Original Paper

e-Driver: A novel method to identify protein regions driving cancer

Eduard Porta-Pardo[1], Adam Godzik[1,*]

[1]Bioinformatics and Systems Biology Program, Sanford-Burnham Medical Research Institute, 10901 North Torrey Pines Road, La Jolla, CA, 92037, USA

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Most approaches to identify cancer driver genes focus, true to their name, on entire genes and assume that a gene, treated as one entity, has (or has not) a specific role in cancer. This approach may be correct to describe effects of gene loss or changes in gene expression, however mutations, splicing variants or post-translational modifications of proteins coded by a given gene may have different effects, including their relevance to cancer, depending on which region of the gene they affect. Except for rare and well-known exceptions, there is not enough data for reliable statistics for individual positions, but an intermediate level of analysis, between an individual position and the entire gene, may give us better statistics than the former and better resolution than the latter approach.

Results: We have developed e-Driver, a method that exploits the internal distribution of somatic missense mutations between the protein's functional regions (domains or intrinsically disordered regions) to find those that show a bias (an enrichment or depletion) in their mutation rate as compared to other regions of the same protein, providing evidence of positive selection and suggesting that these proteins are actual cancer drivers. We have applied e-Driver to a large cancer genome dataset from The Cancer Genome Atlas (TCGA) and compared its performance to that of four other methods, showing that e-Driver identifies novel cancer drivers and, because of its increased resolution, provides deeper insights into the mechanism of cancer driver genes identified by other methods. As an example of such insights, we show examples of e-Driver identifying different protein functional regions in the same protein that are relevant to different cancer types.

Availability: A web server to run e-Driver is being implemented. Also, a Perl script with e-Driver as well as the files to reproduce the results described in this publication can be downloaded from https://github.com/eduardporta/e-Driver.git

Contact: adam@godziklab.org or eppardo@sanfordburnham.org

Supplementary information: Supplementary data are available at Bioinformatics online

1 INTRODUCTION

The landscape of cancer somatic mutations revealed by projects such as The Cancer Genome Atlas (TCGA)(Chang *et al.*, 2013) or the International Cancer Genome Consortium (ICGC)(Hudson *et al.*, 2010) is overwhelmingly complex, as hundreds of thousands of different mutations, ranging from large genomic rearrangements to point missense mutations, have been identified in different cancer samples (Ciriello *et al.*, 2013, Kandoth *et al.*, 2013). Several approaches have been developed in order to identify which genes are driving the carcinogenic process (driver genes). Such methods rely on the hypothesis that driver genes should be under positive selection in the cancer environment. Methods in this category include those that try to identify genes with higher than expected by chance mutation rates, such as MuSiC (Dees *et al.*, 2012), or those that tend to accumulate highly damaging mutations, such as OncodriveFM (Gonzalez-Perez and Lopez-Bigas, 2012). More recently, methods that focus on the internal distribution of mutations along a protein have also been developed. For example, OncodriveCLUST (Tamborero, Gonzalez-Perez, and Lopez-Bigas, 2013) looks for regions of proteins with higher than expected mutation rates, which makes it optimal for the identification of gain-of-function sites that, while being key for the carcinogenic process, would otherwise be missed. Another similar idea is ActiveDriver (Reimand and Bader, 2013), which tries to identify phosphorylation sites that are recurrently mutated in cancer. Note that one of the differences between the two methods is that ActiveDriver tests the mutation frequencies of predefined regions (a phosphorylation site and its neighboring aminoacids), while OncodriveCLUST first looks for potential seeds of highly mutated clusters and then tries to extend them.

Here we present e-Driver, a novel method that identifies protein functional regions (PFRs) that show a bias in their mutation rates. In this context, PFRs can be either domains or intrinsically disordered regions. Our method is based on the assumption that different PFRs within the same protein mediate different functions and, thus, might have distinct roles in carcinogenesis. This becomes evident when describing proteins in terms of functional networks. In such networks nodes represent different proteins and edges between nodes represent functional relationships between them, such as physical interactions or post-translational modifications. Different edges leading to the same node/protein are usually mediated by different PFRs within that same protein, and mutations in the PFR mediating one edge will have different consequences than mutations in another PFR mediating a different edge. For example, if an enzyme contains a catalytic domain and an intrinsically disordered region that is phosphorylated, it is likely that the consequences of a missense mutation disrupting the catalytic domain will be different from those of a missense mutation affecting the phosphorylation site or a truncating mutation that disrupts both PFRs at the same time. Our method exploits this idea, which has been previously used to analyze mutations associated with Mendelian disorders(Zhong *et al.*, 2009, Wang *et al.*, 2012), by looking for PFRs that show a bias in their mutation rate.

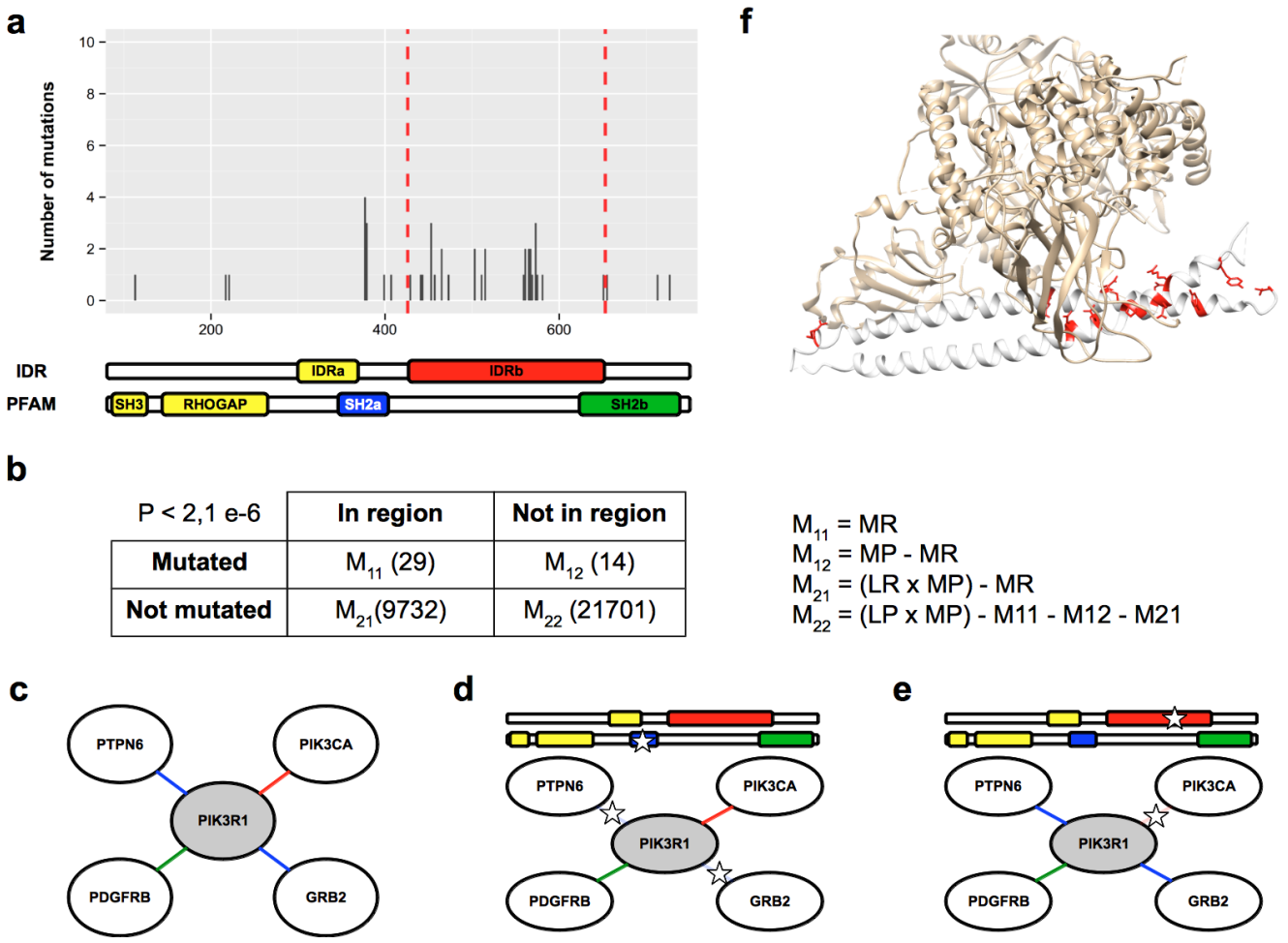

Figure 1 - e-Driver's workflow shown in the example of the analysis of PIK3R1 mutation data from TCGA. (a) e-Driver first retrieves all missense mutations in a protein. It then identifies its PFRs, such as Pfam domains or intrinsically disordered regions (IDRs). For example, in the case of PIK3R1, the protein contains four different Pfam domains (one SH3 domain, one RhoGAP domain and two SH2 domains) and two distinct IDRs. Note that the predictions are independent and, thus, can overlap, as in the case of the second IDR and the second SH2 domain. e-Driver iterates through every functional region, calculating the p values of the mutation distribution using a Fisher's test that takes into account the mutation rates and lengths of both, the region of interest and the protein. (b) Example of the contingency table for the IDRb in PIK3R1. MR is the number of mutations in the region being studied; MP is the total number of mutations in the protein; LR is the length of the region; LP is the length of the protein; MP is the number of mutations in TCGA in this protein. (c) Each of the different PFRs in PIK3R1 performs different functions. For example, the first SH2 domain is responsible for the interactions with GRB2 and PTPN6 (blue edges), while the second SH2 domain mediates the interaction with PDGFRB (green edge) and the second IDR mediates the interaction with PIK3CA (red edge). (d) It is likely that missense mutations in the SH2b domain of PIK3R1 will disrupt, among others, its interaction with PDGFRB without altering the rest of the network. Given that this region is not enriched in cancer somatic mutations the functions/interactions mediated by this domain are unlikely to be oncogenic. (e) On the other hand, IDRb is strongly enriched in somatic mutations, thus, edges mediated by this region, such as the physical interaction with PIK3CA, are likely to be relevant to carcinogenesis. (f) The mutations in PIK3R1 (the white helical protein) IDRb region (shown in red) cluster around the region that interacts with PIK3CA (shown in brown). Representation based on PDB structure

We have applied e-Driver to the cancer genomic dataset from the pan-cancer project of the TCGA. This dataset has also been analyzed with four other methods (MuSiC, OncodriveFM, OncodriveCLUST and ActiveDriver), allowing us to compare the results obtained with e-Driver with those obtained by methods relying on other approaches to identify the signals of positive selection (Tamborero, Gonzalez-Perez, Perez-Llamas, *et al.*, 2013).

2 METHODS

2.1 Identification of the driver PFRs

e-Driver is based on the hypothesis that not all functional regions of a given protein are equally relevant for carcinogenesis. If this is the case it should be reflected in the distribution of missense mutations along the protein, with regions under selection showing an enrichment or depletion of such mutations as compared to regions with random (passenger) mutations.

In order to identify PFRs under selective pressure, e-Driver first retrieves all missense mutations in a cancer cohort located in any given protein as well as the mutation coordinates and maps them to the protein's functional regions. Then, for every PFR we use a two-sided Fisher test to check whether the observed number of muta-

tions in this protein region (MR in Figure 1b) is biased or not. We assume that each mutation is an independent event and that all residues of the protein have the same probabilities of being mutated. Then, given the total number of mutations in the protein, (MP in Figure 1b) we can calculate the number of possible mutated residues by multiplying the length of the protein by the number of mutations in this protein. Similarly, the total number of possible mutated residues of the region is defined as the length of the region times the number of mutations in the protein. Once the p values of all the regions of all mutated proteins in the cohort are obtained, the Benjamini-Hochberg false discovery rate algorithm is applied in order to correct for multiple testing. Those regions with a q value < 0.05 are considered as positive. The whole process is explained in Figure 1 in the example of PIK3R1 and its functional regions.

2.2 PFR annotations

We defined protein functional regions as sections of the protein coding for individual protein domains and intrinsically disordered regions (IDRs). We decided to include intrinsically disordered regions because they can also contain important functional regions such as phosphorylation sites or regions that regulate or mediate protein interactions (Dunker *et al.*, 2005).

To identify protein domains, we assigned, for each protein isoform from ENSEMBL, annotated Pfam domains (Punta *et al.*, 2012) annotated in ENSEMBL and putative novel protein domains located in regions with no previous domain annotations, as predicted using the AIDA server (Xu *et al.*, 2014). We used Foldindex (Prilusky and Felder, 2005) to predict IDRs for each protein, including in our analysis those regions with a predicted unfolded score below –0.1.

Finally, we mapped the different missense somatic mutations of each tumor to these PFRs, giving us a total of 66,492 altered regions in 14,421 genes based on data from 3,205 tumor samples (see below). Among the 66,492 regions we have 36,626 Pfam domain instances, 4,626 putative domains predicted by the AIDA server, and 25,240 IDR. Note that the features can overlap, as the predictions were performed independently and there is no reason why, for example, an intrinsically disordered region cannot overlap with (or even be located within) a Pfam domain. Note also that for the sake of simplicity we only analyzed the longest isoform of each gene.

2.3 TCGA Mutation dataset

We have downloaded the dataset that was used in the TCGA pan-cancer driver analysis (syn1729383). In order to compare our results with the ones obtained in the TCGA pan-cancer analysis we applied the same filters to the dataset, excluding 71 samples that were considered to be hypermutators (Tamborero, Gonzalez-Perez, Perez-Llamas, *et al.*, 2013). After filtering, the final dataset consists of 3,205 tumor samples with 287,822 coding missense mutations.

2.4 Predicted driver genes by the other four methods

In order to assess the value of our method we compared our results with those obtained by four different methods used previously to predict high-confidence gene drivers in the TCGA pan-cancer project: MuSiC, OncodriveFM, OncoCLUST and ActiveDriver (Tamborero, Gonzalez-Perez, Perez-Llamas, *et al.*, 2013). We downloaded the results obtained in this analysis for three of the four methods: OncodriveFM (syn1701498), OncodriveCLUST (syn1701498) and MuSiC (syn1713813). Since no ActiveDriver results for the whole genome were available on the repository describing the pan-cancer analysis, we used ActiveDriver results described in another paper (Reimand *et al.*, 2013) and that, according to their authors, have been obtained with very similar TCGA mutation data (3,185 cancer genomes, syn2237931). Therefore, the results shown here for ActiveDriver are slightly different than those described in the pan-cancer analysis.

2.5 Tissue-specific drivers

We classified the 3,205 tumor genomes into their corresponding 11 tissues of origin, obtaining 11 tissue-specific datasets that were then analyzed individually with e-Driver. We then again corrected for multiple testing by considering as positives only those PFRs with a q value < 0.05.

3 RESULTS

3.1 e-Driver identifies known cancer drivers

In order to assess the validity of our method we reanalyzed the pan-cancer dataset of the TCGA. This dataset contains mutation data for 3,205 tumor samples that come from 11 different types of tumors, and contains 287,822 missense mutations. The dataset has been previously analyzed using four different state-of-the-art methods to predict cancer drivers from mutation data (MuSiC, OncodriveFM, OncoCLUST and ActiveDriver).

When applying our method to this dataset, we identified 74 protein regions in 51 genes, showing a bias in their mutation rate when compared to the rest of the protein (Figure 2a). Among these 51 genes, 23 are included in the Cancer Gene Census, a curated list of 512 cancer drivers (Futreal *et al.*, 2004). This represents a strong enrichment in CGC genes in our list of candidate drivers when compared to random expectation (Figure 2b, odds ratio >25, p value < 1e-16). As shown in Figure 2a, 31 of the 51 genes predicted by e-Driver (61%) are also identified by other methods. The highest overlap of e-Driver predictions is with predictions from OncodriveFM and MuSiC, with 21 of 51 genes (41%) being common. Regarding genes included in the Cancer Gene Census, 22 of the 23 genes identified by e-Driver (96%) that belong to this list have also some other signal of positive selection, as they are also predicted by other methods.

Interestingly, there is one gene in the CGC, CREBBP, that has not been identified by any of the other four methods, but was picked up by e-Driver. CREBBP protein does not show any specific cluster of mutations nor is it recurrently mutated in cancer, which could explain why it is not recognized as cancer driver by the other methods. Nevertheless, its mutation pattern shows a strong bias as the acetyltransferase domain, located between aminoacids 1345 and 1639 (12% of the protein's length) contains 20 of the 60, or 30%, of all the mutations found in this gene (q value < 0.02).

There is one other acetyltransferase domain, in the EP300 gene, which is also enriched in somatic mutations and identified by e-Driver. This gene is also included in the CGC and is also identified by MuSiC and OncodriveFM, but not by OncodriveCLUST or ActiveDriver. This observation suggests that, while EP300 is frequently mutated in cancer, its mutations show no particular clustering. However, by using e-Driver we can identify the specific region of the protein that is enriched in mutations.

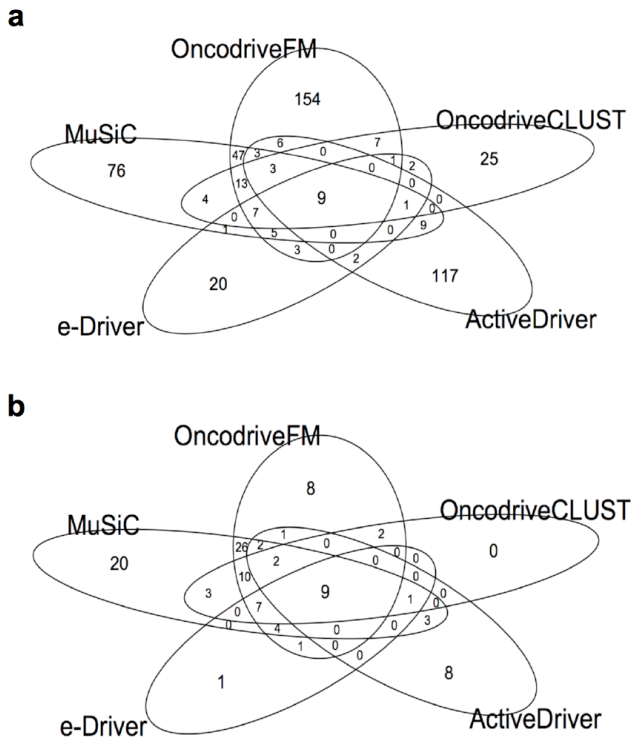

Figure 2 - e-Driver identifies known cancer driver genes. (a) Venn diagram showing the overlap between the five different methods in their predictions. (b) Venn diagram showing the overlap between the five different methods at predicting genes included in the Cancer Gene Census.

3.2 e-Driver finds potential novel drivers

We then reviewed the remaining 28 genes that are identified as potential drivers by our method, but that are not included in the CGC. Eight of them had also been identified by, at least, one other method, supporting their potential role as cancer drivers. For example, our method, as well as OncodriveFM, identified the MGA gene as a potential driver. This gene encodes a dual-specificity transcription factor that regulates expression of Myc/MAX target genes. It suppresses the transcriptional activation by Myc and inhibits Myc-dependent cell transformation. The domain identified by e-Driver is the Helix-loop-helix domain between positions 2425 and 2474 that contains 8 of the 46 mutations identified in this protein (odds ratio 13, q value < 0.001) and that mediates the binding of the protein to E-boxes in the DNA. Additional evidence in favor of the carcinogenic role of MGA comes from a recent study (Lawrence *et al.*, 2014) using a larger genomic dataset, with 4,742 cancer samples, where thanks to an increase in sample size and statistical power, MGA could be identified by MuSiC. As for the other seven genes that were also predicted by other methods, five of them were included in the list of 258 high-confidence drivers described in the pan-cancer driver analysis: FRG1B, NBPF10, DHX9, POTEF and RPSAP58. This result agrees with previous observations that genes predicted by more than a single method are likely to be true cancer drivers (Tamborero, Gonzalez-Perez, Perez-Llamas, *et al.*, 2013) and confirms the power of our method to identify genes relevant to the disease.

Among the remaining 20 genes that are not part of the CGC and that are not identified by any other method we have found several potential drivers. For example, we identified two members of the neuroblastoma breakpoint family, NBPF12 and NBPF20, as having regions with strong enrichment in mutations. These two genes belong to the same family as NBPF10, one of the genes included in the list of high-confidence drivers of the pan-cancer analysis. Interestingly, the disordered regions identified by e-Driver from NBPF12 and NBPF20 have a 94% identity, suggesting that their potential driver role might be achieved through similar mechanisms. Other interesting genes identified uniquely by our method include POTEM, a protein that belongs to the same family as the high-confidence driver POTEF. As in NBPF proteins, the regions identified in POTEM and POTEF are IDR, however in this case they do not show any homology. Another interesting fact about POTEF is that the region identified by e-Driver does not show an enrichment in cancer somatic mutations but instead a depletion, suggesting that the conservation of this PFR is important for the survival of cancer cells and for POTEF's role as driver.

3.3 Tissue-specific PFR

Cancer is a very heterogeneous disease and it is known that mutations driving one type of cancer might be completely irrelevant for another. Thus, while the pan-cancer dataset has more statistical power due to its larger size, it is possible that there are tissue-specific drivers that cannot be detected in the pan-cancer dataset. To explore that possibility we divided the pan-cancer genomes into 11 tissue-specific smaller datasets and analyzed each one of them using e-Driver.

Table 1. Tissue-specific drivers identified by e-Driver

Gene symbol	PFR	Start	End	Pancan qval	Tissue qval	Tissue
CTCF	Pf00096	266	288	0.66	0.02	brca
SPOP	Pf00917	39	162	0.09	0.03	ucec
PIK3CA	Pf02192	32	108	1	1.9 e-4	ucec
EGFR	Pf07714	714	965	1	2.1 e-7	luad
EGFR	Pf00069	712	964	1	2.1 e-7	luad
BAP1	Pf01088	4	214	0.6	0.02	kirc
CTNNB1	Pf05804	334	484	0.12	0.009	ucec
ANKRD36C	IDR	543	632	0.37	0.003	hnsc
ZNF479	Pf00096	437	459	0.1	3.6 e-4	blca
FLNA	Pf00630	1158	1244	1	0.009	gbm
MTOR	IDR	1442	1492	0.07	2.7 e-4	kirc

While most PFRs have stronger signal in the pan-cancer dataset than in any tissue dataset (Figure 3a, black dots), others have stronger tissue-specific signals (Figure 3a, gray dots). This is the case, for example, of FLT3's kinase domain, which is mostly mutated in acute myeloid leukemia (17/23 mutations in this domain happen in this type of cancer). Another example is EGFR, which has two clearly different mutation patterns in glioblastoma and lung adenocarcinoma (Figure 3b). In glioblastoma it is Domain II of EGFR's extracellular region that is mostly affected by missense mutations (Domain IV seems to be also strongly mutated, although since it is not annotated in Pfam it has not been analyzed by e-

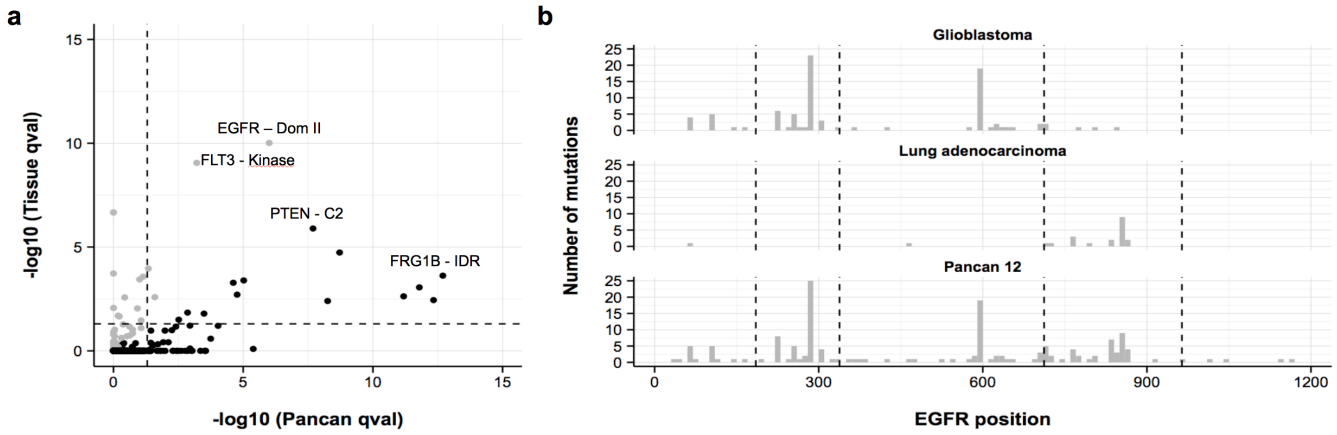

Figure 3 – Tissue-specific drivers identified by e-Driver. (a) Correlation plot showing the q values obtained for each region in the pan-cancer dataset compared to the lowest q value obtained for that region in the 11 different tissues. Dots in gray represent regions with lower tissue-specific than pan-cancer q values whereas black dots have lower pan-cancer than tissue-specific q values. Dashed lines are located in the q = 0.05 threshold that we established to consider a region as a potential driver (b) Histograms showing the mutation distribution of EGFR in three different datasets: pan-cancer (lower histogram, black), lung adenocarcinoma (middle, light gray) and glioblastoma (top, dark gray). In the pan-cancer and glioblastoma datasets only EGFR's extracellular Domain II (positions 185-338, between dashed lines) is enriched in mutations, while the kinase domain (positions 714-965, between dashed lines) shows no bias in its mutation rate. However, in lung adenocarcinoma it seems that only the kinase domain that is relevant, as most mutations (19/21, 90%) are located in such domain.

Driver), and there are almost no mutations in the kinase domain. On the other hand, in lung adenocarcinoma there are almost no mutations in the extracellular region and most mutations are located in the kinase domain of this protein.

There are 11 PFRs in 10 different proteins that can only be identified in the tissue-specific datasets (Pancan qval > 0.05, Tissue qval < 0.05, Table 1). These tissue-specific drivers are strongly enriched in cancer drivers, as 8 out of 10 proteins are part of the CGC. Besides the identifications of EGFR's kinase domain (Pf07714) in lung adenocarcinoma, which have been explained above, there are other interesting examples in the list. For example, while most PIK3CA mutations are located in the Pf00613 domain (including the well studied E545K) and happen in a variety of cancer types, the Pf02192 domain, also known as ABD domain, is mostly mutated in endometrial cancer.

4 DISCUSSION AND CONCLUSIONS

Here we evaluated the hypothesis that some cancer driver genes might accumulate mutations only in those functional regions (domains or disordered regions) that are relevant to the disease. In order to test this idea, we have developed a novel approach, e-Driver, and applied it to one of the largest available datasets of cancer genomic data, the TCGA's pan-cancer project. Our method checks, for each PFR, whether it shows a bias in its mutation rate when compared to the rest of the protein. Since it uses only mutation data for individual proteins, e-Driver, unlike other methods that compare mutation rates of whole genes, does not need to compensate for variations in mutation rates across the entire genome (De and Michor, 2011). Another novelty of our method is that protein domains and IDRs are usually larger than the clusters identified by other methods. This feature is important, as small-clusters of mutations are usually located in oncogenes rather than in tumor-suppressor genes. By using larger functional regions we can identify tumor-suppressors whose contribution to carcinogenesis depends solely on the mutation status of specific regions.

The advantages of our method are exemplified in the identification of MGA using the TCGA dataset. This gene was not mutated in enough samples to be identified by methods that rely on the mutation frequency of the whole gene (note that in a recent study with more cancer samples these methods where able to identify MGA as a potential cancer driver). Because this gene acts as a tumor suppressor the range of positions that can be mutated in order for it to drive the tumor's growth is too large to be identified by OncodriveCLUST. However, its mutations tend to accumulate in its Helix-loop-helix domain rather than in the rest of the protein, allowing e-Driver to find it.

One drawback that comes from the use of predefined regions is that if the gene has no such regions, or the regions cover the whole gene, the gene cannot by identified using our method. This is the case, for example, in IDH1 and IDH2 (Yan *et al.*, 2009). These two known cancer driver genes encode single-domain proteins. In this scenario, even though their only PFRs are frequently mutated in cancer and show clusters of mutations, e-Driver can not identify them. In the case of unannotated proteins, since there are no known PFRs, our method has nothing to compare, so they can't be analyzed. However these represent less than 10% of the human genome (and less than 3% of the proteins with at least one mutation in TCGA, supplementary Table 3). It is important to notice also, that just like most other methods that rely on mutation frequencies to identify potential drivers, e-Driver will also benefit from the increase in number of sequenced cancer genomes, as the statistical power will be larger allowing it to identify novel regions (Supplementary Figure 2).

Another scenario are proteins enriched in mutations in an unannotated region (such as EGFR's extracellular Domain IV) e-Driver will not be able to identify that specific region. In this latter case, however, as long as the protein contains an annotated PFR, e-Driver should be able to find the protein, as it will pick up the annotated PFR because of its lack of missense mutations. Another interesting feature of e-Driver is that, since it detects which PFRs are relevant for each type of cancer, it might also help in defining

strategies to design and administer drugs. For example, it has been recently shown that the two different patterns of mutations that we observed in EGFR for glioblastoma and lung adenocarcinoma have therapeutic implications as to which type of EGFR inhibitors work in each case, as they deregulate EGFR's activity through different mechanisms (Vivanco *et al.*, 2012). Another example are PIK3CA's Pf02192 and Pf00613 domains, which are also driving different subsets of cancer and that determine the response to the IGF1R inhibitor AEW541 (Porta-Pardo and Godzik, submitted).

Overall, we have shown that our approach can identify both, well known oncogenes, as well as novel cancer drivers. Moreover, because of a direct connections between protein regions and specific elements of the protein function, it can also provide further hypotheses the mechanisms of driver genes. Given the complexity of the problem of identifying cancer drivers it is likely that the combination of multiple approaches looking for distinct signals of positive selection is going to be needed in order to get to the final answer. For example, neither e-Driver of any of the other methods discussed here work with data regarding somatic copy number variations, a type of mutation that can be driving several subsets of cancer (Ciriello *et al.*, 2013). Here we have demonstrated that e-Driver can provide a novel, insightful and complementary view of the problem, contributing to its solution.

ACKNOWLEDGEMENTS

We want to thank our colleagues from the SBMRI bioinformatics group, specifically Lukasz Jaroszewski, for providing information and prediction for novel human protein domains and Thomas Hrabe for his help in preparing some of the figures.

Funding: This work has been supported by the Human Frontiers Science Program grant RGP0027/2011

Conflict of Interest: None declared.

REFERENCES

Chang,K. *et al*. (2013) The cancer genome atlas pan-cancer analysis project. *Nat. Genet.*, **45**, 1113–1120.

Ciriello,G. *et al*. (2013) Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.*, **45**, 1127–1133.

De,S. and Michor,F. (2011) DNA replication timing and long-range DNA interactions predict mutational landscapes of cancer genomes. *Nat. Biotechnol.*, **29**, 1103–8.

Dees,N.D. *et al*. (2012) MuSiC: identifying mutational significance in cancer genomes. *Genome Res.*, **22**, 1589–98.

Dunker, a K. *et al*. (2005) Flexible nets. The roles of intrinsic disorder in protein interaction networks. *FEBS J.*, **272**, 5129–48.

Futreal,P. *et al*. (2004) A census of human cancer genes. *Nat. Rev. Cancer*, **4**, 177–183.

Gonzalez-Perez,A. and Lopez-Bigas,N. (2012) Functional impact bias reveals cancer drivers. *Nucleic Acids Res.*, 1–10.

Hudson,T. *et al*. (2010) International network of cancer genome projects. *Nature*, **464**, 993–998.

Kandoth,C. *et al*. (2013) Mutational landscape and significance across 12 major cancer types. *Nature*, **502**, 333–9.

Lawrence,M.S. *et al*. (2014) Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*.

Prilusky,J. and Felder,C. (2005) FoldIndex©: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics*, **21**, 3435–3438.

Punta,M. *et al*. (2012) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290–301.

Reimand,J. *et al*. (2013) The mutational landscape of phosphorylation signaling in cancer. *Sci. Rep.*, **3**, 2651.

Reimand,J. and Bader,G.D. (2013) Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Mol. Syst. Biol.*, **9**, 637.

Tamborero,D., Gonzalez-Perez,A., Perez-Llamas,C., *et al*. (2013) Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci. Rep.*, **3**, 2650.

Tamborero,D., Gonzalez-Perez,A., and Lopez-Bigas,N. (2013) OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics*, **29**, 2238–44.

Vivanco,I. *et al*. (2012) Differential sensitivity of glioma- versus lung cancer-specific EGFR mutations to EGFR kinase inhibitors. *Cancer Discov.*, **2**, 458–71.

Wang,X. *et al*. (2012) Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nat. Biotechnol.*, **30**, 159–64.

Xu,D. *et al*. (2014) AIDA: ab initio domain assembly server. *Nucleic Acids Res.*, 1–6.

Yan,H. *et al*. (2009) IDH1 and IDH2 mutations in Gliomas. *N. Engl. J. Med.*, **360**, 765–73.

Zhong,Q. *et al*. (2009) Edgetic perturbation models of human inherited disorders. *Mol. Syst. Biol.*, **5**, 321.