Journal of Bioinformatics and Computational Biology
Vol. 14, No. 6 (2016) 1650031 (16 pages)

DOI: 10.1142/S0219720016500311

World Scientific
www.worldscientific.com

Identifying cancer type specific oncogenes and tumor suppressors using limited size data

Ana B. Pavel

Graduate Program in Bioinformatics, Boston University
24 Cummington Mall, MA 02215, Boston, USA
anapavel@bu.edu

Cristian I. Vasile

Laboratory for Information and Decision Systems
Massachusetts Institute of Technology
77 Massachusetts Avenue, Cambridge, MA 02139, USA
cvasile@mit.edu

Received 9 February 2016
Accepted 24 August 2016
Published 7 October 2016

Cancer is a complex and heterogeneous genetic disease. Different mutations and dysregulated molecular mechanisms alter the pathways that lead to cell proliferation. In this paper, we explore a method which classifies genes into oncogenes (ONGs) and tumor suppressors. We optimize this method to identify specific (ONGs) and tumor suppressors for breast cancer, lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC) and colon adenocarcinoma (COAD), using data from the cancer genome atlas (TCGA). A set of genes were previously classified as ONGs and tumor suppressors across multiple cancer types (Science 2013). Each gene was assigned an ONG score and a tumor suppressor score based on the frequency of its driver mutations across all variants from the catalogue of somatic mutations in cancer (COSMIC). We evaluate and optimize this approach within different cancer types from TCGA. We are able to determine known driver genes for each of the four cancer types. After establishing the baseline parameters for each cancer type, we identify new driver genes for each cancer type, and the molecular pathways that are highly affected by them. Our methodology is general and can be applied to different cancer subtypes to identify specific driver genes and improve personalized therapy.

Keywords: Oncogene; tumor suppressor; driver gene; driver mutation; cancer.

1. Introduction

Cancer is a complex disease driven by different genetic, genomic or epigenetic mechanisms. A cancer driver gene is activated by driver mutations, but may also contain

passenger mutations with no effect in cancer. Tumor genomes may contain up to thousands of somatic mutations. However, only few of them drive tumorigenesis.[1–5]

Several approaches for estimating driver mutations using an unified analysis of several tumor types were previously proposed, such as MuSiC,[6] OncodriveFM12,[7] OncodriveCLUST[8] and ActiveDriver.[9] A comparison of these methods is presented in Ref. 1. MutSig[10] is a widely used tool which identifies genes that were mutated more often than expected by chance given background mutation processes.

Although the aforementioned methods identify significant variants in cancer, they do not assess the general activity of a gene, such as its role as an oncogene or tumor suppressor. An ONG is a gene that has the potential to cause cancer and it is often mutated in tumor cells by gain-of-function mutations. A tumor suppressor is a gene that protects a cell from cancer and may be mutated in cancer by loss-of-function mutations. Well studied (ONGs) were found to be recurrently mutated at the same amino acid position, while altered tumor suppressors were found to be mutated through protein-truncating alterations throughout their length.[2] A recurrent missense or in-frame indel usually indicate an oncogenic driver with a gain-of-function, while a nonsense, nonstop or frame-shift indel usually indicate a tumor suppressor with a loss-of-function.[2]

To the authors knowledge, there are not many general methods in literature which classify cancer genes as ONGs or tumor suppressors. We are aware of one such method which was proposed and validated in Ref. 2 and further applied to characterize gene activity in cancer.[11,12] The study proposed in Ref. 2 classifies a set of 125 cancer genes as ONGs or tumor suppressors, using data from the catalogue of somatic mutations in cancer (COSMIC).[13] COSMIC database v72 provides over four million variants across different types of cancers. In this paper, we validate and optimize the method proposed in Ref. 2, to accurately identify ONGs and tumor suppressor drivers in reduced size data sets of specific cancer types from the cancer genome atlas (TCGA).[14] Most of cancer studies using patient data are limited in sample size, therefore currently available cancer data sets which provide somatic mutation data may benefit from the proposed general approach of identifying cancer specific driver genes. In this study, we use somatic mutation data generated via DNA sequencing for the following TCGA data sets: breast cancer,[15] lung adenocarcinoma (LUAD),[16] lung squamous cell carcinoma (LUSC)[17] and colon adenocarcinoma (COAD).[18]

We identify potentially active ONGs and tumor suppressors in each of the four cancer types. These genes could serve as potential drug targets to improve therapeutic strategies. In order to develop target specific drugs, it is crucial to understand the activity of the target genes, such as gain-of-function for the oncogenes or loss-of-function for the tumor suppressors. Example of clinically available drugs that target gain-of-function mutations are Erlotinib (EGFR gain-of-function mutation) and BYL719 (PIK3CA gain-of-function mutation).[19,20] Therefore, we propose the following methodology to identify novel cancer targets and better understand the activity of known driver genes in different cancer types.

2. Method

2.1. *Classifying genes into oncogenes or tumor suppressors*

The *20/20 rule* proposed in Ref. 2 is an heuristic algorithm to classify genes into oncogenes or tumor suppressors. The rule takes into account both the mutation categories and their frequencies. First, for a given gene, the total number of variants is computed. Then, each gene is assigned an ONG score and a tumor suppressor gene (TSG) score which are computed based on the frequency of gain-of-function or loss-of-function mutations, respectively. Gain-of-function mutations include missense or in-frame indels which are recurrently mutated at the same aminoacid position, while loss-of-function mutations include nonstop, nonsense and frame-shift indels.[2] For each gene, the ONG score is the frequency of gain-of-function mutations out of the total number of variants, while the TSG score is the frequency of all loss-of-function mutations out of the total number of variants. To validate the method, the authors in Ref. 2 used data from the COSMIC[13] which is a large collection of somatic mutations from multiple studies of different cancers. They computed the ONG and TSG rule for 125 known cancer drivers and concluded that if ONG is greater than 20%, then the gene is an oncogene. Similarly, if TSG score is higher than 20%, then the gene is a tumor suppressor. The *20/20 rule* is illustrated in Fig. 1 for an ONG and a tumor suppressor.

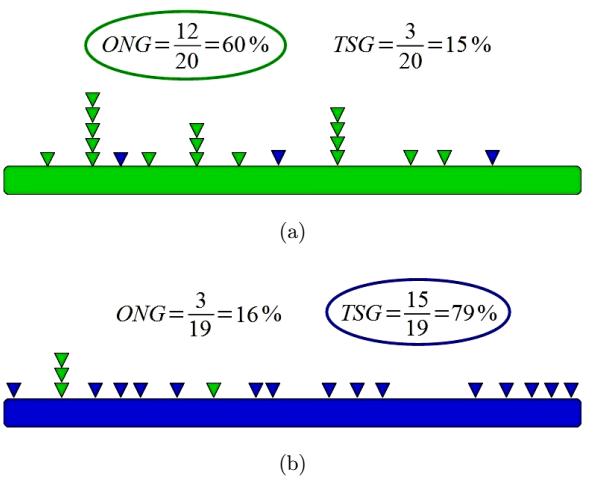


Fig. 1. Computation of ONG/TSG scores based on the method proposed in Ref. 2. Missense or in-frame indels are represented in green, while nonstop, nonsense and frame-shift indels are represented in blue. (a) Example of an oncogene: the frequency of gain-of-function mutations (recurrent missense/in-frame indels) is greater than 20% (ONG > 20%), while the frequency of loss-of-function mutations (nonstop/nonsense/frame-shift indels) is lower than 20% (TSG < 20%). (b) Example of a tumor suppressor: the frequency of loss-of-function mutations (nonstop/nonsense/frame-shift indels) is greater than 20% (TSG > 20%), while the frequency of gain-of-function mutations (recurrent missense/in-frame indels) is lower than 20% (ONG < 20%).

Some genes may be assigned both ONG and TSG scores greater than 20%. For example, TP53 tumor suppressor presents similarly elevated values for both ONG and TSG scores. For TP53 case, missense mutations generally drive the loss-of-function of the gene,[2,21–24] not its oncogenic activity as considered by the *20/20 rule*. However, the *20/20 rule* is an heuristics that was successfully validated for most of the well known ONGs and tumor suppressors.[2] In this paper, we will use the scores for the 125 known ONGs and tumor suppressors[2] as a baseline to further explore and optimize the rule in reduced size data sets of different cancer types.

The results of this paper are based upon data generated by the TCGA Research Network. Publicly available somatic mutation data (level 2) was downloaded from TCGA[14] for four cancer types: breast cancer (BRCA),[15] LUAD,[16] LUSC[17] and COAD.[18]

2.2. *Testing the 20/20 rule*

First, we compared our implementation of the *20/20 rule* using COSMIC mutation data v72 against the published values from Ref. 2. The authors in Ref. 2 used v61, however COSMIC data has been updated and it is currently available for download as v72.

As expected we obtained high correlation values (above 0.9) for both TSG and ONG scores across the 125 driver genes in Ref. 2 (Fig. 2). We explored different thresholds for the recurrence level of gain-of-function mutations: more than 2, 5, 10, 100 and 200 variants at the same aminoacid position (Figs. 2(b)–2(f)). The highest correlation for the ONG scores is obtained for a recurrence threshold of 2. We will further use this threshold in our analysis.

We did not obtain a perfect correlation of 1 between our computed scores and the previously published values due to the differences in the COSMIC database versions. Moreover, the authors in Ref. 2 may have filtered out some of the cancer studies from COSMIC, while we included all the currently available data. However, the correlation values are significantly high (0.92 for TSG scores and 0.95 for ONG scores) to validate our implementation of the *20/20 rule*. Next, we will evaluate the rule in four cancer specific data sets.

2.3. *Optimizing the 20/20 rule*

We hypothesized that the 20% frequency threshold of the *20/20 rule*, which indicates the active state of a gene,[2] could be optimized for each data set. Active cancer genes are specific to each type of cancer, therefore this threshold may be different from one type of cancer to another. This percentage may also depend on the dimensions of the data set and may change when a reduced sample size is used. Moreover, the gene activity may be defined by different thresholds for ONGs than for tumor suppressors.

Therefore, we considered the ONG and TSG scores obtained by the *20/20 rule* as baseline, since they were computed and validated on a large scale data such as

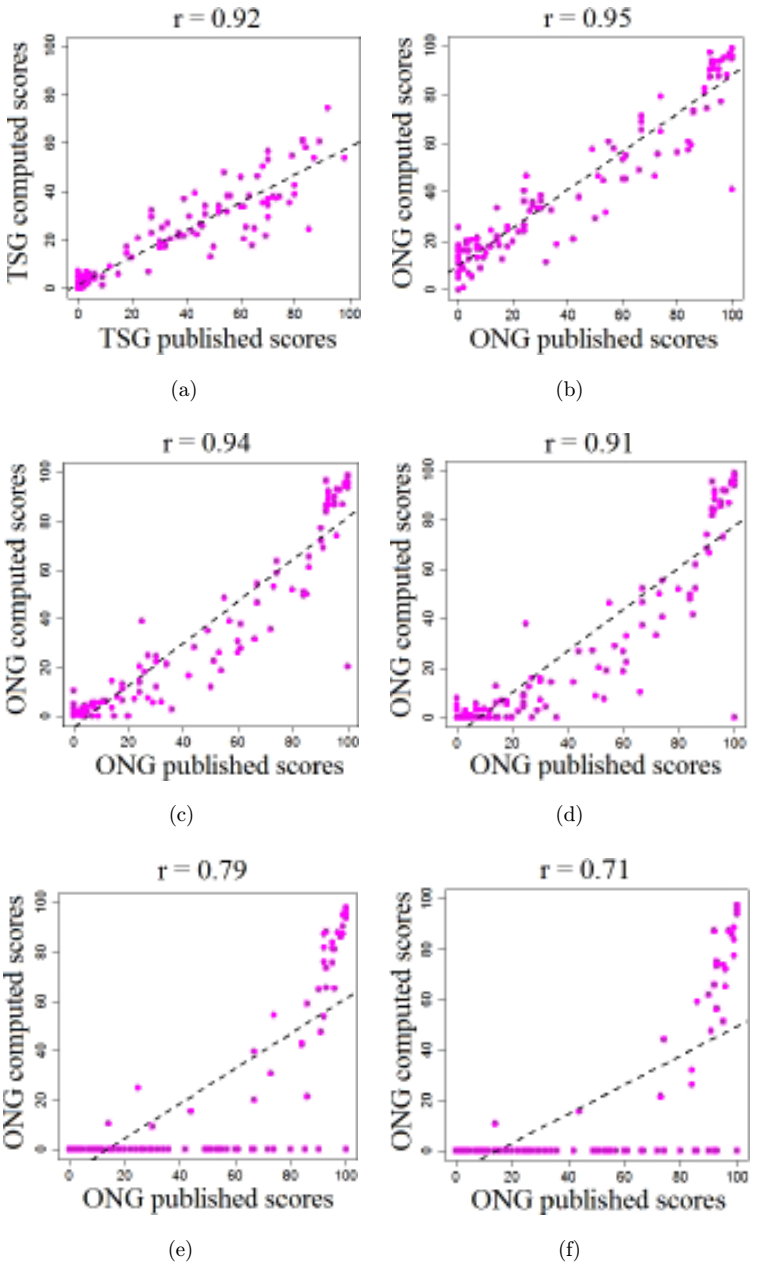


Fig. 2. Correlation between the scores computed in COSMIC data and the published scores in Ref. 2 for 125 known cancer drivers. (a) Loss-of-function (TSG) gene scores (r = 0.92). (b) Gain-of-function (ONG) gene scores for a recurrence threshold of minimum two samples ($r = 0.95$). (c) Gain-of-function (ONG) gene scores for a recurrence threshold of minimum five samples ($r = 0.94$). (d) Gain-of-function (ONG) gene scores for a recurrence threshold of minimum 10 samples ($r = 0.91$). (e) Gain-of-function (ONG) gene scores for a recurrence threshold of minimum 100 samples ($r = 0.79$). (f) Gain-of-function (ONG) gene scores for a recurrence threshold of minimum 200 samples ($r = 0.71$).

COSMIC. Then, we optimized the rule for each cancer type independently. We tested all possibly frequency thresholds between 1% to 100% and filtered out the inactive cancer genes from the 125 genes set. We computed the correlation of the remaining genes with the published scores and compared the results obtained using the default threshold of 20%. We then identified the ONG and TSG frequency thresholds which produced the highest correlation with the selected baseline genes.

To make sure we kept enough genes to estimate the correlation, we restricted our filtering to a minimum of 10% of the baseline genes. Correlation values on less than 12 genes were not considered. Moreover, we only included in the analysis those genes with a total number of variants higher than the background (the mean of variants per gene across all genes in the data set). This minimum value is around five variants in all the four data sets.

3. Results

3.1. *Results of the optimized 20/20 rule in each data set*

We optimized the *20/20 rule* in four different cancer types: TCGA breast cancer, LUAD, LUSC and COAD. For each data set, we selected the baseline genes which validate best. We improved the correlation with the published scores by selecting the most representative baseline genes. Therefore, we obtained correlation coefficients of around 0.9 for both ONG and TSG scores which are greater than the ones obtained by applying the default *20/20 rule* in each data set. For each of the four cancer types, the selected baseline genes are well known cancer genes that have been previously associated with:

- breast invasive carcinoma (BRCA), Fig. 3; the genes selected as the baseline (PIK3CA, KRAS, SF3B1, RET, ERBB2, TP53, MAP3K1, GATA3, CDH1, RUNX1, RB1, ARID1A) have been previously found as mutated in breast cancer[15,26–28];
- LUAD, Fig. 4; the genes selected as the baseline (EGFR, BRAF, KRAS, PIK3CA, U2AF1, CTNNB1, ARID1A, RB1, NF1, SETD2, STK11, SMARCA4, ATM) have been previously found as mutated in LUAD[16];
- LUSC, Fig. 5; the genes selected as the baseline (EGFR, PIK3CA, HRAS, NFE2L2, TP53, PTEN, RB1, TSC1, MLL2, SMAD4, CDKN2A, FUBP1) have been previously found as mutated in LUSC (except for FUBP1)[17];
- COAD, Fig. 6; the genes selected as the baseline (PIK3CA, KRAS, BRAF, NRAS, AR, TP53, APC, SOX9, FAM123B, ARID1A, CASP8, RNF43) have been previously found as mutated in COAD.[18,29,30]

Using the optimized frequency thresholds we identified novel cancer type specific ONGs and tumor suppressors which are available in Supplementary file 1. The correlation coefficients between the computed and the published scores, along with the optimized thresholds are shown for each TCGA data set in Table 1.

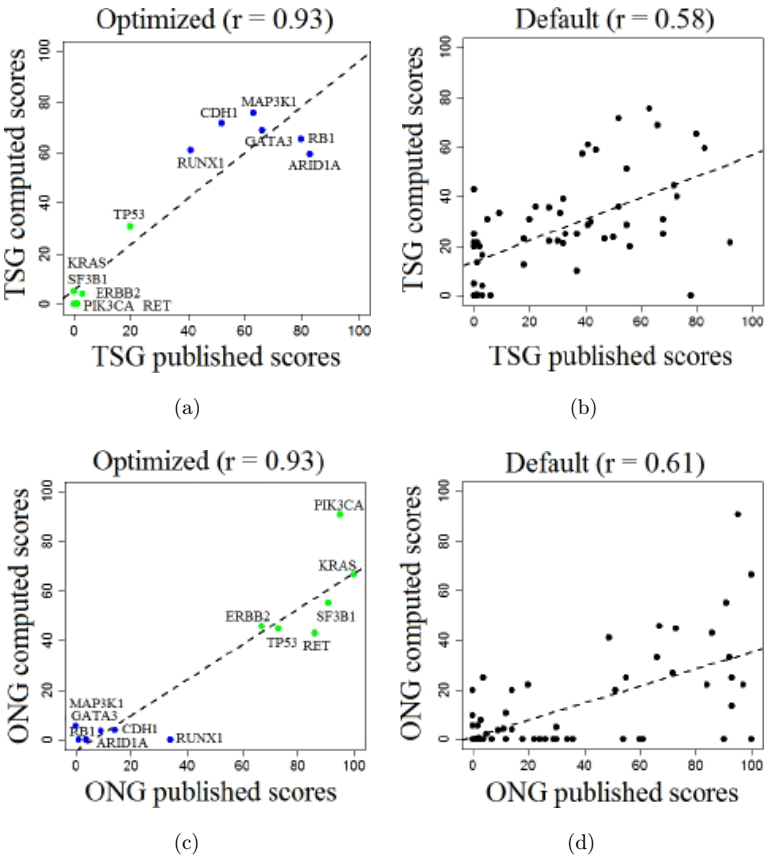


Fig. 3. Comparison between the optimized rule and the default *20/20 rule* in TCGA BRCA data set for the cancer genes in Ref. 2 (green illustrates a higher ONG score, while blue, a higher TSG score). Note that we exclude from the analysis those genes with a lower number of variants than the background (minimum six variants in this case). The correlation between the computed scores and the published scores is higher for the optimized rule. (a) Loss-of-function (TSG) gene scores computed in TCGA BRCA by the optimized rule. (b) Loss-of-function (TSG) gene scores computed in TCGA BRCA by the default *20/20 rule*. (c) Gain-of-function (ONG) gene scores computed in TCGA BRCA by the optimized rule. (d) Gain-of-function (ONG) gene scores computed in TCGA BRCA by the default *20/20 rule*.

Table 1 also provides information about the four TCGA data sets. Note that the TCGA data sets are much reduced in size compared to the entire COSMIC data that was used to test the default *20/20 rule*. COSMIC v72 provides a total number of 292 571 samples and over four million variants, while each TCGA data set provides hundreds of patients and thousands of variants.

Moreover, we included in Table 1 a comparison between the driver genes obtained by the optimized *20/20 rule* and MuSiC [6] or MutSig.[10] These two methods combined with manual curation were used by the TCGA papers to identify significantly mutated genes (MuSiC for BRCA[15] and MutSig for LUAD,[16] LUSC[17] and COAD[18]). Both MuSiC and MutSig analyze the mutations of each gene to identify genes that

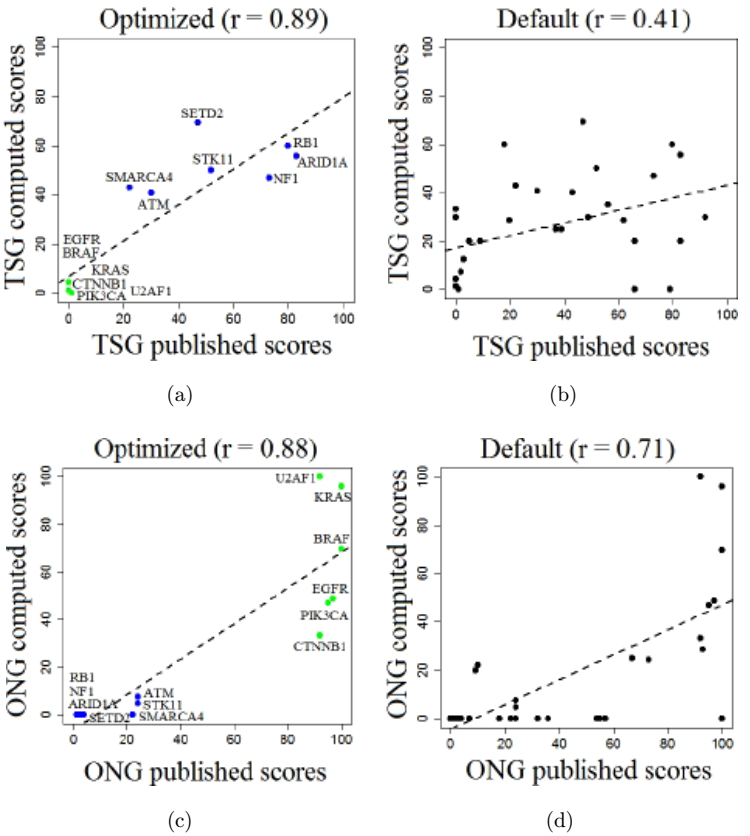


Fig. 4. Comparison between the optimized rule and the default *20/20 rule* in TCGA LUAD data set for the cancer genes in Ref. 2 (green illustrates a higher ONG score, while blue, a higher TSG score). Note that we exclude from the analysis those genes with a lower number of variants than the background (minimum five variants in this case). The correlation between the computed scores and the published scores is higher for the optimized rule. (a) Loss-of-function (TSG) gene scores computed in TCGA LUAD by the optimized rule. (b) Loss-of-function (TSG) gene scores computed in TCGA LUAD by the default *20/20 rule*. (c) Gain-of-function (ONG) gene scores computed in TCGA LUAD by the optimized rule. (d) Gain-of-function (ONG) gene scores computed in TCGA LUAD by the default *20/20 rule*.

were mutated more often than expected by chance, compared to background. They do not take into account the semantics of the mutation variants, such as gain-of-function or loss-of-function activity. Therefore, we do not expect to obtain identical results by the *20/20 rule* compared to MuSiC or MutSig. In addition, the rule provides a metric to classify the driver genes into ONGs or tumor suppressors.

3.2. *Gene drivers are cancer type specific*

Figure 7 shows the number of active driver genes identified in each cancer type. Interestingly, most of the genes we found to be important ONGs or tumor suppressors are specific to each cancer type. However, we found PIK3CA gene, which is known to be an active oncogene in many cancers, to present a high activity in all four

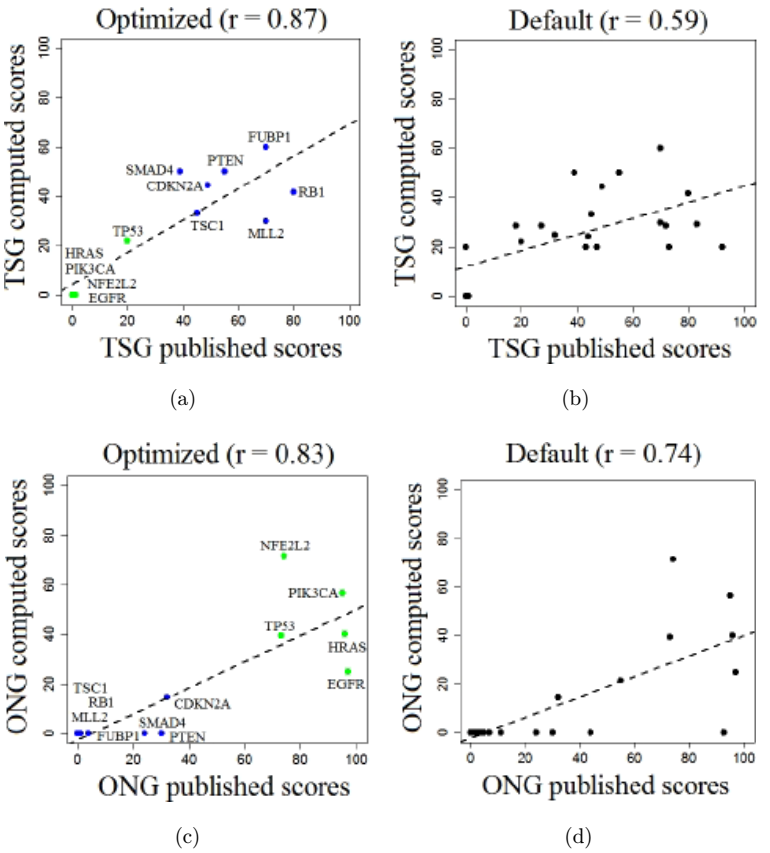


Fig. 5. Comparison between the optimized rule and the default *20/20 rule* in TCGA LUSC data set for the cancer genes in Ref. 2 (green illustrates a higher ONG score, while blue, a higher TSG score). Note that we exclude from the analysis those genes with a lower number of variants than the background (minimum five variants in this case). The correlation between the computed scores and the published scores is higher for the optimized rule. (a) Loss-of-function (TSG) gene scores computed in TCGA LUSC by the optimized rule. (b) Loss-of-function (TSG) gene scores computed in TCGA LUSC by the default *20/20 rule*. (c) Gain-of-function (ONG) gene scores computed in TCGA LUSC by the optimized rule. (d) Gain-of-function (ONG) gene scores computed in TCGA LUSC by the default *20/20 rule*.

cancer types. Moreover, EGFR is highly active in both lung cancer types (LUSC and LUAD), as expected.

TP53 is a gene known to present both oncogenic and tumor suppressor activities, most of the times contributing to cancer as a tumor suppressor with loss-of-function. TP53 is an exception to the *20/20 rule* because missense mutations in TP53 often cause its loss-of-function instead of gain-of-function as considered by the rule. This is also shown in Ref. 2, where the ONG score is higher than the TSG score for TP53. As expected, we found TP53 to have elevated scores in BRCA, LUSC and COAD. However, based on its tumor biology and the fact that it is a known exception to the *20/20 rule*, we classified TP53 as a gene with loss-of-function activity.

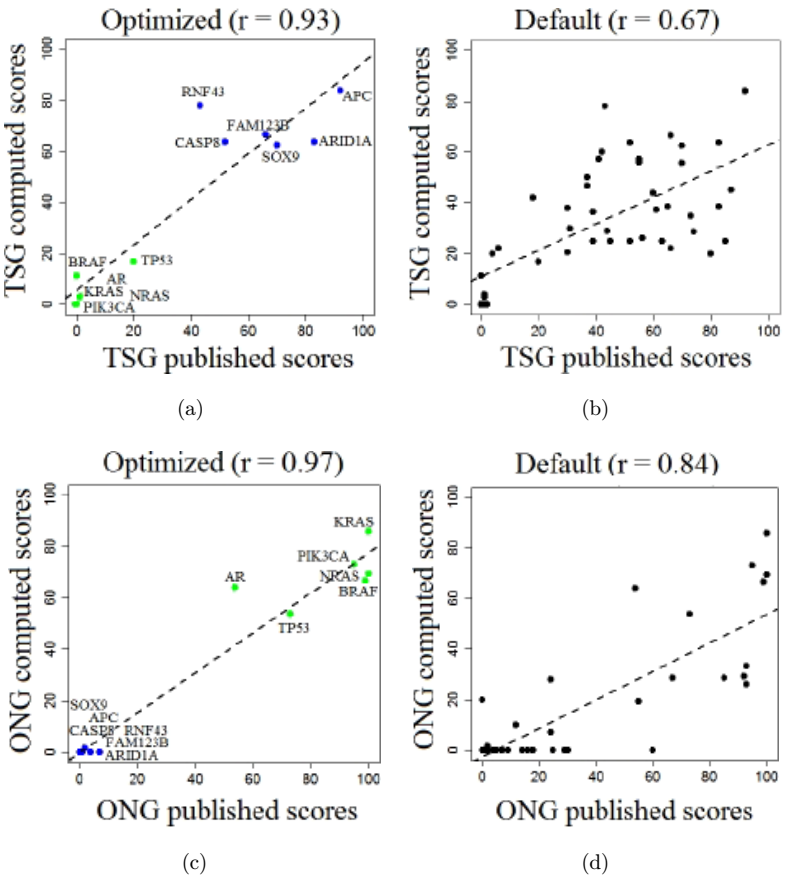


Fig. 6. Comparison between the optimized rule and the default *20/20 rule* in TCGA COAD data set for the cancer genes in Ref. 2 (green illustrates a higher ONG score, while blue, a higher TSG score). Note that we exclude from the analysis those genes with a lower number of variants than the background (minimum seven variants in this case). The correlation between the computed scores and the published scores is higher for the optimized rule. (a) Loss-of-function (TSG) gene scores computed in TCGA COAD by the optimized rule. (b) Loss-of-function (TSG) gene scores computed in TCGA COAD by the default *20/20 rule*. (c) Gain-of-function (ONG) gene scores computed in TCGA COAD by the optimized rule. (d) Gain-of-function (ONG) gene scores computed in TCGA COAD by the default *20/20 rule*.

Based on the ONG/TSG scores classification, we found seven oncogenenes (PIK3CA, EGFR, KRAS, BRAF, KRTAP5-5, KRTAP9-1, KCNN3) and nine tumor suppressors with loss-of-function (RB1, ARID1A, SOX9, RNF43, JPH4, HLA-A, RASA1, ATAD5, TP53) to be active in more than one cancer types. All of the other identified driver genes may be cancer specific as shown in Fig. 7.

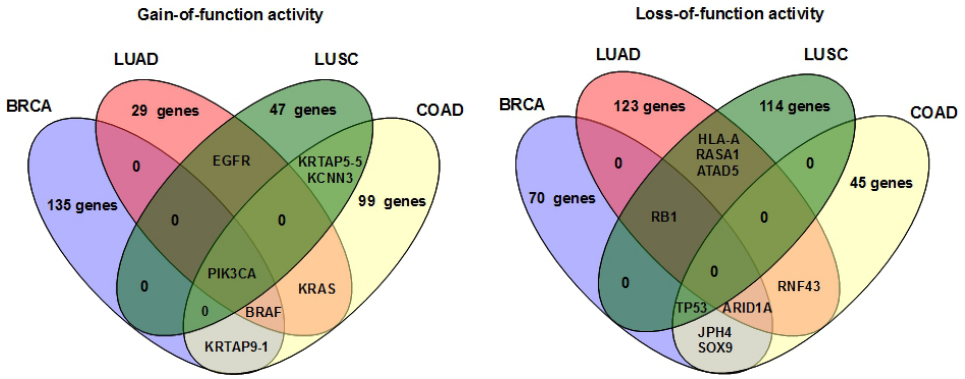
Next, we evaluated the most mutated pathways by an enrichment analysis using DAVID.[25] For each cancer type, we ran DAVID (v6.7) against the union of the two lists of ONGs and tumor suppressors. We considered all significantly enriched pathways ($Benjamini\ corrected\ p - value < 0.05$), from any available database

Table 1. Results of the optimized rule in TCGA data sets.

	BRCA	LUAD	LUSC	COAD
Number of patients	993	230	178	219
Number of variants	90 491	72 542	65 306	114 470
Optimized TSG threshold	59%	40%	30%	62.5%
Optimized ONG threshold	42%	33%	25%	53%
Correlation coefficient of the baseline tumor suppressors obtained by the optimized rule	0.93	0.89	0.87	0.93
Correlation coefficient of the baseline tumor suppressors obtained by the default *20/20 rule* from Ref. 2	0.58	0.41	0.59	0.67
Correlation coefficient of the baseline oncogenes obtained by the optimized threshold	0.93	0.88	0.83	0.97
Correlation coefficient of the baseline oncogenes obtained by the default *20/20 rule* from Ref. 2	0.61	0.71	0.74	0.84
Total number of tumor suppressors in each data set using the optimized threshold	75	129	119	50
Total number of oncogenes in each data set using the optimized threshold	138	33	51	105
Total number of driver genes by MuSiC/MutSig	21	18	22	32
Commonly identified driver genes by MuSiC/MutSig and the optimized rule	CCND3 CDH1 CDKN1B GATA3 MAP3K1 PIK3CA RB1 RUNX1 SF3B1 TP53	ARID1A BRAF EGFR KRAS MGA NF1 PIK3CA RB1 RBM10 RIT1 SETD2 STK11 SMARCA4 U2AF1	CDKN2A EGFR HLA-A HRAS KEAP1 MLL2 NFE2L2 PIK3CA PTEN SMAD4 TP53 TSC1	APC BRAF CASP8 FAM123B KRAS NRAS PIK3CA SOX9 TP53

used by DAVID, such as KEGG, PANTHER, etc. The significantly enriched pathways are related to multiple different cancer types. This suggests that different combination of driver genes may activate the same cellular programs in many cancer types. The significantly enriched pathways, identified for all the four gene sets, were: colorectal cancer, nonsmall lung cancer, myeloid leukemia, endometrial cancer, glioma, prostate cancer, pancreatic cancer, bladder cancer, melanoma cancer, etc. The most relevant pathways for each of the four cancer types were the following:

- breast invasive carcinoma (BRCA): endometrial cancer ($q-\text{value} = 0.0014$); progesterone-mediated oocyte maturation ($q-\text{value} = 0.027$); apoptosis ($q-\text{value} = 0.026$); ErbB signaling pathway ($q-\text{value} = 0.026$); p53 pathway feedback loops 2 ($q-\text{value} = 0.039$);



(a) (b)

Fig. 7. The overlap of the cancer genes between the four cancer types. (a) Very few active ONGs are present in multiple cancer types. Most active ONGs are found to be cancer type specific. PIK3CA is the only active oncogene in all four cancer types. (b) Very few tumor suppressors present loss-of-function in multiple cancer types. Most tumor suppressors with loss-of-function are found to be cancer type specific. Note that TP53 is an exception to *20/20 rule* because most of missense mutations in TP53 produce the gene's loss-of-function, not gain-of-function as considered by the rule. Therefore, TP53 is classified as a gene with loss-of-function activity.

- LUAD: nonsmall cell lung cancer ($q-\text{value}=0.0034$); regulation of actin cytoskeleton ($q-\text{value}=0.024$); ErbB signaling pathway ($q-\text{value}=0.05$);
- LUSC: nonsmall cell lung cancer ($q-\text{value}=0.00016$); p53 pathway feedback loops 2 ($q-\text{value}=0.025$);
- COAD: colorectal cancer ($q-\text{value}=0.0079$); insulin signaling pathway ($q-\text{value}=0.021$); Ras pathway ($q-\text{value}=0.023$); p53 pathway feedback loops ($q-\text{value}=0.045$); ErbB signaling pathway ($q-\text{value}=0.023$); PDGF signaling pathway ($q-\text{value}=0.034$); VEGF signaling pathway ($q-\text{value}=0.039$); signaling by EGFR ($q-\text{value}=0.049$).

Therefore, we identified driver genes by assessing the oncogenic and tumor suppressor activity within each data set separately, even under the limitation of the sample size. Optimizing the *20/20 rule* within each data set plays an important role in identifying the most relevant cancer type specific oncogenes and tumor suppressors. Moreover, we found the mostly mutated cancer pathways in each cancer type.

4. Limitations of the Optimized *20/20 rule*

The *20/20 rule* proposed in Ref. 2 provides a way to classify genes into ONGs or tumor suppressors based on the frequency of their gain-of-function or loss-of-function variants. This method is an heuristics and it is based on the idea that the gain-of-function mutations generally include missense/in-frame indels, while the loss-of-function mutations generally include nonsense/nonstop/frame-shift indels. The rule

has been shown to correctly classify most of the known ONGs and tumor suppressors (120 of 125 genes, 96%), with few exceptions such as TP53, NOTCH1, FBXW7, PAX5 and TRAF7.[2] To explain these exceptions, the authors in Ref. 2 point out that it is less likely for ONGs to harbor stop codons. Therefore, if a gene has a TSGscore $> 5\%$ and the ONG is also elevated (ONGscore $> 20\%$), it is more likely to be a tumor suppressor than an ONG. This assumption is based on the fact that ONGs rarely harbor stop codons and therefore some of the missense mutations may actually be loss-of-function mutations. This idea could be used for the optimized rule as well. In the case of elevated values for both ONG and TSG score, the gene is more likely to be a tumor suppressor. In this paper, we provided both ONG and TSG scores for each gene. The final status of a gene could be determined based on the ONG/TSG values combined with further assessment of certain types of mutations or other knowledge from literature. However, besides the TP53 known tumor suppressor, we have not found in the TCGA other exceptions to the optimized rule. Most genes had either a high ONG and a low TSG, or a high TSG and a low ONG (Supplementary file 1). NOTCH1, PAX5 and TRAF7 have not been selected as driver genes in either of the four TCGA data sets; FBXW7 was detected in LUAD and it was correctly classified as a tumor suppressor (TSG $= 0.6$ and ONG $= 0$).

By improving the thresholds of the *20/20 rule* in each cancer type, we selected those ONGs and tumor suppressors based on previously validated data. It is possible that some true ONGs or tumor suppressors may be missed by our approach due to prior knowledge. However, the optimized rule increases confidence in the newly identified oncogenes and tumor suppressors, by tuning the selection on previously validated genes.

5. Conclusions

In this paper, we evaluate the *20/20 rule* for classifying ONGs and tumor suppressors (TSGs) proposed in Ref. 2, and optimize it for limited size data sets of specific cancer types. To the authors knowledge, there are no other general methods in literature which are able to classify genes into ONGs and tumor suppressors. Most of the existing methods are focused on identifying driver mutations but they do not provide an overall ONG/TSG status for each gene. Although the *20/20 rule* is a general approximation of the absolute gene status, the authors in Ref. 2 prove its value and validate it for most of known driver genes. The impact of this approach is to identify novel ONGs and tumor suppressors that can be used to further understand cancer mechanisms and improve targeted therapies.

Therefore, in this paper we propose a new way of optimizing the *20/20 rule* by comparing the results with the baseline gene scores. The scores of well known cancer drivers were validated using a large sample size from COSMIC database, therefore these genes were used as baseline. We show that the best ONG and TSG frequency thresholds are not always equal to 20%, as proposed by Ref. 2, and may depend on

the cancer type and data size. We conclude that the *20/20 rule* validates in reduced size data sets if it is properly tuned.

We validated our approach by comparing the active baseline driver genes in each data set with known cancer genes from literature. By using the optimized rule, new ONGs and tumor suppressors were identified in each of the four cancer types (breast cancer, LUAD, LUSC and COAD). Moreover, we found that these driver genes are implicated in cancer type specific pathways. This explains part of the heterogeneity between each cancer type, showing that different mutations may lead to cancer development through different molecular mechanisms.

We plan to further investigate the activity of these genes and the pathways-level mechanisms by which they contribute to the initiation and proliferation of cancer. Designing *in-vitro* experiments would be further necessary to confirm the activity of these genes in each cancer type.

Supporting Information

Supplementary file 1: *Active driver genes*

The supplementary spreadsheet Drivers.xlsx contains 8 tabs corresponding to the oncogenes and tumor suppressors identified within each of the four cancer types. The computed ONG and TSG scores, as well as the number of gain-of-function mutations, loss-of-function mutations and the total number of mutations are available for each gene.

Acknowledgments

The authors would like to acknowledge the Graduate Program in Bioinformatics at Boston University for support with the current work.

The results published here are in part based upon publicly available data generated by the TCGA Research Network: http://cancergenome.nih.gov/

We also use publicly available data from COSMIC database: cancer.sanger.ac.uk

References

1. Tamborero D, Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Kandoth C, Reimand J *et al.*, Comprehensive identification of mutational cancer driver genes across 12 tumor types, *Sci Rep* **3**:2650, 2013.
2. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW, Cancer genome landscapes, *Science* **339**(6127):1546–1558, 2013.
3. Garraway L, Lander E, Lessons from the cancer genome, *Cell* **153**:17–37, 2013.
4. Stratton M, Campbell P, Futreal P, The cancer genome, *Nature* **458**:719–724, 2013.
5. Gonzalez-Perez A, Mustonen V, Reva B, Ritchie G, Creixell P, Karchin R *et al.*, Computational approaches to identify functional genetic variants in cancer genomes, *Nat Methods* **10**(9):723–729, 2013.
6. Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC *et al.*, MuSiC: Identifying mutational significance in cancer genomes, *Genome Res* **22**(8):1589–1598, 2012.

7. Gonzalez-Perez A, Lopez-Bigas N, Functional impact bias reveals cancer drivers, *Nucleic Acids Res* **40**(21):e161, 2012.
8. Tamborero D, Gonzalez-Perez A, Lopez-Bigas, OncodriveCLUST: Exploiting the positional clustering of somatic mutations to identify cancer genes, *Bioinformatics* **29**(18):2238–2244, 2013.
9. Reimand J, Bader G, Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers, *Mol Syst Biol* **9**(637), 2013, doi: 10.1038/msb.2012.68.
10. Lawrence M, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A *et al.*, Mutational heterogeneity in cancer and the search for new cancer-associated genes, *Nat Lett* **499**(7457):214–218, 2013.
11. Fleck JL, Pavel AB, Cassandras CG, Integrating mutation and gene expression cross-sectional data to infer cancer progression, *BMC Syst Biol* **10**(12), 2016, doi: 10.1186/s12918-016-0255-6.
12. Pavel AB, Sonkin D, Reddy A, Integrative modeling of multi-omics data to identify cancer drivers and infer patient-specific gene activity, *BMC Syst Biol* **10**(16), 2016, doi: 10.1186/s12918-016-0260-9.
13. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H *et al.*, COSMIC: Exploring the world's knowledge of somatic mutations in human cancer, *Nucleic Acids Res.* **43**(D1):D805–D811, 2014.
14. Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K *et al.*, The cancer genome atlas pan-cancer analysis project, *Nat Genet* **45**(10):1113–1120, 2013.
15. The Cancer Genome Atlas Network TCGA *et al.*, The Comprehensive molecular portraits of human breast tumours, *Nature* **490**(11412):61–70, 2012.
16. The Cancer Genome Atlas Research Network TCGA *et al.*, Comprehensive molecular profiling of lung adenocarcinoma, *Nature* **511**(7511):543–550, 2014.
17. The Cancer Genome Atlas Research Network TCGA *et al.*, Comprehensive genomic characterization of squamous cell lung cancers, *Nature* **489**(7417):519–525, 2012.
18. The Cancer Genome Atlas Network TCGA *et al.*, Comprehensive molecular characterization of human colon and rectal cancer, *Nature* **487**(7407):330–337, 2012.
19. Fritsch C, Huang A, Chatenay-Rivauday C, Schnell C, Reddy A, Liu M *et al.*, Characterization of the Novel and Specific PI3Ka Inhibitor NVP-BYL719 and Development of the Patient Stratification Strategy for Clinical Trials, *Mol Cancer Therapeuti* **13**(5):1117–1129, 2014.
20. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S *et al.*, The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity, *Nature* **483**(7391):603–607, 2012.
21. Sonkin D, Hassan M, Murphy DJ, Tatarinova TV, Tumor suppressors status in cancer cell line encyclopedia, *Mol Oncol* **7**(4):791–798, 2013.
22. Ponder BA, Cancer genetics, *Nature*, **411**(6835):336–341, 2001.
23. Hollstein M, Hainaut P, Massively regulated genes: The example of TP53, *J Pathol* **220**(2):164–173, 2010.
24. Payne SR, Kemp CJ, Tumor suppressor genetics, *Carcinogenesis* **26**(12):2031–2045, 2005.
25. Huang DW, Sherman BT, Lempicki RA, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources, *Nat Protocols* **4**(1):44–57, 2009.
26. Mamo A, Cavallone L, Tuzmen S, Chabot C, Ferrario C, Hassan S *et al.*, An integrated genomic approach identifies ARID1A as a candidate tumor-suppressor gene in breast cancer, *Oncogene* **31**(16):2090–2100, 2012.

27. Morandi A, Isacke C, Targeting RETinterleukin-6 crosstalk to impair metastatic dissemination in breast cancer, *Breast Cancer Res* **16**(1):301, 2014.
28. Pang B, Cheng S, Sun SP, An C, Liu ZY, Feng X *et al.*, Prognostic role of PIK3CA mutations and their association with hormone receptor expression in breast cancer: A meta-analysis, *Sci Rep* **4**(6255), 2014, doi: 10.1038/srep06255.
29. Slattery M, Sweeney C, Murtaugh M, Ma K, Wolff R, Potter J *et al.*, Associations between ERalpha, ERbeta, and AR genotypes and colon and rectal cancer, *Cancer Epidemiol Biomarkers Prev* **14**(12):2936–2942, 2005.
30. Koo B, Spit M, Jordens I, Low T, Stange D, van de Wetering M *et al.*, Tumour suppressor RNF43 is a stem-cell E3 ligase that induces endocytosis of Wnt receptors, *Nature* **488**(7413):665–669, 2012.

Ana B. Pavel received two Bachelor degrees, one in Computer Engineering from Politehnica University of Bucharest, Romania, and another in Biochemistry from the University of Bucharest, and a Masters in Intelligent Control Systems from Politehnica University of Bucharest. She is currently a PhD candidate in the Graduate Program in Bioinformatics at Boston University, MA, USA. Website: https://anabrandusa.wordpress.com

Cristian I. Vasile received his Bachelor degree in Computer Engineering and Masters in Intelligent Control Systems from Politehnica University of Bucharest, Romania, and his PhD in Systems Engineering from Boston University, MA, USA. He is currently working as a post-doctoral associate at Massachusetts Institue of Technology, MA, USA. Website: http://www.cristianvasile.com