al., 2008; Wray et al., 2007; Wray et al., 2008). This includes various cancers and common diseases (Kraft and Hunter, 2009; Gail, 2008; Morrison et al., 2007; Kathiresan et al., 2008; Paynter et al., 2009). Most studies employ a two-fold machine learning approach. First they identify variants from a set of *training* individuals that consist of both case and controls. This is usually a set of single nucleotide polymorphisms (SNPs) that pass a significance test or a number of top ranked SNPs given by a univariate ranking method. In the second part they learn a model with the reduced set of variants on the training data and predict the case and control of a *validation* set of individuals.

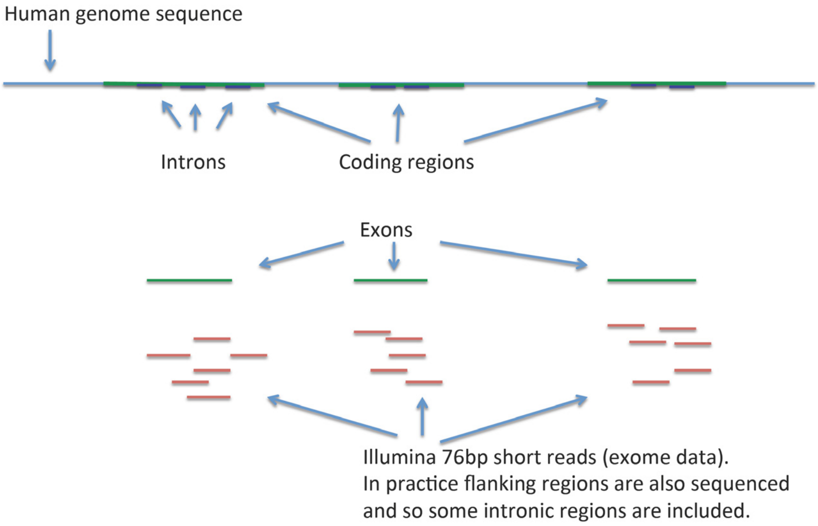
For diseases of low and moderate frequency SNPs have been shown to be more accurate than family history under a theoretical model of prediction (Do et al., 2012). However, for diseases with high frequency and heritability family history based models perform better (Do et al., 2012). Clinical factors with SNPs yields an area under curve (AUC) of 0.8 in a Japanese type 2 diabetes dataset but their SNPs have a marginal contribution of 0.01 to the accuracy (Shigemizu et al., 2014). With a large sample size the highest known AUC of 0.86 and 0.82 for Crohn's disease and ulcerative colitis were reported (Wei et al., 2013). There the authors contend this may be a peak or considerably larger sample sizes would be needed for higher AUCs. Bootstrap methods have given AUCs of 0.82 and 0.83 for type 2 diabetes and bipolar disease on the Wellcome Trust Case Control Consortium (2007) datasets, considerably higher than previous studies. Some studies have also used interacting SNPs in GWAS to boost risk prediction accuracy (Okser et al., 2013; Eleftherohorinou et al., 2009).

Many of these studies are cross-validation. They split the original dataset into training and validation several times randomly and for each split predict case and controls in the validation. Recent work has shown that this may not necessarily generalise to data from different studies (Bernau et al., 2014). Thus, in any risk prediction study it is now essential to include cross-study validation on an independent dataset.

While continuing efforts are made to improve risk prediction accuracy with GWAS datasets the AUCs are still below clinical risk prediction particularly for cancer. The reasons posed for this failure include lack of rare variants, insufficient sample size, and low coverage (.1% of the genome sequenced) (Schrodi et al., 2014; Manolio, 2013; Visscher et al., 2012). In this paper we detect variants from whole exome data that has a much larger coverage. We seek to determine the cross-validation and cross-study prediction accuracy achieved with variants detected in whole exome data and a machine learning pipeline.

For our study we obtained a chronic lymphocytic leukaemia 140X coverage whole exome dataset (Wang et al., 2011; Landau et al., 2013) of 186 tumour and 169 matched germline controls from the NIH dbGaP (Mailman et al., 2007). The whole exome dataset is composed of short next generation sequence reads of exomes as shown in Figure 1. This is one of the largest datasets available there and is an adult leukaemia with an onset median age of 70 (Shanshal and Haddad, 2012). There is currently no known early SNP based detection test for this cancer. Current tests include physical exam, family history, blood count, and other tests given by the National Cancer Institute (see http://www.cancer.gov/cancertopics/pdq/treatment/cll/patient).

Figure 1 Whole exome sequences are short reads of exomes obtained by next generation sequencing (see online version for colours)


Human genome sequence
Introns
Coding regions
Exons
Illumina 76bp short reads (exome data).
In practice flanking regions are also sequenced
and so some intronic regions are included.

We mapped short read exome sequences to the human genome reference GRCh37 with the popular BWA (Li and Durbin, 2009) short read alignment program. We then used GATK (McKenna et al., 2010; DePristo et al., 2011; Auwera et al., 2013) and the Broad Institute exome capture kit (bundle 2.8 b37) in a rigorous quality control procedure to obtain SNP and indel variants. We excluded case and controls that contained excessive missing variants and in the end obtained 122,392 SNPs and 2200 indels across 153 cases and 144 controls.

To better understand the risk prediction value of these variants we perform a cross-validation study on the total 153 cases and 144 controls by creating random training validation splits. We compare the same cross-validation accuracy to that on an Affymetrix 6.0 panel genome wide association study for the same subjects to see the improvement given by our exome analysis. We also obtained exome sequences from three different studies from dbGaP for independent external validation (also known as cross-study validation; Bernau et al., 2014). We rank SNPs in our training set with the Pearson correlation coefficient (Guyon and Elisseeff, 2003) and predict labels of cases and controls with the support vector machine classifier in external validation dataset. We also study the biological significance of top Pearson ranked SNPs in our data. Below we provide details of our experimental results followed by discussion and future work.

2 Materials and methods

We performed a rigorous analysis on raw exome sequences. We first mapped them to the human genome and obtained variants. We then encoded the variants into integers and created feature vectors for each case and control sample.

2.1 Chronic lymphocytic leukaemia whole exome sequences and human genome reference

We obtained whole exome sequences of 169 chronic lymphocytic leukaemia patients (Wang et al., 2011; Landau et al., 2013) from the NIH dbGaP website (Mailman et al., 2007) with dbGaP study ID phs000435.v2.p1. For each of the 169 patients we have exome tumour sequences and germline controls taken from the same patient. In addition we also obtained exome tumour sequences of 17 patients deposited into dbGaP after publication of the original study (Wang et al., 2011; Landau et al., 2013). This gives us a total of 186 cases and 169 controls. The ancestry of the patients is not given in the publications or in the dbGaP files except that we know they were obtained from the Dana Farber Cancer Institute in Boston, Massachusetts, USA. The data comprises of 76 base pair (bp) paired-end reads produced by Illumina Genome Analyser II and Hiseq2000 machines and the Agilent SureSelect capture kit by the Broad Institute (Wang et al., 2011). The data was sequenced to obtain mean coverage of approximately 140X.

We also obtained the human genome reference sequence version GRCh37.p13 from the Genome Reference Consortium (http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/). At the time of writing this paper version 38 of the human genome sequence was introduced. However, our mapping process started well before its release and demands considerable computational resources. Therefore we continue with version 37 for the work in this paper.

2.2 Next generation sequence analysis pipeline

Our pipeline includes mapping short reads to the reference genome, post processing of alignment, variant calling, and filtering candidate variants. The total exome data was in approximately 3 Terabytes (TB) and required high performance computing infrastructure to process. We automated the analysis pipeline using Perl and various bioinformatics tools as described below.

2.2.1 Mapping reads

As the first step of our analysis we mapped exome short read sequences of 186 tumour cases and 169 matched germline controls to the human genome reference GRCh37 with the BWA MEM program (Li and Durbin, 2009). We excluded 6 cases and 14 controls due to excessively large dataset size and downloading problems and we removed reads with mapping quality (MAPQ) below 15.

The read mapping is a process where we align a short read DNA sequences to a reference genome. There are many different programs available for this task and each one differs in mapping methodology, accuracy, and speed. In our pipeline we used the popular program BWA MEM program (version 0.7a-r405) (Li and Durbin, 2009) that implements Burrows-Wheeler transform. This is fairly accurate for its fast processing speed while mapping against huge reference genome such as humans (Fonseca et al., 2012; Hatem et al., 2013). We used the default parameters for mapping reads to the human reference genome.

BWA MEM produces its output in a standard format called Sequence Alignment Map (SAM). We use SAMtools version 0.1.18 (Li et al., 2009) for further analysis of the SAM output files. We convert each alignments in SAM format into its binary format (BAM), sort the alignment with respect to their chromosomal position, and then index it. We also used SAMtools to generate mapping statistics and merging alignments of the same patient across different files. We used PICARD tool (version 1.8, http://broadinstitute.github.io/picard/) to add read groups which connects the reads to the patient subject. We also removed duplicates reads introduced by PCR amplification process to avoid artefacts using the PICARD MarkDuplicates program. Finally, using SAMtools we removed unmapped reads and ones with mapping score (given in the MAPQ SAM field) smaller than 15.

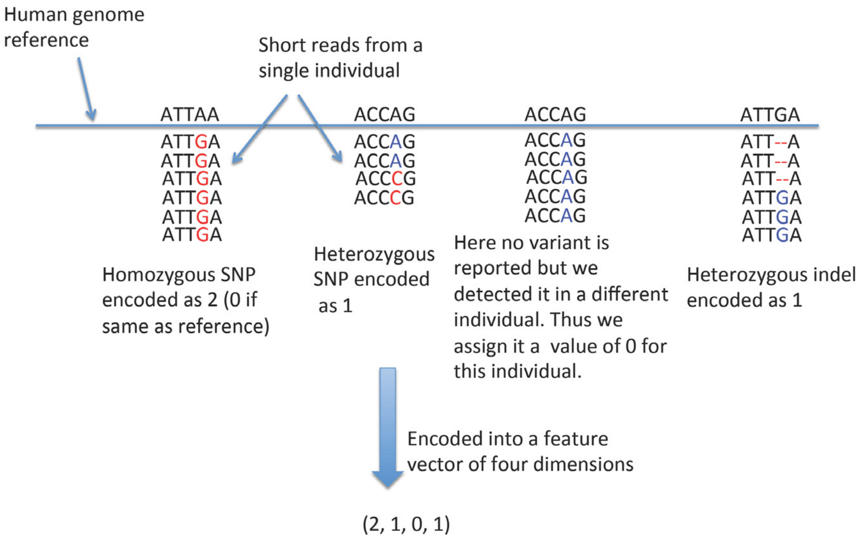
2.2.2 Variant detection

We then used GATK (McKenna et al., 2010; DePristo et al., 2011; Auwera et al., 2013) version 3.2-2 with the Broad Institute exome capture kit (bundle 2.8 b37 available from ftp://ftp.broadinstitute.org/) to detect SNPs in the alignments. We also refer to these SNPs as variants. These variants pass a series of rigorous statistical tests (McKenna et al., 2010; DePristo et al., 2011). If a variant does not pass the quality control or no high quality alignment of a read to the genome was found in that region then GATK reports a missing value.

We found 38 individuals that contain at least 10% missing values. We removed them from our data and recomputed the variants for the remaining 153 cases and 144 controls again with GATK and the exome capture kit. We then removed all variants that have at least one missing value and so eliminate the need for imputation. Note that if we do this before removing the 38 individuals with many missing values we get just a few variants with limited predictive value.

The variant detection procedure gave us a total of 122392 SNPs and 2200 indels. We then encode variants into integers thus obtaining feature vectors for each case and control (as described below).

Figure 2 Toy example depicting the naive 0 1 2 encoding of SNPs and indels. The homozygous and heterozygous genotypes are given by the GATK program (McKenna et al., 2010) when there is a mutation or insertion deletion. For individuals where a SNP is not reported but found in a different individual we use a value of 0 (see online version for colours)


Human genome reference
Short reads from a single individual
ATTAA
ATTGA
ATTGA
ATTGA
ATTGA
ATTGA
ATTGA
ACCAG
ACCAG
ACCAG
ACCCG
ACCCG
ACCAG
ACCAG
ACCAG
ACCAG
ACCAG
ACCAG
ATTGA
ATT--A
ATT--A
ATT--A
ATTGA
ATTGA
ATTGA
Homozygous SNP encoded as 2 (0 if same as reference)
Heterozygous SNP encoded as 1
Here no variant is reported but we detected it in a different individual. Thus we assign it a value of 0 for this individual.
Heterozygous indel encoded as 1
Encoded into a feature vector of four dimensions
(2, 1, 0, 1)

2.3 Machine learning pipeline

After completing the variant analysis in the previous step we proceed with our machine learning analysis. Machine learning methods are widely used to learn models from classified data to make predictions on unclassified data. They consider each data item as a vector in a space of dimension given by the number of features. In our case each data item is a case or control set of exome sequences. By mapping each set to the human genome we obtained variants which represent features. Thus the number of variants determines the number of dimensions in our feature space.

2.3.1 Data encoding

Since the input to machine learning programs must be feature vectors we converted each SNP and indel into an integer. The variants reported by GATK are in standard genotype form A/B where both A and B denote the two alleles found in the individual. The GATK output is in VCF file format whose specifications (available from http://samtools.github.io/hts-specs/VCFv4.1.pdf) provide details on the reported genotypes. When $A = 0$ this denotes the allele in the reference. Other values of 1 through 6 denote alternate alleles and gaps. We kept the max alternative allele option to 6 which is also the default value in GATK. We perform the encoding $7A + B$ to represent all possible outputs.

Each feature vector represents variants from a human individual and is labelled –1 for case and 1 for control. The labels +1 and –1 are standard in the machine learning literature (Alpaydin, 2004).

2.3.2 Feature selection

We rank features with the Pearson correlation coefficient (Guyon and Elisseeff, 2003):

$$\frac{\sum_{i}^{n}(x_{i,j}-x_{i,mean})(y_i-y_{mean})}{\sqrt{\sum_{i}^{n}(x_{i,j}-x_{i,mean})^2}\sqrt{\sum_{i}^{n}(y_i-y_{mean})^2}}$$

where $x_{i,j}$ represents the encoded value of the j-th variant in the i-th individual and y_i is the label (+1 for case and –1 for control) of the i-th individual. The Pearson correlation ranges between +1 and –1 where the extremes denote perfect linear correlation and 0 indicates none. We rank the features by the absolute value of the Pearson correlation.

2.3.3 Classifier

We use the popular soft margin support vector machine (SVM) method (Cortes and Vapnik, 1995) implemented in the SVM-light program (Joachims, 1999) to train and classify a given set of feature vectors created with the above encoding. In brief, the SVM finds the optimally separating hyperplane between feature vectors of two classes (case and control in our case) that minimises the complexity of the classifier plus a regularisation parameter C times error on the training data. For all experiments we use the default regularisation parameter given by $C=\frac{1}{\sum_i x_i^T x_i}$ where n are the number of vectors in the input training (case and control individuals in this study) and x_i is the feature vector of the i^{th} individual (Joachims, 1999). In other words we set C to the inverse of the average squared length of feature vectors in the data.

2.3.4 Measure of accuracy

We define the classification accuracy as $1-BER$ where BER is the balanced error (Guyon, 2004). The balanced error is the average misclassification rate across each class and ranges between 0 and 1. For example suppose class $CASE$ has 10 individuals and $CONTROL$ has 100. If we incorrectly predicted 3 cases and 10 controls then the balanced error is $\left(\frac{3}{10}+\frac{10}{100}\right)/2$ = 0.2.

2.4 High performance computing

We use the Kong computing cluster at NJIT and the condor distributed computing system to speedup our computations.

3 Results

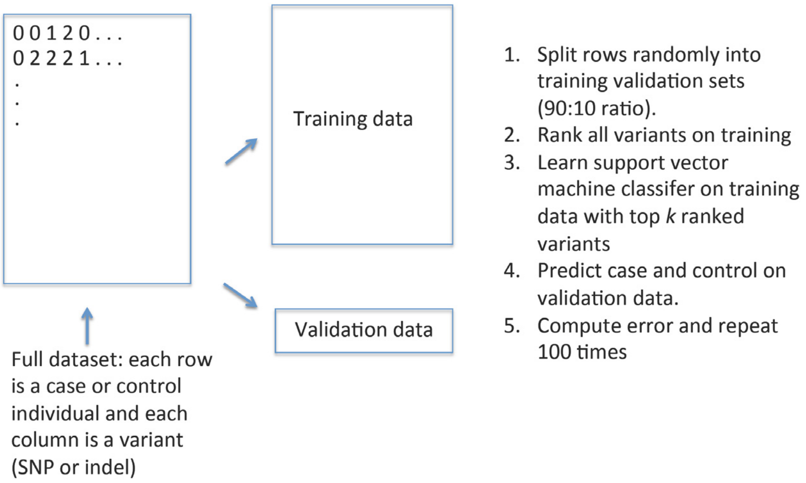
Our next generation sequence pipeline and data encoding gives us feature vectors each representing a case or control dataset and each dimension representing a SNP or indel variant. We can now employ a machine learning procedure to understand the predictive value of the variants.

3.1 Cross-validation

This is a standard approach to evaluate the accuracy of a classifier from a given dataset (Alpaydin, 2004). We randomly shuffle our feature vectors and pick 50% for training and leave the remaining for validation. On the training we rank the variants with the Pearson correlation coefficient. This step is key to performing feature selection in a cross-validation study. Alternatively one may perform feature selection on the whole dataset and then split it into 50% training. However, this method is unrealistic because in practice test labels are not available. In the cross-validation study we simulate that setting by using a validation dataset in place of the test data. The validation labels are only to evaluate the accuracy of the classifer and should not be used for any model training including feature selection. Some studies make this mistake (as previously identified; Smialowski et al., 2010) but we have taken ample care and performed all SNP selection only on the training data.

We then learn a support vector machine (Cortes and Vapnik, 1995) with the SVM-light software (Joachims, 1999) and default regularisation on the training set with k top ranked SNPs (see Figure 3). We consider increments of 10 variants up to 100 and increments of 100 up to 1000. Thus our values of $k = 10, 20, 30, ..., 100, 200, ..., 1000$. For each value of k we predict the case and control status of the validation samples and record the accuracy. We repeat this for 100 random splits and graph the average with standard deviations.

Figure 3 Illustration of cross-validation (see online version for colours)



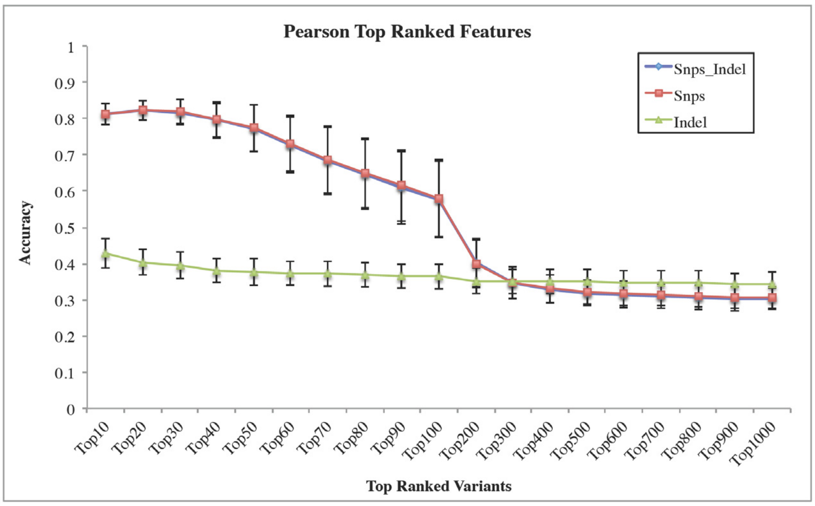
Full dataset: each row is a case or control individual and each column is a variant (SNP or indel)

1. Split rows randomly into training validation sets (90:10 ratio).
2. Rank all variants on training
3. Learn support vector machine classifer on training data with top *k* ranked variants
4. Predict case and control on validation data.
5. Compute error and repeat 100 times

In Figure 4 we show the mean cross-validation accuracy of the support vector machine on 50% training data across 100 random splits. We see that indels alone have much poorer accuracy than SNPs alone and contribute marginally to the SNPs. We achieve a top accuracy of about 82% with the top 20 SNPs. The accuracy drops once we pass the top 20 SNP threshold.

Recall that the accuracies shown in Figure 4 are averaged across 100 training validation splits. In each split we first rank the SNPs and compute prediction on validation with top *k* ranked ones. Thus there is no one set of 20 SNPs to be identified here and this is certainly not the same as the top 20 SNPs from the ranking on the full dataset (although there are some in common with top ranked ones from different splits). Alternatively one may consider the intersection of the top 20 SNPs from all 100 split and use them for prediction on an independent external dataset. The drawback there is that not all of the SNPs in the intersection may pass the GATK quality control filtering thresholds. This is why we choose to rank SNPs on the full dataset and consider the first top 100 that are found in the external dataset.

Figure 4 Average cross-validation accuracy of support vector machine with top Pearson ranked SNPs and indels together and separately on 100 50:50 training validation splits. Also shown are error bars indicating the standard deviation (see online version for colours)

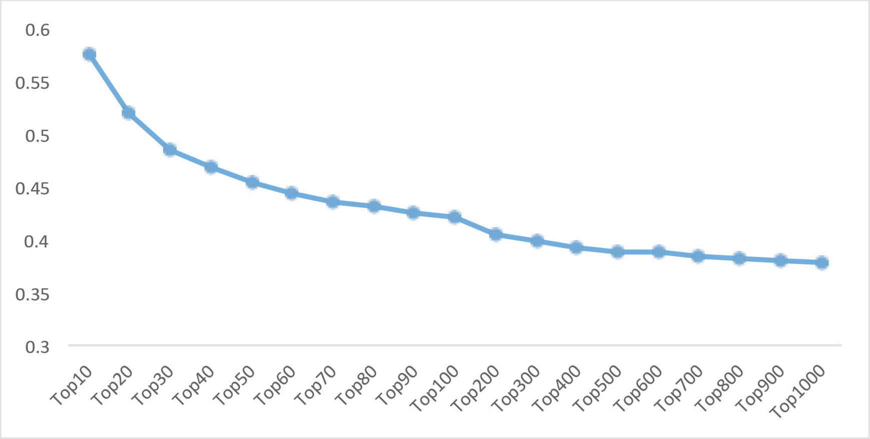


3.2 *Comparison to cross-validation on GWAS*

To better understand the cross-validation results from SNPs obtained in our whole exome analysis we examine a GWAS for the same subjects. This is an Affymetrix 6.0 genome-wide human SNP array of the same disease and subjects that we obtained from the dbGaP site for the whole exome study. We first removed SNPs with more than 10% missing entries and excluded samples that didn't pass the quality control test with 0.4 threshold in the Affymetrix Genotyping Console. The quality control test measures the differences in contrast distributions for homozygote and heterozygote genotypes in each cel file. Following this we ranked the SNPs with the Pearson correlation coefficient. We then created one hundred random 50:50 train and validation splits and determined the average prediction accuracy of the support vector machine in the same manner as described above for the whole exome study.

In Figure 5 we see that the GWAS SNPs give the highest prediction accuracy of 57% in the top 10 SNPs but then it gradually decreases. Thus the SNPs given by our exome analysis which give a higher prediction accuracy may serve as better markers that are not found in the GWAS. Upon closer examination we see that there is no overlap between the top 1000 ranked SNPs from the exome and GWAS datasets except for four that have low Pearson correlation values.

Figure 5 Average cross-validation accuracy of support vector machine with top Pearson ranked SNPs on 100 50:50 training validation splits of the GWAS dataset (see online version for colours)



3.3 *Cross-study validation*

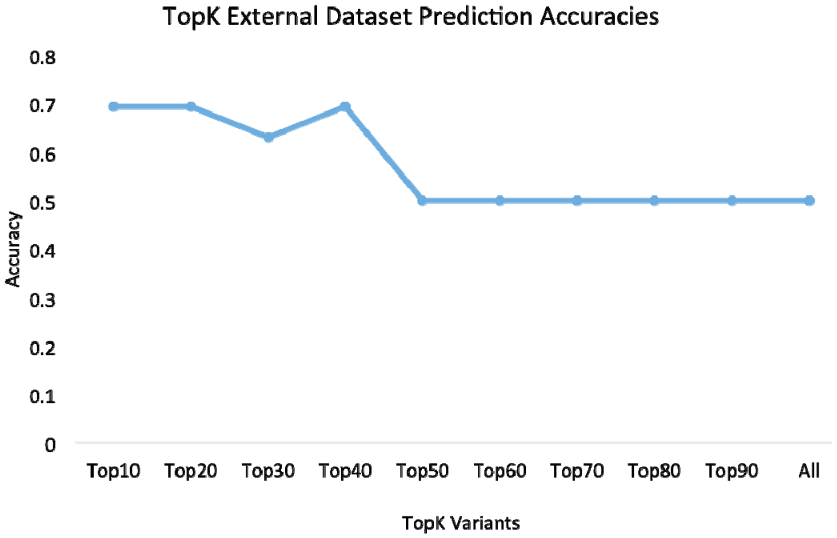
For cross-study validation on an independent dataset we consider a lymphoma whole exome study that has case subjects for lymphocytic leukaemia as well as a few controls. We also consider controls from a head and neck cancer and a breast cancer study.

- Fifteen cases and three controls from a lymphoma whole exome study with dbGaP study ID phs000328.v2.p1 (Pasqualucci et al., 2014). Reads are 101bp length produced from Illumina HiSeq 2000 machine and have 3.4X coverage. The ancestry or origin of data in this study are unavailable in the publication and the dbGaP site.
- Three controls from neck and head cancer whole exome study with dbGaP study ID phs000328.v2.p1 (Stransky et al., 2011). Reads are 77bp length produced from Illumina HiSeq 2000 and have 6.9X coverage. Individuals in this study are from the University of Pittsburgh Head and Neck Spore neoplasm virtual repository.
- Five controls from breast cancer whole exome study with dbGaP study phs000369.v1.p1 ID (Banerji et al., 2012). Reads are 77bp length produced from Illumina HiSeq 2000 of coverage 5.9X. Individuals in this study have Mexican and Vietnamese ancestry.

In all three datasets we followed a similar procedure that we used for the chronic lymphocytic leukaemia exome dataset. We mapped the short reads to the human genome with the BWA program and detect variants with GATK using the same software and parameters as for the lymphocytic leukaemia dataset.

Since this is a validation dataset we cannot use the labels to perform any feature selection or model training. Instead we learn the support vector machine model from the full original dataset. We refer to that as the training set here. We don't consider all SNPs from the training dataset to build a model. We first obtain the top 1000 Pearson correlation coefficient ranked SNPs in the full training. Many of these SNPs don't pass the GATK quality control tests on some of the external validation samples. One reason for this is the much lower coverage (<10X) of the external datasets. Amongst the ones that were detected we consider just the top 100 ranked ones. For each top k ranked ones (for $k = 10, 20, 30, ..., 100$) we learn a support vector machine model on the training and use it to predict labels of the validation data. As discussed above the top k ranked SNPs here are not the same as the top k ranked SNPs in the earlier cross-validation study.

Figure 6 Accuracy of support vector machine with top Pearson ranked SNPs on just the external independent samples. Since this is a validation dataset we cannot use the labels for any type of model training including ranking of features. Thus the ranking of SNPs is obtained from the original full dataset (see online version for colours)



In Figure 6 we see that only the top ranked SNPs give a prediction accuracy above 0.5. We examine the number of cases and controls predicted correctly by the top 20 ranked SNPs in Table 1. Note that the imbalanced accuracy from the table is 64.5%. But in our study we use the balanced accuracy that accounts for different sizes of each class and that value (that we also plot in Figure 6) is 69.4%. In Table 1 we see that the controls for the head and neck cancer are correctly predicted. In the lymphoma dataset also all controls are correctly classified but more than half cases are incorrectly classified as controls.

Table 1 Number of correctly predicted case and controls in three external datasets

Study	*Cases*	*Controls*	*Correct cases*	*Correct controls*
Lymphoma	18	3	7	3
Head and neck cancer	0	3	0	3
Breast cancer	0	7	0	7

3.4 Biological significance of top ranked SNPs

We consider the top 200 ranked SNPs in the Pearson correlation ranking of all SNPs in the full dataset. We run them through the popular ANNOVAR program (Wang et al., 2010) to determine genes and genomic regions they lie on.

We found SNPs in genes SF3B1 and MYD88 both of which were reported as significant genes in the original study of the dataset (Wang et al., 2011). We also found SNPs in genes STRN4 and HLA-DRB5 both of which have been show to be previously associated with this disease in genome wide association studies (Di Bernardo et al., 2008; Berndt et al., 2013; Speedy et al., 2014; Slager et al., 2011). In Table 2 we provide additional details of the SNPs in these genes. All four are exonic but don't necessarily rank high in Pearson correlation coefficient.

Table 2 Details of four variants that are found in genes previously known to be associated with chronic lymphocytic leukaemia. The first column gives the Pearson correlation coefficient value, followed by chromosome number, position in chromosome, SNP rank given by the Pearson correlation coefficient, genomic region, gene, reference nucleotide, alternate nucleotide, and the type

Pearson	*Chr*	*Pos*	*Rank*	*Region*	*Gene*	*Ref*	*Alt*	*Type*
0.19	19	47230736	93	Exonic	STRN4	G	T	Hom
0.19	3	38182641	98	Exonic	MYD88	T	C	Hom
0.19	2	198266834	98	Exonic	SF3B1	T	C	Hom
0.17	6	32497985	159	Exonic	HLA-DRB5	A	G	Hom

We also provide the SNP info from the top three high ranking genes in Table 3. There we see that the Pearson correlation of the top ranked SNPs is considerably higher than the SNPs in known genes identified above. While their direct association with lymphocytic leukaemia is unknown they are well implicated in many different cancers. The highest rank is the Aminoacyl tRNA synthetases (AARS) gene that is known to be associated with various cancers (Park et al., 2008). Following this is the valyl-tRNA synthetase (VARS) gene that is also known to be associated with cancer (Kim et al., 2014). The WD repeat domain 89 (WDR89) is associated with many cancers as given by the Human Protein Atlas http://www.proteinatlas.org/ENSG00000140006-WDR89/cancer and The Cancer Network Galaxy http://tcng.hgc.jp/index.html?t=gene id=112840.

Table 3 Details of top ranking variants given the Pearson correlation coefficient ranking on the full dataset. See Table 2 for more caption details

Pearson	*Chr*	*Pos*	*Rank*	*Region*	*Gene*	*Ref*	*Alt*	*Type*
0.72	16	70305806	1	exon	AARS	G	A	Hom
0.71	16	70305809	2	exon	AARS	G	A	Hom
0.59	16	70305812	3	exon	AARS	C	A	Hom
0.36	6	31749930	5	exon	VARS	C	G	Hom
0.33	14	64066352	9	exon	WDR89	T	A	Hom

In Table 4 we list top ranking SNPs from the GWAS with previous association to this disease and that lie on known genes. Some of these genes are previously linked to leukaemia. For example EML1 (De Keersmaecker et al., 2005), KDM4C (Cheung et al., 2016), NEBL (Emerenciano et al., 2013), BNC2 (Wu et al., 2016), and ANO10 (Lou and Xu, 1997) are all known to be associated with leukaemia while RGS20 and ZNF25 are known to be expressed in leukaemia. However, none of the top 1000 ranked SNPs in the GWAS overlap with the ones from the exome study except for four that lie far down in the rankings.

Table 4 Details of top ranking variants given the Pearson correlation coefficient ranking on the GWAS dataset. See Table 2 for more caption details

dbSNP ID	*Pearson*	*Chr*	*Pos*	*Rank*	*Gene*
rs1951574	0.33	14	100346664	4	EML1
rs1905359	0.33	8	54851272	5	RGS20
rs2792228	0.32	9	6976680	6	KDM4C
rs11011415	0.31	10	38264389	7	ZNF25
rs3900922	0.31	10	21287528	8	NEBL
rs3739714	0.31	9	16435848	9	BNC2
rs9844641	0.31	3	43476335	10	ANO10

4 Discussion

In addition to the results shown here we studied two variations in our machine learning pipeline to see if they would increase prediction accuracy. First we looked at a naive encoding where we convert homozygous alleles to 0 and 2 and the heterozygous to 1. This marginally lowered the accuracy. Second we considered the chi-square ranking of SNPs instead of Pearson correlation and this also marginally lowered the accuracy.

One main challenge in our study is the size of our training set that is considerably smaller than sample sizes (of several thousand) used in GWAS based risk prediction studies. Our primary source of data is the NIH dbGaP repository and so our sample sizes are limited to the data accumulated there.

Another challenge is the quality and coverage of data in dbGaP. For the three external studies we aimed to predict case and control of many samples. Yet for several of the downloaded datasets coverage was insufficient and we found our top ranked variants only in a few samples.

Finally, differences in ancestry can affect risk prediction (Carlson et al., 2013; Pino-Yanes et al., 2015; Freedman et al., 2013). In our case we learnt a model from data obtained in patients at the Dana Farber Cancer Institute in Boston, Massachusetts. In the three external datasets one is of Mexican and Vietnamese ancestry whose genetics are likely to be different from patients at the Dana Farber Institute.

5 Conclusion

Starting from raw exome sequences we obtained a model for predicting chronic lymphocytic leukaemia after a rigorous next generation sequence and machine learning pipeline. We evaluated the model in cross-validation studies as well as on three independent external datasets as part of cross-study validation. In cross-validation we achieve a mean prediction of 82% compared to 57% obtained on an Affymetrix 6.0 panel genome wide association study. In the external cross-study validation we obtain 70% accuracy with a model learnt entirely from the original dataset. Finally we show biological significance of top ranking SNPs in the dataset. Our study shows that even with a small sample size we can obtain moderate to high accuracy with exome sequences and is thus encouraging for future work.

Acknowledgements

We thank Advance Research and Computing Services at NJIT, in particular David Perel, Kevin Walsh, and Gedaliah Wolosh who assisted us with condor and the Kong computing cluster.

References

Abraham, G., Kowalczyk, A., Zobel, J. and Inouye, M. (2013) 'Performance and robustness of penalized and unpenalized methods for genetic prediction of complex human disease', *Genetic Epidemiology*, Vol. 37, No. 2, pp.184–195.

Alpaydin, E. (2004) *Machine Learning*, MIT Press.

Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Angel, G., Levy-Moonshine, A. et al. (2013) 'From Fastq data to high-confidence variant calls: The genome analysis toolkit best practices pipeline', *Current Protocols in Bioinformatics*, pp.11–10.

Banerji, S., Cibulskis, K., Rangel-Escareno, C., Brown, K.K., Carter, S.L., Frederick, A.M. et al. (2012) 'Sequence analysis of mutations and translocations across breast cancer subtypes', *Nature*, Vol. 486, No. 7403, pp.405–409.

Bernau, C., Riester, M., Boulesteix, A., Parmigiani, G., Huttenhower, C., Waldron, L. and Trippa, L. (2014) 'Cross-study validation for the assessment of prediction algorithms', *Bioinformatics*, Vol. 30, No. 12, pp.i105–i112.

Berndt, S.I., Skibola, C.F., Joseph, V., Camp, N.J., Nieters, A., Wang, Z. et al. (2013) 'Genome-wide association study identifies multiple risk loci for chronic lymphocytic leukemia', *Nature Genetics*, Vol. 45, pp.868–876.

Carlson, C.S., Matise, T.C., North, K.E., Haiman, C.A., Fesinmeyer, M.D., Buyske, S. et al. (2013) 'Generalization and dilution of association results from European gwas in populations of non-European ancestry: the page study', *PLoS Biol*, Vol. 11, No. 9, e1001661.

Chatterjee, N., Wheeler, B., Sampson, J., Hartge, P., Chanock, S.J. and Park, J-H. (2013) 'Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies', *Nature Genetics*, Vol. 45, pp.400–405.

Cheung, N., Fung, T.K., Zeisig, B.B., Holmes, K., Rane, J.K., Mowen, K.A. et al. (2016) 'Targeting aberrant epigenetic networks mediated by prmt1 and kdm4c in acute myeloid leukemia', *Cancer Cell*, Vol. 29, No. 1, pp.32–48.

Cortes, C. and Vapnik, V. (1995) 'Support-vector networks', *Machine Learning*, Vol. 20, No. 3, pp.273–297.

De Keersmaecker, K., Graux, C., Odero, M., Mentens, N., Somers, R., Maertens, J. et al. (2005) 'Fusion of eml1 to abl1 in t-cell acute lymphoblastic leukemia with cryptic t(9;14)(q34;q32)', *Blood*, Vol. 105, No. 12, pp.4849–4852.

DePristo, M., Banks, E., Poplin, R., Garimella, K., Maguire, J., Hartl, C. et al. (2011) 'A framework for variation discovery and genotyping using next-generation DNA sequencing data', *Nature Genetics*, Vol. 43, No. 5, pp.491–498.

Di Bernardo, M.C., Crowther-Swanepoel, D., Broderick, P., Webb, E., Sellick, G., Wild, R. Et al. (2008) 'A genome-wide association study identifies six susceptibility loci for chronic lymphocytic leukemia', *Nature Genetics*, Vol. 40, No. 10, pp.1204–1210.

Do, C.B., Hinds, D.A., Francke, U. and Eriksson, N. (2012) 'Comparison of family history and snps for predicting risk of complex disease', *PLoS Genet*, Vol. 8, No. 10, e1002973.

Eleftherohorinou, H., Wright, V., Hoggart, C., Hartikainen, A-L., Jarvelin, M-R., Balding, D. et al. (2009) 'Pathway analysis of gwas provides new insights into genetic susceptibility to 3 inflammatory diseases', *PLoS ONE*, Vol. 4, No. 11, e8068.

Emerenciano, M., Kowarz, E., Karl, K., de Almeida Lopes, B., Scholz, B., Bracharz, S. et al. (2013) 'Functional analysis of the two reciprocal fusion genes mll-nebl and nebl-mll reveal their oncogenic potential', *Cancer Letters*, Vol. 332, No. 1, pp.30–34.

Evans, D.M., Visscher, P.M. and Wray, N.R. (2009) 'Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk', *Human Molecular Genetics*, Vol. 18, No. 18, pp.3525–3531.

Fonseca, N., Rung, J., Brazma, A. and Marioni, J. (2012) 'Tools for mapping high-throughput sequencing data', *Bioinformatics*, Vol. 28, No. 24, pp.3169–3177.

Freedman, B.I., Divers, J. and Palmer, N.D. (2013) 'Population ancestry and genetic risk for diabetes and kidney, cardiovascular, and bone disease: modifiable environmental factors may produce the cures', *American Journal of Kidney Diseases*, Vol. 62, No. 6, pp.1165–1175.

Gail, M.H. (2008) 'Discriminatory accuracy from single-nucleotide polymorphisms in models to predict breast cancer risk', *N Engl J Med*, Vol. 100, No. 14, pp.1037–1041.

Guyon, I. and Elisseeff, A. (2003) 'An introduction to variable and feature selection', *J. Mach. Learn. Res.*, Vol. 3, pp.1157–1182.

Guyon, I., Gunn, S., Ben-Hur, A. and Dror, G. (2004) 'Result analysis of the nips 2003 feature selection challenge', *Advances in Neural Information Processing Systems*, pp.545–552.

Hatem, A., Bozdag, D., Toland, A. and Catalyurek, U. (2013) 'Benchmarking short sequence mapping tools', *BMC Bioinformatics*, Vol. 14, No. 1, p.184.

Janssens, A.C.J.W. and van Duijn, C.M. (2008) 'Genome-based prediction of common diseases: advances and prospects', *Human Molecular Genetics*, Vol. 17(R2), pp.R166–R173.

Joachims, T. (1999) 'Making large-scale svm learning practical', in Schölkopf, N., Burges, C. and Smola, A. (Eds): *Advances in Kernel Methods – Support Vector Learning*, MIT Press.

Kathiresan, S., Melander, O., Anevski, D., Guiducci, C., Burtt, N.P., Roos, C. et al. (2008) 'Polymorphisms associated with cholesterol and risk of cardiovascular events', *New England Journal of Medicine*, Vol. 358, pp.1240–1249.

Kim, D., Kwon, N. and Kim, S. (2014) 'Association of aminoacyl-trna synthetases with cancer', in Kim, S. (Ed): *Aminoacyl-tRNA Synthetases in Biology and Medicine*, Vol. 344 of *Topics in Current Chemistry*, Springer Netherlands, pp.207–245.

Kooperberg, C., LeBlanc, M. and Obenchain, V. (2010) 'Risk prediction using genome-wide association studies', *Genetic Epidemiology*, Vol. 34, No. 7, pp.643–652.

Kraft, P. and Hunter, D.J. (2009) 'Genetic risk prediction – are we there yet?' *New England Journal of Medicine*, Vol. 360, No. 17, pp.1701–1703.

Kruppa, J., Ziegler, A. and König, I.R. (2012) 'Risk estimation and risk prediction using machine-learning methods', *Human Genetics*, Vol. 131, No. 10, pp.1639–1654.

Landau, D.A., Carter, S.L., Stojanov, P., McKenna, A., Stevenson, K., Lawrence, M.S. et al. (2013) 'Evolution and impact of subclonal mutations in chronic lymphocytic leukemia', *Cell*, Vol. 152, No. 4, pp.714–726.

Li, H. and Durbin, R. (2009) 'Fast and accurate short read alignment with burrows wheeler transform', *Bioinformatics*, Vol. 25, No. 14, pp.1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N. et al. (2009) 'The sequence alignment map format and SAM tools', *Bioinformatics*, Vol. 25, No. 16, pp.2078–2079.

Lou, L. and Xu, B. (1997) 'Induction of apoptosis of human leukemia cells by α-anordrin', *Chinese Journal of Cancer Research*, Vol. 9, No. 1, pp.1–5.

Mailman, M.D., Feolo, M., Jin, Y., Kimura, M., Tryka, K., Bagoutdinov, R. (2007) 'The ncbi dbgap database of genotypes and phenotypes', *Nature Genetics*, Vol. 39, No. 10, pp.1181–1186.

Manolio, T.A. (2013) 'Bringing genome-wide association findings into clinical use', *Nature Reviews Genetics*, Vol. 14, pp.549–558.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A. et al. (2010) 'The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data', *Genome Research*, Vol. 20, No. 9, pp.1297–1303.

Morrison, A.C., Bare, L.A., Chambless, L.E., Ellis, S.G., Malloy, M., Kane, J.P. et al. (2007) 'Prediction of coronary heart disease risk using a genetic risk score: the atherosclerosis risk in communities study', *Am. J. Epidemiol*, Vol. 166, No. 1, pp.28–35.

Okser, S., Pahikkala, T. and Aittokallio, T. (2013) 'Genetic variants and their interactions in disease risk prediction – machine learning and network perspectives', *BioData Mining*, Vol. 6, No. 1, p.5.

Park, S.G., Schimmel, P. and Kim, S. (2008) 'Aminoacyl trna synthetases and their connections to disease', *Proceedings of the National Academy of Sciences*, Vol. 105, No. 32, pp.11043–11049.

Pasqualucci, L., Khiabanian, H., Fangazio, M., Vasishtha, M., Messina, M., Holmes, A.B. et al. (2014) 'Genetics of follicular lymphoma transformation', *Cell Reports*, Vol. 6, No. 1, pp.130–140.

Paynter, N.P., Chasman, D.I., Buring, J.E., Shiffman, D., Cook, N.R. and Ridker, P.M. (2009) 'Cardiovascular disease risk prediction with and without knowledge of genetic variation at chromosome 9p21.3', *Annals of Internal Medicine*, Vol. 150.

Pino-Yanes, M., Thakur, N., Gignoux, C.R., Galanter, J.M., Roth, L.A., Eng, C. et al. (2015) 'Genetic ancestry influences asthma susceptibility and lung function among latinos', *Journal of Allergy and Clinical Immunology*, Vol. 135, No. 1, pp.228–235.

Roshan, U., Chikkagoudar, S., Wei, Z., Wang, K. and Hakonarson, H. (2011) 'Ranking causal variants and associated regions in genome-wide association studies by the support vector machine and random forest', *Nucleic Acids Research*, Vol. 39, No. 9, e62.

Sandhu, M., Wood, A. and Young, E. (2010) 'Genomic risk prediction', *The Lancet*, Vol. 376, pp.1366–1367.

Schrodi, S.J., Mukherjee, S., Shan, Y., Tromp, G., Sninsky, J.J., Callear, A.P. et al. (2014) 'Genetic-based prediction of disease traits: prediction is very difficult, especially about the future', *Frontiers in Genetics*, Vol. 5, No. 162.

Shanshal, M. and Haddad, R.Y. (2012) 'Chronic lymphocytic leukemia', *Disease-a-Month*, Vol. 58, pp.153–167.

Shigemizu, D., Abe, T., Morizono, T., Johnson, T.A., Boroevich, K.A., Hirakawa, Y. et al. (2014) 'The construction of risk prediction models using gwas data and its application to a type 2 diabetes prospective cohort', *PLoS ONE*, Vol. 9, No. 3, e92549.

Slager, S.L., Rabe, K.G., Achenbach, S.J., Vachon, C.M., Goldin, L.R., Strom, S.S. et al. (2011) 'Genome-wide association study identifies a novel susceptibility locus at 6p21.3 among familial cll', *Blood*, Vol. 117, No. 6, pp.1911–1916.

Smialowski, P., Frishman, D. and Kramer, S. (2010) 'Pitfalls of supervised feature selection', *Bioinformatics*, Vol. 26, No. 3, pp.440–443.

Speedy, H.E., Di Bernardo, M.C., Sava, G.P., Dyer, M.J.S., Holroyd, A., Wang, Y. (2014) 'A genome-wide association study identifies multiple susceptibility loci for chronic lymphocytic leukemia', *Nat Genet*, Vol. 46, pp.56–60.

Stransky, N., Egloff, A.M., Tward, A.D., Kostic, A.D., Cibulskis, K., Sivachenko, A. et al. (2011) 'The mutational landscape of head and neck squamous cell carcinoma', *Science*, Vol. 333, No. 6046, pp.1157–1160.

Visscher, P.M., Brown, M.A., McCarthy, M.I. and Yang, J. (2012) 'Five years of GWAS discovery', *The American Journal of Human Genetics*, Vol. 90, No. 1, pp.7–24.

Wang, K., Li, M. and Hakonarson, H. (2010) 'ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data', *Nucleic Acids Research*, Vol. 38, No. 16, e164.

Wang, L., Lawrence, M.S., Wan, Y., Stojanov, P., Sougnez, C., Stevenson, K. et al. (2011) 'SF3B1 and other novel cancer genes in chronic lymphocytic leukemia', *New England Journal of Medicine*, Vol. 365, No. 26, pp.2497–2506.

Wei, Z., Wang, W., Bradfield, J., Li, J., Cardinale, C., Frackelton, E. et al. (2013) 'Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease', *The American Journal of Human Genetics*, Vol. 92, No. 6, pp.1008–1012.

Welcome Trust Case Control Consortium (2007) 'Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls', *Nature*, Vol. 447, pp.661–678.

Wray, N.R., Goddard, M.E. and Visscher, P.M. (2007) 'Prediction of individual genetic risk to disease from genome-wide association studies', *Genome Research*, Vol. 17, pp.1520–1528.

Wray, N.R., Goddard, M.E. and Visscher, P.M. (2008) 'Prediction of individual genetic risk of complex disease', *Current Opinion in Genetics and Development*, Vol. 18, pp.257–263.

Wu, U., Zhang, X., Liu, Y., Lu, F. and Chen, X. (2016) 'Decreased expression of bnc1 and bnc2 is associated with genetic or epigenetic regulation in hepatocellular carcinoma', *International Journal of Molecular Sciences*, Vol. 17, No. 2, p.153.