BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 20 2013, pages 2539–2546
doi:10.1093/bioinformatics/btt445

Genome analysis

Advance Access publication August 16, 2013

Oncofuse: a computational framework for the prediction of the oncogenic potential of gene fusions

Mikhail Shugay, Iñigo Ortiz de Mendíbil, José L. Vizmanos and Francisco J. Novo*

Department of Genetics, University of Navarra. 31008 Pamplona, Spain

Associate Editor: John Hancock

ABSTRACT

Motivation: Gene fusions resulting from chromosomal aberrations are an important cause of cancer. The complexity of genomic changes in certain cancer types has hampered the identification of gene fusions by molecular cytogenetic methods, especially in carcinomas. This is changing with the advent of next-generation sequencing, which is detecting a substantial number of new fusion transcripts in individual cancer genomes. However, this poses the challenge of identifying those fusions with greater oncogenic potential amid a background of ‘passenger’ fusion sequences.

Results: In the present work, we have used some recently identified genomic hallmarks of oncogenic fusion genes to develop a pipeline for the classification of fusion sequences, namely, Oncofuse. The pipeline predicts the oncogenic potential of novel fusion genes, calculating the probability that a fusion sequence behaves as ‘driver’ of the oncogenic process based on features present in known oncogenic fusions. Cross-validation and extensive validation tests on independent datasets suggest a robust behavior with good precision and recall rates. We believe that Oncofuse could become a useful tool to guide experimental validation studies of novel fusion sequences found during next-generation sequencing analysis of cancer transcriptomes.

Availability and implementation: Oncofuse is a naive Bayes Network Classifier trained and tested using Weka machine learning package. The pipeline is executed by running a Java/Groovy script, available for download at www.unav.es/genetica/oncofuse.html.

Contact: fnovo@unav.es

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

Received on March 28, 2013; revised on July 4, 2013; accepted on July 29, 2013

1 INTRODUCTION

One of the most common types of somatic mutation in the human cancer gene census involves chromosomal translocations that fuse two different genes, resulting in a chimeric transcript (Futreal *et al.*, 2004). It is estimated that ~20% of all cancers are caused by gene fusions driven by chromosomal translocations (Nambiar *et al.*, 2008), and the number of reported clinical cases is growing rapidly. There is compelling evidence that fusions represent an initial event in oncogenesis (Mitelman *et al.*, 2007) and play an important role in aggressive cancer cases (Villanueva, 2012). Many gene fusions are translated as fusion proteins with oncogenic potential due to the presence of protein domains that are normally located in separate proteins. Examples of recurrent fusions are BCR-ABL1, which is present in chronic myeloid leukemia cases (Ren, 2005), or TMPRSS2-ERG, found in prostate cancers (Rosen *et al.*, 2012). Most chromosomal translocations involved in cancer cases are described in several databases. For example, hundreds of gene fusions associated with clinical reports are included in Mitelman database (http://cgap.nci.nih.gov/Chro-mosomes/Mitelman), although it does not contain sequence data. Our database of fusion sequences TICdb (Novo *et al.*, 2007; http://www.unav.es/genetica/TICdb/) allows the precise mapping of breakpoint positions and facilitates analysis of the structure of fusion genes. Another database, chimerDB2.0 (Kim *et al.*, 2010; http://ercsb.ewha.ac.kr:8080/FusionGene/), contains several thousand putative fusions predicted solely from sequences present in publicly available databases.

Owing to the complex type of rearrangements sustained by the genomes of solid tumors, most gene fusions have been identified and characterized in hematological (HEM) neoplasms. However, novel next-generation sequencing (NGS) technologies coupled with sophisticated algorithms allow the detection of gene fusions with extraordinary sensitivity even in solid tumors: some reliable predictions detect fusion junctions at <0.05 sequencing reads per million (Kim and Salzberg, 2011; Sakarya *et al.*, 2012). It is reasonable to expect that this unprecedented sensitivity will lead to the discovery of hundreds of novel and rare fusion sequences in various cancer types, raising the question of how many of these mutations are important for cancer development and not just passenger events. This is even more relevant in the context of emerging data obtained from high-throughput studies (Frenkel-Morgenstern *et al.*, 2012), which demonstrate the presence of fusion transcripts in normal cells, as had been suggested in earlier studies (Akiva *et al.*, 2006; Parra *et al.*, 2006).

To solve this issue, a deep understanding of fusion-driven oncogenesis is required. The enrichment of certain combinations of functional domains and domain families in gene fusions has been highlighted by previous studies (Frenkel-Morgenstern *et al.*, 2012; Ortiz de Mendíbil *et al.*, 2009). Likewise, we have recently identified some genomic hallmarks of oncogenic fusion genes using information from public databases (Shugay *et al.*, 2012). In addition to promoter or untranslated region (UTR) swapping that leads to well-known expression changes in fusion proteins, these hallmarks include the combination of specific protein domains and protein interaction interfaces (PIIs) with novel oncogenic functions, as well as features related to replication timing. We reasoned that this set of features could be used to predict the oncogenic potential of novel fusion sequences found in tumor

*To whom correspondence should be addressed.

samples. In this work, we present Oncofuse, a robust Bayesian classifier that identifies fusion genes that could behave as 'drivers' of the oncogenic process, based on their shared properties with known oncogenic fusions.

A previous work by Wang *et al.* (2009) represented the first attempt to distinguish 'driver' and 'passenger' mutations by incorporating interactome, pathway and Gene Ontology (GO) information to rank the genes under consideration. Here, we have followed a different approach, as our goal is to predict whether a given fusion sequence is 'driver'. In this regard, our tool is focused on fusion sequences, whereas previous strategies were gene-centered (predicting whether a gene is likely to undergo a fusion). Notably, a recent article deals with the problem of distinguishing 'driver' and 'passenger' missense point mutations (not gene fusions) using a classifier trained on a large number of protein sequence features from positive and negative sets (Tan *et al.*, 2012). Although not directly comparable with our pipeline, that study highlights the potential of classification strategies for the identification of 'driver' mutations in cancer.

2 METHODS

2.1 Gene sets used in study

Supplementary Table S1 provides details of the datasets used in this study. Fusions reported in cancer patients with mapped breakpoint positions were taken from TICdb3.0, which included 1134 gene fusion records as of March 2012 and were used as positive training set. These fusions were filtered so as to select unique fusions (with respect to domain composition), thus creating a non-redundant set for classifier training. Mitelman database of fusions in cancer (http://cgap.nci.nih.gov/Chromosomes/Mitelman) was used to make an estimate of the recurrence of specific fusions. Only unique FPG pairs were selected in this case.

Two datasets were used as negative training set: fusion genes (Frenkel-Morgenstern *et al.*, 2012) and read-through transcripts (Nacu *et al.*, 2011) found in normal cells. Two positive datasets were used for validation purposes: fusions discovered in NGS studies in cancer (Asmann *et al.*, 2012; Benelli *et al.*, 2012; Edgren *et al.*, 2011; Sakarya *et al.*, 2012) and fusions predicted in chimerDB2.0 (Kim *et al.*, 2010) based on combined mRNA, expressed sequence tag (EST) and Sequence Read Archive (SRA) predictions of confidence Class A (i.e. predictions where both partner genes show reliable alignments). We filtered out fusions in which genomic data for one FPG were not available.

All RefSeq genes (taken from UCSC Genome Browser repository) and known oncogenes (according to Sanger Cancer Genome Census, CGC) were used to test the classifier on non-translocated (unbroken) genes.

2.2 Classification procedure

A Naive Bayes Network Classifier was trained and tested using Weka machine learning package (Frank *et al.*, 2004). We chose a bayesian classifier over other algorithms, such as Support Vector Machines (SVM), because it is simple yet robust, and it allows native handling of missing data values, which is always the case when integrating several high-throughput biological datasets from different sources. Moreover, scores are easily interpretable as *P*-values owing to the probabilistic nature of the classifier. The classifier was trained using 10-fold cross-validation, and feature values were discretized using Supervised Discretization algorithm in Weka package.

We selected 24 features for classification of 'driver' fusions based on our previous analysis of fusion gene hallmarks (Shugay *et al.*, 2012). These include 12 features 'retained' and 12 features 'lost' in the resulting fusion gene (e.g. the promoter of the 5′ FPG is retained, whereas the promoter of the 3′ FPG is lost, and so on). Both 'retained' and 'lost' features belong to three categories: six functional profile features (generated as described in the next section), two promoter features, one 3′-UTR feature and three protein interaction features. The complete list of features, as well as additional information on source datasets, is available in Supplementary Table S2.

2.3 Functional profiling pipeline

To generate functional profiles for fusion sequences, we first had to identify which functional families are enriched in the set of 'driver' fusions. To this end, we used all fusion sequences from TICdb to query InterPro database (Hunter *et al.*, 2012) and extract protein features using a local copy generated and downloaded via BioMart. Protein domains were 'lost' or 'retained' after fusion depending on their position relative to fusion junctions. For each protein domain, a representative set of genes was created carrying at least one instance of that domain. These gene sets were then subjected to functional analysis using Functional Clusterization tool at DAVID web server (Huang *et al.*, 2009). This pipeline yielded an extended set of 4907 genes containing the protein domains present in the fusions from TICdb (359 unique genes). Manual inspection of the results led to the identification of six categories with E-score >3, representing six functional families that are significantly enriched in the set of 'driver' fusion genes: transcription factor (TF), kinase activity, transcription co-factor, GTPase (G), helicase/histone modifiers (H) and protein binding (P). For each of these functional families, we obtained associated GO molecular function (GO:MF) terms at a false-discovery rate (FDR) <10%. The full list of GO:MF-functional family associations is presented in Supplementary Table S3.

The generation of functional profiles for new fusion genes follows the pipeline described in Supplementary Figure S1. For each new fusion sequence, protein domains are extracted, and a representative set of genes containing those domains is generated as mentioned previously. Then, a functional family association score (FFAS) is calculated as the number GO:MF terms contained in this gene set that overlap with GO:MF terms associated with each of the six functional families described before. These scores are the six functional profile features used in the classifier.

2.4 Software packages used

Gene annotation analysis was carried out using DAVID web server. Classification was performed using Weka package. GraphPad Prism was used for statistical tests. All other tasks were performed manually or using custom Java/Groovy code, available on request.

2.5 Implementation and availability

The pipeline is executed by simply running a Java/Groovy script with some parameters on a standardized input file (all required packages are installed automatically via Groovy/Grape). Users should provide IDs of fusion partner genes, as well as location of breakpoint (Intron/Exon ID and coordinate) within the major Refseq transcript of each gene. All necessary files with features are included as part of the package. Users could also specify custom data files, e.g. a file with expression data in a tissue of interest. The output is a simple tab-delimited file with classifier *P*-values (probability of belonging to Class 0 or to Class 1), domains retained in the fusion and association scores for each of the six functional profiles. Scripts and documentation are available at www.unav.es/genetica/oncofuse.html. Oncofuse is straightforward to use and supports multiple input formats to handle the output of several existing fusion detection software.

3 RESULTS

3.1 Functional profiling of fusion proteins

In our previous work, we classified proteins into functional categories using Interpro domain annotations. Such approach is not practical in the present study, as the process would have to be performed manually every time that a new list of candidate fusion sequences is analyzed by the classifier. Therefore, we have developed an automated method that infers functional profiles of proteins based on indirect evidence from GO:MF annotations of the genes containing the same domains (see Section 2), which allows data to be digitized and mapped to a finite set of features. This strategy identified six functional families that are significantly enriched in the set of 'driver' fusion sequences from TICdb: TF, kinase activity, transcription cofactor, G, helicase/histone modifiers (H) and protein binding (P).

We then applied this same pipeline to fusion genes used as positive ('driver') and negative ('passenger') sets during training and validation, to obtain, for each fusion, a functional profile that could be sent to the classifier. Figure 1 shows FFAs for fusions in the two negative datasets (RefSeq genes and fusions found in normal cells) and in the three positive datasets (fusions predicted by ChimerDB and from NGS data, CGC oncogenes and TICdb fusion sequences). As expected, the highest enrichment is obtained for TICdb sequences (as they had been used to define the functional families). For all functional families, positive datasets (presumably enriched in 'driver' fusions) show significant differences relative to RefSeq genes and fusion sequences from normal cells. This suggests that the pipeline for the generation of functional profiles is capturing the basic properties of protein domains retained or lost in 'driver' gene fusions.

3.2 Classifier training

Having a high number of known fusions (437 non-redundant fusion entries in TICdb), a robust classifier could be trained if a suitable negative set is provided. For this, we selected fusion transcripts detected in normal cells from a recent high-throughput study (Frenkel-Morgenstern *et al.*, 2012). To balance the number of positive and negative instances, we extended the negative training set with read-through transcripts detected in normal cells (Nacu *et al.*, 2011).

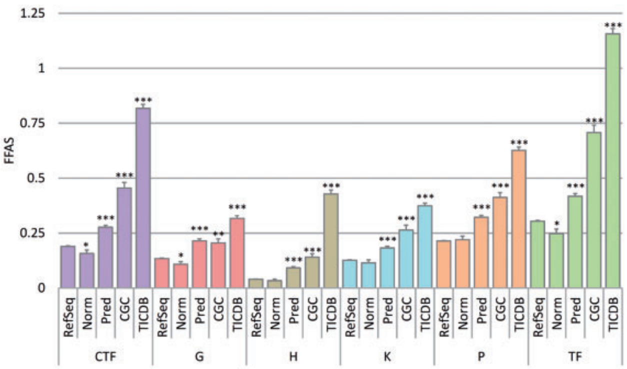


Fig. 1. FFAS of each of the six functional families in different gene sets. $^{*}P<0.05$, $^{**}P<10^{-3}$, $^{***}P<10^{-8}$ compared with RefSeq set, two-tailed *t*-test. Data presented as mean ± SEM. Abbreviations: transcription factor, TF; kinase activity, K; transcription cofactor, CTF; GTPase, G; helicase/histone modifiers, H; protein binding, P. 'RefSeq' are RefSeq genes. 'Norm' are fusion genes and read-through transcripts from normal cells. 'Pred' are gene fusions predicted by ChimerDB 2.0 and from NGS datasets, 'CGC' are known oncogenes from Cancer Genome Census and 'TICDB' are fusion sequences from TICdb

Training the classifier with the complete set of features (k = 24) yielded high precision and recall rates (>85%) after 10-fold cross-validation. Using various subsets of features also yielded reliable results, although overall classification accuracy decreased (Table 1).

To further ensure that data were not overfitted, we performed several classifier training runs with the full set of features by random partitioning data into training and test sets of equal size. This led to only a minor drop in accuracy (<5%). Notably, classifier accuracy was not dependent on the source of tissue of fusions and was equally high for gene fusions from epithelial (EPI), HEM and mesenchymal (MES) tissues included in the training set (data not shown).

The actual information gain afforded by each feature used in classification can be seen in Supplementary Figure S2, with functional profile features providing the largest information gain. As expected, expression and replication features are most informative in 5′ FPGs. Interestingly, features related to lost functional profiles, lost PIIs and self-PIIs provide more information than their retained counterparts.

3.3 Classifier performance

To test our classifier on independent data sources, we selected >700 fusions detected in recent NGS studies and >6000 predicted fusions from chimerDB2.0 (see Supplementary Table S1 for details). Data from chimerDB2.0 include mRNA, EST and SRA sequences and are partitioned into three confidence classes (A—highest confidence, C—lowest confidence). We also tested protein-coding genes from RefSeq as well as a subset of known oncogenes from the Cancer Genome Census. For Refseq and CGC genes, we set the features related to 'lost upon fusion' as unknown variables, as these genes are not taking part in fusions, and their proteins have not lost any domains.

A table with classification results in all these datasets can be found as additional file. Overall classification statistics are presented in Figure 2. As expected, a small number of RefSeq genes (1%) were classified as 'driver' fusion genes. Interestingly, about

Table 1. Precision and recall of Naïve Bayes classifier trained on several sets of features

Feature set	Precision (%)	Recall (%)
All (k = 24)	88	86
Retained (k = 12)	84	81
FFAS-related (k = 12)	78	82
Non-FFAS (k = 12)	85	79
Retained FFAS-related (k = 6)	75	74

Note: Values are given relative to driver fusions (Class 1). Classifier performance tested using 10-fold cross-validation.

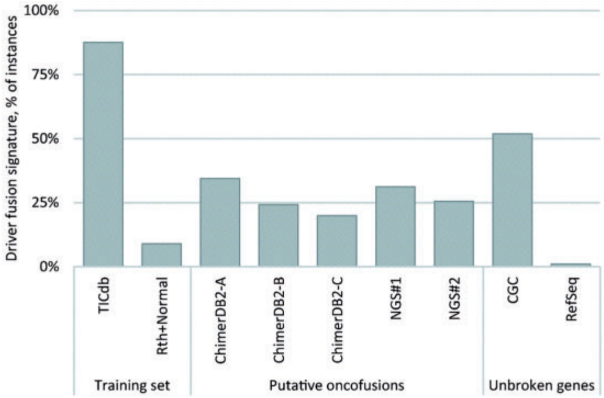


Fig. 2. Percentage of 'driver' fusions identified by the classifier in datasets used in the current study. See Supplementary Table S1 for details about the source of datasets

half of CGC proto-oncogenes showed the pattern characteristic of 'driver' fusions. In addition, Class A fusions from chimerDB2.0 (high confidence) were predicted to contain more 'driver' fusions than less confidence sets B and C. NGS#1 and NGS#2 sets were predicted to have a similar number of driver fusions to chimerDB2.0 A and B classes, respectively.

Although these results suggest a good performance of the classifier, the real number of 'driver' fusions in NGS and chimerDB2.0 datasets is unknown. Therefore, we conducted additional tests to validate these predictions. First, we performed KEGG pathway analysis on predicted 'driver' fusions and compared them to predicted 'passenger' fusions from the same datasets (NGS#1, NGS#2 and chimerDB2.0-A). We performed pathway analysis instead of GO to avoid potential biases owing to the fact that we had used GO:MF categories in the construction of the classifier. KEGG annotations clearly showed that genes involved in fusions predicted to be 'driver' are enriched in 'cancer' and 'cell-cell adhesion' pathways (including 'adherens junction formation' and 'actin cytoskeleton remodeling'), whereas no cancer pathways were enriched in the set of fusions classified as 'passenger' (Supplementary Fig. S3). We also note that genes from fusions classified as 'driver' are enriched in EPI-specific genes ($n = 299$ or 32%, $\mathrm{FDR} = 1.3 \times 10^{-44}$) according to manual InterPro classification (UP_TIS-SUE track in DAVID). This could be due to the fact that samples from NGS datasets are mostly of EPI origin. The same is found (although to a less extent, 23% of genes) for genes involved in 'passenger' fusions. As expected, fusions predicted to be 'driver' are more enriched (8.6-fold, $\mathrm{FDR} = 1.2 \times 10^{-37}$) in genes associated with 'chromosomal rearrangements' (SP_PIR_KEY-WORDS track in DAVID) than predicted 'passenger' fusions (3.1-fold, $\mathrm{FDR} = 5.6 \times 10^{-6}$).

Typically, mutations are said to be 'driver' if they are recurrent, appearing in different samples or in different tumor types. As NGS#1 and NGS#2 datasets include sequencing data from several cancer cell lines (BT474, MCF7, SKBR3, KPL4 and NCIH660), their analysis could identify recurrent fusion genes. We detected 11 and 3 unique fusion gene pairs that were present in two or in three different cell lines, respectively. Although Oncofuse only predicted a 'driver' status for 24% of non-recurrent fusions, it classified as 'driver' 7 of 14 fusions (50%) found in more than one cell line ($P = 0.017$, hypergeometric test). This supports that Oncofuse is correctly identifying those fusions which are likely to be oncogenic. The list of recurrent fusions and predicted drivers is provided in Supplementary Table S4. However, it is worth noting that four other recurrent fusions from these datasets (ARPC4-TTLL3, C15orf38-AP3S2, RPL17L-C18orf32 and RPS10-NUDT3) were also found as read-through transcripts in normal cells; therefore, their true oncogenic potential is unclear. Interestingly, all of them were classified as 'passenger' by Oncofuse.

To further estimate the confidence of the predictions that a fusion is 'driver', we checked whether the *P*-values assigned by the classifier are correlated with the clinical frequency of fusions included in TICdb (i.e. those used as positive dataset during training). Taking the number of reports in Mitelman database as an estimate of recurrence (clinical frequency), we observed that fusions reported two or more times (i.e. recurrent) were assigned significantly lower *P*-values than fusions reported only once (Fig. 3A). Simple correlation analysis of (log frequency) versus (-log *P* value) yields $R = 0.22$, which is a significant dependence with $P = 3 \times 10^{-5}$. We note that this correlation is not due to the fact that more frequent fusions are used repeatedly during training (overfitting) because the classifier was trained on a translocation set in which redundancy had been removed (up to domain composition). When we calculated partial correlations for two of the three variables (*P*-value, clinical frequency and number of times that a fusion appeared in the training set) while controlling for the remaining one, we obtained a partial correlation coefficient $R = 0.16$ ($P = 0.002$) between *P*-value and clinical frequency, but no significant dependence ($R = 0.08$, $P > 0.1$) between *P*-value and number of instances in the training set. Furthermore, we investigated the performance of the classifier with reciprocal fusion sequences, which are occasionally detected at low levels alongside fusion transcripts. Reciprocal fusion transcripts usually lack the potential to code for the protein domains that are required for oncogenic activity, so they are generally regarded as 'passenger' events. In agreement with this, we found that *P*-values given by the classifier to original fusions are significantly lower than *P*-values assigned to their reciprocal counterparts (Fig. 3B).

Finally, we investigated whether the *P*-values given by the classifier might also reflect the tissue-specificity of fusions, as they are partially based on gene expression data. In this analysis, we omitted fusion genes for which no expression data were available. The classifier assigned more significant *P*-values in the correct tissue of origin of gene fusions compared with other tissues (Fig. 3C).

This is interesting because the original features used to construct the classifier were identified on fusions mostly from HEM malignancies; therefore, its performance might decrease in other tissue types. To address this issue, we analyzed the expression levels of 5′ fusion partner genes in three tissue types (EPI, MES and HEM). As shown in Figure 4, expression of 5′ fusion partner genes is significantly higher in the tissue where the fusion was identified, compared with the other two tissues. This suggests that the inclusion of some gene fusions found in MES and EPI cancers will enable the classifier to perform well in new fusion sequences detected in these tissue types.

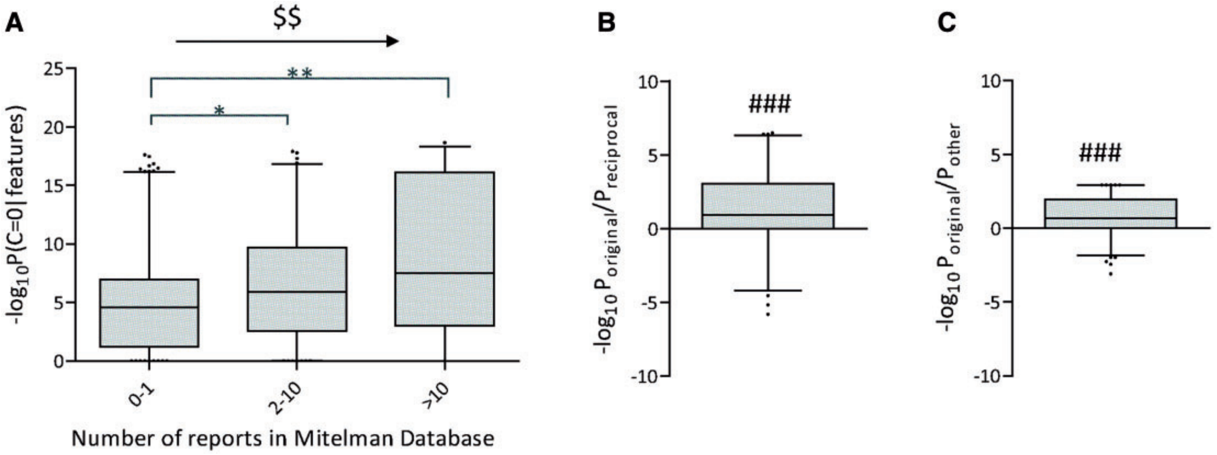
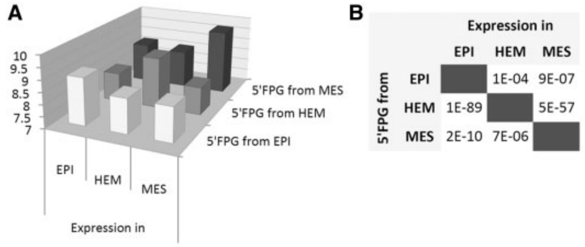


Fig. 3. (**A**) Magnitude of *P*-values is correlated with fusion recurrence. Fusions classified as ‘drivers’ with higher confidence (lower *P*-values) have higher clinical frequency as estimated by the number of reports in Mitelman database. $$, classifier *P*-values and number of reports are dependent; $P = 0.002$, Kruskall–Wallis test; $*P < 0.05$ and $**P < 0.01$, Mann–Whitney test. (**B**) Reciprocal fusions transcripts, which generally have no oncogenic potential, display higher *P*-values. (**C**) Fusions from TICdb are assigned lower *P*-value in the tissue of origin ($P_{\text{originsal}}$) than in other tissues (P_{other}). All possible pairs of tissues compared (HEM, EPI and MES). $^{###}P < 0.0001$, Wilcoxon paired rank test



5'FPG from	Expression in EPI	Expression in HEM	Expression in MES
EPI	■	1E-04	9E-07
HEM	1E-89	■	5E-57
MES	2E-10	7E-06	■

Fig. 4. (**A**) Average expression value of 5′ FPGs from fusions discovered in three tissue types: EPI, HEM and MES. Expression in a given tissue is an average of four samples coming from different tissue subtypes, see Section 2 for details. (**B**) *P*-values for two-tailed paired *t*-test of expression of 5′ FPG from tissue X in tissue X versus its expression in tissues Y and Z

3.4 Validation on independent datasets

Validation of this novel classifier poses a real challenge because the true number of ‘driver’ gene fusions in available datasets is unknown. As comprehensive experimental validation of hundreds of predictions is not feasible, the real predictive potential of our algorithm will only become evident a posteriori, as new oncogenic gene fusions are described in the literature. To test this, we asked whether Oncofuse would have correctly assigned ‘driver’ status to fusion sequences published after the construction of the classifier.

We extracted form the literature gene fusions recently described in cancer, as well as all fusions included in the latest update of TICdb (from April 2012 to March 2013). Table 2 lists these fusions (with PubMed IDs to their references) and their classification by Oncofuse. We found that 79% of the new fusions (23 of 29) were classified as ‘driver’, a recall rate similar to that obtained in cross-validation analysis of the classifier.

As Oncofuse is intended to help researchers to identify the ‘true’ fusion genes in a list of candidates that are obtained

Table 2. A validation set of 29 new fusion genes published since the construction of the classifier, which therefore had not been used to train it

FUSION	Tissue type	PubMed ID	Driver
VTI1A-TCF7L2	EPI	21892161	4.2E-06
KIF5B-RET	EPI	22194472	5.6E-07
INPP5D-ABL1	HEM	21625237	6.2E-02
PAX3-NCOA1	HEM	15313887	No
CD44-SLC1A2	EPI	21471434	No
SCL45A3-FLI1	EPI	22081504	No
MLL-ABI2	HEM	22304832	5.1E-18
MLL-KIAA1524	HEM	23637631	2.8E-17
MLL-PDS5A	HEM	22230297	1.9E-15
FIP1L1-RARA	HEM	21750086	3.1E-04
NAV2-TCF7L1	EPI	22810696	No
POTEF-GLI1	MES	22575261	6.3E-03
PEX5-LCP1	EPI	22782350	2.1E-01
CEP85L-PDGFRB	HEM	21938754	6.4E-08
ETV6-PRDM16	HEM	22050763	9.5E-14
IKZF2-BCL11B	HEM	22867996	7.8E-04
CIC-DUX4	MES	22072439	No
FZD6-SDC2	EPI	22553170	4.5E-01
AHRR-NCOA2	MES	22337624	2.0E-08
CEP85L-ROS1	MES	23637631	7.2E-07
APIP-SLC1A2	EPI	23637631	No
ATG7-RAF1	EPI	23637631	1.7E-04
BCL6-RAF1	EPI	23637631	4.2E-10
EWSR1-CREM	EPI	23637631	2.0E-07
FAM133B-CDK6	HEM	23637631	3.5E-04
CLTC-VMP1	EPI	23637631	8.8E-02
NAB2-STAT6	MES	23313954; 23313952	3.1E-05
CBFA2T3-GLIS2	HEM	23153540	4.3E-03
MAGI3-AKT3	EPI	22722202	7.2E-03

Note: The column at the right shows the Bayesian probability assigned by the classifier for inclusion in Class 0 (‘passenger’); values <0.5 are classified as ‘driver’.

from gene fusion detection algorithms, we also checked the performance of the classifier on novel gene fusion data derived out of NGS platforms. For instance, we ranked all fusions from a set of predictions made in four breast cancer cell lines (BT474, KPL4, MCF7 and SKBR3, data available from http://code.google.com/p/fusion-catcher/). In this dataset, FusionCatcher identified 29 unique significant FPG pairs. Of those, Oncofuse assigned 'driver' status to five pairs with a $P < 0.01$. We then performed manual literature mining to determine which of the 29 fusions have been shown to have oncogenic potential. As shown in Supplementary Table S5, the only proven oncogenic fusions were RPS6KB1-VMP1 and VAPB-IKZF3, which were ranked by Oncofuse #3 and #5 with high significance.

We also explored whether the *P*-values given by Oncofuse could be used to rank fusions in the case when multiple 'drivers' are present in one sample. We analyzed all fusions identified by FusionFinder in RNA-Seq data from chronic myeloid leukemia cell line K562 (Francis *et al.*, 2012). Running Oncofuse on this dataset led to the identification of BCR-ABL1 ($P = 5 \times 10^{-9}$), NUP98-FKBP5 ($P = 5 \times 10^{-6}$), NUP98-JHDM1D ($P = 6 \times 10^{-4}$), NUP98-IGF2BP2 ($P = 7 \times 10^{-4}$) and DDB2-ALS2 ($P = 1 \times 10^{-3}$) as putative 'drivers' with a threshold of $P < 10^{-2}$. It is interesting that BCR-ABL1, which plays a crucial role in CML (Ren, 2005) and is known to be present in this cell line, was ranked #1 in this list. Although the other three fusions have not been tested experimentally, fusions involving NUP98 are known drivers of HEM malignancies (Gough *et al.*, 2011). These examples illustrate how Oncofuse could help to prioritize for experimental validation lists of candidate fusion genes detected in RNA-seq studies.

4 DISCUSSION

It is estimated that ~20% of all cancers are caused by gene fusions driven by chromosomal translocations (Mitelman *et al.*, 2007; Nambiar *et al.*, 2008), but it is reasonable to predict that this figure will increase during the next years, as more samples from solid tumors, notably EPI, will be analyzed with NGS technologies. Sequencing of cancer transcriptomes will yield large lists of chimeric fusion genes, and it will be a challenge to identify which of those are actually contributing to the initiation or progression of cancer.

The oncogenic role of gene fusions, like other types of mutations, is usually judged by their recurrence: if they are found in different samples or in different tumor types they are considered to be important for cancer development. However, this view is probably over-simplistic, considering recent findings showing that many mutations important for cancer development are 'private' (i.e. non-recurrent) because the same biological pathways can be inactivated by mutations in different genes (Bozic *et al.*, 2010; Vogelstein *et al.*, 2013). Furthermore, it is unclear that all recurrent fusion sequences must necessarily be oncogenic. For instance, we found four recurrent fusions in NGS data from cancer cells that were also present as natural read-through transcripts in normal cells. It is reasonable to predict that the increasing use of RNA-seq to interrogate cancer genomes will yield 'recurrent' fusions that will turn out to be passenger events, rendering recurrence a weaker argument for oncogenic potential. Therefore, it is necessary to develop tools to reliably predict the oncogenic potential of gene fusions that play a significant role in cancer development, regardless of whether they are recurrent.

In this article, we present Oncofuse, a novel algorithm that provides an accurate classification of 'driver' and 'passenger' fusion genes with probability estimates, which could be of great help to rank fusions detected by NGS in tumor samples and thus prioritize experimental validation studies. A summary of the analysis pipeline used in this study is presented in Supplementary Figure S4. Building on our previous identification of several genomic hallmarks of oncogenic ('driver') fusion sequences, we generated a carefully selected set of genomic features that were used to train a Bayesian classifier showing good precision and recall rates in cross-validation studies. Thus, when a researcher is faced with a list of fusion genes detected by NGS studies in cancer genomes, Oncofuse will extract the same set of genomic features and will return a *P*-value for the probability that each fusion in the list is 'driver', as well as some additional information about its functional profile.

We observed that functional profiles of fusion proteins provide a substantial amount of information on whether a fusion is likely to be 'driver'. Molecular functions related to TF, transcription co-factor, kinase, histone modification, protein isomerization and GTPase-related are enriched in 'driver' fusion proteins. This is consistent with our previous observations in HEM malignancies (Shugay *et al.*, 2012) and seems to be a common hallmark of all 'driver' gene fusions, regardless their tissue of origin (NGS#1 and NGS#2 datasets used in this study are exclusively of EPI origin). This is a important point because it justifies the use of the same classifier for fusions from different sources by adjusting only gene expression features. The fact that tissue-specific expression patterns of 5' fusion partner genes are robustly replicated for the three tissues examined in this study helps to explain that Oncofuse assigns more significant *P*-values in the tissue where a fusion was identified compared with other tissues. Thus, even though the original features used to construct the classifier were based on insights taken mostly from HEM malignancies, Oncofuse will also classify gene fusions from MES and EPI cancers if expression data of both partner genes are provided.

Although there are a few algorithms for the identification of chimeric fusion sequences in NGS data, the only attempt (to the best of our knowledge) to distinguish 'driver' from 'passenger' rearrangements was performed by Chinnaiyan and colleagues in 2009 (Wang *et al.*, 2009). However, our strategy is different in several important aspects. First, Wang *et al.* made their predictions based on the properties of entire partner genes, such as GO annotations and their position in the interactome, whereas we use features retained or lost in fusion genes (or in the proteins encoded by them). These features, like the loss of specific protein domains, the co-occurrence of domain combinations or promoter and 3′-UTR swapping, could yield information that is not correctly described by simply summarizing the annotations of the parent genes (Frenkel-Morgenstern and Valencia, 2012; Hegyi *et al.*, 2009). Moreover, we show that the analysis of protein domains yields meaningful functional signatures that allow the distinction between driver and passenger gene fusions.

Another important difference is that our pipeline takes into account expression data, as changes in the expression of partner genes are common in gene fusions (Rabbitts, 1994). Notably, among the top 60 genes identified by Wang *et al.* as those more likely to undergo a fusion, there is a 13-fold enrichment for CGC proto-oncogenes not reported to be involved in translocations, whereas proto-oncogenes present in Mitelman database (i.e. involved in translocations) are enriched only 10-fold. This suggests that their predictions (some of which have been validated experimentally) are biased towards known oncogenes, whether involved in fusions or not. Although our approach might also be affected by this bias in favor of known oncogenes (50% of CGC oncogenes are predicted to be 'driver', Fig. 2), we have tried to minimize this problem by including features that are retained or lost as a result of the fusion; therefore, the predictive accuracy of our pipeline should be even better.

Even though these results support a robust behavior of the classifier, we set out to validate it in various ways using results from several independent datasets containing several thousand fusions reported in recent NGS and EST/mRNA studies. We found that they were moderately but significantly enriched in predicted 'driver' fusions, whereas only a small number of false positives were found among chimeric transcripts identified in normal cells. Likewise, those datasets with more confident gene fusion predictions showed a higher percentage of fusions classified as 'driver' by Oncofuse. Furthermore, pathway analysis showed that fusions predicted to be 'driver' in these datasets involve genes that participate in key oncogenic pathways. For example, we found that many predicted 'driver' fusions could lead to oncogenesis by disrupting normal cell adhesion, which is one of the hallmarks of cancer (Hanahan and Weinberg, 2000, 2011). This is an interesting finding, especially given the fact that the datasets analyzed were obtained mostly from EPI tissues in which this pathway plays a significant role in carcinogenesis (Martin-Belmonte and Perez-Moreno, 2012).

Finally, we were able to test the performance of Oncofuse in new datasets that have been made available since its construction. First, Oncofuse would have correctly classified 79% of the gene fusions published in cancer samples in this period (until April 2013). Second, we tested the classifier on the lists of candidates detected by gene fusion detection algorithms in RNA-seq studies. Although the real number of gene fusions with oncogenic properties in such datasets is unknown, because only a few candidates are tested experimentally, fusions with support from published experimental studies were given higher probabilities of being 'driver'.

In summary, we present a robust Bayesian classifier supported by extensive validation studies. We believe that Oncofuse could become a useful computational tool to aid in the prediction of the oncogenic potential of novel fusion genes detected by NGS in the transcriptomes of cancer samples. In this way, it should help to prioritize which candidates should be tested experimentally and thus speed up the identification of novel fusion genes with a relevant role in the diagnosis, classification and treatment of cancer patients.

Funding: This work has been funded with the help of the Spanish Ministry of Science and Innovation (SAF 2007-62473), the PIUNA Program of the University of Navarra, the Caja Navarra Foundation through the Program 'You choose, you decide' (Project 10.830) and ISCIII-RTICC (RD06/0020/0078). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Conflict of interest: none declared.

REFERENCES

Akiva,P. *et al.* (2006) Transcription-mediated gene fusion in the human genome. *Genome Res.*, **16**, 30–36.

Asmann,Y.W. *et al.* (2012) Detection of redundant fusion transcripts as biomarkers or disease-specific therapeutic targets in breast cancer. *Cancer Res.*, **72**, 1921–1928.

Benelli,M. *et al.* (2012) Discovering chimeric transcripts in paired-end RNA-seq data by using Ericscript. *Bioinformatics*, **28**, 3232–3239.

Bozic,I. *et al.* (2010) Accumulation of driver and passenger mutations during tumor progression. *Proc. Natl Acad. Sci. USA*, **107**, 18545–18550.

Edgren,H. *et al.* (2011) Identification of fusion genes in breast cancer by paired-end RNA-sequencing. *Genome Biol.*, **12**, R6.

Francis,R.W. *et al.* (2012) FusionFinder: a software tool to identify expressed gene fusion candidates from RNA-Seq data. *PLoS One*, **7**, e39987.

Frank,E. *et al.* (2004) Data mining in bioinformatics using Weka. *Bioinformatics*, **20**, 2479–2481.

Frenkel-Morgenstern,M. and Valencia,A. (2012) Novel domain combinations in proteins encoded by chimeric transcripts. *Bioinformatics*, **28**, i67–i74.

Frenkel-Morgenstern,M. *et al.* (2012) Chimeras taking shape: potential functions of proteins encoded by chimeric RNA transcripts. *Genome Res.*, **22**, 1231–1242.

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

Gough,S.M. *et al.* (2011) NUP98 gene fusions and hematopoietic malignancies: common themes and new biologic insights. *Blood*, **118**, 6247–6257.

Hanahan,D. and Weinberg,R.A. (2000) The hallmarks of cancer. *Cell*, **100**, 57–70.

Hanahan,D. and Weinberg,R.A. (2011) Hallmarks of cancer: the next generation. *Cell*, **144**, 646–674.

Hegyi,H. *et al.* (2009) Intrinsic structural disorder confers cellular viability on oncogenic fusion proteins. *PLoS Comput. Biol.*, **5**, e1000552.

Huang,D.W. *et al.* (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.

Hunter,S. *et al.* (2012) InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.*, **40**, D306–D312.

Kim,D. and Salzberg,S.L. (2011) TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol.*, **12**, R72.

Kim,P. *et al.* (2010) ChimerDB 2.0—a knowledgebase for fusion genes updated. *Nucleic Acids Res.*, **38**, D81–D85.

Martin-Belmonte,F. and Perez-Moreno,M. (2012) Epithelial cell polarity, stem cells and cancer. *Nat. Rev. Cancer*, **12**, 23–38.

Mitelman,F. *et al.* (2007) The impact of translocations and gene fusions on cancer causation. *Nat. Rev. Cancer*, **7**, 233–245.

Nacu,S. *et al.* (2011) Deep RNA sequencing analysis of readthrough gene fusions in human prostate adenocarcinoma and reference samples. *BMC Med. Genomics*, **4**, 11.

Nambiar,M. *et al.* (2008) Chromosomal translocations in cancer. *Biochim. Biophys. Acta*, **1786**, 139–152.

Novo,F.J. *et al.* (2007) TICdb: a collection of gene-mapped translocation breakpoints in cancer. *BMC Genomics*, **8**, 33.

Ortiz de Mendíbil,I. *et al.* (2009) Signatures of selection in fusion transcripts resulting from chromosomal translocations in human cancer. *PloS One*, **4**, e4805.

Parra,G. *et al.* (2006) Tandem chimerism as a means to increase protein complexity in the human genome. *Genome Res.*, **16**, 37–44.

Rabbitts,T.H. (1994) Chromosomal translocations in human cancer. *Nature*, **372**, 143–149.

Ren,R. (2005) Mechanisms of BCR-ABL in the pathogenesis of chronic myelogenous leukaemia. *Nat. Rev. Cancer*, **5**, 172–183.

Rosen,P. *et al.* (2012) Clinical potential of the ERG oncoprotein in prostate cancer. *Nat. Rev. Urol.*, **9**, 131–137.

Sakarya,O. *et al.* (2012) RNA-Seq mapping and detection of gene fusions with a suffix array algorithm. *PLoS Comput. Biol.*, **8**, e1002464.

Shugay,M. *et al.* (2012) Genomic hallmarks of genes involved in chromosomal translocations in hematological cancer. *PLoS Comput. Biol.*, **8**, e1002797.

Tan,H. *et al.* (2012) A novel missense-mutation-related feature extraction scheme for “driver” mutation identification. *Bioinformatics*, **28**, 2948–2955.

Villanueva,M.T. (2012) Genetics: gene fusion power. *Nat. Rev. Clin. Oncol.*, **9**, 188.

Vogelstein,B. *et al.* (2013) Cancer genome landscapes. *Science*, **339**, 1546–1558.

Wang,X.S. *et al.* (2009) An integrative approach to reveal driver gene fusions from paired-end sequencing data in cancer. *Nat. Biotechnol.*, **27**, 1005–1011.