ABSTRACT

CANCER RISK PREDICTION WITH NEXT GENERATION SEQUENCING DATA USING MACHINE LEARNING

by
Nihir Patel

The use of computational biology for next generation sequencing (NGS) analysis is rapidly increasing in genomics research. However, the effectiveness of NGS data to predict disease abundance is yet unclear. This research investigates the problem in the whole exome NGS data of the chronic lymphocytic leukemia (CLL) available at dbGaP. Initially, raw reads from samples are aligned to the human reference genome using burrows wheeler aligner. From the samples, structural variants, namely, Single Nucleotide Polymorphism (SNP) and Insertion Deletion (INDEL) are identified and are filtered using SAMtools as well as with Genome Analyzer Tool Kit (GATK). Subsequently, the variants are encoded and feature selection is performed with the Pearson correlation coefficient (PCC) and the chi-square 2-df statistical test. Finally, 90:10 cross validation is performed by applying the support vector machine algorithm on sets of top selected features. It is found that the variants detected with SAMtools and GATK achieve similar prediction accuracies. It is also noted that the features that are ranked with the PCC yield better accuracy than the chi-square test. In all of the analyses, the SNPs are identified to have superior accuracy as compared to the INDELs or the full dataset. Later, an exome capture kit is introduced for analysis. The SNPs, ranked with the PCC, along with the exome capture kit yield prediction accuracy of 85.1% and area under curve of 0.94. Overall, this study shows the effective application of the machine learning methods and the strength of the NGS data for the CLL risk prediction.

CANCER RISK PREDICTION WITH NEXT GENERATION SEQUENCING DATA USING MACHINE LEARNING

by
Nihir Patel

A Thesis
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Master of Science in Bioinformatics

Department of Computer Science

January 2015

APPROVAL PAGE

CANCER RISK PREDICTION WITH NEXT GENERATION SEQUENCING DATA USING MACHINE LEARNING

Nihir Patel

Dr. Usman Roshan, Thesis Advisor — Date
Associate Professor of Computer Science, NJIT

Dr. Jason Wang, Committee Member — Date
Professor of Bioinformatics and Computer Science, NJIT

Dr. Zhi Wei, Committee Member — Date
Associate Professor of Computer Science, NJIT

BIOGRAPHICAL SKETCH

Author: Nihir Patel

Degree: Master of Science

Date: January 2015

Undergraduate and Graduate Education:

- Master of Science in Bioinformatics,
  New Jersey Institute of Technology, Newark, NJ, 2015

- Bachelor of Science in Biotechnology,
  Rutgers University, New Brunswick, NJ, 2013

Major: Bioinformatics

I dedicate this thesis to my beloved family.
For their eternal love, continuous inspiration and unconditional support.

ACKNOWLEDGMENT

I would like to express my deepest gratitude to my thesis advisor, Dr. Usman Roshan for his consistent support, valuable time, exceptional patience and impeccable guidance, not limited to the research but throughout the curriculum. I am really grateful to Dr. Jason Wang and Dr. Zhi Wei for serving in my master's thesis committee as well for their outstanding teaching that shaped my skills and helped me prepare for the research.

I would like to thank Bharati Jadhav for providing excellent guidance to initiate the research and for her frequent valuable feedbacks. I also wish to thank my friend, Akhila Nagula, who was always willing to help and give her best suggestions.

I would also like to thank all my wonderful friends and classmates. Their presence in my life is invaluable and continuously inspires me to keep going in this work.

I am really thankful to NJIT for providing excellent research facilities and Dr. David Perel, Dr. Kevin Walsh, and Dr. Gedaliah Wolosh for assisting with Condor high throughput computing.

Finally, I heartily wish to thank my grandmother, parents and brother for their moral support and for standing by me in my ups and downs. Without that I would never have been able to accomplish this goal.

TABLE OF CONTENTS

Chapter **Page**

TABLE OF CONTENTS
(Continued)

Chapter **Page**

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AUC	Area Under Curve
BAM	Binary version of SAM
bp	Base pairs
CLL	Chronic Lymphocytic Leukemia
DNA	Deoxyribonucleic Acid
GATK	Genome Analyzer Tool Kit
GWAS	Genome Wide Association Studies
NGS	Next Generation Sequencing
SAM	Sequence Alignment/Map
WES	Whole Exome Sequencing

CHAPTER 1

INTRODUCTION

1.1 Objective

Advances in genomic sequencing techniques have opened a wide range of opportunity to observe genetic variations and diseases more closely. Nevertheless, the massive Next Generation Sequencing (NGS) data demanding an intense computing makes it a challenging task to extract meaningful information that can be correlated with the fatal diseases such as the cancer. Many researchers have adapted the supervised machine-learning methods to analyze such enormous data. Since such methods take the advantage of relationally structured biological data and utilize only partial information from the data to learn a model. Ultimately, the model can either be applied to the complete dataset or to any relevant independent dataset to classify sensible biological information.

This thesis seeks to investigate a multiple aspects associated with the cancer genomics and the cancer risk predictions with the aid of the supervised machine learning method. The primary purpose of the analysis is to develop a strategy that can effectively classify chronic lymphocytic leukemia (CLL) subjects into tumor and non-tumor, by applying machine-learning algorithm on the key structural variants. The variants were extracted from the whole exome sequencing (WES) data of the CLL. The study, first, compared the performance of two popular variant calling tools, namely, (1) SAMtools (Li et al., 2009) and (2) Genome Analyzer Tool Kit (McKenna et al., 2010). Afterwards, a novel, genotypes based variant encoding method was introduced and the effectiveness of the method was compared with the previous encoding method used in genome wide association studies (GWAS) (Roshan et al., 2011). In subsequent step, an investigation was carried out to evaluate the performance of two statistical variant ranking strategies, namely, (1) Pearson correlation coefficient (PCC) and (2) chi-square test. In the final phase of the analysis, an exome kit was introduced for variant detection, and the improvement in the classification accuracy was assessed and is presented.

1.2 Background

The Chronic Lymphocytic Leukemia (CLL) is a cancer of white blood cells. According to the Cancer Facts and Figures 2014 (distributed by American Cancer Society), in USA itself, 4600 deaths associated with CLL were reported, and about 15720 new cases of CLL are expected for the year. Existing methods can only identify the CLL after its occurrence, but in many cases it is too late before the disease can be diagnosed. Previously, for Crohn's disease and ulcerative colitis diseases, the predictions with an Area Under Curve (AUC) of 0.86 and an AUC of 0.82 were reported, respectively (Wei et al., 2013). Likewise, an AUC of 0.82 for type-2 diabetes, and an AUC of 0.83 for bipolar disease, using bootstrap method has been previously conveyed (Burton et al., 2010). But yet to date, there is no known effective pre-diagnostic method have been implemented, that can predict the CLL or any other cancers. Traditional methods, such as, microarray expression analysis and Genome Wide Association Studies (GWAS) were unable to produce prediction accuracy significant enough to utilize it for clinical purposes. And hence, it is important to initiate a study for a better risk prediction of the CLL that incorporates a distinct approach. Since, the CLL disease is associated with the genetic mutations, the genetic data obtained from the Next Generation Sequencing (NGS) techniques, may lead to an efficient diagnostics scheme. In recent years, use of the computational science to extract important biological information from the NGS data has been increased, dramatically. This is due to the fact that, the computational methods provide a cost efficient reproducibility of an investigation, which can serve as an alternative to the expensive wet lab experiments. The study seeks to take an advantage of such computational approach to assessed risk prediction accuracy in the CLL by using the WES data, targeted to cover the human exome regions.

CHAPTER 2

METHODS

This study involves a substantial analysis of the WES data, which includes a large number of transitional steps, producing a various types of files by utilizing several publically available tools. Hence, in the following sections, the details of the intermediate files and tools involved in the analysis will thoroughly be discussed, simultaneously, with experimental procedure.

2.1 Datasets

For the analysis, raw short reads sequence data were obtained from the database of Genotypes and Phenotypes (dbGaP) from the study phs000435.v2.p1 (Wang et al., 2011). All the samples were produced to achieve a mean coverage of 140X of human exome regions. It contained 76 Base Pair (bp) long, pair ended, exome data generated using Illumina Genome Analyzer-II and Illumina hiseq 2000. The dataset contains, 355 samples, which includes 186 tumor samples, and 169 non-tumor samples produced from the matched germ line non-cancerous cells of 169 tumor patients (the 169 sample from 186 tumor) (Wang et al., 2011). For this analysis, 153 tumor samples and 144 non-tumor samples were considered. The rest of the samples were excluded from the analysis due to one of three reasons, which includes (1) excessive size (greater than 20 gigabytes) (2) missing data (3) erroneous samples. Additional required material beside the sample data, such as the human exome region coordinates (the exome kit) and publically available standard variant datasets were obtained from GATK bundle (2.8 b37) available through ftp://gsapubftp-anonymous@ftp.broadinstitute.org/bundle (McKenna et al., 2010). The Table 2.1 shows all the files with brief information of content and their roles in Variant Score Quality Recalibration (VSQR). VSQR process is a GATK protocol used in the analysis to filter variants (See Section 2.3). Beside that, the human reference genome (version GRCh37.p13) for mapping reads was obtained from the Genome Reference Consortium accessible at http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/.

Table 2.1 Resources Used in GATK for Variant Filtration

Resource Used	Description	Role in VSQR process
Exome intervals	Contains putative exome region coordinates	N/A
dbSNP variants	SNPs found in dbSNP databases	Contains Known sites and not used as training resource
HapMap	HapMap genotypes and VCFs sites for SNPs	Contains True sites and used as training resource to filter SNPs
OMNI 2.5 Genotypes	OMNI 2.5 genotypes for 1000 Genomes samples and SNPs VCF sites	Contains True sites and used as training resource to filter SNPs
1000 Genome Phase-I SNPs	1000 Genomes Phase I SNPs calls	Contains Non-True sites and used as training resource to filter SNPs
Mills_and_1000G _gold_standards	INDELs calls validated with high degree of confidence	Known and True sites and used as training resource to filter INDELs

Source: (Auwera et al., 2013, DePristo et al., 2011 and McKenna et al., 2010)

2.2 Data Pre-Processing

Figure 2.1 represents, the experimental steps followed to obtain variants from the raw reads. For the major pre-processing, SAMtools (version 0.1.18) was considered. SAMtools comprised of multiple utilities, to perform sorting, merging, indexing and filtering of the sequencing data (Li et al., 2009). The samples were received as Sequence Archive Read (SRA) format files. The SRA files were converted to the fastq format using the fastq-dump utility of SRAtoolkit provided by the National Institute of Health (NIH) and is accessible at http://eutils.ncbi.nih.gov/Traces/sra/?view=software. The fastq format comprise of the reads containing nucleotides and the quality scores associated with the reads. Subsequently, the reads were mapped against the human reference genome using BWA-MEM (version 0.7a-r405) by applying all default parameters (Li & Durbin, 2009). BWA is a popular mapping tool that implements the Burrows-Wheeler transform algorithm. BWA was chosen as it can align the sample reads to the massive human reference genome quickly and efficiently (Fonseca et al., 2012 and Hatem et al., 2013). As shown in Figure 2.1, the BWA constructs output files in the Sequence Alignment/Map (SAM) format. Using SAMtools, the SAM files were converted to Binary SAM (BAM) files and then the files were indexed and sorted. Such BAM files are consist of reads alignments between sample sequence and reference sequence. Indexing and sorting procedures, allows a quick access of the massive alignment data within BAM files. An AddOrReplaceReadGroups utility of PICARD tool (Version 1.8), accessible at http://broadinstitute.github.io/picard/, was then used to add read group information to the BAM files. Afterwards, using SAMtools, the unmapped reads and the reads with mapping quality score (MAPQ) below 15 were eliminated. The resulting filtered BAM files were then sorted and indexed. Thereafter, the duplicate reads were removed from the filtered BAM files using the MarkDuplicates utility of PICARD tool. Both of the above steps are important, as they remove the low quality reads and the duplicate reads that largely contributes to the false positive variants.

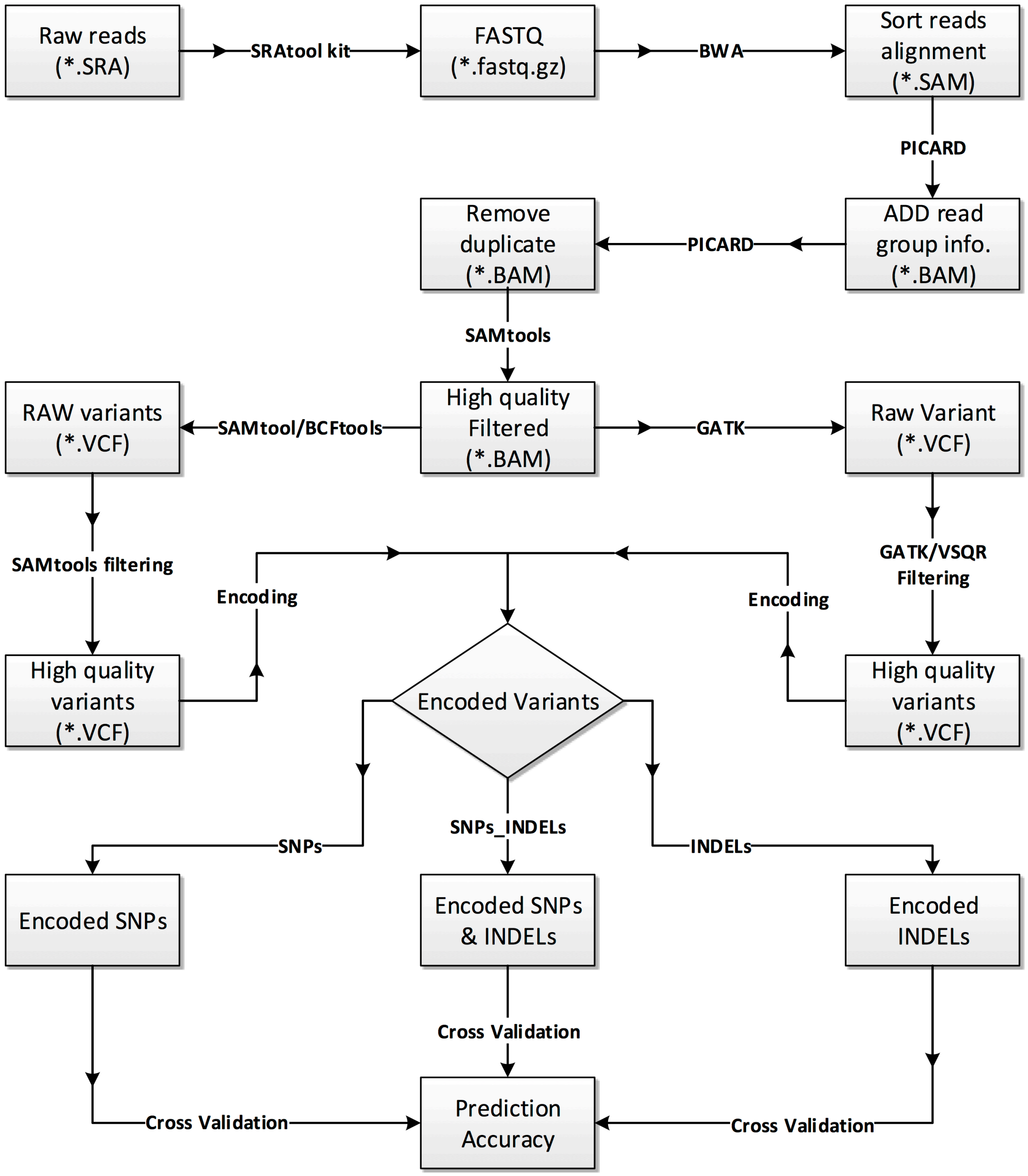


Figure 2.1 Flow chart showing experimental steps.

2.3 Variant Calling and Filtering

The variants were detected jointly, using consensus calling for 297 samples with SAMtools as well as with GATK. Initially, the raw variants were generated using

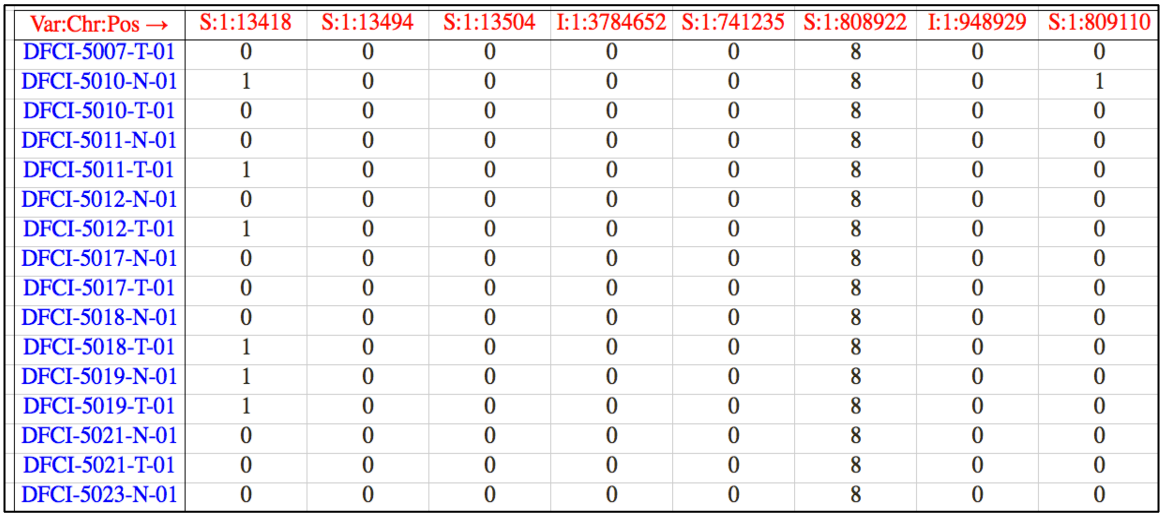
SAMtools together with BCFtools. The process yielded a Variant Calling Format (VCF) file containing raw variants. The VCF file includes two types of structural variants, namely, Single Nucleotide Polymorphism (SNP) and Insertion Deletion (INDEL). The specifications for the VCF format are provided at http://samtools.github.io/hts-specs/VCFv4.1.pdf. Likewise, a raw VCF file was also generated using GATK, following the best practice guidelines provided by the Broad Institute (Auwera et al., 2013). Even though SAMtools and GATK serve a common purpose of calling variants, they both follow distinct steps when it comes to filtering variants. For the variants obtained with SAMtools, filtering was simply done by applying all the default parameters of "vcfutil.pl varFilter" utility provided under the SAMtools. On the other hand, to filter the variant attained with GATK, the Variant Score Quality Recalibration (VSQR) protocol was executed. The VSQR procedure involves two steps, (1) variant recalibration step and (2) apply recalibration step. Briefly, the first step generates a Gaussian mixture model using true sites from the datasets discussed in the Table 2.1 and outputs a recalibration file (DePristo et al., 2010). In the second step, the model created in the previous step is applied to the variants in VCF files and the variants are collected into a new VCF file, with a VQSLOD scores added to them. For a given variant, the VQSLOD score is the log odd ratio for the variant to be true versus it to be false (DePristo et al., 2010). Subsequently, a filtering threshold, namely, 'tranche sensitivity' is applied to the variants (DePristo et al., 2010). If the tranche sensitivity threshold is X%, then GATK considers, the VQSLOD score of X% of the variants from training set and calculates the VQSLOD threshold. If a given variant has the VQSLOD score above the threshold, it is considered a true variant and flagged as PASS in QUAL field (in VCF files). In contrast each variant with VQSLOD score below the threshold is treated as false positive (DePristo et al., 2010). Using the VQSLQD threshold, the low quality variants were flagged, and then they were removed using SelectVariants utility of GATK. The high quality variants were collected in to yet another VCF file and the file was used in succeeding analyses. Herein, tranche sensitivity threshold was kept 99.9% for SNPs and 99.0% for INDELs.

2.4 Variant Encoding and Depth Filtering

Since the input matrix of the supervised machine learning methods (see Section 2.5) must be in the form of feature vectors, each variant was encoded into an integer using corresponding genotypes from the VCF files and a feature vector was generated for each individual sample. In the VCF files, the genotypes are assigned as X/Y format. If X (or Y) is reference allele, then it is always represented using 0. The other representations of X and Y varies and the representing number can go as high as the maximum number of the alternative alleles allowed by the variant detector tool. For SAMtools/BCFtools default number of maximum alternative alleles is 2 and for GATK the number is 6.

Variant encoding was done using two different methods. Let the first encoding method be P and the second method be Q. For the method P, the zygosity of the genotype was considered, where a variant was represented as 0 if it is a homozygous reference allele (i.e. 0/0), 1 if it is a heterozygous allele (i.e. 0/1, 0/2, etc.), and 2 if it is a homozygous alternate allele (i.e. 1/1, 2/2, 3/3 etc.). For the method Q, the variants were simply encoded using 7(X) + Y. The encoding was done such that each genotype is mapped to a distinct integer. Using both the methods, the variants were encoded into a data matrix. A fraction of the resulting data matrix is shown in Figure 2.4, where the sample's IDs are shown in the first column and the feature (variant) names are indicated in the first row. The feature names are represented with three parameters (1) variant type where 'S' represents SNPs and 'I' represents INDELs (2) chromosome number and (3) position of variant in the chromosome. The rows of the matrix correspond to feature vectors and the numbers in the data matrix resembles the encoded genotype.

While encoding the variants, the read depth (DP in INFO field) filtering was executed, and the variants holding DP score above 300 and below two were filtered out. These variants come from the distribution that has an average coverage of 140X. Thus, it is very implausible to have a true variants containing DP above twice the size of the coverage (i.e. 300). Alternatively, a variant with the DP score below two is very unlikely to be a true variant because it does not have enough supporting reads. Both of these types of variants have quite high chances of being false positive and hence it is valid to exclude the variants from the analyses.



Var:Chr:Pos →	S:1:13418	S:1:13494	S:1:13504	I:1:3784652	S:1:741235	S:1:808922	I:1:948929	S:1:809110
DFCI-5007-T-01	0	0	0	0	0	8	0	0
DFCI-5010-N-01	1	0	0	0	0	8	0	1
DFCI-5010-T-01	0	0	0	0	0	8	0	0
DFCI-5011-N-01	0	0	0	0	0	8	0	0
DFCI-5011-T-01	1	0	0	0	0	8	0	0
DFCI-5012-N-01	0	0	0	0	0	8	0	0
DFCI-5012-T-01	1	0	0	0	0	8	0	0
DFCI-5017-N-01	0	0	0	0	0	8	0	0
DFCI-5017-T-01	0	0	0	0	0	8	0	0
DFCI-5018-N-01	0	0	0	0	0	8	0	0
DFCI-5018-T-01	1	0	0	0	0	8	0	0
DFCI-5019-N-01	1	0	0	0	0	8	0	0
DFCI-5019-T-01	1	0	0	0	0	8	0	0
DFCI-5021-N-01	0	0	0	0	0	8	0	0
DFCI-5021-T-01	0	0	0	0	0	8	0	0
DFCI-5023-N-01	0	0	0	0	0	8	0	0

Figure 2.2 Encoded data matrix showing distribution of samples and features.

2.5 Supervised Machine Learning and Feature Selection

Following the variant identification and the encoding procedure, a supervised machine learning analysis was performed. Supervised machine learning is a popular approach that utilizes labeled (classified) data to learn a model, and predicts the labels of unclassified data by applying the model. In the supervised method, the sample rows of a data matrix represents feature vectors in a space dimensions given by features from the columns (See Figure 2.4). Here, in the analysis, the data were separated into train and validation sets, so that the supervised model can be learned using the train data and the predictions can be made on the validation data using the model.

Cohort analysis, such as the one performed here yields a very large number of features (variants), which reduces the classifier's ability to separate data. Hence, it is required to extract few significant features that aid to classify data more efficiently. Thus top-K features were selected by arranging them in the decreasing order of the absolute values of the Pearson correlation coefficient (PCC) (Guyon et al., 2003). The K was incremented by 10, up to 100 features, and from there it was incremented by 100, for maximum of 1000 features. For j^{th} variant, PCC_j can be represented as equation 2.1, where the $X_{i,j}$ is the encoded value of the genotype for the i^{th} sample and the j^{th} variant, and the Y_i is the label for the i^{th} sample. The tumor (case) and non-tumor (control) samples were labeled as -1 and 1, respectively.

$$PCC_j = \frac{\sum_{\mathrm{i}}^{\mathrm{n}}\left(X_{i,j} - \bar{X}_i\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{\mathrm{i}}^{\mathrm{n}}\left(X_{i,j} - \bar{X}_i\right)^2} \times \sqrt{\sum_{i}^{\mathrm{n}}\left(Y_i - \bar{Y}\right)^2}} \qquad (2.1)$$

For comparison purposes, in some analyses, a chi-square df-2 test was considered for the feature selection procedure. For the chi-square test, top-K feature sets were extracted by arranging variants in the decreasing order of chi-square P-values.

2.6 Cross-Validation and Accuracy Assessment

For cross-validation, the binary classification approach was considered. At first, the data matrix (discussed in Section 2.4) was randomly separated by rows (subjects), into 90 % training and 10% validation set. Secondly, the feature selection was performed onto training set and the top features were extracted as explained in the Section 2.5. For each feature set, a supervised learning model was created using the support vector machine (SVM) algorithm (Cortes et al., 1995), implemented in the SVM-light program (Joachims, 1999). The model created in training phase was applied to the validation set and the predictions were made. All the above steps were repeated for the multiple random splits and the average classification accuracy was assessed using 1- balanced error rate (BER) (Guyon et al., 2004). As shown in equation 2.2 the BER is an average of the error rate of the controls and the error rate of the cases. Herein, the error rate is calculated by dividing misclassified labels with true labels.

$$BER = \frac{1}{2} \times \left(\frac{Misclassified\ labels_{control}}{True\ labels_{control}}\right) + \left(\frac{Misclassified\ labels_{cases}}{True\ labels_{cases}}\right) \qquad (2.2)$$

CHAPTER 3

RESULTS AND DISCUSSION

Throughout the study, multiple methods (or tools) were used to perform identical analysis. In the succeeding sections, the experimental outputs and the performance of each method (or tool) will be addressed.

3.1 Extracting Analysis Ready Sample

During the data pre-processing, 23 samples were removed from the analysis as discussed in the Section 2.1, which left 332 samples for the remaining analyses. Initially, the variants were detected with SAMtools using the 332 samples, without using any exome kit. Subsequently, the variants were filtered and encoded into the data matrix. The resulting data matrix found to have only 6000 variants. Without the filtering step, the data matrix would have had 22 million variants. Thus, considering this suspiciously low number of variants, the data matrix was re-analyzed, which revealed that all the 6000 identified variants are resided only in the first chromosome, and the rest of the variants were removed, since, they were absented (or have low coverage) in at least one sample. Meaning if a variant is missing (or have low coverage), even in one sample out of 332, then the variant will be removed. Presence of all variants in single chromosome, suggests that, it is a sequencing artifacts, which likely occurs due to systematic bias in sequencing. Therefore to avoid the bias, samples missing an excessive number of variants were identified. To do that, at first, the variants present in a large number of samples were extracted, followed by the identification of the samples, missing the large number of variants. More precisely, high occurring variants that present in at least 330 samples (90% of samples) were extracted. After that, the samples, missing more than 10% of the variants were identified. The procedure identified 35 more samples that need to be removed from the investigation. In the later steps, the same procedure was repeated with GATK while including the exome kit in the analysis. The method identified the same 35 samples for both SAMtools and GATK, and hence mutually validating the sample removal procedure.

3.2 SAMtools vs. GATK

After removing the samples, all the steps discussed in the Sections 2.3 to 2.6 were repeated using both SAMtools and GATK, without providing an exome kit and by applying the method P for encoding (Section 2.5). Figure 3.2 compares an average cross-validation accuracy of 10 random splits between SAMtools and GATK. Herein, the feature selection (Section 2.5) was performed using the PCC and the SNPs were encoded using the method P. As it can be seen from in the Figure 3.1, almost in all cases, GATK performed better. Even though GATK clearly yielded a better prediction accuracy, the comparison between SAMtools and GATK in not quite rational as both uses distinct types of variant filtering strategy. Nonetheless, the comparison can provides a basic idea of how well two popular variant detecting tools perform. The further analyses were performed using only GATK because (1) SAMtools tends to assign random genotypes values to the variants when they have zero DP value. Ideally, such variants indicate missing information and should be avoided from the analysis. On the other hand, GATK efficiently identifies such variants and assigns ('./.') to their genotypes. (2) The VSQR procedure of GATK utilize high confidence known variants calls and provides a single VQSLOD score for filtering, whereas with SAMtools user have to define multiple filtering parameters, and tuning such parameters is a challenging task for such a massive study. (3) SAMtools lacks the variant annotation function provided by GATK.

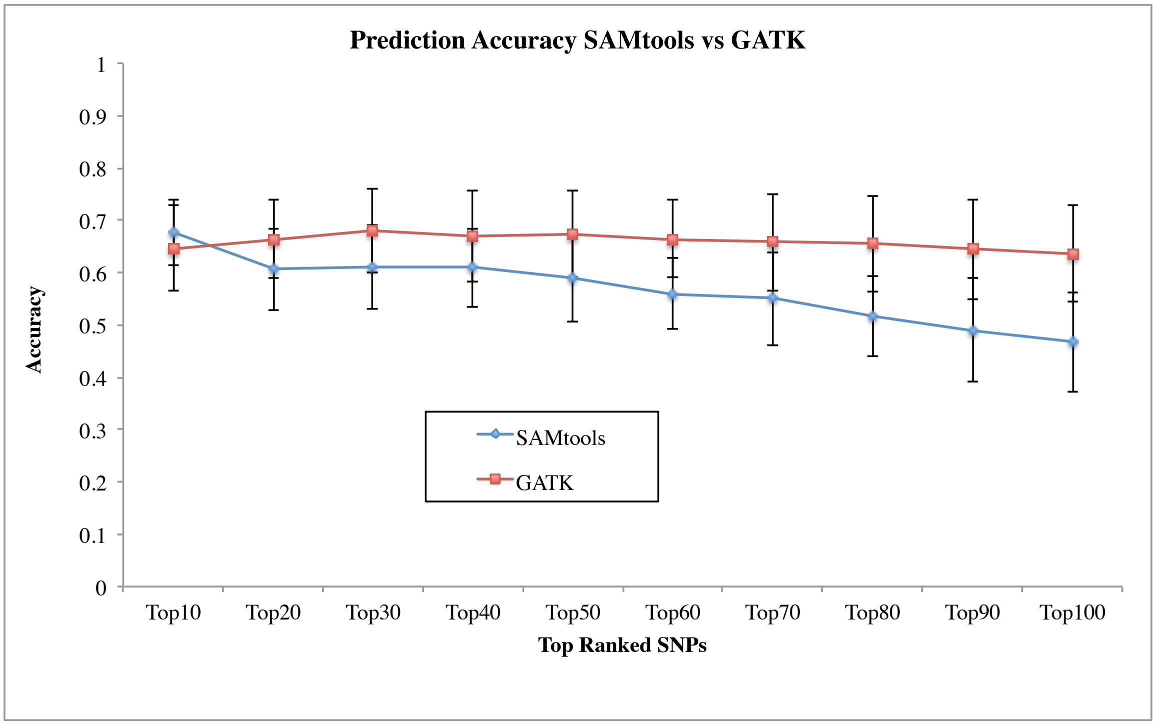


Figure 3.1 Average cross-validation accuracy comparison of SAMtools and GATK over 10 random 90:10 training validation splits. The SNPs were ranked with the PCC and were encoded with the method P. The Error bars represents standard deviations.

3.3 Comparison of Feature Selection Methods

As discussed in the Section 3.2, the remaining experiments were carried out with GATK. At first, the effect of two different types of feature ranking methods (1) PCC and (2) chi-square was analyzed by comparing the classification accuracy on all features (SNPs + INDELs) with 100 random splits. Exome kit was excluded for this procedure, which yielded 296860 features, which were then encoded with the method P (encoding with 0, 1 and 2). Here, only the method P was considered, because it was not plausible to apply chi-square df-2 test to the variants encoded with the method Q. The method Q produces the data matrix with integers ranging from 0 to 48, and that too with the columns (of the encoded matrix) containing varying sets of integers. Thus, in the situation, if one wants to apply the chi-square test, then for each column, an appropriate degree of freedom must be calculate separately, which is a quite complicated task especially, for a large data matrix such as the one used here.

For statistical consistency, the labels for each of 100 splits were kept same, meaning for an individual experiment, train set and validation set were identical for both the PCC and the chi-square ranking. Figure 3.2 shows a comparison of these two ranking methods. For the PCC ranking, top-30 variants yielded the highest accuracy of 66.4%, and for the chi-square ranking, top-20 variants achieved the highest accuracy of 65.1%. As it can be seen from Figure 3.2, there is a negligible difference in the classification accuracy between two methods for first 100 variants, but as the number of variants increases, the PCC constitutively performs better then the chi-square. Therefore, based on the preliminary analysis, only the PCC was used for the feature selection purposes in the remainder of the studies.

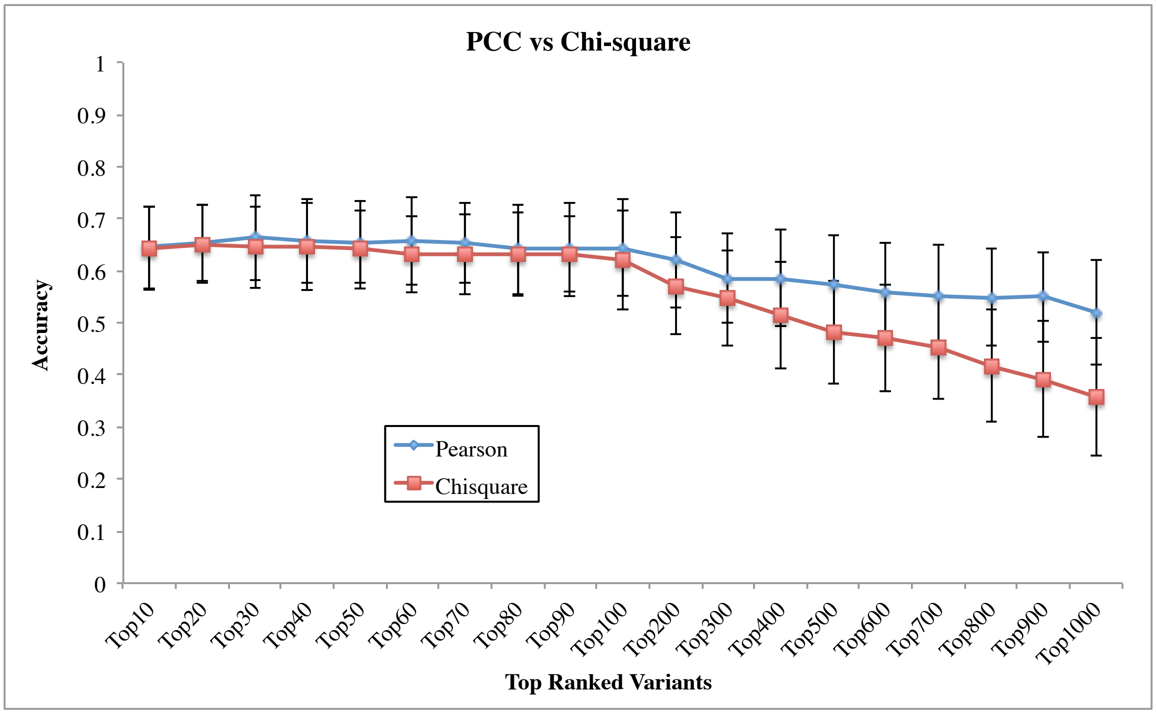
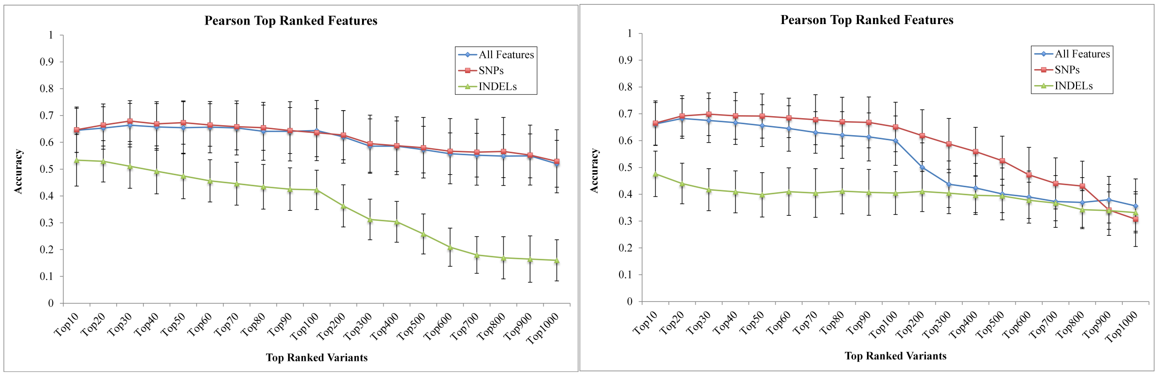


Figure 3.2 Average cross-validation accuracy comparison of top PCC and chi-square ranked features on 100 random 90:10 training validation splits with GATK. The variants were encoded using the method P and the error bars represents standard deviations.

3.4 Comparison of Encoding Methods

Here, in this section, the performances of two different encoding methods, discussed in Section 2.4, are compared. It was not feasible to encode all the variants using the method P and thus, the method Q had a higher number of variants. The variants missed in the method P, are the variants that contains genotypes such as 1/2, 2/3, 3/4 etc. In practice, zygosity of such genotypes cannot be correctly inferred into three categories discussed in Section 2.4, and that is why, such variants were excluded from the analysis. On the other hand, the method Q (encoding using 7X + Y) is designed to include all variants. In fact the sole purpose of introducing a novel encoding method (the method Q) was to avoid the loss of high quality variants that were already passed through the rigorous filtering. As expected, inclusion of all variant did make difference in the classification accuracy. As it can be seen in Figure 3.3, when the classification accuracy was compared, the method Q performed better than the method P, with SNPs alone and also with all variants but with the INDELs, the method P yielded higher accuracy. However for the INDELs, the classification accuracy with both the methods was close to the random guess and hence the performance of the methods was not evaluated based on results obtained with the INDELs. For further studies, the Method Q was considered, because overall it performed better then the method P. Table 3.1 shows the number of encoded genotypes with each method and the highest accuracies associated with it.



Method P Method Q

Figure 3.3 Average cross-validation accuracy with top PCC ranked features using the encoding method P and Q, on 100 90:10 training validation splits. Error bars represents standard deviations.

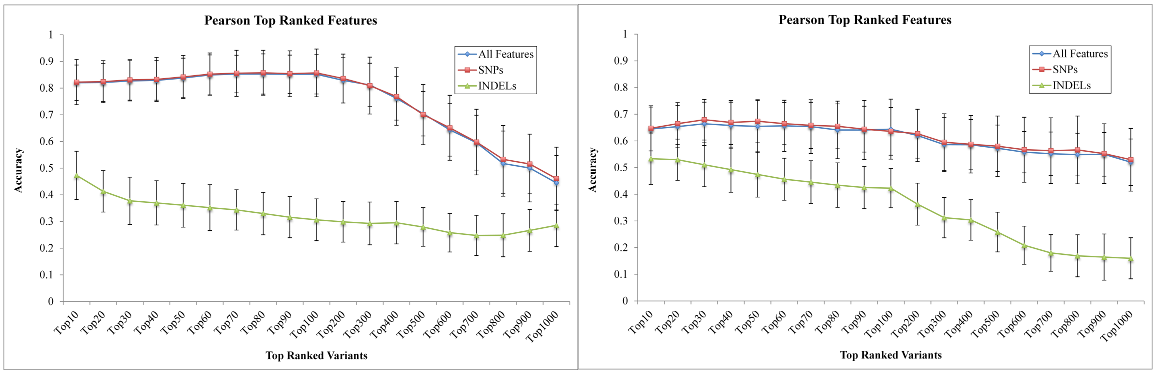
Table 3.1 Numbers of Variants and Obtained Highest Accuracies

Variant Type	Method P		Method Q	
	Number of Variants	Highest Accuracy	Number of Variants	Highest Accuracy
All Features	296860	66.4 % (Top-30)	297924	68.3 % (Top-20)
SNPs	284468	67.9 % (Top-30)	284591	69.9 % (Top-20)
INDELs	12153	53.6 % (Top-10)	13333	47.6 % (Top-10)

3.5 Effect of Capture Kit

Previously, all of analyses were done without any exome kit. The exome kit restricts variant identifications to the only genomic regions that can be encoded into genes. The primary reason to exclude the exome kit from the analysis was to include all putative variants placed near the exonic regions, which otherwise, would not have been taken into consideration. But the filtering recommendation provided by GATK were specifically designed for exome specific regions and it was not clear how effective they will be, if the exome kit is excluded from the analysis. Also, previously, no known study found that can provide guidance for proper filtering parameters, and it was beyond the scope of this investigation to tune all the parameters and then to choose the best one. Hence, it was decided to repeat previous experiments including an exome kit. For the analysis with the exome kit, the variants were identified using GATK and the encoding was done using the method Q (Section 2.4). For the features selection, the PCC was considered. Figure 3.4 compares an average classification accuracy of with and without the use of an exome kit. Unexpectedly, inclusions of the exome kit yielded a quite higher classification accuracy of 85.1%, as compare to 69.9% obtained without the exome kit. With exome kit, 122392 SNPs and 2200 INDELs were identified. As it can be seen in Figure 3.4 that the exome kit analysis achieved significantly higher accuracy with all three forms of datasets. In successive steps, area under curve (AUC) value was calculated for various numbers of top-ranked variants. The AUC values obtained with and without the exome kit were compared, and are reported in Figure 3.5. The AUC values were only calculated for the SNPs because the SNPs gave the highest classification accuracy, regardless the use of an exome kit. The highest AUC observed was 0.94 (Top70 SNPs) with the exome kit and

0.75 (Top40 SNPs) without the exome kit. To further verify the consistency of the results, the classification accuracy for the encoding method P was also assessed, the results are shown in Figure 3.6. Even with the encoding method P, the exome kit yielded quite higher accuracy of 84.7% (Top90) as compare to 67.9% (Top-30) obtained without the exome kit. With the exome kit, just as the previous results (Figure 3.3), the method Q achieved a slightly higher accuracy (85.1%) than the method P (84.7%). One more experiment was performed, where data were split into 50:50 training validation sets, instead of splitting into 90:10. For the data, feature selection was performed using the PCC method and the encoding was done using the method Q. The average classification accuracy with the SNPs alone, with the INDELs alone and the full dataset, are shown in Figure 3.7. The highest accuracy of 82.2% was observed with the SNPs alone as well as with the full dataset. As it can be seen in Figure 3.7 (left), the accuracy curves for the SNPs and the full dataset almost exactly overlaps, which suggests that INDELs have negligible effect on the classification.



With exome kit

Without exome kit

Figure 3.4 Average cross-validation accuracy comparison between with and without the exome kit on 100 90:10 training validation splits. The variants were ranked using PCC and were encoded using the method Q. The error bars represents standard deviations.

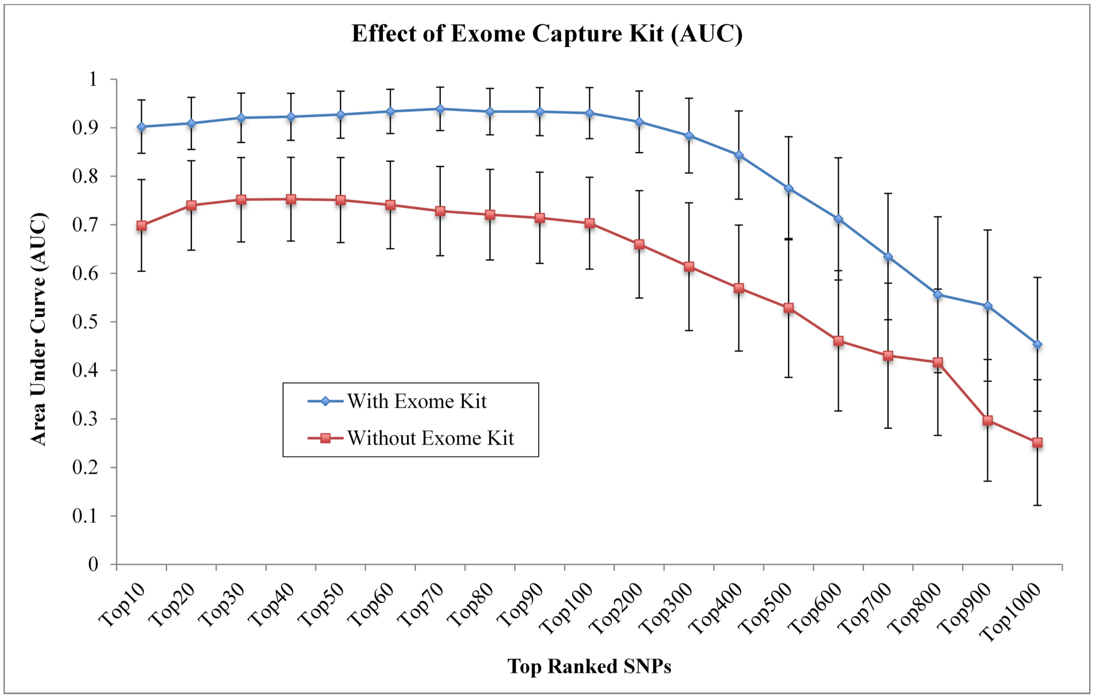


Figure 3.5 Area under curve values for the Q encoding and top PCC ranked SNPs, obtained with and without the exome kit. Error bars represents standard deviations.

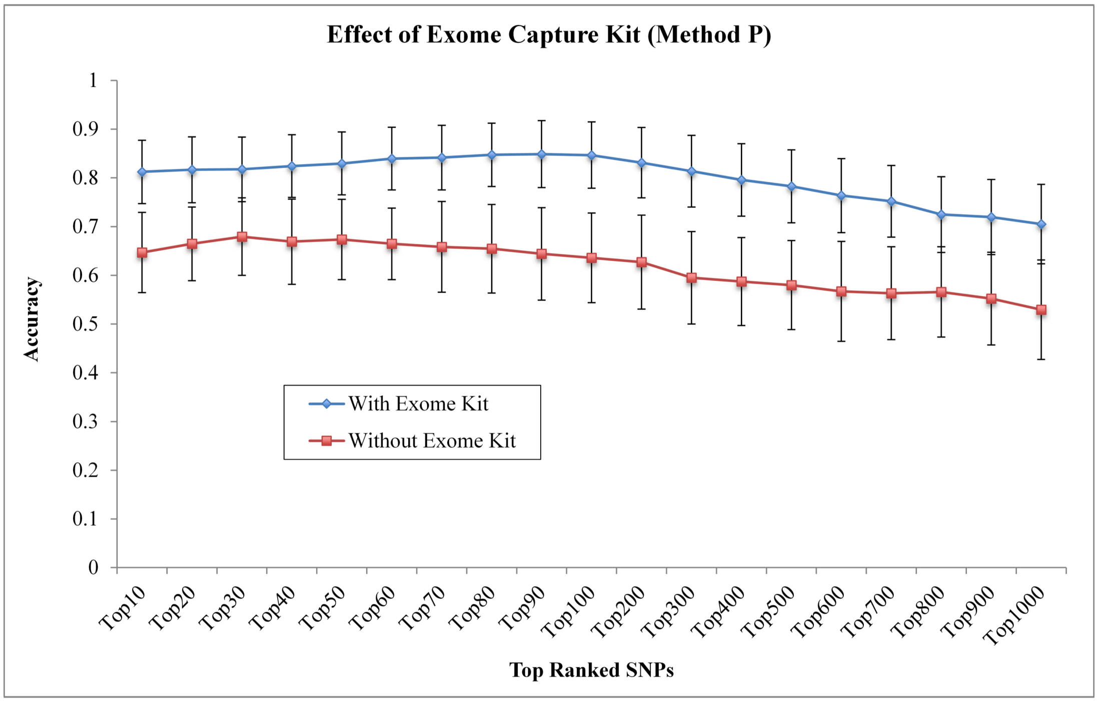


Figure 3.6 Average cross-validation accuracy for the encoding P and top PCC ranked SNPs obtained with and without the exome kit on 100 90:10 training validation splits. The error bars represents standard deviations.

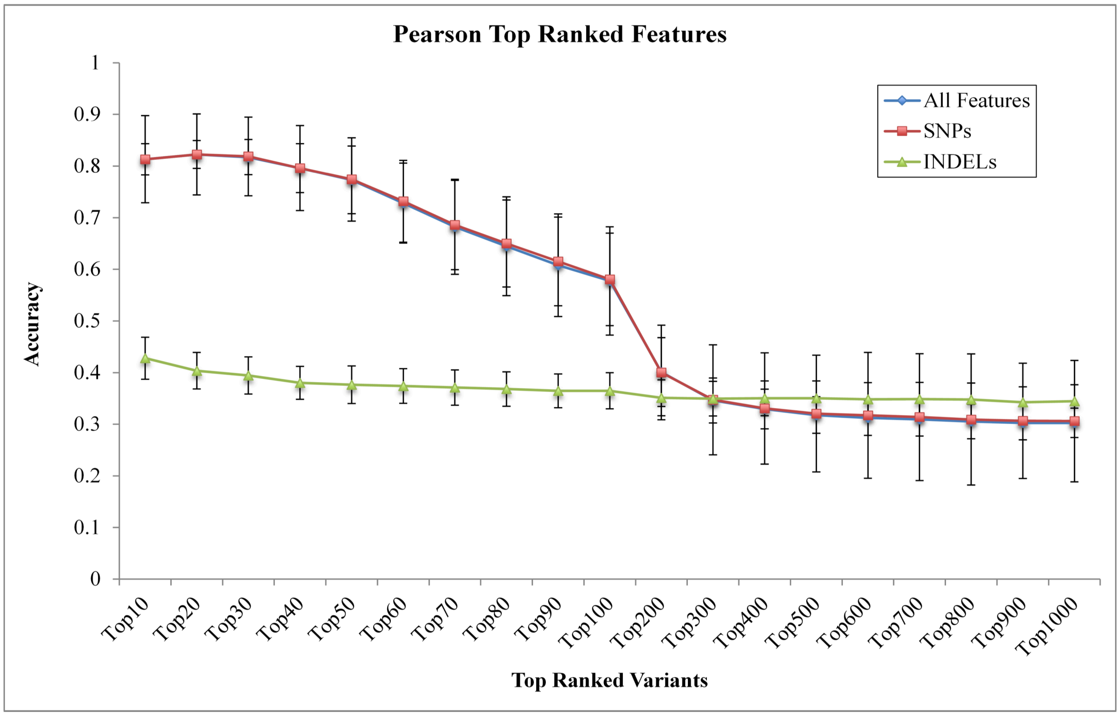
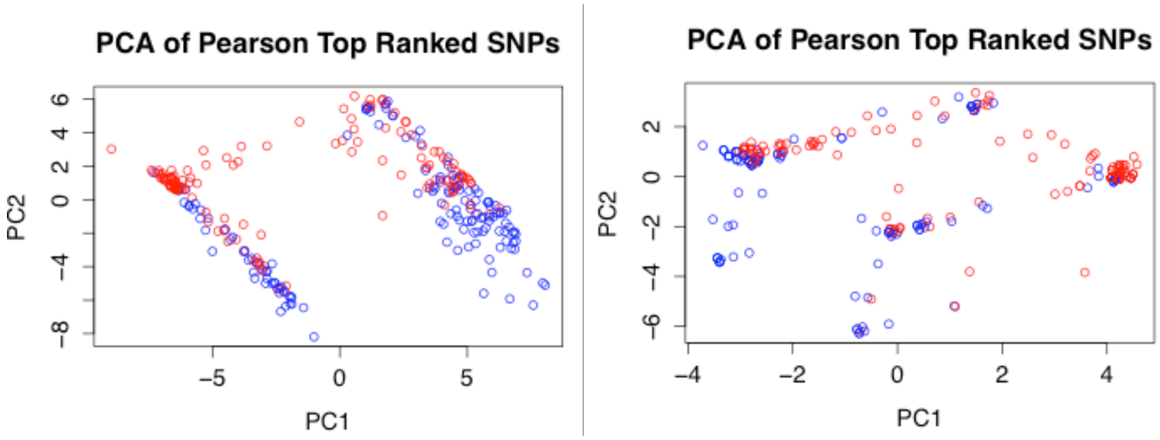


Figure 3.7 Average cross-validation accuracy with top PCC ranked features with exome kit on 100 50:50 training validation splits. Encoding was done using method Q and the error bars represents standard deviations.

3.6 Principle Component Analysis

Two-dimensional plots of top-80 (left) and top-30 (right) SNPs are shown in Figure 3.8. The plots were obtained with principle component analysis (PCA) (Alpaydin, 2004), where X-axis represents first principle component and Y-axis represents second principle component. PCA is used to project the high dimensional data to the lower dimensions so that the data separation can visually be observed. As it can be seen in Figure 3.8, in both experiments, the data are not clearly separated, which suggests that the data are challenging, and hence, supporting the choice to use supervised leaning approach for the data classification. Even though, the data are not quite separated, the data obtained with the exome kit shows a better separation as compare to the data obtained without the exome kit (Figure 3.8).



With Exome Kit (Top-80)

Without Exome Kit (Top-30)

Figure 3.8 PCA plots for top-80 SNPs obtained with exome kit and top-30 SNPs obtained without exome kit.

3.7 Predictive SNPs

With the exome kit, top-80 SNPs, and without the exome kit, top-30 SNPs yielded the highest average classification accuracy over 100 random splits. An intersection of top-80 SNPs (with the exome kit) across the 100 sets yielded 48 common SNPs. Likewise, An intersection of top-30 SNPs (without the exome kit) across the 100 sets yielded nine common SNPs. The SNPs were tagged as predictive SNPs. The 48 predictive SNPs obtained with the exome kit, achieved an average AUC of 0.93 with standard deviation of 0.04, whereas the nine predictive SNPs obtained without the exome kit, yielded an average AUC of 0.72 with standard deviation of 0.09. Additional information regarding these predictive SNPs was obtained using a tool called wANNOVAR (Chang & Wang, 2012). Table 3.1 and Table 3.2 provides the additional details regarding the predictive SNPs identified with the exome kit, in only chromosome 14, and Table 3.2, shows the same for the rest of the chromosomes. Similarly, Table 3.3 shows the additional information about the nine predictive SNPs identified without the exome kit. As it can be seen from the Table 3.1 and 3.2, that many predictive SNPs were found in exonic regions with few exceptions. Hypothetically, if the exome intervals are used to call variants, then all the variants should only be in the exonic regions but this was not observed here. The phenomenon can be explained by the fact, that the exome kit are designed to cover some extra areas on the both end of putative exonic regions, this is done purposely to include flanking regions in the end of the associated gene. It is likely that the non-exonic predictive SNPs belong to the flanking regions. If the exome kit is not provided, then the SNPs can be found in any regions, and that explains why there are many SNPs in the non-exonic regions in Table 3.3 but not in Table 3.1 or Table 3.2. In all three tables the last two columns were added manually along with the information obtained with wANNOVAR. From the two columns, the first column contains number of case subjects containing the particular mutation and the second column represents the same for the control subjects. Theoretically, a cancer leads to the genetic alterations, which suggests that the mutations shown in Table 3.1–3.3, should have high occurrence in the case subjects as compare to the control subjects. But, surprisingly, the higher mutation rate was observed in control subjects. Previously, a study have identified that non-mutated IGHV gene is associated with more aggressive form of the CLL (Ferrer et al., 2004). Also, the gene was used to measure the CLL progression, based on the gene's non-mutated status (Rassenti et al., 2004). These studies suggest that non-mutated IGHV (chromosome 14) may contribute to the CLL, and thus explaining the higher mutation rate in control samples of the predictive SNPs.

Table 3.2 Additional Information of Predictive SNPs in Chromosome 14 (With the Exome kit)

Chr:Pos	Ref	Alt	dbSNP ID	Region	Gene	Case	Control
14:106494153	T	C	.	exonic	IGHV2-5	42	1
14:106494221	T	A	.	exonic	IGHV2-5	30	1
14:106733287	A	C	.	exonic	IGHV1-24	26	5
14:106733289	C	A	.	exonic	IGHV1-24	25	4
14:106733290	C	G	.	exonic	IGHV1-24	24	4
14:107034846	T	G	rs199610746	exonic	IGHV5-51	68	97
14:107034863	C	T	rs199809351	exonic	IGHV5-51	68	97
14:107034873	G	C	rs199524561	exonic	IGHV5-51	59	87
14:107034967	T	C	rs72686844	exonic	IGHV5-51	70	98
14:107113763	A	G	rs377318229	exonic	IGHV3-64	45	81
14:107113780	G	A	rs200164853	exonic	IGHV3-64	61	93
14:107113785	C	T	rs201264785	exonic	IGHV3-64	61	96
14:107113855	A	G	rs111637096	exonic	IGHV3-64	68	97
14:107113858	A	G	rs111853090	exonic	IGHV3-64	68	97
14:107113968	C	A	rs113324720	exonic	IGHV3-64	65	97
14:107179022	C	A/G/T	rs2157615	exonic	IGHV2-70	21	0
14:107282791	A	C	.	downstream	IGHV7-81	6	39
14:107282809	T	C	rs201928713	exonic	IGHV7-81	10	59
14:107282813	C	A	rs199801132	exonic	IGHV7-81	11	56
14:107282814	A	G	rs200749603	exonic	IGHV7-81	11	55
14:107282836	T	A	rs201902530	exonic	IGHV7-81	30	70
14:107282846	G	A	rs200859769	exonic	IGHV7-81	51	100
14:107282852	T	C	rs201336503	exonic	IGHV7-81	54	104
14:107282859	G	A	rs201095197	exonic	IGHV7-81	60	108
14:107282872	T	A	rs61741319	exonic	IGHV7-81	62	109
14:107282909	T	C	rs202202987	exonic	IGHV7-81	65	112
14:107282926	A	T	rs149038822	exonic	IGHV7-81	57	111
14:107282935	A	T	rs201762529	exonic	IGHV7-81	55	107
14:107282973	G	C	rs200848671	exonic	IGHV7-81	49	95
14:107282988	C	G	.	exonic	IGHV7-81	16	54

Source: (Chang & Wang, 2012)

Table 3.3 Additional Information of Predictive SNPs in All Chromosomes except Chromosome 14 (With the Exome kit)

Chr:Pos	Ref	Alt	dbSNP ID	Region	Gene	Case	Control
2:90139116	G	A	rs201820003	exonic	IGKV1D-16	86	136
2:169780261	G	A	.	exonic	ABCB11	0	29
2:169780287	T	A	.	exonic	ABCB11	0	13
6:30553070	G	C	.	exonic	ABCF1	0	15
6:30553073	T	C	.	exonic	ABCF1	0	16
6:31749930	C	G	.	exonic	VARS	0	26
10:82034884	C	A	.	exonic	MAT1A	0	11
12:11286309	G	C	.	intergenic	TAS2R19PRB1	8	29
15:22489958	C	T	rs111826301	intergenic	RP11-2F9.1 TUBGCP5	66	111
15:22489966	A	C	rs72687799	intergenic	RP11-2F9.1 TUBGCP5	68	111
15:22489988	G	A	rs72687801	intergenic	RP11-2F9.1 TUBGCP5	70	112
15:22490019	T	C	rs112521162	intergenic	RP11-2F9.1 TUBGCP5	69	113
16:70305806	G	A	.	intergenic	EXOSC6 DDX19B	0	81
16:70305812	C	A/T	.	intergenic	EXOSC6 DDX19B	0	82
18:43669558	T	C	.	exonic	ATP5A1	0	12
19:1390897	C	T	.	intergenic	AC005330.1 PCSK4	0	12
19:34884932	T	C	.	intergenic	CTD-2518G19.1 CTD-2588C8.1	0	18

Source: (Chang & Wang, 2012)

Table 3.4 Additional Information of Predictive SNPs in All Chromosomes (Without the Exome kit)

Variant	Ref	Alt	dbSNP ID	Region	Gene	Case	Control
14:107034967	T	C	rs72686844	exonic	IGHV5-51	70	98
14:107179022	C	A/G/T	rs2157615	exonic	IGHV2-70	21	0
15:22473106	G	A	rs72687776	intergenic	RP11-2F9.1 TUBGCP5	78	117
15:22489900	T	C	rs113115466	intergenic	RP11-2F9.1 TUBGCP5	66	97
15:22489958	C	T	rs111826301	intergenic	RP11-2F9.1 TUBGCP5	66	111
15:22489966	A	C	rs72687799	intergenic	RP11-2F9.1 TUBGCP5	68	111
15:22489988	G	A	rs72687801	intergenic	RP11-2F9.1 TUBGCP5	70	112
15:22490019	T	C	rs112521162	intergenic	RP11-2F9.1 TUBGCP5	69	113

Source: (Chang & Wang, 2012)

3.8 Genes Associated with CLL

From various chromosomes, multiple genes associated with the predictive SNPs were identified. Remarkably, 30 out of 48 predictive SNPs were found alone in the chromosome 14 and almost all of them were associated with IGHV gene. Previously, IGHV gene was identified to be associated with the CLL, which justify the high occurrence of predictive SNPs in the gene (Damle et al., 1999, Ghia et al., 2003, Ghia et al., 2007 and Kr¨ober et al., 2002). However, the original paper that published these CLL data has also reported few significant genes (Wang et al., 2011), but none of them were appeared as predictive SNPs in this study. Recently, two large-scale GWAS studies were done, which identified significant SNPs associated with the CLL (Berndt et al., 2013 and Speedy et al., 2013), and even those genes were insignificant in this analysis. This study has incorporated a supervised machine learning approach, which was not used in previous studies. Also, above mentioned studies were not aimed to predict CLL, hence the significant SNPs identified here may possibly remained hidden in their investigation.

3.9 AUC with Set of Predictive SNPs

In the end, the AUC values for the 48 predictive SNPs (with the exome kit) were calculated individually, for few chromosomes. Chromosome 2, 6, 14, 15, 16 had more than two predictive SNPs (Table 3.2) and so the AUC values were calculated for only those chromosome. The AUC values are reported in Table 3.4. As it can be seen from Table 3.4, individually, most of the predictive SNPs have significantly low AUC values as compare to their combined AUC value. But interestingly, when predictive SNPs from chromosome 14 were excluded from the analysis, the AUC value for the rest of the SNPs was almost same as the AUC of top70 SNPs (The highest AUC value). Which suggests that the SNPs from the chromosome 14 have very minor contribution in the data classification.

Table 3.5 Average AUC Values of Predictive SNP for Different Chromosome (With the Exome kit)

Chromosome	AUC	Standard Deviation
2	0.77	0.10
6	0.64	0.12
14	0.79	0.09
15	0.67	0.10
16	0.80	0.08
All except 14	0.93	0.04
AUC top70 SNPs	0.94	0.04

CHAPTER 4

CONCLUSION

This investigation shows that jointly, the WES data and the supervised learning method can effectively be utilized to implement a decent CLL risk prediction strategy.

Initially, a novel method to investigate samples containing missing information was introduced. The effective application of the method identified 22 million structural variants across all chromosomes. It was significantly higher compare to the 6000 variants, when the sample removal procedure is not performed. Discovering those samples is an essential step, as single bad samples may contribute to significant amount of data loss in downstream analysis, possibly leaving the data worthless.

Thereafter, the performance of SAMtools and of GATK was compared and it was shown that the variants obtained with both the tools achieve almost similar prediction accuracies. Only GATK was used in the remaining study, considering a minor better performance and its ability to efficiently represent the missing information.

Subsequently, the effectiveness of feature selection through the PCC and the chi-square test was compared, and it was identified that the variants ranked with the PCC obtained a classification accuracy of 66.4 % as compared to 65.1% obtained with the chi-square. Similarly, the classification accuracy for two distinct encoding methods was compared. Where the method Q (encoding with 7X+Y) outperforms the method P (encoding with 0,1 and 2), it was found that with full dataset, the method Q yielded about 4% higher accuracy than the method P. With the SNPs alone, the method Q achieved about 2% higher accuracy than the method P. It was also shown that the method P tends to exclude few variants while encoding genotypes whereas the method Q efficiently encodes all the variants.

Eventually, an exome kit was taken into consideration and it led to about a 15% rise in the classification accuracy. For the exome kit analysis, the parameters that performed well in the prior steps were applied, which gave an average classification accuracy of 85.7% and an AUC of 0.94. That was a significant increase from an average classification accuracy of 70% and an AUC of 0.75 obtained without the exome kit. A PCA analysis of top-30 SNPs (without the exome kit) and top-80 SNPs (with the exome kit) did not show a clear separation when plotted in two dimensions, this suggests that the CLL data are hard to classify.

During the final phase of the investigation, the predictive SNPs were identified from an intersection of 100 training splits. The 48 predictive SNPs, obtained with the exome kit achieved an AUC of 0.93, and the nine predictive SNPs, retrieved without the exome kit attained an AUC of 0.75. Although the predictive SNPs achieved higher AUC values in this study, the SNPs or the associated genes were not found to be reported in any of the previous CLL related studies.

Overall, this study demonstrated the effective implementation of the supervised machine-learning scheme for the CLL risk prediction. The outcome of the experiments created the foundation for the NGS-based CLL prognostics. Since the method is fully reproducible, it can also be applied to other diseases. However, a lack of previous occurrences of predictive SNPs (and associated genes) in the CLL suggest that there is a strong need of a replication study with an independent dataset to fully validate these findings.

REFERENCES

Alpaydin, E. Introduction to machine learning. Cambridge, Massachusetts: MIT Press; 2004.

Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D. and Thibault, J. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. Current Protocols in Bioinformatics 2013:11.10. 11-11.10. 33.

Berndt, S.I., Skibola, C.F., Joseph, V., Camp, N.J., Nieters, A., Wang, Z., Cozen, W., Monnereau, A., Wang, S.S. and Kelly, R.S. Genome-wide association study identifies multiple risk loci for chronic lymphocytic leukemia. Nature genetics 2013;45(8):868-876.

Burton, P.R., Clayton, D.G., Cardon, L.R., Craddock, N., Deloukas, P., Duncanson, A., Kwiatkowski, D.P., McCarthy, M.I., Ouwehand, W.H. and Samani, N.J. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007;447(7145):661-678.

Chang, X. and Wang, K. wANNOVAR: annotating genetic variants for personal genomes via the web. Journal of medical genetics 2012:jmedgenet-2012-100918.

Cortes, C. and Vapnik, V. Support-vector networks. Machine learning 1995;20(3):273-297.

Damle, R.N., Wasil, T., Fais, F., Ghiotto, F., Valetto, A., Allen, S.L., Buchbinder, A., Budman, D., Dittmar, K. and Kolitz, J. Ig V Gene Mutation Status and CD38 Expression As Novel Prognostic Indicators in Chronic Lymphocytic Leukemia Presented in part at the 40th Annual Meeting of The American Society of Hematology, held in Miami Beach, FL, December 4-8, 1998. Blood 1999;94(6):1840-1847.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A. and Hanna, M. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature genetics 2011;43(5):491-498.

Eleftherohorinou, H., Wright, V., Hoggart, C., Hartikainen, A.-L., Jarvelin, M.-R., Balding, D., Coin, L. and Levin, M. Pathway analysis of GWAS provides new insights into genetic susceptibility to 3 inflammatory diseases. PloS one 2009;4(11):e8068.

Ferrer, A., Ollila, J., Tobin, G., Nagy, B., Thunberg, U., Aalto, Y., Vihinen, M., Vilpo, J., Rosenquist, R. and Knuutila, S. Different gene expression in immunoglobulin-mutated and immunoglobulin-unmutated forms of chronic lymphocytic leukemia. Cancer genetics and cytogenetics 2004;153(1):69-72.

Fonseca, N.A., Rung, J., Brazma, A. and Marioni, J.C. Tools for mapping high-throughput sequencing data. Bioinformatics 2012:bts605.

Ghia, P., Guida, G., Stella, S., Gottardi, D., Geuna, M., Strola, G., Scielzo, C. and Caligaris-Cappio, F. The pattern of CD38 expression defines a distinct subset of chronic lymphocytic leukemia (CLL) patients at risk of disease progression. Blood 2003;101(4):1262-1269.

Guyon, I. and Elisseeff, A. An introduction to variable and feature selection. The Journal of Machine Learning Research 2003;3:1157-1182.

Guyon, I., Gunn, S., Ben-Hur, A. and Dror, G. Result analysis of the nips 2003 feature selection challenge. In, Advances in Neural Information Processing Systems. 2004. p. 545-552.

Hatem, A., Bozdağ, D., Toland, A.E. and Çatalyürek, Ü.V. Benchmarking short sequence mapping tools. BMC bioinformatics 2013;14(1):184.

Joachims, T. Making large-scale support vector machine learning practical. In, Advances in kernel methods. Cambridge, Massachusetts: MIT Press; 1999. p. 169-184.

Kröber, A., Seiler, T., Benner, A., Bullinger, L., Brückle, E., Lichter, P., Döhner, H. and Stilgenbauer, S. VH mutation status, CD38 expression level, genomic aberrations, and survival in chronic lymphocytic leukemia. Blood 2002;100(4):1410-1416.

Li, H. and Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 2009;25(14):1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. The sequence alignment/map format and SAMtools. Bioinformatics 2009;25(16):2078-2079.

Mailman, M.D., Feolo, M., Jin, Y., Kimura, M., Tryka, K., Bagoutdinov, R., Hao, L., Kiang, A., Paschall, J. and Phan, L. The NCBI dbGaP database of genotypes and phenotypes. Nature genetics 2007;39(10):1181-1186.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S. and Daly, M. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research 2010;20(9):1297-1303.

Okser, S., Pahikkala, T. and Aittokallio, T. Genetic variants and their interactions in disease risk prediction-machine learning and network perspectives. BioData mining 2013;6(1).

Rassenti, L.Z., Huynh, L., Toy, T.L., Chen, L., Keating, M.J., Gribben, J.G., Neuberg, D.S., Flinn, I.W., Rai, K.R. and Byrd, J.C. ZAP-70 compared with immunoglobulin heavy-chain gene mutation status as a predictor of disease progression in chronic lymphocytic leukemia. New England journal of medicine 2004;351(9):893-901.

Roshan, U., Chikkagoudar, S., Wei, Z., Wang, K. and Hakonarson, H. Ranking causal variants and associated regions in genome-wide association studies by the support vector machine and random forest. Nucleic acids research 2011;39(9):e62-e62.

Speedy, H.E., Di Bernardo, M.C., Sava, G.P., Dyer, M.J., Holroyd, A., Wang, Y., Sunter, N.J., Mansouri, L., Juliusson, G. and Smedby, K.E. A genome-wide association study identifies multiple susceptibility loci for chronic lymphocytic leukemia. Nature genetics 2013.

Wang, L., Lawrence, M.S., Wan, Y., Stojanov, P., Sougnez, C., Stevenson, K., Werner, L., Sivachenko, A., DeLuca, D.S. and Zhang, L. SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. New England journal of medicine 2011;365(26):2497-2506.

Wei, Z., Wang, W., Bradfield, J., Li, J., Cardinale, C., Frackelton, E., Kim, C., Mentch, F., Van Steen, K. and Visscher, P.M. Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease. The American journal of human genetics 2013;92(6):1008-1012.