Cross-validation and cross-study validation of kidney cancer with machine learning and whole exome sequences from the National Cancer Institute

Abdulrhman Aljouie*
Department of Biostatistics and Bioinformatics
King Abdullah International Medical Research Center
King Saud bin Abdulaziz University for Health Sciences
Riyadh, Saudi Arabia 11426
Email: aljouieab@ngha.med.sa

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

Usman Roshan†
Department of Computer Science*†
New Jersey Institute of Technology
Newark, New Jersey 07102
Email: usman@njit.edu†, aa547@njit.edu*

***Abstract*—Accurate cancer risk prediction from genetic and environment variables is a key problem in medicine. One approach is to use somatic mutations which could potentially be used in early detection and prevention. SNP based studies are the most common ones utilizing this approach, however most studies lack a cross-study validation component across at least two independent studies. Here we explore the cross-validation and cross-study validation of predicting kidney cancer case and controls with SNPs obtained from whole exome sequences at the National Cancer Institute. From the Genomics Data Commons portal we obtained aligned whole exome sequences of two different kidney cancer studies: 110 cases and controls of KIRP for renal papillary cell carcinoma and 34 cases and controls of KICH for kidney chromophobe cell carcinoma. We performed a rigorous quality control procedure to obtain SNPs and rank them with feature selection. On top ranked SNPs we find the support vector machine to obtain a cross-validation accuracy of 0.71 (with 10 SNPs) and 0.72 (with 20 SNPs) in KIRP and KICH respectively. We then learn a model on KIRP and with 10 SNPs achieve an accuracy of 0.66 on the KICH samples. Our work shows that we can predict kidney chromophobe carcinoma from a kidney papillary carcinoma dataset with better than a random classification which would have 0.5 accuracy. In continuing work we are expanding these sample sizes and extending cross-study to other kidney cancer datasets in the NCI GDC portal.**

I. Introduction

Cancer risk prediction from one's DNA is of considerable interest in modern medicine [1], [24]. One way to achieve this is to determine mutations by comparing DNA in tumor cells to healthy ones. Such mutations are called somatic and could potentially be used for early detection and prevention of cancer [23], [16], [26].

The majority of efforts on predicting cancer are focused on using SNPs obtained from genome-wide assocation studies and from whole exome sequences [28], [2], [18], [22], [10]. However, there are also dangerous pitfalls associated with SNP-based cancer risk prediction [27]. The most common one is lack of validation on an independent dataset, also known as cross-study validation [5]. Most studies focus on the cross-validation accuracy which is obtained by splitting a given dataset randomly into training and validation several times and obtaining the average accuracy on the validation. In a cross-study validation we want to see how well SNPs determined on data for one disease from a specific study generalizes to the same disease or a related one from a different study.

Here we explore the accuracy of predicting kidney cancer case and controls with somatic mutations across two different whole exome sequence dataset obtained from the National Cancer Institute Genomic Data Commons database [12]. We consider datasets of renal papillary cell carcinoma and chromophobe renal cell carcinoma. Our data are pre-aligned short read sequences from which we determine variants. We study three quality control methods of variant detection and show that the most rigorous one gives the most parsimonious model with the highest accuracy.

Our main result is the cross-study validation between the two datasets. We show that we achieve an accuracy of 66.2% when predicting chromophobe individuals after learning a model of 10 SNPs from the renal papillary dataset. Our work here suggests that we can predict kidney chromophobe carcinoma with high quality SNPs obtained from a kidney papillary carcinoma dataset. We are continuing this work include all remaining samples in the two datasets and cross-study across other datasets in the NCI GDC portal. Below we describe our methods in detail followed by experimental results.

II. Methods

We describe here our data along with our quality control protocol. We then describe our machine learning pipeline.

A. Data

There are several kidney cancer whole exome datasets at the National Cancer Institute (NCI) Genomic Data Commons (GDC) portal from across three different projects: The Cancer Genome Atlas (TCGA), TARGET, and Foundation Medicine Adult Cancer Clinical Dataset (FM-AD). We obtained authorization to the TCGA project from where we downloaded two of their three datasets.

- Kidney Renal Papillary Cell Carcinoma (KIRP): total of 291 individuals
- Kidney Chromophobe (KICH): total of 113 individuals

Both datasets contain individuals of European, Afrian, and Asian ancestry and have older patients between stage I and III cancer. For each individual in each dataset exome sequences of the affect cell and a healthy cell (from the same person) are made available.

In order to avoid mutations that occur from ancestry differences we considered just individuals of European ancestry (which is also the majority ancestry). Due to time constraints and checksum/download errors we were able to download only some male subjects from each study. In Table I we give the number of case and controls that we downloaded successfully for each study.

TABLE I
Kidney cancer datasets used in our study

Dataset	Cases	Controls
TCGA-KIRP	110	110
TCGA-KICH	34	34

Each case and control file that we downloaded are prealigned exome sequences to the human genome reference (build 38, version GRCh38.d1.vd1) with the BWA program [13]. Thus, from the GDC portal we obtained BAM files [14] for each individual's tumor and healthy exome sequences. These are binary files of the SAM format that show the alignment of each short read to the reference genome.

The NCI GDC portal also contains files with already detected variants for each individual. However, those variants were obtained by comparing each individuals healthy exome sequences to their tumor ones. In our analysis we do a collective analysis of all the individuals at the same time to determine variants so we can detect missing values as explained below.

B. Quality control for determining SNPs

We combine all the case and controls and perform a collective variant analysis with the popular Genome Analysis Toolkit (GATK) software [17], [8], [4]. In the collective analysis we were able to identify SNPs that are not reported across samples. For example if a SNP does not pass quality control it is not reported and is thus a missing value.

We study three filtering methods of obtaining SNPs with the popular Genome Analysis Toolkit (GATK) software [17], [8], [4]. By default any reads with a MAPQ quality score (which is a measure of the alignment quality) below 25 is eliminated in the analysis.

- Soft filtering: This is the GATK Variant Quality Score Recalibration which uses machine learning to identify good variants from bad ones.
- Hard filtering: Any SNP with a genotype quality score below 30 and a depth below 5 is ignored. The genotype quality score is a statistical quantity the gives us the accuracy of the SNP and the depth is the minimum of reads that contain the SNP. These are default values used in the GATK program.
- Soft and hard filtering: Both of the above are applied.

After each filtering we remove any SNP that is missing (not reported by GATK) in at least one sample, thus eliminating the need for imputation. After filtering we obtain number of SNPs given in Table II. In the same table we also show the number of SNPs common in the two studies, we use these for the cross-study validation.

TABLE II
Number of SNPs in our datasets after filtering

	Filtering		
Dataset	Soft	Hard	Soft+hard
TCGA-KIRP	264858	131141	109700
TCGA-KICH	246290	135937	111394
Intersection of KIRP and KICH	131157	44426	36029

C. SNP encoding

Once a dataset of SNPs is obtained after the quality control described above we perform encoding. The GATK program outputs variants in the VCF format [7] which encodes the reference allele as 0 and alternate alleles (including gap) from 1 onwards. For example a genotype of 0/0 means the individual is homozygous in the reference allele, 0/2 means heterozygous in the second alternate allele, and 1/1 means homozygous in the second alternate allele. We can encode these to unique numbers with the simple formula 4A+B for a SNP enconded as A/B.

D. Experimental performance study

To evaluate the predictive capability of SNPs we perform cross-validation and a cross-study validation experiment. In the cross-validation we split a given dataset into train and test and evaluate the error of predicting the test. A high accuracy doesn't necessarily mean the SNPs would generalize to other datasets or related diseases. Thus we also perform a cross-study validation to determine generalization to another dataset.

1) Cross-validation and machine learning: We describe our step-wise cross-validation procedure below.

1) First we randomly split the number of samples into two random disjoint sets of 90:10 ratio. We call the 90% set our *training* data and the 10% *test* data.
2) We rank the SNPs according to the Pearson correlation coefficient [3] as implemented in the Python scikit-learn machine learning library [19]. The Pearson correlation coefficient is the sample correlation coefficient that measures the covariance between two variables divided by their variances to normalize. A value close to 1 or -1 indicates a linear correlation whereas 0 means the variables are uncorrelated [3].
3) We consider the top k ranked SNPs for increasing values of k and train a support vector machine (SVM) [6] also with the linear SVM in the Python scikit-learn library [19]. We cross-validate the regularization parameter C of the SVM by cross-validating on the training set (as opposed to the whole dataset as we are doing here).
4) Briefly the support vector machine is a linear classifier with known powerful generalization capabilities. It is also easy and fast to train.
5) With the trained model we predict case and controls of the individuals in the test dataset and determine the error (since we know their true case and control status).
6) We repeat the above steps 10 times and take the average error.

2) Cross-study validation: Here we wish to determine the error of predicting case and controls across two independently obtained studies. We measure the error of a predicting case and control in the KICH dataset, which contain individuals with renal chromophobe carcinoma, with a model trained on the KIRP dataset, which are renal papillary carcinoma individuals.

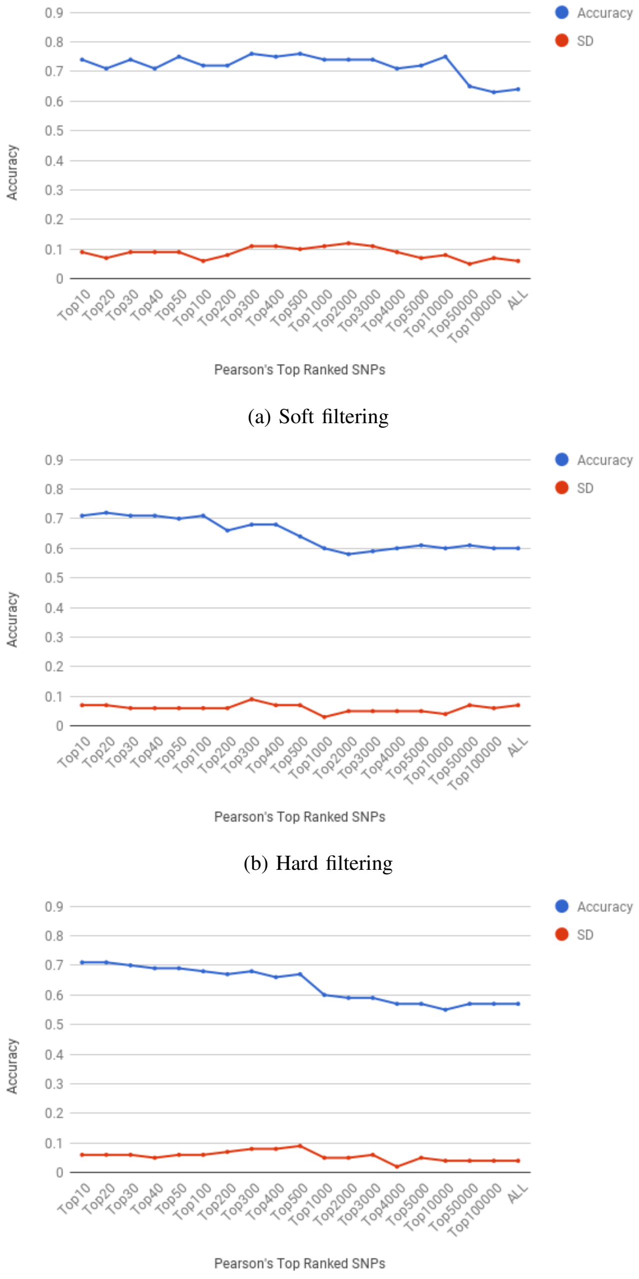
III. Results

We begin with cross-validation accuracy of each dataset.

A. Cross-validation

In Figure III we show the average accuracy of the support vector across 10 random 90:10 train test splits of the KIRP dataset. We make several interesting observations consistent with previous findings.

- Top ranked SNPs with the Pearson correlation coefficient give a higher accuracy than lower ranked ones and all SNPs. This is consistent with previous findings on predicting cancer and disease risk with genomic SNPs [18], [2], [20]
- The soft filtering gives a slightly higher accuracy (reaching 0.76 with 500 SNPs) and fluctuating curve compared to hard and combined filtering.
- The hard and combined filtering achieve their top accuracy of 0.72 and 0.71 with just top 20 and 10 ranked SNPs respectively.
- The combined filtering gives us the most parsimonious model, it achieves its highest accuracy with the fewest number of SNPs (10).



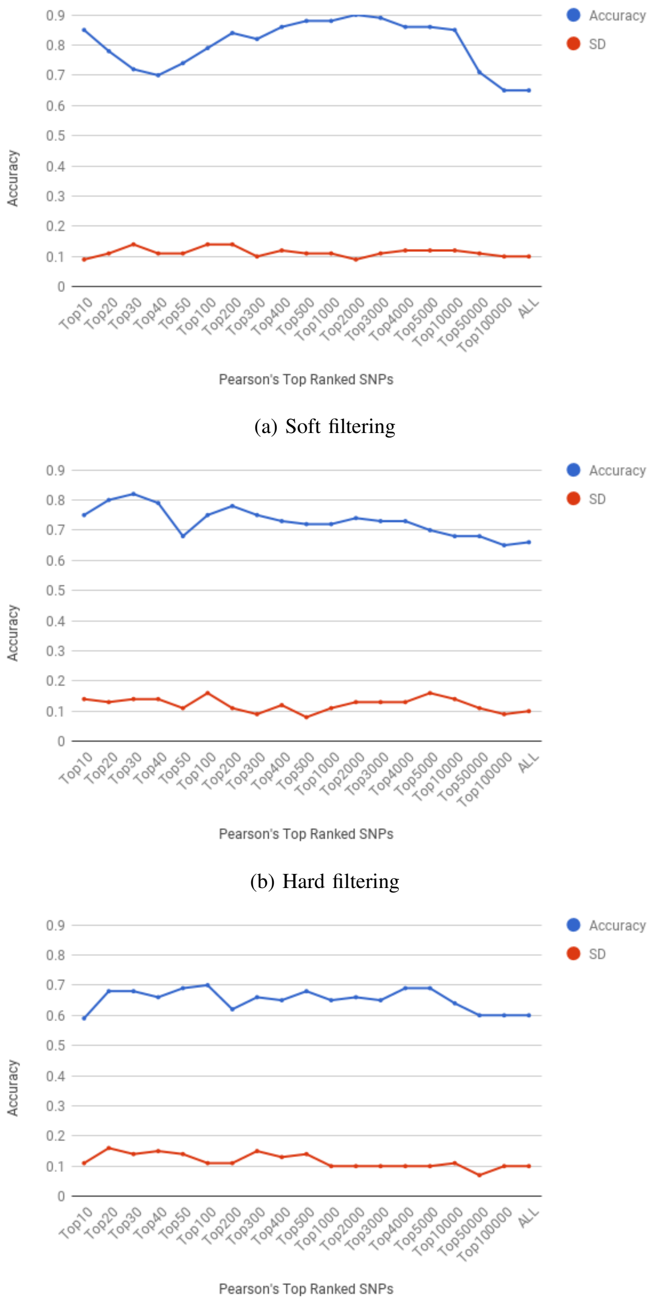
(a) Soft filtering

(b) Hard filtering

(c) Soft and hard combined

TABLE III

Average cross-validation accuracy of the support vector machine on top ranked SNPs obtained after three filterings on the KIRP dataset



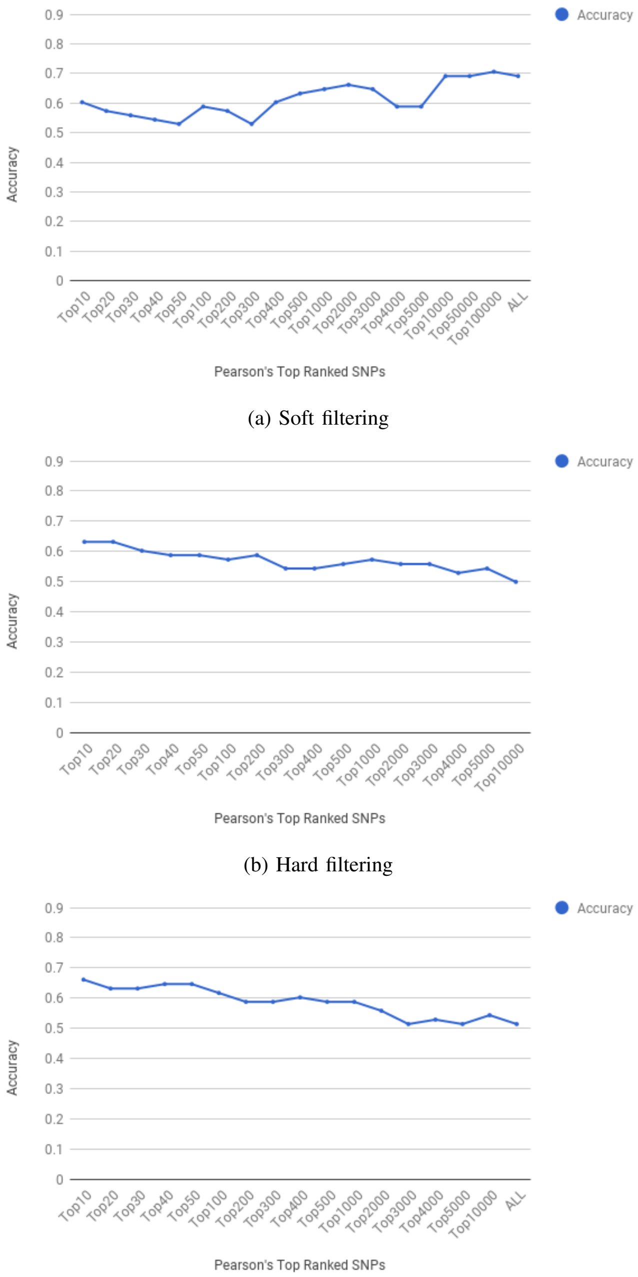
(a) Soft filtering

(b) Hard filtering

(c) Soft and hard combined

TABLE IV
AVERAGE CROSS-VALIDATION ACCURACY OF THE SUPPORT VECTOR MACHINE ON TOP RANKED SNPS OBTAINED AFTER THREE FILTERINGS ON THE KICH DATASET

We now look at the average accuracy of the support vector across 10 random 90:10 train test splits of the KICH dataset (Figure IV). This dataset is less than one third the size of the KICH dataset and so we see different trends. Due to its small sample size the accuracy fluctuates in all three filterings and peaks equally with a few and many SNPs. We use this dataset primarily as an independent set.



(a) Soft filtering

(b) Hard filtering

(c) Soft and hard filtering

TABLE V
ACCURACY OF SUPPORT VECTOR MACHINE ON THE KICH DATASET AFTER TRAINED ON TOP RANKED SNPS IN THE KIRP DATASET

B. Cross-study validation

For the cross-study validatio Here we learn a support vector machine model on top ranked SNPs in the KIRP dataset and predict individuals in the KICH dataset. In Figure V we see that the most parsimonious model is given by the soft and hard filtering. There we see an accuracy of 0.66 with just 10 SNPs. In comparison the hard filtering peaks at 0.63 with 10 SNPs and soft peaks at 0.7 with 100,000 SNPs.

While our focus is on the cross-study prediction accuracy,

SNP	Ref	Alt	Pearson	Gene/Region	Chromosome
1	A	G,C	0.35	ANO2	12
2	A	C	0.07	ADAMTS9	3
3	G	A	0.07	Non-coding	X
4	A	C	0.07	GORAB	1
5	T	C	0.06	NR2C2	3
6	G	A	0.06	SELP	1
7	A	T	0.06	Non-coding	
8	C	T	0.06	LOC100421093	6
9	A	C	0.06	C9orf171	9
10	G	A	0.06	FBXL4	6

TABLE VI
REFERENCE AND ALTERNATE ALLELE, PEARSON CORRELATION COEFFICIENT VALUE, GENE, AND CHROMOSOME NUMBER OF TOP 10 RANKED SNPS USED IN THE CROSS-STUDY (SOFT PLUS HARD FILTERING)

we show in Table VI the top 10 ranked SNPs in the cross-study validation. These SNPs are present in both studies but the ranking is performed on just the KIRP (training) dataset. We see that most of the SNPs are in coding regions except for two. The SNP in the ANO2 gene has the highest Pearson correlation whereas the others are lower by a large margin. The same SNP is also highly ranked in both the datasets separately.

The ANO2 gene belongs to the family of anoctamins that are known to be expressed in gastrointestinal stromal tumors and neck and head carcinoma [25]. This gene is known to have a functional role in calcium activated chloride currents [9] but it is unclear how that relates kidney cancer. The ANO1 gene that comes from the same family, however, is known to be expressed in pancreatic cancer [21]. In continuing work as we add more samples and explore other kidney cancer datasets we expect to make more interesting biological findings.

C. Ranking of previously known kidney cancer genes in our data

In Table VII we show the ranking of SNPs present in genes previously known to be associated with kidney cancer [15]. We show the rankings as well as the Pearson correlation coefficients in the KIRP and KICH datasets separately and the intersection of their SNPs (as in the cross-study). The MET gene is the highest ranked in the KIRP study in this dataset and is also a drug target for clinical treatment of renal papillary carcinoma [11]. We see that these genes have low Pearson correlation coefficients indicating that while they are associated with kidney cancer from previous studies their predictive value is limited here.

Gene	KIRP	KICH	KIPR & KICH
VHL	28296 (5.39e-18)	58117 (1.026e-17)	15831 (5.39e-18)
FH	107511 (0)	6437 (0.07)	35682 (0)
FLCN	15889 (1.88e-17)	2777 (0.096)	13308 (7.26e-18)
MET	1975 (0.026)	20909 (3.86e-17)	799 (0.026)
TSC1	5732 (0.009)	4327 (0.088)	7430 (2.05e-17)
TSC2	16060 (1.85e-17)	4295 (0.088)	7816 (1.85e-17)

TABLE VII
RANK OF SNPS (AND THEIR PEARSON CORRELATION VALUES IN PARENTHESIS) IN KNOWN KIDNEY CANCER GENES IN EACH OF OUR DATASETS GIVEN BY THE HARD AND SOFT FILTERING COMBINED

IV. CONCLUSION

We perform an initial cross-validation and cross-study validation across two kidney cancer datasets obtained from the NCI GDC database. Our results show that we can predict kidney chromphobe carcinoma case and controls with 66% accuracy with SNPs learnt from a kidney papillary cell carcinoma dataset. We are extending our study to include more samples from the existing datasets and other datasets from the NCI GDC database.

ACKNOWLEDGMENT

We thank the NJIT Academic and Research Computing Systems Group (ARCS) for their support in running computational experiments necessary for this study.

REFERENCES

[1] Gad Abraham and Michael Inouye. Genomic risk prediction of complex human disease and its clinical application. *Current opinion in genetics & development*, 33:10–16, 2015.

[2] Abdulrhman Aljouie, Nihir Patel, Bharati Jadhav, and Usman Roshan. Cross-validation and cross-study validation of chronic lymphocytic leukaemia with exome sequences and machine learning. *International Journal of Data Mining and Bioinformatics*, 16(1):47–63, 2016.

[3] Ethem Alpaydin. *Machine Learning*. MIT Press, 2004.

[4] Geraldine A Auwera, Mauricio O Carneiro, Christopher Hartl, Ryan Poplin, Guillermo del Angel, Ami Levy-Moonshine, Tadeusz Jordan, Khalid Shakir, David Roazen, Joel Thibault, et al. From Fastq data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics*, pages 11–10, 2013.

[5] Christoph Bernau, Markus Riester, Anne-Laure Boulesteix, Giovanni Parmigiani, Curtis Huttenhower, Levi Waldron, and Lorenzo Trippa. Cross-study validation for the assessment of prediction algorithms. *Bioinformatics*, 30(12):i105–i112, 2014.

[6] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[7] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A Albers, Eric Banks, Mark A DePristo, Robert E Handsaker, Gerton Lunter, Gabor T Marth, Stephen T Sherry, et al. The variant call format and vcftools. *Bioinformatics*, 27(15):2156–2158, 2011.

[8] Mark A DePristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, Guillermo del Angel, Manuel A Rivas, Matt Hanna, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, 43(5):491–498, 2011.

[9] Charity Duran and H Criss Hartzell. Physiological roles and diseases of tmem16/anoctamin proteins: are they all chloride channels? *Acta pharmacologica Sinica*, 32(6):685–692, 2011.

[10] HC Erichsen and SJ Chanock. Snps in cancer research and treatment. *British journal of cancer*, 90(4):747–751, 2004.

[11] André P Fay, Sabina Signoretti, and Toni K Choueiri. Met as a target in papillary renal cell carcinoma. *Clinical Cancer Research*, 20(13):3361–3363, 2014.

[12] Robert L Grossman, Allison P Heath, Vincent Ferretti, Harold E Varmus, Douglas R Lowy, Warren A Kibbe, and Louis M Staudt. Toward a shared vision for cancer genomic data. *New England Journal of Medicine*, 375(12):1109–1112, 2016.

[13] Heng Li and Richard Durbin. Fast and accurate short read alignment with burrowswheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

[14] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[15] W Marston Linehan, Ramaprasad Srinivasan, and Laura S Schmidt. The genetic basis of kidney cancer: a metabolic disease. *Nature reviews urology*, 7(5):277–285, 2010.

[16] Iñigo Martincorena and Peter J Campbell. Somatic mutation in cancer and normal cells. *Science*, 349(6255):1483–1489, 2015.

[17] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, and Mark A. DePristo. The Genome Analysis Toolkit: A Mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–1303, 2010.

[18] Nihir Patel, Bharati Jhadav, Abdulrhman Aljouie, and Usman Roshan. Cross-validation and cross-study validation of chronic lymphocytic leukemia with exome sequences and machine learning. In *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on*, pages 1367–1374. IEEE, 2015.

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[20] Usman Roshan, Satish Chikkagoudar, Zhi Wei, Kai Wang, and Hakon Hakonarson. Ranking causal variants and associated regions in genome-wide association studies by the support vector machine and random forest. *Nucleic Acids Research*, 39(9):e62, 2011.

[21] Yohan Seo, Kunhi Ryu, Jinhong Park, Dong-kyu Jeon, Sungwoo Jo, Ho K Lee, and Wan Namkung. Inhibition of ano1 by luteolin and its cytotoxicity in human prostate cancer pc-3 cells. *PloS one*, 12(3):e0174935, 2017.

[22] Doug Speed and David J Balding. Multiblup: improved snp-based prediction for complex traits. *Genome research*, 24(9):1550–1557, 2014.

[23] Cristian Tomasetti, Lu Li, and Bert Vogelstein. Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science*, 355(6331):1330–1334, 2017.

[24] Juliet Usher-Smith, Jon Emery, Willie Hamilton, Simon J Griffin, and Fiona M Walter. Risk prediction tools for cancer in primary care. *British journal of cancer*, 113(12):1645–1650, 2015.

[25] Podchanart Wanitchakool, Luisa Wolf, Gudrun E Koehl, Lalida Sirianant, Rainer Schreiber, Sucheta Kulkarni, Umamaheswar Duvvuri, and Karl Kunzelmann. Role of anoctamins in cancer and apoptosis. *Phil. Trans. R. Soc. B*, 369(1638):20130096, 2014.

[26] Ian R Watson, Koichi Takahashi, P Andrew Futreal, and Lynda Chin. Emerging patterns of somatic mutations in cancer. *Nature reviews Genetics*, 14(10):703–718, 2013.

[27] Naomi R Wray, Jian Yang, Ben J Hayes, Alkes L Price, Michael E Goddard, and Peter M Visscher. Pitfalls of predicting complex traits from snps. *Nature Reviews Genetics*, 14(7):507–515, 2013.

[28] Robert P Young, Fenghai Duan, Erin Greco, Raewyn J Hopkins, Caroline Chiles, Greg D Gamble, and Denise Aberle. Snp-based risk score out performs a clinical model for dying of lung cancer in the nlst-acrin sub-study (n= 10,054). In *C30. LUNG CANCER SCREENING: WHO, WHY, WHERE, AND HOW MUCH*, pages A5172–A5172. Am Thoracic Soc, 2017.