Genome analysis

Somatic selection distinguishes oncogenes and tumor suppressor genes

Pramod Chandrashekar1,2, Navid Ahmadinejad1,2, Junwen Wang1,3, Aleksandar Sekulic3, Jan B. Egan3, Yan W. Asmann4, Sudhir Kumar5,6, Carlo Maley2 and Li Liu1,2,3,*

1College of Health Solutions, Arizona State University, Phoenix, AZ, 85004, USA, 2Center for Personalized Diagnostics, Biodesign Institute, Arizona State University, Tempe, AZ, 85281, USA, 3Department of Health Sciences Research & Center for Individualized Medicine, Mayo Clinic Arizona, Scottsdale, AZ, 85259, USA, 4Department of Health Sciences Research, Mayo Clinic Florida, Jacksonville, AZ, 32224, USA, 5Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, PA, 19122, USA and 6Department of Biology, Temple University, Philadelphia, PA, 19122, USA

*To whom correspondence should be addressed.

Associate Editor: Russell Schwartz

Received on August 30, 2019; revised on October 22, 2019; editorial decision on November 7, 2019; accepted on November 12, 2019

Abstract

Motivation: Functions of cancer driver genes vary substantially across tissues and organs. Distinguishing passenger genes, oncogenes (OGs) and tumor-suppressor genes (TSGs) for each cancer type is critical for understanding tumor biology and identifying clinically actionable targets. Although many computational tools are available to predict putative cancer driver genes, resources for context-aware classifications of OGs and TSGs are limited.

Results: We show that the direction and magnitude of somatic selection of protein-coding mutations are significantly different for passenger genes, OGs and TSGs. Based on these patterns, we develop a new method (genes under selection in tumors) to discover OGs and TSGs in a cancer-type specific manner. Genes under selection in tumors shows a high accuracy (92%) when evaluated via strict cross-validations. Its application to 10,172 tumor exomes found known and novel cancer drivers with high tissue-specificities. In 11 out of 13 OGs shared among multiple cancer types, we found functional domains selectively engaged in different cancers, suggesting differences in disease mechanisms.

Availability and implementation: An R implementation of the GUST algorithm is available at https://github.com/liliulab/gust. A database with pre-computed results is available at https://liliulab.shinyapps.io/gust.

Contact: liliu@asu.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

In tumor development, oncogenes (OGs) and tumor-suppressor genes (TSGs) work complementarily to promote and maintain abnormal cell growth (Morris and Chan, 2015; Weinberg, 1994). OGs cause cancers through gain-of-function variants, whereas TSGs operate by loss of function. While there are a few well-known OGs (e.g. RAS) and TSGs (e.g. TP53), it is fast becoming clear that the tumor-enabling activities of a gene is not the same for all types of cancers. Activities of driver genes depend strongly on their cellular contexts because of tissue-specific organizations of cancer pathways (Schaefer and Serrano, 2016; Schneider et al., 2017; Visvader, 2011). Prediction of functional status of genes in different cancer types and cellular contexts is critical for not only understanding tumor biology, but also informing targeted therapies and drug-repurposing (Morris and Chan, 2015; Schneider et al., 2017; Sleihe et al., 2017).

Interestingly, only one computational method (20/20+) are available to predict OGs and TSGs (Tokheim et al., 2016). The 20/20+ is an extension of the 20/20 rule in which OGs have >20% mutations causing missense changes at recurrent positions and TSGs have >20% mutations causing inactivating changes (Vogelstein et al., 2013). However, recurrent missense mutations are not a deterministic feature of OGs because these events can cluster at functionally neutral positions due to high mutational rates (Schaub et al., 2018), and many TSGs harbor hotspots of inactivating missense mutations (Iacobuzio-Donahue et al., 2004; Miller et al., 2015). Meanwhile, random mutational processes may introduce protein-truncating mutations (i.e. nonsense and frame-shifting mutations) into OGs, which increase in frequency via genetic drift with no significant impact on tumor fitness and mislead annotations (Lipinski et al., 2016; Mort et al., 2008; Schaub et al., 2018). Therefore,
conventional ratiometric measures are inadequate to distinguish these two groups of genes.

Because tumor development is an evolutionary process, cells carrying somatic mutations are under natural selection within tumors. The positive selection promotes advantageous genotypes that confer higher fitness to a tumor. The negative selection eliminates genotypes with adverse effects. Neutral evolution lets insignificant genotypes to drift up or down in frequency. In OGs, gain-of-function mutations may be advantageous, while in TSGs, which are expected to be positively selected. In contrast, protein-truncating mutations (e.g., nonsense mutations and frame-shifting mutations) often inactivate an OG and are detrimental to tumor fitness, resulting in negative selection. In TSGs, both protein-truncating mutations and missense mutations can be positively selected when they result in the loss of function. Otherwise, they may drift neutrally or be even under negative selection if they disrupt essential biological functions. For passenger genes (PGs) that do not have significant impact on tumor fitness, we expect that all mutations are under neutral selection (Sun et al., 2017; Williams et al., 2016).

In this study, we tested whether the difference in evolutionary dynamics of missense and truncating mutations has sufficient signal and power to improve the detection of OGs and TSGs beyond that of ratiometric measures. Such contrast is essential to distinguish TSGs deactivated by missense mutations from OGs activated by missense mutations, which is a challenging task for conventional ratiometric measures because hotspots of missense mutations are present in both cases. Furthermore, when activities of a gene vary across cancer types, the direction and magnitude of somatic selection will change accordingly, enabling contextual classification of driver genes.

Our analysis of 10 172 tumor exomes from The Cancer Genome Atlas (TCGA) (Cancer Genome Atlas Research Network et al., 2013) project revealed significant differences in selective patterns of OGs, TSGs and PGs. Based on these patterns, we developed a computational method, named genes under selection in tumors (GUST) that integrates somatic selection of genes in tumor development, molecular conservation during species evolution and conventional ratiometric measures to classify cancer genes in different tissues and organs.

2 Materials and methods

Curation of cancer-type specific functions of driver genes: to test our hypothesis and to train a random forest model, we needed cancer-type specific functional annotations of cancer genes. Because these annotations are not currently available, we conducted manual curation of driver genes, using two lists of genes with complementary information. The first list consisted of 36 OGs, 48 TSGs and 21 genes with dual OG/TSG roles annotated in the cancer gene consensus (CGC, version 87) (Sondka et al., 2018). The tumor-activating or -suppressing roles of these genes have been confirmed with cancer hallmarks in experimental assays and are attributable to coding substitutions or indels (Hanahan and Weinberg, 2000). The second list consisted of 235 computationally predicted driver genes assigned to specific cancer types (Bailey et al., 2018). These predictions were based on a meta-analysis of the TCGA samples with multiple computational programs. These two lists shared 70 genes. We then removed somatic mutations of these 70 genes from the TCGA project (Cancer Genome Atlas Research Network et al., 2013). For a gene to qualify as an OG in a specific cancer type, it needs to be annotated as an OG or a dual-role gene in the CGC, predicted as a driver in the meta-analysis and have an abundance of truncating mutations or missense mutations in the corresponding TCGA tumor samples. For a gene to qualify as a TSG in a specific cancer type, it needs to be annotated as an OG or a dual-role gene in the CGC, predicted as a driver in the meta-analysis, and have an abundance of truncating mutations or missense mutations in the corresponding TCGA tumor samples. For a gene to qualify as a PG in a specific cancer type, it needs to be predicted as a PG in the meta-analysis and shows no mutational hotspots or overabundance of truncating mutations in corresponding TCGA tumor samples. Genes that did not meet these requirements were removed. The final collection consisted of 55 OG annotations, 174 TSG annotations and 304 PG annotations that involved a total of 50 known driver genes and 33 cancer types (Supplementary Table S1).

Somatic selection features: given a gene with somatic mutations reported in a collection of tumor samples, we denote the selection coefficient of missense mutations as \( a \) and the selection coefficient of protein-truncating (nonsense and frame-shifting) mutations as \( b \). To account for differences in mutational rates, we consider seven substitution types: \( 1: A \rightarrow T \) or \( T \rightarrow G \); \( 2: A \rightarrow G \) or \( T \rightarrow C \); \( 3: A \rightarrow T \) or \( T \rightarrow A \); \( 4: C \rightarrow A \) or \( G \rightarrow T \); \( 5: C \rightarrow G \) or \( G \rightarrow C \); \( 6: C \rightarrow T \) or \( T \rightarrow G \) at non-CpG sites, and \( 7: C \rightarrow T \) or \( G \rightarrow A \) at CpG sites, one insertion type and one deletion type. Based on the statistical framework proposed by Greenman et al. (2006), the probability of observing these mutations is a product of multinomial distributions

\[
L(\{s^k, m^k, n^k, i^k, f^k\}) = \prod_{s^k} \left( \frac{(s^k)^{m^k} (a s^k)^{n^k}}{(m^k + n^k + i^k + f^k)!} \right)
\]

where \( s^k, m^k, n^k, i^k \) and \( f^k \) are the observed numbers of synonymous, missense, nonsense, in-frame indel and frame-shifting indel mutations in the \( k \)th rate category, respectively; \( S^k, M^k, N^k, I^k \) and \( F^k \) are the corresponding expected numbers of changes computed by saturated mutations, in which we introduced each possible single nucleotide mutation at one time; and \( s^k = s^k + m^k + n^k + i^k + f^k \) is the total number of observed mutations. The values of \( \log(a) \) and \( \log(b) \) are determined by maximizing the log likelihood \( L \) and constrained within the range of \( \{-5, 5\} \). The sign and absolute value of \( \log(a) \) and \( \log(b) \) indicate the direction and magnitude of somatic selection. Values around 0 indicate neutral somatic evolution. Details of parameter tuning are available in Supplementary Methods and Supplementary Figure S1.

GUST algorithm: GUST is a random forest model that predicts the class label (OG, TSG or PG) of a gene based on 10 features (Supplementary Table S2 and Fig. S2). In addition to the \( \log(a) \) and \( \log(b) \) values, we also compute ratiometric measures to detect mutational hotspots and conservational measures to estimate substitutional rate across species. Specifically, given a gene and a set of somatic missense mutations detected in tumor samples, we applied density estimates with a rectangular kernel and a bandwidth of five protein positions to aggregate closely-spaced mutations into peaks and denoted the highest peak as the summit. To estimate evolutionary conservation of a gene, we downloaded multiple sequence alignments of 100 vertebrate species from the UCSC Genome Browser (Knutson et al., 2002), and computed the substitution rate of each protein position (Kumar et al., 2012; Liu and Kumar, 2013). The average substitution rate over all positions measures the gene-level conservation. The average substitution rate over positions in a summit measures the conservation of a mutational hotspot. For each gene/cancer-type pair in the curated annotations, we retrieved somatic mutations from the corresponding TCGA tumor samples and computed values of the 10 features. Using these training data, we constructed a random forest classifier with 200 trees. For each gene, this model produces three probability scores of it being an OG, a TSG or a PG, respectively. It assigns the class label based on the highest probability score. For all predictions, GUST reports random forests probability score, sensitivity and specificity. For OG or TSG predictions, GUST also reports false discovery rate. Detailed information of data processing, feature selection and false discovery rate calculation is available in the Supplementary Materials.

3 Results

3.1 Different selection patterns of cancer genes

For each gene/cancer-type pair in our manual annotations, we retrieved somatic mutations in the matching tumor samples from the TCGA project, and computed the somatic selection coefficients. We found that missense mutations in OGs were under stronger positive selection than in TSGs, as the mean \( \log(a) \) was 4.18 and 1.68, respectively (\( P < 10^{-16} \), Fig. 1A). In contrast, protein-truncating
Classify OGs and TSGs

mutations showed positive selection in TSGs \(\text{mean } \log(\phi) = 4.08\) but negative selection in OGs \(\text{mean } \log(\phi) = -3.25\). The effect size of the differences observed is very large, and the \(P\)-values highly significant \((P < 10^{-5})\). The selection measures observed on PGs were close to zero \(\text{mean } \log(\phi) = 0.60\) for missense and \(\text{mean } \log(\phi) = -0.28\) for nonsense mutations. Therefore, TSGs, OGs and PGs show significant evolutionary differences.

The distribution of \(\log(\phi)\) values of PGs had two peaks. The largest peak located close to 0, consistent with the expected neutral selection of PGs. The second peak located close to -5, indicating that loss of function of these PGs is detrimental to tumor growth. Interestingly, many genes in the second peak are established TSGs in other cancer types where loss of their functions is beneficial to tumors. For example, the \(BCOR\) gene regulates apoptosis in stomach cancer and had overabundant truncating mutations \((\text{Cancer Genome Atlas Research Network}, 2014)\). However, this gene was depleted of truncating mutations in melanoma \((\text{Fig. 1B})\). Such contrast suggested that although disabled TSGs promote tumor growth in certain cellular contexts, maintaining their activities may be essential for tumor development in other contexts. We then examined the joint distributions of \(\log(\omega)\) and \(\log(\phi)\) values and found that somatic selection patterns reflected the contextual activities of a gene \((\text{Fig. 1C})\). For example, the \(PIK3CA\) gene had high \(\log(\omega)\) values and low \(\log(\phi)\) values in bladder cancers, breast cancers and colorectal cancers, consistent with its well-known OG role. The \(\log(\omega)\) and \(\log(\phi)\) values of this gene were close to zero in melanoma, indicating lack of a role resulting in neutral patterns. Recently, the passenger role of \(PIK3CA\) in melanoma has been proposed in a study that shows \(PIK3CA\)-mutated melanoma cells rely on cooperative signaling to promote cell proliferation and PI3K inhibitors do not repress tumor growth in the absence of other activating driver genes in melanoma \((\text{Silva et al.}, 2017)\).

For TSGs, such as \(TP53\), their high \(\log(\phi)\) values occupied spaces distant from OGs in the distribution plot \((\text{Fig. 1C})\). As discussed earlier, TSGs with hotspots of missense mutations, such as the \(FBXW7\) gene in uterine carcinosarcoma \((\text{Fig. 1D})\) are challenging to distinguish from OGs using ratiometric methods. Based selection measures \(\log(\omega) = 5.0, \log(\phi) = 3.9\), this gene is unambiguously separated from OGs \(\log(\phi) < 0\), consistent with our expectations.

\[\begin{align*}
\text{Table 1. Performance of GUST and 20/20-} & \\
\text{Binary classes} & \text{Three classes} & \\
\text{Positive} & \text{Negative} & \text{Positive} & \text{Negative} & \text{Positive} & \text{Negative} & \text{Positive} & \text{Negative} \\
\text{OG, TSG} & \text{PG} & \text{OG} & \text{PG, TSG} & \text{TSG} & \text{PG, OG} \\
\text{GUST} & TPR 0.93 & 0.84 & 0.93 & — & — \\
 & TNR 0.94 & 0.98 & 0.95 & — & — \\
 & PPV 0.92 & 0.85 & 0.9 & — & — \\
 & NPV 0.95 & 0.98 & 0.96 & — & — \\
 & ACC 0.94 & 0.97 & 0.94 & 0.92 & — \\
 & AUC 0.97 & 0.99 & 0.97 & 0.98 & — \\
\text{20/20-} & TPR 0.90 & 0.85 & 0.86 & — & — \\
 & TNR 0.85 & 0.97 & 0.90 & — & — \\
 & PPV 0.82 & 0.78 & 0.81 & — & — \\
 & NPV 0.92 & 0.99 & 0.93 & — & — \\
 & ACC 0.88 & 0.97 & 0.89 & 0.86 & — \\
 & AUC 0.94 & 0.97 & 0.93 & 0.95 & — \\
\end{align*}\]

\(^a\text{Macro-AUC values were calculated by averaging three one-vs-rest ROC curves. Linear interpolation was used between points of ROC (Wei and Wang, 2018).}\)

\(\text{TPR, true positive rate; TNR, true negative rate; PPV, positive predictive value; NPV, negative predictive value; ACC, accuracy; AUC, area under the ROC curve.}\)

3.2 Performance of the GUST method

We trained a random forest classifier (GUST) using the 10 features of the curated genes. Via 10-fold gene-holdout cross-validations, the testing accuracy of GUST was 0.92. As a comparison, the accuracy of 20/20+ on the entire training dataset was 0.86. To calculate traditional performance metrics, we converted three-class predictions to binary predictions by contrasting one class with the other two classes combined, i.e. one-vs-rest predictions. In all categories, GUST showed better or comparable performance than 20/20+. The largest improvements were on the precision of identifying OGs and TSGs, which increased from 0.78–0.82 in 20/20+ to 0.85–0.92 in GUST \((\text{Table 1})\). The receiver operating characteristic (ROC) curves confirmed the superior performance of GUST \((\text{Fig. 2A})\). Compared to 20/20+, GUST had a significantly higher area under the curve \(\text{(AUC) value of the PG-vs-rest ROC curve (0.97 versus 0.94, DeLang’s test } P = 0.0008)\), and a significantly higher AUC value of the TSG-vs-rest ROC curve \(0.97 versus 0.93, P = 0.0011)\). However, the AUC values of the OG-vs-rest ROC curves were not significantly different between these two methods \(0.99 versus 0.97, P = 0.21)\.

To evaluate the concordance of GUST classifications with other methods that predict cancer drivers but do not distinguish OGs and TSGs, we computed a driver score by adding the OG and TSG scores of each gene. The TCGA PancanAtlas consortium reported a collection of putative driver genes based on consensus predictions from 12 computational methods \((\text{Bailey et al.}, 2018)\). We first examined the 510 gene/cancer-type pairs \(204 \text{ unique genes) predicted as drivers by } \ge 2 \text{ methods}\). In this permissive list, GUST predicted 373 pairs \(73.1\%, 145 \text{ unique genes) as drivers. We then examined the } 283 \text{ gene/cancer-type pairs (109 unique genes) predicted as drivers by } \ge 3 \text{ methods. In this more stringent list, GUST predicted 254 pairs (89.8\%, 96 unique genes) as drivers. These results showed that drivers predicted by GUST had a high agreement with existing methods while providing additional OG/TSG classifications.}\)

To measure the importance of each predictor in the random forest model, we computed the mean decreased Gini index by permuting out-of-bag samples \((\text{Louppe et al.}, 2013)\). The most informative predictors are the selection coefficients and fraction of truncating mutations, followed by the selection coefficient and fraction of missense mutations \((\text{Fig. 2B})\). Interestingly, evolutionary conservation was not very informative, which may be because a vast majority of drivers are known to occur at highly conserved positions.
TSGs with TSG-vs-rest specificities consisting of 22 OGs with OG-vs-rest specificity the ROC curves. We then derived a list of high-confidence drivers the sensitivity and specificity of each one-vs-rest prediction based on scores. To estimate the confidence of each prediction, we computed were annotated in only one cancer type and had low probability (Dudley et al., 2012), providing a limited power to discriminate OGs and TSG.

Although recurrence among patients has been taken as a surrogate of mutations under functional selection, recent investigations have shown that passenger hotspot mutations are common (Buisson et al., 2019; Hess et al., 2019). For example, multiple samples of various cancer types harbored a C>T or C>G mutation at position 931 of the MB21D2 gene (Fig. 2C, Supplementary Fig. S3). Buisson et al. discovered that this mutational hotspot is due to its location in a hairpin loop susceptible to mutagenesis and functions as a passenger (Buisson et al., 2019). GUST analysis confirmed that the selection pattern of this gene was consistent with neutral evolution in individual cancer types and in the combined samples (Fig. 2D). Thus, GUST predicted the MB21D2 gene as a PG correctly. This demonstrated the effectiveness of quantifying the contribution of genetic alterations to tumor fitness in cancer gene classifications.

3.3 Application to TCGA data

We retrieved somatic mutations from whole-exome sequencing data of 10 172 TCGA tumor samples spanning 33 cancer types. We then removed low-quality mutations, hyper-mutated or hypo-mutated samples, genes with fewer than four protein-altering mutations and genes mutated in <2% of tumors (Supplementary Materials). We applied GUST to the remaining 9663 samples. We predicted 161 OGs of which 98 were unique genes in 29 cancer types. We also predicted 331 TSGs of which 179 were unique genes in 33 cancer types (Fig. 3A, Supplementary Tables S3 and S4).

3.3.1 Novel driver genes

The GUST-predicted drivers consisted of 55 putative OGs and 97 putative TSGs that were classified as PGs in the CGC database (Sondka et al., 2018). Most (81.7%) of these new putative drivers were annotated in only one cancer type and had low probability scores. To estimate the confidence of each prediction, we computed the sensitivity and specificity of each one-vs-rest prediction based on the ROC curves. We then derived a list of high-confidence drivers consisting of 22 OGs with OG-vs-rest specificity ≥0.99 and 74 TSGs with TSG-vs-rest specificities ≥0.99, all of which had a PG-vs-rest sensitivity ≥0.99. This short list of high-confidence drivers included two novel OGs and 28 novel TSGs not annotated in the CGC. The two novel OGs (CNOT9 in melanoma and GTF2I in thymoma) had single mutational hotspots disrupting highly conserved protein positions (Fig. 3B and C). The GTF2I mutant stimulates cell proliferation in vitro and has been associated with favorable prognosis of thymoma (Roy, 2017).

All of the novel TSGs had an overabundance of truncating mutations (Supplementary Fig. S4). For example, frame-shifting mutations in SOX9 were observed in 40 colon cancers (Fig. 3D). As an atypical tumor suppressor, SOX9 has been shown to interact with nuclear β-catenin. Inactivation of SOX9 causes loss of inhibition of the oncogenic Wnt/β-catenin signaling pathway and is associated with patient survivals (Prevostel et al., 2016). Some novel TSGs harbor mutational hotspots. For instance, the N583fs frame-shifting mutation in BMP2R introduced premature stops of protein synthesis and was observed in nine stomach adenocarcinomas (Fig. 3E).
searched the literature and found supporting evidence of the tumor suppressing functions of 22 (78.6%) novel TSGs (Supplementary Table S5). Many of these novel TSGs were also annotated as putative drivers by other computational methods (Bailey et al., 2018).

As an independent assessment of the validity of these predicted drivers, we examined how many of their mutations were in major clones and compared with PGs. The rationale is that genes frequently mutated in sub-clones may not suggest a selective advantage, but rather other mechanisms, such as increased background mutational rates. Specifically, we used SciClone (Miller et al., 2014) to cluster mutations in each tumor based on variant allele frequencies. We considered the cluster with the highest variant allele frequencies as the major clone and the remaining clusters as sub-clones. For the 30 novel drivers, 93.2% of protein-altering mutations were in major clones, which was similar to the percentage (93.7%) for the 96 known drivers (Fisher’s exact test \( P = 0.54 \)). For the 40 most frequently mutated PGs, a significantly lower percentage (89.9%) of protein-altering mutations were in major clones (Fisher’s exact test \( P = 10^{-7} \)). Therefore, these predicted drivers highly likely promote tumorigenesis.

3.3.2 Spectrum of tissue specificity

Even after removing low-confidence predictions, most of the drivers annotated by GUST were engaged in only one cancer type, showing high tissue-specificities. Only 13 (59.1%) OGs and 23 (33.8%) TSGs in this high-confidence set are broad-spectrum drivers, promoting tumorigenesis in two or more cancer types (Fig. 3F). The most prevalent OG was the PIK3CA gene found in 15 cancer types with high confidence, followed by the KRAS/NRAS/HRS genes found in 13 cancer types. The most prevalent TSG was the TP53 gene found in 18 cancer types, followed by the ARID1A gene found in 10 cancer types.

Furthermore, 11 out of the 13 broad-spectrum OGs possessed multiple hotspots (one-sided proportional test \( P < 0.05 \) after Bonferroni corrections, Supplementary Fig. S5 and Supplementary Methods). For each significant hotspot, we examined the affected functional domains as annotated in the NCBI Gene database. A representative example is the EGF gene. In lung adenocarcinoma, 48% of missense mutations clustered at a single mutational hotspot affecting the tyrosine kinase activation loop (Fig. 3G). In glioma, only one mutation hit this loop (chi-square test \( P < 10^{-15} \)), and 69.3% of all missense mutations clustered at two hotspots affecting the extracellular domains independent of kinase activities. The context-dependent selection of mutations altering the kinase catalytic domain in glioma suggests an alternate path of activating EGF signaling. In fact, several studies have reported the associations of these hotspot mutations with different levels of EGF activities (Kamburov et al., 2015; Niu et al., 2016; Porta-Pardo et al., 2017). For cancer management, although tyrosine kinase inhibitors blocking EGF are common in the therapeutic armamentarium of lung cancer (Grigoriu et al., 2015; Takeda and Nakagawa, 2019), these agents have not been successful in treating glioma even with improved drug delivery techniques to penetrate the blood–brain barrier (Bethune et al., 2010; Vivanco et al., 2012; Westphal et al., 2017). These findings suggest a potential direction to investigate and enhance current treatment regimen.

Interestingly, each of the 33 cancer types engaged at least one broad-spectrum driver and multiple tissue-specific drivers, implicating the synchrony of convergent and divergent disease pathways. Clustering of cancers based on broad-spectrum driver genes grouped cancer types largely matching their tissue and cellular origins (Fig. 3H).

4 Discussions

Distinguishing OGs and TSGs in individual cancer types is critical to understanding cancer etiology and pinpointing clinically actionable targets. In this study, we proved that protein-coding mutations in OGs and TSGs are under different somatic selection, and subsequent development of the GUST method to discover cancer-type specific functions of cancer driver genes. We compared GUST with the 20/20+ method that is the only available method to classify OGs and TSGs. Both GUST and 20/20+ employ a random forest model to integrate features extracted from tumor exomes. Despite that GUST uses only 10 features compared to 24 features in 20/20+, the accuracy of GUST is consistently higher. In the GUST model, selection measures contribute the most information content. In 20/20+, the \( P \)-value of enrichment of inactivating mutations is the most informative feature. Interestingly, this feature is also related to selection, although it is not a strict evolutionary measure (Kryazhimskiy and Plotkin, 2008; Temko et al., 2018). These results suggest that using a small number of features engineered on evolutionary mechanisms is more powerful than feeding a large number of raw features to machine learning models. Furthermore, given the scarcity of known drivers for specific cancer types, reducing the number of features in predictive models helps mitigate overfitting problems.

We acknowledge that a driverMAPS (Zhao et al., 2019) method has been recently developed that estimates selection coefficients of a gene under three competing models (i.e. a PG, an OG and a TSG model). However, this method later combines the OG model and the TSG model into a driver model and contradicts it with the PG model to predict driver genes. Consequently, the reported posterior likelihood and false discovery rate are for the purpose of distinguishing drivers and passenger, but not OGs versus TSGs. Via personal communications with the authors of driverMAPS, we confirmed that this method does not provide statistical significance of OG and TSG classifications. Therefore, we did not compare GUST with driverMAPS.

While we discovered many known and novel cancer driver genes, none of them showed dual OG/TSG roles with high confidence in our analysis. A straightforward explanation is that GUST makes predictions based on protein-altering substitutions and indels, thus it is unable to capture genes acting through other mechanisms, such as noncoding regulatory variants, copy number variants, translocations, fusions, differential expressions, post-translational modifications and epigenetic regulations. Further investigations will shed light on key switches that divert paths of dual-role drivers. We also note that genes with only a small number of mutations may cause non-convergence problems during maximum likelihood estimations of selection coefficients, which limits the application of GUST to discovering rare drivers.

For practical use, we have built an online database (https://lliliu.lab.shinyapps.io/gust) with pre-computed results of analyzing TCGA samples. Users can query the database and visually inspect somatic selection patterns and consensual patterns of selected genes. Combined with information showing if a gene has been annotated by CGC as a driver or a drug target, users can make informed decisions on prioritizing candidate genes for further investigations. The R implementation of the GUST algorithm is available on Github (https://github.com/llilulab/gust).

5 Conclusions

Somatic selection is a quantitative measure of the impact of mutated genes on tumor fitness. The GUST method estimates these features directly from whole-exome sequencing or targeted sequencing data and pinpoints to genes and functional domains driving tumorigenesis in different cellular contexts. As gene-centered treatment and drug-repurposing attracts increasing interest, we expect this new method and the online database will facilitate discoveries of clinically actionable targets.

Acknowledgements

We thank Parween Wang for helpful discussions. L.L., C.M., and S.K. designed this study. L.L. developed the method. L.L. and N.A. performed the analysis. P.C. developed the database. All authors interpreted the results.

Funding

This work was supported by grants from the National Institutes of Health [LM012487 to S.K., U54CA217376 to C.M.]; the Finn Foundation [2088 to
Conflict of Interest: none declared.

References


