Using Transcriptional Signatures to Find Cancer Drivers with LURE

David Haan, Ruikang Tao, Verena Friedl, Ioannis N Anastopoulos, Christopher K Wong, Alana S Weinstein, Joshua M Stuart[†]

Dept. of Biomolecular Engineering and UC Santa Cruz Genomics Institute
University Of California, Santa Cruz, Santa Cruz, CA 95064, USA
[†]*E-mail: jstuart@ucsc.edu*

Cancer genome projects have produced multidimensional datasets on thousands of samples. Yet, depending on the tumor type, 5-50% of samples have no known driving event. We introduce a semi-supervised method called Learning UnRealized Events (LURE) that uses a progressive label learning framework and minimum spanning analysis to predict cancer drivers based on their altered samples sharing a gene expression signature with the samples of a known event. We demonstrate the utility of the method on the TCGA Pan-Cancer Atlas dataset for which it produced a high-confidence result relating 59 new connections to 18 known mutation events including alterations in the same gene, family, and pathway. We give examples of predicted drivers involved in TP53, telomere maintenance, and MAPK/RTK signaling pathways. LURE identifies connections between genes with no known prior relationship, some of which may offer clues for targeting specific forms of cancer. Code and Supplemental Material are available on the LURE website: https://sysbiowiki.soe.ucsc.edu/lure.

Keywords: Machine Learning, Cancer, Genomics, Drivers, Gene Expression

1. Introduction

Cancer is a genetic disease caused by mutation and selection in somatic cells. Mutations in normal cells are usually repaired or result in apoptosis, whereas in cancer cells mutations accumulate, leading to uncontrolled growth and tumorigenesis. There are two broadly defined types of mutations: drivers and passengers. Tumors contain around 2-5 driver mutations that cause and accelerate cancer, and about 10-200 passenger mutations which are byproducts of thwarted DNA repair mechanisms.[1] Driver mutations define some characteristics of the tumor and may offer therapeutic targets, yet identifying them amid the myriad of passengers remains a challenge.

The Cancer Genome Atlas (TCGA) is a publicly accessible dataset of cancer samples cataloguing mutations, mRNA, miRNA, DNA methylation, copy number variation, and protein expression for roughly 11,000 patients across 33 cancer types.[2] The identification of driver mutations played a key role in many TCGA analyses. For example, the TCGA study of papillary thyroid carcinoma identified two tumor subtypes characterized by different driver mutations. One subtype harbored mutations in BRAF and the other subtype in RAS genes such as KRAS, NRAS, or HRAS. This study identified at least one driver mutation in about

95% of the samples, leaving about 5% of the samples without known driver mutations.[3] To address situations like this, we present a semi-supervised pattern recognition tool, Learning UnRealized Events (LURE), that finds new drivers sharing molecular signatures with known drivers.

There are several existing methods that distinguish driver from passenger mutations[4] including those that make use of mutual exclusivity of events (e.g. CoMEt and MEMo), and those that use copy number and/or transcriptional profiles together with prior pathway information (e.g. Paradigm-Shift and DriverNet). Three approaches most similar to LURE—REVEALER, Onco-GPS, and one based on pathway-level expression signatures—are discussed in Supplemental Methods. In contrast, LURE uses mRNA data to iteratively identify mutations in "driver-unknown" samples with similar expression signatures to known drivers, thereby implicating a novel set of mutations as possible drivers.

Several studies have built gene expression signatures to identify samples with certain driver events. For example, studies have identified a TP53 gene expression signature as a reliable and independent predictor of disease outcome in breast cancer.[5,6] In addition, in patients with epithelial ovarian cancer, a BRCAness gene expression signature is just as predictive of chemotherapy responsiveness and outcome as mutation status.[7] While creating signatures as prognostic markers to guide treatment is important in a clinical setting, there has been little work using such gene expression signatures to find related mutational events. LURE identifies gene expression signatures across the 723 COSMIC cancer genes[8] and then uses iterative semi-supervised learning to discover potentially related events.

2. Method

LURE associates events by finding similar molecular signatures among the samples in which the events occur. An event here is a particular type of alteration that affects a single gene, such as a focal deletion, a missense mutation, a truncating mutation, a gene fusion, and so forth. In this study, mRNA sequencing-based expression data is used to generate signatures, although other data choices are possible (e.g. microRNA expression or DNA methylation). LURE finds related events by training a classifier using the samples containing a known driver mutation (the "bait"). It then applies the classifier to find "target" samples defined as high-scoring samples lacking bait alterations. In classic machine learning, the target samples would be considered false-positives. However, in this case, they offer an opportunity to find drivers because the comprehensive collection of TCGA mutation calls can be searched to find other events that significantly coincide with the target samples. Any such event (the "catch") is associated with the starting bait and provides a new set of labels for retraining a more accurate classifier in subsequent rounds. Thus, the approach can be viewed as a specific kind of semi-supervised learning in which the possible labels are constrained to the event samples. LURE concomitantly identifies related events, expands the labels, and improves the classification accuracy using the iterative procedure described below.

LURE's first step establishes a known driver mutation as the initial bait (Step 0; Fig. 1A). Next, LURE trains a logistic regression classification model using gene expression as features and bait mutation status as the label to be predicted (Step 1). The method retains only those baits that yield accurate models determined by cross-validated area under the precision-recall curve (PR AUC). LURE then uses the classification model to score each sample in the dataset (Fig. 1A,B). Despite the inherent bias towards overfitting because the classifier is applied to samples used for training, some samples without the bait alteration may still receive a high classifier score, thus producing a sufficiently large set of target samples to identify and associate new catch events.

To leverage this new information, LURE runs a Sample Set Enrichment Analysis (SSEA) to test whether the samples of an event are associated with the target samples (Step 2, Fig. 1A). For SSEA, the GSEAPreranked tool[9] is used, which performs a Kolmogorov-Smirnov test to determine if samples with an event significantly sort toward the top of the list when all non-bait samples are ranked by their classification scores. LURE establishes a set of "Catch" events as those with significant SSEA association scores ($p < 0.05$, $FDR < .25$, number of events > 3). For each catch event, LURE then combines the positive samples for both bait and catch events into a new, intermediate bait event and trains a new classifier for these samples (Return to Step 1; Fig. 1A). Cross-validation is run for the new classification model and the PR AUC results are compared to the initial classifier to ensure the model improves when including the new positive samples (Student's t-test, $t > 0$). In addition, LURE tests the new classifier against a null model created by adding the same number of randomly chosen catch samples to the true positives and running cross-validation (Student's t-test, $p < 0.05$). After establishing that the new event both improves the original classifier and significantly outperforms a random background distribution, the new classifier is rerun on all samples to search for the next set of catch events (Fig. 1A,C). In this manner, LURE builds an "Event Discovery Tree" (EDT) by connecting catch events to bait classifiers that discovered them. It recursively builds classifiers at each node (Fig. 1D). The tree recursion stops when no further events are found by SSEA or the classifier's performance no longer increases.

The above procedure can lead to the association of numerous events at every level of the EDT. By chance, mutations in passenger events can occur in the target samples, especially as a result of highly unstable genomes. Our task then is to distinguish such passengers from true, but as-yet unknown, drivers. To do this, LURE identifies a set of events, where each event occurs in at least some target samples that contain no other event. Intuitively, these events may be the sole explanation for driving the signature in the target samples. To make this concrete, consider the hypothetical scenario in which two events—X and Y—have been included in the EDT. Further, imagine that several samples have mutations only in event X. In contrast, no such samples exist for event Y; i.e. Y only occurs in samples that have at least one other event from the EDT. LURE assumes that X is more likely to be a driver than Y because X is the only event that can explain the presence of the signature in several samples.

Thus, in the last step (Step 3; Fig. 1A), LURE computes a minimal set of events in the catch that best span the catch samples. To this end, LURE builds a final classification model from the union of all events in the EDT and uses this model to identify a final set of samples scored as positive by the final classifier (score ¿ 0.5). It then takes all events in the EDT that span these positive-scoring samples (Fig. 1E) and runs the Vazirani approximation to the set cover algorithm[10] as implemented in the RcppGreedySetCover R Package. This step identifies

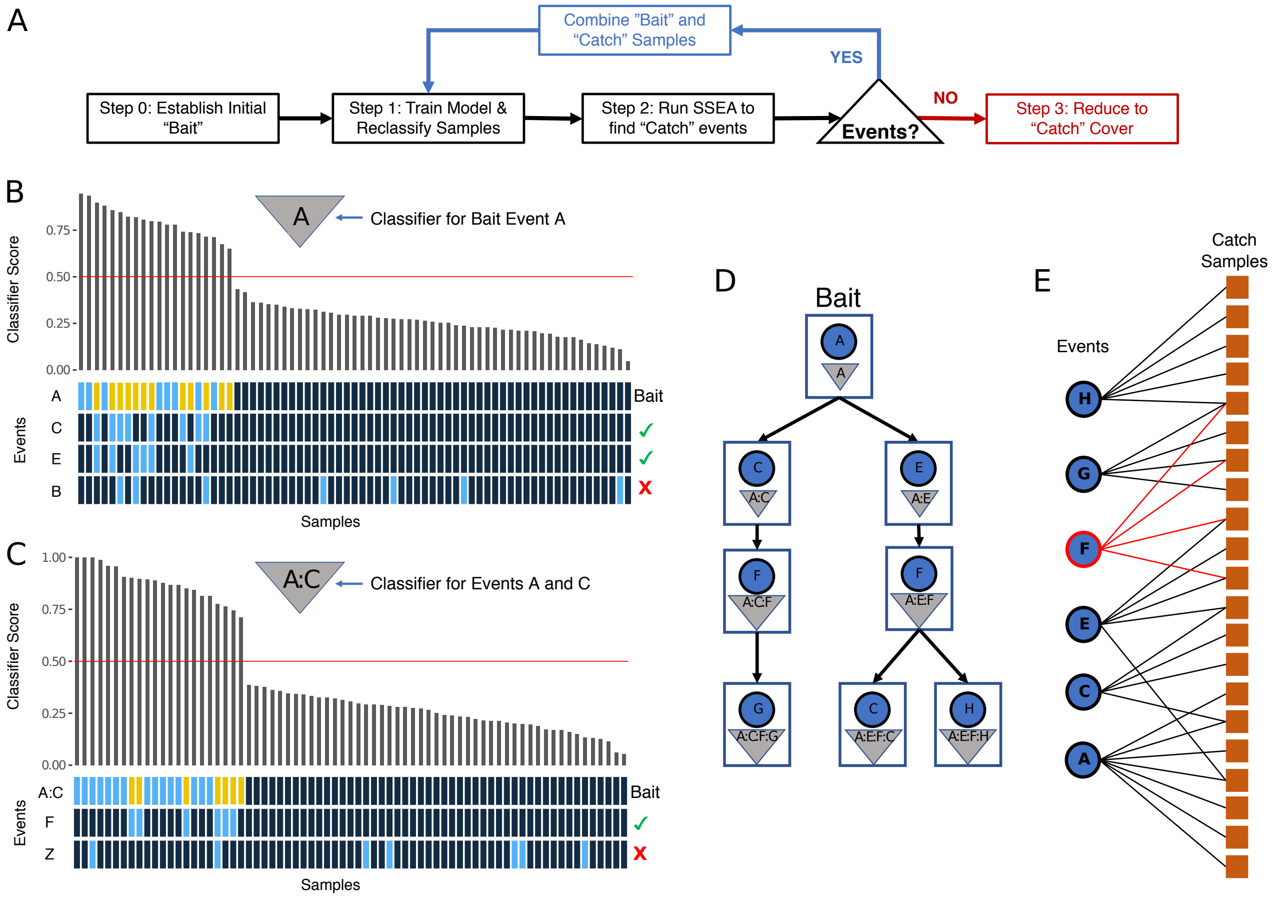
A
Combine "Bait" and "Catch" Samples
YES
NO
Step 0: Establish Initial "Bait"
Step 1: Train Model & Reclassify Samples
Step 2: Run SSEA to find "Catch" events
Events?
Step 3: Reduce to "Catch" Cover
B
Classifier Score
0.75
0.50
0.25
0.00
A
Classifier for Bait Event A
Events
A
C
E
B
Bait
✓
✓
X
Samples
C
1.00
0.75
0.50
0.25
0.00
A:C
Classifier for Events A and C
A:C
F
Z
Bait
✓
X
Samples
D
Bait
A
A
C
A:C
E
A:E
F
A:C:F
F
A:E:F
G
A:C:F:G
C
A:E:F:C
H
A:E:F:H
E
Catch Samples
Events
H
G
F
E
C
A

Fig. 1. **LURE Method Overview. (A) LURE Method Flowchart. (B) LURE Graph of Bait Event A.** Scores (y-axis) assigned to samples (x-axis) by the logistic regression model trained to recognize samples harboring event A (triangle). Red line indicates 0.5 score cutoff for positive predictions. Presence of mutations for each sample are indicated below the barplot (blue, present; black, absent). Positive samples lacking alterations in the bait event define the set of "Target" samples (yellow). Events C and E are found by Sample Set Enrichment Analysis (SSEA) to be present in samples with high classifier scores (green check marks on right). **(C) LURE Graph of Combined Bait Event A:C.** Results from a classification model trained to recognize samples with either event A or C. New event F is found by SSEA to be associated with the A:C classifier. **(D) LURE Event Discovery Tree.** Each node represents an event (node label) discovered using the classification model (triangle) of its parent. The new classification model is trained using samples with the new event as well as any samples with events along the path to the root of the tree. The blue circles within each node represent the newest event added to the model. **(E) LURE Catch Cover.** A bipartite graph is constructed using all events from the Event Discovery Tree (left nodes) and all samples found to have scores greater than 0.5 with the final combined classification model as well as an alteration in one of the catch events ("Catch Samples," right nodes). If an event is present in a sample, their nodes are connected in the graph. The set cover algorithm is run to obtain the minimal set of events that span all samples. Event F illustrates an example of an event that would be removed by the algorithm because the samples with event F contain other events that explain the observed signature. The final event set is retained as the "Catch Cover."

the minimum set of events, the "Catch Cover," that collectively occur in all catch samples, removing events with completely overlapping samples in order to enrich for events that are uniquely responsible for the expression signature.

More details about LURE and its implementation are available in the Supplemental Methods section on the LURE website: https://sysbiowiki.soe.ucsc.edu/lure.

3. Results

3.1. *Datasets*

All data were obtained from the TCGA Pan-Cancer Atlas collection.[2] These data included batch-corrected and normalized gene expression data, as well as single nucleotide variant (SNV) calls determined from the TCGA's MC3 consensus analysis, and gene-associated copy number data from GISTIC2. To consider all transcripts that might increase prediction accuracy, LURE uses the full set of features in the mRNA expression data.

3.2. *Positive Controls for LURE*

Isocitrate Dehydrogenase 1 (IDH1) is one of three IDH enzymes, which when mutated causes hypermethylation and subsequent altered gene expression in gliomas.[11] In order to test LURE's ability to discover known "catch" events, we created a test set of bait and catch events using IDH1 missense mutations in the TCGA Lower Grade Glioma (LGG) samples. Of the 210 LGG samples with an IDH1 missense mutation, we created an initial bait with 150 samples and three sets of 20 samples as potential catch events. LURE identified all three of the held-out events in the Catch Cover, as well as the IDH2 missense mutation event. IDH2 is another IDH isozyme, and a mutation in IDH2 has the same oncogenic effect as an IDH1 mutation[12] (Supp. Fig. 1A). We also tested a similar method called REVEALER on this positive control set. We found that REVEALER was unable to identify the held-out samples (Supp. Fig. 2).

Splicing Factor 3b Subunit 1 (SF3B1) is a well-known splicing factor which is recurrently mutated in many tumor types, including Uveal Melanoma (UVM).[13] Missense mutations in SF3B1 lead to aberrant splicing and a unique gene expression signature.[14] We used SF3B1-mutated tumors in the TCGA UVM dataset as a positive control. Of the 80 UVM samples, 18 samples have missense mutations in SF3B1. We created an initial bait using 8 of these SF3B1 mutant samples and left out the remaining two sets of 5 samples for discovery. LURE re-discovered both withheld sets correctly, collecting all of the SF3B1 missense events in the Catch Cover (Supp. Fig. 1B).

3.3. *LURE on the TCGA Pan-Cancer Atlas Dataset*

In order to look for novel associations among genes already implicated in cancer, we ran LURE for each tumor type in the TCGA Pan-Cancer Atlas dataset, restricting both baits and catches to mutation events involving one of the 723 COSMIC genes.[2,8] We created tumor type-specific classification models to avoid any confounding effects due to tissue-specific expression and mutation patterns, we created tumor type-specific models by considering each bait for classification within each tumor type. Bait events were created for distinct mutation types including missense mutations, truncating mutations, homozygous focal point copy number deletions, and gene fusions for all COSMIC genes.

We trained logistic regression models on the resulting 3,053 bait/tumor type combinations. We also tested both random forest and neural network models, but found the logistic regression model to consistently score higher for the majority of mutations (Supp. Fig. 3). Due to the imbalanced nature of the data, precision and recall, rather than overall accuracy, was used to measure classifier accuracy to emphasize the ability of a classifier to detect bait samples that are often highly underrepresented.[15] For the TCGA data, we found that using a minimum of five bait samples was just enough to indicate that an accurate classifier represents a robust signature rather than due to chance (Supp. Fig. 4A). However, in order to limit the number of putative false positive results, we restricted the number of bait classifiers by considering only those with at least 10 positive samples. To limit the results yet further and produce a confident set of results amenable for interpretation, we also limited baits to those with PR AUC > 0.5, precision > 0.4, and recall > 0.75 (Supp. Fig. 4B). Since the objective of LURE is to identify and reclassify target samples, the precision is set to be more lenient than the recall. Among the bait classifiers passing these thresholds, the most common bait gene across all tumor types was TP53, and the tumor types with the highest number of passing bait classifiers were Lower Grade Glioma (LGG), Thyroid Carcinoma (THCA), and Prostate Adenocarcinoma (PRAD) (Supp. Fig. 5).

After filtering the models, we ran LURE with the 81 remaining classifiers as bait, using missense mutations, truncating mutations, splice site mutations, gene fusions, focal point copy number amplifications, and homozygous deletions of COSMIC genes as possible catches. LURE found significant bait-catch associations for 35 of the 81 baits tested. Adding catches to each initial bait event increased the classifier PR AUC by various amounts across the different baits (Supp. Fig. 7A). Tumor type was evenly distributed among the baits and SNVs dominated the bait mutation type (Supp. Fig. 7D). The most common gene among the 35 baits was again TP53 (Supp. Fig. 7E).

Among the high-confidence results with a final classifier PR AUC > 0.8, 14 of 59 bait-catch associations were between events involving the same gene, such as TP53 truncating, splice site, and missense mutations, in various tumor types (Fig. 2A). There were four associations within the same gene families, e.g. IDH1/2 or the RAS protein family. In addition, we identified gene fusion event partners in BRAF and RET that associated with a BRAF missense mutation in THCA. For 20 of the 59 high confidence bait-catch associations, both genes were members of the same signaling pathway (excluding pathway gene sets with > 1000 genes).[16]

Superimposing the EDTs obtained within each tissue type into a compendium of bait-catch associations we refer to as the "Event Net," reveals several pathway-oriented findings with support in multiple tumor types (Supp. Fig. 8). In particular, four canonical pathway-oriented findings emerge: EGFR signaling, TP53/TP63/SMAD4, PI3K, and MAPK/RKT (Fig. 2B). LURE identified interesting associations for PTEN, in particular between PTEN and CTNNB1, a connection supported by recent research which suggests PTEN plays a role in regulating the subcellular localization of β-catenin.[18] Another striking LURE association is between PTEN and EGFR, which is consistent with recent findings suggesting that PTEN

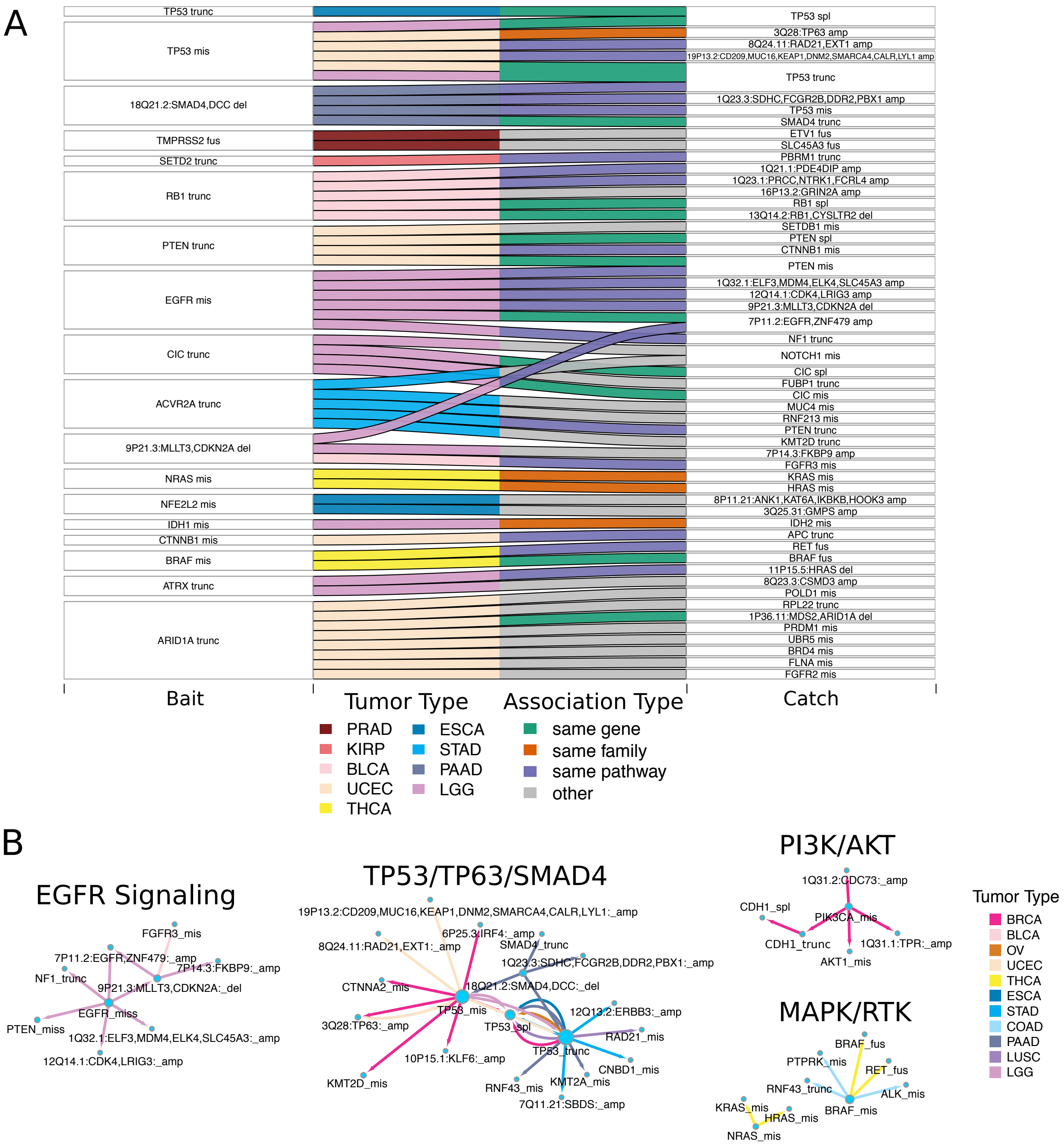

Fig. 2. **LURE Results for the Pan-Cancer Atlas Dataset. (A) High Confidence Bait-Catch Associations.** Sankey diagram shows the high confidence bait-catch associations for the 18 baits with a final classifier PR AUC > 0.8. Bait gene and mutation type are shown on the left; catch gene and mutation type on the right. Each horizontal flow connection represents an association between the bait and catch event found by LURE. The left half of each flow bar is colored by tumor type in which the association was found. The right half of each flow bar is colored by the association type. **(B) LURE "Event Net" Showing Selected Pathway Associations.** Cytoscape[17] visualization of a subset of LURE associations in the TCGA Pan-Cancer dataset grouped by pathway. Each node represents an event. Directed edges represent an association and the direction of the LURE discovery (bait to catch). The color of each edge represents the tumor type in which the association was found. Refer to Supp. Fig. 8 for full Cytoscape visualization of all TCGA Pan-Cancer Atlas dataset results.

regulates EGFR signaling.[19] These LURE associations for PTEN reveal crosstalk between pathways and provide further evidence that alterations in PTEN influence EGFR signaling and β-catenin signaling.

LURE's TCGA Pan-Cancer Atlas analysis associated a total of 1,216 samples with 35 bait events through 2,845 connections to 53 catch events. Because baits were created for different alteration types in the same gene, as opposed to one bait for any alteration in a gene, the method was able to identify associations between different alteration types within the same gene as well as infer the functional impact of alterations. For example, α-thalassemia mental retardation X-linked (ATRX), a gene recurrently mutated in LGG, only has an oncogenic effect with a loss-of-function mutation—either a truncating mutation or copy number loss—whereas a missense mutation may not have an oncogenic effect.[12] Different mutation types showed different performance as baits, with gene fusions generally performing strongest followed by deletions, truncating mutations, and missense mutations (Supp. Fig. 6). This indicates that many fusion genes have a consistent effect on gene transcription, whereas missense mutations often exhibit inconsistent or no phenotypic effects.

We attempted to measure the degree to which LURE expands the list of known drivers in the TCGA dataset. While estimating this quantity is difficult, we approximated it by counting up the frequency with which LURE connects catch events to baits between genes without any prior functional relationship. To this end, we found that 6.9% (10,271) of the total events considered were associated with a transcriptional signature either as a bait, catch, or both (Supp. Fig. 9). Over a third of these events (34%; 3497) are from catches that were not themselves used as a bait event. We categorized these catch events according to whether they connect the same gene (20%, 683), family (2%; 77), pathway (25%; 890), or no known relationship (53%; 1847) with the original bait. Thus, the numbers demonstrate that among these 6.9% of events with a transcriptional signature, 18% (1847) are newly implicated by LURE linking one signature to another.

3.4. *LURE Finds New Drivers of the Alternative Lengthening of Telomeres Pathway in Sarcomas*

Tumor cells must employ Telomere Maintenance Mechanisms (TMMs) to extend their telomeres in order to multiply rapidly and avoid senescence.[20] To date, there are two known mechanisms tumor cells use to avoid telomere erosion: the overexpression of telomerase, an enzyme which extends telomeres, and the Alternative Lengthening of Telomeres (ALT) pathway. The vast majority of tumors overexpress telomerase in some way, whereas a small portion (10-15%) use ALT.[21] ALT-positive (ALT$^+$) samples lengthen telomeres through homologous recombination, mediated by loss-of-function mutations in the ATRX and DAXX genes.[22] Approximately 80% of tumors with ALT harbor mutations in ATRX or DAXX,[23] leaving 20% with no known driver. Using LURE, we sought to identify new driver mutations of the ALT pathway using gene expression signatures of samples harboring ATRX loss-of-function mutations. Sarcomas and Lower Grade Gliomas (LGG) have the highest prevalence of ALT$^+$ samples, and ATRX is recurrently mutated in these diseases. We therefore chose these tumor types in which to search for new drivers of the ALT pathway.[22] We restricted our gene set to a manually-curated set of genes associated with telomere maintenance derived from the TelNet database[24] and focused on looking for associations with ALT. Since TP53 is commonly mutated in ALT$^+$ samples and TP53 mutations are not known to be sufficient to cause ALT,[25] we excluded TP53 events from the possible catches in order to identify novel ALT drivers.

Using ATRX truncating mutations as bait, LURE identified four associated mutations in sarcomas (Fig. 3A). While we expect ATRX truncating mutations to associate with an ATRX copy number deletion,[12] the deletions found in RB1 and SP100 are novel. We suggest that the expression signature LURE identified in this analysis is classifying ALT$^+$ TMM samples, and that the associated alterations are possibly driving the TMM. Previous work has associated RB1 alterations with long telomeres in the absence of TERT mutations and ATRX inactivation.[26] In addition, mouse models have revealed that the knock-out of Rb-family proteins causes elongated telomeres.[27] LURE also identified SP100 deletions as an ALT driver, and while SP100 deletions have not been directly reported to be involved in ALT, overexpression of SP100 in ALT$^+$ cell lines has resulted in suppression of ALT characteristics.[28] We therefore suggest that an SP100 deletion may lead to unhindered ALT TMM activity. To further investigate the subset of LURE-classified ALT$^+$ samples, we performed a survival analysis and found that the ALT$^+$ samples show a worse prognosis (p = 0.068) (Fig. 3B), consistent with recent research.[29]

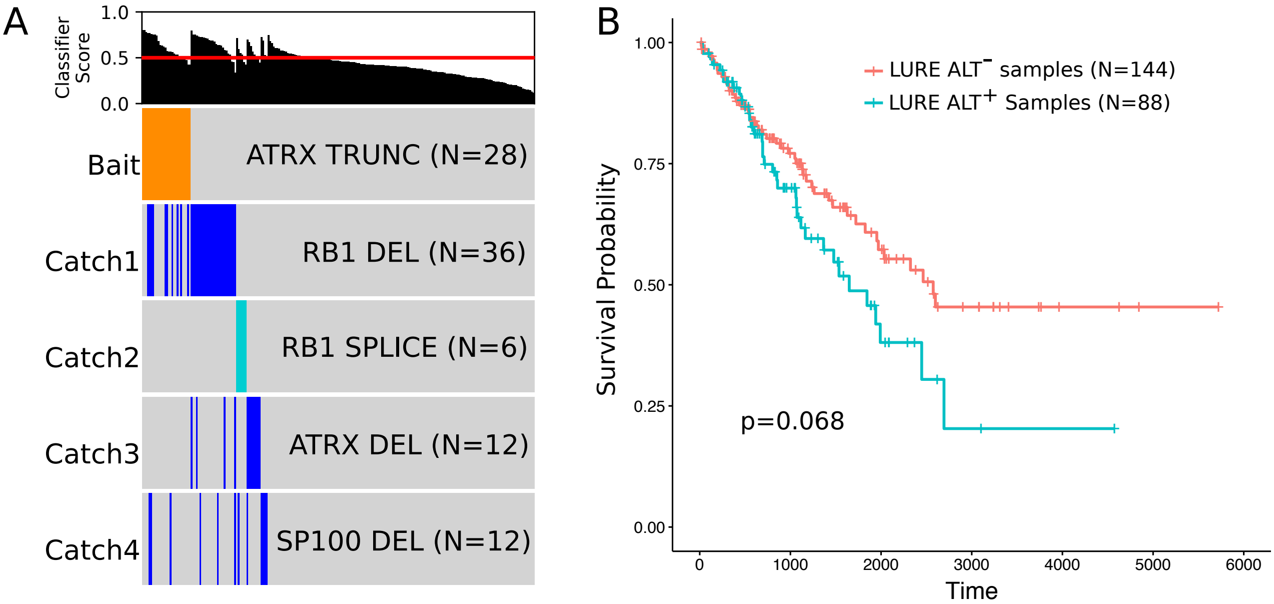

Fig. 3. **LURE Results for the Alternative Lengthening of Telomeres (ALT) in Sarcomas. (A) LURE ALT Graph** Using a known driver for ALT (ATRX truncating mutations) as bait in Sarcomas (SARC), LURE finds four catch events: copy number deletions of ATRX, RB1, and SP100, as well as RB1 splice site mutations. **(B) LURE ALT Survival Plot** Kaplan-Meier survival plot dividing SARC samples by ALT classification using the final Catch Cover classifier from LURE. LURE ALT$^-$ samples have classification scores < 0.5, ALT$^+$ samples ≥ 0.5.

In LGG, again using ATRX truncating mutations as bait, LURE found similar results identifying associations with other ATRX alterations such as ATRX deletions, splice site mutations, and missense mutations (Supp. Fig. 10). While ATRX missense mutations are not generally thought to be ALT drivers, we found that ATRX missense mutations were in fact associated with truncating mutations, suggesting a loss of function role for some of these missense mutations.[12] Together, these findings implicate new single and combinations of driver mutations required for the initiation of the ALT telomere maintenance mechanism and suggest potential therapeutic targets.

3.5. *LURE Identifies Associations within the MAPK/RTK Signaling Pathway*

Oncogenic mutations of the HRAS, NRAS, or KRAS genes are frequently found in human tumors, altering the control of cellular proliferation, differentiation, and survival. Oncogenic mutations in a number of other upstream or downstream components of the MAPK/RTK signaling pathway, including membrane receptor tyrosine kinases (RTKs) and cytosolic kinases, have been recently detected in a variety of cancer types.[30] Oncogenic RAS mutations and other mutation events within the MAPK/RTK signaling pathway are often mutually exclusive, indicating that the deregulation of Ras-dependent signaling is essential for tumorigenesis.[30] Previous studies have shown that tumor samples harboring Ras protein mutations have a unique gene expression signature, and Ras-dependent samples can be more accurately defined by this signature than by mutation status alone.[31] Building on this knowledge, we were able to use LURE not only to train an accurate Ras-dependent classifier, as was done in Way et al.,[32] but also to identify new alterations which may be activating the MAPK/RTK signaling pathway in samples without a Ras protein mutation. We ran LURE using genes known to be involved in the MAPK/RTK signaling pathway[33] as baits. We considered baits with the following alterations: missense mutations, truncating mutations, homozygous focal point copy number deletions, splice site mutations, and gene fusions. We did not restrict our catch set to COSMIC genes as we were looking for genes not previously implicated in cancer. We ran LURE for the 23 bait event classifiers which scored greater than 0.5. The resulting Event Net revealed known as well as novel associations (Fig. 4A). One interesting association found by LURE in Head and Neck Squamous Cell Carcinomas (HNSC) is between HRAS missense mutations and focal deletions of the 2q23.3 locus (Fig. 4B). The samples were almost mutually exclusive for these events, with only one sample having both a 2q23.3 deletion and an HRAS alteration and no samples having alterations in either KRAS or NRAS. Among the 61 genes in the 2q23.3 locus is CHST11, which has been shown to regulate MAPK/RTK pathway activity in hepatocellular carcinoma.[34] We suggest that in the absence of an HRAS mutation, MAPK signaling may be activated by a deletion of the 2q23.3 locus in HNSC.

4. Discussion

We described the Learning UnRealized Events (LURE) method that leverages the availability of multi-omic cancer datasets to identify driver events using signatures constructed from other feature data (in this study, mRNA expression). LURE is related to semi-supervised machine learning approaches, but has a slight shift in emphasis. Rather than focusing on labeling

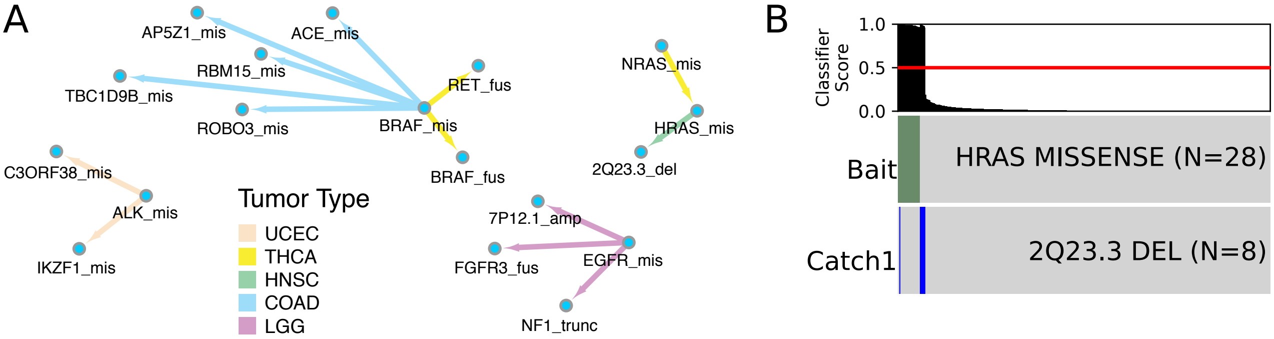

Fig. 4. **LURE RAS/MAPK Signaling Pathway Analysis. (A) LURE "Event Net" Showing Associations with MAPK/RTK signaling pathway genes.** Each node represents an event. Directed edges represent associations and the direction of the LURE discovery (bait to catch). The color of each edge represents the tumor type in which the association was found. **(B) LURE Graph of HRAS in HNSC.** Using HRAS missense mutations as bait in Head and Neck Squamous Cell Carcinomas (HNSC), LURE found a deletion catch event in the 2q23.3 locus.

unknown samples to optimize performance, LURE instead uses the mutual compatibility of label sets to form a coherent classifier as an indicator that the label sets themselves are related. Thus, the accuracy of the classifier is only important to the extent that it produces a signature that can detect strong associations. LURE takes advantage of the "dark matter" (false-positives in gene expression-based classifiers) to search for associated events. LURE outperformed REVEALER, another signature-based method, on the IDH1 positive control test. A comparison to other methods can also be found in the supplement (Supp. Methods).

Some limitations of this study could be addressed to expand the set of predicted drivers. First, we used a highly conservative set of parameters to analyze the TCGA dataset in order to control the set of false positives and to highlight the most confident connections including the minimum number of bait samples, a baseline accuracy for bait classifiers, a cutoff for the SSEA tests, the allowed sample overlap between events, and restricting catches to COSMIC genes. One could run a less stringent LURE analysis to produce many more catch events that could be further filtered using criteria such as the recurrence of the same catch found in multiple tumor types. Second, we only considered gene expression data as features for building signatures, which could be driven by the presence of confounding covariates such as subtypes and microenvironment factors. Multiple other data types are available, such as DNA methylation, to increase the ability to find associations among events. Third, the use of SSEA restricts the consideration of catch events to those occurring in at least four target samples in order to pass the multi-test FDR cutoff. However, it might be worthwhile to add events to the catch even if they cover only one additional target sample. Methods that consider sets of mutually exclusive events simultaneously might offer some advantage.

Using signatures to associate events will extend the list of known drivers that may ultimately identify targets for precision medicine. LURE identified an intriguing set of new candidate drivers based on the Pan-Cancer Atlas dataset. Most of the found associations were between events of the same gene (e.g. TP53), gene family (e.g. IDH, RTK), or biological pathway (e.g. PI3K). In addition to associating known events across the Pan-Cancer Atlas dataset, LURE identified 1847 events of unknown significance to known pathways such as putative drivers of telomere maintenance and MAPK/RTK signaling. The collection of these events provides possible actionable clues for cancer patients. For example, LURE found that deletions in 2q23.3 in head-and-neck cancers are strongly associated with RTK signaling. Whether such a focal deletion could be used as a novel biomarker for treating patients with an RTK inhibitor, such as gefitinib for EGFR, remains to be seen.

Acknowledgments

D.H., V.F., I.N.A., C.K.W., A.S.W., and J.M.S. were supported by grants from NCI (U24-CA143858, 1R01CA180778, NHGRI (5U54HG006097), and NIGMS (5R01GM109031).

References

1. B. Vogelstein et al., *Science* **339**, p. 1546 (2013).
2. K. A. Hoadley et al., *Cell* **173**, p. 291 (2018).
3. N. Agrawal et al., *Cell* **159**, p. 676 (2014).
4. A. Gonzalez-Perez et al., *Nature Methods* **10**, p. 723 (2013).
5. S. Takahashi et al., *Cancer Science* **99**, p. 324 (2008).
6. S. Yamaguchi et al., *Oncotarget* **9**, p. 14193 (2018).
7. P. A. Konstantinopoulos et al., *Journal of Clinical Oncology* **28**, p. 3555 (2010).
8. J. G. Tate et al., *Nucleic Acids Research* **47**, p. D941 (2018).
9. A. Subramanian et al., *Proceedings of the National Academy of Sciences* **102**, p. 15545 (2005).
10. V. V. Vazirani, *Approximation Algorithms* (Springer Science & Business Media, 2013).
11. S. Turcan et al., *Nature* **483**, p. 479 (2012).
12. Cancer Genome Atlas Research Network, *New England Journal of Medicine* **372**, p. 2481 (2015).
13. A. G. Robertson et al., *Cancer Cell* **32**, p. 204 (2017).
14. S. Alsafadi et al., *Nature Communications* **7**, p. 10615 (2016).
15. T. Saito and M. Rehmsmeier, *PLOS One* **10**, p. e0118432 (2015).
16. D. Merico et al., *PLOS One* **5**, p. e13984 (2010).
17. P. Shannon et al., *Genome Research* **13**, p. 2498 (2003).
18. A. Persad et al., *Genes & Cancer* **7**, p. 368 (2016).
19. S. R. Shinde and S. Maddika, *Nature Communications* **7**, p. 10689 (2016).
20. J. W. Shay and W. E. Wright, *Nature Reviews Molecular Cell Biology* **1**, p. 72 (2000).
21. J. D. Henson and R. R. Reddel, *FEBS Letters* **584**, p. 3800 (2010).
22. C. M. Heaphy et al., *The American Journal of Pathology* **179**, p. 1608 (2011).
23. C. A. Lovejoy et al., *PLOS Genetics* **8**, p. e1002772 (2012).
24. D. M. Braun et al., *BMC Genetics* **19**, p. 32 (2018).
25. Y.-J. Chen et al., *Cancer Research* **66**, p. 6473 (2006).
26. F. P. Barthel et al., *Nature Genetics* **49**, p. 349 (2017).
27. M. García-Cao et al., *Nature Genetics* **32**, p. 415 (2002).
28. W.-Q. Jiang et al., *Molecular and Cellular Biology* **25**, p. 2708 (2005).
29. R. T. Lawlor et al., *BMC Cancer* **19**, p. 232 (2019).
30. A. Fernández-Medarde and E. Santos, *Genes & Cancer* **2**, p. 344 (2011).
31. A. Loboda et al., *BMC Medical Genomics* **3**, p. 26 (2010).
32. G. P. Way et al., *Cell Reports* **23**, p. 172 (2018).
33. F. Sanchez-Vega et al., *Cell* **173**, p. 321 (2018).
34. H. Zhou et al., *Digestive Diseases and Sciences* **61**, p. 1972 (2016).