Machine learning based prediction of gliomas with germline mutations obtained from whole exome sequences from TCGA and 1000 Genomes Project

1st Abdulrhman Aljouie*, 2nd Michael Schatz†, 3rd Usman Roshan‡
*Department of Biostatistics and Bioinformatics, King Abdullah International Medical Research Center**
Riyadh, Saudi Arabia, Email: aljouieab@ngha.med.sa
Department of Computer Science and Department of Biology, Johns Hopkins University†
Baltimore, MD, Email: mschatz@cs.jhu.edu
Cold Spring Harbor Laboratory Cold Spring Harbor, NY†
Department of Computer Science, New Jersey Institute of Technology‡
Newark, NJ, Emails: aa547@njit.edu* , usman@njit.edu‡

***Abstract*—Germline variants can be early useful predictors of cancer risk. Here we present cross-study validation and cross-validation of two brain cancers: Gliobastoma Multiforme (GBM) and Lower Grade Glioma (LGG). We obtained whole exome germline sequences of European ancestry individuals with these cancers from The Cancer Genome Atlas and of European ancestry control individuals from the 1000 Genomes Project. We performed a rigorous quality controlled GATK procedure to obtain variants with which we perform cross-study and cross-validation experiments. We find our germline variants to be highly predictive of both cancers in cross-study as well as in cross-validation. Predicting LGG+controls from GBM+controls gives an 89% accuracy and predicting vice versa is 88% accurate both with the linear support vector machine classifier. We find that the main bulk of accuracy comes from the SNP rs10792053 that lies on gene OR9G1. We see that this SNP is in Hardy Weinberg equilibrium and allele frequencies similar to previously published in controls but not so in our cases. Our manual inspection of alignments reveals nothing unusual in the cases. We find our other top ranked SNPs to lie in genes known to be connected to brain cancer and cancer in general. Our study here shows a highly discriminative germline SNP for GBM and LGG cancer but requires replication studies to further verify.**

***Index Terms*—GBM, LGG, 1000 Genomes, Whole Exome Sequencing, Prediction**

I. Introduction

Estimating susceptibility to cancer from germline variants is important for recommending regular screening that helps in early cancer detection, and enhances patient chances of successful treatment. Linkage analysis studies show that gliomas may cluster within families [1]–[4]. Also, many genome-wide association studies have identified germline genomic loci that increase glioma risk [5]–[7].

In our work, we look into the collective germline SNPs predictive ability for brain cancer predisposition. We preform a Genome Analysis Toolkit (GATK) joint germline SNPs discovery workflow for TCGA Gliobastoma Multiforme (GBM) and Lower-Grade Glioma (LGG) white individual cases and 1000 Genomes Project white individual controls . We discarded SNPs that failed GATK Variant Quality Score Recalibration soft filtering or hard filtering (genotype quality $\leq$20, depth $\leq$ 5, or missing genotype) quality control from further machine learning analysis.

On the training set, we excluded SNPs with zero variance and scaled each SNP so that it remains between zero and one. Then, we selected the best K SNPs based on chi-squared test value. For cross-validation, we combined 1K Genomes Project, GBM, and LGG samples and their common SNPs. We split the data into 10-fold (90% for training and 10% for testing) and learned a predictive model with SVM and random forest classifiers. In each training fold, we cross-validated SVM cost hyperparameter and the number of trees to grow for RF with 3-fold for each top K selected SNPs. We then measured the predictive ability with average balanced accuracy across all folds. For cross-study, we ran linear SVM on best K selected SNPs on 50% randomly selected samples from 1K Genomes Project and GBM, and predict LGG and the remaining half of 1K Genomes Project samples.

To confirm that all samples came from the same population and that the classification of cases and controls is not due to ethnicity differences, we performed principles component analysis on the entire dataset (before feature selection) and projected the first two principal components. There we see that the individuals are related. SNPs departure from Hardy-Weinberg Equilibrium (HWE) can be a sign of genotyping error or population stratification. Top SNPs in controls that violate HWE are removed from further machine learning analysis. We used Plink to perform HWE with exact test since using chi-squared test is not suitable for multi-allelic sites.

We show that we can predict GBM and LGG white individual cases and 1000 Genomes Project white individual controls with 90% mean balanced accuracy of 10-fold cross-validation (CV) when learning in best 10 germline variants selected by chi-squared value with support vector machine (SVM) and random forest (RF). In cross-study, learning with GBM+controls and predicting LGG achieved 89% balanced accuracy, and 88% balanced accuracy the other way around.

The most contribution to the accuracy comes from SNP rs10792053. When we removed this SNP cross-validation mean balanced accuracy drops to 54% with top 10 SNPs, and 50% in cross-study. We looked into the original alignments of SNP rs10792053 in cases and controls samples with the Integrative Genomics Viewer (IGV). In both cases and controls, reads coverage and mapping quality at this locus were high.

II. Methods

A. Data

For case individuals, we obtained white normal samples (germline) whole-exome sequencing (WES) data pre-aligned to Genome Reference Consortium Human Build 38 (GRCh38) in binary alignment map (BAM) format from The Cancer Genome Atlas (TCGA) through National Cancer Institute's Genomic Data Commons (GDC) portal for two brain cancer studies (males: 477, females: 331, mean age: 52.08). For control individuals, we downloaded Europeans samples WES pre-alined to GRCh38 in CRAM format from 1000 Genomes Project phase 3 (males: 250, females: 297). In our analysis, we considered only white individuals, to reduce race differences effect on phenotype occurrence. We then performed a variant calling workflow followed by a machine learning pipeline on these samples. Table I summarizes cohort studies used in our analysis. In Table II, we show the number of SNPs for 1K, GBM, and LGG as well as common SNPs after applying soft+hard filtering.

TABLE I
Samples population

Population (sub-population)	Count
1K Genomes Project (CEU)	102
1K Genomes Project (FIN)	105
1K Genomes Project (GBR)	102
1K Genomes Project (IBS)	108
1K Genomes Project (TSI)	112
1K Genomes europeans (all)	**529**
GBM white (not hispanic)	274
GBM white (hispanic)	5
GBM white (not reported)	58
GBM white (all)	**337**
LGG white (not hispanic)	421
LGG white (hispanic)	27
LGG white (not reported)	23
LGG white (all)	**471**

B. Joint genotyping

For germline variant discovery, we used the Genome Analysis Toolkit (GATK) version 4 [8]. GATK HaplotypeCaller variant calling walker produces an intermediate Genomic Variant Call Format (GVCF) file for each sample. We pooled the intermediate GVCF files of all samples together for genotyping by passing it to GATK genotypeGVCFs to obtain a VCF file for samples cohort. Passing samples GVCFs with the whole-exome regions is computationally intensive, to speed up the variants calling workflow we divide each chromosome into roughly 10 equal intervals in a scatter and gather fashion and run it simultaneously on a cluster. Figure 2 illustrates the joint variant discovery workflow. After obtaining the final VCF file, we applied quality control measures to reduce sequencing artifacts and false-positive genotypes.

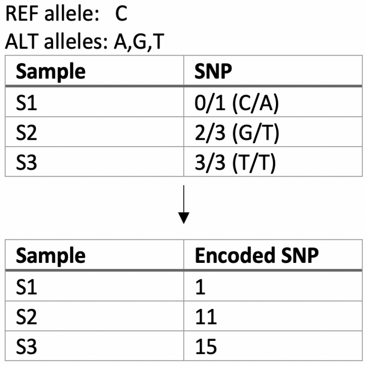
C. SNPs encoding

The output of the GATK GenotypeGVCFs tool is in a VCF format. In the header, it has the reference base (REF), one of A, C, G, T, N bases, and alternate non-reference alleles (ALT) base(s). It is possible but not common to have a multiallelic site (two or more ALT bases). We considered all permutations of genotypes to use as input features to learn a predictive model. A SNP encoding to a numerical value is an essential pre-processing step to machine learning. We encoded each SNP as follow:

$$4 \times A + B \tag{1}$$

where A and B are the two alleles (copies) for a given sample at a particular locus of the genome.

Fig. 1. A toy example for encoding a multiallelic site



REF allele: C
ALT alleles: A,G,T

Sample	SNP
S1	0/1 (C/A)
S2	2/3 (G/T)
S3	3/3 (T/T)

Sample	Encoded SNP
S1	1
S2	11
S3	15

TABLE II
SNPs count after applying soft+hard filtering

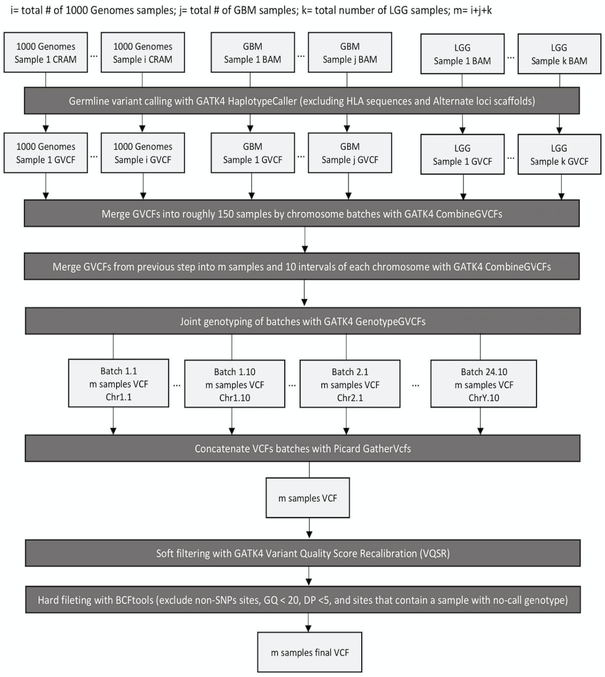
	Number of SNPs
1000 Genomes Project	184690
GBM	297106
LGG	485115
Common SNPs[a]	118439

[a] Intersection of 1000 Genomes+GBM+LGG SNPs.

D. Missing genotypes

In GATK, a genotype with low supporting reads is encoded as "./." to denote no variant call was made at that site for a given sample. Imputation is a method that is commonly used in GWA studies to increase the number of genotypes in the association analysis. Imputation algorithms predict ungenotyped loci in individuals that were genotyped on a subset of loci of SNPs chip to boost SNPs array coverage utilizing haplotype information across samples and HapMap data as an imputation reference panel [9]–[11]. In our study, we excluded column features that have a missing genotype from any sample from further analysis. Thus, we eliminated the need for imputation.

Fig. 2. Germline SNPs calling pipeline


i= total # of 1000 Genomes samples; j= total # of GBM samples; k= total number of LGG samples; m= i+j+k
1000 Genomes Sample 1 CRAM
1000 Genomes Sample i CRAM
GBM Sample 1 BAM
GBM Sample j BAM
LGG Sample 1 BAM
LGG Sample k BAM
Germline variant calling with GATK4 HaplotypeCaller (excluding HLA sequences and Alternate loci scaffolds)
1000 Genomes Sample 1 GVCF
1000 Genomes Sample i GVCF
GBM Sample 1 GVCF
GBM Sample j GVCF
LGG Sample 1 GVCF
LGG Sample k GVCF
Merge GVCFs into roughly 150 samples by chromosome batches with GATK4 CombineGVCFs
Merge GVCFs from previous step into m samples and 10 intervals of each chromosome with GATK4 CombineGVCFs
Joint genotyping of batches with GATK4 GenotypeGVCFs
Batch 1.1 m samples VCF Chr1.1
Batch 1.10 m samples VCF Chr1.10
Batch 2.1 m samples VCF Chr2.1
Batch 24.10 m samples VCF ChrY.10
Concatenate VCFs batches with Picard GatherVcfs
m samples VCF
Soft filtering with GATK4 Variant Quality Score Recalibration (VQSR)
Hard fileting with BCFtools (exclude non-SNPs sites, GQ < 20, DP <5, and sites that contain a sample with no-call genotype)
m samples final VCF

E. Variants calling quality control

GATK HaplotypeCaller by default excludes sites with mapping quality (MAPQ) $\leq$ 20. In our analysis, we used two layers of quality controls: SNPs soft filtering followed by hard filtering to minimize false-positive SNPs. To confirm that the samples came from the same population, we ran principal components analysis (PCA) on the whole dataset before SNPs selection. In Figure 3, the projection of the first two components shows that the samples are related. We removed the two outlier samples and replotted PCA projections of the first two components in Figure 4, and case and control individuals don't form distinct clusters. We used Plink version (1.9) [12] to test for departure from Hardy Weinberg equilibrium with an exact test in control samples. We excluded SNPs that deviate from HWE. Only top SNPs in HWE are included in the analysis, Table III shows the exact test p-values of top 10 SNPs in control individuals from 1000 Genomes Project dataset

TABLE III
HARDY WEINBERG EQUILIBRIUM EXACT TEST PVALUES ON TOP SELECTED 10 SNPS IN CONTROL INDIVIDUALS FROM 1000 GENOMES PROJECT

SNP	Observed het	Expected het	P-value
rs80356578	0.06049	0.05866	1
rs150707706	0.03592	0.03527	1
rs143139551	0.03214	0.03162	1
rs145172249	0.04159	0.04072	1
rs148782546	0.02268	0.02243	1
rs10792053	0.2042	0.2069	0.6774
rs144518683	0.02268	0.02243	1
rs140561687	0.03025	0.02979	1
rs138772802	0.03403	0.03345	1
rs147042091	0.02836	0.02795	1

Fig. 3. Projection of principal component analysis first two components

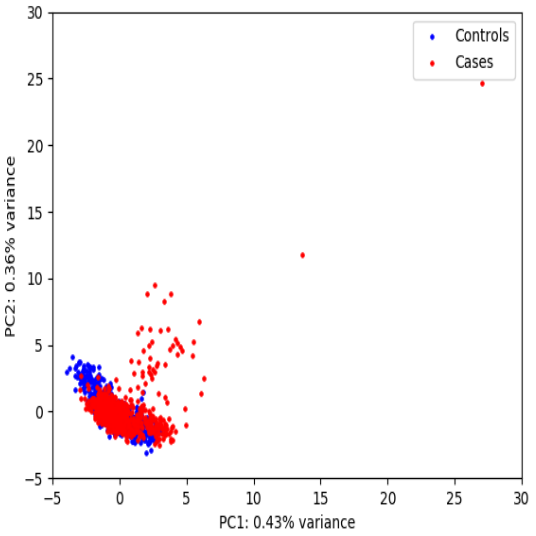
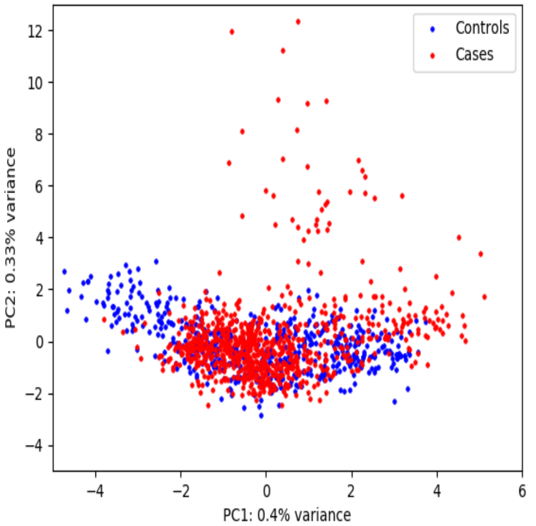


Fig. 4. Projection of principal component analysis first two components after excluding the two outliers



F. Soft filtering

We used the GATK variant quality recalibration score (VQSR) that uses machine learning by training on external databases with known variant sites, and then it assigns a probability score to each variant in the cohort. We set the truth sensitivity filter for VQSR to a 99.0% threshold. We used the following VCF annotations with VQSR to build a recalibration model: InbreedingCoeff, QD, MQ, MQRankSum, ReadPosRankSum, FS, SOR. variants that failed soft filtering are removed from further analysis.

G. *Hard filtering*

At the sample level, variant sites that have genotype quality (GQ) $>$ 20 and depth (DP) $>$ 5 for all samples are considered. DP is the number of reads to support the genotyping, and GQ is a confidence score between 0 and 99, the higher the more confident the program in its assigned genotype. We used BCFtools (version 1.3) [13] for hard filtering and to extract VCF fields into table format.

H. *Soft+hard filtering*

We included only SNPs that passed both soft filtering and hard filtering for further machine learning analysis.

I. *Feature scaling*

Features with zero variance in training split were removed. The remaining features were linearly transformed based on the training subset using Min-Max normalization to keep the data between 0,1 while preserving distance. We used scikit-learn [14] minMaxScaler and the implementation is as follow:

$$X' = \frac{X - X_{j(min)}}{X_{j(max)} - X_{j(min)}} \times (X_{min} - X_{max} + X_{j(min)}) \tag{2}$$

Where X' is the transformed training data, $X_{j(min)}$ and $X_{j(max)}$ is the minimum and maximum values at the j-th SNP in the original data, $X_{min} - X_{max}$ is the SNP range. We applied the exact same transformation to validation data where we determined SNPs min and max from training data only.

J. *Chi-squared features selection*

Top SNPs are selected based on the chi-squared statistic between each SNP and the label. In the chi-squared test, a higher value is an indicator of dependence between the SNP and the label. We ranked SNPs based on their chi-squared value using the scikit-learn chi2 function.

$$\chi^2 = \sum_{i}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{3}$$

where n is the number of classes. O_i is the sum of SNP alleles encoding for the i^{th} class. $f = \sum_i^n O_i$, and $E_i = 1/n \times f$ Table IV shows top chi2-ranked 1K+GBM+LGG common SNPs .

K. *Classifiers*

We used support vector machines (SVM) with linear kernel [15] and random forest (RF) [16] classifiers using scikit-learn package [14].

1) *Support vector machine:* SVM finds a hyperplane that maximizes the distance between classes:

$$\min_{w,w_0} \frac{||w||^2}{2} + C\ max(0, 1 - y_i(w^T x_i + w_0) \tag{4}$$

were x_i is a genotype vector of the i^{th} individual, y_i is the label, w is a weight vector, C is a regularization parameter. $max(0, 1 - y_i(w^T x_i + w_0)$ is the hinge loss and the sign of $(w^T x_i + w_0)$ assigns the input x into class -1 or $+1$. We cross-validated the C hyperparameter with 3-fold cross-validation from the list $(0.1, 1)$.

TABLE IV
TOP SNPS FOR 1K GENOMES, GBM AND LGG

1000 Genomes, GBM and LGG			
Alt allele frequency			
1K Genomes	GBM+LGG	SNP rs ID	Chi2 score
0.0302	0.0068	rs80356578	21.84
0.0180	0.0019	rs150707706	20.15
0.0161	0.0006	rs143139551	22.67
0.0208	0.0006	rs145172249	30.26
0.0113	0	rs148782546	18.33
0.1096	0.4963	rs10792053	50.63
0.0113	0	rs144518683	18.33
0.0151	0	rs140561687	24.44
0.0170	0	rs138772802	27.49
0.0142	0.0006	rs147042091	19.65

TABLE V
TOP SNPS FOR 1K GENOMES AND GBM

1K Genomes and GBM			
Alt allele frequency			
1K Genomes	GBM	SNP rs ID	Chi2 score
0.0047	0.0237	rs140717526	12.28
0.0038	0.0341	rs782010133	12.78
0.0076	0.0312	rs779492064	13.69
0	0.0134	rs202040378	14.13
0.0009	0.0148	rs146032550	12.51
0.0019	0.0386	rs759512484	34.27
0.0009	0.0163	rs76672487	14.05
0.0009	0.0148	rs148088117	12.51
0.1096	0.4926	rs10792053	42.67
0.0019	0.0341	rs768904765	24.25

TABLE VI
TOP SNPS FOR 1K GENOMES AND LGG

1K Genomes and LGG			
Alt allele frequency			
1K Genomes	LGG	SNP rs ID	Chi2 score
0.0302	0.0074	rs80356578	13.30
0.0076	0.0308	rs12721607	14.53
0.0009	0.0159	rs35723440	13.97
0.0208	0	rs145172249	19.59
0.0076	0.0297	rs2232449	13.60
0.0236	0.0032	rs61734485	14.88
0.0085	0.0329	rs2069548	14.84
0.1096	0.4989	rs10792053	45.18
0.0151	0	rs140561687	14.25
0.0170	0	rs138772802	16.03

2) *Random forest:* RF is an ensemble method that builds decision trees by selecting random samples with replacement to construct each tree and randomly generating a subset of features to choose from for each candidate split, the one with the highest Gini impurity or entropy, then it takes the majority vote of trees predictions to output a class prediction. We used the default parameters for the quality measure of the split, and 3-fold cross-validation from the list $(100, 1000)$ for the number of trees to construct.

L. *Performance metrics*

Since classes are imbalanced in the studies included in our analysis, it is inappropriate to used accuracy as a measure of classifiers performance. We used balanced accuracy for performance evaluation. Balanced accuracy is the average of true positive rate and true negative rate.

$$Balanced\ accuracy = \frac{(\frac{true\ positive}{postiive} + \frac{true\ negative}{negative})}{2} \qquad (5)$$

III. Results

A. Cross-validation

With chi-squared statistic best 10 SNPs, linear SVM and RF achieved 90% mean balanced accuracy of 10-fold cross-validation when predicting 1K Genomes controls and GBM+LGG cases. The predictive ability deteriorates when we consider all SNPs to 65% and 54% for SVM and RF, respectively. Figure 6 shows results for predicting three-classes of 1K Genomes, GBM, and LGG with linear SVM, one-vs-one, mean balanced accuracy attained is 68% on top 10 SNPs, however, the accuracy drops to 46% with all SNPs. The accuracy declines as we add more SNPs in both binary and three class classification of glioma subtypes individuals and control individuals.

Fig. 5. 10-fold cross-validation learning and classifying binary labels

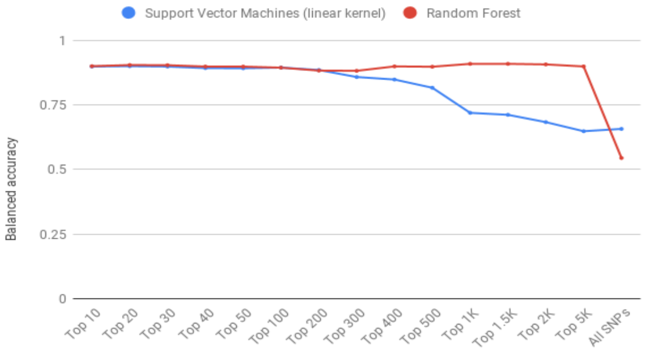
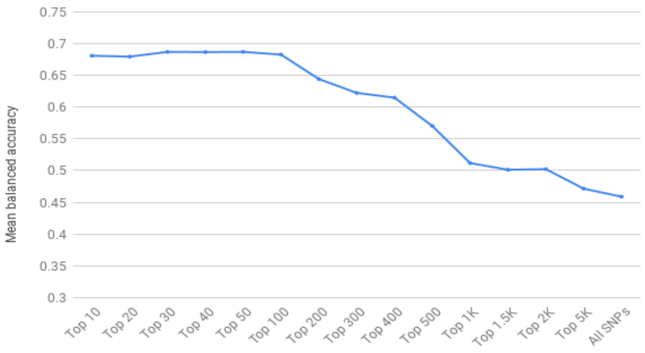


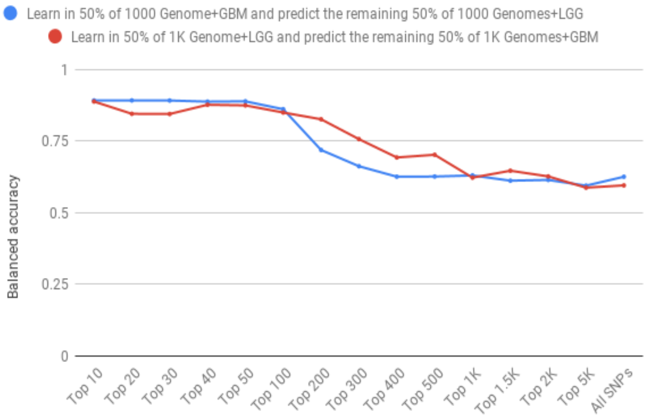
Fig. 6. 10-fold cross-validation of learning and classifying 3-class labels



B. Cross-study validation

To test the generalization of the model, we trained the data on GBM and randomly selected 50% of 1K Genomes samples, and predict the labels of the unseen LGG dataset and the remaining 50% of 1K Genomes samples. Top 10 ranked SNPs obtained the highest balanced accuracy of 89%, again we see the advantage of ranking the SNPs with chi-squared as we included more SNPs to learn the model, where the worst accuracy attained by considering all SNPs with 63%. We also tested the accuracy the other way around, where we learned in LGG and 50% randomly selected samples from 1K Genomes and predict the labels of GBM and the remaining 50% samples of 1K Genomes. We observed the same thing where ranking the SNPs with chi-squared boost the balanced accuracy from 60% with all SNPs to 88% with only 10 SNPs. As expected, ranking SNPs by their dependence on labels improved the balanced accuracy greatly on all cross-validation and cross-study validation experiments.

Fig. 7. Cross-study validation



C. Cancer significance of top ranked SNPs

A point mutation could be nonsynonymous (missense, or nonsense) or synonymous (silent). Missense mutations, which is a change in a single nucleotide that substitutes amino acid encoding and influences protein function [17] [18], are heavily investigated in cancer research because it can alter protein function. Synonymous mutations are often called silent mutations due to their inability to change the amino acid sequence, therefore, these mutations usually are disregarded in cancer research [18]. However, synonymous variants can affect protein folding, and thus it plays a role in cancer [19]. In this work, we investigated both synonymous and nonsynonymous variants. SNPs rs76672487 (in gene ABCC2) and rs2069548 (in gene TG) are cancer-related genes according to The Human Atlas Protein. SNP rs76672487 ranked fifth on the selected SNPs by chi-squared from the GBM+1K dataset, while SNP rs2069548 ranked fourth on 1K+GBM top SNPs. In the 1K+GBM+LGG dataset's top 10 ranked SNPs, six genes are reported by The Human Atlas Protein to be prognostic markers for survival in glioma, liver, renal, cervical, urothelial, pancreatic, and endometrial cancers based on gene expression

FPKM values. Tables VII through IX show top-ranked genes in the 1K+GBM+LGG, 1K+GBM, and 1K+LGG datasets that are prognostic for survival time in cancer. five genes in top-ranked in the 1K+GBM+LGG dataset are expressed in all cancers according to the The Human Atlas Protein. KCNC2 gene is expressed in breast and prostate cancers. P4HA3 gene is expressed in pancreatic, breast, renal, glioma, and lung cancers. Genes OR9G1 and OTOF are not expressed in cancer. Tables X through XII show the genes that are expressed in cancer in top-ranked SNPs in the 1K+GBM+LGG, 1K+GBM, and 1K+LGG datasets.

TABLE VII
1K, LGG, and GBM top ranked SNPs genes expression with significant ($p < 0.001$) association with patient survival

Gene	Survival prognostic marker in cancer
OTOF	No
EAF2	Prognostic marker.
ALPK1	Prognostic marker.
LOC108783645, HFE	No
PTPRJ	Prognostic marker.
OR9G1	No
P4HA3	Prognostic marker.
ATF7IP	Prognostic marker.
PLBD1	Prognostic marker.
KCNC2	No

TABLE VIII
1K and GBM top ranked SNPs genes expression with significant ($p < 0.001$) association with patient survival

Gene	Survival prognostic marker in cancer
SARS	Prognostic marker
CA14	No
LHX9	No
DGKG	No
OSMR	Prognostic marker.
DMXL1	Prognostic marker.
ABCC2	Prognostic marker.
OR56B4	No
OR9G1	No
ZNF641	Prognostic marker.

TABLE IX
1K and LGG top ranked SNPs genes expression with significant ($p < 0.001$) association with patient survival

Gene	Survival prognostic marker in cancer
OTOF	No
NR1I2	No
IGSF10	No
LOC108783645, HFE	No
MICAL1, ZBTB24	Prognostic marker.
CA1	No
TG	No
OR9G1	No
ATF7IP	Prognostic marker.
PLBD1	Prognostic marker.

TABLE X
Top SNPs for 1K Genomes, GBM and LGG genes and functional consequences

1K Genomes, GBM, and LGG			
rs ID	Gene	Functional consequence	Cancer mRNA expression
rs80356578	OTOF	synonymous	Not detected
rs150707706	EAF2	missense	Expressed in all
rs143139551	ALPK1	missense	Expressed in all
rs145172249	HFE	intron variant	Expressed in all
rs148782546	PTPRJ	synonymous	Expressed in all
rs10792053	OR9G1	synonymous	Not detected
rs144518683	P4HA3	synonymous	Mixed
rs140561687	ATF7IP	missense	Expressed in all
rs138772802	PLBD1	intron	Expressed in all
rs147042091	KCNC2	missense	Group enriched

TABLE XI
Top SNPs for 1K Genomes, GBM genes and functional consequences

1K Genomes and GBM			
rs ID	Gene	Functional consequence	Cancer mRNA expression
rs140717526	SARS	missense	Expressed in all
rs782010133	CA14	missense	Group enriched
rs779492064	LHX9	intron	Mixed
rs202040378	DGKG	intron	Tissue enhanced
rs146032550	OSMR	synonymous	Expressed in all
rs759512484	DMXL1	missense	Expressed in all
rs76672487	ABCC2	intron	Tissue enhanced
rs148088117	OR56B4	missense	Not detected
rs10792053	OR9G1	synonymous	Not detected
rs768904765	ZNF641	intron	Expressed in all

TABLE XII
Top SNPs for 1K Genomes, LGG genes and functional consequences

1K Genomes and LGG			
rs ID	Gene	Functional consequence	Cancer mRNA expression
rs80356578	OTOF	synonymous	Not detected
rs12721607	NR1I2	missense	Group enriched
rs35723440	IGSF10	synonymous	Mixed
rs145172249	HFE	intron	Expressed in all
rs2232449	ZBTB24	synonymous	Expressed in all
rs61734485	CA1	missense	Group enriched
rs2069548	TG	missense	Tissue enriched
rs10792053	OR9G1	synonymous	Not detected
rs140561687	ATF7IP	missense	Expressed in all
rs138772802	PLBD1	intron	Expressed in all

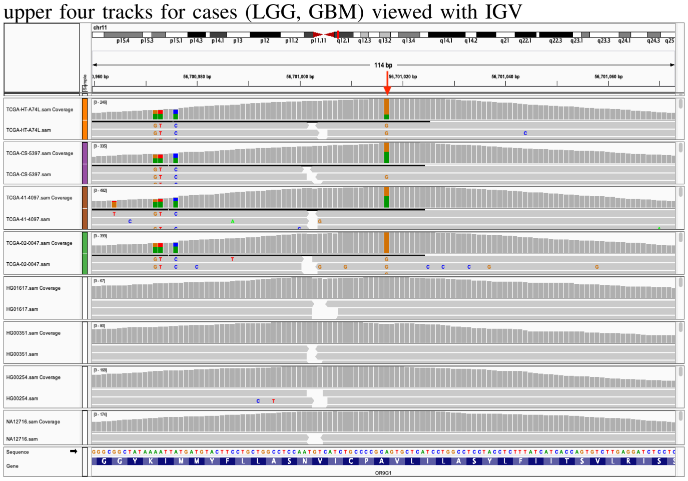
D. SNP rs10792053 mapping quality

To confirm that there is no issue with reads mapping quality or coverage, we inspected eight individuals from both cases and controls alignments with Integrative Genomics Viewer (IGV) with the original reads mapping at locus 11:56701017 and its adjacent loci.

Figure 8 shows alignments with IGV of four samples from cases vs four from controls against the GRCh38 reference genome. The red arrow in Figure 8 points SNP rs10792053 position. The tangerine color in the tracks at the position refers to allele C and the green refers to reference allele A.

In IGV, if both allele copies in the sample is homozygous reference, then it is shown in gray. Three of the four cases viewed are heterozygous and the remaining one is homozygous alternate allele. All controls in the figure are homozygous reference. In IGV we set the mapping quality threshold to 1 since GATK HaplotypeCaller discards reads with a mapping quality of 0. The original alignments of both cases and controls have high coverage at this location. Although GATK HaplotypeCaller reassembles alignments at active regions and discards original alignments, the final VCF is consistent with what we observed in original alignments. For SNP rs10792053, the average depth across all cases is 407.15 and across all controls is 63.58. These average depths are after running the GATK germline variant discovery workflow. We tested for Hardy Weinberg equilibrium exact test in controls individual and the p-value is 0.677, which confirms that this SNP is in HWE, however, it is out of HWE in cases.

Fig. 8. Alignments of four cases vs four controls at SNP rs10792053 the upper four tracks for cases (LGG, GBM) viewed with IGV



E. Alternate allele frequency of top SNPs

Table XIII shows the alternate allele frequency of dbSNP 1K Europeans samples, GBM, LGG and 1K samples that are considered for our study, which is slightly larger than 1K Genomes sample size in dbSNPs since we downloaded samples from 1000 Genomes Project phase 3. For example, SNP rs80356578 sample size in dbSNP is 503 and the sample size for our 1K Genomes is 526. Our alternate allele frequency is close to what is reported by dbSNP for 1000 Genomes Project Europeans samples.

IV. Conclusion

We show that we can predict glioma cases with few germline SNPs selected based on the chi-squared statistics with 90% mean balanced accuracy in cross-validated TCGA GBM and LGG white individual cases and 1000 Genomes Project Europeans controls whole-exome sequences with linear SVM and random forest. We also show that in cross-study linear SVM achieves 89% predictive accuracy when learning with GBM+controls and predicting LGG and 88% contrariwise on the top-ranked germline SNPs. Most of the accuracy comes from SNP rs10792053, a replication study is needed to verify its discriminative power in glioma.

TABLE XIII
1K, GBM, and LGG top ranked SNPs alternate allele frequencies

rs ID	dbSNP (EUR)	Controls	Cases
rs80356578	A=0.029	0.0302	0.0068
rs150707706	C=0.019	0.0179	0.0018
rs143139551	A=0.017	0.0160	0.0006
rs145172249	C=0.019	0.0207	0.0006
rs148782546	T=0.012	0.0113	0
rs10792053	G=0.116	0.1096	0.4962
rs144518683	C=0.012	0.0113	0
rs140561687	T=0.016	0.0151	0
rs138772802	C=0.017	0.0170	0
rs147042091	C=0.013	0.0141	0.0006

Acknowledgment

We thank NJIT Academic and Research Computing Services for computational support. We thank Professor Zhi Wei for helpful discussion and feedback on our results.

References

[1] S. Sadetzki, R. Bruchim, B. Oberman, G. N. Armstrong, C. C. Lau, E. B. Claus, J. S. Barnholtz-Sloan, D. Ilyasova, J. Schildkraut, and C. e. a. Johansen, "Description of selected characteristics of familial glioma patients results from the gliogene consortium," *European Journal of Cancer*, vol. 49, no. 6, pp. 1335–1345, 2013.

[2] M. L. Goodenberger and R. B. Jenkins, "Genetics of adult glioma," *Cancer Genetics*, vol. 205, no. 12, pp. 613–621, 2012.

[3] B. Malmer, P. Adatto, G. Armstrong, J. Barnholtz-Sloan, J. L. Bernstein, E. Claus, F. Davis, R. Houlston, D. Il'yasova, and R. e. a. Jenkins, "Gliogene an international consortium to understand familial glioma," 2007.

[4] N. Paunu, P. Lahermo, P. Onkamo, V. Ollikainen, I. Rantala, P. Heln, K. Simola, J. Kere, and H. Haapasalo, "A novel low-penetrance locus for familial glioma at 15q23-q26.3," 2002. [Online]. Available: http://cancerres.aacrjournals.org/content/62/13/3798.article-info

[5] K. Labreche, B. Kinnersley, G. Berzero, A. L. Di Stefano, A. Rahimian, I. Detrait, Y. Marie, B. Grenier-Boley, K. Hoang-Xuan, and J.-Y. e. a. Delattre, "Diffuse gliomas classified by 1p/19q co-deletion, tert promoter and idh mutation status are associated with specific genetic risk loci," *Acta Neuropathologica*, vol. 135, no. 5, pp. 743–755, 2018.

[6] T. Rice, D. H. Lachance, A. M. Molinaro, J. E. Eckel-Passow, K. M. Walsh, J. Barnholtz-Sloan, Q. T. Ostrom, S. S. Francis, J. Wiemels, and R. B. e. a. Jenkins, "Understanding inherited genetic risk of adult glioma - a review," *Neuro-Oncology Practice*, vol. 3, pp. 10–16, 2016.

[7] Y. Liu, S. Shete, F. J. Hosking, L. B. Robertson, M. L. Bondy, and R. S. Houlston, "New insights into susceptibility to glioma," *Archives of Neurology*, vol. 67, no. 3, 2010.

[8] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, and M. e. a. Daly, "The genome analysis toolkit: A mapreduce framework for analyzing next-generation dna sequencing data," *Genome Research*, vol. 20, no. 9, pp. 1297–1303, 2010.

[9] J. Marchini and B. Howie, "Genotype imputation for genome-wide association studies," *Nature Reviews Genetics*, vol. 11, no. 7, pp. 499–511, 2010.

[10] C. C. A. Spencer, Z. Su, P. Donnelly, and J. Marchini, "Designing genome-wide association studies: Sample size, power, imputation, and the choice of genotyping chip," *PLoS Genetics*, vol. 5, no. 5, p. e1000477, 2009.

[11] B. N. Howie, P. Donnelly, and J. Marchini, "A flexible and accurate genotype imputation method for the next generation of genome-wide association studies," *PLoS Genetics*, vol. 5, no. 6, p. e1000529, 2009.

[12] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, and M. J. e. a. Daly, "Plink: A tool set for whole-genome association and population-based linkage analyses," *The American Journal of Human Genetics*, vol. 81, no. 3, pp. 559–575, 2007.

[13] H. Li, "A statistical framework for snp calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data," *Bioinformatics*, vol. 27, no. 21, pp. 2987–2993, 2011.

[14] F. Pedregosa, G. Buitinck, R. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, and R. e. a. Weiss, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, p. 28252830, 2011.

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

[16] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[17] V. Gotea, J. J. Gartner, N. Qutob, L. Elnitski, and Y. Samuels, "The functional relevance of somatic synonymous mutations in melanoma and other cancers," *Pigment Cell & Melanoma Research*, vol. 28, no. 6, pp. 673–684, 2015.

[18] N. Deng, H. Zhou, H. Fan, and Y. Yuan, "Single nucleotide polymorphisms and cancer susceptibility," *Oncotarget*, vol. 8, no. 66, 2017.

[19] F. Supek, B. Miana, J. Valcrcel, T. Gabaldn, and B. Lehner, "Synonymous mutations frequently act as driver mutations in human cancers," *Cell*, vol. 156, no. 6, pp. 1324–1335, 2014.