Rapid Identification of Genomic Alterations in Tumors affecting lymphocyte Infiltration (RIGATonI)

Running title: Identification of genomic variants and altered immune phenotypes in cancer

Raven Vella[1,2,3,4], Emily L. Hoskins[1,2,4], Lianbo Yu[4,5], Julie W. Reeser[1,4], Michele R. Wing[1,4], Eric Samorodnitsky[1,4], Leah Stein[1,2,4], Anoosha Paruchuri[1,4], ORIEN investigators[5], Michelle Churchman[5], Nancy Single[4], Wei Chen[4,7], Aharon G. Freud[4,7], Sameek Roychowdhury[1,4]

[1] Department of Internal Medicine, Division of Medical Oncology, The Ohio State University

[2] Biological Sciences Graduate Program, The Ohio State University

[3] Medical Scientist Training Program, The Ohio State Univer

[4] Comprehensive Cancer Center, The Ohio State University

[5] Department of Biomedical Informatics, The Ohio State University

[6] Aster Insights

[7] Department of Pathology, The Ohio State University

Corresponding author: Sameek Roychowdhury, Department of Internal Medicine, Division of Medical Oncology, Comprehensive Cancer Center and The James Cancer Hospital, Ohio State University, Columbus, OH 43210, USA. Tel: +1 614-685-5842; email: Sameek.roychowdhury@osumc.edu

Key words: Genomics, immune infiltration, computational staining, tumor immunity

Word count: 4471 words

Abstract

Recently identified tumor genomic alterations are associated with altered tumor immune microenvironments and therapeutic outcomes. These studies beg the question: are there other genomic variations altering the immune microenvironment in tumors that can provide insight into mechanisms of immune evasion? Current approaches to estimate immunity in bulk RNAseq from tumors include gene set enrichment and cellular deconvolution. In contrast, our new software package RIGATonI utilizes a gene agnostic approach to classify immune cell infiltration in tumors. Using pathologist reviewed histology slides and paired bulk RNA sequencing expression data, we trained a machine learning algorithm to detect inflamed and non-inflamed immune microenvironments. RIGATonI enables identifcation of candidate genomic alterations associated with immune infilration in subsets of cancer through evaluation of various features including gene expression, genomic alterations, pathologist-classified tumors, clinical outcomes, and protein validation. In addition to immune infiltrate classification, RIGATonI leverages a novel algorithm using known protein-protein interactions for prediction of gain-of-function and loss-of-function for genomic alterations to improve filtering of genomic alteration candidats. Applying these two novel methods, we analyzed all genomic alterations present in The Cancer Genome Altas and visualized promising results with R shiny for public use. Thus, we present our R tool and online database for Rapid Identification of Genomic Alterations in Tumors affecting lymphocyte Infiltration (RIGATonI).

Introduction

Immune surveillance is crucial for the eradication of cancer and is driven through understanding factors that impact the tumor immune microenvironment[1]. The tumor microenvironment is composed of tumor cells, along with surrounding immune cells, stroma cells and tissue matrix, and microbiota[1]. Each of these elements vary across patients and tumor types producing a spectrum of immune phenotypes and consequently clinical outcomes[1]. Although scRNA-seq and spatial transcriptomic technologies have dramatically increased our understanding of tumor immunology, there are substantially more tumor samples with bulk RNA sequencing and clinical data available for analysis. Thus, there is a need to develop computational tools specifically designed to assess immunity in large databases of tumor transcriptomes. In particular, there is a growing need to better understand the interconnected relationships between tumor genomics and their associated microenvironments[2, 3].

Although there are many techniques to assess immunity in cancer using bulk RNA sequencing data, there are alternative approaches that have yet to be explored. Current approaches (**Figure 1A**) include gene signature analyses (ImSig[4], Thorsson immune landscape of cancer[5]), cellular deconvolution (MCP-counter[6], QuantiseqR[7], CIBERSORT[8]), and ensemble results provided in databases (TIMER2.0[9], TIMEDB[10]). Gene set enrichment involves assessing the relative expression of gene sets related to immunity and then clustering cancers based on these results. This technique is particularly useful for understanding the drivers of immunity within one cancer type. This is routinely employed in large studies of immunotherapy outcomes to reveal what tumor features underscore interpatient differences[11, 12]. On the other hand, cellular deconvolution tools attempt to emulate flow cytometry or immunohistochemistry by estimating the number of specific immune cells in the sample[6-8]. This technique is extremely valuable for small subsets of tumors and for researchers who have a good understanding of tumor immunity. Databases of cellular deconvolution results have emerged which include ensemble analyses

across The Cancer Genome Atlas (TCGA) and other sources[9, 10]. As an exploratory tool for researchers without strong bioinformatic resources, this is exceptionally useful. We developed a novel computational approach to connect functional genomic variants and altered immune phenotypes in an unbiased manner. Our approach explores new frontiers in precision immuno-oncology not addressed by current methods: 1) model training with pathologist-defined immune infiltration as gold standard, 2) consideration for immune cell histologic features (tertiary lymphoid structures, dispersion characteristics, and degree of infiltration), 3) rapid, robust, and precise immune phenotype classification across big data resources, and 4) candidate filtering that leverages protein function prediction rather than simply gene mutation status.

We developed rapid identification of genomic alterations in tumors affecting lymphocyte Infiltration (RIGATonI) to accurately predict both the protein-level function and associated immune phenotype of genomic alterations. To build RIGATonI, we combined histologic features found by computational staining[13], pathologist-defined immune infiltration, protein-protein interaction networks[14], genomic data, and transcriptomic data (**Figure 1B**). First, we trained a machine learning algorithm to predict immune phenotypes using bulk RNAseq expression. We validated and fine-tuned our approach using manually reviewed histology by pathologists, computational staining, immunohistochemistry, and flow cytometry**.** We also built a novel modeling algorithm which can accurately predict the function of genomic variants (copy number alterations, single nucleotide variations, and structural variations) from bulk RNAseq expression. These two methods were combined to uncover connections between all the genomic alterations in TCGA and the immune phenotypes of samples harboring these alterations. We created a visualization interface (https://rigatoni.osc.edu) to help researchers access our TCGA analysis results on a gene-by-gene basis. Additionally, we built an R package which can be used to perform the analyses described. Finally, we provide examples of how RIGATonI identified genomic alterations associated with altered immune infiltration in subsets of cancer. In summary, RIGATONI will allow

researchers to investigate and shortlist genes of interest that associate with immune phenotypes in cancer.

Results

Building RIGATonI algorithm based on histology and bulk RNAseq

We used 403 tumors which were classified in a blinded fashion by two pathologists for low, intermediate, or high immune infiltration (**Figure 2A**). Each H&E slide was manually reviewed for amount of space occupied by lymphocytes (not occupied by tumor or stroma cells) and the distribution of these lymphocytes throughout the tumor (deeply penetrating, semi-penetrating, or peripheral). Further the pathologist also considered the quality of inflammation such as evidence of tumor cell killing or presence of tertiary lymphoid structures. All bulk RNAseq data for these samples were extracted and subsequent analyses utilized only RNAseq expression. 334/403 tumors were used for training the algorithm with XgBoost[15] machine learning in R to classify tumors as low, medium, or high immune infiltration. We assessed the accuracy of the algorithm by using the remaining 69 pathologist reviewed slides (not used for training), through comparison to computational staining results from TCGA[13], and finally an independent cohort of paired multiplex immunohistochemistry (mIHC) and flow cytometry data[16] (**Figure 2B-E**).

Assessment of the RIGATonI algorithm on tumors reserved for testing

We evaluated the RIGATonI model's assessment of 69 tumors to classify tumors high, medium, low infiltration and compared to pathologist review as the ground truth (**Figure 2B**). The balanced accuracy for high was 82.04%, the sensitivity was 73.33%, and the specificity was 90.74%. The balanced accuracy for medium was 63.4%, the sensitivity was 44.44%, and the specificity was 82.35%. The balanced accuracy for low was 82.58%, the sensitivity was 83.33%, and the specificity was 81.82%. The overall accuracy of the model was 71.01% (95% confidence interval 58.84% - 81.31%). The accuracy was significantly different from the no information rate ($p < .01$) and we failed to reject the null hypothesis of Mcnemar's test which indicates there is not significant evidence that the predictions made by the algorithm are different from the true

phenotypes (p > .05). Detailed statistics as well as a confusion matrix of the results are available in **Supplementary Table 1**.

Protein level validation of RIGATonI assessment of immune infiltrate

We compared RIGATonI-classified low, medium, and high immune phenotypes based on RNAseq from a study of 32 gastric cancer tumors which also included matching mIHC and flow cytometry[16]. First, we found that highly infiltrated tumors detected by RIGATonI display higher counts of immune cells collectively detected by mIHC (**Figure 2C)** including T cells and macrophages. In *post hoc* analysis of these results, we discovered this was predominantly due to increased CD8+ T cells in high samples compared to low samples (**Figure 2D**). In flow cytometry results from the same study, we found that RIGATonI classsified highly infiltrated tumors display significantly increased lymphocytes compared to RIGATonI-low tumors (**Figure 2E**).

Comparison of RIGATonI annotations to those of a computational staining

We applied RIGATonI to 5,202 tumors from TCGA and used an orthogonal approach for computational staining of the H&E images as previously described[13]. Saltz et al., developed a convolutional neural network to detect the percentage and distribution of lymphocytes on digital histology images. First, we evaluated 5 types of spatial attributes of tumors classified by RIGATonI as high, medium, and low infiltrates (**Figure 2F**). There is a significant difference in the composition of spatial arrangements of lymphocytes between these three groups. Highly infiltrated tumors had a higher proportion of brisk diffuse lymphocytes than the population (p <.01). Brisk diffuse arrangements indicate a broad distribution of many lymphocytes throughout the histology slide. Medium infiltrated tumors display brisk band-like arrangements more often than the population (p <.01). Brisk band-like infiltration patterns mean that the lymphocytes are clustered in a band across the slide, but that there are many lymphocytes. Tumors with low infiltration displayed a higher proportion of non-brisk focal lymphocyte arrangements (p <.01).

Non-brisk focal arrangements indicate that there are very few lymphocytes in a handful of locations across the image. Finally, we compared the number of lymphocytes across each group: tumors with RIGATonI high status have increased lymphocyte infiltration compared to both medium and low, and medium tumors display increased lymphocytes percentages to low tumors **(Figure 2G)**.

Developing an algorithm for functional assessment and shortlisting of tumor genomic alterations

RIGATonI seeks to connect genomic alterations with features of immune infiltration. There are many genomic alterations to evaluate; therefore, a functional measure of these alterations can help to prioritize alterations that are more likely to have a biologic impact. Currently, there is no bioinformatic tool to assess the functional impact of a genomic alteration from corresponding RNAseq. To address this gap, we developed RIGATonI's function assessment algorithm using parallel modelling and machine learning approaches. In contrast to DNA-based annotation of gene alteration pathogenicity, this approach has several benefits: 1) RIGATonI's assessment method is blind to the cause of altered function, so although we apply this method to groups with different DNA mutation status, RIGATonI could be applied to differing methylation status, alternative splicing conditions, or even treatment. 2) RIGATonI does not assess the mutation change itself, rather it investigates the RNA-based changes associated by that genomic alteration. This allows RIGATonI to assess novel variants and novel mechanisms of altered expression. 3) Not all patients with a genomic alteration experience the same molecular effects. RIGATonI can make predictions about alterations as a whole and sample level predictions based on the sample's specific molecular characteristics. This allows for users to investigate either by mutation status or functional status.

We sought to annotate alterations as gain-of-function (GOF) or loss-of-function (LOF). First, using the STRING[14] protein-protein interaction database, the algorithm identifies proteins which have direct and validated interactions with the gene of interest (**Figure 3A**). Both proteins which act on the gene of interest (upstream proteins) and proteins on which the protein of interest acts (downstream proteins) are considered. These two gene sets serve different purposes: upstream genes allow us to assess alterations which impact the expression level of the gene of interest; downstream genes allow us to assess the impact of alterations which may change the activity level of the gene of interest. We build a generalized linear model over a Poisson distribution using the extracted RNAseq counts of upstream and downstream proteins using samples with no alteration in the gene of interest. These two models are then saved and used to predict the RNAseq counts of the gene of interest within the mutant samples. 95% prediction intervals are created for both the upstream and downstream models. If the true counts of a mutant sample fall below the 95% prediction interval of either model, we annotate the sample LOF. If the true counts fall above the 95% prediction interval of either model, we annotate the sample GOF. If the true counts are within the 95% prediction interval for both models, the sample is annotated "unknown" (**Figure 3A**). To assess the function algorithm, we analyzed genomic alterations from 1008 oncogenes and tumor suppressor genes annotated by OncoKB [17]. OncoKB is the premiere source for clinically actionable genomic variants in cancer that are annotated based on curation and review of publications with corresponded experimental validation[17]. When applied to 10,464 tumors from the TCGA, the algorithm successfully identified 400 GOF alterations and 966 LOF genomic alterations (SNVs, structural variations, and copy number variations) within these 1008 genes. Genomic alterations harbored by fewer than 5 tumors were not assessed. Fusions were excluded mainly due to the high potential for false positives for fusions as many tumors harbored multiple fusions and imprecise breakpoints.

Validation of algorithm to classify known GOF and LOF genomic alterations

Of the 400 candidate GOF alterations uncovered, 62 were annotated in the OncoKB[17] database. The algorithm correctly categorized 60 of 62 as GOF, and 2 were incorrectly categorized as LOF, yielding an accuracy for GOF calling of 96% (**Figure 3B**). Notably, 84.5% of the alterations annotated by the algorithm had no information available in OncoKB[17]; most of these variants were whole gene amplifications (**Supplementary Table 2).** Similarly, we validated the accuracy of this algorithm for LOF alterations. Of the 966 LOF alterations uncovered, only 229 were annotated in the OncoKB[17] database. The algorithm correctly categorized 207 as LOF, while 22 were incorrectly categorized as GOF, yielding an accuracy of 90% (**Figure 3B**). Further, 76% of the candidate LOF alterations annotated by the algorithm had no information available in OncoKB[17]; most of these variants were whole gene deletions or premature terminations (**Supplementary Table 2)**. Altogether, these results suggest that this algorithm can successfully classify known GOF and LOF mutations and uncover novel candidate alterations that have not been well described.

The TCGA analysis yielded 7410 genomic alterations with possible immune effects among 5746 genes

These genes have been visualized in the RIGATonI web app. The front page of the website allows the user to choose a gene of interest and subset the analysis to only include certain cancer types or alterations (**Figure 4A**). After setting these parameters, there are two pages which allow for analysis of the mutant vs control tumors. First, the user may analyze the tumors' expression of specific genes using the transcriptome tab (**Figure 4B**). The user can choose to analyze using mutation status, function, or immune status. Axis labels, colors, and graph type can be changed for the user's convenience (**Figure 4B**). Across the website, non-parametric statistical tests are performed, and their p-values reported. We uncovered alterations of interest in 22 cancer types; most alterations discovered were associated with a highly infiltrated immune microenvironment (**Figure 4C**). To verify our results and allow for more detailed investigation, we

have analyzed these cohorts with quantiseqR[7], a popular cellular deconvolution tool, in the immune tab (**Figure 4D**). The user can analyze the same subsets as on the transcriptome tab and make the same adjustments to the graphs. Here we also show the composition of immune cells as a whole (left) and allow the user to select cell-types to investigate (right) (**Figure 4D**). To determine the number of novel results among the RIGATonI output, we investigated 226,093 original research works between 6/12/2010 and 6/12/2023 mentioning cancer and immunity in their abstracts via text mining. We discovered that 2773 (48%) of the RIGATonI output genes had not been previously connected to cancer immunity (**Figure 4E**). Only 72 genes (1%) had been mentioned in cancer immunity abstracts more than 100 times (**Figure 4E**).

RIGATonI identifies genomic alterations associated with altered immune infiltration in cancer

We sought to explore how RIGATonI can identify genomic alterations associated altered immune infiltration. We chose to investigate two genomic alterations of interest which had been previously reported in cancer immunity literature: *IDH1* and *MED16*. *MED16* has been recently identified as a potential immunotherapy target in melanoma via a CRISPR mouse screen [18]. Additionally, *MED16* has been connected to cancer progression and treatment resistance in breast cancer[19] and papillary thyroid cancer[20]. RIGATonI identified a connection between *MED16* copy number loss and immunity in melanoma. *MED16*-mutant tumors display a highly infiltrated phenotype more often than wildtype tumors ($p < .05$) (**Figure 5A**) and occurred in 12% of melanoma. These alterations were annotated LOF ($p < .01$). Using quantiseqR[7], it is evident that *MED16*-mutant tumors display increased immune infiltration compared to wildtype (**Figure 5B**). On *post hoc* analysis, this difference can be attributed to CD8+ T cell infiltration. *IDH1* alterations have been studied for their neoplastic, metabolic, or immunologic effects in glioma, cholangiocarcinoma, and leukemia[21]. RIGATonI identified a novel connection between *IDH1* copy number loss and immunity in lung adenocarcinoma. *IDH1*-mutant tumors display a highly

infiltrated phenotype more often than wildtype tumors ($p < .05$) (**Figure 5C**) and occurred in 2.5% of lung adenocarcinoma. These alterations were annotated LOF ($p < .01$). Using quantiseqR[7], it is evident that *IDH1*-mutant tumors display increased immune infiltration compared to wildtype (**Figure 5D**). This difference can be attributed to CD8+ T cell infiltration. Thus, *MED16* and *IDH1* alterations are examples of how RIGATonI can discover genomic alterations with altered immune infiltrates in subsets of cancer.

Discussion

RIGATonI's immune phenotyping algorithm leverages tumor RNAseq with two key approaches and offers a novel means to discover genetic subsets of cancer with altered immunity. Currently, gene set enrichment, cellular deconvolution, and ensemble approaches for estimating immunity from bulk RNAseq have advantages such as deeply understanding the immunological features of small cohorts, specificity of immune cell predictions, and increased access to big data for research teams lacking bioinformatic expertise, respectively. In contrast, RIGATonI utilizes a different approach from these methods by incorporating histologic features into our algorithm's development and validation. RIGATonI provides an additional layer of information not investigated by any other bulk RNAseq method: spatial characteristics of lymphocyte infiltrates. Although absolute infiltration is important, it is crucial to understand whether the infiltrates are deeply penetrating the tumor itself, displaying signs of active cell killing, and organize into tertiary lymphoid structures. This information, to our knowledge, is not captured in any other approach. RIGATonI is specifically designed to analyze thousands of samples and to categorize genomic alterations and their association with altered immune infiltrates.

RIGATonI's prediction of protein-level effects of genomic alterations from bulk RNAseq expression data is also novel. To our knowledge, only one other method has been devised to determine the function of genomic variants using bulk RNAseq: eVIP2[22]. Importantly, eVIP2[22] was validated using only 2 alterations in the same gene of interest whereas our validation set

demonstrated remarkable accuracy across 291 different alterations across 207 different genes (**Figure 3B**). Additionally, although we use the protein connections in STRING[14], if a user wanted to use a list of their own protein connections, they could do so with our technique thus identifying switch of function or specific pathways effected by a genomic alteration. This flexibility is not offered by methods which rely on DNA sequencing or by eVIP2[22] itself. VIPER is an R package which utilizes cell-type specific regulons built from bulk RNAseq expression data to predict the protein activity of different samples[23]. However, VIPER was not developed to determine the mutation level effects of genomic alterations themselves and was not validated against a robust set of known genomic alterations with molecularly confirmed protein-level effects[23]. RIGATonI's use of parallel modelling using two gene sets is highly innovative (**Figure 3A**). These two gene sets provide different information to the user and can be exploited separately or together. The upstream gene list uses expression counts of transcription factors which directly impact the expression of the gene of interest. Using this model, we can uncover genomic alterations which increase or decrease the expression of the gene directly. The downstream model however is build using any proteins on which the gene of interest acts. Using this model, we can understand how a genomic alteration impacts the function of the protein when the expression of the protein is not affected. We attempted to combine these models but found that doing so worsens the performance of the model. The reason for this is unclear and an area for future exploration.

Although user specific gene lists can be used, we believe the use of the STRING[14] database is a powerful feature of our tool. Initially, our group attempted to perform traditional pathway analysis (using both Kegg[24-26] and GO[27, 28]) to determine the functional outcomes of genomic alterations; however, we were not successful. Protein-protein interaction networks from STRING[14] proved the most effective choice because of the ability to filter for proteins with direct interactions with the gene of interest (**Figure 3A**). Individual research groups with specific

knowledge of a particular protein can bypass the use of STRING[14] and input their own upstream and downstream gene lists.

The only other approach we are aware of which systematically assesses protein effects on cancer immunity is CRISPR screening in mouse models[29]. This approach also has various benefits: the ability to assess proteins which are not frequently mutated in human cancer, molecular validation of the immunologic effects, and the ability to test different treatment protocols. The RIGATonI software suite is strictly an in-silico approach, and all results certainly require experimental validation to confirm. Doing so is outside the scope of this analysis, but we did investigate the results of *IDH1* copy number loss in lung adenocarcinoma (**Figure 5C-D**). Mutant *IDH1* has been connected to cancer immunity in glioma with improved survival benefit in numerous cancer types[21]. We see a pattern of increased immune cell infiltration in IDH1 mutant tumors compared to wildtype (**Figure 5C-D**). Additionally, we identified 72 genes with immunologic affects which have been discussed more than 100 times in the literature (**Figure 4E**). These results together are encouraging there is evidence that these alterations have legitimate immunological impacts. Despite this clear limitation, RIGATonI does have key advantages over a CRISPR screen. RIGATonI assesses gain of function variants, alterations that already occur in human cancers, and immunity in human patients not mice. Furthermore, RIGATonI is inexpensive and instantaneous. Coupling RIGATonI output and those of a CRISPR screen would be a powerful combination to filter and assess which variants merit further investigation. In fact, we explored this possibility via a target recently identified by a CRISPR screen, *MED16*[18]. Loss of function alterations in *MED16* associate with increased immune infiltration and increased CD8+ T cells in melanoma (**Figure 5A-B**). A CRISPR screen performed in a melanoma cell line specifically mentioned *MED16* as a promising target[18]. This result not only gives credence to RIGATonI's output, but also extends the CRISPR screen results to include a

rapid analysis of human tumors. We hope to explore candidates from this CRISPR screen and other in the future.

RIGATonI is a powerful approach for identifying novel genomic alterations with immune effects. That said, RIGATonI suffers from several limitations including sample size limitations, discordance between pathologists' reviews, and finally exclusion of genomic fusions from the function annotation algorithm validation.

By using pathologist review rather than computational staining as a ground truth for our algorithm, we do limit the sample size of our training data. However, the quality of sample annotations is far more important that the size of the training data (in this case 403 tumors). For instance, although the Saltz et al approach considers the digital slides as a whole and does not differentiate between lymphocytes deeply penetrating the tumor and lymphocytes outside the tumor boundaries[13]. Our initial model trained with samples selected using Saltz et al annotations was discordant with pathologist annotations. New approaches like Lunit SCOPE IO also use deep learning techniques to estimate immune cells on digital pathology slides[30]. Lunit SCOPE IO was trained on >17,000 H&E images across 24 cancer types[30]. This scale of training data would never be possible using manual pathologist review. Additionally, Lunit SCOPE IO considers the boundaries of the tumor on the histology slide[30]. Importantly, the goal of this software is to predict immunotherapy response not to quantify or qualify inflammation more broadly[30]. Lunit SCOPE IO is primarily a clinical tool and does not make available its software for research use on big data repositories. In contrast, RIGATonI is a bioinformatic research tool developed specifically to analyze connections between genomic alterations and their associated immune microenvironments. Although it could be argued that Lunit SCOPE IO could be adapted for this purpose, not all large databases of tumors make digital slides available. For these reasons, RIGATonI offers a distinct benefit over other methods to estimate spatial features of immunity in big data despite smaller sample size of the training data.

Pathologist assessment introduces some subjectivity into the training data annotations. This subjectivity is clear from the heterogeneity of annotations made within the medium group (**Figure 2B**). Although this may limit the calculated accuracy of our model, it reflects the reality that independent physician judgement of the same slides may yield different results. In fact, two pathologists initially reviewed slides for this study and disagreed with each other's annotations ~50% of the time. We built models using each pathologist's annotations separately and together using two different machine learning approaches (**Supplementary Table 3**). The best model was yielded using the annotations from pathologist two (**Supplementary Table 3**). We speculate that pathologist two's annotations are more internally consistent and thus easier to model. It is widely agreed that physician opinion is the gold standard against which AI and machine learning approaches should be measured[31]; however, doing so demonstrates that human judgement is more malleable than computational modelling. In this case, our extensive validation using two orthogonal datasets with protein and spatial characteristics reveals that although the model statistics may suffer from human error, qualitatively the model performs exceptionally well regardless.

Validation of RIGATonI's functional analysis only took place on alterations with >5 instances of the alteration to ensure that statistical tests could be performed to determine the functional status. The ability of RIGATonI to classify alterations declines with declining sample count. Additionally, co-mutations are not considered in the analysis of each sample because, again, of sample size concerns. Finally, in our pan cancer analysis of TCGA, we did not consider whether samples contained more than one alteration in each gene. Unfortunately, such distinctions would not be computationally scalable. Because of this sample size limitation, in the pan-TCGA analysis, we annotate copy number alterations more frequently than any other alteration type. Copy number alterations are extremely common compared to specific point mutations, gene fusions, or structural variations. Additionally, we did not group together early

terminations or different point mutations occurring at the same locus out of an abundance of caution (early terminations at different loci may have different effects; different point mutations at the same locus may have different effects). Without detailed knowledge of all 17 thousand genes analyzed, we approached the first version of RIGATonI conservatively.

We were not able to include gene fusions in our function annotation algorithm validation. Gene fusions are genomic events wherein two genes translocate together and sometimes produce a functional gene product which behaves differently than the parent genes[32]. Not all gene fusions have functional impacts on the gene of interest[32]. Additionally, although two samples might harbor a fusion of the same two genes, the breakpoints of this fusion could be vastly different and the resulting protein product entirely distinct[32]. Fusion calling bioinformatic methods are very accurate for determining the breakpoints and read support for a fusion event; however, these tools are intended to be manually assessed by a genetics expert who can determine whether the resulting gene product is likely to yield altered gene function[32]. Developing a computational approach to carry out this manual assessment for any gene of interest and any fusion partner with any breakpoint is outside the scope of this analysis. Without such a method to provide ground truth for our validation, we were not able to perform validation with fusion samples. Despite this limitation, the RIGATonI method is blind to the genomic cause of altered function, and therefore we don't believe there is any reason fusion samples would perform differently (Figure 3A).

Despite these limitations, RIGATonI is an innovative and novel approach for identifying genomic alterations with immunological impacts in cancer. Overall, RIGATonI is a novel tool specifically designed for the emerging field of precision immuno-oncology. We believe RIGATonI can be used for a vast variety of purposes including biomarker discovery and identification of novel drug targets. RIGATonI will assist researchers in their search for targetable genes which impact tumor immunity.

Methods

Tumor sequencing data. All data shown here were previously published or made available by a large database of tumor RNA sequencing (RNAseq) and whole exome sequencing data. Secondary analyses were performed within Ohio Supercomputer Center. Data sources include The Cancer Genome Atlas and the Oncology Research Information Exchange Network. RNAseq count data were downloaded from TCGA using GenomicDataCommons[33], and batch corrected with ComBat-seq[34] using institution of origin as the batch. Genomic variant calling was downloaded in the form of combined .maf files. Copy number alterations were downloaded in the form of gene based raw copy number. Finally, all whole genome .bam files were downloaded from TCGA and then processed with Delly[35], Manta[36], and GRIDSS[37] to assess for structural variants. These results were combined using SURVIVOR[38] with default settings. RNA sequencing data was also obtained from ORIEN and batch corrected with ComBat-seq[34] according to the RNAseq batch information made available through Aster Insights. Copy number alterations were downloaded from ORIEN in the form of gene based raw copy number. Demographic information was downloaded from ORIEN as well. ORIEN data is managed by Aster Insights, requests for this data should be sent to Aster Insights.

RIGATONI immune phenotyping algorithm validation with pathologist review. Two pathologists independently reviewed 403 slides assessing lymphocyte infiltration characteristic. Pertinent characteristics included percentage of space not occupied by tumors or stroma which was occupied by lymphocytes, dispersion of lymphocytes within tumor, and presence or absence of tertiary lymphoid structures. Taking into consideration all these characteristics, each pathologist annotated the slide either high, medium, or low infiltration. Paired RNA sequencing was collected and used along with annotations to build a series of models predicting immune phenotypes from bulk RNAseq counts. First, important features were selected using multinomial ElasticNet[39] via the R package glmnet[40]. Next, these features were used as predicted for six different machine

learning models. Three models were built with XgBoost[15] via the R package with the same name and three with the R package ordinalForest[41]. Two models were built considering both pathologist predictions, two considering just pathologist one and two considering pathologist two. Each algorithm uses the same cohort of 334 samples for training and 69 samples for testing. The algorithm with the best performance on the testing data was selected to be used going forward. Please see supplemental table 4 for model specific information. Accuracy of the testing data for the model selected were visualized using ggplot2[42].

Gastric cancer multiplex immunohistochemistry (mIHC) and flow cytometry. RNAseq count data was downloaded from Saito et al along with flow cytometry and IHC outputs. These include 30 patients with gastric cancer in Tokyo[16]. These results were processed as previously described in Saito et al. Using this resource, we used a MANOVA[43] to compare the IHC counts to determine if there were any significant differences between groups. Next, we used a Wilcox test[44] to assess the counts of each subset of IHC-marked cells to find significant differences. We performed pairwise comparisons of flow cytometry results using Wilcox tests[44]. Results were visualized using ggplot2[42].

Determine resulting immune phenotype of each genomic alteration using RNAseq data. An R function was created which converts RNAseq count data to TPM using the R packages DESeq2[45] and DGE.obj.utils[46], and then filters the data down to only the genes selected by ElasticNet[39]. The immune phenotype of each samples was predicted using XgBoost[15]. The proportion of high and low tumors for each cancer type were calculated. To analyze a genomic alteration, all mutant samples provided to the function are compiled, and a 1-proportion z-test is performed where the null hypothesis is that the proportion of hot samples will be equal to the population proportion within that cancer type and the alternative hypothesis is that the proportion of high samples is greater than the population proportion. If the null hypothesis is rejected ($P<.05$), the genomic alteration is annotated high. If not, a second 1-proportion z-test is performed where

the null hypothesis is that the proportion of low samples will be equal to the population proportion and the alternative hypothesis is that the proportion of low samples is greater than the population proportion within that cancer type. If this null hypothesis is rejected, the genomic alteration is annotated low. If not, the alteration is annotated Unknown.

Determining functional status of each genomic alteration using RNA-seq data. The STRING[14] database's protein action version 10.5 was downloaded. The list of protein actions is subset to include only actions on or by the protein of interest (POI). These actions are further filtered into two lists: an upstream protein list with only proteins that act to affect the expression of the POI, and a downstream protein list including all genes the POI activates, inhibits, or alters expression. The upstream gene list is used to model the RNAseq counts of the POI using modulators of the POI's expression. The downstream gene list is used to model the RNAseq counts of the POI using downstream genes as indicators of its activity.

To model typical expression patterns of the POI, all samples with no alteration (control samples) in the gene of interest (GOI) are collected, and two generalized linear models are created over a Poisson distribution to predict the RNA counts of the POI/GOI. One model uses the upstream protein list as predictors, and another using the downstream protein list. Both models predict the RNAseq counts of the POI.

Next, the RNAseq counts of the GOI within samples harboring mutations (mutant sample) predicted separately with each regression model. If, in either model, the expression of the GOI is lower than the lower bound of the 95% prediction interval (created by ciTools[47]), the mutant sample is annotated loss of function (LOF). On the other hand, if, in either model, the expression of the GOI is higher than the upper bound of the 95% prediction interval, the mutant sample is annotated gain-of-function (GOF). Falling outside the bounds of these prediction intervals indicates that the mutant sample's GOI expression or activity is more different than that of a

control sample than we would expect from random chance. If the mutant sample's GOI expression falls within the bounds of both these prediction intervals, we can conclude that there is not enough evidence to indicate the mutation is causing abnormal expression or activity in that sample.

Next, all mutant samples provided are compiled, and a 2-proportion z-test is performed (null hypothesis z = .5) comparing the proportion of LOF and GOF annotations. If the null hypothesis is rejected ($P<.05$), the genomic alteration is annotated with the more frequent sample level annotation. If not, the alteration is annotated Unknown.

Function annotation algorithm validation. OncoKB[17] provides a list of oncogenes with either targeting drugs or known oncogenic mutations. All alterations in these 1008 genes were analyzed in parallel. The results were filtered to only include GOF and LOF calls from the algorithm. Various alterations are called Unknown due to low sample count, confounding variables, or lack of known connections in STRING[14]. We did not include these samples in the pan-cancer analysis of TCGA, however users can elect to include them in their own analysis. Additionally, gene fusions were removed due to complexity of their calling. This left 1366 genomic alterations to investigate. We manually searched OncoKB[17] for information about each alteration and, if available, recorded the true function of the variant. Results were visualized with ggplot2[42].

Building the R package. Functions were created with the R package devtools[48] which create an upstream and downstream gene list from STRING[17], determine the function of a group of samples using RNA expression data from bulk RNAseq, and the sample level immune phenotype using RNA expression data from bulk RNAseq. These functions are described in detail above. The R package along with relevant documentation is available at https://github.com/OSU-SRLab/RIGATONI.

Comprehensive analysis of genomic alterations in TCGA. All mutation information in TCGA was downloaded and compiled. A sample is said to have a copy number variation (CNV) if the total number of copies is >= 6 or <2. We also considered the sex of the patient in question if the gene of interest was on the X or Y chromosome. For male patients, genes on the X chromosome were said to be deleted if there were 0 copies. For female patients, no genes on the Y chromosome were considered deleted. For any patients without biological sex available, we excluded them from the copy number analysis of genes on the X or Y chromosome. Next, all genomic alterations were analyzed in parallel through RIGATonI R package described above. We ensured to analyze each alteration within each cancer type for both functional annotations and immune phenotyping. The population proportions used for immune phenotype were those of the cancer type being analyzed. Finally, we stored the significant results from TCGA to be visualized by our online tool. The RIGATonI website is located at https://rigatoni.osc.edu/ and is managed by the Roychowdhury lab group and the Ohio Supercomputing Center. The website is built with the R package shiny[49], all graphs are visualized with ggplot2[42]. The website input is a user-provided GOI, and the website compiles gene level copy number, fusion, and simple nucleotide variation results from the pan TCGA analysis. Next, we provide various visualizations to the user (quantiseqR[7] cell type proportions, RNA seq count data, and primary site prevalence) along with a table describing the different genomic alterations which were both functional and immunogenic within the GOI.

Text mining of PubMed abstracts to estimate novelty of RIGATONI TCGA output. PubMed abstracts containing the words “cancer” and “immunity” or “immunology” or “immune” since 6/12/2010 were downloaded using the R package pubmed.mineR[50]. The function gene_atomization was used to perform text mining annotation of each gene mentioned. The RIGATonI results were extracted and each gene was annotated with their frequency of

appearance. Preprint publications were excluded from this analysis. A total of 226,093 abstracts were analyzed. Results were visualized with ggplot2[42].

Gene level analysis of *MED16* and *IDH1*. These genes were analyzed as part of the pan-TCGA analysis described above. Results were extracted to highlight here.

Acknowledgements. The results published here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

1. Binnewies M, Roberts EW, Kersten K, Chan V, Fearon DF, Merad M, et al. Understanding the tumor immune microenvironment (TIME) for effective therapy. Nature Medicine. 2018;24(5):541-50.
2. Haslam A, Prasad V. Estimation of the Percentage of US Patients With Cancer Who Are Eligible for and Respond to Checkpoint Inhibitor Immunotherapy Drugs. JAMA network open. 2019;2(5):e192535.
3. Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Ou Yang TH, et al. The Immune Landscape of Cancer. Immunity. 2018;48(4):812-30.e14.
4. Nirmal AJ, Regan T, Shih BB, Hume DA, Sims AA-O, Freeman TC. Immune Cell Gene Signatures for Profiling the Microenvironment of Solid Tumors. (2326-6074 (Electronic)).
5. Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Ou Yang TH, et al. The Immune Landscape of Cancer. (1097-4180 (Electronic)).
6. Becht E, Giraldo NA, Lacroix L, Buttard B, Elarouci N, Petitprez F, et al. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. Genome biology. 2016;17(1):218.
7. Finotello F, Mayer C, Plattner C, Laschober G, Rieder D, Hackl H, et al. Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data. Genome Medicine. 2019;11(1):34.
8. Chen B, Khodadoust MS, Liu CL, Newman AM, Alizadeh AA. Profiling Tumor Infiltrating Immune Cells with CIBERSORT. Methods in molecular biology (Clifton, NJ). 2018;1711:243-59.
9. Li T, Fu J, Zeng Z, Cohen D, Li J, Chen Q, et al. TIMER2.0 for analysis of tumor-infiltrating immune cells. (1362-4962 (Electronic)).
10. Wang X, Chen L, Liu W, Zhang Y, Liu D, Zhou C, et al. TIMEDB: tumor immune micro-environment cell composition database with automatic analysis and interactive visualization. Nucleic acids research. 2023;51(D1):D1417-D24.
11. Motzer RJ, Robbins PB, Powles T, Albiges L, Haanen JB, Larkin J, et al. Avelumab plus axitinib versus sunitinib in advanced renal cell carcinoma: biomarker analysis of the phase 3 JAVELIN Renal 101 trial. Nature Medicine. 2020;26(11):1733-41.
12. Motzer RJ, Powles T, Atkins MB, Escudier B, McDermott DF, Alekseev BY, et al. Final Overall Survival and Molecular Analysis in IMmotion151, a Phase 3 Trial Comparing Atezolizumab Plus Bevacizumab vs Sunitinib in Patients With Previously Untreated Metastatic Renal Cell Carcinoma. (2374-2445 (Electronic)).
13. Saltz J, Gupta R, Hou L, Kurc T, Singh P, Nguyen V, et al. Spatial Organization and Molecular Correlation of Tumor-Infiltrating Lymphocytes Using Deep Learning on Pathology Images. (2211-1247 (Electronic)).

549 14. Szklarczyk D, Kirsch R, Koutrouli MA-O, Nastou KA-O, Mehryary FA-O, Hachilif R, et al. The
550 STRING database in 2023: protein-protein association networks and functional enrichment analyses for
551 any sequenced genome of interest. (1362-4962 (Electronic)).
552 15. Chen T, Guestrin C, editors. Xgboost: A scalable tree boosting system. Proceedings of the 22nd
553 acm sigkdd international conference on knowledge discovery and data mining; 2016.
554 16. Saito N, Sato Y, Abe H, Wada I, Kobayashi Y, Nagaoka K, et al. Selection of RNA-based evaluation
555 methods for tumor microenvironment by comparing with histochemical and flow cytometric analyses in
556 gastric cancer. Scientific Reports. 2022;12(1):8576.
557 17. Chakravarty D, Gao J, Phillips SM, Kundra R, Zhang H, Wang J, et al. OncoKB: A Precision
558 Oncology Knowledge Base. LID - 10.1200/PO.17.00011 [doi] LID - PO.17.00011. (2473-4284 (Print)).
559 18. Dubrot J, Du PP, Lane-Reticker SK, Kessler EA, Muscato AJ, Mehta A, et al. In vivo CRISPR screens
560 reveal the landscape of immune evasion pathways across cancer. Nature Immunology.
561 2022;23(10):1495-506.
562 19. Li H, Li K, Shu D, Shen M, Tan Z, Zhang W, et al. MED16 Promotes Tumour Progression and
563 Tamoxifen Sensitivity by Modulating Autophagy through the mTOR Signalling Pathway in ER-Positive
564 Breast Cancer. LID - 10.3390/life12101461 [doi] LID - 1461. (2075-1729 (Print)).
565 20. Gao H, Bai P, Xiao L, Shen M, Yu Q, Lei Y, et al. Mediator complex subunit 16 is down-regulated
566 in papillary thyroid cancer, leading to increased transforming growth factor-β signaling and radioiodine
567 resistance. (1083-351X (Electronic)).
568 21. Ni Y, Shen P, Wang X, Liu H, Luo H, Han X. The roles of IDH1 in tumor metabolism and immunity.
569 Future Oncology. 2022;18(35):3941-53.
570 22. Thornton AM, Fang L, Lo A, McSharry M, Haan D, O'Brien C, et al. eVIP2: Expression-based
571 variant impact phenotyping to predict the function of gene variants. PLOS Computational Biology.
572 2021;17(7):e1009132.
573 23. Alvarez MJ, Shen Y, Giorgi FM, Lachmann A, Ding BB, Ye BH, et al. Functional characterization of
574 somatic mutations in cancer using network-based inference of protein activity. Nature genetics.
575 2016;48(8):838-47.
576 24. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research.
577 2000;28(1):27-30.
578 25. Kanehisa M. Toward understanding the origin and evolution of cellular organisms. Protein
579 science : a publication of the Protein Society. 2019;28(11):1947-51.
580 26. Kanehisa M, Furumichi M, Sato Y, Kawashima M, Ishiguro-Watanabe M. KEGG for taxonomy-
581 based analysis of pathways and genomes. Nucleic acids research. 2023;51(D1):D587-d92.
582 27. Aleksander SA, Balhoff J, Carbon S, Cherry JM, Drabkin HJ, Ebert D, et al. The Gene Ontology
583 knowledgebase in 2023. Genetics. 2023;224(1).
584 28. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for
585 the unification of biology. The Gene Ontology Consortium. Nature genetics. 2000;25(1):25-9.
586 29. Holcomb EA, Pearson AN, Jungles KM, Tate A, James J, Jiang L, et al. High-content CRISPR
587 screening in tumor immunology. (1664-3224 (Electronic)).
588 30. Shen JA-O, Choi YL, Lee T, Kim H, Chae YK, Dulken BW, et al. Inflamed immune phenotype
589 predicts favorable clinical outcomes of immune checkpoint inhibitor therapy across multiple cancer
590 types. LID - 10.1136/jitc-2023-008339 [doi] LID - e008339. (2051-1426 (Electronic)).
591 31. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. (2514-6645
592 (Print)).
593 32. Taniue K, Akimitsu N. Fusion Genes and RNAs in Cancer Development. LID -
594 10.3390/ncrna7010010 [doi] LID - 10. (2311-553X (Electronic)).
595 33. Martin Morgan SD, Marcel Ramos. GenomicDataCommons. 1.26.0 ed2023.

34. Zhang Y, Parmigiani G, Johnson WE. ComBat-seq: batch effect adjustment for RNA-seq count data. NAR Genomics and Bioinformatics. 2020;2(3):lqaa078.
35. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics (Oxford, England). 2012;28(18):i333-i9.
36. Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics (Oxford, England). 2016;32(8):1220-2.
37. Cameron DL, Schröder J, Penington JS, Do H, Molania R, Dobrovic A, et al. GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. Genome research. 2017;27(12):2050-60.
38. Jeffares DA-O, Jolly C, Hoti M, Speed DA-O, Shaw L, Rallis C, et al. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. (2041-1723 (Electronic)).
39. Zou H, Hastie T. Regularization and Variable Selection Via the Elastic Net. Journal of the Royal Statistical Society Series B: Statistical Methodology. 2005;67(2):301-20.
40. Friedman J, Hastie T, Tibshirani R, Narasimhan B, Tay K, Simon N, et al. glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models. cran2023.
41. Hornung R. Ordinal Forests. Journal of Classification. 2020;37(1):4-17.
42. Wickham H. ggplot2: Elegant Graphics for Data Analysis: Springer-Verlag New York; 2016. Available from: https://ggplot2.tidyverse.org.
43. Smith H, Gnanadesikan R, Hughes JB. Multivariate Analysis of Variance (MANOVA). Biometrics. 1962;18(1):22-41.
44. Wilcoxon F. Individual comparisons by ranking methods. Biom. Bull., 1, 80–83. 1945.
45. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology. 2014;15(12):550.
46. Law CW, Alhamdoosh MA-O, Su S, Dong X, Tian LA-O, Smyth GA-O, et al. RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR. LID - ISCB Comm J-1408 [pii] LID - 10.12688/f1000research.9005.3 [doi]. (2046-1402 (Electronic)).
47. John Haman MA, Institute for Defense Analyses. ciTools: Confidence or Prediction Intervals, Quantiles, and Probabilities for Statistical Models. 2020.
48. Hadley Wickham JH, Winston Chang, Jennifer Bryan, RStudio. devtools: Tools to Make Developing R Packages Easier. 2.4.5 ed2022.
49. Winston Chang JC, JJ Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert, Barbara Borges, Posit Software, PBC, jQuery Foundation, jQuery contributors, jQuery UI contributors, Mark Otto, Jacob Thornton, Bootstrap contributors, Twitter, Inc, Prem Nawaz Khan, Victor Tsaran, Dennis Lembree, Srinivasu Chakravarthula, Cathy O'Connor, PayPal, Inc, Stefan Petre, Andrew Rowls, Brian Reavis, Salmen Bejaoui, Denis Ineshin, Sami Samhuri, SpryMedia Limited, John Fraser, John Gruber, Ivan Sagalaev, R Core Team. shiny: Web Application Framework for R. 1.8.0 ed2023.
50. Rani J, Shah Ab Fau - Ramachandran S, Ramachandran S. pubmed.mineR: an R package with text-mining algorithms to analyse PubMed abstracts. (0973-7138 (Electronic)).

A

	Gene Set Enrichment (ImSig, Thorsson et al, JAVELIN)	Cellular Deconvolution (TIMER, CIBERSORT, quanTIseq)	RIGATONI immune phenotyping
Goal	Find subgroups of cancer	Predict number of each immune cells and their subtype	Determine degree of inflammation in tumor sample
Feature Selection	Known immunity gene sets	Known immunity gene sets	Agnostic gene selection with machine learning
Solid Tumor Validation	Immunotherapy response	Correlation with flow cytometry, liquid cell suspensions, and in-silico mixtures	Histology slide review

B

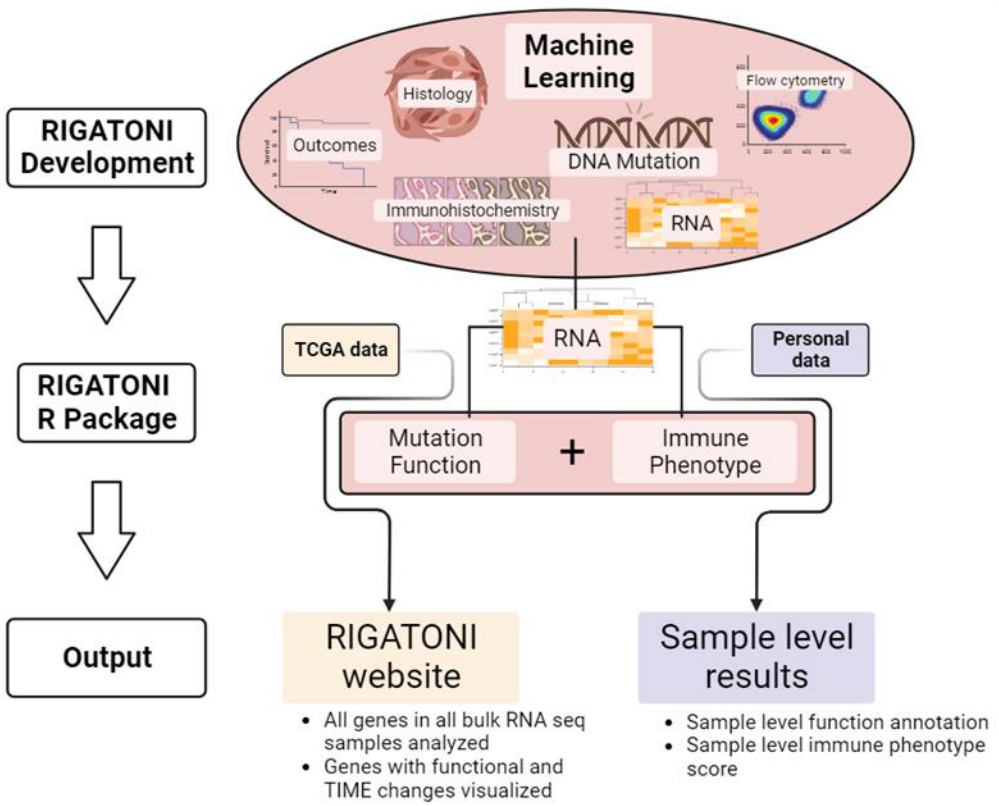

Figure 1: A. Table of current approaches to estimate immunity in big data bulk RNA seq. Unlike other big data approaches for measuring immunity, RIGATonI is not developed to quantify the number of cells or the activity of those cells. **B.** RIGATonI was developed using both the degree of infiltration and the spatial organization of the lymphocytes. Instead of using canonical markers of immune activation, we allowed unsupervised selection of predictors to best predict pathologist review.

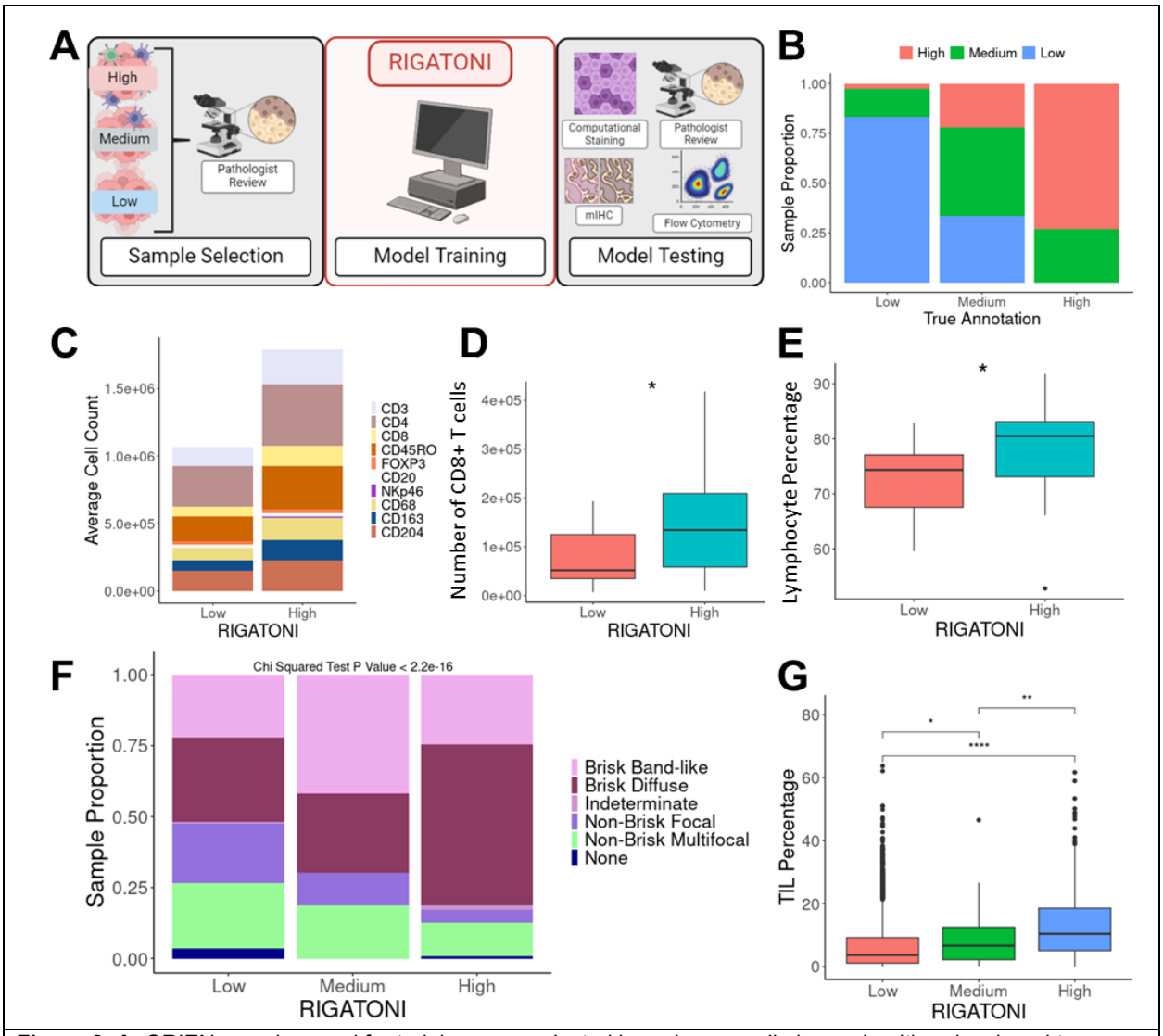

Figure 2: A. ORIEN samples used for training were selected based on a preliminary algorithm developed to predict strongly hot and cold phenotypes. Medium samples were selected at random from the ORIEN dataset. A pathologist reviewed each slide and annotated high, medium or low degrees of infiltration. We built a machine learning model to predict these annotations using XgBoost. We validated the accuracy of the algorithm using computational staining, a subset of pathologist reviewed slides not used for training, multiplex IHC, and flow cytometry of an independent dataset. **B.** The accuracy of the model (Y-axis) was determined using a subset of pathologist reviewed slides (X-axis) not using for training data. **C/D/E**. To confirm our algorithm was robust outside of the ORIEN training and testing database, we investigated an independent dataset of gastric tumors with paired RNAseq, mIHC and Flow cytometry. **C.** The overal composition of the immune microenvironment of high and low samples**. D.** The counts of CD8+ T cells detected by mIHC are significantly increased in high samples. **E**. Flow cytometry of these tumors demonstrated an increase in lymphocytes in RIGATonI high vs. low samples. **F/G.** We also investigated our algorithms association with histologic features detected by convolutional neural networks in TCGA. The overal distribution of spatial characteristics is significantly different across low medium and high subsets **(F)**. High samples more often display brisk diffuse lymphocyte patterns shown in maroon **(F)**. Additionally, the percentage of cells on the slide estimated to be lymphocytes is significantly greater in Medium samples compared to Low samples, High samples compared to Medium samples, and High samples compared to low samples **(G)**.

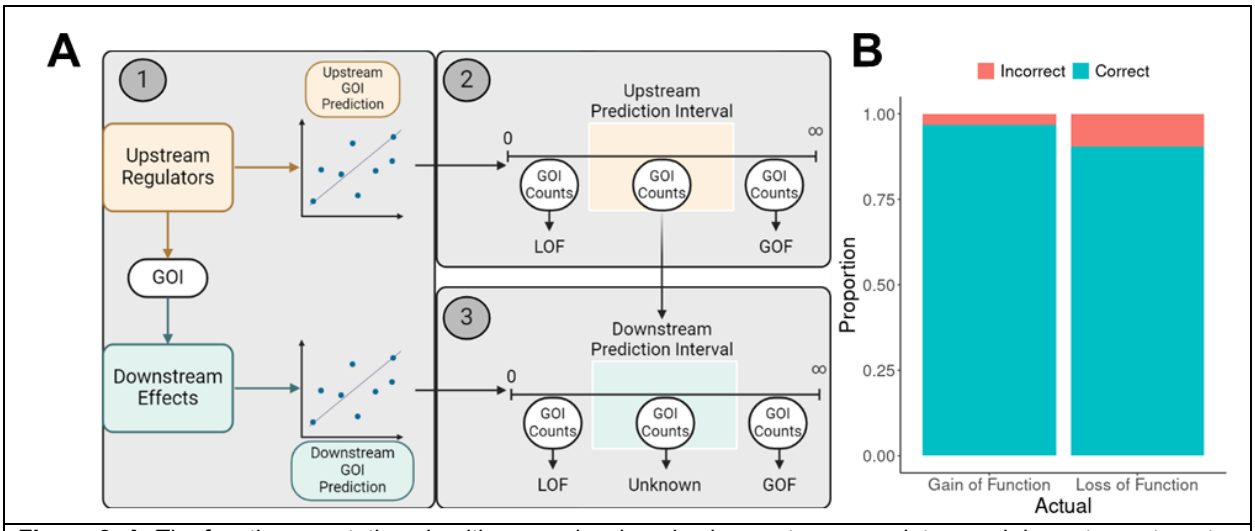

Figure 3: A. The function annotation algorithm was developed using upstream regulators and downstream targets of a given gene of interest from the STRING database. Parrallel linear models over a poisson distribution are built. First the model of the upstream regulators is assessed. Next, if the expression of the gene of interest falls within the prediction interval, the expression gene predicted by the downstream targets is assessed. **B.** The function annotation algorithm was assessed using OncoKB. 291 genomic alterations with annotations in OncoKB were called, 96.8% of the gain of funciton calls (60/62) were accurate and 90.4% of the loss of function calls (207/229).

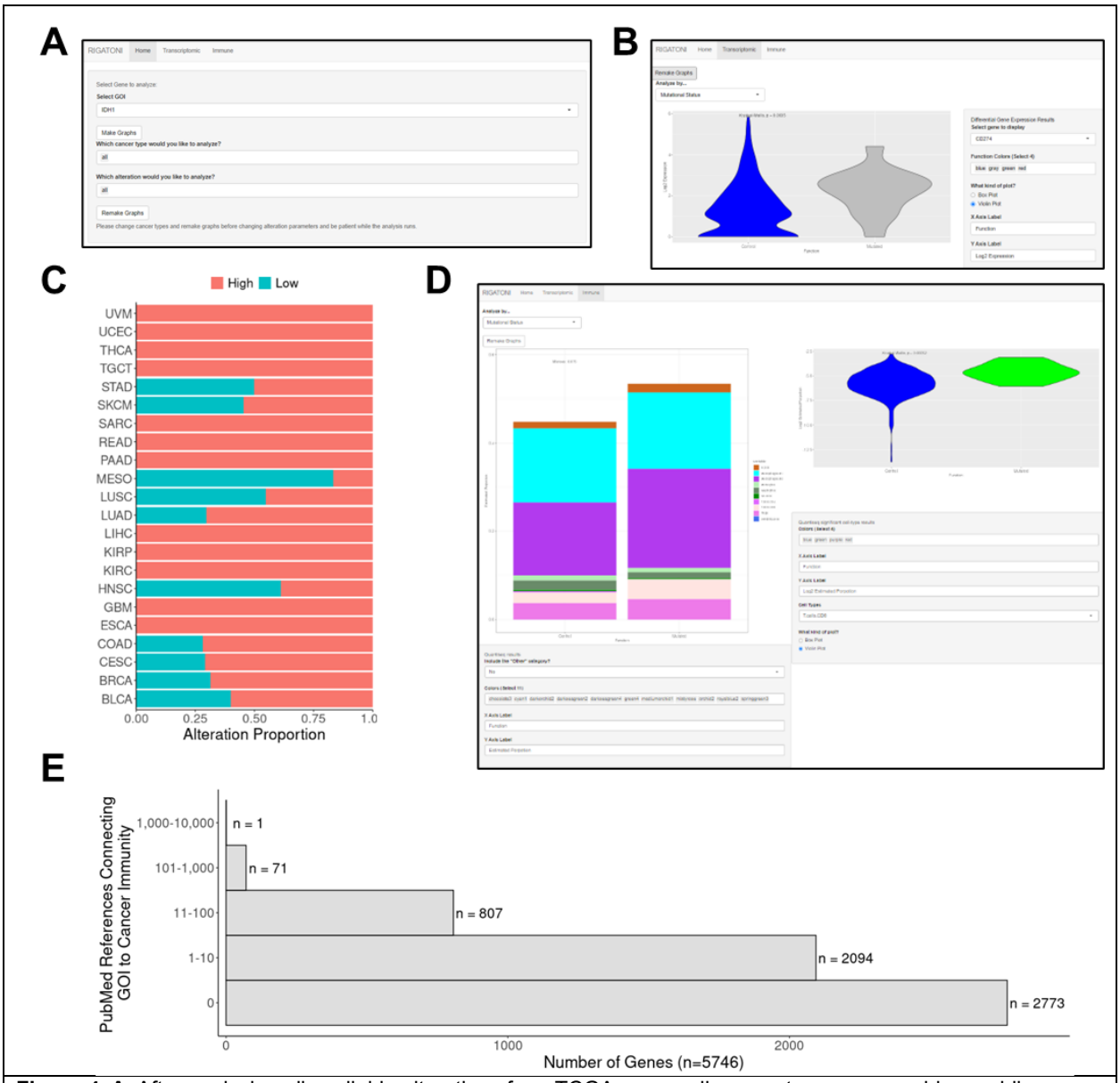

Figure 4. A. After analyzing all available alterations from TCGA across all cancer types, we provide a public website to visualize and explore these results. The user selects a gene of interest and can filter results to types of alterations and cancers. Only positive results will be displayed. **B/D.** Once an alteration is selected for analysis, there are two methods to explore the characteristics of the mutant group. The first is to explore the expression of specific genes under the transcriptome tab. We provide relevant statistical tests and the functionality to change characteristics of the graph such as colors, axis labels, and graph type **(B).** We also analyzed each group using the cellular deconvolution software quanitseqR. These results are available under immune tab. We provide the same functionalities as the transcriptome tab. Estimated proportions of specific cell types can be explored (right) or the wholistic immune microenvironment can be displayed (left) **(D). C.** We discovered alterations in 22 cancer types which are associated with altered immunity. Most of these alterations are associated with a highly infiltrated immune microenvironment. **E.** To assess the novelty of these results, we performed a text mining study of cancer immunity abstracts in PubMed. 84% of the genes with suspected immunologic impacts from TCGA have been mentioned in cancer immunity literature less than 10 times. 72 genes or 1% of genes discovered were mentioned over 100 times.

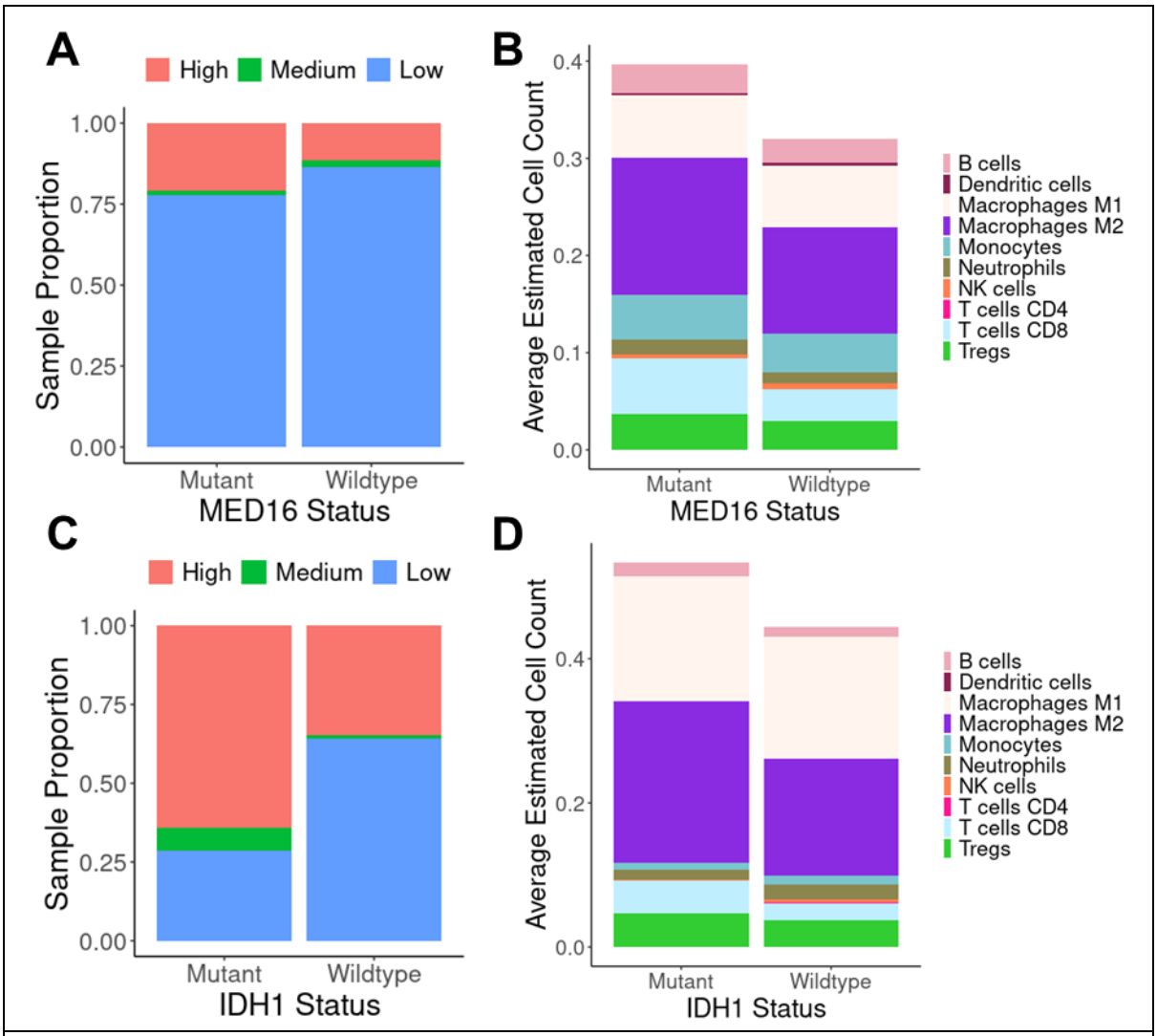

Figure 5. A/B. In TCGA, we evaluated the immune impact of *MED16* loss in melanoma. We see mutant patients display significantly enriched highly infiltrated phenotypes compared to wildtype tumors **(A).** Using quanitseqR, we see enrichment of CD8+ T cells and overall increase in immune cells in mutant samples compared to wildtype **(B). C/D.** In TCGA, we evaluated the immune impact of *IDH1* loss in lung adenocarcinoma. We see mutant patients display significantly enriched highly infiltrated phenotypes compared to wildtype tumors **(C).** Using quanitseqR, we see enrichment of CD8+ T cells and overall increase in immune cells in mutant samples compared to wildtype **(D).** We also see decreased NK cells and neutrophils in IDH1 mutant tumors **(D).**