Classifying cancer genome aberrations by their mutually exclusive effects on transcription

Jonathan B. Dayton[1], Stephen R. Piccolo[1,2,*]

[1] Department of Biology, Brigham Young University, Provo, UT 84602, United States of America

[2] Department of Biomedical Informatics, University of Utah, Salt Lake City, UT 84108, United States of America

[*] Corresponding Author

Email Addresses:

JBD: jonathandayton23@gmail.com

SRP: stephen_piccolo@byu.edu

Abstract

Background

Malignant tumors are typically caused by a conglomeration of genomic aberrations—including point mutations, small insertions, small deletions, and large copy-number variations. In some cases, specific chemotherapies and targeted drug treatments are effective against tumors that harbor certain genomic aberrations. However, predictive aberrations (biomarkers) have not been identified for many tumor types and treatments. One way to address this problem is to examine the downstream, transcriptional effects of genomic aberrations and to identify characteristic patterns. Even though two tumors harbor different genomic aberrations, the transcriptional effects of those aberrations may be similar. These patterns could be used to inform treatment choices.

Methods

We used data from 9300 tumors across 25 cancer types from The Cancer Genome Atlas. We used supervised machine learning to evaluate our ability to distinguish between tumors that had mutually exclusive genomic aberrations in specific genes. An ability to accurately distinguish between tumors with aberrations in these genes suggested that the genes have a relatively different downstream effect on transcription, and vice versa. We compared these findings against prior knowledge about signaling networks and drug responses.

Results

Our analysis recapitulates known relationships in cancer pathways and identifies gene pairs known to predict responses to the same treatments. For example, in lung adenocarcinomas, gene-expression profiles from tumors with somatic aberrations in EGFR or MET were negatively correlated with each other, in line with prior knowledge that MET amplification causes resistance to EGFR inhibition. In breast carcinomas, we observed high similarity between PTEN and PIK3CA, which play complementary roles in regulating cellular proliferation. In a pan-cancer analysis, we found that genomic aberrations in BRAF and VHL exhibit downstream effects that are clearly distinct from other genes.

Conclusion

We show that transcriptional data offer promise as a way to group genomic aberrations according to their downstream effects, and these groupings recapitulate known relationships. Our approach shows potential to help pharmacologists and clinical trialists narrow the search space for candidate gene/drug associations, including for rare mutations, and for identifying potential drug-repurposing opportunities.

Keywords

Cancer genomics, integrative omics, machine learning, drug repurposing, pan cancer.

Background

Typically, a single tumor contains anywhere from tens to millions of genomic aberrations—including point mutations, small insertions, small deletions, and large copy number variations—that differ from the patient’s normal cells[1–4]. Knowledge of these aberrations may be useful in guiding therapeutic decisions. In some cases, a genomic aberration is the target of an existing therapy and thus may indicate that the therapy is a good match for that patient. For example, *Trastuzumab* is a targeted therapy for HER2-amplified breast cancers[5]. In other cases, a genomic aberration may be a biomarker for an existing therapy, even though the therapy was not explicitly designed to target that aberration[6]. Many such relationships have been identified for combinations of genomic aberration and therapy[7]. However, in many cases, tumors contain no therapeutic biomarker. Furthermore, few gene/drug associations have been made for the “long tail” of genomic aberrations that occur infrequently at the population level[8]. Although it may be economically infeasible to develop targeted therapies for every rare mutation, we may be able to repurpose existing cancer treatments by identifying similarities in tumor biology between tumors that harbor rare and common aberrations.

By disrupting signaling cascades—or pathways—within tumor cells, genomic aberrations can cause the tumor to grow, divide, or dedifferentiate in an uncontrolled manner[9]. Genomic aberrations within tumors are highly variable across cancer patients—each tumor carries a unique panoply of genomic aberrations. However, a much smaller number of signaling cascades is affected. Even though different genes are mutated in two different tumors, these mutations may affect common signaling cascades (e.g., Ras → Raf → MEK → ERK)[10]. We may be able to better understand the effects of genomic aberrations by considering such downstream effects.

Although it is possible to place genomic aberrations in the context of biological pathways, it may be difficult to decipher whether two aberrations have a similar effect on tumor biology, even though they occur within the same signaling cascade. This observation may be especially true for rare mutations, because little is known about the roles they play in tumorigenesis or therapeutic responses, and samples sizes are small. An alternative approach for understanding the effects of these mutations and their potential as biomarkers is to evaluate the transcriptional effects of the mutations. Using existing, high-throughput technologies (e.g., microarrays and RNA-Sequencing), it is possible to quantify gene-expression levels across the entire transcriptome for a modest cost. Two tumors may have similar gene-expression profiles, even though they have no genomic aberrations in common, ostensibly because the aberrations in either tumor have led to similar downstream effects. Therefore, the tumors may respond similarly to drug treatments. When this approach is applied to many tumors, it may be possible to identify transcriptional patterns that can be used as biomarkers of treatment response, independent of the genomic aberrations that occur within these tumors.

We evaluated this idea using publicly available data from The Cancer Genome Atlas (TCGA). We acquired data representing mutations (SNVs, insertions, or deletions), copy-number variations (large amplifications or deletions), and transcription for 9300 tumors across 25 cancer types available in TCGA. We identified tumors that carried mutations in frequently mutated genes (e.g., KRAS, EGFR, and ERBB2) and in genes that are mutated relatively rarely. We made the simplifying assumption that mutations at different genomic loci within a given gene exert a similar effect on tumor biology. We also assumed that mutations in individual genes have a characteristic effect on tumor transcription, despite the presence of additional mutations within each of these tumors.

Initially focusing on lung adenocarcinomas and then extending our analysis to other tumor types, we identified relatively common mutations and filtered the data to include only tumors where these mutations occurred in a mutually exclusive manner—harboring only one of the mutations of interest. Having classified the tumors by mutation status, we used a supervised, machine-learning algorithm to predict mutation status based on transcriptional patterns observed in the tumors. In many cases, the transcriptional patterns were highly predictive of mutation status, thus indicating that individual genomic aberrations influence transcription in distinct ways. Finally, using lower-frequency genes that had been excluded from the initial analysis, we identified genes that, when mutated, resulted in transcriptional patterns that were similar to some of the genes from our initial set. In several cases, these similarities coincided with prior knowledge about cancer pathways as well as with known therapeutic biomarkers.

Our findings show promise as a way to identify pairs of genes that, when mutated, may serve as biomarkers for the same treatment. These observations promise to be useful in guiding drug-repurposing efforts.

Methods

Somatic mutation data

In July 2016, we downloaded all available TCGA somatic mutation data via the National Cancer Institute's Genomic Data Commons[11]. These data had been generated using high-throughput, exome-sequencing technologies. Using the *MuTect* tool[12], these data had been compared against germline variation using human reference genome *GRCh38* to make somatic calls. Subsequently, the variants had been annotated with the *Variant Effect Predictor* (VEP) tool[13] to record population frequencies and to predict effects of the mutations on gene function. We used these annotations to filter the somatic variant data. We excluded variants that did not pass MuTect's quality-control criteria or that had a minor allele frequency greater than 0.01 in the ExAC database. Any mutation predicted by VEP to have a "LOW" or "MODIFIER" impact on protein function was excluded. We retained variants that had been predicted by SIFT[14] to be "deleterious" or "deleterious (low confidence)" or by Polyphen-2[15] to be "probably damaging" or "possibly damaging." Although some of these variants were likely false positives, we preferred to err on the side of inclusion rather than exclusion, to minimize the chances of false negatives. In addition, we retained variants that contained no predictions for SIFT or Polyphen-2.

Copy-number variation data

We downloaded copy-number variation (CNV) data that had been preprocessed and stored in the University of California Santa Cruz Xena database[16]. These data had been produced using whole-genome microarrays. The CNVs had been called using the GISTIC2.0 algorithm[17] and summarized to gene-level values. In addition, the gene-level values were thresholded to estimate whether each sample carried a homozygous deletion, a single-copy deletion, a low-level amplification, or a high-level amplification. We considered tumors with either a homozygous deletion or a high-level amplification to be “mutated.” We considered the remaining samples either to have been in a normal state or to have exerted only a modest effect, if any, on tumor transcription.

Finally, we aggregated the somatic mutation and CNV data to identify tumor samples that carried at least one genomic aberration in a given gene. We merged the somatic-mutation and CNV data to a single value per gene and tumor sample, using Boolean values to indicate whether the tumor harbored an aberration in a given gene.

RNA-sequencing data

We downloaded gene-level, RNA-Sequencing data that had been preprocessed and aligned using the Rsubread package[18, 19] and had been summarized using the transcripts per million (TPM) method. For tumor samples that had been sequenced multiple times, we averaged expression values across these samples. We log2-transformed the RNA-Sequencing values to mitigate the effects of extremely high expression values and to enable easier visualization.

Disease-drug-gene mapping data

We downloaded disease-drug-gene mappings that had been curated via a crowdsourcing effort and had been made available via the CIViC database[7]. We used the October 1, 2016 version of this database and focused solely on evidence where the relationship between gene and drug was "Supported."

Machine-learning analysis

Our analysis focused exclusively on TCGA samples for which data were available for all three molecular types (somatic mutation, CNV, and RNA-Sequencing). To reduce noise and computational complexity, we limited the data to 325 genes considered to play a role in canonical cancer pathways, according to the *Pathways in Cancer* diagram in the Kyoto Encyclopedia of Genes and Genomes (KEGG) [20].

Using the genome-aberration data, we categorized each tumor according to cancer type and whether an aberration had been identified in a given gene. We generated a "training set" by identifying samples that had at least one mutation in one of the frequently mutated genes (frequency threshold varied; see Results). We then excluded tumor samples that had a mutation in more than one of these genes. We used the resulting data set for a classification analysis, using genes as class labels. The remaining samples were set aside as a "test set." As a way to assess our therapeutic predictions, we limited genes in the test set to those that were described in the CIViC database.

Using the training set, we performed 5-fold cross validation to evaluate our ability to predict mutation status for each gene. Later, we trained a model on the entire training set and made predictions for the test set. In both cases, we used the Random Forests classification algorithm[21] with default parameters, other than that we requested probabilistic predictions.

Because mutation frequencies varied considerably across genes, we subsampled the data. For example, if the minimum number of mutated samples across all selected genes were 15, we would randomly select 15 samples for each gene. We repeated this process 10 times and averaged the results across the various subsampled results. This approach ensured that class imbalance would not bias our results.

Analysis pipeline

We wrote scripts in the *Python* programming language[22] to parse, filter, and summarize the input data. To enable easier analyses in subsequent steps, we restructured the data into "tidy data" format[23]. In performing the analysis steps and producing graphics, we used the *R* programming language[24]. These steps were aided by the following packages: *readr*, *dplyr*, *magrittr*, *ggplot2*, *RColorBrewer*, *randomForest*, *mlr*, *coin*, and *AUC*[25–33]. The entire pipeline executes in 5-15 minutes on a laptop computer with 4 cores and 16 GB RAM. We placed all analysis code and the tidy data in an open-access repository at https://osf.io/ndjkg.

To evaluate differences in expression for individual genes, we calculated p-values using Student's t-test and then performed a Bonferroni correction to account for all possible gene pairs (n = 52,650) in our data.

Results

We sought to identify genes whose transcriptional profiles were similar to each other when mutated in tumors. From TCGA, we obtained somatic-mutation data for 10,391 tumor samples. The filtering steps (see Methods) reduced the number of somatic mutations by 93.5%, mostly due to the removal of synonymous variants and common variants. Within our cancer-related genes of interest (n=325), we observed a total of 45,950 somatic mutations (5.57 per sample). We observed 52,012 high-level amplifications (8.63 per sample) and 19,037 large-scale, homozygous deletions (4.13 per sample) within these genes. After we removed samples that lacked data for at least one type of aberration, data for 9,300 patients across 25 distinct cancer types remained.

As might be expected, strong, tissue-specific patterns were apparent in the gene-expression data (Fig. 1A). In particular, expression levels for Acute Myeloid Leukemia (LAML) and Lower Grade Glioma (LGG) were clearly distinct from the other cancer types, likely in part because these cancers affect myeloid blood cells and glioma cells, respectively, whereas most other cancer types affect epithelial cells. Although in some contexts, these tissue-specific differences may provide valuable insights about how tumor biology differs among cell types, our preliminary analyses revealed that these differences caused a strongly confounding effect that would make it difficult to derive pan-cancer insights. Therefore, to correct for these effects, we applied the *ComBat* algorithm[34], using cancer type as batch. This adjustment reduced cancer-type effects (Fig. 1B). We used these corrected data in our analyses whenever we aggregated data across multiple cancer types.

Single-gene analysis

For the first analysis phase, we examined the relationship between genomic aberrations and expression levels of individual genes. In lung adenocarcinomas (n=515), the most frequently mutated genes were TP53 (n=233), CDKN2A (n=109), KRAS (n=95), CDKN2B (n=91), and EGFR (n=82). For decades, mutations in KRAS and EGFR have been recognized to play a critical role in lung adenocarcinoma and have been used to classify these tumors into subtypes based on mutation status[35]. Codons 12, 13, or 61 are typically mutated in KRAS [36]; similar mutations affect homologous genes HRAS and NRAS, though infrequently for lung adenocarcinomas. EGFR is often affected by point mutations and amplifications. Due to its role as a cell-surface receptor, EGFR is the target of antibody-based therapies, including *Gefitinib*[37], and vaccine-based immunotherapies[38]. We examined the relationship between mutations and expression levels of these genes. Tumors that harbored an EGFR aberration—a mutation or a copy-number variation—expressed EGFR at significantly higher levels than tumors that harbored a KRAS aberration but no EGFR aberration (Fig. 2A). In addition, tumors with a KRAS aberration expressed KRAS at significantly higher levels than samples with an EGFR aberration but no KRAS aberration (Fig. 2B). Although genomic aberrations often do not correlate strongly with gene-expression levels of the affected gene, we observed strong patterns for these genes. EGFR interacts with KRAS indirectly, via the Ras-Raf-MEK-ERK signaling cascade; thus EGFR aberrations may have led to suppression of KRAS activity in some samples. Indeed, mutations in these genes are often mutually exclusive, and KRAS mutations have been associated with a lack of sensitivity to EGFR-targeted therapies, such as *Gefitinib* and *Erlotinib*[39]. These observations suggest that, despite some likely overlap, the downstream effects of aberrations in these genes are distinct. In contrast, we examined differences in CDKN2A expression between samples that harbored either an EGFR or a KRAS aberration. No significant difference in expression was observed (Fig. 2C), as might be expected, given that CDKN2A is far downstream from these genes. We observed similar patterns after merging data across all 25 cancer types (Fig. 3).

Supervised-learning analysis

Although it is interesting to examine the effects of genomic aberrations on individual genes, most aberrations cause transcriptional responses across a broad range of genes expressed throughout tumor cells. To account for these broad effects, we used a supervised-learning approach (see Methods). Initially for lung adenocarcinomas, we identified genes that were mutated in at least 10 tumor samples. Then we excluded tumor samples that had a mutation in more than one of these genes. Finally, we retained genes that still had mutations in at least 10 tumor samples, resulting in a set of tumors with mutually exclusive mutations. We used cross validation to evaluate the ability of the Random Forests classification algorithm to predict gene-mutation status based on the gene-expression levels. We interpreted relatively high accuracy levels to mean that the downstream, transcriptional effects of mutations in these genes were distinct from each other. Fig. 4 illustrates predictions for the five genes selected. Across these genes, the average classification accuracy was 0.42—whereas 0.20 would be expected by chance in this scenario. For each gene, we calculated the area under the receiver operating characteristic curve (AUROC), an alternative measure of classification accuracy that accounts for the probabilistic nature of predictions. The highest AUROC was 0.81 for EGFR (0.50 expected by chance). The lowest AUROC was 0.58 for KRAS. These findings suggest that EGFR mutations have a particularly distinct effect upon transcription levels in lung adenocarcinomas. To get a sense for similarities and differences among genes, we calculated the Spearman correlation coefficient for probabilistic predictions between each pair of genes. We observed no strong positive correlations for lung adenocarcinoma, providing evidence that the transcriptional effects of mutations in KRAS, EGFR, MET, NTRK1, and PIK3CA are distinct from each other (Fig. 5). We saw a modest negative correlation (rho = -0.18) between EGFR and KRAS—consistent with our single-gene analysis—and a strong negative correlation (rho = -0.47) between EGFR and MET. Both EGFR and MET are from the receptor tyrosine kinase family, encode for cell-membrane receptors, and act as oncogenes when mutated. Although crosstalk mediated by microRNAs may occur between these proteins [40], the negative correlation we observed may indicate that these proteins cause very different downstream effects and thus should be addressed differently from a therapeutic standpoint. Indeed, it has been shown that MET amplification causes resistance to *Gefinitib*, an EGFR inhibitor[41].

We applied the same methodology to breast tumors. Seven genes passed the threshold for inclusion in the training set, and the average accuracy was again 0.42 (0.14 expected by chance). The genes with the highest individual AUROC were BIRC5 and TP53. Both play a role in tumor cells' ability to evade apoptosis, but our results suggest that mutations in these genes operate via distinct mechanisms. The strongest positive correlation (rho = 0.49; see Figure S1) was between PTEN and PIK3CA, two genes known to play important roles in breast tumorigenesis (Fig. 6). PIK3CA is an oncogene, activated by receptor tyrosine kinases and plays an important role in cellular proliferation via the PKB/Akt signaling cascade. PTEN tightly regulates PIK3CA; thus, PKB/Akt activity is expected to increase when PTEN loss (via mutation or deletion) has occurred[42]. Our results coincide with the expectation that the downstream effects of genomic aberrations in these genes are similar. Indeed, aberrations in these genes have been shown to correlate similarly with prognostic factors[43].

We repeated this process for other cancer types—bladder carcinoma, head/neck squamous carcinoma, ovarian cystadenocarcinoma, metastatic skin melanoma, stomach adenocarcinoma—that had enough data; we required that at least three genes be mutated at least ten times in a mutually exclusive manner. *Additional files 1-8* indicate correlation coefficients, nominal p-values, and Bonferroni-adjusted p-values for each pairwise comparison. The strongest negative correlation (rho = 0.83; see Figure S2) was between FGFR1 and FGFR3 in bladder carcinoma. Experimental work has demonstrated that both genes are activated via mutations and the genes play distinct roles in regulating bladder-tumor growth[44]. The genes that showed the most significant differences in expression between FGFR1- and FGFR3-mutated tumors were SMAD3, FGFR3, FN1, LAMA1, and FGFR1 (Figures S3-S7). In addition to growth signaling, these genes play roles in regulating extracellular matrix adhesion and intracellular signaling.

The correlations we observed between pairs of mutated genes were often tissue specific; for example, a strong relationship between PTEN and PIK3CA was observed in breast carcinomas but not in head/neck squamous carcinomas, stomach adenocarcinomas, or ovarian cystadenocarcinomas.

Next we extended the same methodology to all 25 cancer types in a pan-cancer approach. We increased the threshold to 50 for the minimum number of mutually exclusive mutations per gene. Nine genes passed this threshold (see Fig. 7). Perhaps due to the larger sample sizes per gene, accuracy increased to 0.50 (0.11 expected by chance). The AUROC was extremely high for BRAF (0.97) and VHL (0.93), suggesting that the transcriptional effects of aberrations in these genes are particularly distinct. However, mutation status was predicted accurately for all nine genes, with a minimum AUROC of 0.73 (APC gene). Several interesting relationships between genes were apparent, including a strong correlation (rho = 0.44) between TP53 and RB1, which play critical roles in regulation of the cell cycle and DNA repair[45, 46]. FGFR1 was strongly correlated (0.42) with PIK3CA, likely due to FGFR1's role as an upstream activator of PIK3CA. Perhaps surprisingly, FGFR1 predictions were also correlated strongly with RB1 predictions; however, it has been shown recently that CDK6, a key regulator of RB1, is regulated by MIR9, a microRNA that also regulates FGFR1 activity[47].

For the last stage of our analysis, we trained the Random Forests classification algorithm on the full training set and made predictions for tumor samples in the test set. We interpreted that probabilistic predictions coinciding with the actual mutation status of genes in the test set would indicate that the expression profiles of genes in the training set were similar to the expression profiles of genes from the test set. We hoped this analysis would provide insights about genes that are mutated rarely. Using a minimum threshold of five mutated samples (test set) and focusing initially on breast cancer, we found that MMP2 predictions were strongly associated with predictions for PIK3CA (0.47), PTEN (0.49), and AKT1 (0.28), which operate via the same pathway. Like MMP2, PIK3CA and AKT1 play important roles in cancer cell migration and metastasis[48–50].

We also observed a strong positive correlation (0.33) between TP53 and AKT2 and a modest correlation between ERBB2 and AKT2 (0.20). AKT2 mediates TP53 activity via MDM2[51]. *Trastuzumab* was originally developed as a targeted therapy for ERBB2 (Her2) amplification; more recently, aberrant AKT2 expression has been associated with longer time to progression and overall patient survival in Her2-positive patients[52]. We also observed a slightly negative correlation (-0.20) between AKT1 and AKT2 predictions; although the proteins encoded by these genes operate within the same signaling cascade, their roles in regulating cell migration and differentiation are distinct[48].

When we made test-set predictions for all 25 cancer types, we noted a few modestly positive correlations. The first was between VHL and MTOR1. VHL inactivation leads to constitutive activation of HIF-2 and/or HIF-1. In clear-cell renal carcinomas, the downstream effects of HIF activation are inhibition of *mTor Complex 1*[53]. The second positive correlation was between NRAS and BRAF. These genes interact directly with each other and operate via the Ras → Raf → MEK → ERK cascade. Indeed, the CIViC database indicates that several antibody-based treatments—including *Cetuximab*, *Selumetinib*, and *Vemurafenib*—target tumors with mutations in either of these genes.

Although we used the ComBat software to correct for tissue-type effects and used a principal component analysis to visually verify the effects of this correction (Figure 1), we conducted a follow-up evaluation to assess whether tissue specificity might still have influenced our results, because, in many cases, mutations occur in a tissue-specific manner[54]. Initially, we used the

Random Forest algorithm to predict tumor type based on the expression data that had not been adjusted using ComBat. Then we repeated this process using the ComBat-adjusted data. Using this approach, we could predict tumor type with >90% accuracy for both versions of the expression data. Even though ComBat had adjusted for tissue specificity, a subtle footprint remained, which the Random Forest algorithm was able to detect. In a second follow-up evaluation, we used the Random Forest algorithm to predict gene-mutation status using only tumor type (instead of gene-expression data). Although tumor type predicted mutation status less accurately than the gene-expression data (47% and 50%, respectively), these results confirm that tumor type is confounded with mutation status and thus that our pan-cancer results—and results from other pan-cancer studies that examine the relationship between genomic and transcriptomic variation—should be interpreted with caution. However, it is difficult to distinguish between correlation and causation; certain tissue environments (driven by gene expression) may select for somatic variants in certain genes, and/or certain somatic variants may strongly influence the tissue specificity of cancer.

Discussion

We have developed a computational approach that uses publicly available, molecular-profiling data to identify genes that have similar (or different) effects on gene expression in human tumors. Our overarching goal was to develop a methodology that can be used to guide drug-repurposing efforts and, more generally, to help cancer researchers make sense of the vast complexity and heterogeneity of tumors. Using data from The Cancer Genome Atlas, we observed relationships that recapitulate what was previously known about canonical cancer pathways and treatment responses. Alternative methods have primarily evaluated molecular data in a low-throughput manner, have examined one type of molecule at a time, or have considered the expression of individual genes associated with mutations; in contrast, our method accounts for broad-ranging effects of genomic aberrations on gene expression within cells. Using a supervised-learning approach, we found that we could predict mutation status, often with high accuracy. This provides evidence that many mutations confer a clear and distinct effect on transcriptional responses within downstream genes.

To increase interpretability, we made simplifying assumptions and simplified our approach in several ways. We limited our analysis to genes known to play a role in cancer, as described in KEGG's *Pathways in Cancer* diagram. We assumed that the effects of genomic aberrations are mutually exclusive. Even though this assumption may not hold in every case, it reflects current approaches that are used to prioritize targeted cancer therapies. For example, even though it is clearly understood that tumors with mutations in EGFR harbor many variants in genes other than EGFR, mutations in EGFR are used as a biomarker for treating lung adenocarcinomas with therapies such as *Gefitinib* and *Erlotinib*. In addition, evidence suggests that EGFR and KRAS mutations occur in a mutually exclusive manner and that tumors with KRAS mutations fail to respond to these drugs[39]. Such findings suggest that mutations in individual genes can strongly influence treatment responses, despite a background of other mutations. To an extent, our goal was to identify such scenarios.

In addition, we ignored the potential impact of epigenomic factors, such as DNA methylation and miRNA expression, as well as gene fusions. Our approach could be refined in future studies to use such observations to indicate whether a given gene is "mutated." Although genomic aberrations are often thought to be the main drivers of tumorigenesis, epigenomic factors often play a critical role in modulating tumor activity and/or interacting with genomic aberrations. For example, DNA hypermethylation of promoter regions can cause transcriptional silencing of tumor-suppressor genes; BRCA1 hypermethylation has been shown to alter responses to platinum-salt therapies[55]. miRNAs can also play important roles in regulating tumor transcription[56]; in chronic lymphocytic leukemia, two miRNAs on human chromosome 13q14 occur frequently [57] and likely play important roles in regulating tumorigenesis and treatment responses.

We treated all mutations equally within a given gene, regardless of genomic loci or mutation type; but in many cases, there may be considerable heterogeneity across genomic loci and mutation types. As sample sizes increase over time, it will be more feasible to sub-classify mutations in finer detail.

We focused primarily on the genome- and cancer-related aspects of this work rather than on algorithmic aspects. In future work, it would be valuable to compare (and perhaps combine) multiple algorithms as a way to optimize our approach. We used the Random Forests algorithm because it has been shown to deal effectively with high-throughput transcriptomic data, executes quickly, and is more amenable to *post hoc* interpretation than many other algorithms[58].

Finally, although we have demonstrated a potential to learn about the effects of rare variants, our analysis touched only briefly on such variants. To evaluate the reliability of our method, we focused on genes that had been affected by at least a modest number of mutations. However, we believe this methodology can be applied in cases where a given mutation has been observed in only a single tumor, potentially providing insights for “n-of-1” clinical trials.

Conclusions

We have used supervised-machine learning to integrate genomic and transcriptomic data across 9300 tumors and 25 cancer types to aid in deciphering the downstream effects of genomic aberrations on tumor transcription. This approach has potential to guide development of treatment biomarkers and to understand similarities and differences among genes that play a role in specific types of cancer and across multiple cancer types. We hope this approach will be useful to pharmacologists and clinical trialists who seek to identify relationships between genomic aberrations and treatment responses. In particular, we hope this methodology will reduce barriers for drug-repurposing efforts so that existing treatments can be used on tumors with no current treatment biomarker. We believe such approaches will reduce the costs of developing new cancer drugs and increase the number of tumors that can be treated in a targeted manner.

Competing interests

The authors declare that they have no competing interests.

Ethics approval and consent to participate

The Brigham Young University Institutional Review Board granted an exemption for this study (#E14522). This study uses only publicly available data.

Declarations

Availability of data and material

All tidy-data files, analysis code, and analysis outputs (including some figures not included in this paper) are publicly accessible at https://osf.io/ndjkg. We encourage others to extend and refine our methods.

Funding

S.R.P. acknowledges startup funds from Brigham Young University. J.B.D. expresses gratitude for a student mentoring grant from Brigham Young University. Publication costs for this article were funded by Brigham Young University.

Authors' contributions

SRP conceived and designed the study. Both authors developed the analytical pipeline. JBD parsed, filtered, and tidied the data. SRP performed the machine-learning analysis. JBD prepared and organized the code and data. Both authors wrote the paper and prepared figures for the paper. Both authors read and approved the final manuscript.

Acknowledgements

We thank T. James Lee for providing insights on processing data from TCGA. We thank the Fulton Supercomputing Lab at Brigham Young University for providing computational resources that enabled us to complete this study. We thank the patients who donated tumor samples to TCGA Consortium and consented to share the resulting data with the research community.

References

1. Jones S, Zhang X, Parsons DW, Lin JC-H, Leary RJ, Angenendt P, Mankoo P, Carter H, Kamiyama H, Jimeno A, Hong S-M, Fu B, Lin M-T, Calhoun ES, Kamiyama M, Walter K, Nikolskaya T, Nikolsky Y, Hartigan J, Smith DR, Hidalgo M, Leach SD, Klein AP, Jaffee EM, Goggins M, Maitra A, Iacobuzio-Donahue C, Eshleman JR, Kern SE, Hruban RH, et al.: **Core signaling pathways in human pancreatic cancers revealed by global genomic analyses.** *Science* 2008, **321**:1801–6.

2. Parsons WW, Lin JC-HC, Jones S, Siu I-MM, Zhang X, Rasheed AA, Marie SKNK, Leary RJ, Shinjo SMOM, Angenendt P, Mankoo P, Carter H, Gallia GL, Olivi A, McLendon R, Keir S, Nikolskaya T, Nikolsky Y, Busam DA, Tekleab H, Diaz LA, Hartigan J, Smith DR, Strausberg RL, Yan H, Riggins GJ, Bigner DD, Karchin R, Papadopoulos N, Parmigiani G, et al.: **An**

integrated genomic analysis of human glioblastoma multiforme. *Science* 2008, **321**:1807–1812.

3. Network TCGAR: **Comprehensive molecular profiling of lung adenocarcinoma.** *Nature* 2014, **511**:543–50.

4. The Cancer Genome Atlas Research Network: **Comprehensive genomic characterization of squamous cell lung cancers.** *Nature* 2012, **489**:519–25.

5. Molina MA, Codony-Servat J, Albanell J, Rojo F, Arribas J, Baselga J: **Trastuzumab (Herceptin), a humanized anti-HER2 receptor monoclonal antibody, inhibits basal and activated HER2 ectodomain cleavage in breast cancer cells**. *Cancer Res* 2001, **61**:4744–4749.

6. Garnett MJ, Edelman EJ, Heidorn SJ, Greenman CD, Dastur A, Lau KW, Greninger P, Thompson IR, Luo X, Soares J, Liu Q, Iorio F, Surdez D, Chen L, Milano RJ, Bignell GR, Tam AT, Davies H, Stevenson J a., Barthorpe S, Lutz SR, Kogera F, Lawrence K, McLaren-Douglas A, Mitropoulos X, Mironenko T, Thi H, Richardson L, Zhou W, Jewitt F, et al.: **Systematic identification of genomic markers of drug sensitivity in cancer cells**. *Nature* 2012, **483**:570–575.

7. Griffith M, Spies NC, Krysiak K, Coffman AC, McMichael JF, Ainscough BJ, Rieke DT, Danos AM, Kujan L, Ramirez CA, Wagner AH, Skidmore ZL, Liu CJ, Jones MR, Bilski RL, Lesurf R, Barnell EK, Shah NM, Bonakdar M, Trani L, Matlock M, Ramu A, Campbell KM, Spies GC, Graubert AP, Gangavarapu K, Eldred JM, Larson DE, Walker JR, Good BM, et al.: *CIViC: A Knowledgebase for Expert-Crowdsourcing the Clinical Interpretation of Variants in Cancer*. 2016.

8. Garraway LA, Lander ES: **Lessons from the cancer genome**. *Cell* 2013, **153**:17–37.

9. Hanahan D, Weinberg RAA: **Hallmarks of cancer: the next generation.** *Cell* 2011,

144:646–74.

10. Kolch W: **Meaningful relationships: the regulation of the Ras/Raf/MEK/ERK pathway by protein interactions.** *Biochem J* 2000, **351 Pt 2**:289–305.

11. **Genomic Data Commons** [https://gdc.cancer.gov]

12. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, Getz G: **Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples**. *Nat Biotechnol* 2013, **31**:213–219.

13. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, Cunningham F, Eisenstein M, Weil M, Chen A, Visscher P, Brown M, McCarthy M, Yang J, Pierre A Saint, Génin E, Zuk O, Schaffner S, Samocha K, Do R, Hechter E, Kathiresan S, Hindorff L, Sethupathy P, Junkins H, Ramos E, Mehta J, Collins F, Puente X, et al.: **The Ensembl Variant Effect Predictor**. *Genome Biol* 2016, **17**:122.

14. Kumar P, Henikoff S, Ng PC: **Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm.** *Nat Protoc* 2009, **4**:1073–81.

15. Adzhubei I, Jordan DM, Sunyaev SR: **Predicting functional effect of human missense mutations using PolyPhen-2.** *Curr Protoc Hum Genet* 2013, **Chapter 7**:Unit7.20.

16. Goldman M, Craft B, Swatloski T, Cline M, Morozova O, Diekhans M, Haussler D, Zhu J: **The UCSC Cancer Genomics Browser: update 2015**. *Nucleic Acids Res* 2015, **43**:D812–D817.

17. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G: **GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers**. *Genome Biol* 2011, **12**:R41.

18. Rahman M, Jackson LK, Johnson WE, Li DY, Bild AH, Piccolo SR: **Alternative**

preprocessing of RNA-Sequencing data in the Cancer Genome Atlas leads to improved analysis results. *Bioinformatics* 2015, **31**:3666–3672.

19. Liao Y, Smyth GK, Shi W: **FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features**. *Bioinformatics* 2014, **30**:923–930.

20. Tanabe M, Kanehisa M: **Using the KEGG database resource**. *Curr Protoc Bioinforma* 2012(SUPPL.38).

21. Breiman L: **Random Forests**. *Mach Learn* 2001, **45**:5–32.

22. Python Software Foundation: **Python Language Reference, version 2.7**. *Python Software Foundation* 2013.

23. Wickham H: **Tidy Data**. *J Stat Softw* 2014, **59**.

24. R Core Team: **R: A Language and Environment for Statistical Computing**. 2016.

25. Wickham H, Hester J, Francois R: **readr: Read Tabular Data**. 2016.

26. Wickham H, Francois R: **dplyr: A Grammar of Data Manipulation**. 2016.

27. Bache SM, Wickham H: **magrittr: A Forward-Pipe Operator for R**. 2014.

28. Wickham H: *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York; 2009.

29. Neuwirth E: **RColorBrewer: ColorBrewer Palettes**. 2014.

30. Bischl B, Lang M, Kotthoff L, Schiffner J, Richter J, Jones Z, Casalicchio G: **mlr: Machine Learning in R**. 2016.

31. Liaw A, Wiener M: **Classification and Regression by randomForest**. *R News* 2002, **2**:18–22.

32. Ballings M, Van den Poel D: **AUC: Threshold independent performance measures for probabilistic classifiers.** 2013.

33. Hothorn T, Hornik K, van de Wiel M a, Zeileis A: **A Lego System for Conditional Inference**. *Am Stat* 2006, **60**:257–263.

34. Johnson WE, Li C, Rabinovic A: **Adjusting batch effects in microarray expression data using empirical Bayes methods.** *Biostatistics* 2007, **8**:118–27.

35. Vargas AJ, Harris CC: **Biomarker development in the precision medicine era: lung cancer as a case study**. *Nat Rev Cancer* 2016, **16**:525–537.

36. Rodenhuis S, van de Wetering ML, Mooi WJ, Evers SG, van Zandwijk N, Bos JL: **Mutational activation of the K-ras oncogene. A possible pathogenetic factor in adenocarcinoma of the lung.** *N Engl J Med* 1987, **317**:929–35.

37. Paez JG, Jänne PA, Lee JC, Tracy S, Greulich H, Gabriel S, Herman P, Kaye FJ, Lindeman N, Boggon TJ, Naoki K, Sasaki H, Fujii Y, Eck MJ, Sellers WR, Johnson BE, Meyerson M: **EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy.** *Science* 2004, **304**:1497–500.

38. Sampson JH, Archer GE, Mitchell DA, Heimberger AB, Bigner DD: **Tumor-specific immunotherapy targeting the EGFRvIII mutation in patients with malignant glioma.** *Semin Immunol* 2008, **20**:267–75.

39. Pao W, Wang TY, Riely GJ, Miller V a, Pan Q, Ladanyi M, Zakowski MF, Heelan RT, Kris MG, Varmus HE: **KRAS mutations and primary resistance of lung adenocarcinomas to gefitinib or erlotinib.** *PLoS Med* 2005, **2**:e17.

40. Acunzo M, Romano G, Palmieri D, Laganá A, Garofalo M, Balatti V, Drusco A, Chiariello M, Nana-Sinkam P, Croce CM: **Cross-talk between MET and EGFR in non-small cell lung cancer involves miR-27a and Sprouty2.** *Proc Natl Acad Sci U S A* 2013, **110**:8573–8.

41. Engelman J a, Zejnullahu K, Mitsudomi T, Song Y, Hyland C, Park JO, Lindeman N, Gale

CM, Zhao X, Christensen J, Kosaka T, Holmes a J, Rogers a M, Cappuzzo F, Mok T, Lee C, Johnson BE, Cantley LC, Janne P a: **MET amplification leads to gefitinib resistance in lung cancer by activating ERBB3 signaling**. *Science (80-)* 2007, **316**:1039–1043.

42. Karakas B, Bachman KE, Park BH: **Mutation of the PIK3CA oncogene in human cancers**. *Br J Cancer* 2006, **94**:455–459.

43. Pérez-Tenorio G, Alkhori L, Olsson B, Waltersson MA, Nordenskjöld B, Rutqvist LE, Skoog L, Stål O: **PIK3CA mutations and PTEN loss correlate with similar prognostic factors and are not mutually exclusive in breast cancer**. *Clin Cancer Res* 2007, **13**:3577–3584.

44. Cheng T, Roth B, Choi W, Black PC, Dinney C, McConkey DJ: **Fibroblast Growth Factor Receptors-1 and -3 Play Distinct Roles in the Regulation of Bladder Cancer Growth and Metastasis: Implications for Therapeutic Targeting**. *PLoS One* 2013, **8**:e57284.

45. Smith ML, Seo YR: **p53 regulation of DNA excision repair pathways.** *Mutagenesis* 2002, **17**:149–156.

46. Cook R, Zoumpoulidou G, Luczynski MT, Rieger S, Moquet J, Spanswick VJ, Hartley JA, Rothkamm K, Huang PH, Mittnacht S: **Direct involvement of retinoblastoma family proteins in DNA repair by non-homologous end-joining**. *Cell Rep* 2015, **10**:2007–2019.

47. Rodriguez-Otero P, Román-Gómez J, Vilas-Zornoza A, José-Eneriz ES, Martín-Palanco V, Rifón J, Torres A, Calasanz MJ, Agirre X, Prosper F: **Deregulation of FGFR1 and CDK6 oncogenic pathways in acute lymphoblastic leukaemia harbouring epigenetic modifications of the MIR9 family**. *Br J Haematol* 2011, **155**:73–83.

48. Irie HY, Pearline R V., Grueneberg D, Hsia M, Ravichandran P, Kothari N, Natesan S, Brugge JS: **Distinct roles of Akt1 and Akt2 in regulating cell migration and epithelial-mesenchymal transition**. *J Cell Biol* 2005, **171**:1023–1034.

49. Weber GL, Parat MO, Binder ZA, Gallia GL, Riggins GJ: **Abrogation of PIK3CA or PIK3R1 reduces proliferation, migration, and invasion in glioblastoma multiforme cells**. *Oncotarget* 2011, **2**:833–849.

50. Mendes O, Kim HT, Stoica G: **Expression of MMP2, MMP9 and MMP3 in breast cancer brain metastasis in a rat model**. *Clin Exp Metastasis* 2005, **22**:237–246.

51. Abraham AG, O'Neill E: **PI3K/Akt-mediated regulation of p53 in cancer.** *Biochem Soc Trans* 2014, **42**:798–803.

52. Grell P, Fabian P, Khoylou M, Radova L, Slaby O, Hrstka R, Vyzula R, Hajduch M, Svoboda M: **Akt expression and compartmentalization in prediction of clinical outcome in HER2-positive metastatic breast cancer patients treated with trastuzumab.** *Int J Oncol* 2012, **41**:1204–12.

53. Kucejova B, Peña-Llopis S, Yamasaki T, Sivanand S, Tran TAT, Alexander S, Wolff NC, Lotan Y, Xie X-J, Kabbani W, Kapur P, Brugarolas J: **Interplay Between pVHL and mTORC1 Pathways in Clear-Cell Renal Cell Carcinoma.** *Mol Cancer Res* 2011, **9**:1255–1265.

54. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin A V, Bignell GR, Bolli N, Borg A, Børresen-Dale A-L, Boyault S, Burkhardt B, Butler AP, Caldas C, Davies HR, Desmedt C, Eils R, Eyfjörd JE, Foekens JA, Greaves M, Hosoda F, Hutter B, Ilicic T, Imbeaud S, Imielinski M, Imielinsk M, Jäger N, Jones DTW, Jones D, Knappskog S, et al.: **Signatures of mutational processes in human cancer.** *Nature* 2013, **500**:415–21.

55. Stefansson O a, Villanueva A, Vidal A, Martí L, Esteller M: **BRCA1 epigenetic inactivation predicts sensitivity to platinum-based chemotherapy in breast and ovarian cancer.** *Epigenetics* 2012, **7**:1225–9.

56. Lujambio A, Lowe SW: **The microcosmos of cancer.** *Nature* 2012, **482**:347–55.

57. Calin GA, Dumitru CD, Shimizu M, Bichi R, Zupo S, Noch E, Aldler H, Rattan S, Keating M, Rai K, Rassenti L, Kipps T, Negrini M, Bullrich F, Croce CM: **Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia.** *Proc Natl Acad Sci U S A* 2002, **99**:15524–9.

58. Díaz-Uriarte R, Alvarez de Andrés S: **Gene selection and classification of microarray data using random forest.** *BMC Bioinformatics* 2006, **7**:3.

Figure legends

Figure 1

Principal component plots illustrating the relationship between cancer type and gene-expression levels. Values for the first two principal components are shown. Each point represents data across all genes for a given tumor sample. The colors and legend indicate the cancer type of each tumor sample. A) Distinct, tissue-specific patterns were present in the original data. B) After a regression-based adjustment for cell-type effects, these patterns were no longer prominent in the data. Cancer type abbreviations are listed on the TCGA web site.

Figure 2

Associations between genomic aberration status and expression levels of individual genes in lung adenocarcinoma samples. A) EGFR expression levels for tumors with an aberration in EGFR, KRAS, or neither of these genes. B) KRAS expression levels for the same tumors. C) CDKN2A expression levels for the same tumors. Expression levels are log-transformed, transcripts-per-million values. P-values were calculated using Welch's t-test and adjusted using a Bonferroni correction.

Figure 3

Associations between genomic aberration status and expression levels of individual genes across 25 cancer types. A) EGFR expression levels for tumors with either an aberration in EGFR, KRAS, or neither of these genes. B) KRAS expression levels for the same tumors. C) CDKN2A expression levels for the same tumors. Expression levels are log-transformed, transcripts-per-million values. P-values were calculated using Welch's t-test and adjusted using a Bonferroni correction.

Figure 4

Random Forest predictions, trained on gene-expression data, of gene-mutation status for lung adenocarcinoma. We identified genes that had been mutated in at least ten tumor samples and that contained no mutation in other genes that had been mutated in at least ten samples. Then using cross validation, we evaluated how well the Random Forest classification algorithm could identify which gene was mutated in a given sample. The algorithm produced a probabilistic prediction for each gene (class), and we evaluated these predictions using the area under the receiver operating characteristic curve (AUROC). The x-axis labels indicate AUROC values for each gene. Relatively high AUROC values indicate that gene-expression levels are highly predictive of gene-mutation status and thus suggest that mutations in these genes exert a characteristic effect on gene-expression levels.

Figure 5

Heatmap showing Spearman correlation coefficients between gene pairs for lung adenocarcinomas. The data values and colors represent correlation coefficients between probabilistic predictions for pairs of genes that exhibited mutually exclusive mutations. Warmer colors represent higher levels of correlation, while cooler colors represent lower levels of correlation.

Figure 6

Heatmap showing Spearman correlation coefficients between gene pairs for breast carcinomas. The data values and colors represent correlation coefficients between probabilistic predictions for pairs of genes that exhibited mutually exclusive mutations. Warmer colors represent higher levels of correlation, while cooler colors represent lower levels of correlation.

Figure 7

Random Forest predictions of gene-mutation status across 25 cancer types. We identified genes that had been mutated in at least 50 tumor samples and that contained no mutation in other genes that had been mutated in at least 50 samples. Then using cross validation, we evaluated how well the Random Forest classification algorithm could identify which gene was mutated in a given sample. The algorithm produced a probabilistic prediction for each gene (class), and we evaluated these predictions using the area under the receiver operating characteristic curve (AUROC). The x-axis labels indicate AUROC values for each gene. Relatively high AUROC values indicate that gene-expression levels are highly predictive of gene-mutation status and thus suggest that mutations in these genes exert a characteristic effect on gene-expression levels.

Additional files

Additional file 1

Multi-class Correlations – BLCA (.tsv). Correlation coefficients and measures of significance for each gene pair in our analysis of bladder carcinoma.

Additional file 2

Multi-class Correlations – BRCA (.tsv). Correlation coefficients and measures of significance for each gene pair in our analysis of breast carcinoma.

Additional file 3

Multi-class Correlations – HNSC (.tsv). Correlation coefficients and measures of significance for each gene pair in our analysis of head/neck squamous carcinoma.

Additional file 4

Multi-class Correlations – LUAD (.tsv). Correlation coefficients and measures of significance for each gene pair in our analysis of lung adenocarcinoma.

Additional file 5

Multi-class Correlations – OV (.tsv). Correlation coefficients and measures of significance for each gene pair in our analysis of ovarian cystadenocarcinoma.

Additional file 6

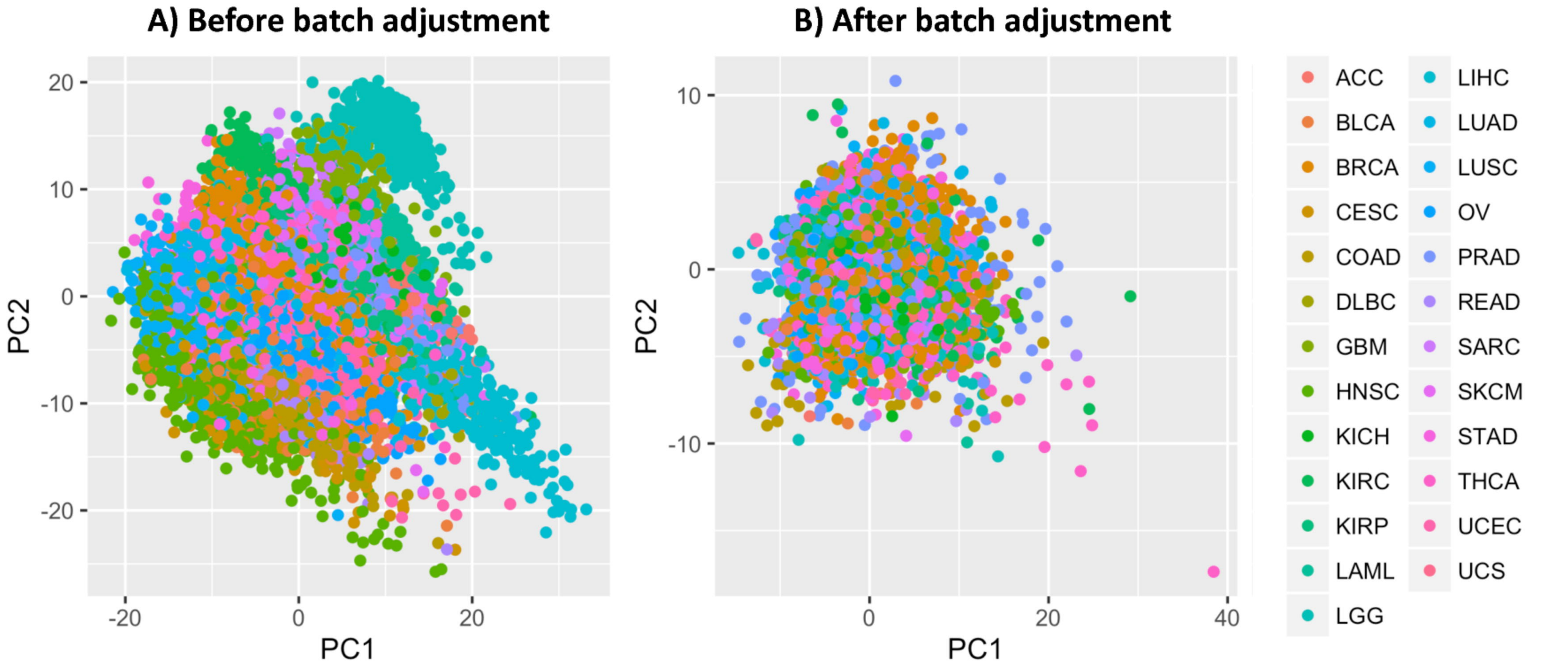
Multi-class Correlations – SKCM (.tsv). Correlation coefficients and measures of significance for each gene pair in our analysis of metastatic skin melanoma.

Additional file 7

Multi-class Correlations – STAD (.tsv). Correlation coefficients and measures of significance for each gene pair in our analysis of stomach adenocarcinoma.

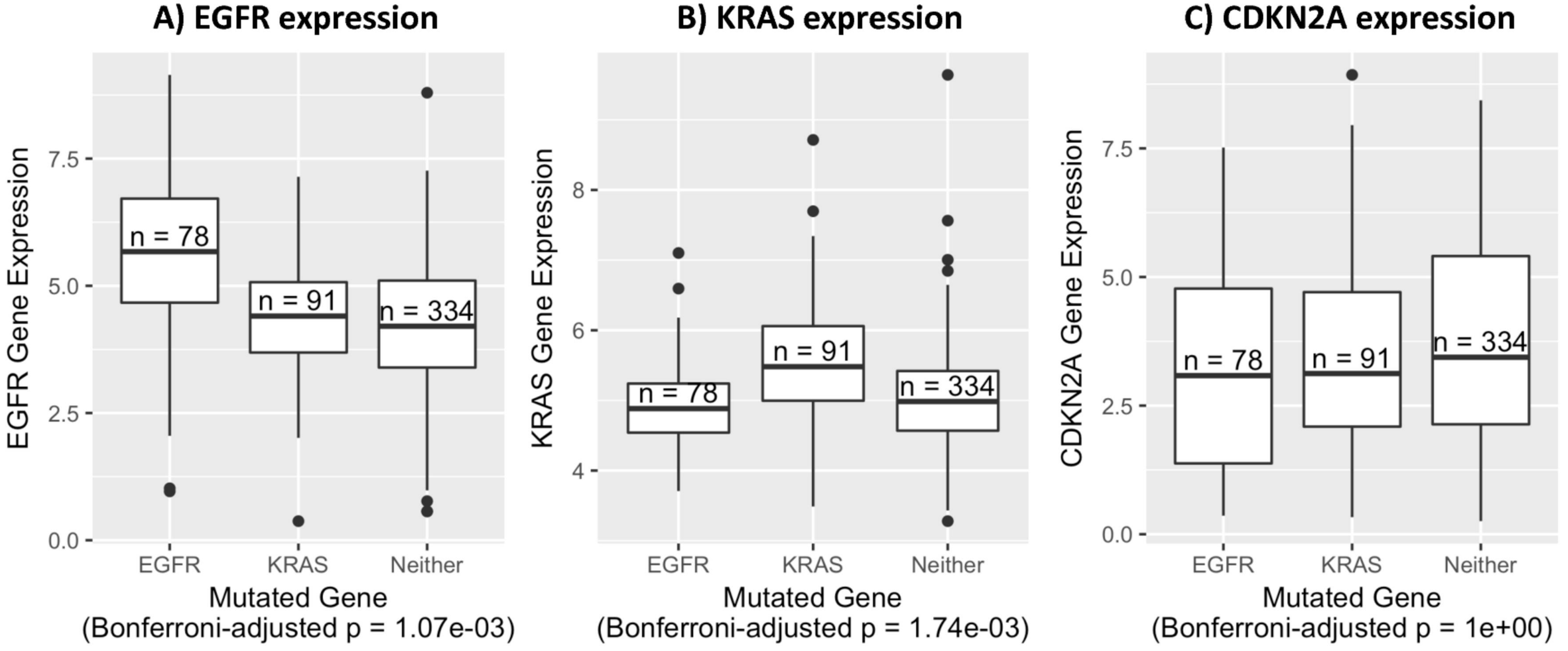
Additional file 8

Multi-class Correlations – All Cancer Types (.tsv). Correlation coefficients and measures of significance for each gene pair in our pan-cancer analysis.



A) Before batch adjustment

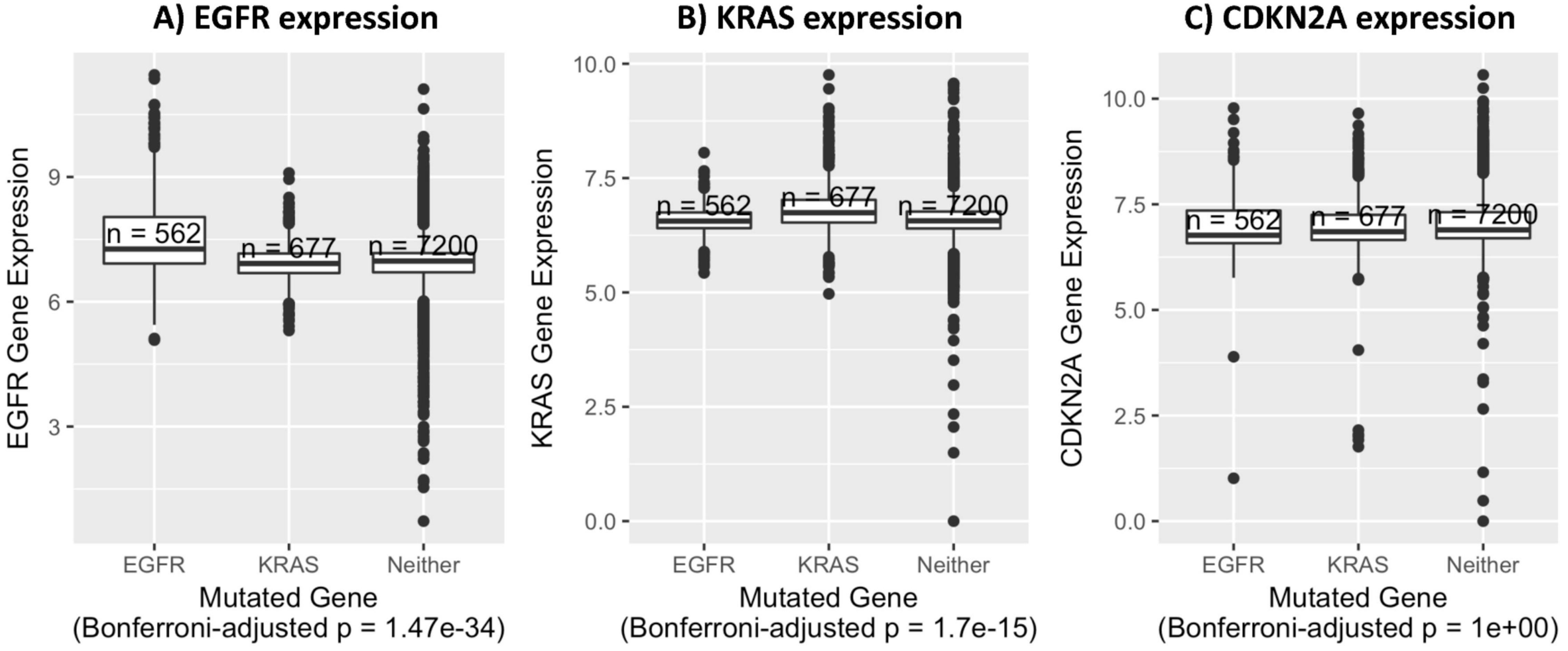
B) After batch adjustment



A) EGFR expression

B) KRAS expression

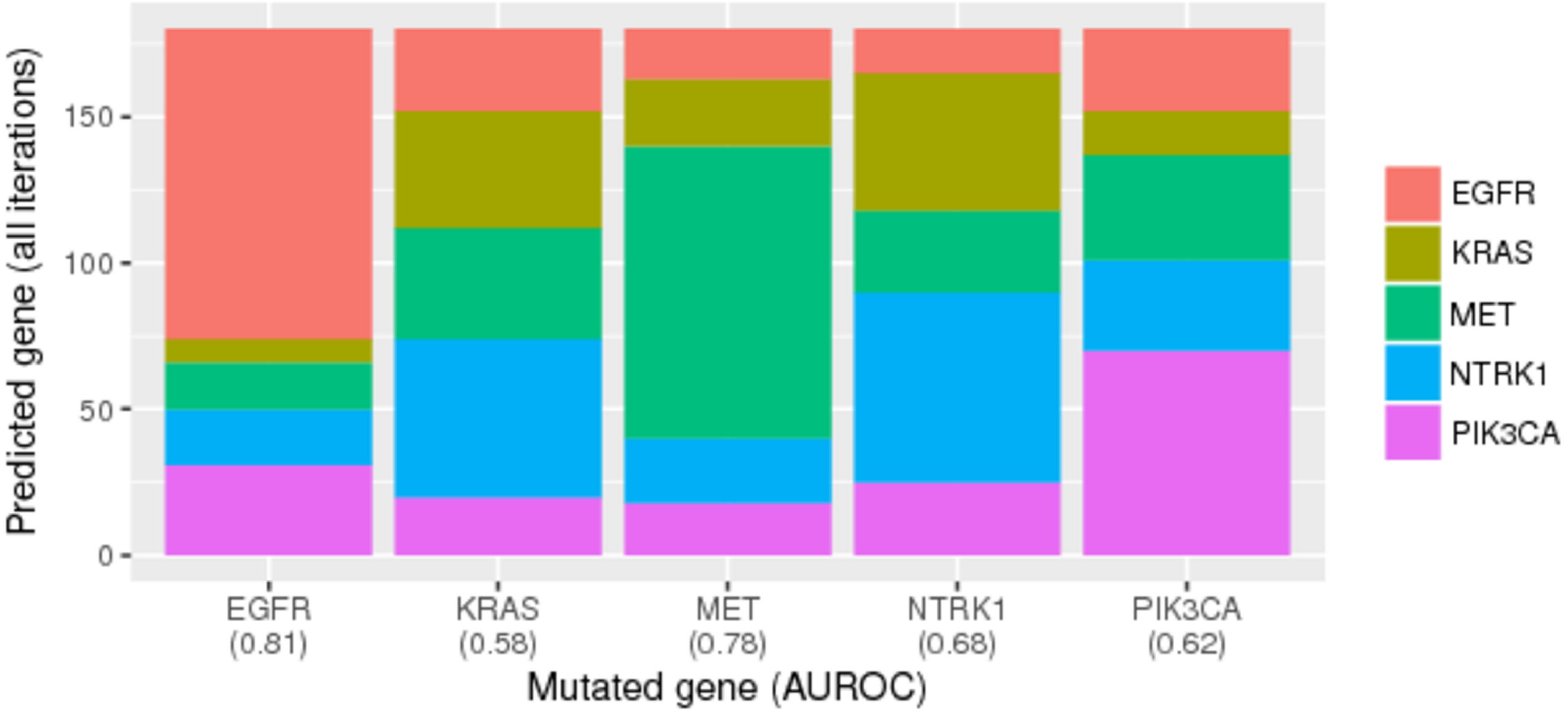
C) CDKN2A expression

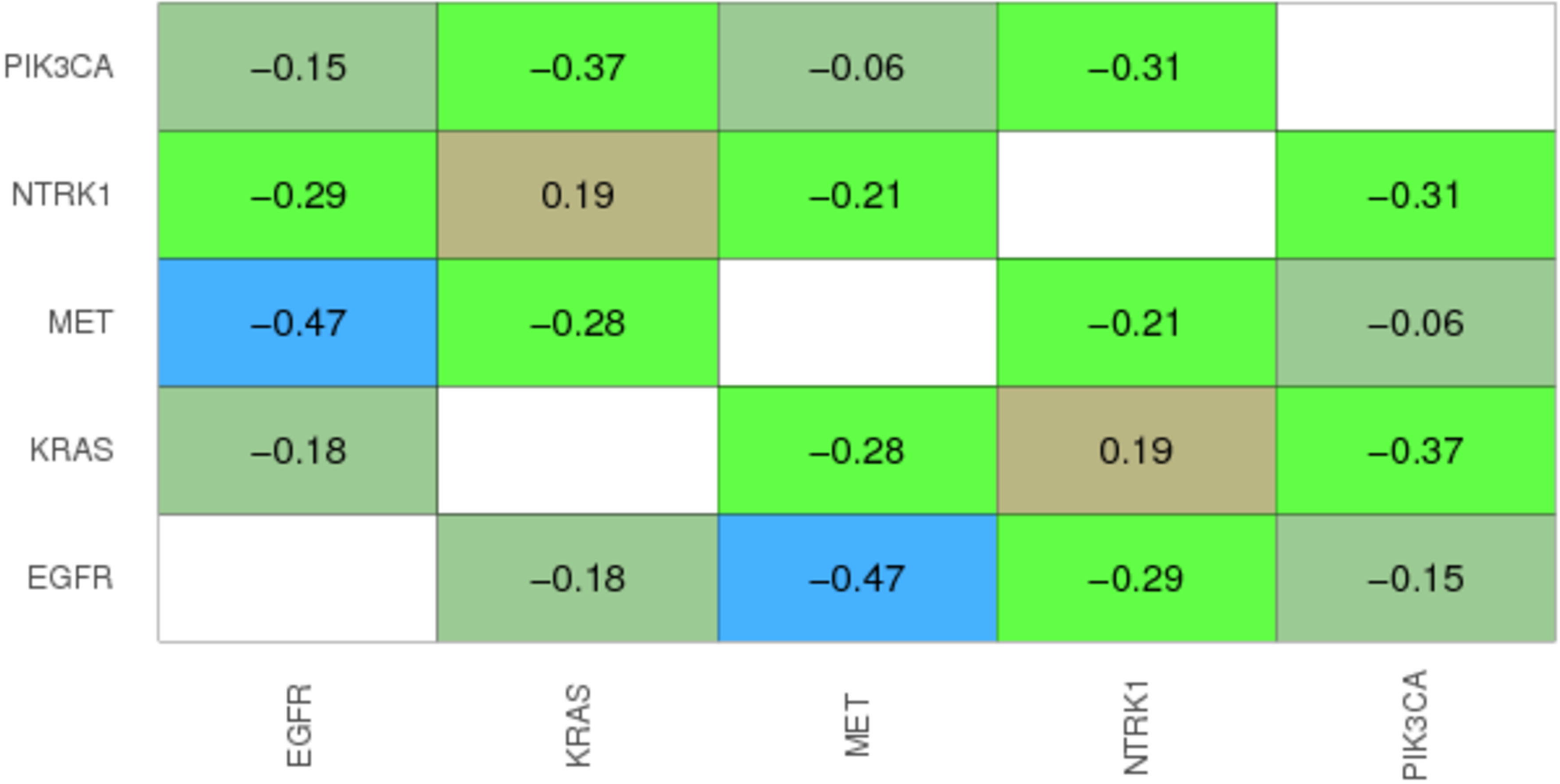


A) EGFR expression

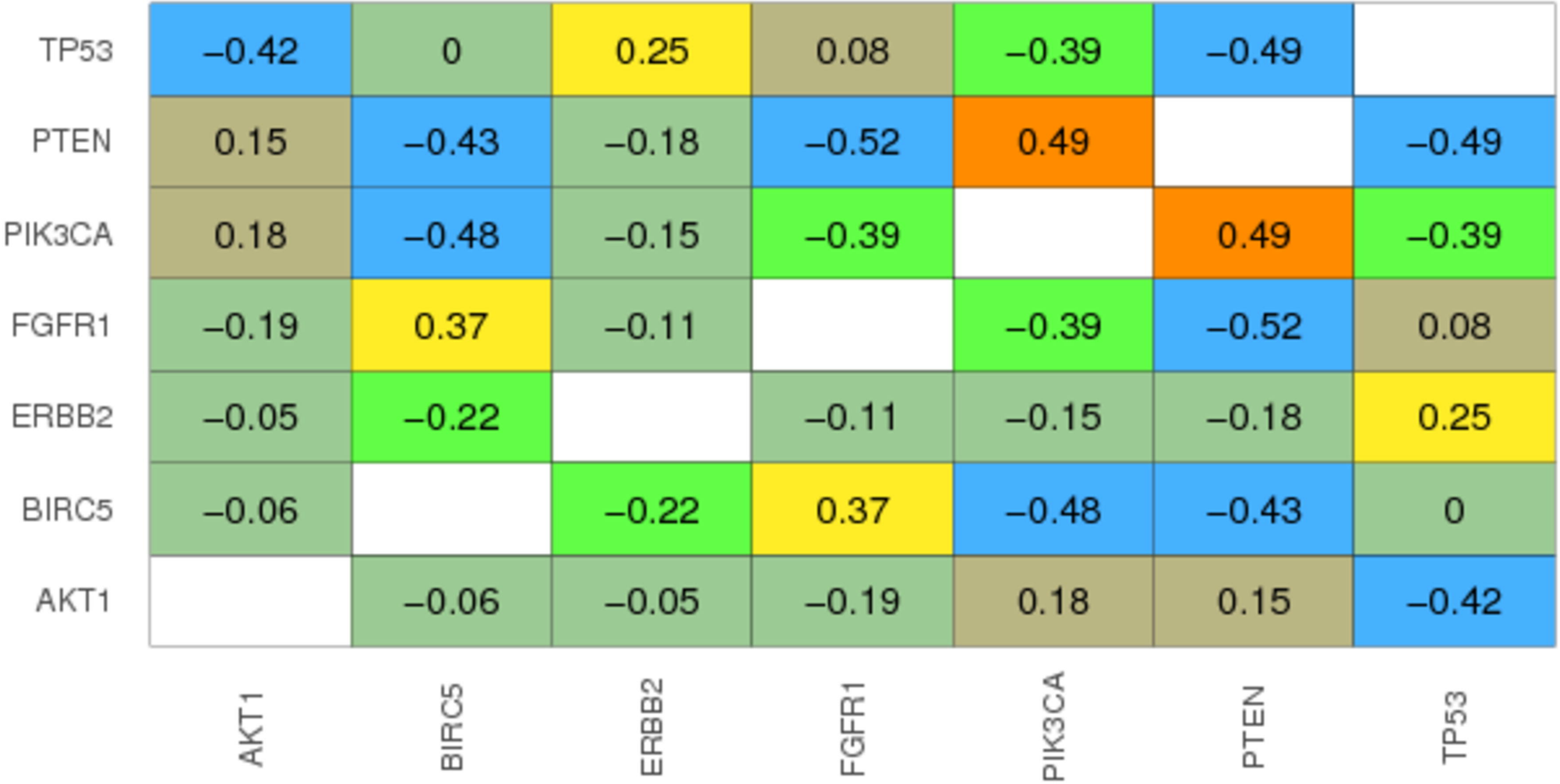
B) KRAS expression

C) CDKN2A expression

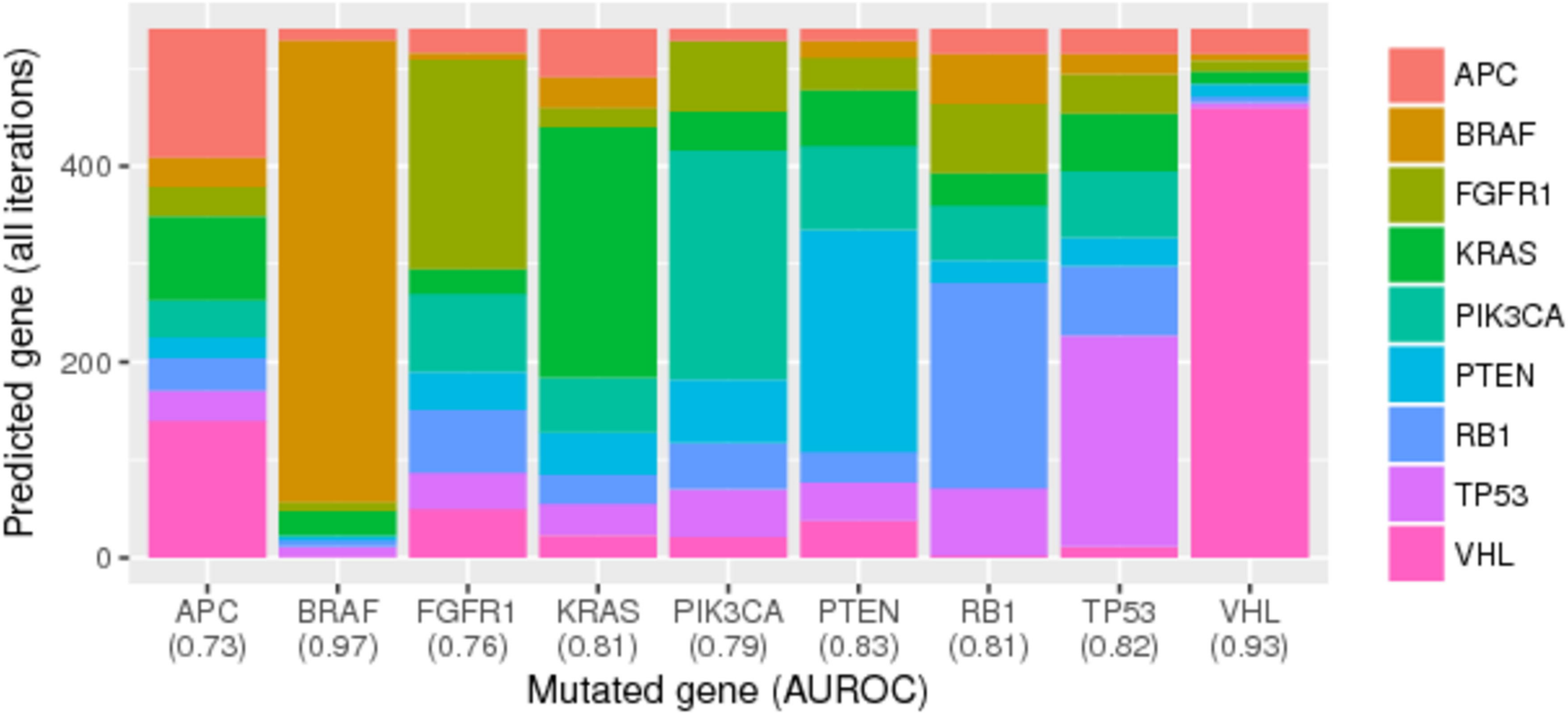

Predicted gene (all iterations)
0
50
100
150
EGFR
(0.81)
KRAS
(0.58)
MET
(0.78)
NTRK1
(0.68)
PIK3CA
(0.62)
Mutated gene (AUROC)
EGFR
KRAS
MET
NTRK1
PIK3CA



	EGFR	KRAS	MET	NTRK1	PIK3CA
PIK3CA	−0.15	−0.37	−0.06	−0.31	
NTRK1	−0.29	0.19	−0.21		−0.31
MET	−0.47	−0.28		−0.21	−0.06
KRAS	−0.18		−0.28	0.19	−0.37
EGFR		−0.18	−0.47	−0.29	−0.15



	AKT1	BIRC5	ERBB2	FGFR1	PIK3CA	PTEN	TP53
TP53	−0.42	0	0.25	0.08	−0.39	−0.49	
PTEN	0.15	−0.43	−0.18	−0.52	0.49		−0.49
PIK3CA	0.18	−0.48	−0.15	−0.39		0.49	−0.39
FGFR1	−0.19	0.37	−0.11		−0.39	−0.52	0.08
ERBB2	−0.05	−0.22		−0.11	−0.15	−0.18	0.25
BIRC5	−0.06		−0.22	0.37	−0.48	−0.43	0
AKT1		−0.06	−0.05	−0.19	0.18	0.15	−0.42


Predicted gene (all iterations)
0
200
400
APC
(0.73)
BRAF
(0.97)
FGFR1
(0.76)
KRAS
(0.81)
PIK3CA
(0.79)
PTEN
(0.83)
RB1
(0.81)
TP53
(0.82)
VHL
(0.93)
Mutated gene (AUROC)
APC
BRAF
FGFR1
KRAS
PIK3CA
PTEN
RB1
TP53
VHL