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

Nihir Patel and Bharati Jhadav
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
Email: nihir.patel@mssm.edu and bharati.jadhav@mssm.edu

Abdulrhman Aljouie and Usman Roshan
Department of Computer Science
New Jersey Institute of Technology
Newark, NJ 07102
Email: aa547@njit.edu and usman@njit.edu

***Abstract*—The era of genomics brings the potential of better DNA based risk prediction and treatment. While genome-wide association studies are extensively studied for risk prediction, the potential of using whole exome data for this purpose is unclear. We explore this problem for chronic lymphocytic leukemia that is one of the largest whole exome dataset of 186 case and 169 controls available from the NIH dbGaP database. We perform a standard next generation sequence procedure to obtain SNP variants on 153 cases and 144 controls after exclusion of samples with missing data. To evaluate their predictive power we first conduct a 50% training and 50% test cross-validation study on the full dataset with the support vector machine as the classifier. There we obtain a mean accuracy of 82% with top 20 ranked SNPs obtained by the Pearson correlation coefficient. We then perform a cross-study validation on case and controls from a lymphoma external study and just controls from head and neck cancer and breast cancer studies (all obtained from NIH dbGaP). On the external dataset we obtain an accuracy of 70% with top ranked SNPs obtained from the original dataset. We also find our top Pearson ranked SNPs to lie on previously implicated genes for this disease. 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.**

I. 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) [1], [2], [3], [4], [5], [6], [7], [8], [9], [10]. This includes various cancers and common diseases [11], [12], [13], [14], [15]. 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 [16]. However, for diseases with high frequency and heritability family history based models perform better [16]. 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 [17]. With a large sample size the highest known AUC of 0.86 and 0.82 for Crohn's disease and ulcerative colitis were reported [18]. 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 [19] datasets, considerably higher than previous studies. Some studies have also used interacting SNPs in GWAS to boost risk prediction accuracy [20], [21].

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 generalize to data from different studies [22]. 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) [23], [24], [25]. 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 leukemia 140X coverage whole exome dataset [26], [27] of 186 tumor and 169 matched germline controls from the NIH dbGaP [28]. 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 leukemia with an onset median age of 70 [29]. 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 [30].

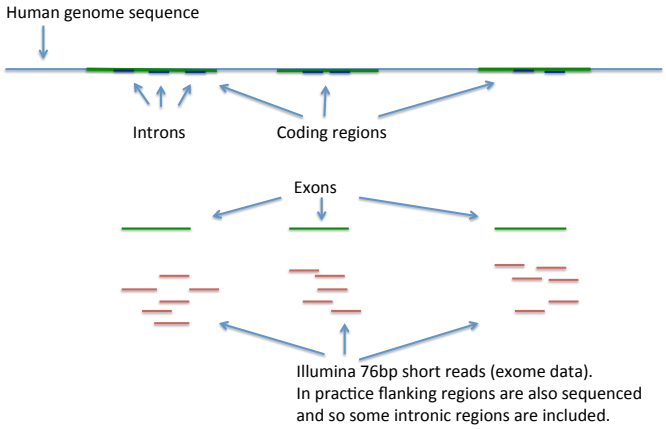

Fig. 1. Whole exome sequences are short reads of exomes obtained by next generation sequencing.

We mapped short read exome sequences to the human genome reference GRCh37 with the popular BWA [31] short read alignment program. We then used the GATK tool [32], [33], [34] 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 122392 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 also obtained exome sequences from three different studies from dbGaP for independent external validation (also known as cross-study validation [22]). We rank SNPs in our training set with the Pearson correlation coefficient [35] 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.

II. 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.

A. Chronic lymphocytic leukemia whole exome sequences and human genome reference

We obtained whole exome sequences of 169 chronic lymphocytic leukemia patients [26], [27] from the NIH dbGaP website [28] with dbGaP study ID phs000435.v2.p1. For each of the 169 patients we have exome tumor sequences and germline controls taken from the same patient. In addition we also obtained exome tumor sequences of 17 patients deposited into dbGaP after publication of the original study [26], [27]. 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 Analyzer II and Hiseq2000 machines and the Agilent SureSelect capture kit by the Broad Institute [26]. 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 processed started well before its release and demands considerable computational resources. Therefore we continue with version 37 for the work in this paper.

B. 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.

1) Mapping reads: As the first step of our analysis we mapped exome short read sequences of 186 tumor cases and 169 matched germline controls to the human genome reference GRCh37 with the BWA MEM program [31]. 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 program 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) [31] that implements Burrows-Wheeler transform. This is fairly accurate for its fast processing speed while mapping against huge reference genome such as humans [36], [37]. 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 [38] 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://picard.sourceforge.net) to add read groups which connects the reads to the patient subject. We also removed duplicates reads introduced by PCR amplification process to avoid artifacts 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) Variant detection: We then used GATK [32], [33], [34] 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 [32], [33]. 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).

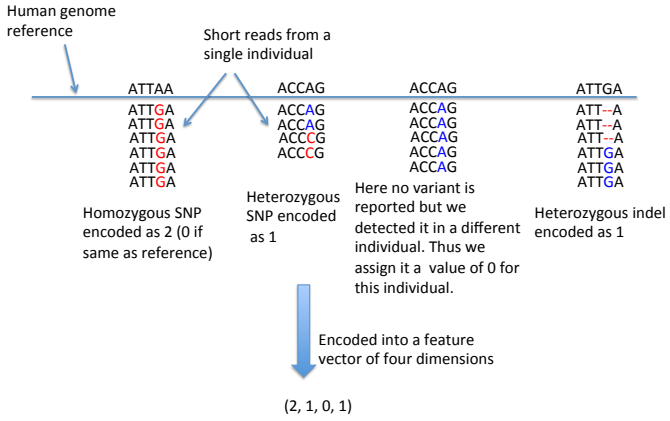

Fig. 2. Toy example depicting the naive 0 1 2 encoding of SNPs and indels. The homozygous and heterozygous genotypes are given by the GATK program [32] when there is a mutation or insertion deletion. For individuals where a SNP is not reported but found in a different individual we use a value of 0.

C. 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.

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 perform the encoding $7A + B$ to represent all possible outputs.

Each feature vector represents variants from a human individual and is labeled -1 for case and 1 for control. The labels +1 and -1 are standard in the machine learning literature [39].

2) Feature selection: We rank features with the Pearson correlation coefficient [35]:

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

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.

3) Classifier: We use the popular soft margin support vector machine (SVM) method [40] implemented in the SVM-light program [41] 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 minimizes the complexity of the classifier plus a regularization parameter C times error on the training data. For all experiments we use the default regularization 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 [41]. In other words we set C to the inverse of the average squared length of feature vectors in the data.

4) Measure of accuracy: We define the classification accuracy as $1 - BER$ where BER is the balanced error [42]. 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 $(\frac{3}{10} + \frac{10}{100})/2 = 0.2$.

D. High performance computing

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

III. 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.

A. Cross-validation

This is a standard approach to evaluate the accuracy of a classifier from a given dataset [39]. 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 [43]) but we have taken ample care and performed all SNP selection only on the training data.

We then learn a support vector machine [40] with the SVM-light software [41] and default regularization on the training set with k top ranked SNPs (see Figure 3). We consider increments of 10 variants upto 100 and increments of 100 upto 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.

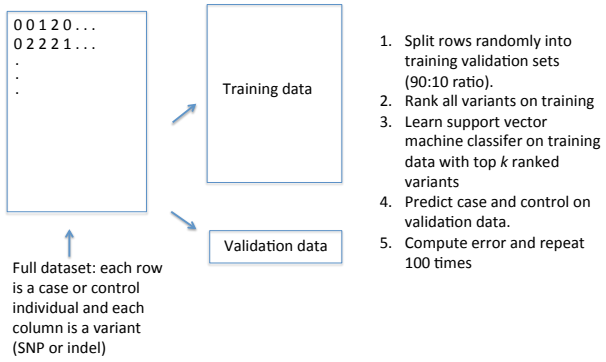

Fig. 3. Illustration of cross-validation.

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.

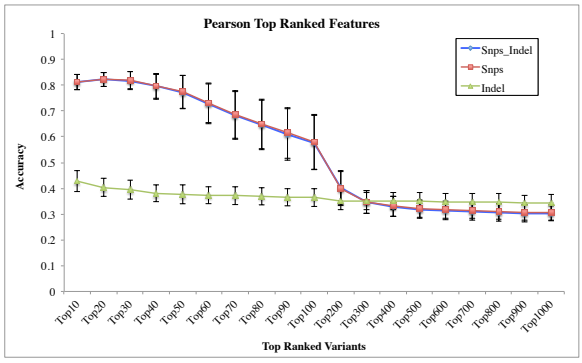

Fig. 4. Average cross-validation accuracy of support vector machine with top Pearson ranked SNPs and indels together and separately on 100 50:50 training validation splits. Also shown are error bars indicating the standard deviation.

B. *Cross-study validation*

For cross-study validation on an independent dataset we consider a lymphoma whole exome study that has case subjects for lymphocytic leukemia 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 [44]. 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 [45]. 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 [46]. 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 leukemia 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 leukemia 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. 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.

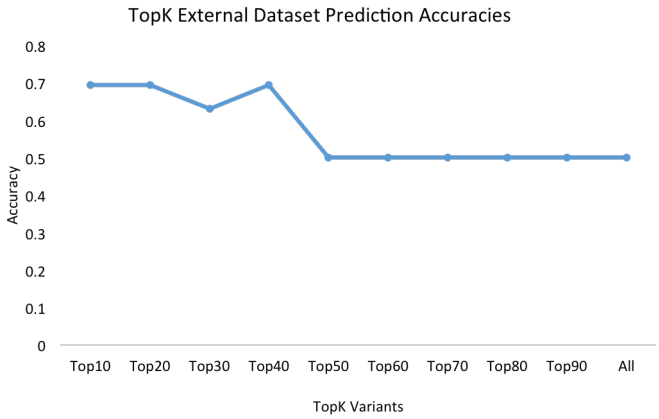

Fig. 5. Accuracy of support vector machine with top Pearson ranked SNPs on just the external independent samples. Since this is a validation dataset we cannot use the labels for any type of model training including ranking of features. Thus the ranking of SNPs is obtained from the original full dataset.

In Figure 5 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 I. 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 5) is 69.4%. In Table I 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.

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

TABLE I. NUMBER OF CORRECTLY PREDICTED CASE AND CONTROLS IN THREE EXTERNAL DATASETS

C. Biological significant 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 [47] 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 [26]. 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 [48], [49], [50], [51]. In Table II we provide additional details of the SNPs in these genes. All four are exonic but don't necessarily rank high in Pearson correlation coefficient.

We also provide the SNP info from the top three high ranking genes in Table III. 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 leukemia 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 [52]. Following this is the valyl-tRNA synthetase (VARS) gene that is also known to be associated with cancer [53]. 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.

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

TABLE II. DETAILS OF FOUR VARIANTS THAT ARE FOUND IN GENES PREVIOUSLY KNOWN TO BE ASSOCIATED WITH CHRONIC LYMPHOCYTIC LEUKEMIA. 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
0.72	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

TABLE III. DETAILS OF TOP RANKING VARIANTS GIVEN THE PEARSON CORRELATION COEFFICIENT RANKING ON THE FULL DATASET. SEE TABLE II FOR MORE CAPTION DETAILS.

IV. 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 [54], [55], [56]. 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.

V. Conclusion

Starting from raw exome sequences we obtained a model for predicting chronic lymphocytic leukemia 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% whereas 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.

Acknowledgment

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

[1] G. Abraham, A. Kowalczyk, J. Zobel, and M. Inouye, "Performance and robustness of penalized and unpenalized methods for genetic prediction of complex human disease," *Genetic Epidemiology*, vol. 37, no. 2, pp. 184–195, 2013. [Online]. Available: http://dx.doi.org/10.1002/gepi.21698

[2] N. Chatterjee, B. Wheeler, J. Sampson, P. Hartge, S. J. Chanock, and J.-H. Park, "Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies," *Nature Genetics*, vol. 45, p. 400405, 2013.

[3] J. Kruppa, A. Ziegler, and I. Knig, "Risk estimation and risk prediction using machine-learning methods," *Human Genetics*, vol. 131, no. 10, pp. 1639–1654, 2012. [Online]. Available: http://dx.doi.org/10.1007/s00439-012-1194-y

[4] U. Roshan, S. Chikkagoudar, Z. Wei, K. Wang, and H. Hakonarson, "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, p. e62, 2011.

[5] M. Sandhu, A. Wood, and E. Young, "Genomic risk prediction," *The Lancet*, vol. 376, pp. 1366–1367, 2010.

[6] C. Kooperberg, M. LeBlanc, and V. Obenchain, "Risk prediction using genome-wide association studies," *Genetic Epidemiology*, vol. 34, no. 7, pp. 643–652, 2010.

[7] D. M. Evans, P. M. Visscher, and N. R. Wray, "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, 2009.

[8] A. C. J. Janssens and C. M. van Duijn, "Genome-based prediction of common diseases: advances and prospects," *Human Molecular Genetics*, vol. 17, no. R2, pp. R166–R173, 2008.

[9] N. R. Wray, M. E. Goddard, and P. M. Visscher, "Prediction of individual genetic risk of complex disease," *Current Opinion in Genetics and Development*, vol. 18, pp. 257–263, 2008.

[10] ——, "Prediction of individual genetic risk to disease from genome-wide association studies," *Genome Research*, vol. 17, pp. 1520–1528, 2007.

[11] P. Kraft and D. J. Hunter, "Genetic Risk Prediction – Are We There Yet?" *N Engl J Med*, vol. 360, no. 17, pp. 1701–1703, 2009. [Online]. Available: http://content.nejm.org

[12] M. H. Gail, "Discriminatory accuracy from single-nucleotide polymorphisms in models to predict breast cancer risk," *N Engl J Med*, vol. 100, no. 14, pp. 1037–1041, 2008.

[13] A. C. Morrison, L. A. Bare, L. E. Chambless, S. G. Ellis, M. Malloy, J. P. Kane, J. S. Pankow, J. J. Devlin, J. T. Willerson, and E. Boerwinkle, "Prediction of Coronary Heart Disease Risk using a Genetic Risk Score: The Atherosclerosis Risk in Communities Study," *Am. J. Epidemiol.*, p. kwm060, 2007.

[14] S. Kathiresan, O. Melander, D. Anevski, C. Guiducci, N. P. Burtt, C. Roos, J. N. Hirschhorn, G. Berglund, B. Hedblad, L. Groop, D. M. Altshuler, C. Newton-Cheh, and M. Orho-Melander, "Polymorphisms associated with cholesterol and risk of cardiovascular events," *New England Journal of Medicine*, vol. 358, pp. 1240–1249, 2008.

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

[16] C. B. Do, D. A. Hinds, U. Francke, and N. Eriksson, "Comparison of family history and snps for predicting risk of complex disease," *PLoS Genet*, vol. 8, no. 10, p. e1002973, 10 2012. [Online]. Available: http://dx.doi.org/10.1371%2Fjournal.pgen.1002973

[17] D. Shigemizu, T. Abe, T. Morizono, T. A. Johnson, K. A. Boroevich, Y. Hirakawa, T. Ninomiya, Y. Kiyohara, M. Kubo, Y. Nakamura, S. Maeda, and T. Tsunoda, "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, p. e92549, 03 2014. [Online]. Available: http://dx.doi.org/10.1371%2Fjournal.pone.0092549

[18] Z. Wei, W. Wang, J. Bradfield, J. Li, C. Cardinale, E. Frackelton, C. Kim, F. Mentch, K. V. Steen, P. M. Visscher, R. N. Baldassano, and H. Hakonarson, "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, 2013.

[19] Welcome Trust Case Control Consortium, "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls." *Nature*, vol. 447, pp. 661–678, 2007.

[20] S. Okser, T. Pahikkala, and T. Aittokallio, "Genetic variants and their interactions in disease risk prediction - machine learning and network perspectives," *BioData Mining*, vol. 6, no. 1, p. 5, 2013. [Online]. Available: http://www.biodatamining.org/content/6/1/5

[21] H. Eleftherohorinou, V. Wright, C. Hoggart, A.-L. Hartikainen, M.-R. Jarvelin, D. Balding, L. Coin, and M. Levin, "Pathway analysis of gwas provides new insights into genetic susceptibility to 3 inflammatory diseases," *PLoS ONE*, vol. 4, no. 11, p. e8068, 11 2009. [Online]. Available: http://dx.doi.org/10.1371%2Fjournal.pone.0008068

[22] C. Bernau, M. Riester, A.-L. Boulesteix, G. Parmigiani, C. Huttenhower, L. Waldron, and L. Trippa, "Cross-study validation for the assessment of prediction algorithms," *Bioinformatics*, vol. 30, no. 12, pp. i105–i112, 2014.

[23] S. J. Schrodi, S. Mukherjee, Y. Shan, G. Tromp, J. J. Sninsky, A. P. Callear, T. C. Carter, Z. Ye, J. L. Haines, M. H. Brilliant, P. K. Crane, D. T. Smelser, R. C. Elston, and D. E. Weeks, "Genetic-based prediction of disease traits: Prediction is very difficult, especially about the future," *Frontiers in Genetics*, vol. 5, no. 162, 2014. [Online]. Available: http://www.frontiersin.org/applied_genetic_epidemiology/10.3389/fgene.2014.00162/abstract

[24] T. A. Manolio, "Bringing genome-wide association findings into clinical use," *Nature Reviews Genetics*, vol. 14, pp. 549–558, 2013. [Online]. Available: http://dx.doi.org/10.1038/nrg3523

[25] P. M. Visscher, M. A. Brown, M. I. McCarthy, and J. Yang, "Five years of GWAS discovery," *The American Journal of Human Genetics*, vol. 90, no. 1, pp. 7 – 24, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0002929711005337

[26] L. Wang, M. S. Lawrence, Y. Wan, P. Stojanov, C. Sougnez, K. Stevenson, L. Werner, A. Sivachenko, D. S. DeLuca, L. Zhang, W. Zhang, A. R. Vartanov, S. M. Fernandes, N. R. Goldstein, E. G. Folco, K. Cibulskis, B. Tesar, Q. L. Sievers, E. Shefler, S. Gabriel, N. Hacohen, R. Reed, M. Meyerson, T. R. Golub, E. S. Lander, D. Neuberg, J. R. Brown, G. Getz, and C. J. Wu, "SF3B1 and other novel cancer genes in chronic lymphocytic leukemia," *New England Journal of Medicine*, vol. 365, no. 26, pp. 2497–2506, 2011.

[27] D. A. Landau, S. L. Carter, P. Stojanov, A. McKenna, K. Stevenson, M. S. Lawrence, C. Sougnez, C. Stewart, A. Sivachenko, L. Wang *et al.*,

"Evolution and impact of subclonal mutations in chronic lymphocytic leukemia," *Cell*, vol. 152, no. 4, pp. 714–726, 2013.

[28] M. D. Mailman, M. Feolo, Y. Jin, M. Kimura, K. Tryka, R. Bagoutdinov, L. Hao, A. Kiang, J. Paschall, L. Phan *et al.*, "The ncbi dbgap database of genotypes and phenotypes," *Nature genetics*, vol. 39, no. 10, pp. 1181–1186, 2007.

[29] M. Shanshal and R. Y. Haddad, "Chronic lymphocytic leukemia," *Disease-a-Month*, vol. 58, p. 153167, 2012.

[30] "National cancer institute website (http://www.cancer.gov/cancertopics/pdq/treatment/cll/patient)." [Online]. Available: http://www.cancer.gov/cancertopics/pdq/treatment/cll/patient

[31] H. Li and R. Durbin, "Fast and accurate short read alignment with burrowswheeler transform," *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, 2009.

[32] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo, "The Genome Analysis Toolkit: A Mapreduce framework for analyzing next-generation DNA sequencing data," *Genome Research*, vol. 20, no. 9, pp. 1297–1303, 2010.

[33] M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna *et al.*, "A framework for variation discovery and genotyping using next-generation DNA sequencing data," *Nature genetics*, vol. 43, no. 5, pp. 491–498, 2011.

[34] G. A. Auwera, M. O. Carneiro, C. Hartl, R. Poplin, G. del Angel, A. Levy-Moonshine, T. Jordan, K. Shakir, D. Roazen, J. Thibault *et al.*, "From Fastq data to high-confidence variant calls: The genome analysis toolkit best practices pipeline," *Current Protocols in Bioinformatics*, pp. 11–10, 2013.

[35] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Mar. 2003. [Online]. Available: http://dl.acm.org/citation.cfm?id=944919.944968

[36] N. A. Fonseca, J. Rung, A. Brazma, and J. C. Marioni, "Tools for mapping high-throughput sequencing data," *Bioinformatics*, vol. 28, no. 24, pp. 3169–3177, 2012.

[37] A. Hatem, D. Bozdag, A. Toland, and U. Catalyurek, "Benchmarking short sequence mapping tools," *BMC Bioinformatics*, vol. 14, no. 1, p. 184, 2013. [Online]. Available: http://www.biomedcentral.com/1471-2105/14/184

[38] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and . G. P. D. P. Subgroup, "The Sequence Alignment Map format and SAMtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009.

[39] E. Alpaydin, *Machine Learning*. MIT Press, 2004.

[40] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[41] T. Joachims, "Making large-scale svm learning practical," in *Advances in Kernel Methods - Support Vector Learning*, B. Schölkopf, C. Burges, and A. Smola, Eds. MIT Press, 1999.

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

[43] P. Smialowski, D. Frishman, and S. Kramer, "Pitfalls of supervised feature selection," *Bioinformatics*, vol. 26, no. 3, pp. 440–443, 2010.

[44] L. Pasqualucci, H. Khiabanian, M. Fangazio, M. Vasishtha, M. Messina, A. B. Holmes, P. Ouillette, V. Trifonov, D. Rossi, F. Tabbò *et al.*, "Genetics of follicular lymphoma transformation," *Cell reports*, vol. 6, no. 1, pp. 130–140, 2014.

[45] N. Stransky, A. M. Egloff, A. D. Tward, A. D. Kostic, K. Cibulskis, A. Sivachenko, G. V. Kryukov, M. S. Lawrence, C. Sougnez, A. McKenna, E. Shefler, A. H. Ramos, P. Stojanov, S. L. Carter, D. Voet, M. L. Corts, D. Auclair, M. F. Berger, G. Saksena, C. Guiducci, R. C. Onofrio, M. Parkin, M. Romkes, J. L. Weissfeld, R. R. Seethala, L. Wang, C. Rangel-Escareo, J. C. Fernandez-Lopez, A. Hidalgo-Miranda, J. Melendez-Zajgla, W. Winckler, K. Ardlie, S. B. Gabriel, M. Meyerson, E. S. Lander, G. Getz, T. R. Golub, L. A. Garraway, and J. R. Grandis, "The mutational landscape of head and neck squamous cell carcinoma," *Science*, vol. 333, no. 6046, pp. 1157–1160, 2011.

[46] S. Banerji, K. Cibulskis, C. Rangel-Escareno, K. K. Brown, S. L. Carter, A. M. Frederick, M. S. Lawrence, A. Y. Sivachenko, C. Sougnez, L. Zou *et al.*, "Sequence analysis of mutations and translocations across breast cancer subtypes," *Nature*, vol. 486, no. 7403, pp. 405–409, 2012.

[47] K. Wang, M. Li, and H. Hakonarson, "ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data," *Nucleic Acids Research*, vol. 38, no. 16, p. e164, 2010.

[48] M. C. Di Bernardo, D. Crowther-Swanepoel, P. Broderick, E. Webb, G. Sellick, R. Wild, K. Sullivan, J. Vijayakrishnan, Y. Wang, A. M. Pittman *et al.*, "A genome-wide association study identifies six susceptibility loci for chronic lymphocytic leukemia," *Nature genetics*, vol. 40, no. 10, pp. 1204–1210, 2008.

[49] S. I. Berndt, C. F. Skibola, V. Joseph, N. J. Camp, A. Nieters, Z. Wang, W. Cozen, A. Monnereau, S. S. Wang, R. S. Kelly, Q. Lan, L. R. Teras, N. Chatterjee, C. C. Chung, M. Yeager, A. R. Brooks-Wilson, P. Hartge, M. P. Purdue, B. M. Birmann, B. K. Armstrong, P. Cocco, Y. Zhang, G. Severi, A. Zeleniuch-Jacquotte, C. Lawrence, L. Burdette, J. Yuenger, A. Hutchinson, K. B. Jacobs, T. G. Call, T. D. Shanafelt, A. J. Novak, N. E. Kay, M. Liebow, A. H. Wang, K. E. Smedby, H.-O. Adami, M. Melbye, B. Glimelius, E. T. Chang, M. Glenn, K. Curtin, L. A. Cannon-Albright, B. Jones, W. R. Diver, B. K. Link, G. J. Weiner, L. Conde, P. M. Bracci, J. Riby, E. A. Holly, M. T. Smith, R. D. Jackson, L. F. Tinker, Y. Benavente, N. Becker, P. Boffetta, P. Brennan, L. Foretova, M. Maynadie, J. McKay, A. Staines, K. G. Rabe, S. J. Achenbach, C. M. Vachon, L. R. Goldin, S. S. Strom, M. C. Lanasa, L. G. Spector, J. F. Leis, J. M. Cunningham, J. B. Weinberg, V. A. Morrison, N. E. Caporaso, A. D. Norman, M. S. Linet, A. J. De Roos, L. M. Morton, R. K. Severson, E. Riboli, P. Vineis, R. Kaaks, D. Trichopoulos, G. Masala, E. Weiderpass, M.-D. Chirlaque, R. C. H. Vermeulen, R. C. Travis, G. G. Giles, D. Albanes, J. Virtamo, S. Weinstein, J. Clavel, T. Zheng, T. R. Holford, K. Offit, A. Zelenetz, R. J. Klein, J. J. Spinelli, K. A. Bertrand, F. Laden, E. Giovannucci, P. Kraft, A. Kricker, J. Turner, C. M. Vajdic, M. G. Ennas, G. M. Ferri, L. Miligi, L. Liang, J. Sampson, S. Crouch, J.-H. Park, K. E. North, A. Cox, J. A. Snowden, J. Wright, A. Carracedo, C. Lopez-Otin, S. Bea, I. Salaverria, D. Martin-Garcia, E. Campo, J. F. Fraumeni Jr, S. de Sanjose, H. Hjalgrim, J. R. Cerhan, S. J. Chanock, N. Rothman, and S. L. Slager, "Genome-wide association study identifies multiple risk loci for chronic lymphocytic leukemia," *Nat Genet*, vol. 45, pp. 868–876, Aug 2013, article. [Online]. Available: http://dx.doi.org/10.1038/ng.2652

[50] H. E. Speedy, M. C. Di Bernardo, G. P. Sava, M. J. S. Dyer, A. Holroyd, Y. Wang, N. J. Sunter, L. Mansouri, G. Juliusson, K. E. Smedby, G. Roos, S. Jayne, A. Majid, C. Dearden, A. G. Hall, T. Mainou-Fowler, G. H. Jackson, G. Summerfield, R. J. Harris, A. R. Pettitt, D. J. Allsup, J. R. Bailey, G. Pratt, C. Pepper, C. Fegan, R. Rosenquist, D. Catovsky, J. M. Allan, and R. S. Houlston, "A genome-wide association study identifies multiple susceptibility loci for chronic lymphocytic leukemia," *Nat Genet*, vol. 46, pp. 56–60, Jan 2014, letter. [Online]. Available: http://dx.doi.org/10.1038/ng.2843

[51] S. L. Slager, K. G. Rabe, S. J. Achenbach, C. M. Vachon, L. R. Goldin, S. S. Strom, M. C. Lanasa, L. G. Spector, L. Z. Rassenti, J. F. Leis *et al.*, "Genome-wide association study identifies a novel susceptibility locus at 6p21.3 among familial cll," *Blood*, vol. 117, no. 6, pp. 1911–1916, 2011.

[52] S. G. Park, P. Schimmel, and S. Kim, "Aminoacyl trna synthetases and their connections to disease," *Proceedings of the National Academy of Sciences*, vol. 105, no. 32, pp. 11 043–11 049, 2008.

[53] D. Kim, N. Kwon, and S. Kim, "Association of aminoacyl-trna synthetases with cancer," in *Aminoacyl-tRNA Synthetases in Biology and Medicine*, ser. Topics in Current Chemistry, S. Kim, Ed. Springer Netherlands, 2014, vol. 344, pp. 207–245. [Online]. Available: http://dx.doi.org/10.1007/128_2013_455

[54] C. S. Carlson, T. C. Matise, K. E. North, C. A. Haiman, M. D. Fesinmeyer, S. Buyske, F. R. Schumacher, U. Peters, N. Franceschini, M. D. Ritchie, D. J. Duggan, K. L. Spencer, L. Dumitrescu, C. B. Eaton, F. Thomas, A. Young, C. Carty, G. Heiss, L. Le Marchand, D. C. Crawford, L. A. Hindorff, C. L. Kooperberg, and for the PAGE Consortium, "Generalization and dilution of association results from european gwas in populations of non-european ancestry: The page study," *PLoS Biol*, vol. 11, no. 9, p. e1001661, 09 2013.

[55] M. Pino-Yanes, N. Thakur, C. R. Gignoux, J. M. Galanter, L. A. Roth, C. Eng, K. K. Nishimura, S. S. Oh, H. Vora, S. Huntsman *et al.*,

"Genetic ancestry influences asthma susceptibility and lung function among latinos," *Journal of Allergy and Clinical Immunology*, vol. 135, no. 1, pp. 228–235, 2015.

[56] B. I. Freedman, J. Divers, and N. D. Palmer, "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, 2013.