The Germline Variants rs61757955 and rs34988193 are Predictive of Survival in Low Grade Glioma Patients

Ajay Chatrath[1], Manjari Kiran[1], Pankaj Kumar[1], Aakrosh Ratan[2], and Anindya Dutta[1]

[1] Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, Virginia

[2] Center for Public Health Genomics, University of Virginia, Charlottesville, Virginia

Running Title:
Germline Variants are Predictive of Survival in LGG Patients

Keywords: Low Grade Glioma, Germline Variant, Single Nucleotide Polymorphism, Survival Analysis, *GRB2*, *ANKDD1a*

Financial Support:
This work was supported by grants from the NIH R01 CA166054 and CA60499 and a Cancer Genomics Cloud Collaborative Support grant. The Seven Bridges Cancer Genomics Cloud has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, Contract No. HHSN261201400008C and ID/IQ Agreement No. 17X146 under Contract No. HHSN261201500003I.

Corresponding Author:
Anindya Dutta
University of Virginia School of Medicine
1240 Jordan Hall, Charlottesville, VA 22908.
Phone: 434-924-1227
Fax: 434-924-5069
E-mail: ad8q@virginia.edu

Potential Conflicts of Interest:
The authors do not have any conflicts of interest to disclose.

Abstract Word Count: 250
Manuscript Word Count: 5,660
Number of Figures/Tables: 7
Number of References: 50
Number of Supplementary Figures/Tables: 8

Abstract

Low grade gliomas are invasive brain tumors that are difficult to completely resect neurosurgically. They often recur following resection and progress, resulting in death. Although previous studies have shown that specific germline variants increase the risk of tumor formation, no previous study has screened many germline variants to identify variants predictive of survival in glioma patients. In this study, we present an approach to identify the small fraction of prognostic germline variants from the pool of over four million variants that we variant called in The Cancer Genome Atlas whole exome sequencing and RNA sequencing datasets. We identified two germline variants that are predictive of poor patient outcomes by Cox regression, controlling for eleven covariates. rs61757955 is a germline variant found in the 3' UTR of *GRB2* associated with increased *KRAS* signaling, *CIC* mutations, and 1p/19q co-deletion. rs34988193 is a germline variant found in the tumor suppressor gene *ANKDD1a* that causes an amino acid change from lysine to glutamate. This variant was found to be predictive of poor prognosis in two independent low grade glioma datasets and is predicted to be within the top 0.06% of deleterious mutations across the human genome. The wild type residue is conserved in all 22 other species with a homologous protein.

Implications: This is the first study presenting an approach to screening many germline variants to identify variants predictive of survival and our application of this methodology revealed the germline variants rs61757955 and rs34988193 as being predictive of survival in low grade glioma patients.

Introduction

Grade II and grade III (low grade) gliomas are primary brain tumors that are derived from glial cells and include astrocytomas and oligodendrogliomas. They are most commonly found in the cerebral hemispheres. They are highly invasive and therefore difficult to completely resect neurosurgically without significant patient morbidity. Following surgery, patients are typically treated with chemotherapy and radiation, though these tumors typically recur or progress to grade IV gliomas and are fatal.[1] The median survival following low grade glioma diagnosis is around 7 years.[2]

While the 2007 World Health Organization's (WHO) classification of central nervous system neoplasms differentiated between neoplasms primarily based on histological features, the updated 2016 WHO classification system now utilizes both molecular and histological parameters.[1] Isocitrate dehydrogenase mutation (*IDH*) status, 1p/19q co-deletion status, telomerase reverse transcriptase (*TERT*) promoter mutation status, *MGMT* promoter methylation, *TP53* mutation status, and *ATRX* mutation status may be used to molecularly characterize gliomas.[1,3] The availability of genomic data from patient glioma samples from groups such as The Cancer Genome Atlas (TCGA), the Chinese Glioma Genome Atlas (CGGA), and the Ivy Glioblastoma Atlas Project has substantially contributed to our understanding of these tumors.[4,5]

Many studies have utilized these datasets to identify gene expression signatures, microRNA expression patterns, somatic mutation status, and imaging characteristics that are predictive of survival in low grade gliomas.[6–8] While studies have shown that germline mutations can increase an individual's susceptibility for specific cancers,[9–12] including a recent study that identified 853 pathogenic or likely pathogenic germline

93 variants found in 8% of 10,389 cancer patients,[13] no study has comprehensively screened
94 all of the germline variants in a given cancer type to discover the prognostic variants in
95 that cancer type. Although germline mutations have been shown to be prognostic in
96 breast cancer[14] and medulloblastoma[9] in genes that have been well-characterized in the
97 context of these cancers, these variants were not identified using an unbiased approach
98 that screened a large number of germline variants. Identifying prognostic germline
99 variants is challenging due to the limited effect size of germline variants, the large
100 number of germline variants, and confounding clinical factors that may be associated
101 with germline variants. Here we present a novel methodology for identifying prognostic
102 germline variants and report two germline variants that we have found to be associated
103 with survival in low grade glioma patients.

104 **Methods**

105 **Glioma Datasets**

106 491whole exome sequenced normal blood samples (WXS normal), 503 whole
107 exome sequenced tumor samples (WXS tumor), and 501 RNA sequenced tumor samples
108 (RNA tumor) from TCGA low grade glioma[4] patients available on the Cancer Genomics
109 Cloud (CGC)[15] platform were used as part of this analysis. The clinical information was
110 downloaded directly from the TCGA data portal using the GenomicDataCommons
111 (https://bioconductor.org/packages/release/bioc/html/GenomicDataCommons.html) R
112 package available through Bioconductor. Additional molecular characteristics about these
113 TCGA patients were acquired by downloading the supplement from Ceccarelli et. al.[16]
114 The raw sequencing data from the Chinese Glioma Genome Atlas patients was
115 downloaded using accession number SRP027383 from the Sequence Read Archive.

116 Clinical information for these patients was downloaded directly from the project's
117 website (http://www.cgga.org.cn/).

118 **Variant Calling**

119 Variant calling was performed on the TCGA low grade glioma whole exome
120 sequenced normal blood samples (WXS normal), whole-exome sequenced tumor samples
121 (WXS tumor), and RNA sequenced tumor samples (RNA tumor) using VarDict[17] on
122 CGC. The VarDict settings were set at default except for requiring mapping quality
123 greater than 30, base quality greater than 25, a minimum of 3 variant reads, minimum
124 allele frequency of 5%, and the removal of duplicate reads. We compiled a list of all of
125 the unique variants and ran 'samtools[18] depth' on all sequencing files requiring a
126 mapping quality greater than 30. We determined the status of each variant in each patient
127 from the three datasets (WXS normal sample, WXS tumor sample, and RNA tumor
128 sample). The variant status at positions with fewer than ten reads for a given patient was
129 changed to unknown. We used the WXS tumor samples to insert variant calls into the
130 WXS normal samples at positions at which a variant status was listed as unknown in the
131 WXS normal samples. If the variant status was still missing in a given patient, we then
132 used the RNA tumor sample to insert variant calls into the combined WXS variant call
133 set, allowing us to create the combined set of variant calls.

134 The same program parameters and approach were used to variant call and process
135 the CGGA RNA sequencing dataset. All computation on the CGGA dataset was
136 performed locally and not on CGC.

137 **Quality Control**

138 We used annovar[19] to determine the allele frequencies of the variants called by
139 VarDict as listed in gnomAD (http://gnomad.broadinstitute.org/). We calculated the allele
140 frequency of the variants in our study using the following formula:
141

$$\frac{2 * Number\ of\ Minor\ Allele\ Homozygotes + Number\ of\ Heterozygotes}{2 * Total\ Number\ of\ Patients}$$

142
143 The R package GGally (https://cran.r-project.org/web/packages/GGally/index.html) was
144 used to calculate the correlation between the four variant call sets and to display their
145 correlations with each other. Only variants with an allele frequency of greater than 5% in
146 gnomAD and found in 15 or more of the TCGA low grade glioma patients were tested for
147 an association with survival by Cox regression.
148 Because we used the WXS tumor and RNA tumor samples to fill in missing
149 variant calls, we evaluated whether somatic mutations were affecting the validity of our
150 results. We first determined the percentage of variants called in the WXS tumor sample
151 that were somatic mutations. To do this, we downloaded the set of somatic mutations
152 generated by the TCGA Research Network.[20] We then calculated the number of somatic
153 mutations called in each patient in this variant call set and divided that number by the
154 total number of variants called in that patient's WXS normal sample. To assess whether
155 somatic mutations were affecting the integrity of our results, we counted the number of
156 times that a somatic mutation called by the TCGA Research Network overlapped with the
157 set of germline variants that we were testing for an association with survival.
158 Since we used the RNA tumor sample to fill in missing variant calls, we evaluated
159 whether RNA editing was having a significant impact on our analysis. To do this, we
160 downloaded the set of over 2.5 million known RNA editing sites from a rigorously

161 annotated database of RNA editing sites, RADAR.[21] We counted the number of times
162 that the germline variants that we were testing for an association with survival overlapped
163 with any of the known 2.5 million RNA editing sites.

164 **Principal Component Analysis**

165 In order to calculate principal components that could separate patients on the basis
166 of race, we used PLINK[22] to create a pruned set of germline variants to avoid bias from
167 variants in linkage disequilibrium. Pruning was performed using a window size of 50
168 variants and a variance inflation factor of 2. These variants were used to calculate
169 principal components using base R.

170 **Cox Regression and Receiver Operator Characteristic Curves**

171 Lasso in the R package glmnet[23] was run on 17 covariates (**Table 1**). Information
172 about patient age, gender, tumor location, grade, treatment site, and TP53 mutation status
173 was acquired from the TCGA data portal, while data for patient somatic mutation count,
174 percent aneuploidy, TERT expression, IDH mutation status, 1p/19q co-deletion status,
175 MGMT promoter methylation status, and chromosome 7 gain with chromosome 10 loss
176 status was acquired from Ceccarelli et. al.[16] The principal components were calculated as
177 described above. 11 of these 17 covariates were selected for inclusion in the final model
178 for survival prediction. The R packages survival[24] and survminer[25] were used to run Cox
179 regression and create Kaplan-Meier curves. For each minor allele, we our model tested
180 whether the minor allele was associated with a difference in survival outcomes with
181 respect to the reference allele. False discovery rate correction was performed through
182 Bonferroni correction.

Receiver operator characteristic (ROC) curves were created and evaluated using the survivalROC (https://cran.r-project.org/web/packages/survivalROC/survivalROC.pdf) and pROC (https://cran.r-project.org/web/packages/pROC/pROC.pdf) R packages. In order to test whether rs61757955 significantly improves the survival model consisting of the eleven covariates selected by Lasso, we compared the two ROC curves using the bootstrap method with 1000 iterations. We also used this bootstrapping approach to determine whether ANKDD1a expression levels, GRB2 expression levels, rs61757955, and rs34988193 together improve the survival model with respect to the eleven covariates selected by Lasso.

RNA-Sequencing Data Processing

We downloaded the HTSeq FPKM quantification files for each patient from the Genomic Data Commons data portal. We only used gene quantification files from primary tumor samples as part of this analysis. Replicate samples from a single patient were averaged.

Variant Correlation to Covariates and Somatic Mutations

In order to test for associations between the germline variants and genomic and histological tumor characteristics, we divided patients based on their germline variant status. We used the Wilcoxon rank-sum test to test for significant differences in each of the continuous variables between patients with and without a given variant. We used Fisher's exact test to test for differences in each of the discrete variables using a similar approach. Somatic mutation calls were downloaded from Ellrott et. al.[20]

Gene Set Enrichment Analysis

206 Gene set enrichment analysis (GSEA) of mRNA changes associated with
207 rs61757955 and rs34988193 was performed by dividing the patients into two groups for
208 each variant based on whether or not they had the reference allele at the position of the
209 variant. For each germline variant, we calculated the log fold change for all genes
210 expressed greater than one fragment per kilobase per million mapped reads (FPKM)
211 between patients with the variant and without the variant. For each gene, fold change was
212 calculated by dividing the median expression of the gene in patients with the variant by
213 the median expression of the gene in patients without the variant. We used the log fold
214 change to rank the genes from greatest log fold change to smallest log fold change. This
215 file was used as input for GSEA.[26]

216 **Variant Annotation**

217 In order to identify deleterious mutations, we annotated all variants by combined
218 annotation dependent depletion (CADD) scores and only analyzed the variants predicted
219 to be within the top 0.1% of all deleterious variants (CADD > 30).[27] This led us to
220 identify rs34988193 in *ANKDD1a* as a potentially deleterious variant predictive of
221 survival. Because rs34988193 causes an amino acid change from positively charged
222 lysine to negatively charged glutamate, we ran a BLASTp
223 (httpps://blast.ncbi.nlm.nih.gov/Blast.cgi) search so that we could determine how many
224 species have a protein homologous to ANKDD1a and how consistently the wild type
225 lysine residue was conserved. We identified homologous sequences in 22 other species.
226 These sequences were aligned using ClustalW in MEGA.[28] We also annotated this
227 variant with its PhyloP score.[29] Because the crystal structure for *ANKDD1a* was not
228 available, we downloaded the predicted model for this protein from Modbase

229 (https://modbase.compbio.ucsf.edu/modbase-cgi/index.cgi) and calculated the Gribskov
230 score using prophecy on EMBOSS.[30] We retrieved linked variants from Ensembl using
231 the population of Utah residents with Northern and Western European ancestry which is
232 demographically similar to the TCGA low grade glioma patient population.

233 **Results**

234 **Identification of High Quality Germline Variants**

235 Our variant calling pipeline is shown in **Figure 1**. Briefly, we used the variant
236 caller VarDict on Cancer Genomics Cloud to identify variants from whole exome
237 sequencing (WXS) and RNA sequencing samples in about 500 low grade glioma
238 patients. In total, we found 4,453,701 unique variants. We used 'samtools depth' to
239 determine the sequencing depth at each of these variants for each patient and changed the
240 variant status to 'unknown' for patients with sequencing coverage less than 10 reads at a
241 given position. We created a set of combined variant calls by using the WXS and RNA
242 tumor samples to fill in unknown values in the whole exome sequenced normal samples
243 that resulted from having a sequencing coverage of less than 10 reads at a given position.
244 This approach increased our sample size and enabled us to include many more variants in
245 our analysis than if we had solely used variant calls from the whole exome sequenced
246 normal blood samples. Ultimately, this left us with four sets of variants – WXS normal,
247 WXS tumor, RNA tumor, and a combined set that resulted from merging the other three
248 variant call sets, giving preference to the WXS normal and then WXS tumor variant calls.
249 We used the combined variant call set when testing variants for an association with
250 survival. We only tested variants found in 15 or more low grade glioma TCGA patients
251 and listed in gnomAD as having an allele frequency of greater than 5%.

252 **Tumor Variant Calls are not Significantly Affected by Somatic Mutations or RNA**
253 **Editing After Filtering**

254 Because we used sequencing data from the WXS tumor and RNA tumor samples
255 to fill in missing calls in the WXS normal samples, we evaluated our variant calls for
256 contributions from somatic mutations and RNA editing. We first showed that the
257 majority of variant calls in the tumor sample are germline variant calls. To do this, we
258 counted the number of somatic mutations called by the TCGA Research Network's
259 analysis in each patient and divided that number by the number of variants that we called
260 in the WXS normal sample.[20] The median number of somatic mutations called per patient
261 was 39. The median number of variants called in the WXS normal sample was 95,794.
262 We therefore estimated that over 99.9% of variants called in the WXS tumor sample
263 consisted of germline variants and that the percentage of somatic mutations in the WXS
264 tumor sample across all patients was quite small (**Figure S1**). Because somatic mutations
265 rarely occur at the same position, we suspected that the number of somatic mutations
266 included in our study was extremely small since we limited our analysis to variants found
267 in 15 or more of the low grade glioma patients and found in gnomAD with an allele
268 frequency of greater than 5%. Indeed, only one of the 196,022 variants that we tested
269 overlapped with a somatic mutation. This somatic mutation occurred in only a single
270 patient (**Table S1**). Ultimately, we did not find any evidence to suggest that somatic
271 mutations were impacting the quality of our analysis.

272 We next determined whether RNA editing was affecting our analysis by
273 downloading the 2.5 million known RNA editing sites from the rigorously annotated
274 RNA editing database, RADAR.[21] Only 215 of the 196,022 variants that we tested were

275 located at a position that overlapped with a known RNA editing site. We did not find any
276 of these variants to be prognostic as part of our analysis. We therefore did not find any
277 empirical evidence to suggest that somatic mutations or RNA editing impacted our
278 findings (**Table S1**).

279 Finally, we established that our four variant call sets (WXS normal, WXS tumor,
280 RNA tumor, and combined) were concordant with each other by calculating the allele
281 frequency of each variant called in the four sets and demonstrating a very strong
282 correlation between all pairs of variants ($r > 0.98$ for all pairs, **Figure S2**). To further
283 evaluate the quality of our variants calls, we calculated the frequency of each allele and
284 compared it to the frequency of these alleles as listed in gnomAD. Our alleles frequencies
285 were well correlated with gnomAD ($r > 0.93$ for all four variant sets, **Table S2**). As
286 expected, the distribution of allele frequencies is negatively skewed as the majority of the
287 identified variants are rare (**Figure S2**). We used the variants from the WXS normal
288 samples to determine the principal components. As expected, these principal components
289 effectively separate patients on the basis of reported race (**Figure S3**).

290 **Identification of 271 Prognostic Germline Variants that are Independent of Clinical**
291 **Covariates**

292 In order to identify clinically relevant germline variants, we restricted our analysis
293 to variants found in at least 15 patients in the TCGA dataset and found in gnomAD with
294 an allele frequency of greater than five percent. This restricted our analysis to 196,022
295 testable variants (**Figure 2A**). In order to reduce the risk of identifying variants that are
296 prognostic because they are confounded by other covariates known to be associated with
297 survival, we used the machine learning algorithm Lasso to determine which of 17

298 covariates should be controlled for in our Cox regression model. Lasso regression was
299 useful in the screening of these 17 covariates because it penalizes models based on the
300 number of coefficients, allowing for the elimination of less predictive coefficients from
301 the model. The algorithm selected 10 covariates known to be associated with differences
302 in survival in low grade glioma (age, somatic mutation count, percent aneuploidy,
303 histological subtype of astrocytoma, tumor grade, treatment site, *IDH* mutation status,
304 1p/19q co-deletion status, *MGMT* promoter methylation status, chromosome 7
305 gain/chromosome 10 loss status) along with the third principal component that we
306 calculated (**Table 1**). Although the first two principal components are more effective in
307 stratifying patients on the basis of race than the third principal component, the selection
308 of the third principal component over the first two suggests that the third principal
309 component contributes more information to the survival model than the first two principal
310 components. This third principal component primarily separates African Americans from
311 each other, suggesting that a subpopulation of African Americans experienced worse
312 clinical outcomes in this dataset compared to other groups. We ran Cox regression on all
313 196,022 variants one at a time, controlling for these 11 covariates, to identify germline
314 variants predictive of survival.

315 We identified 271 germline variants that are predictive of survival ($p < 0.001$)
316 (**Figure 2A**). As is the case with germline variants in general, the majority of these
317 germline variants are found in protein-coding genes (**Figure 2B**), are located in introns
318 (**Figure 2C**), and are single nucleotide polymorphisms (**Figure 2D**). Most single
319 nucleotide polymorphisms are transitions (**Figure 2E**).

320 **The Germline Variant rs61757955 in *GRB2* is Associated with Poor Prognosis**

321 We identified two germline variants that are highly predictive of survival after
322 false discovery rate correction (FDR < 0.10) (**Figure 3A, Table 2A**). rs61757955 results
323 in a mutation in the 3' UTR of Growth Factor Receptor Bound Protein 2 (*GRB2*) and is
324 associated with a poor prognosis (p=7.08E-10, hazard ratio(HR)=20.4, **Figure 3B, Table**
325 **2A**). To determine whether rs61757955 enhances the survival model compared to the
326 eleven clinical covariates alone, we calculated a risk score for each patient using a Cox
327 regression model with rs61757955 and the other 11 covariates and a risk score using the
328 11 covariates alone. Using these risk scores, we determined the rate at which a patient
329 would be correctly labeled as alive or dead at 7 years with a given false positive rate to
330 create a receiver operator characteristic curve. The increased area under the curve
331 suggests that rs61757955 enhances the survival model compared to the eleven clinical
332 covariates alone (p=0.0489, **Figure 3C**). The allele frequency of rs61757955 is close to
333 0% according to the 1000 Genomes Project[31] in the Chinese population and, as expected,
334 did not show up in the Chinese Glioma Genome Atlas. We also found rs28672782, a
335 germline variant found in the intron of BRSK2, to be associated with a favorable
336 prognosis, though the testable sample size for this variant was small and the maximum
337 follow up for patients with this variant was only three years. Therefore, we did not
338 investigate this variant further (**Figure S4, Table 2A**).

339 In order to test whether rs61757955 in *GRB2* is associated with an increased risk
340 of other genomic abnormalities, we separated patients on the basis of this variant to see if
341 there was a difference in the incidence of the genomic or histological variables (**Table 3**).
342 We found this variant to be associated with an increased incidence of 1p/19q co-deletions
343 (p=0.038). Because 1p/19q co-deletions are frequently seen in Capicua transcriptional

344 repressor (*CIC*) mutated gliomas[32] and *CIC* aberrations are known to be a driver in low
345 grade glioma tumorigenesis,[33] we tested whether there was a difference in the incidence
346 of *CIC* mutations in patients with this variant. 38% of patients with this variant had *CIC*
347 mutated gliomas, whereas only 16% of patients without the variant had a *CIC* mutation
348 (p=0.0168, **Table 3**). Although the incidence of oligodendrogliomas was elevated in
349 patients with the variant compared to patients without the variant, consistent with reports
350 from the literature that 1p/19q co-deletions and *CIC* mutations are enriched in
351 oligodendrogliomas,[32] this difference was not statistically significant (p=0.475). Since
352 rs61757955 is in a non-coding region, we also tested whether this variant is associated
353 with differences in gene expression. We separated patients based on their variant status
354 and calculated the log fold change of each gene between patients with the variant and
355 patients without the variant. This data was used as the input for gene set enrichment
356 analysis (GSEA). We found rs61757955 to be associated with increased *KRAS* signaling
357 (FDR=0.015) (**Figure 3D**).

358 Because we only have whole exome sequencing and RNA sequencing data from
359 The Cancer Genome Atlas, we do not know whether the upregulation of genes in the
360 *KRAS* signaling pathway and the increased incidence of *CIC* mutations and 1p/19q
361 deletions are due to this variant or a linked variant in a regulatory region that we would
362 be able to analyze with whole genome sequencing data. Therefore, we identified the four
363 other variants that are genetically linked to rs61757955 in the European population, the
364 population which is most similar to the TCGA low grade glioma patient population
365 (**Table S3**). These variants did not pass the criteria to be included within the 196,022

366 testable variants that we had identified at the beginning of this study but could become
367 useful in the future.

368 **rs34988193 is a Deleterious Germline Variant Present in *ANKDD1a* Associated with**
369 **Poor Outcomes**

370 In order to identify prognostic variants that are predicted to be deleterious due to
371 effects on the encoded protein, we repeated our analysis but restricted it to only variants
372 with a combined annotation dependent depletion (CADD) score greater than 30 and
373 expression greater than one FPKM on average. 81 variants met this criteria. These
374 variants correspond to the top 0.1% of deleterious mutations as predicted by this scoring
375 system. We found the germline variant rs34988193 in the tumor suppressor gene
376 *ANKDD1a* to be associated with poor prognosis in the TCGA dataset (p=0.001,
377 HR=1.73, FDR < 0.10, **Figure 4A-B, Table 2B**). Because this variant is found in both
378 the European and Asian populations, we were able to test whether this variant is also
379 predictive of survival in the independent Chinese Glioma Genome Atlas (CGGA) dataset.
380 We found this variant to be predictive of survival in the CGGA dataset and we found the
381 hazard ratio that we calculated in CGGA to be very similar to the hazard ratio calculated
382 in the TCGA dataset (p=0.0743, HR=1.79, **Figure 4C, Table 2B**). rs34988193 is not
383 linked with any other variant in the European population. We did not find any enriched
384 pathways after performing gene set enrichment analysis and this variant was not
385 associated with differences of any of the genomic or histological variables (**Table S4**).

386 *ANKDD1a* contains ten ankyrin repeat domains and one death-like domain. This
387 variant causes a non-synonymous mutation in the last codon of the ninth ankyrin repeat
388 domain. The AAG to GAG codon change results in the incorporation of negatively

389 charged glutamate instead of the wild type positively charged lysine residue in the loop
390 between ankyrin repeats nine and ten (**Figure 4D**). This variant has a CADD score of 32
391 and is therefore predicted to be in the top 0.06% of deleterious mutations across the
392 human genome. We performed a BLASTp search using the *ANKDD1a* protein sequence
393 to identify homologous sequences in 22 other species. We aligned these sequences using
394 ClustalW and found that this lysine residue is conserved in all 22 of these species (**Figure**
395 **4E**). The PhyloP score at this position is 8.42, suggesting that evolution is occurring
396 much more slowly than expected at this residue assuming no selection pressure. We
397 determined the position-specific profile Gribskov's score for a lysine to glutamate amino
398 acid change at this position using the multiple sequencing alignment from 23 species to
399 be 15 to 3, suggesting that this variant is highly unfavorable.

400 **Combined Model Predicts Survival Better Than Clinical Covariates Alone**

401 As a result of this analysis, we found the germline variants rs61757955 in the 3'
402 UTR of GRB2 and rs34988193 in the protein-coding region of ANKDD1a to be
403 predictive of survival in low grade glioma patients. We constructed a survival model
404 consisting of the eleven clinical covariates, rs61757955, rs34988193, GRB2 expression,
405 and ANKDD1a expression and generated a receiver operator characteristic curve by
406 using this model to categorize patients as alive or dead after seven years of follow up.
407 This combined model is significantly better at predicting survival compared to the eleven
408 clinical covariates alone (p=0.0279, **Figure 4F**).

409 **Discussion**

410 Up until this point, the identification of prognostic features in gliomas has been
411 limited to clinical factors, somatic mutations, gene expression changes, and methylation

412 pattern changes.[6–8] Although many studies have commented on how germline variants
413 could enable physicians to better individualize patient care by being able to better predict
414 how a patient might respond to chemotherapeutic treatment,[34–36] most large-scale studies
415 have focused on identifying germline variants that predispose or protect an individual to a
416 disease.[13,37] These studies have not focused on understanding how germline variants can
417 be used to individualize patient care following diagnosis. Identifying prognostic germline
418 variants is difficult due to the large number of germline variants, the limited effect of any
419 single germline variant, and clinical factors that may confound the effect of germline
420 variants. In this study, we have developed a novel method that can be used to identify
421 prognostic germline variants and we have used that method to identify two variants that
422 are predictive of survival in the TCGA dataset. The germline variant rs61757955 in
423 *GRB2* is not found in the Asian population and so could not be confirmed in an
424 independent dataset. In contrast, the germline variant rs34988193 in *ANKDD1a* is found
425 in both the European and Asian populations, and remarkably, was found to be prognostic
426 with very similar hazard ratios in both the TCGA and CGGA datasets.

427 Studies of germline variants using TCGA datasets typically solely utilize the
428 WXS normal blood samples.[13,38] One major disadvantage to this approach is that it limits
429 the analysis to genes within the capture regions of the whole exome sequencing kits used
430 by the study.[4] In this study, we combined the information from both the whole exome
431 sequencing and RNA sequencing datasets for a given patient to identify germline variants
432 outside of the whole exome sequencing capture region. Our approach had the added
433 benefit of providing us with more information for a given variant for variants with low
434 sequencing depth in the whole exome sequencing datasets. We do not believe that this

435 approach significantly affected the accuracy of our variant calls because the allele
436 frequencies calculated from the RNA sequencing dataset were well correlated with the
437 allele frequencies from gnomAD and with the allele frequencies calculated from the
438 whole exome sequencing datasets. We showed that somatic mutations and RNA editing
439 did not affect the integrity of our finding. Only one somatic mutation in a single patient
440 overlapped with the 196,022 variants that we tested in our analysis and only 215 of the
441 196,022 variants that we tested overlapped with the 2.5 million known RNA editing sites.
442 We did not find any of these variants to be predictive of survival. Instead, we feel that the
443 increased sample size resulting from the additional sequencing coverage greatly
444 outweighs any effect that somatic mutations or RNA editing had on our results.
445 We next needed to devise an approach to using these germline variants in a Cox
446 regression model. We first had to decide how to deal with the absence of a variant in the
447 variant call file. The variant could be absent because the patient was wild type for that
448 allele or because the sequencing depth at that position was too low to make the variant
449 call. We therefore determined the sequencing depth of each variant at each position so
450 that we could exclude patients with low sequencing depths for the testing of specific
451 variants. Testing a large number of variants increased the probability of a variant being
452 significant solely because it was confounded with another significant variable. To avoid
453 this issue, we tested each variant while controlling for 11 other covariates that we found
454 to be predictive of survival. In this study, we found rs61757955 to be associated with
455 differences in 1p/19q co-deletion status. By including the 1p/19q co-deletion as a
456 covariate in our model, we were able to estimate the effect of rs61757955 independent
457 from the 1p/19q co-deletion status and the other ten covariates.

458 *GRB2* is a signal transduction adaptor protein that plays an oncogenic role in a
459 variety of cancers.[39–42] *GRB2* plays an important role in the *RAS/RAF/ERK* pathway. Its
460 SH2 domain binds the phosphotyrosine of activated growth factor receptor, while its two
461 SH3 domains bind the guanine nucleotide exchange factor son of sevenless (*SOS*)
462 protein, resulting in *SOS* recruitment to the plasma membrane and subsequent *RAS*
463 activation. *RAS* binds and activates the kinase *RAF*, which phosphorylates the kinase
464 *MEK*. *MEK* phosphorylates and activates extracellular signal-regulated kinase (*ERK*)
465 which transmits the signal to transcription factors in the nucleus. This results in cell
466 proliferation.[43] We found the variant rs61757955 located in the 3' UTR of *GRB2* to be
467 associated with poor prognosis in glioma patients. Separating patients on the basis of this
468 variant revealed that the *KRAS* signaling pathway is upregulated in patients with this
469 variant. As described above, *GRB2* plays a well-characterized role in this pathway.[43] We
470 also found this variant to be associated with an increased incidence of *CIC* mutations and
471 1p/19q co-deletions. *CIC* is a known driver of low grade glioma pathogenesis.[33]
472 Mutations in *CIC* are common in oligodendrogliomas and are associated with poor
473 prognosis.[4,32] Although patients with rs61757955 variant exhibited an elevation in the
474 incidence of oligodendrogliomas which we expected given the increased incidence of
475 *CIC* mutations and 1p/19q co-deletions,[32] this difference was not statistically significant.
476 It is possible that this germline variant or the four other germline variants that it is linked
477 with increase a patient's risk for oligodendrogliomas with the *CIC* mutation and 1p/19q
478 co-deletion.

479 In this study, we were only able to study variants in the whole exome or RNA
480 sequencing data. Although it is possible that the 3' UTR of *GRB2* has regulatory activity

481 or affects *GRB2* protein translation efficiency, it is also possible that one of the variants
482 that rs61757955 is linked to regulates the *KRAS* signaling pathway. None of the four
483 linked variants are in the protein coding sequence of *GRB2* so that if they upregulate
484 RAS activity, like the rs61757955, they likely do so by regulating the expression of
485 *GRB2*. While recent large-scale sequencing studies have published patient whole genome
486 sequences,[44] this data is not yet available for gliomas. We will be able to apply our
487 approach to variants in regulatory regions in the future to specifically identify these
488 prognostic variants when whole genome sequencing data for gliomas is available. Our
489 inability to further study this variant in the CGGA dataset due to this variant being rare in
490 Asian populations is a limitation of this study which could be addressed in the future with
491 the availability of additional glioma sequencing datasets. This result also suggests that the
492 clinical usefulness of specific germline variants is dependent on the frequency of that
493 germline variant in the population.

494 *ANKDD1a* is a tumor suppressor gene that has been shown to inhibit cell
495 autophagy and induce apoptosis in glioblastoma multiforme (GBM). It directly interacts
496 with and upregulates *FIH1*, resulting in inhibition of *HIF1α* activity and decreased
497 *HIF1α* half-life. This induces apoptosis in GBM cell lines in hypoxic microenvironments.
498 Hypermethylation of this gene is common in GBM and leads to decreased *ANKDD1a*
499 expression and increased cell proliferation.[45] We found the germline variant rs34988193,
500 located at the end of the ninth of ten ankyrin repeat domains in this protein, to be
501 associated with a poor prognosis in low grade glioma patients in both the TCGA and
502 CGGA datasets. The hazard ratio independently calculated using the two datasets is
503 remarkably similar. The wild type lysine residue is conserved in all 22 species with a

504 homologue to *ANKDD1a* and this position has a high PhyloP score. This variant is
505 predicted to be within the top 0.06% of deleterious mutations in the human genome by
506 CADD score[27] because it causes a change from a positively charged lysine residue to a
507 negatively charged glutamic acid residue in the loop of this ankyrin repeat. Ankyrin
508 repeats are common domains known for their involvement with protein-protein
509 interactions.[46,47] Previous studies have suggested that mutations in the loops of ankyrin
510 repeats may disrupt protein-protein interactions.[48–50] The change from a positively to
511 negatively charged amino acid resulting from the germline variant rs34988193 in the loop
512 of *ANKDD1a* may disrupt *ANKDD1a*'s protein interaction partners and could explain the
513 poor prognosis associated with this variant seen in two independent datasets. Given the
514 amino acid change, further studies involving rs34988193 in *ANKDD1a* could be directed
515 towards experimentally determining whether or not this variant alters *ANKDD1a*'s
516 protein-protein interactions.

517 rs61757955 in *GRB2* and rs34988193 in *ANKDD1a* could also be used to enhance
518 predictions made by survival models clinically, as we found that these variants are
519 significant predictors of prognosis even after controlling for eleven covariates. The
520 prognostic effect of rs34988193 in *ANKDD1a* seems to be fairly reliable, as we found
521 that this variant had a similar hazard ratio in both the TCGA and CGGA datasets. Our
522 approach could be used in the future to identify sets of germline variants that together
523 enhance the predictions made by survival models, though the current number of low
524 grade glioma sequencing samples is small relative to the large number of possible
525 combinations of germline variants. Focused studies on particular sets of genes or
526 pathways could potentially get around this low sample size problem by drastically

limiting the number of variants studied. We believe that this study provides researchers with an effective approach to identifying biologically significant germline variants and provides clinicians with germline variants that could enhance currently existing survival models.

Acknowledgements

We thank Dr. Wei Min Chen at the University of Virginia for his statistical input on our Cox regression models. We thank Dr. Ana Damljanovic and Dr. Liz Williams for their assistance with computation on the Cancer Genomics Cloud Platform. We thank dbGAP for providing us with access to The Cancer Genome Atlas data. We thank all of the patients and their families that participated in The Cancer Genome Atlas and Chinese Glioma Genome Atlas studies. We thank the Dutta lab members for the valuable feedback during the drafting of this manuscript.

541 **Figure Legends**

542 **Figure 1**. A flowchart describing the steps involved in identifying prognostic germline
543 variants.

544 **Figure 2**. Prognostic germline variants in the TCGA dataset.

545 **(A)** Of the 4.4 million unique variants called in the TCGA dataset, we ran Cox regression
546 on the 196,022 germline variants found in gnomAD with an allele frequency greater than
547 5% and found in 15 or more of the TCGA low grade glioma patients.

548 **(B-E)** Similar to the 196,022 germline variants, the 271 prognostic variants are mostly
549 found in **(B)** protein-coding genes, **(C)** are located in introns, and are **(D)** single
550 nucleotide polymorphisms (SNP). **(E)** Most single nucleotide polymorphisms cause
551 transitions.

552 **Figure 3.** rs61757955 is a highly prognostic germline variant identified in the TCGA
553 dataset.

554 **(A)** Manhattan plot showing the p-values resulting from testing each germline variant by
555 Cox regression, controlling for the 11 variables in red in Table 1.Two variants passed the
556 FDR threshold in the TCGA dataset.

557 **(B)** A Kaplan-Meier plot depicting the deleterious outcome associated with rs61757955,
558 adjusting for the eleven covariates.

559 **(C)** Receiver operator characteristic curve at 7 years. rs61757955 increases the area under
560 the curve compared to the 11 covariates alone, suggesting that it improves the clinical
561 model.

562 **(D)** Separation of patients on the basis of whether or not they have this germline variant
563 to determine which genes are induced (red) or repressed (blue) in patients with

rs61757955. Subsequent gene set enrichment analysis reveals that patients with this germline variant exhibit upregulation of the genes involved with *KRAS* signaling.

Figure 4. rs34988193 is a prognostic variant predicted to be highly deleterious.

(A) A Manhattan plot with the p-values resulting from testing each germline variant by Cox regression, controlling for the eleven covariates in red in Table 1. rs34988193 is prognostic (FDR<0.10) in the TCGA when restricting the analysis to the top 0.1% most deleterious variants by combined annotation dependent depletion (CADD).

(B-C) Kaplan-Meier plots depicting the deleterious outcome associated with rs34988193 in the **(C)** TCGA and **(D)** CGGA datasets, adjusted for the eleven covariates.

(D) A schematic showing that this variant is located in the ninth ankyrin repeat of *ANKDD1a*. The predicted protein structure of *ANKDD1a* reveals that this variant leads to an amino acid change from lysine to glutamate on the loop of an ankyrin repeat.

(E) Multiple sequence alignment of *ANKDD1a* in 22 species showing that lysine is conserved at this position in all of the species with this protein.

(F) Receiver operator characteristic curves comparing the ability of two survival models to label patients as alive or dead after seven years of follow up. The inclusion of rs61757955 variant status, rs34988193 variant status, GRB2 expression and ANKDD1a expression significantly improves the survival prediction compared to the eleven covariates bolded in red from table 1 alone (p=0.0279).

587 **Table Legends**

588 **Table 1.** List of variables that are known to be associated with differences in survival in
589 low grade glioma patients. 11 variables (red) were selected by Lasso for inclusion in the
590 survival model. We used these 11 variables as covariates in our Cox regression model
591 when testing each germline variant.

592 **Table 2.** Description of the prognostic germline variants identified in this study.
593 **(A)** A description of the two prognostic germline variants (FDR<0.10) in the TCGA
594 dataset identified when testing all 196,022 germline variants.
595 **(B)** A description of the prognostic germline variant (FDR<0.10) rs34988193 in
596 *ANKDD1a* identified when the analysis was restricted to only germline variants with a
597 combined annotation dependent depletion (CADD) score greater than 30 in the TCGA
598 and CGGA datasets.

599 **Table 3.** The association between the germline variant rs61757955 and genomic and
600 histological variables. Patients were divided based on whether or not they had the
601 germline variant rs61757955. Patients with the germline variant rs61757955 were more
602 likely to have *CIC* mutated gliomas and the 1p/19q co-deletion.

603

604

605

606

607

608

609
610

611 **References**
612

613 1. Louis DN, Perry A, Reifenberger G, et al. The 2016 World Health Organization
614 Classification of Tumors of the Central Nervous System: a summary. *Acta*
615 *Neuropathol*. 2016. doi:10.1007/s00401-016-1545-1
616 2. Bauman G, Fisher B, Watling C, Cairncross JG, Macdonald D. Adult
617 Supratentorial Low-Grade Glioma: Long-Term Experience at a Single Institution.
618 *Int J Radiat Oncol Biol Phys*. 2009. doi:10.1016/j.ijrobp.2009.01.010
619 3. Jiao Y, Killela PJ, Reitman ZJ, et al. Frequent ATRX, CIC, FUBP1 and IDH1
620 mutations refine the classification of malignant gliomas. *Oncotarget*.
621 2012;3(7):709-722. doi:10.18632/oncotarget.588
622 4. Network CGAR, Brat DJ, Verhaak RGW, et al. Comprehensive, Integrative
623 Genomic Analysis of Diffuse Lower-Grade Gliomas. *N Engl J Med*. 2015.
624 doi:10.1056/NEJMoa1402121
625 5. Yan W, Zhang W, You G, et al. Molecular classification of gliomas based on
626 whole genome gene expression: A systematic report of 225 samples from the
627 Chinese Glioma Cooperative Group. *Neuro Oncol*. 2012.
628 doi:10.1093/neuonc/nos263
629 6. Qian Z, Li Y, Fan X, et al. Prognostic value of a microRNA signature as a novel
630 biomarker in patients with lower-grade gliomas. *J Neurooncol*. 2018.
631 doi:10.1007/s11060-017-2704-5
632 7. Hu X, Martinez-Ledesma E, Zheng S, et al. Multigene signature for predicting
633 prognosis of patients with 1p19q co-deletion diffuse glioma. *Neuro Oncol*. 2017.
634 doi:10.1093/neuonc/now285
635 8. Zhou H, Vallières M, Bai HX, et al. MRI features predict survival and molecular
636 markers in diffuse lower-grade gliomas. *Neuro Oncol*. 2017.
637 doi:10.1093/neuonc/now256
638 9. Guerrini-Rousseau L, Dufour C, Varlet P, et al. Germline SUFU mutation carriers
639 and medulloblastoma: Clinical characteristics, cancer risk, and prognosis. *Neuro*
640 *Oncol*. 2018. doi:10.1093/neuonc/nox228
641 10. Dudley B, Karloski E, Monzon FA, et al. Germline mutation prevalence in
642 individuals with pancreatic cancer and a history of previous malignancy. *Cancer*.
643 2018. doi:10.1002/cncr.31242
644 11. Zhang J, Walsh MF, Wu G, et al. Germline Mutations in Predisposition Genes in
645 Pediatric Cancer. *N Engl J Med*. 2015. doi:10.1056/NEJMoa1508054
646 12. Carter H, Marty R, Hofree M, et al. Interaction landscape of inherited
647 polymorphisms with somatic events in cancer. *Cancer Discov*. 2017.
648 doi:10.1158/2159-8290.CD-16-1045
649 13. Huang K lin, Mashl RJ, Wu Y, et al. Pathogenic Germline Variants in 10,389
650 Adult Cancers. *Cell*. 2018. doi:10.1016/j.cell.2018.03.039
651 14. Baretta Z, Mocellin S, Goldin E, Olopade OI, Huo D. Effect of BRCA germline
652 mutations on breast cancer prognosis: A systematic review and meta-analysis.
653 *Medicine (Baltimore)*. 2016. doi:10.1097/MD.0000000000004975
654 15. Lau JW, Lehnert E, Sethi A, et al. The cancer genomics cloud: Collaborative,
655 reproducible, and democratized - A new paradigm in large-scale computational
656 research. *Cancer Res*. 2017. doi:10.1158/0008-5472.CAN-17-0387

657 16. Ceccarelli M, Barthel FP, Malta TM, et al. Molecular Profiling Reveals
658 Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma.
659 *Cell*. 2016. doi:10.1016/j.cell.2015.12.028
660 17. Lai Z, Markovets A, Ahdesmaki M, et al. VarDict: A novel and versatile variant
661 caller for next-generation sequencing in cancer research. *Nucleic Acids Res*. 2016.
662 doi:10.1093/nar/gkw227
663 18. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and
664 SAMtools. *Bioinformatics*. 2009. doi:10.1093/bioinformatics/btp352
665 19. Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic
666 variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010.
667 doi:10.1093/nar/gkq603
668 20. Ellrott K, Bailey MH, Saksena G, et al. Scalable Open Science Approach for
669 Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *Cell Syst*.
670 2018. doi:10.1016/j.cels.2018.03.002
671 21. Ramaswami G, Li JB. RADAR: A rigorously annotated database of A-to-I RNA
672 editing. *Nucleic Acids Res*. 2014. doi:10.1093/nar/gkt996
673 22. Purcell S, Neale B, Todd-Brown K, et al. PLINK: A Tool Set for Whole-Genome
674 Association and Population-Based Linkage Analyses. *Am J Hum Genet*. 2007.
675 doi:10.1086/519795
676 23. Friedman AJ, Hastie T, Simon N, Tibshirani R, Hastie MT. Lasso and Elastic-Net
677 Regularized Generalized Linear Models. *Available online https//cran.r-*
678 *project.org/web/packages/glmnet/glmnet.pdf (Verified 29 July 2015)*. 2015.
679 24. Therneau TM, T. Lumley. Package ‘ survival .’ *R Top Doc*. 2015.
680 doi:10.1016/j.jhydrol.2011.07.022.
681 25. *Package “survminer” Type Package Title Drawing Survival Curves Using*
682 *“Ggplot2.”*; 2018. https://cran.r-
683 project.org/web/packages/survminer/survminer.pdf. Accessed August 27, 2018.
684 26. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: A
685 knowledge-based approach for interpreting genome-wide expression profiles. *Proc*
686 *Natl Acad Sci*. 2005. doi:10.1073/pnas.0506580102
687 27. Kircher M, Witten DM, Jain P, O’roak BJ, Cooper GM, Shendure J. A general
688 framework for estimating the relative pathogenicity of human genetic variants. *Nat*
689 *Genet*. 2014. doi:10.1038/ng.2892
690 28. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics
691 Analysis Version 7.0 for Bigger Datasets. *Mol Biol Evol*. 2016.
692 doi:10.1093/molbev/msw054
693 29. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral
694 substitution rates on mammalian phylogenies. *Genome Res*. 2010.
695 doi:10.1101/gr.097857.109
696 30. Rice P, Longden L, Bleasby A. EMBOSS: The European Molecular Biology Open
697 Software Suite. *Trends Genet*. 2000. doi:10.1016/S0168-9525(00)02024-2
698 31. The 1000 Genomes Project Consortium. A global reference for human genetic
699 variation. *Nature*. 2015. doi:10.1038/nature15393
700 32. Gleize V, Alentorn A, Connen De Kérillis L, et al. CIC inactivating mutations
701 identify aggressive subset of 1p19q codeleted gliomas. *Ann Neurol*. 2015.
702 doi:10.1002/ana.24443

703 33. Bailey MH, Tokheim C, Porta-Pardo E, et al. Comprehensive Characterization of
704 Cancer Driver Genes and Mutations. *Cell*. 2018;173(2):371-385.e18.
705 doi:10.1016/j.cell.2018.02.060
706 34. Pesenti C, Gusella M, Sirchia SM, Miozzo M. Germline oncopharmacogenetics, a
707 promising field in cancer therapy. *Cell Oncol*. 2015. doi:10.1007/s13402-014-
708 0214-4
709 35. Pinto N, Cohn SL, Dolan ME. Using germline genomics to individualize pediatric
710 cancer treatments. *Clin Cancer Res*. 2012. doi:10.1158/1078-0432.CCR-11-1938
711 36. Schärfe CPI, Tremmel R, Schwab M, Kohlbacher O, Marks DS. Genetic variation
712 in human drug-related genes. *Genome Med*. 2017. doi:10.1186/s13073-017-0502-5
713 37. Lu C, Xie M, Wendl MC, et al. Patterns and functional implications of rare
714 germline variants across 12 cancer types. *Nat Commun*. 2015.
715 doi:10.1038/ncomms10086
716 38. Koire A, Katsonis P, Lichtarge O. REPURPOSING GERMLINE EXOMES OF
717 THE CANCER GENOME ATLAS DEMANDS A CAUTIOUS APPROACH
718 AND SAMPLE-SPECIFIC VARIANT FILTERING. *Pac Symp Biocomput*. 2016.
719 39. Gay B, Suarez S, Caravatti G, Furet P, Meyer T, Schoepfer J. Selective Grb2 SH2
720 inhibitors as anti-Ras therapy. *Int J Cancer*. 1999. doi:10.1002/(SICI)1097-
721 0215(19991008)83:2<235::AID-IJC15>3.0.CO;2-B
722 40. Giubellino A, Burke TR, Bottaro DP. Grb2 signaling in cell motility and cancer.
723 *Expert Opin Ther Targets*. 2008. doi:10.1517/14728222.12.8.1021
724 41. Haines E, Saucier C, Claing A. The adaptor proteins p66Shc and Grb2 regulate the
725 activation of the GTPases ARF1 and ARF6 in invasive breast cancer cells. *J Biol*
726 *Chem*. 2014. doi:10.1074/jbc.M113.516047
727 42. Yu GZ, Chen Y, Wang JJ. Overexpression of Grb2/HER2 signaling in Chinese
728 gastric cancer: Their relationship with clinicopathological parameters and
729 prognostic significance. *J Cancer Res Clin Oncol*. 2009. doi:10.1007/s00432-009-
730 0574-8
731 43. Lowenstein EJ, Daly RJ, Batzer AG, et al. The SH2 and SH3 domain-containing
732 protein GRB2 links receptor tyrosine kinases to ras signaling. *Cell*. 1992.
733 doi:10.1016/0092-8674(92)90167-B
734 44. Bolouri H, Farrar JE, Triche T, et al. The molecular landscape of pediatric acute
735 myeloid leukemia reveals recurrent structural alterations and age-specific
736 mutational interactions. *Nat Med*. 2018. doi:10.1038/nm.4439
737 45. Feng J, Zhang Y, She X, et al. Hypermethylated gene ANKDD1A is a candidate
738 tumor suppressor that interacts with FIH1 and decreases HIF1α stability to inhibit
739 cell autophagy in the glioblastoma multiforme hypoxia microenvironment.
740 *Oncogene*. August 2018:1. doi:10.1038/s41388-018-0423-9
741 46. Sedgwick SG, Smerdon SJ. The ankyrin repeat: A diversity of interactions on a
742 common structural framework. *Trends Biochem Sci*. 1999. doi:10.1016/S0968-
743 0004(99)01426-7
744 47. Shi DJ, Ye S, Cao X, Zhang R, Wang KW. Crystal structure of the N-terminal
745 ankyrin repeat domain of TRPV3 reveals unique conformation of finger 3 loop
746 critical for channel function. *Protein Cell*. 2013. doi:10.1007/s13238-013-3091-0
747 48. Sullivan JM, Zimanyi CM, Aisenberg W, et al. Novel mutations highlight the key
748 role of the ankyrin repeat domain in TRPV4-mediated neuropathy. *Neurol Genet*.

749 2015;1(4):e29. doi:10.1212/NXG.0000000000000029
750 49. Lamandé SR, Yuan Y, Gresshoff IL, et al. Mutations in TRPV4 cause an inherited
751 arthropathy of hands and feet. *Nat Genet*. 2011. doi:10.1038/ng.945
752 50. Li J, Mahajan A, Tsai MD. Ankyrin repeat: A unique motif mediating protein-
753 protein interactions. *Biochemistry*. 2006. doi:10.1021/bi062188q
754

Figure 1

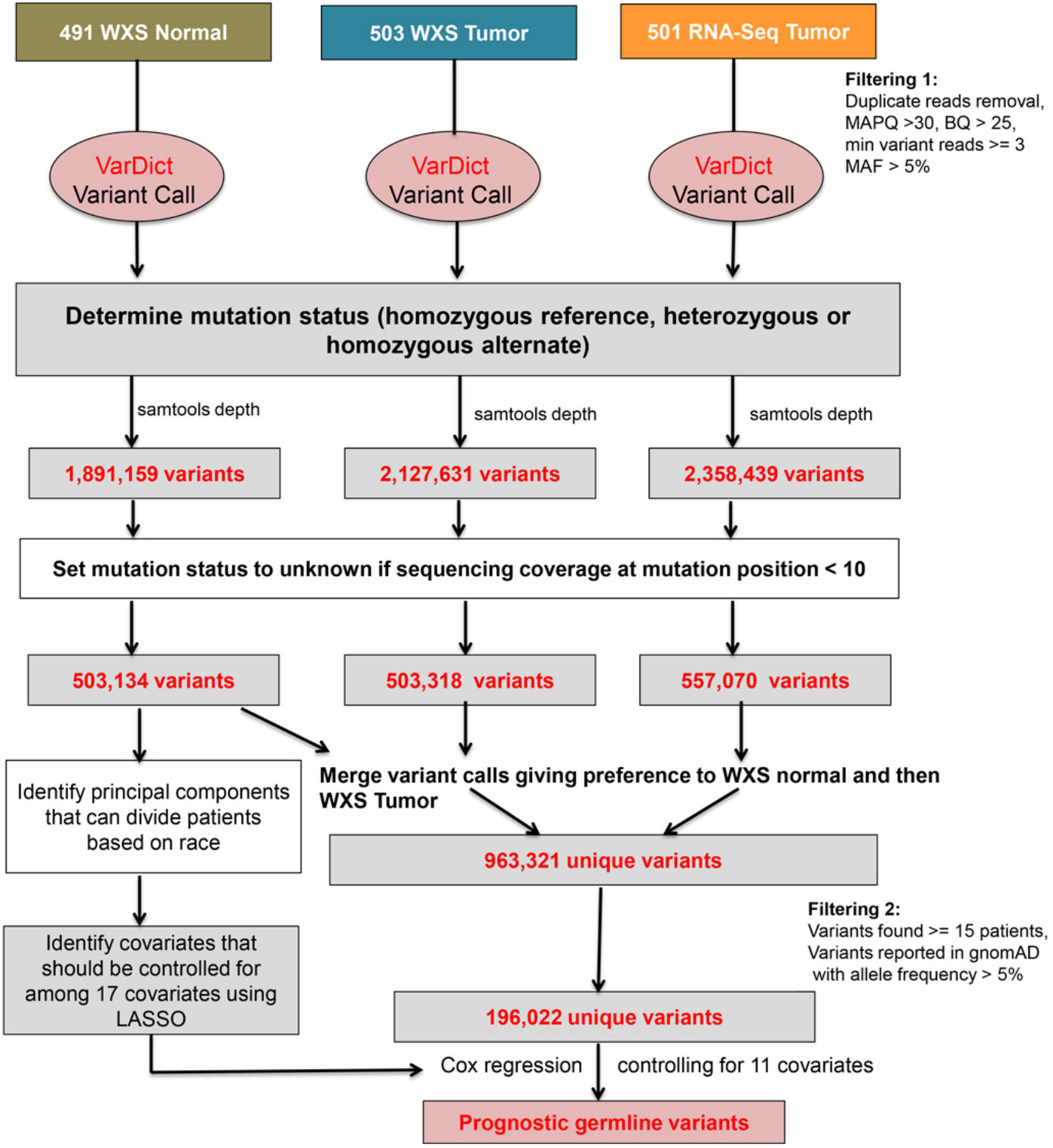

491 WXS Normal
503 WXS Tumor
501 RNA-Seq Tumor
Filtering 1:
Duplicate reads removal,
MAPQ >30, BQ > 25,
min variant reads >= 3
MAF > 5%
VarDict Variant Call
VarDict Variant Call
VarDict Variant Call
Determine mutation status (homozygous reference, heterozygous or homozygous alternate)
samtools depth
samtools depth
samtools depth
1,891,159 variants
2,127,631 variants
2,358,439 variants
Set mutation status to unknown if sequencing coverage at mutation position < 10
503,134 variants
503,318 variants
557,070 variants
Identify principal components that can divide patients based on race
Merge variant calls giving preference to WXS normal and then WXS Tumor
963,321 unique variants
Filtering 2:
Variants found >= 15 patients,
Variants reported in gnomAD with allele frequency > 5%
Identify covariates that should be controlled for among 17 covariates using LASSO
196,022 unique variants
Cox regression
controlling for 11 covariates
Prognostic germline variants

Figure 2

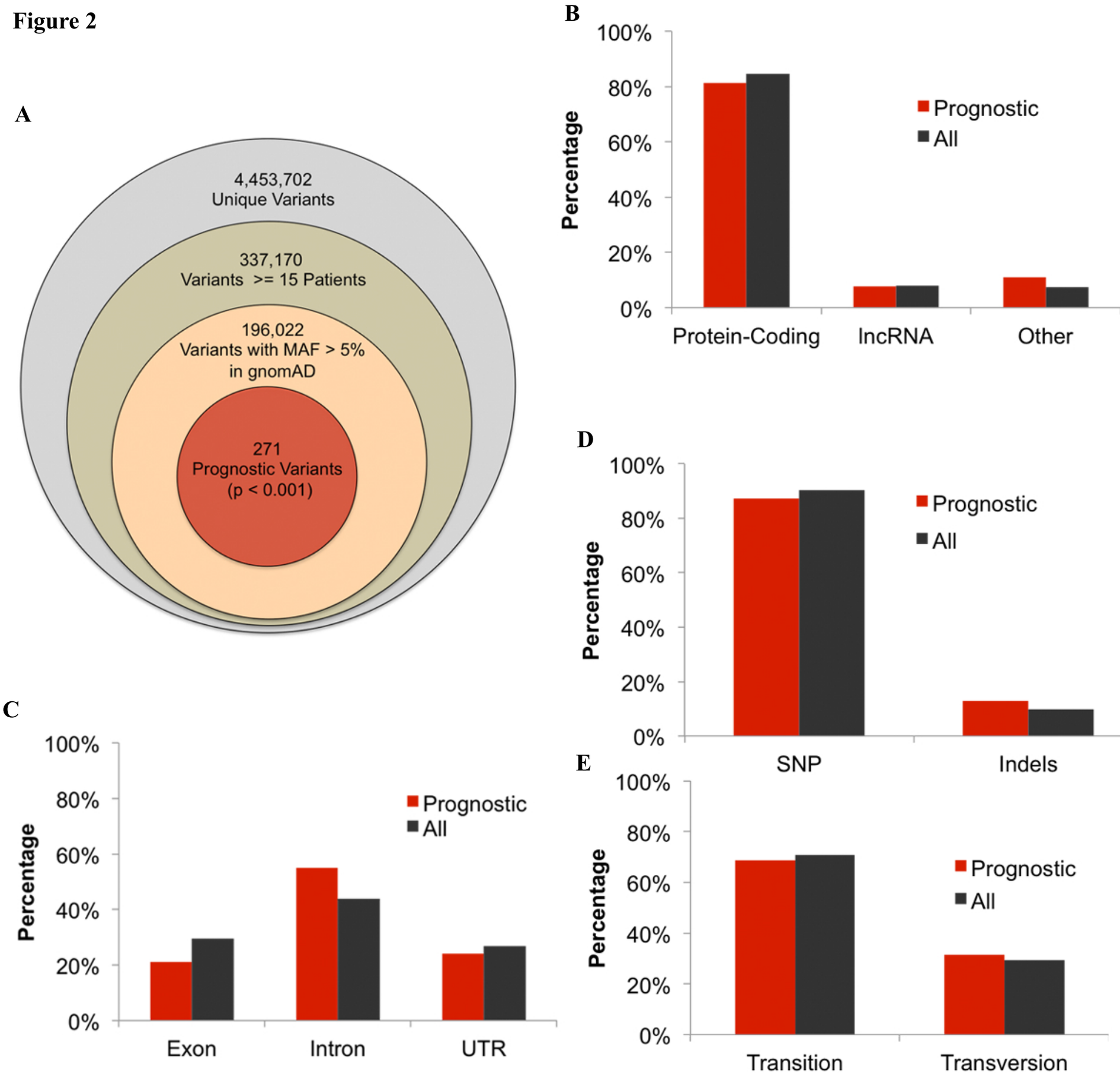
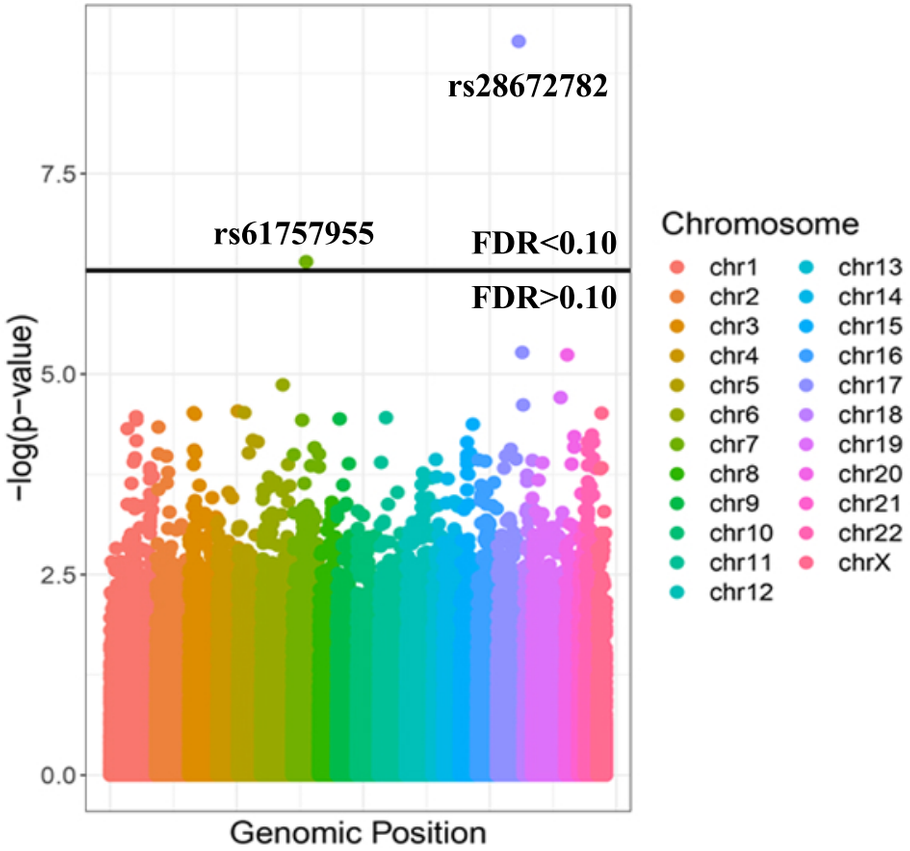
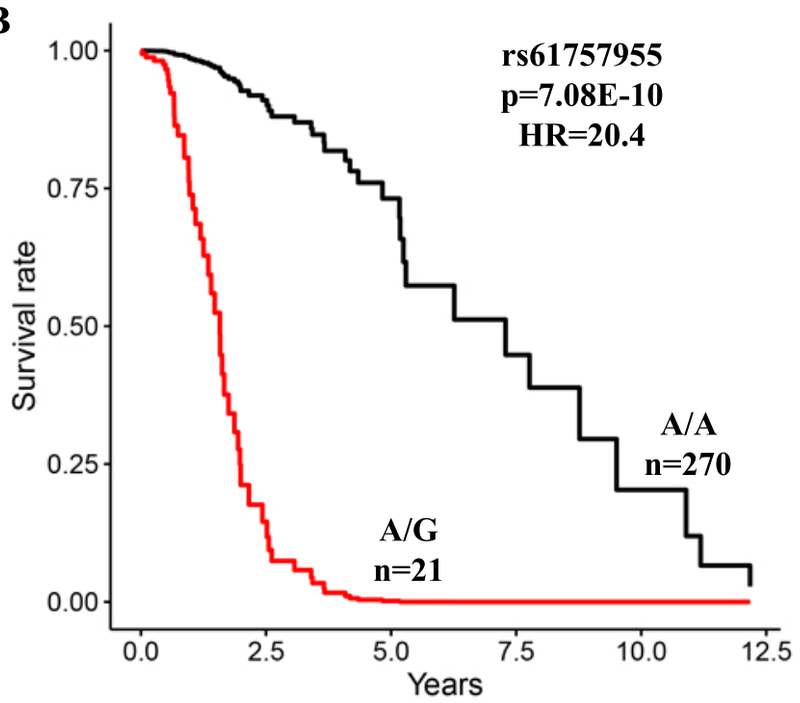

A
4,453,702
Unique Variants
337,170
Variants >= 15 Patients
196,022
Variants with MAF > 5%
in gnomAD
271
Prognostic Variants
(p < 0.001)
B
Percentage
100%
80%
60%
40%
20%
0%
Prognostic
All
Protein-Coding
lncRNA
Other
C
Percentage
100%
80%
60%
40%
20%
0%
Prognostic
All
Exon
Intron
UTR
D
Percentage
100%
80%
60%
40%
20%
0%
Prognostic
All
SNP
Indels
E
Percentage
100%
80%
60%
40%
20%
0%
Prognostic
All
Transition
Transversion

Figure 3

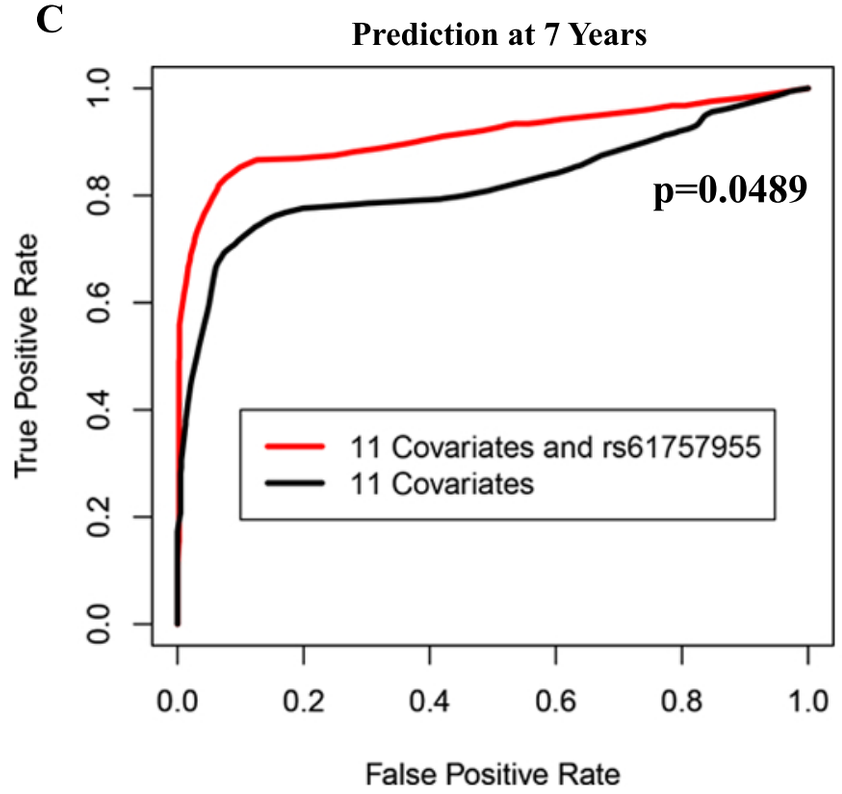
A



B



C



D

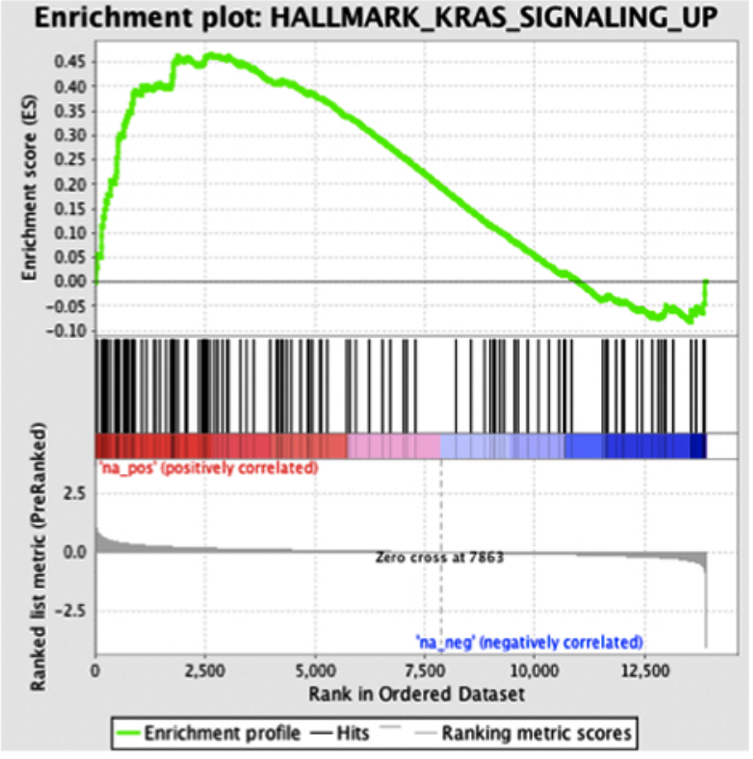


Figure 4

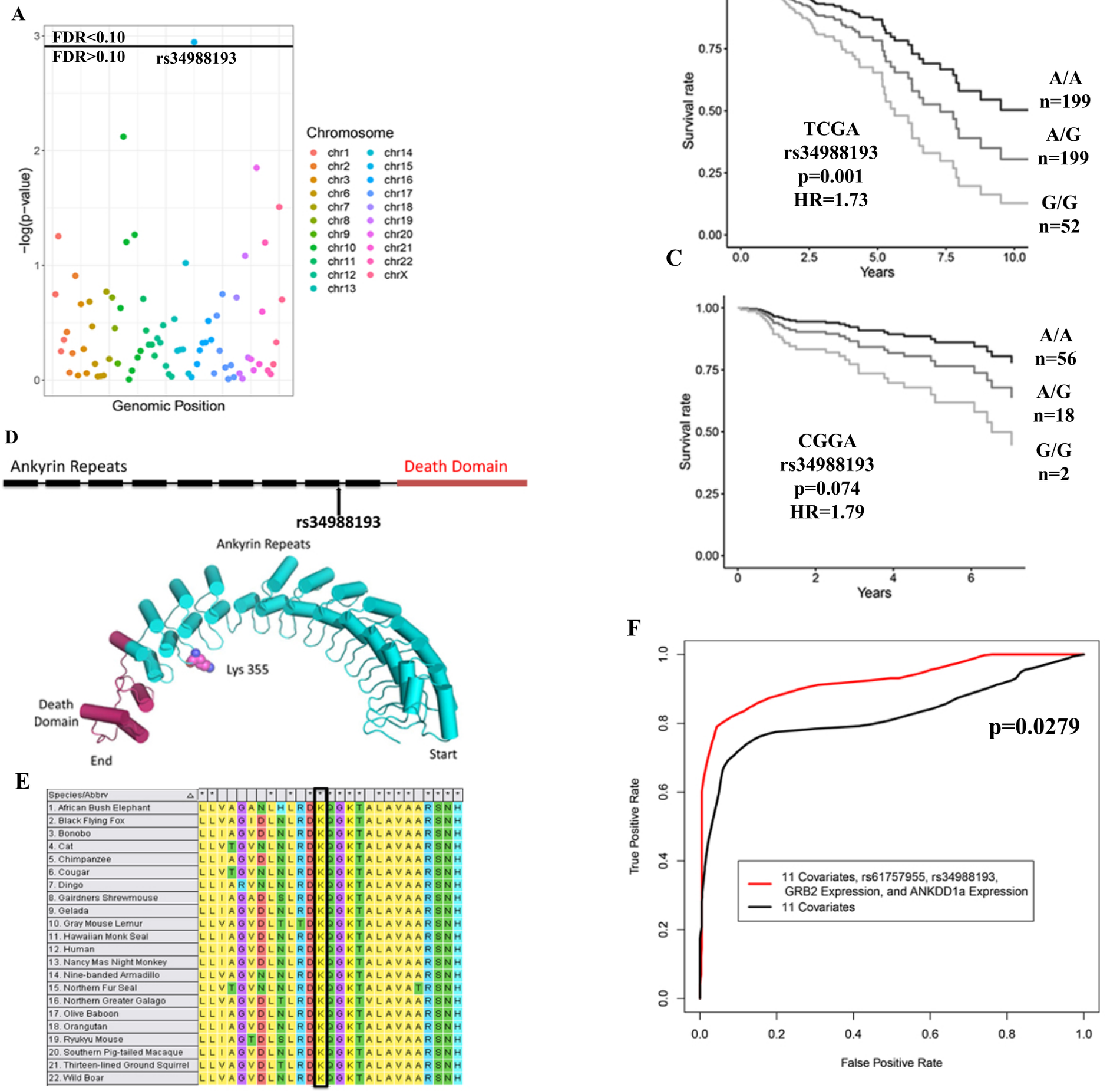

A
FDR<0.10
FDR>0.10
rs34988193
−log(p−value)
Genomic Position
Chromosome
chr1
chr2
chr3
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chrX
B
Survival rate
Years
TCGA
rs34988193
p=0.001
HR=1.73
A/A
n=199
A/G
n=199
G/G
n=52
C
Survival rate
Years
CGGA
rs34988193
p=0.074
HR=1.79
A/A
n=56
A/G
n=18
G/G
n=2
D
Ankyrin Repeats
Death Domain
rs34988193
Ankyrin Repeats
Lys 355
Death Domain
End
Start
E
Species/Abbrv
1. African Bush Elephant
2. Black Flying Fox
3. Bonobo
4. Cat
5. Chimpanzee
6. Cougar
7. Dingo
8. Gairdners Shrewmouse
9. Gelada
10. Gray Mouse Lemur
11. Hawaiian Monk Seal
12. Human
13. Nancy Mas Night Monkey
14. Nine-banded Armadillo
15. Northern Fur Seal
16. Northern Greater Galago
17. Olive Baboon
18. Orangutan
19. Ryukyu Mouse
20. Southern Pig-tailed Macaque
21. Thirteen-lined Ground Squirrel
22. Wild Boar
F
True Positive Rate
False Positive Rate
p=0.0279
11 Covariates, rs61757955, rs34988193, GRB2 Expression, and ANKDD1a Expression
11 Covariates

Table 1

Covariate		Median (Min-Max) or Number of Patients
Age		41 (14 - 87)
Gender	Female	250
	Male	200
Somatic Mutation Count		50 (0 - 12255)
Percent Aneuploidy		11% (5.2E-4% - 95%)
log(TERT Expression)		1.0 (0.0 - 9.1) FPKM
Principle Component 1 (PC1)		0.043 (-0.091 - 0.064)
Principle Component 2 (PC2)		-0.017 (-0.23 - 0.17)
Principle Component 3 (PC3)		-0.53 (1.34E-4 - 0.33)
Histological Type	**Astroctyoma**	172
	Oligoastrocytoma	113
	Oligodendroglioma	165
Tumor Location	Frontal Lobe	265
	Temporal Lobe	125
	Parietal Lobe	42
	Other	18
Grade	G2	212
	G3	237
	Cannot Be Assessed	1
Treatment Site	**Henry Fords Hospital**	82
	Case Western St. Joes	90
	Other	278
IDH Mutant	Wild Type	83
	Mutant	367
1p/19q Co-deletion	Absent	303
	Present	147
MGMT Promoter Methylation	Unmethylated	80
	Methylated	370
Chr 7 gain/Chr 10 loss	Absent	399
	Present	51
TP53 Mutant	Absent	232
	Present	218

Table 2

A

Variant	Chrom	Pos	Ref	Alt	Population Frequency	Sample Size	Number of Heterozygotes	Number of Homozygotes	Gene Name	Median Expression (FPKM)	p-value	Hazard Ratio
rs61757955	17	75318086	A	G	5.01%	291	21	0	GRB2	42.2	7.08E-10	20.4
rs28672782	11	1446622	C	T	16.27%	50	15	5	BRSK2	7.29	< 1E-16	1.15E-10

B

Variant	Chrom	Pos	Ref	Alt	Population Frequency	CADD Score	PhyloP Score	Gene Name	Dataset	Sample Size	Number of Heterozygotes	Number of Homozygotes	p-value	Hazard Ratio
rs34988193	15	64943580	A	G	30.90%	32	8.42	ANKDD1a	TCGA	450	199	52	0.00113	1.73
									CGGA	76	18	2	0.0743	1.79

Table 3

Variable	Mean or Percentage (Wild Type)	Mean or Percentage (Mutant)	p-value
CIC Mutated	**15.9%**	**38.1%**	**0.017**
1p/19q Co-deletion	**25.2%**	**47.6%**	**0.038**
Oligodendroglioma	33.7%	42.9%	0.475
Total Somatic Mutation Count	30.9	30.0	0.766
Percent Aneuploidy	15.1%	11.7%	0.524
Astrocytoma	38.1%	42.9%	0.651
Grade 3	53.0%	42.9%	0.497
IDH Mutated	78.1%	85.7%	0.583
1p/19q Co-deletion	25.2%	47.6%	0.038
MGMT Promoter Methylation	77.8%	81.0%	1.000
Chr 7 Gain/Chr 10 Loss	13.0%	9.5%	1.000
Expression of GRB2 (FPKM)	45.7	44.4	0.636