PREDICTING SIGNIFICANCE OF UNKNOWN VARIANTS IN GLIAL TUMORS THROUGH SUB-CLASS ENRICHMENT

ALEX M. FICHTENHOLTZ†
afichtenholtz@foundationmedicine.com

NICHOLAS D. CAMARDA†
ndc9@duke.edu

ERIC K. NEUMANN†
eneumann@foundationmedicine.com

†*Technology Innovation, Foundation Medicine Inc.*
Cambridge, MA 02141, USA

Glial tumors have been heavily studied and sequenced, leading to scores of findings about altered genes. This explosion in knowledge has not been matched with clinical success, but efforts to understand the synergies between drivers of glial tumors may alleviate the situation. We present a novel molecular classification system that captures the combinatorial nature of relationships between alterations in these diseases. We use this classification to mine for enrichment of variants of unknown significance, and demonstrate a method for segregating unknown variants with functional importance from passengers and SNPs.

1. Introduction

Molecular diagnostics are increasing in importance to clinical oncology as the number of therapies targeting specific molecular alterations and pathways in cancer grows. These new drugs are accompanied by a shift in the tumor classification paradigm away from one based on histopathology to one centered on the molecular drivers of cancer. This has resulted in a proliferation of studies investigating the roles of various tumor suppressors and oncogenes in many types of tumors. The methods by which samples are interrogated have also shifted away from single gene hotspot tests to massively parallel, multiple marker integrated platforms[1]. In addition to genetic alterations in driver genes, gene expression changes, promoter mutations and methylation status have been implicated in cancer progression. This explosion in our ability to measure does not always lead to an increase in understanding, as we struggle to understand the relationships between the many markers we can now observe.

The genomic landscape of brain cancer, in particular tumors of glial origin, is a particularly difficult area for interpretation, for while large-scale sequencing studies of glioblastoma and lower grade astrocytomas have identified multiple targetable oncogenic driver alterations[2], these results have yet to meaningfully impact treatment decisions. Targeted therapies have had limited success in these tumor types, and multiple clinical trials have failed to show benefit with targeted tyrosine kinase inhibitors[3,4]. Existing molecular classification schemes are either based on gene expression[5], performed exclusively in lower grade gliomas like oligodendrogliomas[6], or focused on the 'main three' markers (*IDH* mutation, *TERT* promoter mutation and chromosome 1p/19q loss)[7]. We present a genomic classification for glial tumors based on comprehensive massively

parallel sequencing of over 800 glial cancers of different grades, annotation of the resultant variant calls, and subsequent latent class analysis of the detected genetic alteration landscape.

In our classification, we take care to annotate alterations as either: 'known or likely' drivers of cancer, or 'variants of unknown significance' (VUS), as described in Methods and Materials. The motivation for this is to segregate genomic events that play a role in the tumor mechanism from innocuous alterations (i.e. SNPs and passenger mutations). While there are notable exceptions, in general it is somatic alterations that drive tumors[8]. We filter out suspected germ line variant calls using dbSNP[9], but this database is based on 1000 human genomes, meaning that the rarer SNPs will not be accounted for. 'Passenger' mutations can also confound the alteration landscape. Briefly speaking, passenger mutations are alterations accumulated during clonal expansion that do not currently confer any selective advantage onto the tumor[10]. The clinical significance of labeling these genomic events is paramount: if the oncologist elects to target a passenger mutation, the therapeutic regimen will presumably have no effect.

Any alteration that cannot be explicitly labeled as driver, passenger, or SNP is considered to be a VUS. If we look at the number of variant calls that are considered 'known and likely' versus the number we consider to be VUSs, we find that the VUSs account for the majority of what we detect. Thus we have a scenario in which there are modes of cancer unaccounted for, but this signal is mixed heavily with the noise of germ line variants and passenger mutations, making this excellent territory for variant prioritization approaches. Existing methods to evaluate the significance of unknown alterations in the context of disease range widely in their strategies. Sequence conservation based algorithms make the argument that mutations at heavily conserved residues in oncogenes and tumor suppressors are more likely to be deleterious[11]. Structural biology based methods stratify alterations based on their impact on protein folding energy and solubility[12].

We propose that a parallel method to assign significance to uncharacterized mutations is to align them to existing knowledge. Our classification of glioma samples, which is based on a small number of heavily mutated known cancer drivers, is statistically robust and well supported by existing literature. We use this classification as a reference point representing the current state of knowledge regarding the molecular landscape of glioma and examine how VUSs, which were not included in the definition of the molecular classes, distribute themselves along class partitions. We argue that genes that show skewed distributions towards a specific class participate in the mechanism driving tumors of that class in an as yet previously undescribed manner.

2. Materials and Methods

2.1 Comprehensive genomic profiling

All samples were submitted to a CLIA-certified, New York State and CAP-accredited laboratory (Foundation Medicine, Cambridge MA) for NGS-based genomic profiling, as previously described[13]. Extracted DNA was adaptor-ligated and capture was performed for all coding exons of 287 cancer-related and 47 introns of 19 genes frequently rearranged in cancer (Sup Table S1).

Captured libraries were sequenced to a median exon coverage depth of >500x, and resultant sequences were analyzed for base substitutions, insertions, deletions, copy number alterations (focal amplifications and homozygous deletions) and select gene fusions. Natural germline variants from the 1000 Genomes Project (dbSNP135)[8] were removed, and known confirmed somatic alterations deposited in the Catalog of Somatic Mutations in Cancer (COSMIC v62)[14] were highlighted as biologically significant (i.e. 'known and likely variants'). All inactivating events (i.e. truncations and deletions) in known tumor suppressor genes were also called as significant.

2.2 Data selection and filtering

Analysis was performed on a combined dataset of 847 glial tumor samples from four separate diseases: 76 oligodendrogliomas (BOD), 99 low and mid-grade astrocytomas (LGA), 101 anaplastic astrocytomas (AA), and 571 brain glioblastomas (GBM). For each gene that is altered in the data set, we count how many samples within that set carry an alteration in this gene, not taking into account alteration type (e.g. gene amplification, point mutation, etc.). If a sample has multiple alterations in a given gene (i.e. an indel and an amplification), it is still only counted once. The list of gene counts is then sorted, and all genes not altered in at least six samples are discarded in order to keep statistical power high. Because certain sets of genes tend to occur together in co-amplified vectors (e.g. *CDKN2A* and *CDKN2B*), we added a 'co-amplification' feature: if two genes occur in a data set, and their genetic co-ordinates are within 10 Mb of one another on the same chromosome, a separate variable that indicates their co-occurrence is added to the feature list. For instance, if co-mutations in *CDKN2A* and *CDKN2B* occur with sufficient frequency in a data set, there should be three features: the presence of a *CDKN2A* alteration only, the presence of a *CDKN2B* alteration only, and the presence of an alteration in both genes. This concept can be extended to three and four co-amplified genes. We estimate structure models while changing the 'class number' parameter r for each r=1,…,R, and then progressively increase r until the number of parameters to be estimated in the model exceed the number of samples (i.e. system is underdetermined). The full list of features used for clustering is given in Sup Table S2.

2.3 Latent class analysis

Latent class analysis was performed with the R package **poLCA**, version 1.4[15]. For each tumor sample *i* in our set, there are *J* manifest (observable) variables (genes), each of which can have one of two K_j outcomes (altered|not altered). The latent class model approximates the observed distribution of the manifest variables with a weighted sum of *R* cross-classification tables. The probability that a sample in class $r = 1,\ldots,R$ produces the *k*th outcome on the *j*th variable is represented by π_{jrk}, and the weight for a given class *r* is denoted by p_r. Thus, for each manifest variable within a class, $\sum_{k=1}^{K_j} \pi_{jrk} = 1$, and across all classes $\sum_r p_r = 1$. Denoting the observed value of the j*th* manifest variable for sample *i* having the *k*th outcome as Y_{ijk} (such that if gene *j* in sample *i* is mutated $Y_{ijk} = 1$, otherwise $Y_{ijk} = 0$), the probability that sample *i* in class *r* has any given set of mutations *J* is given by:

$$P(Y_i;\pi_r) = \prod_{j=1}^{J}\prod_{k=1}^{K_j}(\pi_{jrk})^{Y_{ijk}} \quad (1)$$

Across all classes '*r*', this probability is given by:

$$P(Y_i|\pi,p) = \sum_{r=1}^{R} p_r \prod_{j=1}^{J}\prod_{k=1}^{K_j}(\pi_{jrk})^{Y_{ijk}} \quad (2)$$

The parameters estimated by **poLCA** are p_r and π_{jrk}. Denoting the total number of samples in the set as *N*, the latent class model is found by maximizing the log-likelihood function (3) with respect to p_r and π_{jrk} using expectation-maximization:

$$\ln L = \sum_{i=1}^{N} \ln \sum_{r=1}^{R} p_r \prod_{j=1}^{J}\prod_{k=1}^{K_j}(\pi_{jrk})^{Y_{ijk}} \quad (3)$$

The above equation reaches its maximum values with class partitions that best satisfy the conditional independence criterion (i.e. that manifest variables show conditional independence within a class). The number of classes in a model is a parameter adjusted by the user, and a larger value results in a higher log-likelihood score for the model. **poLCA** finds local maxima starting from initial values of p_r and π_{jrk}, thus to ensure the global maximum is found for each model we estimate the model 10,000 times with random initial parameter values each time, ultimately keeping the estimated model with the best log-likelihood score. To select the most probable of the many candidate latent structure models, we use a metric called the Akaike Information Criterion[16], defined as:

$$AIC = 2k - 2\ln L \quad (4)$$

where *k* is the number of estimated parameters specified by the model (a product of the number of features and the number of classes in the model), and *L* is the maximized value of the log-likelihood function we described earlier. Under certain conditions, an alternate information metric called the Bayesian Information Criterion[17] can be used. This is defined as:

$$BIC = -2\ln L + k\ln n \quad (5)$$

where *n* is the total number of sample instances being analyzed. Discussed in greater detail elsewhere[18], the BIC can be the appropriate metric for evaluating multiple models when the model space is dominated by a few major effects and contains likely nested models, which we believe to be the case for our combined dataset. Given a set of candidate models, the most appropriate model is the one that minimizes the AIC or BIC.

2.4 Feature selection approach

We performed an initial LCA on the dataset using a large number of features. For each feature, we calculated the initial entropy of the feature using Shannon's entropy formula[19], where *p* is the probability of seeing that feature in any given sample across the data set:

$$H(X) = -p \log_2 p \tag{6}$$

This allows us to compare the loss of entropy in each feature across classes as we increase the complexity of the models we fit the data to. We use entropy loss as a measure of how well the LCA partitions the alteration occurrences. If the probability of seeing a feature within a class is p_x and the probability of a given sample being a member of that class is given as p_y, then this entropy can be calculated across multiple classes using the conditional version of Shannon's formula[20]:

$$H(X|Y) = p_{x|y} \log_2 \frac{p_y}{p_{x|y}} \tag{7}$$

We calculate the entropy loss for each feature across all models to determine if that feature accounts for a significant portion of the entropy drop for the entire system. We then use that metric as a basis to set a lower threshold for the number of features to be included in the modeling and then performed a 'higher-resolution' LCA using this reduced feature set.

2.5 Association of VUSs with known classes

After computing the most likely classification for each sample using the high-resolution LCA based on a small number of known and likely features, we performed statistical testing of ***all*** of the alterations detected in the samples assigned to each class. We grouped alterations by gene, and separated those annotated as 'known and likely' from ones annotated as 'VUS.' For each feature in each class, we calculated the enrichment within the given class against all other classes of 'known and likely' alterations, 'VUS' alterations, and total alterations using a one-sided Fisher's Exact test. P-values obtained using Fisher's Exact are adjusted for multiple hypothesis testing using the Bonferroni formula[21]:

$$\alpha_{\text{adj}} = 1 - (1 - \alpha)^n \tag{8}$$

where α is the significance level of the result, and n is the total number of independent tests conducted, which in this case is 83 features × 5 tests each = 415 total tests. In the case where a gene's association with two classes simultaneously is tested, the exponent used for Bonferroni correction is 1245.

3. Results

Investigational LCA of the combined glial tumors set was performed using 83 features with up to eight classes being modeled. Analysis of entropic loss per feature during modeling reveals that the majority of information loss is accounted for by the 17 most frequently occurring features (84.1%, 84.1%, 70.8%, 70.2%, 70.2%, 69.2%, 69.1% for 2-class, 3-class, 4-class, 5-class, 6-class, 7-class, 8-class models, respectively; Figure 1). LCA was performed with these 17 features with the 6-class solution deemed most likely given BIC metrics (Table 1). Class 1 is dominated by alterations in *CDKN2A/B* (190/190 samples; p=1) and *EGFR* (161/190; p=0.84) (Figure 2). Class 2 also features alterations in *CDKN2A/B* (116/162; p=0.72), but instead of *EGFR* shows alterations in *NF1* (92/162; p=0.56). Class 3 showcases alterations in *EGFR* (64/114; p=0.56)

and *CDK4*/*MDM2* (49/114; p=0.43). Class 4 is selective for alterations in *IDH1* (137/151; p=0.91), *TP53* (145/151; p=0.96) and *ATRX* (140/151; p=0.93). Class 5 is characterized by *IDH1* mutation (89/89; p=1) and alterations in *CIC* (31/89; p=0.35) and *TP53* (38/89; p=0.43). Class 6 is represented by alterations in *TP53* (111/141; p=0.79), *RB1* (55/141; p=0.39) and *PTEN* (68/141; p=0.48). The full model is given in Table 2.

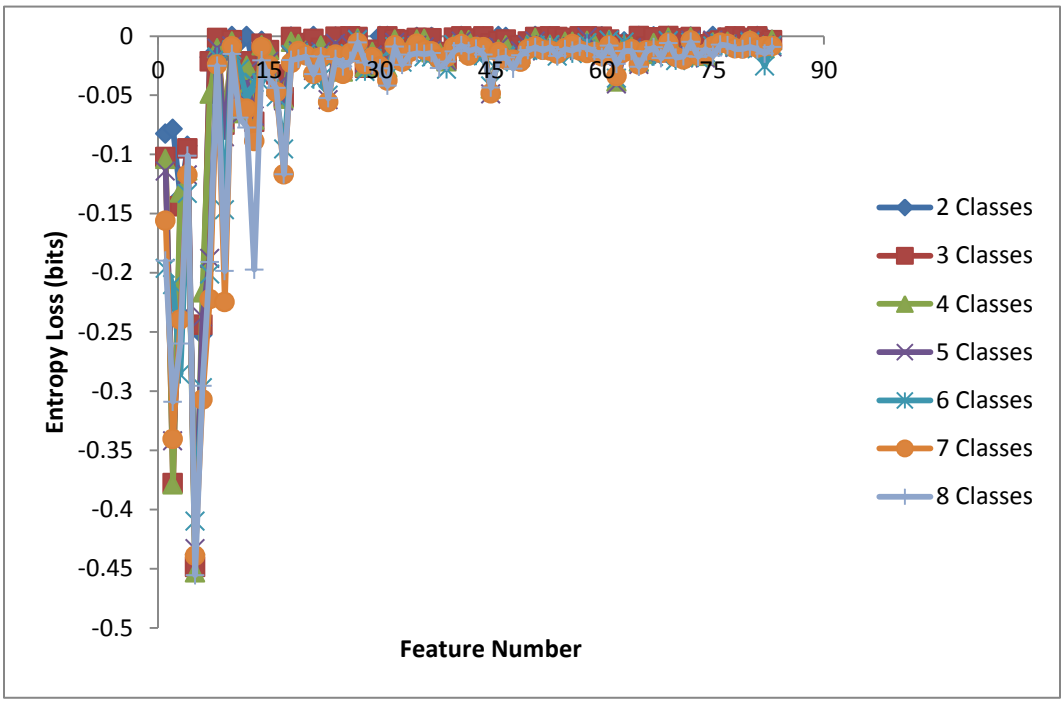

Figure 1. Information theoretic analysis of the clustering behavior of the 83 features most frequently altered in the glial tumor cohort. The bulk (~70%) of entropy loss achieved by class segregation is concentrated in the top 17 features.

Enrichment analysis of all alterations including VUSs for samples within a given class reveals notable contributions from *PTEN* (p=1.2e-8) in Class 1, *PTPN11* (p=0.018) in Class 2, and *NOTCH1* (p=0.018), *NOTCH4* (p=0.043), *ARID1A* (p=5.6e-4), and *SMARCA4* (p=0.09) in Class 4 (Table 3). The full list of enrichment results is listed in Supplemental Table S3.

Classes	LL	AIC	BIC	Parameters	Relative P (AIC)	Relative P (BIC)
1	-5370.53	10775.07	10855.68	17	0	1.20e-261
2	-4829.37	9728.731	9894.69	35	3.30e-146	5.74e-53
3	-4682.99	9471.985	9723.295	53	1.86e-90	9.48e-16
4	-4602.15	9346.296	9682.957	71	3.65e-63	5.44e-07
5	-4528.9	9235.796	9657.807	89	3.61e-39	0.16
6	-4466.37	9146.745	9654.107	107	7.84e-20	1
7	-4422.78	9095.558	9688.27	125	1.02e-08	3.82e-08
8	-4386.38	9058.758	9736.821	143	1	1.10e-18

Table 1. High resolution LCA solutions for the combined glial tumor data set, with class number parameter varying between 1 and 8. LL = log likelihood, AIC = Akaike Information Criterion, BIC = Bayesian Information Criterion. Relative probabilities for the models depend on AIC and BIC, and are calculated using the Akaike weight formula: $p_i = e^{-(\mathrm{AIC}_i - \mathrm{AIC}_{min})/2}$.

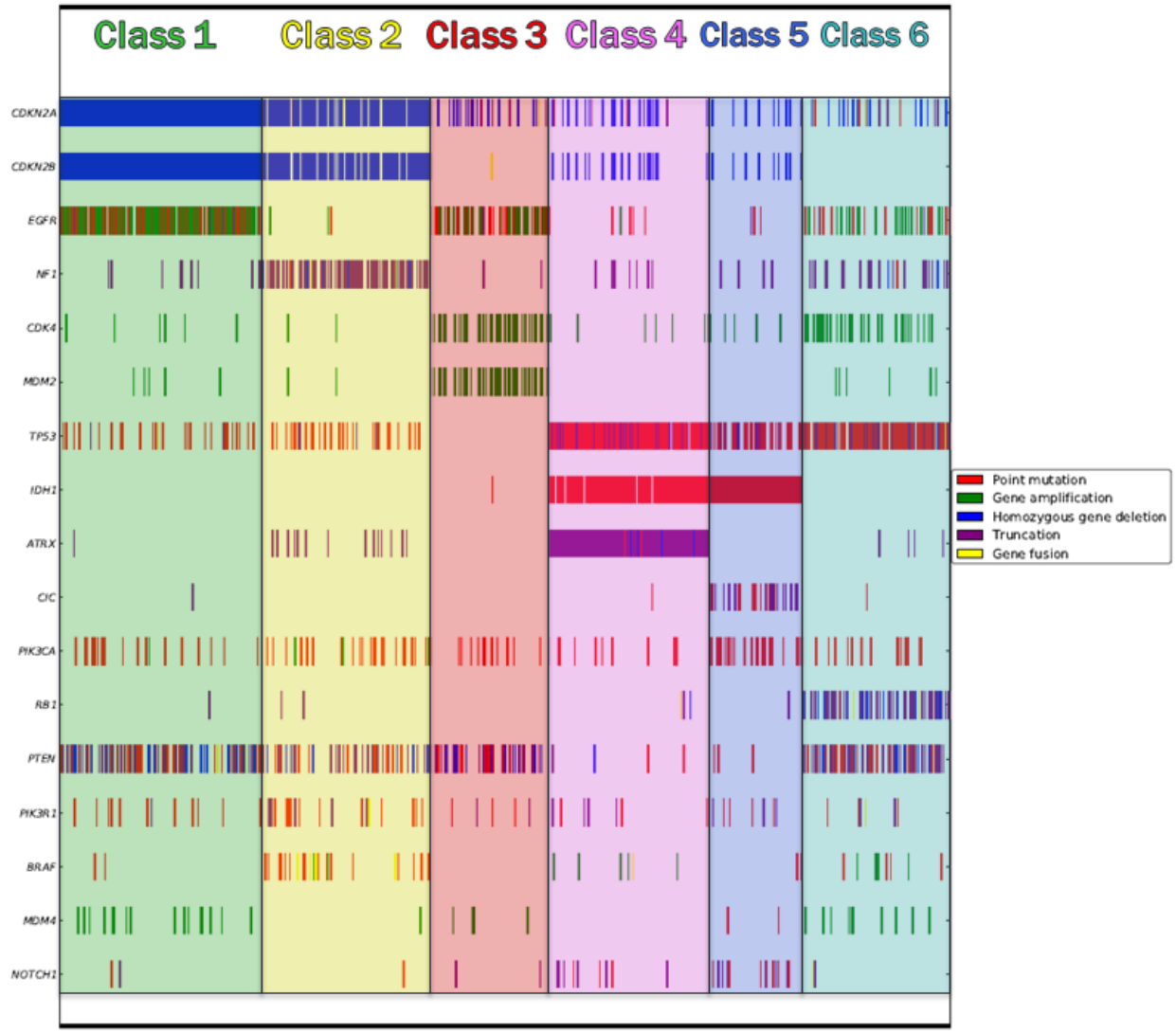

Figure 2. Latent model of the combined glial tumor cohort featuring six classes. The glial tumor cohort includes clinically confirmed cases of oligodendroglioma, low- and mixed-grade astrocytoma, anaplastic astrocytoma, and glioblastoma.

Feature	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6
TP53	0.178	0.197	0	0.963	0.433	0.79
CDKN2A/B	1	0.716	0	0.142	0.106	0
EGFR	0.848	0.033	0.56	0.019	0.038	0.263
PTEN	0.44	0.29	0.37	0.028	0.036	0.483
IDH1	0	0	0.014	0.908	1	0
ATRX	0.005	0.079	0	0.929	0	0.033
NF1	0.047	0.559	0.036	0.049	0.062	0.167
PIK3CA	0.096	0.137	0.109	0.063	0.306	0.114
RB1	0.007	0.012	0	0.019	0.013	0.388
PIK3R1	0.069	0.101	0.041	0.044	0.087	0.045
CDKN2A	0	0.033	0.208	0.028	0	0.163
CDK4	0.025	0	0	0.051	0.049	0.228
CDK4/MDM2	0.007	0.012	0.425	0	0	0

BRAF	0.015	0.139	0	0.019	0.024	0.033
MDM4	0.077	0.012	0.03	0	0.026	0.088
NOTCH1	0.009	0.008	0.017	0.07	0.162	0.015
CIC	0.005	0	0	0.005	0.345	0.008

Table 2. Best LCA model of the brain glioblastoma data set according to BIC metrics presented in the form of a class-conditional probability table. Each class is a column, and the probability of observing a sample of a given class is shown at the top of the column. Each row is a feature used in the model, and each cell value is the probability of observing an alteration in that feature in a sample assigned to that class.

Gene	Class	Known/Likely	VUS	p-Value	Adj. p-Value
PTEN	1	88	15	2.9e-11	1.2e-8
PTPN11	2	13	4	4.3e-5	1.8e-2
ARID2	4/5	6	14	3.1e-5	3.8e-2
SMARCA4	4/5	11	18	1.3e-5	1.5e-2
ARID1A	5	9	12	1.4e-6	5.6e-4
NOTCH1	5	13	12	4.5e-5	1.8e-2
NOTCH4	5	1	12	1.0e-5	4.3e-2

Table 3. Enrichment of selected genes in pre-defined molecular sub-classes taking VUSs into account.

Gene	Class-Specific Variants of Unknown Significance
PTEN	R15del, D24ins, I28T, P30L, Y46C, R47S, N48K, N48S, H61Q, Y65N, V85I, V175M, Y188D, D326Y, D326H
PTPN11	D61A, F285S, T411M, A461T
ARID1A	A139T, A165_166insA, G278D, P289A, G352R, G712A, P870L, P989L, V1082G, K1124N, P1209L, G1222E, G1274R, P1275S, V1834M
ARID2	G16E, S129C, V138A, R231K, S365L, V450I, A858T, V1040I, A1107V, E1249del, P1395S, V1503L, D1703E, N1796K
SMARCA4	G10E, P16S, P24S, G53E, G206S, P324S, L754F, P913S, Q1104R, Q1104R, V1016L, R1157G, G1159R, D1177N, A1186T, R1192H, R1203L, E1287K, T1358I, E1512K
NOTCH1	C344G, F357S, P422S, P447S, C461Y, C467Y, G519S, A958V, R1114H, P1287C, C1467Y, R1664K, G1704E, S2030F, A2035_L2048del, G2169E, V2249M
NOTCH4	E33K, G47R, R113K, G216S, S312F, G337D, P631L, G642N, A647G, R807H, V1457M, G1510S, G1701E

Table 4. Class-specific VUSs in genes found to associate significantly with pre-defined classes.

4. Discussion

Oligodendroglioma (OD), low- and mixed-grade astrocytoma (LGA), anaplastic astrocytoma (AA), and brain glioblastoma (GBM) tumors are all thought to originate from glial precursor cells, and are difficult to segregate using histopathology[22]. Multiple studies have grouped various subsets of these four diseases together for the purposes of molecular profiling[23]. We sequenced 847 of these tumors on a comprehensive massively parallel sequencing platform capable of detecting alterations in several hundred cancer related genes, examined the genetic alteration landscape from an information theoretic perspective using unsupervised classification, and found the genes that contribute most to classification. After stratifying the dataset into the likely molecular classes based on known and likely somatic alterations, we examined the distribution of

variants of unknown significance (VUS) within each class. The highest enrichments of VUSs within a given class as measured by p-value are considered in more detail.

4.1 Analysis of alteration landscape

Recognizing the similarities in histology between each disease type, we felt it was prudent to combine the tumor sets into one large superset consisting of 847 total samples. The initial analysis was done using 83 features. Analysis of the information content in every model showed that the entropic loss is concentrated in the 17 most frequently occurring features, which we consider sufficient to classify a large subset of tumors belonging to any of these four disease types. These features are alterations in *TP53*, *CDKN2A*/*B*, *EGFR*, *PTEN*, *IDH1*, *ATRX*, *NF1*, *PIK3CA*, *RB1*, *PIK3R1*, *CDKN2A*, *CDK4*, *CDK4*/*MDM2*, *BRAF*, *MDM4*, *NOTCH1*, and *CIC*. LCA considering just these 17 features yields a well-delineated six-way mixture model. Class 1 is driven by disruption to the cell-cycle mechanism with every sample in this class showing a co-deletion of *CDKN2A* and *CDKN2B*, as well as a hyper-activated signal transduction network with *EGFR* alterations being found in most samples in this class. Though not unique to this class, *PTEN* alterations are found in nearly half of all tumors in class 1. Class 2 seems to be related to Class 1, with both featuring *CDKN2A*/*B* co-deletion, though with *NF1* alterations in place of *EGFR*. Interestingly, alterations in these genes between these two classes show a nearly completely mutually exclusive relationship, suggesting that the alterations have a similar functional effect. Previous subtyping studies detected both of these classes, but were unable to segregate them, and did not allude to any mutual exclusivity between *EGFR* and *NF1*[4-6]. These two classes are significantly associated with poor prognosis, high tumor grade, and positive response to aggressive therapy.

Classes 4 and 5 are also related, and have also been discovered and confirmed to be clinically significant by prior studies. Class 4 features alterations in *IDH1*, *ATRX* and *TP53*, and is associated with positive prognosis and lower tumor grade. This class is typically found in astrocytoma, though we have seen this profile in samples assigned to both lower (e.g. oligodendroglioma) and higher grade (e.g. glioblastoma) histology categories. Class 5 shows *IDH1* mutation without *ATRX* alteration. Previous studies suggest that this class should also be associated with heterozygous deletion of chromosome arms 1p and 19q; manual inspection of sequencing data confirmed this to be the case. Class 5 should be associated with the best prognosis according to previous work, and is found primarily in oligodendrogliomas.

Class 3, which features alteration of *EGFR* in conjunction with co-amplification of *CDK4* and *MDM2* in a variation of the cell-cycle/signal transduction perturbation theme found in classes 1 and 2, and Class 6, which is driven by the combination of *TP53* and *RB1* alteration represent novel classes that have not been detected in previous cohorts, though a prior investigation of anaplastic oligodendrogliomas found simultaneous disruption of the *RB1* and *TP53* pathways in 9/20 tumors[24].

4.2 Association of VUSs with known classes

Learning the classification of glial tumors by known and likely somatic variants leads us to comparing the distribution of VUSs in genes across those classifications. Class 1, which is driven by the combination of *CDKN2A*/*B* deletion and *EGFR* alteration, shows preference for mutations in *PTEN*, and in particular, VUSs. *PTEN* alterations are known to de-regulate the PI3K pathway, and have been found to synergize with activation of *EGFR* to promote increased tumor growth[25]. The unknown variants we detected in *PTEN* within class 1 tend to cluster towards the front end of the phosphatase tensin-type domain found between residues 14 and 185 in this protein. Inspection of the COSMIC databases shows multiple confirmed somatic events around this same area, with several direct overlaps between a somatic event from COSMIC and a VUS at the residue of interest.

Class 2 features a combination between *CDKN2A*/*B* co-deletion and *NF1* alteration, and is enriched for *PTPN11* alterations, including four previously undetected VUSs. *PTPN11* encodes a cytoplasmic protein tyrosine phosphatase that promotes the activation of the Ras/MAPK pathway, and mutations in this protein lead to it being constitutively active[26]. The association of alterations in these genes with alterations in *NF1*, which encodes a negative regulator of the GTPase HRAS, suggests that tumors of this type rely on perturbation of the Ras and PI3K pathways for their tumorigenicity, and that targeting proteins in these pathways may be a viable treatment paradigm. Comparing the unknown variants we found to the collection of known somatic *PTPN11* variants from COSMIC reveals that there are confirmed somatic variants at D61 and at A461.

While classes 1 and 2 are variants of the cell-cycle/signal transduction co-perturbation mode, class 5 is driven primarily by *IDH1* mutation and some mutations in *CIC*. Previous studies claim that this class should be associated with mutation in *NOTCH1* and *FUBP1*, along with heterozygous deletion of chromosome arms 1p and 19q. These have all been confirmed in this dataset. This class features enrichment of both *NOTCH1* and *NOTCH4* at a statistically significant level. These proteins are transmembrane receptors known for their role in patterning during embryogenesis. Alterations in *NOTCH* genes have been found in a large number of cancers, and are thought to contribute to oncogenesis by promoting angiogenesis and modulating the EMT[27]. Less frequently discussed is the formation of complexes featuring *NOTCH* proteins and histone de-methylases. Histone modification is known to be essential for transcription of *NOTCH*, and may be related to the fact that alterations in the SWI/SNF proteins *ARID1A* and *SMARCA4* are also heavily enriched for in this class. Furthermore, *ARID2*, another known chromatin remodeler associates with classes 4 and 5. The role of the chromatin signaling network in this glioma subtype has not been previously described.

4.3 Extending glial tumor genomics knowledge

An obvious question regarding this approach is whether the ‘known’ LCA classes we use to explore VUS distributions are clinically relevant, or merely statistically significant. Several previous studies have discovered these classes independently, and verified their clinical

significance. A study of low-grade gliomas in the TCGA data set uncovered both a *TP53*/*IDH1*/*ATRX* class (Class 4) as well as a class driven primarily by *IDH1* alterations with no concomitant *ATRX* events, with the occasional alteration in *CIC* (Class 5)[5]. These classes were shown to be clinically significant in terms of event-free and overall survival, age at diagnosis, primary tumor site, and molecular phenotype (i.e. methylation, gene expression, protein expression). Another study of GBM patients in the TCGA data found classes driven by *CDKN2A*/*B* co-deletion and *NF1*/*EGFR* alterations (Classes 1 and 2), though it did not recognize the mutual exclusivity between *NF1* and *EGFR*[4]. The researchers found this class to be significantly associated with poorer prognosis and an older age at diagnosis. The combination of *RB1* and *TP53* alterations that typify Class 2 has been found in pre-clinical studies to generate sarcomas in mesenchymal stem cells, and is generally known to be a transforming combination in cell lines[28]. Class 3, which features *EGFR* alterations in combination with *CDK4*/*MDM2* co-amplifications, is a novel result that has not been described before in glial tumor literature. Our results have added novel associations of tumor suppressor and oncogene alterations with these classes.

Additionally, we can learn from the distribution of variants across the length of the gene. For genes significantly associated with a given class, the pileup of alterations has a conspicuous pattern. A superb example is *SMARCA4*. This SWI/SNF protein is enriched in classes 4 and 5, and the specific VUSs associated with these classes are distributed non-randomly. Specifically, 10 of 20 unknown variants are found in a 200 aa region between AA1100 and AA1300. PROSITE[29] suggests that this area is home to a helicase domain. Further investigation should be conducted as to the role of this domain in this protein in *IDH1* driven brain tumors, as disrupting this mechanism represents a potential avenue to treating these cancers.

5. Acknowledgements

We wish to acknowledge Alex Poliakov, Bryan Lewis, Marylin Matz, and the rest of the Paradigm4 team for their technical support and enhancement of the SciDB platform used during this project.

References

[1]. Heuckmann JM, et al. *Annals of Oncology* **26**: 1830-7 (2015).
[2]. Brennan CW, et al. *Cell* **155**: 462-77 (2013).
[3]. Razis E, et al. *Clinical Cancer Research* **15**: 6258-66 (2009).
[4]. Galanis E, et al. *Clinical Cancer Research* **19**: 4816-23 (2013).
[5]. Verhaak RG, et al. *Cancer Cell* **17**: 98-110 (2010).
[6]. Cancer Genome Atlas Research Network, et al. *New England Journal of Medicine* **372**: 2481-98 (2015).
[7]. Eckel-Passow JE, et al. *New England Journal of Medicine* **372**: 2499-508 (2015).
[8]. Armitage P, et al. *British Journal of Cancer* **8**: 1-12 (1954).

[9]. Sherry ST, et al. *Nucleic Acids Research* **29**: 308-11 (2001).
[10]. Greaves M, et al. *Nature* **481**: 306-13 (2012).
[11]. Ng PC, et al. *Nucleic Acids Research* **31**: 3812-4 (2003).
[12]. Adzhubei I, et al. *Nature Methods* **7**: 248-9 (2010).
[13]. Frampton GM, et al. *Nature Biotechnology* **31**: 1023-31 (2013).
[14]. Forbes SA, et al. *Nucleic Acids Research* **43**: D805-11 (2014).
[15]. Linzer DA, et al. *Journal of Statistical Software* **42**: 1-29 (2011).
[16]. Akaike, H. *IEEE Transactions on Automatic Control* **19**: 716-23 (1974).
[17]. Schwarz GE. *Annals of Statistics* **6**: 461-4 (1978).
[18]. Burnham KP, et al. *Sociological Methods & Research* **33**: 261-304 (2004).
[19]. Shannon, CE. *Bell System Technical Journal* **27**: 379-423 (1948).
[20]. Cover, TM, et al. *Elements of Information Theory (1st Ed.)* New York: Wiley (1991).
[21]. Bland JM, et al. *British Medical Journal* **310**: 170 (1995).
[22]. Maher EA, et al. *Genes and Development* **15**: 1311-33 (2001).
[23]. Vigneswaran, et al. *Annals of Translational Medicine* **3**: 95-108 (2015).
[24].Watanabe T, et al. *Journal of Neuropathology and Experimental Neurology* **60**: 1181-9. (2001).
[25]. Pires MM, et al. *Cancer Biology and Therapy* **14**: 246-53 (2013).
[26]. Bentires-Alj M, et al. *Cancer Research* **64**: 8816-20 (2004).
[27]. Allenspach EJ, et al. *Cancer Biology and Therapy* **1**: 466-76 (2002).
[28]. Rubio R, et al. *Oncogene* **32**: 4970-80 (2013).
[29]. Sigrist CJA, et al. *Nucleic Acids Research* **41**: D344-7 (2012).

Supplementary material is hosted at http://files.fm/u/ozghtas/