**TITLE:**

A machine learning classifier trained on cancer transcriptomes detects NF1 inactivation signal in glioblastoma

**AUTHORS:**

*Gregory P. Way[a,b], *Robert J. Allaway[c], Stephanie J. Bouley[c], Camilo E. Fadul[d], Yolanda Sanchez[c,e,1], Casey S. Greene[b,2]

*these authors contributed equally to this work

**AFFILIATIONS:**

[a]Genomics and Computational Biology Graduate Program, University of Pennsylvania, Philadelphia, PA, USA
[b]Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania, Philadelphia, PA, USA
[c]Department of Molecular and Systems Biology, Geisel School of Medicine at Dartmouth, Dartmouth College, Hanover, NH, USA
[d]Department of Neurology, University of Virginia, Charlottesville, VA, USA
[e]Norris Cotton Cancer Center, Dartmouth-Hitchcock Medical Center, Lebanon, NH, USA

**CO-CORRESPONDING AUTHORS:**

[1]Geisel School of Medicine at Dartmouth, HB 7650, Hanover, NH 03755; Phone: 603-650-1669; Fax: 603-650-1669; yolanda.sanchez@dartmouth.edu
[2]10-131 SCTR 34th and Civic Center Blvd, Philadelphia, PA 19104; Phone: 215-573-2991; Fax: 215-573-9135; csgreene@upenn.edu

**AUTHOR EMAILS:**

GPW: gregway@mail.med.upenn.edu
RJA: Robert.J.Allaway.GR@dartmouth.edu
SJB: Stephanie.J.Bouley.GR@dartmouth.edu
CF: CEF3W@hscmail.mcc.virginia.edu
YS: Yolanda.Sanchez@dartmouth.edu
CSG: csgreene@mail.med.upenn.edu

**KEYWORDS:**

Neurofibromatosis Type I, Glioblastoma, Machine Learning, Cancer, NF1 Inactivation, Classifier

**ABSTRACT (309/350):**

*Background*

We have identified molecules that exhibit synthetic lethality in cells with loss of the neurofibromin 1 (*NF1*) tumor suppressor gene. However, recognizing tumors that have inactivation of the *NF1* tumor suppressor function is challenging because the loss may occur via mechanisms that do not involve mutation of the genomic locus. Degradation of the NF1 protein, independent of *NF1* mutation status, phenocopies inactivating mutations to drive tumors in human glioma cell lines. NF1 inactivation may alter the transcriptional landscape of a tumor and allow a machine learning classifier to detect which tumors will benefit from synthetic lethal molecules.

*Results*

We developed a strategy to predict tumors with low NF1 activity and hence tumors that may respond to treatments that target cells lacking NF1. Using RNAseq data from The Cancer Genome Atlas (TCGA), we trained an ensemble of 500 logistic regression classifiers that integrates mutation status with whole transcriptomes to predict NF1 inactivation in glioblastoma (GBM). On TCGA data, the classifier detected *NF1* mutated tumors (test set area under the receiver operating characteristic curve (AUROC) mean = 0.77, 95% quantile = 0.53 – 0.95) over 50 random initializations. On RNA-Seq data transformed into the space of gene expression microarrays, this method produced a classifier with similar performance (test set AUROC mean = 0.77, 95% quantile = 0.53 – 0.96). We applied our ensemble classifier trained on the transformed TCGA data to a microarray validation set of 12 samples with matched RNA and NF1 protein-level measurements. The classifier's NF1 score was associated with NF1 protein concentration in these samples.

*Conclusions*

72 We demonstrate that TCGA can be used to train accurate predictors of NF1 inactivation in GBM. The
73 ensemble classifier performed well for samples with very high or very low NF1 protein concentrations
74 but had mixed performance in samples with intermediate NF1 concentrations. Nevertheless, high-
75 performing and validated predictors have the potential to be paired with targeted therapies and
76 personalized medicine.
77
78 **BACKGROUND:**
79 Genomic tools allow investigators to devise therapies targeting specific molecular abnormalities in
80 tumors. One such alteration is the loss of neurofibromin 1 (NF1), an important tumor suppressor that
81 regulates the activity of *RAS* GTPases [1,2]. Heterozygous mutation or deletion of *NF1* causes
82 neurofibromatosis type 1 (NF), one of the most frequently inherited genetic disorders [3]. NF patients
83 often develop plexiform neurofibromas (PNs), benign nerve tumors for which the only therapy is
84 surgery. However, resection is often impossible due to the tumor's intimate association with peripheral
85 and cranial nerves [4]. PNs can transform to malignant peripheral nerve sheath tumors (MPNSTs), which
86 are chemo- and radiation-resistant sarcomas with a dismal 20% 5-year survival [5]. In addition, patients
87 with NF are susceptible to a broad spectrum of other tumors including low-grade/pilocytic
88 astrocytomas, pheochromocytomas, optic nerve gliomas, and juvenile myelomonocytic leukemias [6].
89 Many aggressive non-NF associated (sporadic) tumors have recently been shown to harbor *NF1*
90 mutations, including glioblastoma (GBM), neuroblastoma, melanoma, thyroid, ovarian, breast, and lung
91 cancers [7]. Therefore, somatic and inherited loss of *NF1* function is emerging as a driver of tumors from
92 different organ sites.
93 Several groups including our own have been working to develop therapeutic approaches to target
94 tumors with loss of NF1. Previously, our lab developed a high throughput approach using yeast and
95 mammalian screening platforms to identify tool compounds and drug targets for cancer cells in which

96 NF1 loss drives tumor formation. Our pipeline identified small molecules that selectively kill or stop the
97 growth of MPNST cells carrying a mutation in *NF1* or yeast lacking the *NF1* homolog *IRA2 [8]*. We also
98 developed an assay in yeast to identify the targets of our lead tool compounds and found that one of
99 these compounds (UC-1) shares a mechanism (phosphorylation of RNA Pol II CTD Ser2/5) with
100 experimental drugs in clinical trials [8]. UC-1 impacts CTD phosphorylation, which is regulated by the
101 CTD kinase Ctk1, the yeast homolog of human Cdk9. We showed that deletion of *CTK1* was synthetic
102 lethal with loss of the yeast *NF1* homolog *IRA2*. Furthermore, we have found that inhibitors of this
103 process (dinaciclib, SNS-032) can inhibit other types of RAS-dysregulated tumor cells [9].
104 However, relying on genetic data alone to identify tumors that may be susceptible to therapies
105 targeting NF1 loss may leave a proportion of potentially actionable tumors unrecognized. NF1 tumor
106 suppressor activity can be lost via mutation of the genomic locus, proteasome-mediated degradation,
107 inhibition by miRNA, *de novo* insertion of an ALU element, and C→U editing of the *NF1* mRNA [10–14].
108 This complexity presents challenges when trying to identify tumors that will benefit from molecules that
109 exert synthetic lethality with dysregulation of *NF1/RAS* pathways.
110 The Cancer Genome Atlas (TCGA) has released a large volume of data on several cancer tissues
111 measured on a variety of genomic platforms. In the present study, we leverage TCGA GBM RNAseq
112 expression data with matched mutation calls to construct a classifier capable of identifying an NF1
113 inactivation signature. This strategy sidesteps the problem of functional characterization of mutations
114 by evaluating a regulator's downstream gene expression activity. We applied this signature to predict
115 NF1 inactivation in a cohort of biobanked GBMs. In general, this approach can be translatable to any
116 gene producing measurable downstream transcriptome-wide effects.
117
118 **METHODS:**
119 *The Cancer Genome Atlas Data used for building the classifier*

120 We downloaded RNAseq and mutation data from TCGA Pan Cancer project from the UCSC Xena
121 data portal [15] and subset each dataset to only the GBMs [16]. The data consists of 607 GBMs; of which
122 291 have mutation calls, 172 have RNAseq measurements, and 149 have both RNAseq and mutation
123 calls. Of these 149 samples, 15 have inactivating *NF1* mutations (10.1%) and were used as gold standard
124 positives in building the classifier (Supplementary Table S1). Additionally, to reduce dimensionality while
125 avoiding unexpressed and invariant genes, we subset to the top 8,000 most variably expressed genes by
126 median absolute deviation. We z-scored all gene expression measurements. This resulted in the final
127 input matrix with dimension 149 samples by 8,000 genes. For use in platform independent predictions,
128 we used Training Distribution Matching (TDM) to transform the TCGA RNAseq data to match a
129 microarray expression distribution [17].
130 Since we are also aware of the *NF1* mutation status for each of the samples, we form a supervised
131 learning task – predicting when a sample has loss of *NF1* activity. Our "X" matrix is formed by the
132 RNAseq measurements for all 149 samples measured by 8,000 genes, which are the features in the
133 model. Our "y" vector consist of {0, 1} elements where a 1 corresponds to a sample with an inactivating
134 *NF1* mutation and a 0 is an *NF1* wildtype sample. The machine learning task is to find the feature
135 weights, or gene coefficients, that best minimize our objective function. Along with these feature
136 weights corresponding to the genes' importance in the learning task, we also output a probability
137 estimate for each sample that they have loss of *NF1* activity.
138
139 *Hyperparameter optimization of the logistic regression classifier*
140 Using the GBM RNAseq data, we trained logistic regression classifiers with an elastic net penalty
141 using stochastic gradient descent to detect tumors with NF1 inactivation. We chose a penalized
142 regression model because it is simple to train and has easily interpretable outputs including importance
143 scores for each gene (feature weights) associated with the downstream consequences of *NF1* loss of

144 function and a probability for each sample that *NF1* is lost. An elastic net logistic regression model has
145 also been successfully implemented in similar studies [18–20].
146 We identified high-performing alpha and L1 mixing parameters using 5-fold cross validation ensuring
147 balanced membership of *NF1* mutations in each fold. Briefly, alpha controls how weight penalty and the
148 L1 mixing parameter tunes the amount of test set regularization by controlling the sparsity of the
149 features. An L1 mixing parameter value of zero corresponds to the L2 penalty and a value of one
150 corresponds to the L1 penalty, with L1 bringing a sparser solution. We used python 3.5.1 and Sci-kit
151 Learn for machine learning implementations [21].
152
153 *Ensemble classifier construction and application to the validation set*
154 After selecting optimal hyperparameters, we constructed 500 classifiers that would compose our
155 ensemble model. Specifically, across 100 different random initializations, we subset the full TCGA GBM
156 data into 5 folds and trained a single classifier for each training fold.
157 We borrowed terminology from the epidemiology field to describe data partitioning. We trained our
158 models on a "training" partition and assessed model performance on a "test" partition, which refers to
159 the held out cross-validation fold. The independent "validation set" refers to the GBM dataset
160 generated in a different lab (see Figure 1A).
161 Because of the small number of gold standard positive training examples, we were concerned about
162 the stability of our model solutions. Therefore, we constructed an ensemble classifier from the 500
163 models. Specifically, we assigned each classifier a weight using the specific randomization's "test set"
164 cross-validation AUROC. Lastly, for the final *NF1* inactivation prediction, we used the mean of the
165 weighted predictions across all iterations as the *NF1* inactivation prediction. We applied this ensemble
166 classifier to the validation set in which NF1 protein levels were directly measured.
167

168 *Effect sizes and power analysis*

169 We calculated the decision function of each ensemble classifier applied to all samples in the
170 training and testing 5-fold cross validation folds to calculate Cohen's D effect size between predicted
171 NF1 wildtype and NF1 inactive samples [22]. The Cohen's D metric quantifies the difference between
172 NF1 wildtype and NF1 inactive samples according to the mean classifier score and directly demonstrates
173 how different the ensemble model predicts the two groups to be.

174 Moreover, we were also concerned that our relatively small validation set would not provide us
175 with enough power to observe a detectable effect in the ensemble model's final prediction. We
176 performed a one-tailed Welch's two-sample *t*-test comparing the NF1 protein concentration of our
177 validation samples that were predicted to be either NF1 wildtype or NF1 deficient. Using the given
178 sample size, Cohen's D effect size, and a significance threshold of $\alpha = 0.05$, we calculated the power of
179 the prediction scores on the validation set. The power analysis was two-sample, one-tailed and
180 incorporated unequal sample sizes in each group.

181

182 *Validation Sample Acquisition*

183 Thirteen flash-frozen, de-identified GBM samples were obtained from the Maine Medical Center
184 Biobank. Samples were received on dry ice and stored at -80°C until isolation of DNA/RNA/protein. To
185 isolate DNA, tumor fragments of approximately 20 mg in mass were harvested on an aluminum block
186 pre-chilled on dry ice. Samples were then immediately transferred to a mortar and pestle containing a
187 small volume of liquid nitrogen. The fragments were pulverized in the mortar and pestle, and the liquid
188 nitrogen was allowed to evaporate. Next, samples were immediately processed with a
189 DNA/RNA/Protein Purification Plus kit (Norgen Biotek) following the standard operating protocol for
190 animal tissue. DNA concentration and quality were assessed using an ND-1000 (Nanodrop), a Qubit
191 Fluorometer (Thermo Scientific), and a Fragment Analyzer (Advanced Analytical Technologies). To

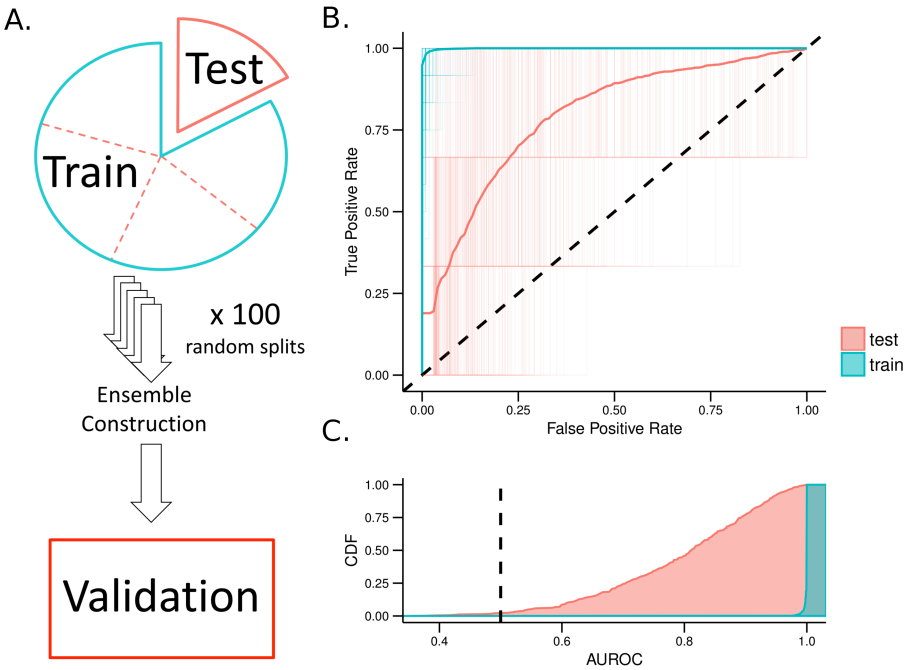
192 isolate RNA, -80 C tumor fragments were placed in 5-10 volumes of RNAlater-ICE Frozen Tissue
193 Transition Solution (Ambion) and placed at -20°C until RNA extraction with a mirVana miRNA isolation
194 kit, without phenol, following the standard operating protocol (Thermo Scientific). Samples were
195 homogenized using a manual homogenizer in the presence of mirVana lysis buffer. RNA concentration
196 and quality were determined using a Qubit Fluorometer (Thermo Scientific) and a Fragment Analyzer
197 (Advanced Analytical Technologies). To isolate protein, small tumor fragments were pulverized and lysed
198 in approximately 3 volumes of ice-cold radioimmunoprecipitation assay (RIPA) buffer (150 mM sodium
199 chloride, 1% v/v nonidet P40, 0.5% w/v sodium deoxycholate, 0.05% w/v sodium dodecyl sulfate, 50 mM
200 Tris pH 8.0) containing 1 mM sodium orthovanadate, 1 mM sodium fluoride, 1 mM
201 phenylmethylsulfonyl fluoride, and 1X protease inhibitor cocktail (0.1 µg/mL leupeptin, 100 µM
202 benzamidine HCl, 1 µM aprotinin, 0.1 µg/mL soybean trypsin inhibitor, 0.1 µg/mL pepstatin, 0.1 µg/mL
203 antipain). Samples were passed through a 25 ⅝ g needle and subsequently sonicated on ice to promote
204 efficient lysis and DNA shearing. After a 30 minute incubation on ice, lysates were cleared by
205 centrifuging at 16100 x g for 20 minutes. HEK293T, U87-MG, and U87-MG cells treated for two hours
206 with 1 micromolar bortezomib (Selleckchem) and 10 micromolar MG132 (Selleckchem) were also
207 prepared in RIPA buffer. Protein samples were stored at -80°C until analysis.
208
209 *Cell Culture*
210 U87-MG and HEK293T cells were purchased from ATCC. Cell lines were regularly passaged and
211 were cultured in Dulbecco's Modified Eagle Medium (Corning) with 10% v/v fetal bovine serum (Gibco)
212 at 37°C in 5% $CO_2$.
213 Recent data regarding the U87MG cell line published by Allen *et al* suggest that the U87MG cell
214 line distributed by ATCC is not from the same tumor as the cell line that was originally isolated in
215 Uppsala. Transcriptome analysis comparing ATCC U87MG cell line to known tumor transcriptomes

216 indicate that the ATCC U87MG cell line is a central nervous system tumor and is likely a glioblastoma cell
217 line [23].
218 In the present study, we employ this cell line as a control representing an NF1-deficient tumor
219 cell line. Previous studies have shown that the U87MG cell line has elevated proteasome-mediated
220 degradation of NF1 and that this cell line required the loss of NF1 protein to promote tumorigenesis in
221 xenograft tumor models [10]. Given that the ATCC U87MG cell line is a well-characterized and broadly-
222 used model of NF1 deficient tumor cells [10,24–26], we propose that the use of the ATCC U87MG cell
223 line is an appropriate control for Figure 2.
224
225 *RNA Microarray*
226 After RNA isolation and QC, samples were labeled for the GeneChip Human Transcriptome Array 2.0
227 (HTA 2.0, Affymetrix). Labeling was performed with Affymetrix Proprietary DNA Label (biotin-linked)
228 using a WT Plus Kit (Affymetrix) provided with the HTA 2.0, following the standard operating protocol for
229 HTA 2.0, including PolyA controls. Hybridization, washing, and staining were performed with the WT Plus
230 Kit, following the standard operating protocol for HTA 2.0. Washing and staining were performed using a
231 GeneChip Fluidics 450. Scanning was performed with a GeneChip Scanner 3000. These data were
232 deposited in the Gene Expression Omnibus under accession GSE85033.
233
234 *Validation Sample Processing*
235 We applied a quality control pipeline [27] to all CEL files generated by the HTA 2.0. All validation
236 samples passed processing quality control, which included an inspection of spatial artifacts, MA plots,
237 probe distributions, and sample comparison boxplots. We summarized transcript intensities using
238 robust multi-array analysis (RMA) [28]. We determined batch normalization was unnecessary after a
239 guided principal components analysis (gPCA) using sample processing date and array plate ID as

240 potential batch effect confounders [29]. Lastly, we collapsed HTA2.0 transcripts into gene level
241 measurements using the `collapseRows()` function with the "maxmean" method from the R package
242 WGCNA [30]. We used the pd.hta.2.0 platform design file (version 3.12.1) and the Bioconductor package
243 "hta20sttranscriptcluster.db" (version 8.3.1) to map manufacturer transcript IDs to genes. We
244 performed all preprocessing steps using R version 3.2.3.
245
246 *Western Blotting*
247 Prior to sodium dodecyl sulfate polyacrylamide gel electrophoresis, protein sample concentration
248 was determined using a Pierce BCA Protein Assay Kit (Thermo Scientific). Protein samples were prepared
249 with 1X Laemmli sample buffer (50 mM Tris pH 6.8, 0.02% w/v bromophenol blue, 2% w/v SDS, 10% v/v
250 glycerol, 1% v/v beta-mercaptoethanol, 12.5 mM EDTA) and 50 µg of tumor protein. Volumes were
251 normalized with RIPA buffer including the protease/phosphatase inhibitors described above. SDS-PAGE
252 was performed using a 4-15% Mini-PROTEAN TGX gel (Bio-Rad) for 1 hour at 120V. The samples were
253 then transferred to a nitrocellulose membrane for 2 hours and 45 minutes at 400 mA in cold transfer
254 buffer (384 mM glycine, 50 mM Tris, 20% methanol, 0.005% w/v sodium dodecyl sulfate. Following this,
255 the blots were then blocked in 5% w/v BSA or 5% w/v nonfat dry milk in Tris-buffered saline (137 mM
256 NaCl, 2.7 mM KCl, 19 mM Tris, 0.05% v/v Tween 20, pH 7.4) for 25 minutes. Immunoblotting was
257 performed with the following antibodies and conditions (vendor, species, diluent, dilution, incubation
258 time, incubation temperature): anti-NF1 D7R7D (Cell Signaling, rabbit, 2% BSA, 1:1000, overnight, 4°C),
259 anti-tubulin B-1-2-5 (Santa Cruz, mouse, 2% milk, 1:10000, 1 hour, RT), anti-EGFR D38B1 (Cell Signaling,
260 rabbit, 2% milk, 1:1000-1:2000, 1h, RT), p-ERK ½ (p44/42 MAPK) #9101 (Cell Signaling, rabbit, 2% BSA,
261 1:2000, overnight, 4°C), SUZ12 D39F6 #3737 (Cell Signaling, rabbit, 2% milk, 1:1000, overnight, 4°C).
262 Anti-NF1 D7R7D was a kind gift from Cell Signaling Technologies, Inc.

263 The binding of the primary antibodies was detected by incubation with secondary antibodies goat
264 anti-rabbit HRP 1:20000 or goat anti-mouse HRP 1:10000 (Jackson Immunoresearch Laboratories Inc.) at
265 room temperature in 2% milk in TBST and detection of HRP activity using Pierce ECL Western Blotting
266 substrate (Thermo Scientific), or in the case of *NF1*, SuperSignal West Femto Maximum Sensitivity
267 Substrate (Thermo Scientific). The chemiluminescent signal was captured with MED-B medical x-ray film
268 (Med X Ray Company Inc.). Between primary antibodies, the membrane was stripped twice for 10
269 minutes at room temperature using a mild stripping buffer containing 1.5% w/v glycine, 0.1% w/v SDS,
270 1% v/v Tween 20 at pH 2.2 (Abcam). One sample was eliminated due to low yield, and apparent
271 degradation as determined by western blotting (all proteins examined were undetectable with the
272 exception of tubulin, not shown). Densitometry was performed using Li-COR Image Studio Lite 5.0.
273 Briefly, intensity measurements for *NF1* and tubulin were taken using equally-sized regions for all bands.
274 The background was subtracted using the local median intensity from the left and right borders (size=2)
275 of each measurement region. *NF1* values were divided by tubulin intensity to adjust for protein loading.
276 All measurement ratios were then normalized by dividing values by the "U87+PI" measurement for each
277 blot, respectively.
278
279 *Reproducibility of Computational Analyses*
280 We provide software with a permissive open source license to reproduce all computational analyses
281 [31]. Ensuring a stable compute environment, we performed all analyses in a Docker image [32]. This
282 image and source code can be used to freely confirm, modify, and build upon this work.
283
284 **RESULTS:**
285 *Classifier performance*

286 Using 5-fold cross validation across a parameter sweep, we identified optimal hyperparameters
287 at alpha = 0.15 and L1 mixing = 0.1 (Supplementary Figure S1). To assess model performance, we
288 performed 100 random initializations of five-fold cross-validation. These models had mean test area
289 under the receiver operating characteristic curve (AUROC) of 0.77 (95% Quantiles: 0.53 – 0.95) and a
290 mean train AUROC of 0.997 (95% Quantile: 0.98 – 1.00) (Supplementary Figure S2). We repeated this
291 procedure after TDM transformation (Supplementary Figure S3) and achieved comparable results with
292 alpha = 0.15 and l1 mixing = 0.1 (mean test AUROC = 0.77, 95% Quantiles: 0.51 – 0.96; mean train
293 AUROC = 0.998, 95% Quantiles: 0.99 – 1.00) (Figure 1B). Because the validation set was measured by
294 microarray, we used the classifier trained on TDM transformed data to construct our ensemble
295 classifier. We also determined the Cohen's D effect size estimate for all training and testing partitions
296 across all 5-fold cross validation iterations of the TDM transformed model (Supplementary Fig S4). The
297 classifier consistently and robustly separated *NF1* wildtype and *NF1* inactivated GBM samples with high
298 effect sizes (Training: mean Cohen's D = 3.07, 95% CI = 2.24 – 4.16; Testing: mean Cohen's D = 1.27, 95%
299 CI = 0.19 – 2.67).
300




301

302 Figure 1: *Logistic regression classifier with elastic net penalty training and testing errors over 100*
303 *iterations for Training Distribution Matching (TDM) transformation of The Cancer Genome Atlas*
304 *Glioblastoma RNAseq data.* (A) Schematic describing the terms used for training, testing, and validating
305 our model. We applied 5-fold cross validation to the full dataset which consists of training and testing
306 splits in each fold. The model is then applied as an ensemble classifier on a set of in-house samples
307 (validation set) (B) Receiver operating characteristic (ROC) curves for all 500 classifiers that make up the
308 ensemble model applied to both training and testing set. Also shown is the aggregate performance of
309 the ensemble classifier. (C) The cumulative density of area under the ROC curve (AUROC) for training
310 and testing partitions.
311

312 *Identification and characterization of NF1 deficient glioblastoma tumor samples*

313 We characterized NF1 protein concentrations as well as other molecules involved in RAS signaling in
314 the 12 GBM samples (Figure 2A). Two samples (CB2, 3HQ) had no apparent NF1 protein. Eight other
315 samples had similar or less NF1 signal than the U87-MG NF1-low control (H5M, LNA, YXL, VVN, R7K,
316 TRM, UNY, W31). Two samples (PBH, RIW) had equal or greater NF1 than the positive control, U87-MG +
317 proteasome inhibitors (preventing NF1 degradation). We also observed variable EGFR content in these
318 samples, with non-existent to low levels (3HQ, YXL, R7K), or medium to large EGFR signal (CB2, H5M,
319 PBH, LNA, YXL, VVN, RIW, TRM, UNY, W31). All GBM samples had high concentrations of phospho-
320 ERK1/2 signal relative to cell line controls. Samples with increased phospho-ERK1/2 may have greater

321 Ras pathway activation. This can be attributed to multiple factors, including increased EGFR expression
322 and/or NF1 inactivation.

323 Our ensemble classifier predicted four samples to have NF1 inactivation (CB2, UNY, R7K, and 3HQ)
324 and eight samples to be NF1 wildtype (W31, TRM, PBH, VVN, LNA, RIW, H5M, and YXL) (Figure 2B).
325 Because two samples, (CB2 and H5M) were measured on both western blots (Figure 2C), we used the
326 mean of their NF1 protein level across both experiments.

327 We performed a one-tailed Welch's t-test to determine if NF1 protein concentrations were
328 significantly higher in *NF1* wildtype versus *NF1* deficient samples based on our classifier predictions
329 (Figure 2D). We did not observe a significant difference across groups ($t$ = -1.38, $p$ = 0.098, effect size =
330 0.699). Additionally, while the effect size was fairly large, a power analysis indicated that 22 samples per
331 group would be required to achieve a power = 0.8 at that effect size. With a lack of glioblastoma
332 samples with quantified NF1 protein available, the trend of less protein present in samples scored as
333 *NF1* inactivated by the classifier nevertheless remains promising.

334 One of the samples predicted to be NF1 inactive contains detectable NF1 protein (R7K), suggesting
335 that this sample may have NF1 inactivation not detectable by assaying protein, have a different
336 alteration that phenocopies NF1 loss, or is incorrectly predicted by the classifier. Conversely, there are
337 three samples predicted to be NF1 wildtype that have low or undetectable protein (YXL, VVN, W31),
338 which either indicates unknown elements that confound the detection of some NF1 dysregulated
339 tumors or a classification error.

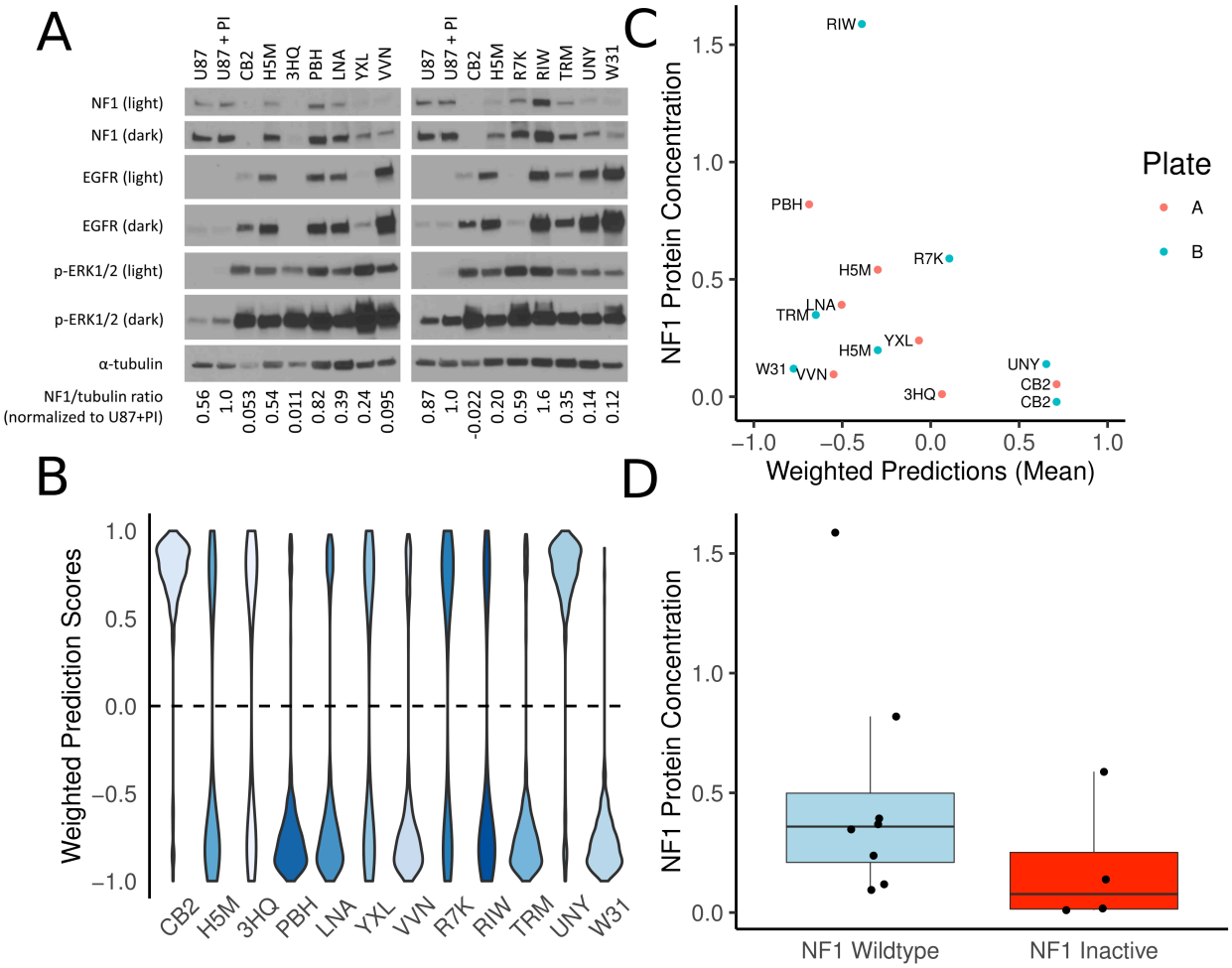
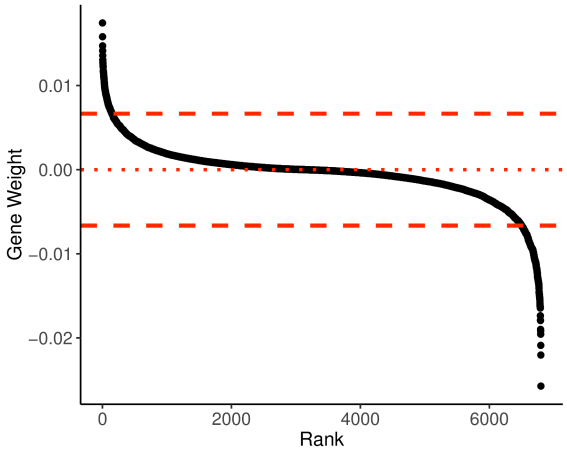



Figure 2: *Performance of our classifier on a validation set.* (A) Two distinct western blots for each of our twelve samples. The controls are U87-MG, an *NF1* WT glioblastoma cell line that exhibits proteasomal degradation of the NF1 protein. U87+PI are U87-MG cells are treated with the proteasome inhibitors (PI) MG-132 and bortezomib to block proteasome-mediated degradation of NF1. We used the NF1/tubulin ratio normalized to U87+PI as our NF1 protein level estimate. (B) Prediction scores for each of the 500 classifiers weighted by cross validation test set AUROC where a negative number indicates *NF1* wildtype and a positive number is indicates NF1 inactivation. Increasing grayscale indicates higher observed NF1 protein concentrations. (C) We quantify protein against U87+PI and provide the mean of the weighted predictions. (D) Based on weighted predictions, we show the abundance of NF1 protein compared to U87+PI.

*Highly Contributing Genes*

We observed several genes that consistently contributed to the ensemble classifier performance (Figure 3). Since we applied several classifiers to the validation set as an ensemble, we took the sum of

all classifier's gene weights across all 500 iterations to define these consistently contributing genes. While the data indicate that these genes have an impact on classifier performance, the data do not indicate whether changes in the expression of these genes are a direct consequence in changes in NF1 signaling. Expression of genes such as *TXNIP, ARRDC4, ISPD, C10orf107*, and *DUSP18* appear to be predictive of intact NF1 signaling. Among the list of genes that appear to be expressed in tumors with loss of NF1 function are *QPRT, ATF5, HUS1B, PEG10, HMGA2, RSL1D1, and NRG1*. A full list of positive and negative weight genes that were two standard deviations beyond the gene weight distribution is provided in Supplementary Table S2.




Top Contributing Genes

| Positive | Negative |
|---|---|
| 1. QPRT | TXNIP |
| 2. KCNK5 | DNAH5 |
| 3. ATF5 | ISPD |
| 4. HUS1B | CUX1 |
| 5. PEG10 | PRSS27 |
| 6. DHDH | C10orf107 |
| 7. HMGA2 | DUSP18 |
| 8. CNGA1 | TMEM52 |
| 9. RSL1D1 | GDF9 |
| 10. PRKCQ | ARRDC4 |

Figure 3: *Genes that contribute to the classifier performance.* Genes are shown ranked by their weighted contribution to the ensemble classifier. Weights are scaled to unit norm. The top 10 positive and top 10 negative contributing high weight genes are given on the right.

We also performed over-representation analysis of the most influential genes in the classifier to identify gene ontology (GO) sets and pathways that may be predictive of NF1 status [33–36]. S3For high-weight genes predictive of intact NF1 signaling, we observed GO sets involved in plasma membrane-localized proteins (GO:0005886, GO:0071944, GO:0016324) and homeostasis (GO:0048871, GO:0001659, GO:0048873, GO:0031224), among others. Annotated pathways associated with genes from this dataset include hematopoietic stem cell differentiation, thyroid cancer, voltage-gated potassium channels, and RHO GTPase functional pathways.

375 For high-weight genes predictive of NF1 loss of function, we observed GO sets related to cellular
376 adhesion (GO:0007155, GO:0098742), negative regulation of signaling (GO:0009968, GO:0023507,
377 GO:0010648), and nervous system development (GO:0051962, GO0007416, GO: 0050808), among
378 others. These genes were also enriched for elements of the phototransduction cascade and thyroxine
379 production pathways.
380
381 **DISCUSSION:**
382 A machine learning classifier, based on gene expression data, can capture signal associated with the
383 inactivation of a tumor suppressor. Our classifier is able to detect subtle downstream changes in gene
384 expression as a result of the tumor responding to NF1 loss of function. This finding supports using mRNA
385 as a summary measurement capable of capturing system-wide responses to molecular events beyond
386 transcription factor alterations. Machine learning has been applied to gene expression in a variety of
387 studies with various goals [37–41]. In a similar study, Guinney *et al.* trained a classifier to model RAS
388 activity in colorectal cancer and demonstrated its clinical utility by predicting response to MEK inhibitors
389 and anti-EGFR based treatments [18]. With a wealth of signal embedded in gene expression and a
390 rapidly growing library of datasets, the performance of machine learning models is likely to rapidly
391 improve. An increase in performance leads to more reliable clinical applications that would potentially
392 predict the effectiveness of pathway-specific targeted therapies.
393 While our classifier was able to predict NF1 inactivation status to an extent, its performance is far
394 from being clinically actionable. A major difficulty in developing a reliable classifier in this case is
395 contamination in gold standard positives and negatives. While we aim to detect NF1 inactivation events,
396 our gold standard positives can only include samples with known *NF1* mutation status. Conversely, we
397 expect that negative samples (about 90% of the data) are also contaminated with NF1 inactivated
398 samples due to protein loss and other mechanisms. We cannot determine scenarios where NF1 is

399 inactivated beyond mutation at scale in the TCGA data. Another challenge with the construction of
400 classifiers from such data is overfitting. Even after hyperparameter optimization we observed
401 substantial overfitting (Figure 2), which has also been observed in competitions (see, for example,
402 supplementary figure S2 of Noren *et al*. 2016 [42] in which the best performing algorithms also overfit).
403 Finally with a small number of positive examples the model performance is unstable, which
404 demonstrates high variability in gold standard samples used to train the model [43]. We employed
405 ensemble classification to mitigate this issue as averaging over heterogeneous models would result in a
406 relatively stable classifier (see Figure 2B). In summary, our results are promising but these challenges are
407 substantial and significant work remains to reach a robust classifier with clinical utility.

408 The performance of the classifier appears to be impacted by many cancer related genes. For
409 example, genes such as *TXNIP* and *ARRDC4*, which are both indicative of lactic acidosis, correlate with
410 better clinical outcomes, and contribute to predicting tumors with intact NF1 signaling [44]. We also
411 observed transcripts that are more highly expressed in brain tissue than either other normal tissue
412 (*ISPD, C10orf107*), or more highly expressed in normal brain tissue than glioma (*EPHA5*) [45–47].
413 *DUSP18* contributes to the prediction of *NF1* wildtype status and is a negative regulator of ERK
414 phosphorylation, possibly by regulating *SHP2* phosphorylation [48]. It is unclear whether the expression
415 of these genes is a direct result of *NF1* expression, the result of signaling downstream of *NF1*, or a
416 consequence of other phenomena (such as expression of *SPRED1*, an NF1 binding partner that is
417 essential for *NF1* signaling). Future studies could elucidate the potential connections between *NF1* and
418 the genes identified as important for the performance of this classifier.

419 Over-representation analysis of these data highlighted changes in potassium channel expression. It
420 was previously demonstrated that *NF1* wild-type Schwann cells have altered K+ channel activity as
421 compared to $NF1^{-/-}$ Schwann cells suggesting that this may be one factor by which *NF1* mutant and wild-
422 type cells can be distinguished [49].

423 Regarding prediction of NF1 inactivated tumors, we observed several genes that have been linked to
424 cancer such as *QPRT*, which is highly expressed in malignant pheochromocytomas as compared to
425 benign; *RSL1D1* (CSIG), which stabilizes *c-myc* in hepatocellular carcinoma; *PPEF*, which is highly
426 expressed in astrocytic gliomas as compared to normal brain tissue [50–52]; and *PEG10*, a poor
427 prognostic marker and regulator of proliferation, migration, and invasion in several tumor types [53–55].
428 We also observed *ATF5,* a gene for which expression in malignant glioma is correlated with poor survival
429 [56]. Knockdown of *ATF5* in GBM cells causes cell death *in vitro* and *in vivo* [57]. Analysis of genes that
430 contribute to the prediction of *NF1* inactivation yielded several GO terms related to neural
431 development. It is well established that loss of *NF1* can result in abnormal neural development and/or
432 tumorigenesis [14,58,59]. We also observed genes associated with the mesodermal commitment
433 pathway, components of which are linked to the epithelial to mesenchymal transition in human cancer
434 cells [60–62]. Analysis of this pathway may be informative in identifying tumors with *NF1* loss because
435 mesenchymal GBMs are enriched for tumors with *NF1* loss [63].
436 Our ensemble classifier was able to robustly detect the samples with the highest and lowest NF1
437 protein concentrations, but it struggled with samples of intermediate NF1 concentrations. This could be
438 a result of an enrichment of mechanisms causing NF1 inactivation beyond protein abundance, an
439 overrepresentation of mesenchymal tumors in NF1 inactivated samples contaminating dataset splits
440 [63], poor classifier generalizability, or incomplete data transformation between RNAseq and microarray
441 data. Because training and testing performance were similar between transformed and non-
442 transformed data (see Figure 1 and Supplementary Fig S2) we don't anticipate performance to be
443 impacted much by platform differences or classifier generalizability. Nevertheless, we demonstrated the
444 ability of system-wide gene expression measurements to capture downstream consequences of a
445 complex biological mechanism that would otherwise require several different types of data acquisition
446 to capture.

447

448 **CONCLUSIONS:**

449 A machine learning classifier for transcriptomic data was able to detect signal associated with the
450 inactivation of *NF1*, a tumor suppressor gene. The gene is an important regulator of the oncogene *RAS*
451 and is inactivated frequently in GBM and in other tumors. The measurement of NF1 inactivity cannot be
452 comprehensively captured by any single genomic characterization such as targeted sequencing or
453 fluorescence in situ hybridization. This difficulty arises from diverse and complex biological mechanisms
454 that inactivate the tumor suppressor in a variety of ways. However, we demonstrated that measuring
455 system-wide RNA can capture subtle downstream changes that occur in response to NF1 inactivation.
456 Improving classification performance is required before transitioning such a model into clinical use, but
457 our method could be used to characterize cell lines or patient-derived xenograft (PDX) models with
458 inactive NF1. Eventually, with more data and improved classification, we expect machine-learning
459 models constructed on system-wide transcriptomics will translate into clinically relevant predictions that
460 will guide targeted therapy.

461

462 **DECLARATIONS:**

463 *Ethics approval and consent to participate*

464 The Dartmouth-Hitchcock Medical Center Committee for the Protection of Human Subjects declared
465 IRB review and approval is not required since the study does not meet the regulatory definition of
466 human subjects.

467

468 *Consent for publication*

469 Not Applicable

470

*Availability of data and materials*

All source code is available under a permissive open source license in the nf1_inactivation GitHub repository, https://github.com/greenelab/nf1_inactivation. We also provide a docker image to replicate the computational environment at https://hub.docker.com/r/gregway/nf1_inactivation/. Additionally, all validation data is available under Gene Expression Omnibus accession number GSE85033.

*Competing interests*

The authors declare no competing interests.

*Funding*

This work was funded by the Genomics and Computational Biology PhD Program at The University of Pennsylvania to GPW, The Gordon and Betty Moore Foundation GBMF4552 to CSG, NINDS R01 NS095411 to YS and CSG, Children's Tumor Foundation Young Investigator Award 2014-01-12 to RJA, a Nancy P. Shea Trust grant to YS, a Prouty Pilot Grant from the Friends of the Norris Cotton Cancer Center to YS, a Synergy grant to YS, CSG and CF funded by NIH NCATS UL1 TR001086 to The Dartmouth Center for Clinical and Translational Science, NCI Cancer Center Support Grant P30 CA023108 to the Dartmouth-Hitchcock Norris Cotton Cancer Center. RJA is an Albert J. Ryan Fellow.

*Authors' contributions*

GPW built, analyzed, and interpreted the classifier, generated the figures, created all source code, and wrote the manuscript. RJA performed all experiments, interpreted the classifier, and wrote the manuscript. SJB interpreted the results and was a major contributor to the manuscript. CEF was a major contributor to the study design and was a major contributor to the manuscript. YS designed

495 the study, interpreted the results, and wrote the manuscript. CSG designed the study, interpreted

496 the results, and wrote the manuscript. All authors read and approved the final manuscript.

497

498 *Acknowledgements*

499 This work was supported by the MMC BioBank, a core facility of the Maine Medical Center Research

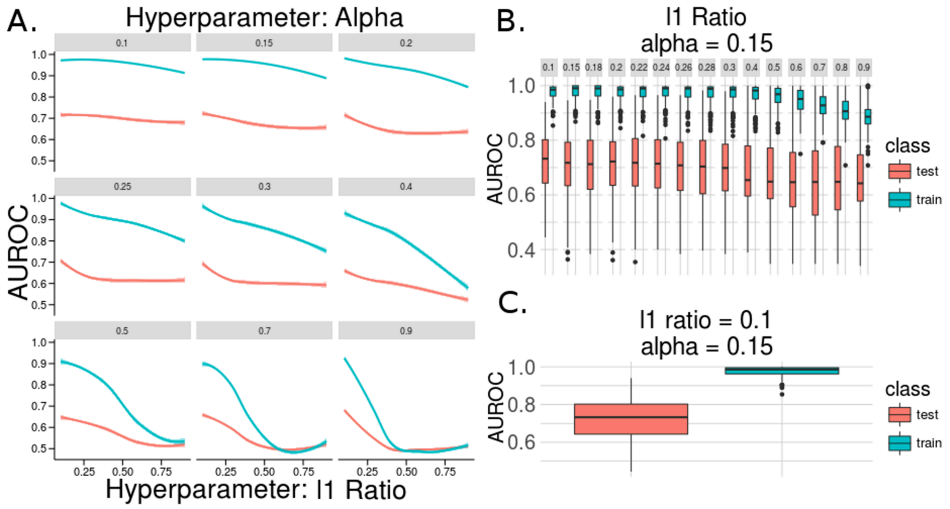
500 Institute, and the Dartmouth Genomics Shared Resource, a core facility of the Norris Cotton Cancer

501 Center.

502

503 **SUPPLEMENTARY FIGURES:**

504




505

506 <u>Supplementary Figure S1:</u> *Non-transformed RNAseq results of The Cancer Genome Atlas Glioblastoma*

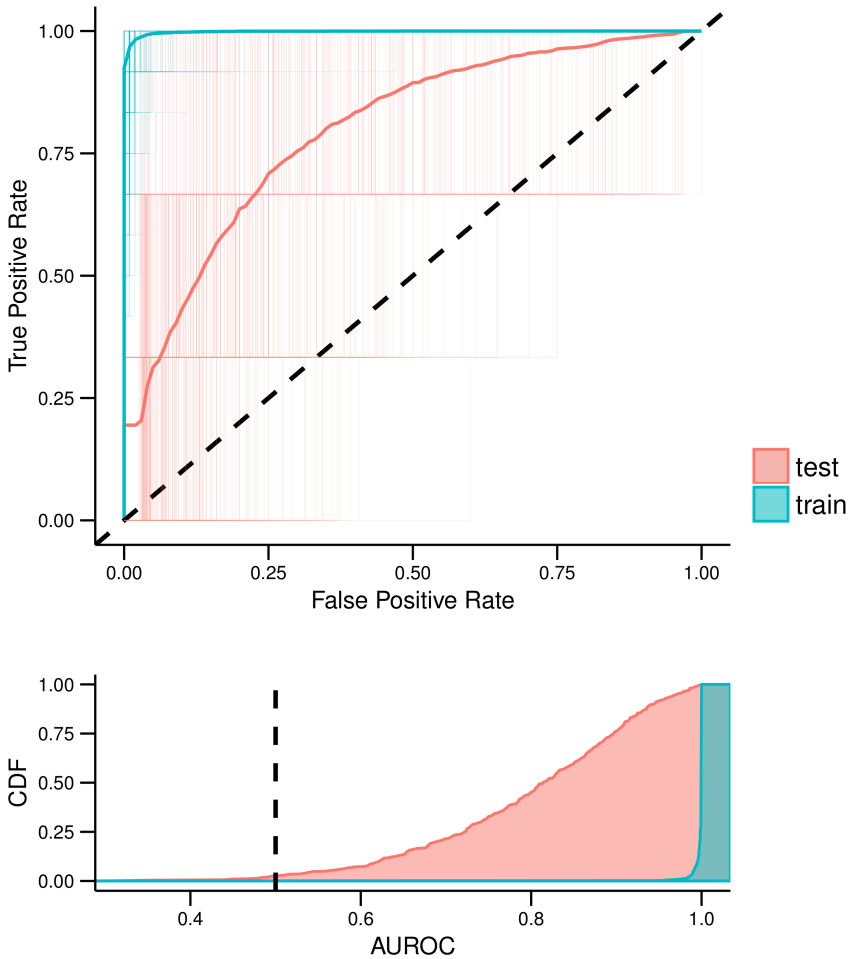
507 *parameter sweep for stochastic gradient descent logistic classifiers with elastic net penalty.* (A) Training

508 and testing area under the receiver operating characteristic curve (AUROC) ass is given for each

509 parameter tested. All accuracies are presented following 5-fold cross validation after 50 random

510 initializations. (B) The l1 mixing parameter with the optimal alpha and (C) the classifier performance

511 across all random starts for the best hyperparameters.




512

513 Supplementary Figure S2: *Logistic regression classifier with elastic net penalty training and testing errors*

514 *over 100 iterations for non-transformed The Cancer Genome Atlas Glioblastoma RNAseq data.* (A)

515 Receiver operating characteristic (ROC) curve and shows the average training and testing performance

516 of 5-fold cross validation over 100 random initializations as well as each individual classifier in the

517 ensemble model. (B) The cumulative density of area under the ROC curve (AUROC) for all training and

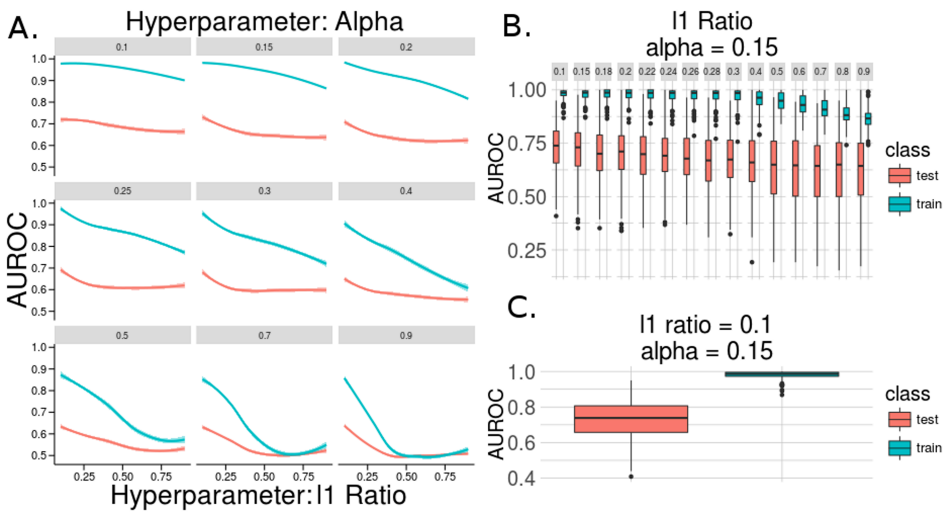
518 testing partitions.

519

520

521

522




523

524

525 Supplementary Figure S3: *Training Distribution Matching (TDM) transformation of RNAseq results of The*

526 *Cancer Genome Atlas Glioblastoma parameter sweep for stochastic gradient descent logistic classifier*

527 *with elastic net penalty.* (A) Training and testing area under the receiver operating characteristic curve

528 (AUROC) is given for each parameter tested. All accuracies are presented following 5-fold cross

529 validation after 100 random initializations. (B) The l1 mixing parameter with the optimal alpha and (C)

530 the classifier performance across all random starts for the best hyperparameters.

531

532

533

534

535

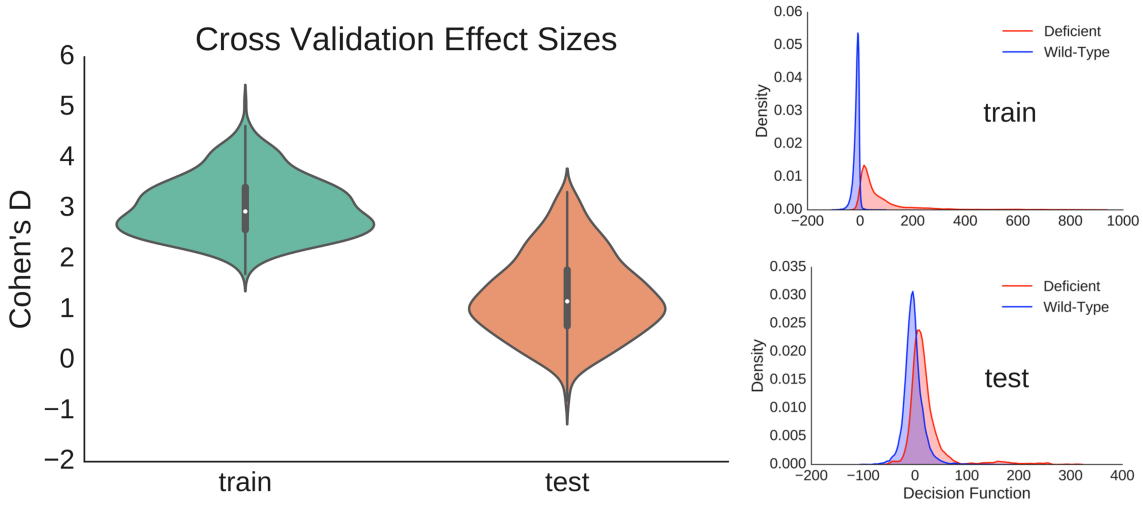
536

537

538

539

540




Supplementary Figure S4: Cohen's D effect size estimates across 5 fold cross validation parameters for all 100 iterations of the TDM transformed ensemble classifier. The effect size for the test set is consistently lower than the training set (left). Additionally, the training and testing decision functions for gold standard NF1 deficient vs. NF1 wildtype samples shows a difference in mean estimates (right). The decision function represents the raw score of all samples as applied to the respective classifiers through each of the 100 iterations of 5 fold cross validation on the TCGA training set.

**REFERENCES:**

1. Martin GA, Viskoohil D, Bollag G, McCabe PC, Crosier WJ, Haubruck H, et al. The GAP-related domain of the neurofibromatosis type 1 gene product interacts with ras p21. Cell. 1990;63:843–9.

2. Xu G, O'Connell P, Viskochil D, Cawthon R, Robertson M, Culver M, et al. The neurofibromatosis type 1 gene encodes a protein related to GAP. Cell. 1990;62:599–608.

3. Boyd KP, Korf BR, Theos A. Neurofibromatosis type 1. J. Am. Acad. Dermatol. 2009;61:1–14.

4. Dogra B, Rana K. Facial plexiform neurofibromatosis: A surgical challenge. Indian Dermatol. Online J. 2013;4:195.

5. Evans DGR, Baser M, McGaughran J, Sharif S, Howard E, Moran A. Malignant peripheral nerve sheath tumours in neurofibromatosis 1. J. Med. Genet. 2002;39:311–4.

6. Rad E, Tee AR. Neurofibromatosis type 1: Fundamental insights into cell signalling and cancer. Semin. Cell Dev. Biol. 2016;52:39–46.

561 7. Ratner N, Miller SJ. A RASopathy gene commonly mutated in cancer: the neurofibromatosis type 1
562 tumour suppressor. Nat. Rev. Cancer. 2015;15:290–301.

563 8. Wood M, Rawe M, Johansson G, Pang S, Soderquist RS, Patel AV, et al. Discovery of a Small Molecule
564 Targeting IRA2 Deletion in Budding Yeast and Neurofibromin Loss in Malignant Peripheral Nerve Sheath
565 Tumor Cells. Mol. Cancer Ther. 2011;10:1740–50.

566 9. Allaway RJ, Fischer DA, de Abreu FB, Gardner TB, Gordon SR, Barth RJ, et al. Genomic characterization
567 of patient-derived xenograft models established from fine needle aspirate biopsies of a primary
568 pancreatic ductal adenocarcinoma and from patient-matched metastatic sites. Oncotarget.
569 2016;7:17087–102.

570 10. McGillicuddy LT, Fromm JA, Hollstein PE, Kubek S, Beroukhim R, De Raedt T, et al. Proteasomal and
571 Genetic Inactivation of the NF1 Tumor Suppressor in Gliomagenesis. Cancer Cell. 2009;16:44–54.

572 11. Subramanian S, Thayanithy V, West RB, Lee C-H, Beck AH, Zhu S, et al. Genome-wide transcriptome
573 analyses reveal p53 inactivation mediated loss of miR-34a expression in malignant peripheral nerve
574 sheath tumours. J. Pathol. 2010;220:58–70.

575 12. Wallace MR, Andersen LB, Saulino AM, Gregory PE, Glover TW, Collins FS. A de novo Alu insertion
576 results in neurofibromatosis type 1. Nature. 1991;353:864–6.

577 13. Skuse GR, Cappione AJ, Sowden M, Metheny LJ, Smith HC. The Neurofibromatosis Type I Messenger
578 RNA Undergoes Base-Modification RNA Editing. Nucleic Acids Res. 1996;24:478–86.

579 14. Cichowski K, Jacks T. NF1 tumor suppressor gene function: narrowing the GAP. Cell. 2001;104:593–
580 604.

581 15. UCSC Xena [Internet]. Available from: http://xena.ucsc.edu/

582 16. Brennan CW, Verhaak RGW, McKenna A, Campos B, Noushmehr H, Salama SR, et al. The Somatic
583 Genomic Landscape of Glioblastoma. Cell. 2013;155:462–77.

584 17. Thompson JA, Tan J, Greene CS. Cross-platform normalization of microarray and RNA-seq data for
585 machine learning applications. PeerJ. 2016;4:e1621.

586 18. Guinney J, Ferte C, Dry J, McEwen R, Manceau G, Kao K, et al. Modeling RAS Phenotype in Colorectal
587 Cancer Uncovers Novel Molecular Traits of RAS Dependency and Improves Prediction of Response to
588 Targeted Agents in Patients. Clin. Cancer Res. 2014;20:265–72.

589 19. Liang Y, Liu C, Luan X-Z, Leung K-S, Chan T-M, Xu Z-B, et al. Sparse logistic regression with a L1/2
590 penalty for gene selection in cancer classification. BMC Bioinformatics. 2013;14:198.

591 20. Sokolov A, Carlin DE, Paull EO, Baertsch R, Stuart JM. Pathway-Based Genomics Prediction using
592 Generalized Elastic Net. Przytycka TM, editor. PLOS Comput. Biol. 2016;12:e1004790.

593 21. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine
594 Learning in Python. CoRR. 2012;

595 22. Cohen J. Statistical power analysis for the behavioral sciences. New York: Academic Press; 1969.

596 23. Allen M, Bjerke M, Edlund H, Nelander S, Westermark B. Origin of the U87MG glioma cell line: Good
597 news and bad news. Sci. Transl. Med. 2016;8:354re3–354re3.

598 24. Daginakatte GC, Gutmann DH. Neurofibromatosis-1 (Nf1) heterozygous brain microglia elaborate
599 paracrine factors that promote Nf1-deficient astrocyte and glioma growth. Hum. Mol. Genet.
600 2007;16:1098–112.

601 25. Hollstein PE, Cichowski K. Identifying the Ubiquitin Ligase Complex That Regulates the NF1 Tumor
602 Suppressor and Ras. Cancer Discov. 2013;3:880–93.

603 26. Tan X, Wang S, Yang B, Zhu L, Yin B, Chao T, et al. The CREB-miR-9 Negative Feedback Minicircuitry
604 Coordinates the Migration and Proliferation of Glioma Cells. Li J, editor. PLoS ONE. 2012;7:e49570.

605 27. Carvalho BS, Irizarry RA. A framework for oligonucleotide microarray preprocessing. Bioinforma. Oxf.
606 Engl. 2010;26:2363–7.

607 28. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip
608 probe level data. Nucleic Acids Res. 2003;31:e15.

609 29. Reese SE, Archer KJ, Therneau TM, Atkinson EJ, Vachon CM, de Andrade M, et al. A new statistic for
610 identifying batch effects in high-throughput genomic data that uses guided principal component
611 analysis. Bioinformatics. 2013;29:2877–83.

612 30. Miller JA, Cai C, Langfelder P, Geschwind DH, Kurian SM, Salomon DR, et al. Strategies for
613 aggregating gene expression data: The collapseRows R function. BMC Bioinformatics. 2011;12:322.

614 31. Greg Way. nf1_inactivation: Pre-Release. 2016 [cited 2016 Aug 1]; Available from:
615 http://dx.doi.org/10.5281/zenodo.58864

616 32. Boettiger C. An introduction to Docker for reproducible research. ACM SIGOPS Oper. Syst. Rev.
617 2015;49:71–9.

618 33. Kamburov A, Wierling C, Lehrach H, Herwig R. ConsensusPathDB--a database for integrating human
619 functional interaction networks. Nucleic Acids Res. 2009;37:D623–8.

620 34. Kamburov A, Stelzl U, Lehrach H, Herwig R. The ConsensusPathDB interaction database: 2013
621 update. Nucleic Acids Res. 2013;41:D793–800.

622 35. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the
623 unification of biology. Nat. Genet. 2000;25:25–9.

624 36. The Gene Ontology Consortium. Gene Ontology Consortium: going forward. Nucleic Acids Res.
625 2015;43:D1049–56.

626 37. Molla M, Waddell M, Page D, Shavlik J. Using Machine Learning to Design and Interpret Gene-
627 Expression Microarrays. AI Mag. 2004;25:23–44.

628 38. Bastani M, Vos L, Asgarian N, Deschenes J, Graham K, Mackey J, et al. A Machine Learned Classifier
629 That Uses Gene Expression Data to Accurately Predict Estrogen Receptor Status. Rogers S, editor. PLoS
630 ONE. 2013;8:e82144.

631 39. Pirooznia M, Yang JY, Yang MQ, Deng Y. A comparative study of different machine learning methods
632 on microarray gene expression data. BMC Genomics. 2008;9:S13.

633 40. Chou W-C, Ma Q, Yang S, Cao S, Klingeman DM, Brown SD, et al. Analysis of strand-specific RNA-seq
634 data using machine learning reveals the structures of transcription units in Clostridium thermocellum.
635 Nucleic Acids Res. 2015;43:e67–e67.

636 41. Tan J, Hammond JH, Hogan DA, Greene CS. ADAGE-Based Integration of Publicly Available
637 *Pseudomonas aeruginosa* Gene Expression Data with Denoising Autoencoders Illuminates Microbe-Host
638 Interactions. Gilbert JA, editor. mSystems. 2016;1:e00025–15.

639 42. Noren DP, Long BL, Norel R, Rrhissorrakrai K, Hess K, Hu CW, et al. A Crowdsourcing Approach to
640 Developing and Assessing Prediction Algorithms for AML Prognosis. Tan K, editor. PLOS Comput. Biol.
641 2016;12:e1004890.

642 43. Yu B. Stability. Bernoulli. 2013;19:1484–500.

643 44. Chen JL-Y, Merl D, Peterson CW, Wu J, Liu PY, Yin H, et al. Lactic Acidosis Triggers Starvation
644 Response with Paradoxical Induction of TXNIP through MondoA. Gibson G, editor. PLoS Genet.
645 2010;6:e1001093.

646 45. Willer T, Lee H, Lommel M, Yoshida-Moriguchi T, de Bernabe DBV, Venzke D, et al. ISPD loss-of-
647 function mutations disrupt dystroglycan O-mannosylation and cause Walker-Warburg syndrome. Nat.
648 Genet. 2012;44:575–80.

649 46. Thierry-Mieg D, Thierry-Mieg J. AceView: a comprehensive cDNA-supported gene and transcripts
650 annotation. Genome Biol. 2006;7 Suppl 1:S12.1–14.

651 47. Almog N, Ma L, Raychowdhury R, Schwager C, Erber R, Short S, et al. Transcriptional Switch of
652 Dormant Tumors to Fast-Growing Angiogenic Phenotype. Cancer Res. 2009;69:836–44.

653 48. Sacco F, Boldt K, Calderone A, Panni S, Paoluzi S, Castagnoli L, et al. Combining affinity proteomics
654 and network context to identify new phosphatase substrates and adapters in growth pathways. Front.
655 Genet. [Internet]. 2014 [cited 2016 Aug 1];5. Available from:
656 http://journal.frontiersin.org/article/10.3389/fgene.2014.00115/abstract

657 49. Xu Y, Chiamvimonvat N, Vázquez AE, Akunuru S, Ratner N, Yamoah EN. Gene-targeted deletion of
658 neurofibromin enhances the expression of a transient outward K+ current in Schwann cells: a protein
659 kinase A-mediated mechanism. J. Neurosci. Off. J. Soc. Neurosci. 2002;22:9194–202.

660 50. Thouënnon E, Elkahloun AG, Guillemot J, Gimenez-Roqueplo A-P, Bertherat J, Pierre A, et al.
661 Identification of Potential Gene Markers and Insights into the Pathophysiology of Pheochromocytoma
662 Malignancy. J. Clin. Endocrinol. Metab. 2007;92:4865–72.

663 51. Cheng Q, Yuan F, Lu F, Zhang B, Chen T, Chen X, et al. CSIG promotes hepatocellular carcinoma
664 proliferation by activating c-MYC expression. Oncotarget. 2015;6:4733–44.

665 52. Bageritz J, Puccio L, Piro RM, Hovestadt V, Phillips E, Pankert T, et al. Stem cell characteristics in
666 glioblastoma are maintained by the ecto-nucleotidase E-NPP1. Cell Death Differ. 2014;21:929–40.

667 53. Deng X, Hu Y, Ding Q, Han R, Guo Q, Qin J, et al. PEG10 plays a crucial role in human lung cancer
668 proliferation, progression, prognosis and metastasis. Oncol. Rep. 2014;32:2159–67.

669 54. Li C-M, Margolin AA, Salas M, Memeo L, Mansukhani M, Hibshoosh H, et al. PEG10 is a c-MYC target
670 gene in cancer cells. Cancer Res. 2006;66:665–72.

671 55. Akamatsu S, Wyatt AW, Lin D, Lysakowski S, Zhang F, Kim S, et al. The Placental Gene PEG10
672 Promotes Progression of Neuroendocrine Prostate Cancer. Cell Rep. 2015;12:922–36.

673 56. Sheng Z, Li L, Zhu LJ, Smith TW, Demers A, Ross AH, et al. A genome-wide RNA interference screen
674 reveals an essential CREB3L2-ATF5-MCL1 survival pathway in malignant glioma with therapeutic
675 implications. Nat. Med. 2010;16:671–7.

676 57. Greene LA, Lee HY, Angelastro JM. The transcription factor ATF5: role in neurodevelopment and
677 neural tumors. J. Neurochem. 2009;108:11–22.

678 58. Zhu Y, Romero MI, Ghosh P, Ye Z, Charnay P, Rushing EJ, et al. Ablation of NF1 function in neurons
679 induces abnormal development of cerebral cortex and reactive gliosis in the brain. Genes Dev.
680 2001;15:859–76.

681 59. Joseph NM, Mosher JT, Buchstaller J, Snider P, McKeever PE, Lim M, et al. The Loss of Nf1 Transiently
682 Promotes Self-Renewal but Not Tumorigenesis by Neural Crest Stem Cells. Cancer Cell. 2008;13:129–40.

683 60. Morishita A, Zaidi MR, Mitoro A, Sankarasharma D, Szabolcs M, Okada Y, et al. HMGA2 is a driver of
684 tumor metastasis. Cancer Res. 2013;73:4289–99.

685 61. de Boeck M, Cui C, Mulder AA, Jost CR, Ikeno S, ten Dijke P. Smad6 determines BMP-regulated
686 invasive behaviour of breast cancer cells in a zebrafish xenograft model. Sci. Rep. 2016;6:24968.

687 62. Salomonis N, Mshel016, Cirillo E, Hanspers K, Kutmon M. Mesodermal Commitment Pathway (Homo
688 sapiens). http://www.wikipathways.org/index.php/Pathway:WP2857.

689 63. Verhaak RGW, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, et al. Integrated genomic
690 analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA,
691 IDH1, EGFR, and NF1. Cancer Cell. 2010;17:98–110.

692