BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 18 2013, pages 2238–2244
doi:10.1093/bioinformatics/btt395

Genome analysis

Advance Access publication July 24, 2013

OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes

David Tamborero[1], Abel Gonzalez-Perez[1,*] and Nuria Lopez-Bigas[1,2,*]

[1]Research Unit on Biomedical Informatics, Department of Experimental and Health Sciences, Universitat Pompeu Fabra, Dr. Aiguader 88, 08003 Barcelona and [2]Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluis Companys, 23, 08010 Barcelona, Spain

Associate Editor: John Hancock

ABSTRACT

Motivation: Gain-of-function mutations often cluster in specific protein regions, a signal that those mutations provide an adaptive advantage to cancer cells and consequently are positively selected during clonal evolution of tumours. We sought to determine the overall extent of this feature in cancer and the possibility to use this feature to identify drivers.

Results: We have developed OncodriveCLUST, a method to identify genes with a significant bias towards mutation clustering within the protein sequence. This method constructs the background model by assessing coding-silent mutations, which are assumed not to be under positive selection and thus may reflect the baseline tendency of somatic mutations to be clustered. OncodriveCLUST analysis of the Catalogue of Somatic Mutations in Cancer retrieved a list of genes enriched by the Cancer Gene Census, prioritizing those with dominant phenotypes but also highlighting some recessive cancer genes, which showed wider but still delimited mutation clusters. Assessment of datasets from The Cancer Genome Atlas demonstrated that OncodriveCLUST selected cancer genes that were nevertheless missed by methods based on frequency and functional impact criteria. This stressed the benefit of combining approaches based on complementary principles to identify driver mutations. We propose OncodriveCLUST as an effective tool for that purpose.

Availability: OncodriveCLUST has been implemented as a Python script and is freely available from http://bg.upf.edu/oncodriveclust

Contact: nuria.lopez@upf.edu or abel.gonzalez@upf.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on April 12, 2013; revised on June 13, 2013; accepted on July 4, 2013

1 INTRODUCTION

One of the current challenges of oncogenomics is to distinguish the genomic alterations that are involved in tumourigenesis (i.e. drivers), from those that give no advantage to cancer cells, but occur stochastically as a by-product of cancer development, (i.e. passengers). In line with this distinction, driver genes (i.e. those bearing driver alterations) confer selective advantage to tumour development and can be identified by detecting signals of positive selection across a cohort of tumours (Dees *et al*., 2012; Getz *et al*., 2007; Gonzalez-Perez and Lopez-Bigas, 2012; Greenman *et al*., 2007; Hodis *et al*., 2012; Reimand and Bader, 2013; Tamborero *et al*., 2013). One of the most intuitive approaches—and the most profusely used at present—to detect positive selection is based on frequency. Briefly, it consists in identifying genes more frequently mutated than the background mutation rate (Dees *et al*., 2012; Getz *et al*., 2007). We recently described a second complementary approach, OncodriveFM (Gonzalez-Perez and Lopez-Bigas, 2012), to identify genes under positive selection in tumour development by assessing their bias towards the accumulation of mutations with high functional impact (FMbias) across a cohort of tumour samples. Here we propose a third alternative and complementary approach aimed at detecting genes that bear mutations significantly clustered in specific regions of the amino acid sequence. Gain-of-function mutations in cancer genes predominantly occur at specific protein residues or active domains; for example, most KRAS mutations are found in residues 12 and 13 of the protein (Malumbres and Barbacid, 2003), and mutations in PIK3CA are predominantly found in the kinase and helical domains of the PIK3CA subunit (Karakas *et al*., 2006). However, the extent of this phenomenon in cancer genes and how it can be used to nominate drivers, not just gain-of-function cancer genes, remains to be clarified.

The clustering of mutations in the amino acid sequence of proteins—mutations clustering, for clarity—has already been suggested as a marker of positive selection (Wagner, 2007), and some approaches have been described to measure it in cancer (Stehr *et al*., 2011; Wagner, 2007; Ye *et al*., 2010; Yue *et al*., 2010). However, the methods described in previous reports assume that mutation probability is homogeneous across the gene sequence, which is likely an oversimplification that introduces a bias in the detection of meaningful events. To solve this problem, we have developed OncodriveCLUST, a novel method that measures the bias of genes towards large mutation clustering with respect to a background model composed of coding-silent mutations, which are in principle under no selective pressure and thus may reflect the baseline clustering of somatic mutations. We have applied OncodriveCLUST to the set of mutations reported in the Catalogue of Somatic Mutations in Cancer (COSMIC) (Forbes *et al*., 2010), with the objective of assessing its performance to select known cancer drivers, i.e. those found in the Cancer Gene Census (CGC) (Futreal *et al*., 2004), and to determine the extent of the mutations clustering phenomenon among known driver genes, both those described as oncogenes and

*To whom correspondence should be addressed.

tumour suppressors. We have also applied OncodriveCLUST to several available large datasets of cancer somatic mutations from The Cancer Genome Atlas (TCGA) initiative (Consortium *et al*., 2008), and compared these results with other methods to identify driver candidates, namely OncodriveFM (Gonzalez-Perez and Lopez-Bigas, 2012) and MutSig (Getz *et al*., 2007). Because all three approaches measure complementary and independent criteria aimed to discover positive selection (i.e. mutation clustering, FMbias and mutation frequency), and each of them likely detects important events, we conclude that the optimal solution is to combine them to retrieve the most reliable and comprehensive list of cancer drivers from datasets of somatic mutations of diverse tumour types.

2 METHODS

2.1 OncodriveCLUST implementation

The OncodriveCLUST method comprises the following five steps, illustrated in Figure 1: First, single-nucleotide protein-affecting mutations (i.e. non-synonymous, stop and splice site mutations) are retrieved (panel I). Second, positions with a number of mutations above a background rate threshold (those with a $\leq$1% probability of occurrence, according to the binomial cumulative distribution function, which takes into account both the gene length and the overall number of gene mutation) are identified as potentially meaningful cluster seeds (panel II). Third, these positions are grouped to form clusters, joining positions that fall within distances of five or less amino-acid residues (panel III). Fourth, once these clusters are obtained, they are completed by including the positions within or adjacent to each cluster that contains mutations in addition to those considered in the second step (panel IV). Finally, a score is computed for each cluster. This score is directly proportional to the percentage of mutations grouped within the cluster and inversely proportional to its length, as shown in the equation,

$$Clustering\ score = \sum_i \frac{fractionMutations}{(\sqrt{2})^{distance}}$$

where i represents protein positions within the cluster, *fractionMutations* is the percentage of mutations falling in that position (out of the total observed in the protein across samples) and *dist* is the number of amino acids spanning between i and the position of the cluster with the largest number of mutations, i.e. its peak. Note that this score ranges between 0 and 1, where 1 means that a single position concentrates all the mutations observed in the gene across samples. The rationale behind considering the fraction of mutations instead of their absolute figure is to avoid overestimating the score of clusters in frequently mutated genes. On the other hand, inversely weighting the clustering score with the cluster length (see denominator) tends to favour genes with mutations concentrated in narrower regions (Supplementary Fig. S1). To complete the final step, a gene clustering score is obtained by summing the scores of all its clusters. Finally, the significance of the observed gene clustering score is estimated by comparing it with the background model distribution (panel V), obtained by calculating all gene clustering scores of the coding-silent mutations. Because the distribution of the null model is sufficiently close to normal, the significance is computed through a Z-score, which is then transformed into a P value and corrected for multiple testing using the false discovery rate approach. Therefore, this corrected P value measures the clustering bias of the protein-affecting mutations of a gene, compared with that of the coding-silent mutations measured across the dataset.

OncodriveCLUST has been implemented as a Python script (downloadable from http://bg.upf.edu/oncodriveclust), which requires a file stating the position of each mutation within the protein sequence as input (see the user's guide for further details). The preparation of this file is at the user's discretion; note that the selection of the protein isoform may cause, among other factors, different consequence types of the mutations. In the present manuscript, we have processed the mutation data by using an in-house pipeline that selected the largest isoform. In addition, the launcher allows the use of several optional arguments to customize the values used by OncodriveCLUST internal calculations. Moreover, the user may also choose whether to output the details of the identified clusters and the protein domains to further explore the occurrence of mutations in known functional regions. Finally, the number of coding-silent mutations of the analysed dataset must ensure the construction of a background model. For those cases in which this is not possible owing to a reduced number of samples and/or a low occurrence of synonymous mutations, the method can be run with an 'external' background model computed from 26 datasets of different large cancer exome sequencing projects (Supplementary Table SI) that are publicly available. However, the results of this analysis have to be carefully considered because it may be subjected to several inaccuracies due to technical (e.g. differences in mutation calling pipelines) and biological (e.g. differences in mutation clustering between tumours) considerations.

2.2 Application to the COSMIC dataset

Somatic mutations were downloaded from the COSMIC v.62 dataset (Forbes *et al*., 2010) and analysed by OncodriveCLUST including only those genes with at least 10 single-nucleotide protein-affecting mutations across all tumour samples in the dataset. Only entries labeled as genome-wide screen were selected to avoid the inclusion of mutations biased towards more intensively studied positions, which are overrepresented in COSMIC. In addition, and also to ensure a fair assessment of mutation clustering, we have also excluded mutations found in metastases, secondary and recurrent tumours, as well as mutations not explicitly declared as somatic, and mutations found in cell lines and xenografts. We obtained 233 775 protein-affecting mutations, whereas the background model was constructed using 37 338 coding-silent mutations.

2.3 Application to whole exome sequencing TCGA datasets

We downloaded somatic mutations of four TCGA datasets whose moratorium period has expired: breast invasive carcinoma (BRCA) (Bell *et al*., 2011), lung squamous cell carcinoma (LUSC) (Hammerman *et al*., 2012), ovarian serous carcinoma (OV) (Bell *et al*., 2011) and uterine corpus endometrioid carcinoma (UCEC) (Getz *et al*., 2013). These datasets possess a large number of silent mutations, thus producing a good background model; in addition, MutSig results were available for all of them. They were downloaded from the latest available run in October 2012 at the Broad Institute Firehose system.

We removed from the OncodriveCLUST analysis samples with abnormally high number of mutations in each dataset, i.e. greater than the median of the entire distribution as compared with the remaining samples. These were identified as outliers of the distribution of mutations per sample in each dataset. Overall, 12 such outliers appeared in UCEC but none in the remaining datasets. OncodriveCLUST assessed significant clustering of all genes with protein-affecting mutations in at least three samples of each cancer dataset. As stated above, the results of OncodriveCLUST in each TCGA dataset were compared with the ones of a frequency-based analysis (MutSig) and a functional impact bias analysis (OncodriveFM). All the analyses aimed to find driver candidates were performed using the same mutation data. OncodriveFM was run for all genes with protein-affecting mutations in more than two samples of the OV and BRCA datasets, and in more than five samples in UCEC and LUSC. Heatmaps depicted in the present manuscript were constructed using Gitools (Perez-Llamas and Lopez-Bigas, 2011). Venn diagrams were created with the BioVenn application (Hulsen *et al*., 2008).

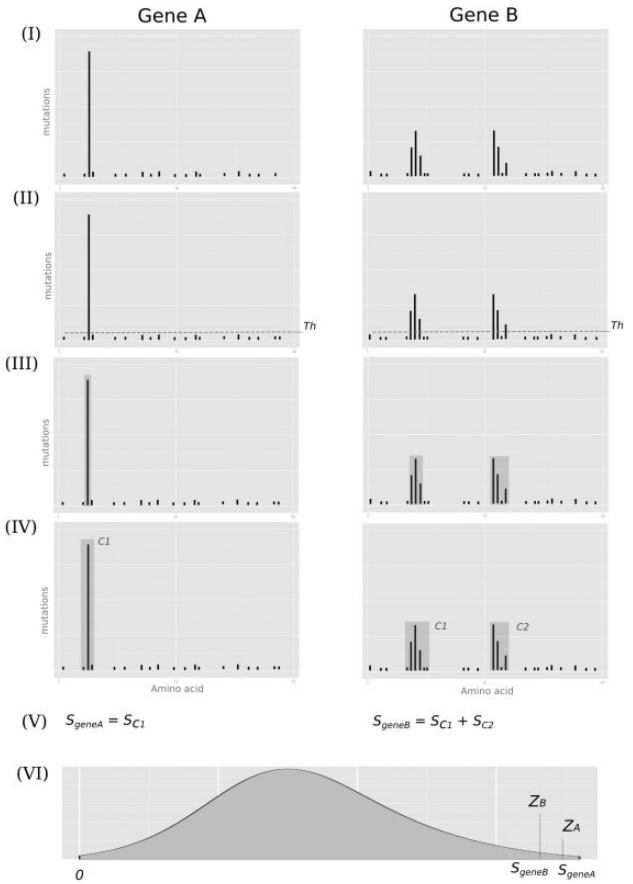

Fig. 1. Schematic representation of the steps performed by OncodriveCLUST to assess the mutation clustering of two dummy genes. (I) Input data represented by the histogram of protein-affecting gene mutations, where the x axis is the protein sequence position and the y axis is the percentage of the gene mutations that occurs in that position. (II) Meaningful positions are selected as those having a number of mutations that is not expected by chance. This is calculated for each gene depending on its length and its number of mutations according to the binomial cumulated distribution function, which is a loose criteria according to our non-uniform distribution of mutations hypothesis (probability <1% by default to state a position as meaningful). In the present panel, the minimum number of mutations calculated for each gene as the meaningful threshold is represented by a dotted line labelled as 'Th'. (III) Meaningful positions found in step (II) are grouped to form mutation clusters (shadowed in grey in the presented panel). Two consecutive positions are grouped within the same cluster if there is a default maximum distance of 5 amino acids between them. This is considered a stringent cut-off to define that two mutations in different residues are affecting the same tumourigenic mechanism, as stressed by the fact that most (95%) of the coding-silent mutations, which are assumed to be stochastically distributed, of the COSMIC dataset are separated by larger distances. Note that if no meaningful position is found across the protein sequence, the gene is considered to have no mutation clusters and therefore no further analysis is performed. In the present example, gene A has a single meaningful position and thus it forms a single cluster; in gene B, there are 6 meaningful positions that are grouped forming two separate clusters. (IV) Mutations that do not occur in the meaningful positions found in (II) but are either enclosed within the boundaries of the clusters defined in (III) or adjacent to them are also accounted for the final figure of the number of mutations enclosed by that cluster. (V) Thereafter, a score is calculated for each cluster. This score is proportional to the number of enclosed mutations and inversely related to the cluster length (see the 'Methods' section for further details about the clustering score calculation). A final value per gene is obtained by summing the scores of each of the clusters found in that gene (C1 for gene A and C1 plus C2 for gene B). (VI) Finally, to estimate the significance of each gene score obtained in (V), this is compared with the distribution of the background model. The background model is obtained by calculating the clustering scores of all the dataset genes following steps (I–V) but analysing the coding-silent mutations. Note that such comparison allows to obtain a standard score (i.e. a Z score) after the normal distribution of the background model has been validated, and a corrected *P* value is thereafter derived

3 RESULTS

3.1 Motivation and hypothesis

Mutation recurrence and the accumulation of functional mutations across samples have both been used to identify genes with positively selected mutations during tumourigenesis, which are thus good candidates to play an important role in this process. On the other hand, the physical grouping of mutations in certain protein regions—although recognized as a signal of positive selection—has not been exploited as thoroughly for this purpose. For instance, the MuSiC tool, which is aimed to identify recurrently mutated genes, also points at genes with a high density of mutations, through the so-called proximity analysis (Dees *et al.*, 2012), but it does not attempt to estimate the significance of such observation. Other approaches developing an analytical framework, base their calculations on the oversimplified assumption of a homogeneous background distribution of mutations (see below) (Ye *et al.*, 2010). In this study, we present a novel computational method able to identify genes biased towards a significant clustering of mutations within regions of the protein sequence and we assess its ability to nominate cancer drivers. Our method defines clusters by grouping together mutations that are in physical proximity. Also, based on the acknowledgment that mutations do not occur with equal probability on all positions of a gene, the OncodriveCLUST background model relies on the degree of clustering of silent mutations observed across genes in the dataset under analysis.

3.2 Silent mutations do not follow a uniform distribution

Previously described methods to assess the significance of mutation clustering assume that the baseline mutation probability is homogeneous across all gene positions (Stehr *et al.*, 2011; Wagner, 2007; Ye *et al.*, 2010; Yue *et al.*, 2010), an assumption that seems oversimplified according to recent evidences of non-random mutation processes along the genome (Amos, 2010; Liu *et al.*, 2013; Martincorena *et al.*, 2012; Roberts *et al.*, 2012). As a first step to overcome this problem, our method measures the clustering of coding-silent mutations, which are assumed not to be under selective pressure and may thus reflect the baseline tendency of somatic mutations to preferentially occur at certain positions of the protein sequence.

We collected COSMIC entries annotated as whole genome screens and we constructed the background model using the synonymous mutations found in this dataset. In general, coding-

silent mutations did not follow a uniform distribution but were instead noticeably grouped (Supplementary Fig. S2). Therefore, protein-affecting mutations that are not clustered beyond this point may reflect spurious findings, stressing the benefit of using a non-homogeneous mutations distribution null model to retrieve more specific results. Interestingly, some of the clustered coding-silent mutations may well represent unknown functional variants positively selected during the tumourigenic process and deserve further study, which is beyond the scope of the present manuscript. In any case, we assume that the vast majority of coding-silent variants are under neutral selection; therefore, we used their general clustering distribution to construct our background model.

3.3 Genes with mutation clustering detected by OncodriveCLUST are enriched for known cancer genes

Analysis of the COSMIC dataset included 9565 genes that contained at least 10 protein-affecting mutations annotated as whole genome screen. Of these, 6024 showed at least one cluster of mutations according to the OncodriveCLUST criteria and were assessed for significance. Overall, 123 genes were biased towards a larger mutation clustering as compared with the aforementioned background model (corrected P value <5%). The output of this analysis is provided in Supplementary File S1. Genes selected by OncodriveCLUST were enriched for known cancer drivers included in the CGC (27/122 = 22% versus 232/9444 = 2.4%). Most of these (19/27 = 70%) possessed a dominant phenotype according to the CGC annotation. Nevertheless, we also detected significant clusters among recessive cancer genes, which were wider, as a rule. These results support the hypothesis that clustering is more pronounced among genes that experience gain-of-function, although mutations leading to loss-of-function in some genes also occur predominantly within certain regions of the protein (Fig. 2).

Next, we used OncodriveCLUST to analyse the four selected TGCA datasets. In all of them, the genes nominated by OncodriveCLUST were enriched for known cancer drivers (Supplementary Fig. S3). The detailed output of these analyses is provided in Supplementary Files S2–S5. The observation that known cancer genes were consistently selected by OncodriveCLUST as bearing significant mutations clusters demonstrate that this method is also useful to nominate novel drivers.

3.4 Using OncodriveCLUST to complement results of other methods to identify drivers

There are crucial cancer genes that cannot be identified by the mutation clustering approach, such as tumour suppressors whose loss-of-function mutations are evenly distributed across the sequence. Moreover, the ability to measure the mutations clustering depends on the number of observed mutations. Therefore, this approach in general is not suitable to detect lowly recurrent drivers. As a consequence, the analysis of somatic mutations in a tumour cohort with OncodriveCLUST may provide a list of reliable but incomplete candidate drivers. The retrieval of a more comprehensive driver list requires the combination of several

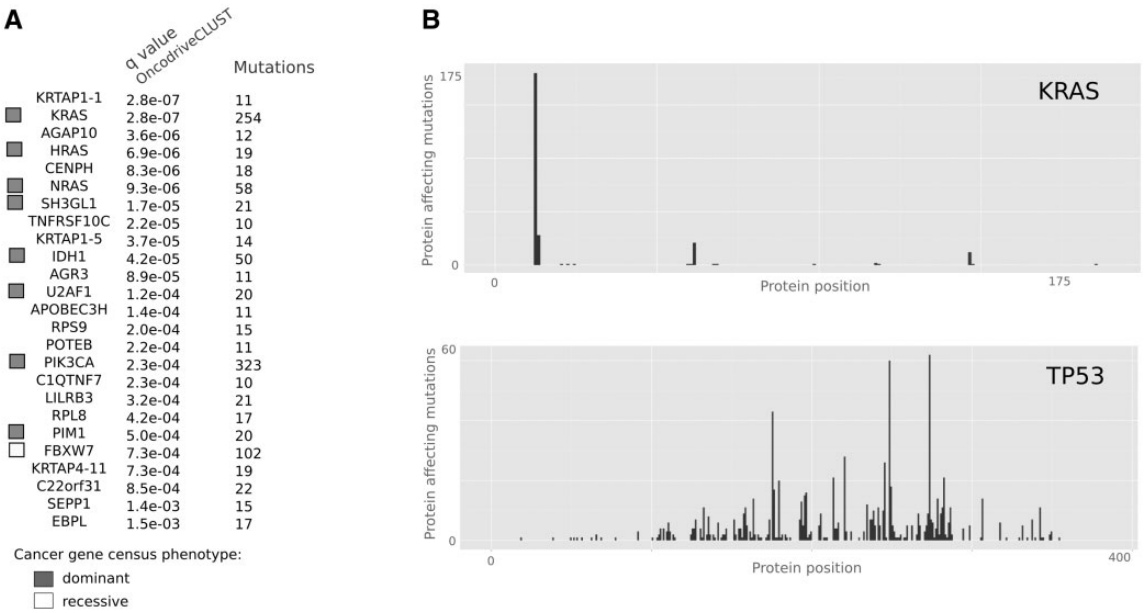

Fig. 2. Selected top-ranking genes from the OncodriveCLUST analysis on COSMIC. Panel A depicts the results for the 25 top-ranking genes obtained by OncodriveCLUST from the analysis of the COSMIC dataset. Gene selection was enriched for genes found in the CGC, predominantly those annotated with a dominant phenotype. The label mutations indicate number of samples with protein-affecting mutations across entries annotated as whole gene screen. Panel B depicts the mutation histograms of two well-known cancer drivers, namely KRAS and TP53. The former illustrates the ideal expected behavior of an oncogene, where mutations occur in specific regions that cause the selected gain-of-function, and this gene obtained the largest cluster bias from the COSMIC dataset. On the other hand, TP53 loss-of-function is the result of mutations that occur in wider regions of the protein sequence; although they are less clustered, mutations in this gene still show a high degree of spatial affinity (corrected OncodriveCLUST P value of 0.03)

methods that use complementary criteria to identify genes bearing signals of positive selection. We anticipate that the analysis of mutations clustering may combine well with OncodriveFM, a method based on the accumulation of mutations with high functional impact within genes as a marker of positive selection. Because there are other methods aimed to identify driver mutations, we also compared the Oncodrive results with those obtained by a current well-established approach, MutSig, which is based on analysing whether a gene is mutated more often than expected by chance across a cohort of tumour samples.

As a result of this comparison, we found some genes identified by OncodriveCLUST that were missed by the other methods. Some of these are found in the CGC, and are thus likely drivers identified by clustering criteria but not by functional impact or recurrence analysis; interestingly, most of them (9 out of 11) exhibit a dominant phenotype (Fig. 3). In addition to genes included in the CGC, OncodriveCLUST is also able to highlight novel driver candidates that are overlooked by other methods. For instance, among the 55 genes with significant mutation clusters in the BRCA dataset, 38 were not identified by other methods—as PLA2G3 (involved in cell growth and death), MMP14 (matrix metalloproteinase), REV1 (involved in DNA repair) and the caspase CASP5. These may be regarded as potential candidates to drive the emergence of the tumour phenotype that were selected only by the OncodriveCLUST method (Supplementary Fig. S3).

4 DISCUSSION

The elucidation of cancer drivers relies on identifying the marks of positive selection that occur during the clonal evolution of tumours. The trend shown by protein-affecting mutations to accumulate predominantly in certain gene regions is a fingerprint that may denote events targeted by the tumourigenesis. However, it is now apparent that the probability of occurrence of somatic mutations is affected by several factors and varies across the genomic sequence (Amos, 2010; Liu *et al*., 2013; Martincorena *et al*., 2012; Roberts *et al*., 2012). This should be taken into account when identifying mutation clusters because the assumption of a uniform mutation distribution as the null model may overestimate the significance of observed clusters. However, the factors determining the occurrence of mutational hotspots are not fully understood, and their inference is not straightforward by mining the data under analysis to obtain a background model per gene and sample. Instead, we propose to construct the background model using the degree of clustering of synonymous mutations, which are assumed not to be under positive selection and may thus reflect the baseline mutation

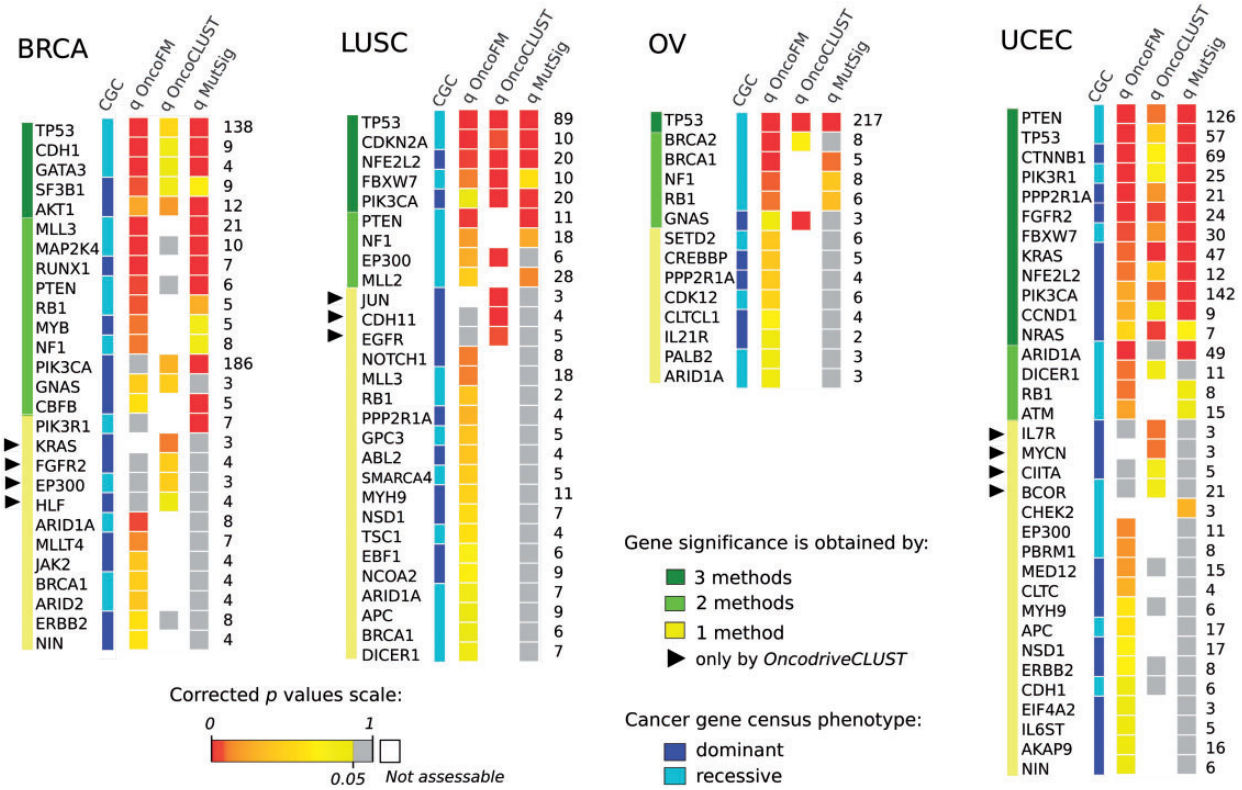

Fig. 3. Selected top-ranking genes from the OncodriveCLUST analysis on four TCGA datasets. Summary of the results obtained by three methods aimed to find driver genes for the four analysed datasets retrieved from TCGA project: OncodriveFM (functional impact criteria), OncodriveCLUST (mutation clustering criteria) and MutSig (mutation frequency criteria). Only genes annotated in the CGC are depicted. Combination of approaches based on different theoretical principles confers different levels of evidence. Note that OncodriveCLUST selected genes likely to be involved in the disease, several of them missed by the other methods. The label 'Not assessable' denotes genes whose functional impact metrics cannot be calculated (OncodriveFM), or genes with no mutation cluster (OncodriveCLUST)

clustering of the tumour. This assumption also probably comprises a simplification because some coding-silent mutations could in principle alter processes such as chromatin remodeling or mRNA processing. Nevertheless, apart from this, they are not in general functionally involved in tumourigenesis. Moreover, and owing to the low-recurrence of synonymous mutations, the background model cannot establish constraints for particular genomic regions. Therefore, the method presents several caveats, which have to be taken into account. However, although limited, this null model shall in principle specifically support the identification of genes whose mutations are grouped above the background. As a matter of fact, this approach was proven successful in prioritizing genes involved in the disease because OncodriveCLUST retrieved a reduced list of genes enriched for cancer drivers in all the datasets analysed. In general, the clustering analysis better detected drivers known to exhibit a dominant phenotype because mutations in these genes occur at particular sites that lead to the favoured gain-of-function. However, recessive genes were also highlighted over the background model because their mutations occur in larger but still delimited regions leading to loss-of-function.

The identification of drivers can profit from the combination of methods based on different theoretical principles, in particular if they use complementary criteria. In this regard, OncodriveCLUST complements OncodriveFM because it better captures the behaviour of gain-of-function cancer genes, whereas OncodriveFM is better at identifying tumour suppressor genes, which bear loss-of-function mutations that usually receive higher functional impact scores. By comparison, mutations conferring gain-of-function are more prone to produce lower scores of functional impact. For instance, mutations that change residue 1047 of PIK3CA are known to be oncogenic but are underestimated by functional impact metrics because they rely largely on conservation criteria and this particular residue is highly variable across species. Therefore, OncodriveFM overlooks this gene in the BRCA dataset because most PIK3CA mutations occur in that specific position. Nevertheless, this scenario in which gene mutations predominantly occur at certain protein positions is well highlighted by the OncodriveCLUST analysis, illustrating how the bias of a method is captured by using complementary criteria.

At present, the most widely used approach to identify driver genes consists in detecting those that are mutated more frequently than expected by chance. Therefore, we compared the OncodriveCLUST results obtained in the TGCA datasets with those obtained by MutSig (Getz *et al*., 2007), a well-established method to detect significantly mutated genes across cohorts of tumour samples. To avoid further interpretation, we limited the comparison with already well described drivers, i.e. genes contained in the CGC. As a result, we demonstrated that OncodriveCLUST selected several cancer drivers that were missed not only by OncodriveFM but also by the recurrence analysis. Of note, this suggests that OncodriveCLUST could identify novel driver candidates (i.e. not included in the CGC) that are nevertheless overlooked by other methods and emphasizes the benefit of using a combination of methods to elucidate driver genes that balances the strengths and weaknesses of each approach. This would allow the retrieval of a more comprehensive list of putative drivers because each may exhibit different signals of positive selection. In addition, it could also help to estimate the reliability of the results because the false positives of a method are unlikely to be picked up by another one based on complementary criteria and therefore with its own sources of bias.

One important advantage of OncodriveCLUST, and of OncodriveFM as well, is that the required input consists only on the list of tumour somatic mutations in the cohort. Raw data, such as BAM files are not required, which reduces the burden of download, storage and computational processing. This is a major issue because the amount of data generated by tumour genome re-sequencing studies continues to increase and the technical requirements of some methods could limit the analysis of large datasets. This is specially true for initiatives aimed at integrating data from different projects [e.g. IntOGen (Gundem *et al*., 2010)].

In summary, the elucidation of genes involved in cancer is a challenging task that requires the combined use of approaches based on different criteria. In this regard, we show that Oncodrive CLUST complements well other existing methods and should be taken into account for the identification of cancer drivers.

ACKNOWLEDGEMENTS

We acknowledge the task of the reviewers, which helped to improve the present manuscript.

Funding: Spanish Ministry of Economy and Competitivity (grant number SAF2009-06954 and SAF2012-36199) and Spanish National Institute of Bioinformatics.

Conflict of Interest: none declared.

REFERENCES

Amos,W. (2010) Even small SNP clusters are non-randomly distributed: is this evidence of mutational non-independence? *Proc. Biol. Sci.*, **277**, 1443–1449.

Bell,D. *et al*. (2011) Integrated genomic analyses of ovarian carcinoma. *Nature*, **474**, 609–615.

Consortium,T.C.G.A. *et al.* (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**, 1061–1068.

Dees,N.D. *et al.* (2012) MuSiC: Identifying mutational significance in cancer genomes. *Genome Res*., **22**, 1589–1598.

Forbes,S.A. *et al*. (2010) COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer. *Nucleic Acids Res*., **38**, D652–D657.

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

Getz,G. *et al.* (2007) Comment on "The consensus coding sequences of human breast and colorectal cancers". *Science*, **317**, 1500.

Getz,G. *et al*. (2013) Integrated genomic characterization of endometrial carcinoma. *Nature*, **497**, 67–73.

Gonzalez-Perez,A. and Lopez-Bigas,N. (2012) Functional impact bias reveals cancer drivers. *Nucleic Acids Res*., **40**, e169.

Greenman,C. *et al.* (2007) Patterns of somatic mutation in human cancer genomes. *Nature*, **446**, 153–158.

Gundem,G. *et al.* (2010) IntOGen: integration and data mining of multidimensional oncogenomic data. *Nat Meth*., **7**, 92–93.

Hammerman,P.S. *et al.* (2012) Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, **489**, 519–525.

Hodis,E. *et al.* (2012) A landscape of driver mutations in melanoma. *Cell*, **150**, 251–263.

Hulsen,T. *et al.* (2008) BioVenn-a web application for the comparison and visualization of biological lists using area-proportional Venn diagrams. *BMC Genomics*, **9**, 488.

Karakas,B. *et al.* (2006) Mutation of the PIK3CA oncogene in human cancers. *Br J. Cancer*, **94**, 455–459.

Liu,L. *et al.* (2013) DNA replication timing and higher-order nuclear organization determine single-nucleotide substitution patterns in cancer genomes. *Nat. Commun.*, **4**, 1502.

Malumbres,M. and Barbacid,M. (2003) RAS oncogenes: the first 30 years. *Nat. Rev. Cancer*, **3**, 459–465.

Martincorena,I. *et al.* (2012) Evidence of non-random mutation rates suggests an evolutionary risk management strategy. *Nature*, **485**, 95–98.

Perez-Llamas,C. and Lopez-Bigas,N. (2011) Gitools: analysis and visualisation of genomic data using interactive heat-maps. *PLoS One*, **6**, e19541.

Reimand,J. and Bader,G.D. (2013) Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Mol. Syst. Biol.*, **93**, 637.

Roberts,S.A. *et al.* (2012) Clustered mutations in yeast and in human cancers can arise from damaged long single-strand DNA regions. *Mol. Cell*, **46**, 424–435.

Stehr,H. *et al.* (2011) The structural impact of cancer-associated missense mutations in oncogenes and tumor suppressors. *Mol. Cancer*, **10**, 54.

Tamborero,D. *et al.* (2013) Oncodrive-CIS: a method to reveal likely driver genes based on the impact of their copy number changes on expression. *PloS One*, **8**, e55489.

Wagner,A. (2007) Rapid detection of positive selection in genes and genomes through variation clusters. *Genetics*, **176**, 2451–2463.

Ye,J. *et al.* (2010) Statistical method on nonrandom clustering with application to somatic mutations in cancer. *BMC Bioinformatics*, **11**, 11.

Yue,P. *et al.* (2010) Inferring the functional effects of mutation through clusters of mutations in homologous proteins. *Hum. Mutat.*, **31**, 264–271.