[Article] Pegasus: a comprehensive annotation and prediction tool for detection of driver gene fusions in cancer

BMC Systems Biology

SOFTWARE Open Access

Pegasus: a comprehensive annotation and prediction tool for detection of driver gene fusions in cancer

Francesco Abate[1,2,3†], Sakellarios Zairis[2†], Elisa Ficarra[3], Andrea Acquaviva[3], Chris H Wiggins[2,6,7], Veronique Frattini[5], Anna Lasorella[5], Antonio Iavarone[5], Giorgio Inghirami[4] and Raul Rabadan[1,2*]

Abstract

Background: The extraordinary success of imatinib in the treatment of BCR-ABL1 associated cancers underscores the need to identify novel functional gene fusions in cancer. RNA sequencing offers a genome-wide view of expressed transcripts, uncovering biologically functional gene fusions. Although several bioinformatics tools are already available for the detection of putative fusion transcripts, candidate event lists are plagued with non-functional read-through events, reverse transcriptase template switching events, incorrect mapping, and other systematic errors. Such lists lack any indication of oncogenic relevance, and they are too large for exhaustive experimental validation.

Results: We have designed and implemented a pipeline, Pegasus, for the annotation and prediction of *biologically functional* gene fusion candidates. Pegasus provides a common interface for various gene fusion detection tools, reconstruction of novel fusion proteins, reading-frame-aware annotation of preserved/lost functional domains, and data-driven classification of oncogenic potential. Pegasus dramatically streamlines the search for oncogenic gene fusions, bridging the gap between raw RNA-Seq data and a final, tractable list of candidates for experimental validation.

Conclusion: We show the effectiveness of Pegasus in predicting new driver fusions in 176 RNA-Seq samples of glioblastoma multiforme (GBM) and 23 cases of anaplastic large cell lymphoma (ALCL). Contact: fa2306@columbia.edu.

Keywords: Gene fusion, Next-generation sequencing, Machine learning

Background

Gene fusions are the result of genetic aberrations (translocations, deletions, amplifications and inversions) involving the juxtaposition of two genes that can generate a single hybrid transcript. Since 1960, gene fusions have been known to play a major role in tumorgenesis. The BCR-ABL1 gene fusion, arising from the Philadelphia chromosome (t(9;22)(q34;q11)), was the first case of a translocation-induced gene fusion associated with the development of a cancer, namely chronic myelogenous leukemia [1]. In this fusion, the N-terminus oligomerization domain of BCR and the tyrosine kinase domain in ABL1 are essential in promoting oncogenic activity [2]. Among the gene fusions associated with tumor development, it is worth mentioning TMPRSS2-ERG, a gene fusion occurring in 40-80% of cases of prostate cancer [3, 4] and fusions involving the ALK gene with different partners in various malignancies [5], such as NPM1-ALK in anaplastic large cell lymphoma (ALCL) [6] and ELM4-ALK in non-small-cell lung cancer [7].

Discovering the relationship between gene fusions and cancer is gaining significant momentum thanks to advances in next generation sequencing (NGS) technology, particularly RNA paired-end sequencing [8]. Recently, the application of this technology allowed the discovery of new chromosomal rearrangements of the CIITA gene with various promiscuous partners in the lymphomagenesis of primary mediastinal B cell lymphomas [9]. In

* Correspondence: rabadan@dbmi.columbia.edu
†Equal contributors
[1]Department of Biomedical Informatics, Columbia University, 1130 St. Nicholas Ave, New York, NY 10032, USA
[2]Center for Computational Biology and Bioinformatics, Columbia University, 1130 St. Nicholas Ave, New York, NY 10032, USA
Full list of author information is available at the end of the article

Singh *et al.* [10], the analysis of RNA-Seq data led to the discovery of the highly oncogenic fusion protein FGFR3-TACC3 in 3% of patients diagnosed with glioblastoma multiforme (GBM). Even though FGFR3-TACC3 occurs at low frequency, the efficacy of FGFR inhibitors in the treatment of these tumors opens the door to personalized therapies for this deadly disease. Moreover, the FGFR3-TACC3 fusion has been found in other cancers such as bladder [11] and lung [12]. These recent discoveries underscore the power of high throughput genomics for the identification of targetable gene fusions, opening the door to personalized cancer therapies.

Several bioinformatics tools are now established for the detection of candidate fusion events from paired-end RNA-Seq data. Generally, the detection of read pairs that discordantly map to two distinct genes generates a first set of gene fusion candidates. Subsequently, the exact fusion junction is determined for each candidate by searching for reads spanning the breakpoint, i.e. reads that partially map to both genes. FusionSeq [13] and deFuse [14] were the earliest examples of software based on this strategy. Detection tools differ in the type and number of cascading filters they apply to reduce the large number of false positive fusions. ChimeraScan [15] implements an algorithm based on trimming reads to increase fusion detection sensitivity. Bellerophontes [16] uses TopHat [17] and Cufflinks [18] to identify gene fusions involving truly expressed genes, and applies a set of modular cascading filters based on an accurate gene fusion model [19]. A comprehensive comparison of fusion detection tools has recently been published [20].

The methods adopted by fusion detection tools to shrink the list of candidates lead to increased specificity but reduced sensitivity. As reported in the comparative analysis performed by Abate *et al.* [16] , the heterogeneity of filtering strategies often yields poorly overlapping sets of candidate transcripts between algorithms. The union of all candidate fusions reported by different detection tools should be considered for further experimental validation, in order to maximize sensitivity. A problem arises however, since the number of putative gene fusions might be on the order of hundreds of candidates per RNA-Seq sample. This is largely due to the presence of read-through events, reverse transcriptase template switching artifacts, and different systematic errors in the analysis of the reads [21]. The naïve approach of considering all candidates from all detection tools quickly overwhelms the capacity of experimental validation procedures, and highlights the need to focus on a reduced number of select *biologically relevant* fusions *driving* the oncogenic progression of disease.

The classification of gene fusions into driver and passenger events is a complex problem that has not been fully explored yet. To address this issue, several databases have collected hundreds of chromosomal translocations involved in cancer cases and reported in the biomedical literature. For instance, Mitelman [22], TICdb [23] and ChimerDB2.0 [24] are manually curated repositories of known gene fusions along with detailed information such as chromosomal breakpoints, reported tissue types, and fusion sequences. New computational approaches to nominate biologically relevant fusions from high-throughput data have been proposed. ConSig assesses driver gene fusions by combining copy number variations (CNV), ontologies and interactomes based on the assumption that fusion events are more likely to arise from genes with similar biological functions [25]. Wu *et al.* have proposed a network based approach relying on relative co-occurrence of protein domains and domain-domain interactions, and location of the gene fusion in a gene network [26]. Recently, Oncofuse has improved the computational analysis with a machine learning approach based on a Naïve Bayes classifier applied to preserved domains after chromosomal rearrangement [27]. Compared to earlier methods, Oncofuse introduces a new level of detail by considering only the domains that are maintained on the resulting fusion transcripts. The domain analysis should be extended, however, by taking into account all possible transcript isoforms as well as the reading frame, which plays a crucial role since frame-shifted fusions imply a loss of the 3'-gene domains. Moreover, Oncofuse relies on a Naïve Bayes classifier that makes a restrictive assumption on the class conditional independence of all features. Taking the FGFR3-TACC3 gene fusion as an example, however, the acquired coiled-coil domain of the TACC3 gene cooperates with tyrosine kinase functionality of FGFR3 to produce the dramatic oncogenic effect [10]. This example illustrates the limitations of a model assumption that ignores interactions between functional protein domains.

In this paper we aim to discern oncogenic driver fusions from the background of passenger events and artifacts by combining 1) functional domain annotation based on accurate fusion sequence analysis and 2) a binary classification algorithm using gradient tree boosting. The implementation of this methodology is Pegasus, a new framework for the functional characterization of RNA-Seq gene fusion candidates and quantification of their oncogenic potential. Pegasus runs on top of multiple state of the art fusion detection tools in order to maximize detection sensitivity and consider the largest possible set of fusion candidates.

The main innovative steps introduced by Pegasus are as follows:

- **Common interface** between several fusion detection tools.
- **Chimeric transcript sequence reconstruction:** a key feature since fusion detection tools do not report whole transcript sequences.

- **Reading frame identification and accurate domain annotation,** including both preserved and lost protein domains within the assembled chimeric transcript.
- **Prediction of fusion oncogenic potential:** high performance ensemble learning technique trained on a feature space of protein domain annotations.
- **Automated workflow** that would otherwise require massive effort if manually executed by the scientist.

We assess the trained Pegasus model's prediction accuracy by applying it to a set of recently discovered gene fusions where it compares quite favorably with the current state of the art, Oncofuse. Beyond curated datasets, we report the results of Pegasus on real RNA-Seq data from three distinct patient cohorts: public GBM samples from TCGA, non-public GBM samples, and non-public ALCL samples. We successfully identify driver gene fusions in both cancer types and demonstrate the utility of coupling our algorithm with experimental analysis.

Implementation

In order to first motivate our feature engineering, we briefly review the main mechanisms hitherto identified in oncogenic gene fusions (see Figure 1). Fusion transcripts can broadly lead to three scenarios: i) enhanced overexpression of an oncogene ii) deregulation of a tumor suppressor gene iii) formation of a new, aberrant protein.

Enhanced overexpression of an oncogene is exemplified by the famous IgH-MYC fusion (Figure 1a), and is the main reason for our explicit annotation of oncogene status and interactions with known oncogenes in our feature space representation of fusion transcripts. In other cases, deregulating properties can be associated with the fused transcript, such as insertion of one or two nucleotides across the junction breakpoint introducing a shift of the reading frame. This scenario is illustrated in the PPP2RA-CHEK2 fusion [28] (Figure 1b) where the introduced frame-shifted sequence prevents the formation of the CHEK2 protein that is a known tumor suppressor gene. Here we see the motivation for our explicit annotation of tumor suppressor status and interactions with known tumor suppressors in the feature space, as well as the need for computing reading frame of each candidate fusion. Finally, fusion transcripts can also yield a completely new chimeric protein. BCR-ABL1 [1] and NPM1-ALK [6] are well studied examples of such in-frame fusions. The new protein is generally larger than the kinase involved and causes an increase of the

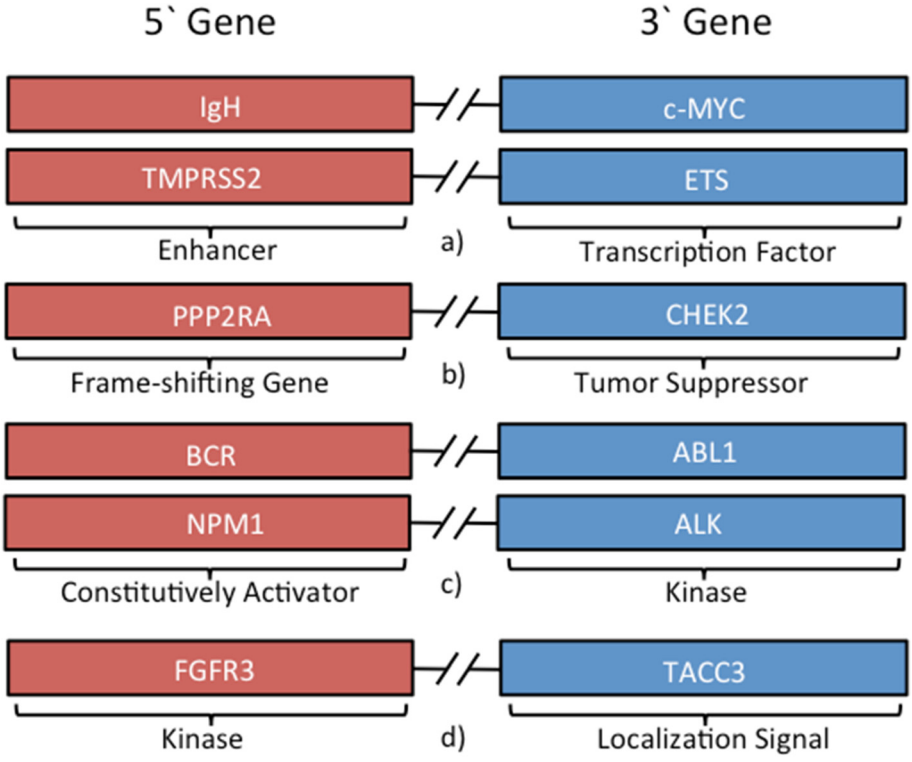

Figure 1 Common gene fusions in cancer biology. Different mechanisms of gene fusions in cancer are shown. Figure 1**(a)** Shows the mechanisms of oncogenic transcription factor activation by means of an endogenous enhancer. Figure 1**(b)** Shows an example of disrupting gene fusion where the 5′ gene leads to the deregulation of the 3′ tumor suppressor gene. Finally, Figure 1**(c)** and 1**(d)** depicts the BCR-ABL1, NPM1-ALK and FGFR3-TACC3 chimeras where a completely new protein is produced.

tyrosine kinase activity (Figure 1c). Moreover, in the recently discovered FGFR3-TACC3 gene fusion, the acquired coiled-coil domain of TACC3 gene drives the localization of the fusion protein to the mitotic spindle through a mechanism that is dependent on tyrosine kinase functionality [10] (Figure 1d). It seems that a reasonable feature space representation for predicting the oncogenic properties of such novel chimeric proteins should maintain knowledge of both preserved and lost functional domains in the partner genes.

The methodology of Pegasus is composed of three phases (see Figure 2): a) integration of candidates from fusion detection tools b) chimeric transcript sequence reconstruction and domain annotation c) classifier training and driver prediction. The first phase, *Fusion Detection Tools Integration,* involves pooling the entire set of unique gene fusion candidates detected by any of the fusion detection tools. The second phase, *Chimeric Transcript Sequence Reconstruction and Domain Annotation,* includes two steps: i) reconstruction of the chimeric transcript using the genomic breakpoint coordinates and the partner gene annotations ii) annotation of the assembled sequence to provide information on the fusion frame and to generate a report of all the protein domains conserved or lost in the gene fusion. The final phase frames *Driver Fusion Prediction* as a binary classification task and fits an ensemble of decision tress via the gradient boosting algorithm.

Fusion detection tools integration

The *Fusion Detection Tools Integration* is the repository of the entire set of fusion candidates detected by any of the fusion detection tools. Several fusion detection algorithms are supported in Pegasus: Bellerophontes, deFuse, and ChimeraScan. Each tool adopts a private formalism for reporting fusion information with different levels of detail. However, some chimeric fusion features are

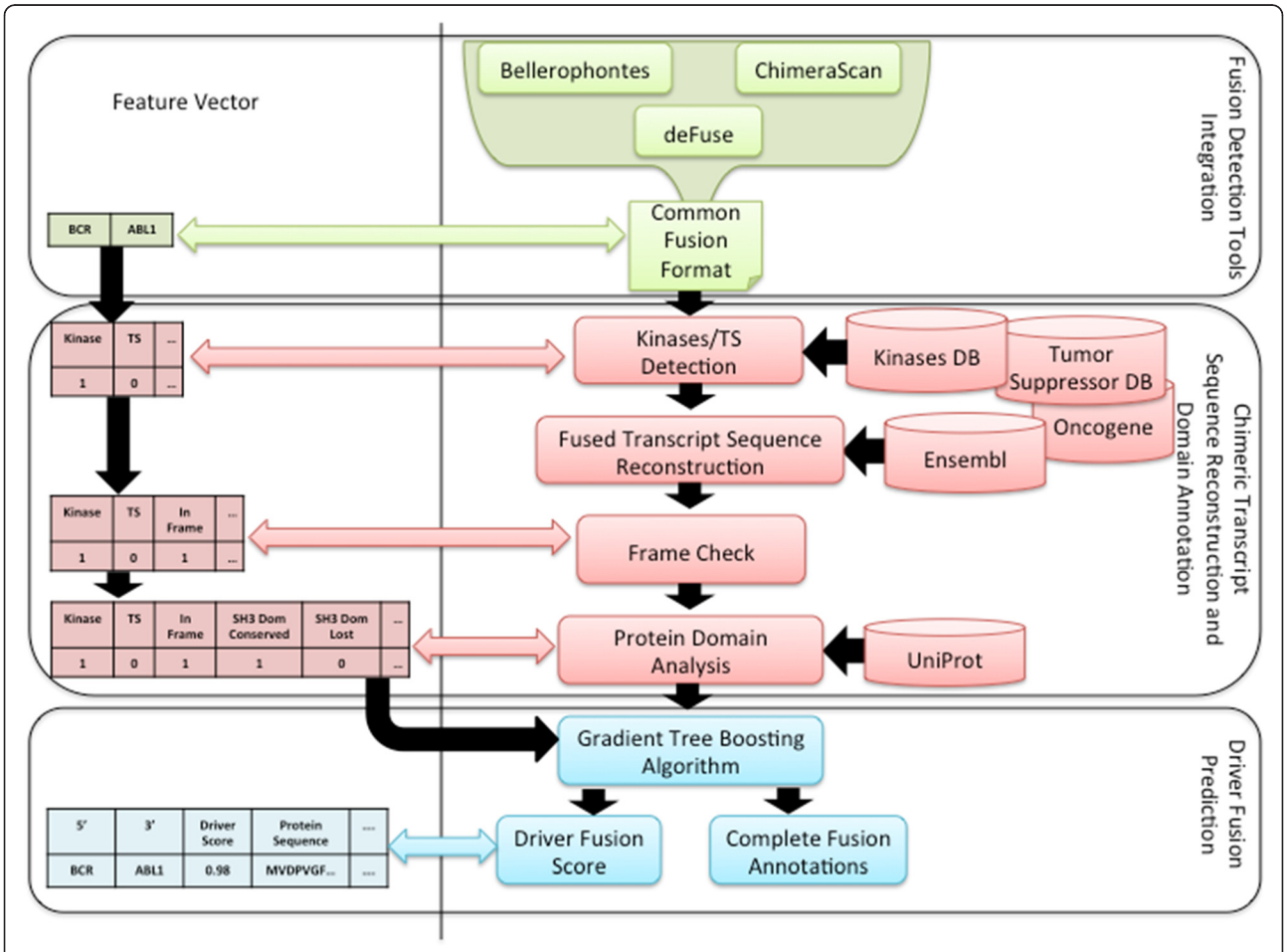

Figure 2 Pegasus fusion annotation flow. For each phase, the figure shows how the feature vector is constructed on the left side of the panel. In Fusion Detection Tools Candidates Integration step, report files from several fusion detection tools are loaded in a unique fusion database. In Chimeric Transcript Sequence Reconstruction and Functional Analysis phase, the fusion transcript is assembled according to the fusion breakpoint coordinates, the reading frame is checked and the protein domain annotation is performed on the resulting fused sequence. Finally, the Driver Fusion Prediction applies machine learning techniques to determine prediction scores.

common to all the fusion detection tool reports (e.g. genes involved in the fusion, genomic breakpoint coordinates, number of reads encompassing and spanning the fusion breakpoint, etc.). Thus, the internal database structure of Pegasus provides a unique point of access for all the information needed to fully describe a gene fusion candidate. Furthermore, experimental analysis might involve the comparison of several RNA-Seq samples per case study. To this end, the common repository embedded in Pegasus provides an organized overview of all the fusions occurring in the entire sample set. This feature allows comparison and the recurrence analysis of the fusion candidates within both the same experimental dataset (samples of the same disease) and within different experimental datasets (samples across different diseases).

Chimeric transcript sequence reconstruction and domain annotation

For each gene fusion candidate, the entire chimeric transcript sequence is first assembled according to publicly available gene annotations and the fused gene breakpoint coordinates. This is the most computationally intensive step in the methodology. For each gene fusion candidate, Pegasus assembles the chimeric sequence based on the possible isoforms and splicing junctions of each gene, as well as the genomic breakpoint coordinates (Figure 3). It is worth specifying that Pegasus reconstructs the fusion sequence exclusively on the basis of gene annotation and fusion breakpoint, and it does not exploit the sequenced reads because they are not an input to the program. Therefore the reconstructed sequence may not reflect the actual sequences especially in case of alternative splicing events.

We adopt the annotation file from ENSEMBL database [29]. Since several distinct isoforms might be available for a specific gene, Pegasus considers the combinations of all possible isoforms reported in the annotations of those genes involved in the fusion (Figure 3). The chimeric transcript sequence is therefore reconstructed combining the 5' gene isoform sequence (from the isoform start codon to the genomic breakpoint) and the 3' gene isoform sequence (from the genomic breakpoint to the isoform stop codon). Different gene isoforms allow for different protein domains to be retained or disrupted during the fusion. If this scenario occurs, Pegasus considers the union of all possible domains that are retained

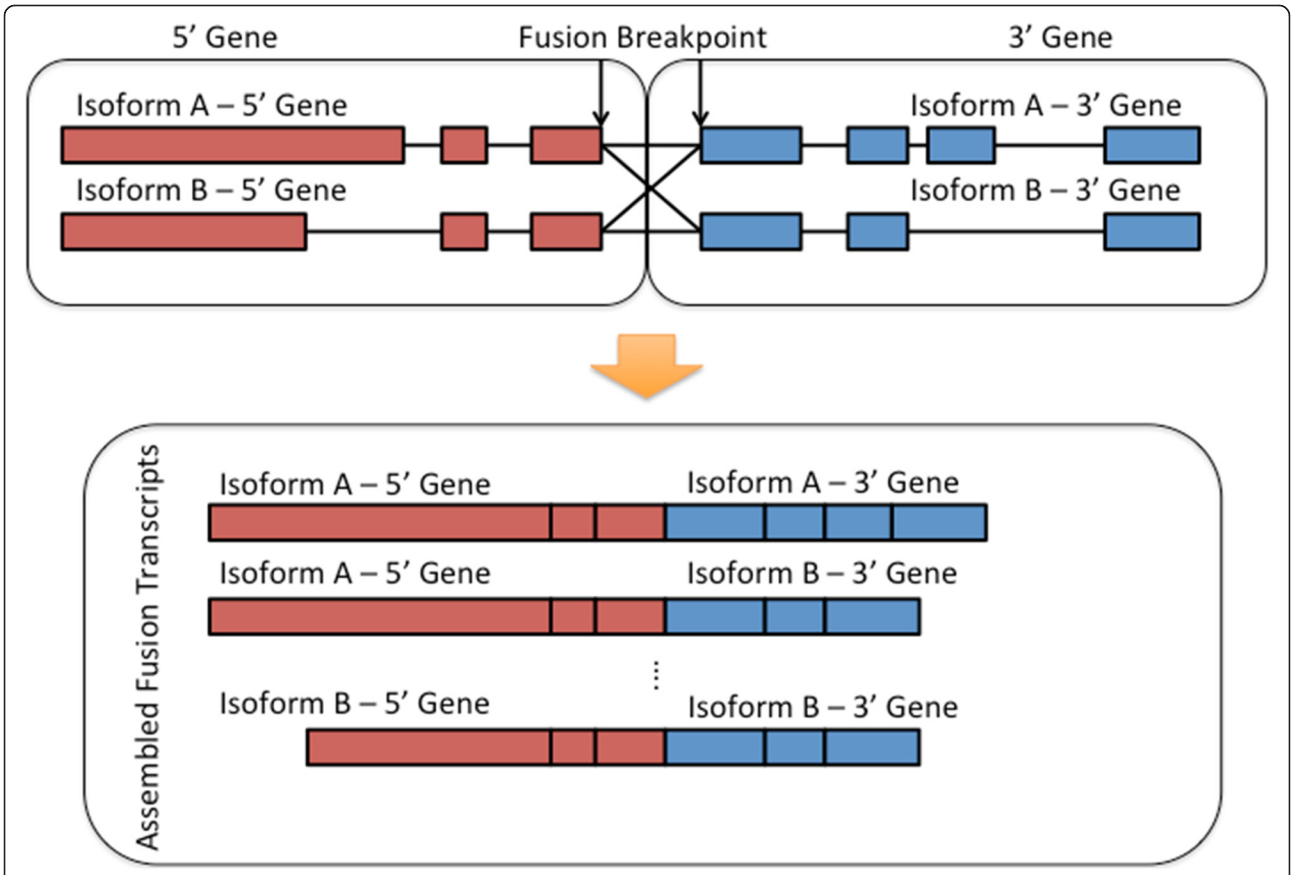

Figure 3 Chimeric transcript sequence reconstruction. Both for the 5' and the 3' gene the annotated isoform sequences are retrieved. All the possible combinations between the isoforms of the 5' and 3' genes are considered as putative fusion transcripts.

and lost and as input features for downstream classification. Furthermore, the fusion breakpoint can fall in either the coding region (exon-exon junction boundaries), or in non-coding regions (exon-intron or intron-intron junction boundaries). Pegasus takes the latter scenario into account and if the fusion breakpoint falls in an intron, the intronic sequence is retained.

After sequence reconstruction we assess the preservation of the reading frame in the chimeric transcript, which enables a great deal of our downstream feature engineering (see Figure 4a). If the gene fusion introduces or deletes a nucleotide in one of the codons, the entire reading frame is shifted and the corresponding amino acid sequence changes. Consequently, the resulting protein sequence is different from the one encoded by the gene involved in the fusion. The gene fusion encodes a protein sequence that either corresponds to a completely unknown protein or contains a premature stop codon (the presence of a premature stop codon in the chimeric sequence interrupts the protein translation resulting in the truncation of the protein encoded by the 5' fused gene). This class of mutations is functionally similar to nonsense point mutations that play a role in many cancers and might imply the loss of functionality of the 5' fused gene. The sequence is labeled as in-frame if the number of nucleotides composing the fusion sequenced is a multiple of three and no premature stop codons are introduced in the chimeric sequence.

The annotation of the preserved and lost protein domains is essential in order to capture the oncogenic potential of a translated chimeric transcript. The nucleotide fusion sequence assembled in the previous step is translated into an amino acid sequence. Subsequently, the UniProt web service [30] is queried for all available annotations of the putative protein encoded by the two genes involved in the fusion (Figure 4b). Leveraging the reading frame information and fusion breakpoint, Pegasus determines the conserved and lost domains associated with both 5' and 3' genes. It is worth emphasizing that both conserved and the lost domains are valuable features of a fusion transcript, with the former more likely to discriminate oncogene related fusions and the latter more likely to discriminate tumor suppressor related fusions.

The domain annotation permits the creation of a detailed feature space for the fusion transcripts, a prerequisite step for posing the ensuing machine learning task. In Pegasus the feature space is composed of:

- Binary information about reading frame and breakpoint region (if the breakpoint falls in coding regions, introns and UTRs);
- Presence or absence of ~1000 protein domains from UniProt. Our selection was based on the domains occurring in the training set from ChimerDB2.0.
- Number of oncogenic or tumor suppressor domains, as defined by association with the keywords "tumor suppressor" or "oncogene" in the UniProt database.
- Number of protein-protein oncogenic interacting domains. We check if one or more domains of the fusion interact with both oncogenic and tumor suppressor domains.

Driver fusion prediction as a binary classification task

We aim to fit a model that can identify oncogenic fusions from the background of passenger events and artifacts. More precisely, we aim to learn a mapping $f: X \rightarrow y$ from the fusion transcript feature space X to a label $y \in \{0, 1\}$ representing oncogenic driver status. Since we desire a biologically interpretable model that is also capable of capturing interactions between features, the decision tree is a natural choice. On the other hand, high dimensional feature spaces predispose to overfitting, and previous driver fusion prediction studies [27] focused a great deal on upfront dimensionality reduction for this reason. A single, large decision tree classifier is not likely to generalize beyond some training depth. An ensemble of shallower decision trees, if learned in a boosting framework, can guard against overfitting because of the iterative nature of the learning and the additive structure of the model [31]. Therefore the balance we strike between the expressive power of decision trees and robustness to overfitting comes in the form of stagewise additive modeling. In Pegasus we employ an additive model $f(x) = \sum_m \alpha_m h_m(x)$ composed of weighted decision trees, an instance of which is shown in Figure 5, and fit via gradient boosting [32]. There is no manual reduction of features or feature space dimension in this strategy, unlike the manual selection of 6 enriched functional categories in the Oncofuse framework [27].

Thus we require neither upfront dimensionality reduction schemes nor the restrictive assumption of class conditional feature independence in the Naïve Bayes model. Gradient tree boosting is an ensemble learning technique, wherein decision trees are used as base learners and the final model is expressed as an expansion in these basis functions. Figure 5 depicts a sample regression tree that would be added to the ensemble in a single round of boosting. Although the base learners are performing regression, appropriate choice of loss function for gradient boosting yields a classification task. Here we use the binomial deviance loss, and enforce a maximum depth of 5 nodes in the individual decision trees. The gradient tree boosting algorithm, originally published in 2000, is outlined below [33] and the implementation we use can be found in the scikit-learn python library [34]. We denote the number of training

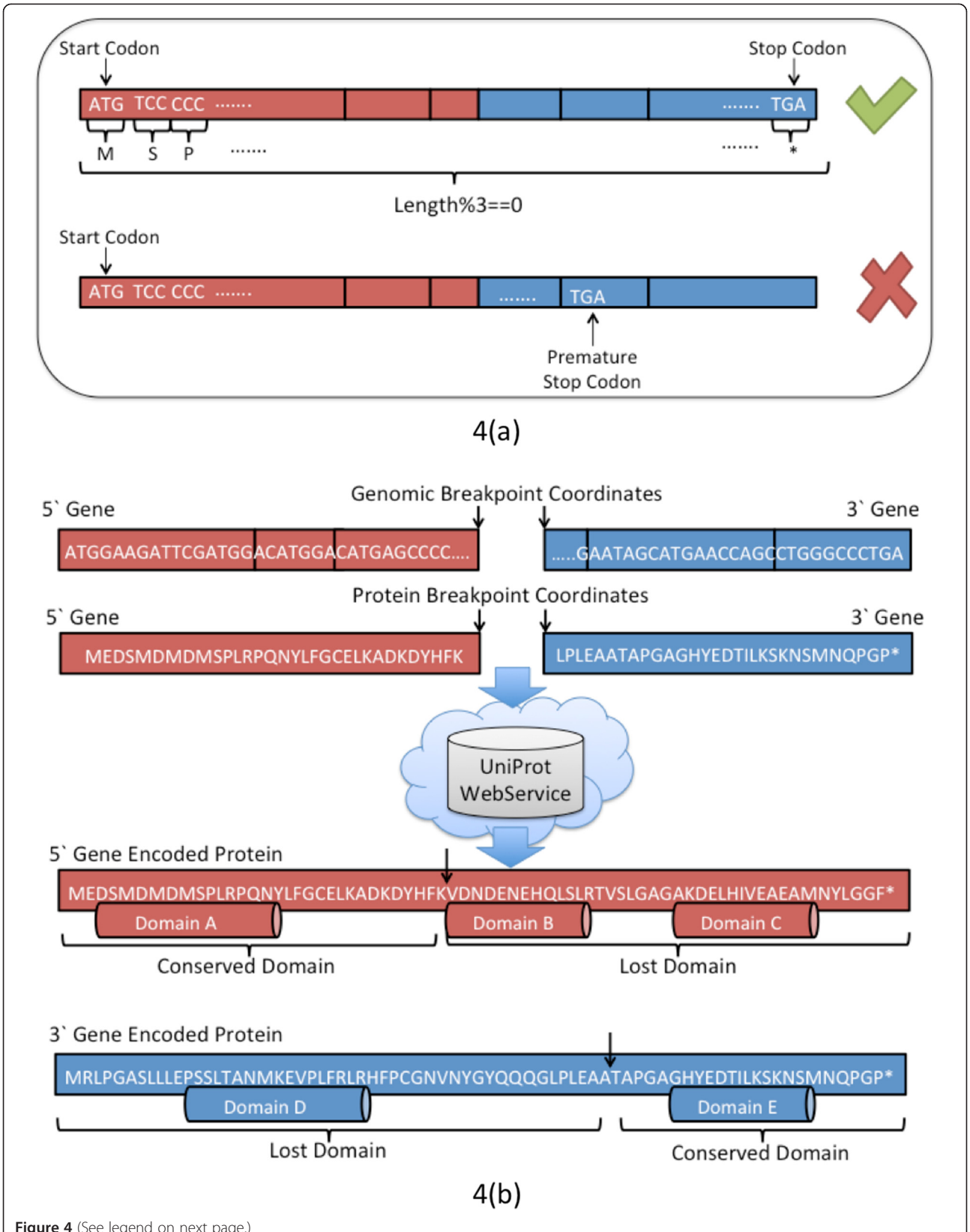

Figure 4 (See legend on next page.)

(See figure on previous page.)
Figure 4 Frame check and domain annotation. Figure 4**(a)** The length of the fusion transcript, from the start to the stop codon, must be multiple of three (three nucleotides per single encoded codon). If the length of the sequence module three is non-zero, the fusion sequence if frame-shifted. A premature stop codon can be introduced in the protein sequence. Figure 4**(b)** The nucleotide sequence resulting from the fusion of the 5′ and 3′ gene is translated into amino acid sequence. Similarly, the genomic breakpoint coordinates are translated into protein amino acid coordinates. UniProt Web Service is queried and the list of the available domains for both the gene is retrieved. On the basis of the protein domain sequence and protein breakpoint, the list of both conserved and lost domains is reported.

examples by N, the number of boosting rounds by M, and the loss function by $\mathscr{L}$.

$$(a)\ \text{For } i = 1, 2, \ldots, N \text{ compute}$$
$$\text{Let } y_i \equiv \text{class label of transcript } i$$
$$\text{Let } f_m(x) \equiv \text{classification function at boosting round } m$$
$$\text{Initialize } f_0(x) = \operatorname{argmin}_\gamma \sum_{i=1:N} \mathscr{L}(y_i, \gamma)$$
$$\text{For } m = 1 \text{ to } M:$$

$$(a)\ \text{For } i = 1, 2, \ldots, N \text{ compute}$$
$$r_{im} = -\left[\partial \mathscr{L}\left(\frac{y_i, f(x_i)}{\partial f(x_i)}\right]_{f=f_{m-1}}\right.$$

(b) Fit a regression tree to r_{im} producing regions R_{jm}, $j = 1, 2, \ldots, J_m$

$$\text{For } j = 1, 2, \ldots, J_m \text{compute}$$
$$(c)\ \gamma_{jm} = \operatorname{argmin}_\gamma \sum_{x_i \in R_{jm}} \mathscr{L}(y_i, f_{m-1}(x_i) + \gamma)$$
$$\text{For } j = 1, 2, \ldots, J_m \text{compute}$$
$$(d)\ \gamma_{jm} = \operatorname{argmin}_\gamma \sum_{x_i \in R_{jm}} \mathscr{L}(y_i, f_{m-1}(x_i) + \gamma)$$

An alternative ensemble classification strategy, the random forest algorithm, demonstrated comparable performance to gradient tree boosting in our experiments. There is recent precedent in the machine learning literature for initializing gradient tree boosting models with rankings learned via random forests for achieving superior performance to either algorithm alone [35]. Interestingly, those authors found that posing web-search rankings as a classification task rather than a regression task increased the performance of a gradient boosted regression tree model, confirming our hypothesis in constructing the current Pegasus classifier.

Results and Discussion

This section highlights the performance of Pegasus in detecting driver gene fusions. First, we examine the performance of the classifier on the training data and compare its effectiveness to a recently published tool, Oncofuse, on a separately curated validation dataset. Next, we run Pegasus on two experimental datasets and demonstrate its role in reducing the search space of potential oncogenic drivers by accurately ranking fusion transcripts from a vast set of putative candidates. The first is the publicly available RNA-Seq data of GBM from TCGA. The second is a non-public set of 23 RNA-Seq samples from a cohort of patients with ALCL, with 2 out of the 23 samples reporting the NPM1-ALK fusion.

We analyze these datasets with ChimeraScan or deFuse and apply Pegasus to the entire set of detected fusions. It is worth specifying that in the reported

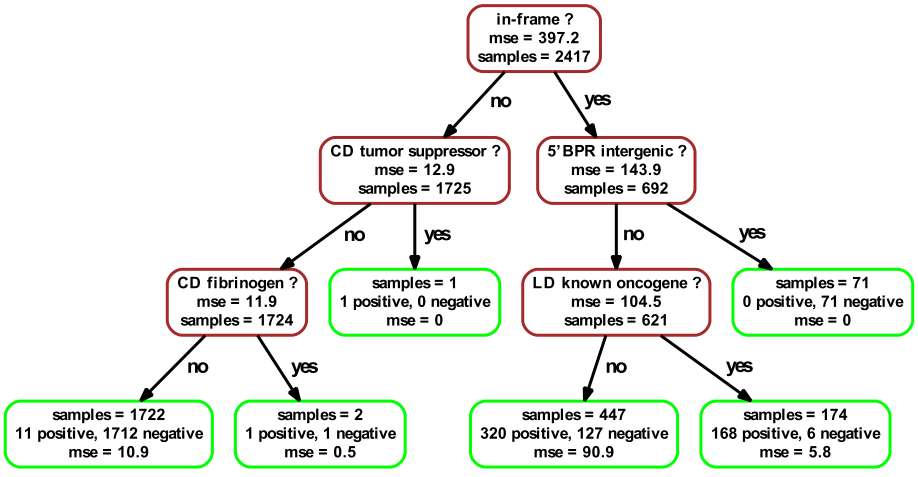

Figure 5 Decision tree base learner. The base learner in the boosted classification model is the decision tree, which defines a recursive partitioning of the feature space into disjoint regions each modeled by a constant. Decision nodes are colored brown while leaf nodes are colored green. At the leaf nodes, the samples have been partitioned into disjoint sets with a higher degree of label homogeneity than at the root.

results about chimeric transcript annotations, if two or more fusions share the same junction breakpoint coordinates, they are counted as a single fusion. The rationale is that according to the Pegasus fusion domain analysis, if two genes fuse in different samples with the same breakpoint they also share exactly the same domain. Conversely, if two genes occur in different samples with different junction breakpoint coordinates, the domain analysis accordingly changes.

Classifier performance on training corpus and independent validation set

The corpus of labeled data used to train the classifier comes from two sources. Positive examples, meaning true oncogenic driver fusions, are drawn from ChimerDB2.0, which contains 501 curated driver fusions. 1500 negative examples are then drawn from an internal collection of reactive lymph node tissue in patients with no clinical history of malignancy. The negative examples contain passenger fusions as well as read-through transcripts. We also supplement the negative training data with 416 deliberately frame-shifted transcripts from ChimerDB2.0 such that the necessary driver domains are lost. In total there are 501 positive examples and 1916 negative examples in the training corpus. The rationale for augmenting the negative set with 416 frame-shifted fusions from ChimerDB2.0 is to include the scenario of chimeric transcripts containing an oncogene at the 3' position that is frame-shifted. Since such events occur at low frequency in normal lymph node tissue, this design choice improves the performance of the classifier (for a detailed discussion please refer to Additional file 1). In summary, the 501 fusions from ChimerDB2.0 form the positive training set and provide mostly in-frame fusions involving oncogenes. The 1500 fusions from normal tissue contribute to the negative set and provide both in-frame and frame-shifted fusions. The 416 deliberately frame-shifted fusions from ChimerDB2.0 complete the negative set and provide frame-shifted gene fusions mostly with an oncogene at the 3' position.

The classifier is trained for 100 rounds of boosting under 10-fold stratified cross validation (CV) and achieves a mean test split AUROC of 0.96. As expected, in Figure 6b the loss on the train split monotonically decreases with increasing model complexity, while we see no sign of overfitting in the form of rising loss on the test split. Since each decision tree base learner implies a hierarchy of informative features, we can average over the boosting rounds to produce an aggregate view of the most important features in the classification task. Specifically, the relative feature importances in Figure 6a are computed via an ensemble average of the single decision tree feature importances as defined in Breiman *et al.* [36]:

$$\text{Let } I_k^2 \equiv \text{squared importance of feature } k$$
$$\text{Let } M \equiv \text{number of boosting rounds.}$$
$$\text{Let } T \text{ be a decision tree with } J-1 \text{ pairs } (t,\ \hat{i_t}^2) \text{ of}$$
$$(\text{internal decision node, estimated improvement in squared error risk}).$$
$$I_k^2(T) = \sum_{t=1:J-1} \hat{i_t}^2 \mathbb{I}(t=k).$$
$$\Rightarrow I_k^2 = \frac{1}{M} \sum_{m=1:M} I_k^2(T_m).$$

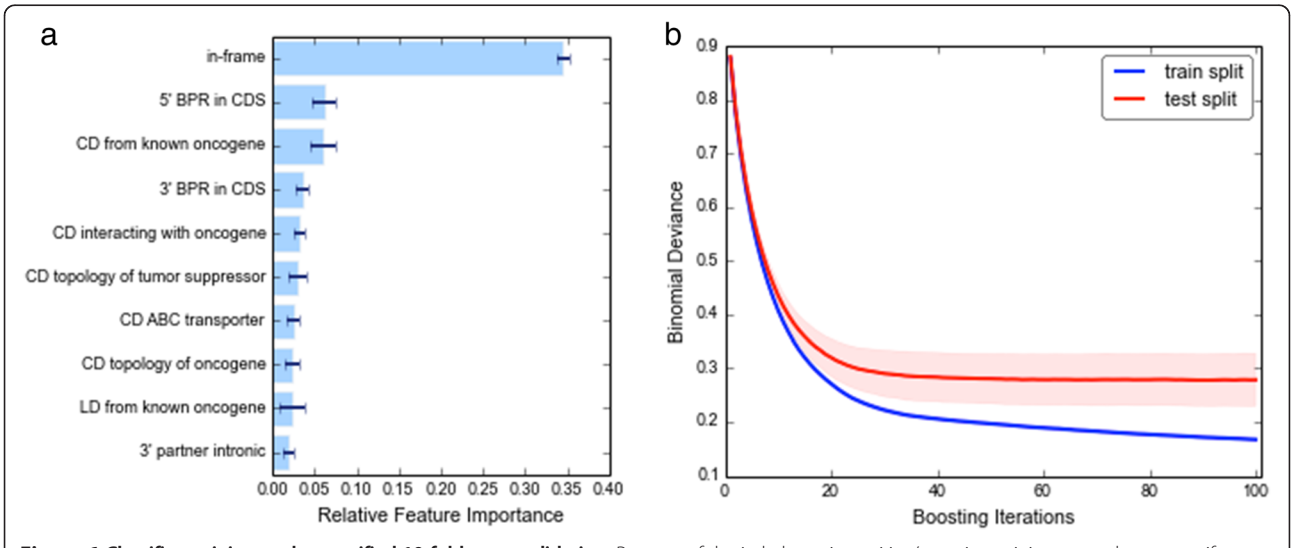

Figure 6 Classifier training under stratified 10-fold cross validation. Because of the imbalance in positive/negative training examples we stratify on the label and ensure that each 10% of data held out as a test set contains equal proportions of both labels. Figure 6**(a)** Relative feature importances averaged over the 10 CV folds. Figure 6**(b)** Binomial deviance loss function plotted against model complexity for both the train and test splits of the data.

We observe that the computationally expensive step of computing the fusion transcript reading frame is justified in the eyes of the classifier, as it is the single most informative feature. Looking a little further down the list we learn other transcript features that are highly informative of driver events, such as having breakpoints in the CDS and conserving domains shared with or interacting with known oncogenes.

Despite strong performance of the model under 10-fold CV on the training corpus, we are interested to see whether the classifier can generalize to new fusion transcripts that were unseen during the training phase. A list of 39 driver fusions, the majority of which are more recent than ChimerDB2.0, and corresponding to the validation set used in [27], is adopted as the positive validation set examples. To balance the label frequencies, we also select 39 transcripts from benign, reactive lymph node tissue as the negative validation set examples. None of the 78 validation examples are included in the training data. The negative examples are selected to contain at least one oncogene or tumor suppressor domain with the rationale that such transcripts more closely resemble driver fusions and would be most challenging for a classification function. In Figure 7a we demonstrate the favorable performance of the trained Pegasus classifier versus Oncofuse, the current state of the art in data-driven prediction of driver fusions. Since the ROC curve does not necessarily reveal how well separated the Pegasus scores are for the two class labels, we include Figure 7b to illustrate the remarkable resolution the classifier achieves between positive and negative examples. We also verify that Pegasus outperforms Oncofuse on randomly drawn sets of 39 non-oncogenic transcripts, though by a smaller margin (ROC curve in Additional file 2: Figure S1). This is to be expected because the majority of non-oncogenic fusions are very easily classified, whereas our curated subset represents a more challenging task. Such robust performance on manually curated data sets naturally leads to the next proving ground, applying Pegasus to the enormous candidate lists generated from real RNA-Seq samples.

Pegasus driver fusion predictions in non-public GBM data

In order to demonstrate the effectiveness of Pegasus in predicting driver fusions, we analyze 15 samples of short-term glioblastoma stem cells freshly isolated from individuals with GBM. RNA-Seq samples were first analyzed with ChimeraScan and deFuse [14,15] for fusion detection. Next, we apply Pegasus to the set of gene fusion candidates and consider as driver events all those fusions having a number of supporting reads greater than 10 and a Pegasus Driver Score (PDS) greater than 0.8. As shown in Figure 7(b), a threshold of PDS > 0.8 promises a good trade-off between specificity and sensitivity. Table 1 reports the 4 detected driver fusions. All fusions have been validated with RT-PCR (see Additional file 3: Figure S2) yielding a 100% rate of transcript validation. And while recurrence is often the surrogate measure of functional importance, the four unique

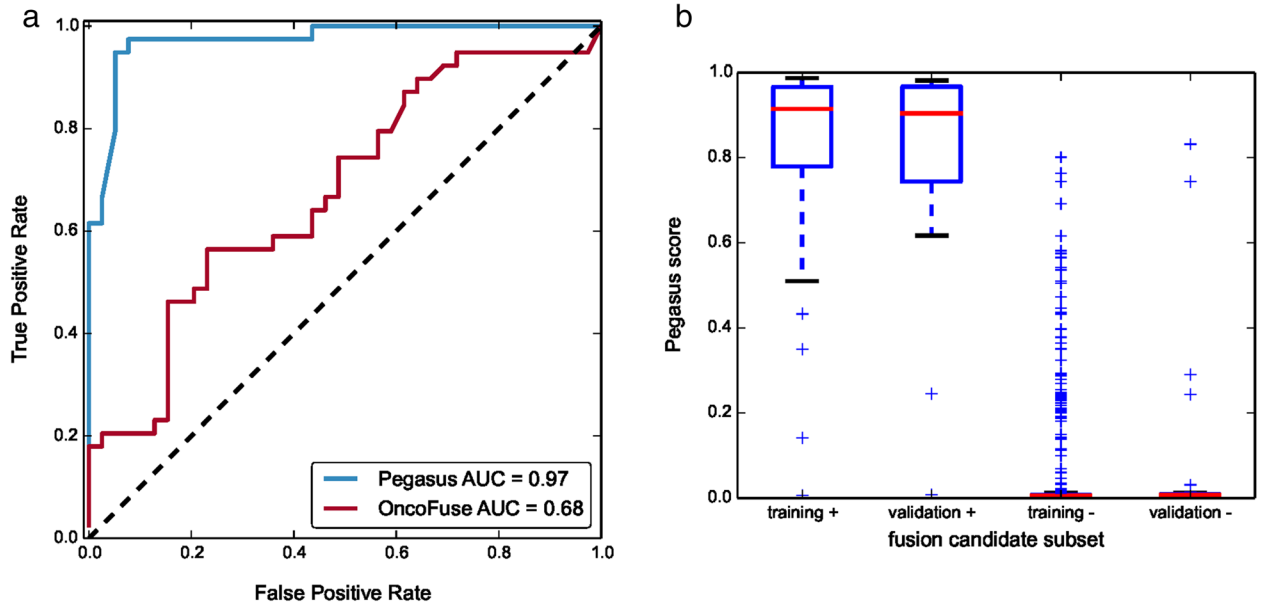

Figure 7 Classification performance of trained model. A curated set of 39 recently reported fusions that were not present in the ChimerDB2.0 training corpus are now used as an independent validation set. Figure 7**(a)** Superior classification performance of Pegasus compared to Oncofuse on the validation dataset. Figure 7**(b)** Boxplot demonstrating the high resolution of the Pegasus driver score in discerning the class boundary. Positive examples are consistently scored near 1 and negative examples consistently near 0.

Table 1 Pegasus predictions on 15 private GBM RNA-Seq data

5′ Gene partner	3′ Gene partner	Spanning reads	Split reads	Pegasus driver score	Validated
CAND1	EGFR	17	14	0.9437	YES
MAPK1	FAM119B	145	96	0.9426	YES
ADCK4	NUMBL	11	4	0.9426	YES
VOPP1	IL22	48	35	0.8243	YES

Pegasus top driver scores (PDS > 0.8) 4 new driver fusions in GBM data. Number of supporting reads, Pegasus Driver Score and RT-PCR validation status are shown.

driver candidates from this 15 patient sample still contain features associated with oncogenic gene fusions. In fact, CAND1-EGFR has been reported in [37] and the EGFR gene has been demonstrated to have an oncogenic role in GBM. Moreover, fusions involving MAPK1 and VOPP1 are reported as frequent in GBM with different gene partners [37,38]. These results show that Pegasus can successfully detect relevant driver fusion candidates from RNA-Seq data and that a threshold of PDS > 0.8 and number of supporting reads greater than 10 provide a strong transcript validation rate.

Pegasus driver fusion predictions in public TCGA GBM data

As the most common and deadly primary brain cancer, GBM has recently undergone a deep investigation by the multi-institutional consortium, the cancer genome atlas (TCGA). TCGA makes its collected RNA-Seq data available to the larger scientific community, and we analyze a set of 161 samples from their GBM cohort. We first analyze the 161 RNA-Seq samples with ChimeraScan (default parameters) [15], detecting a total of 9349 unique fusions across the entire dataset. Next, we apply Pegasus to the set of candidates and consider as driver events all fusion transcripts having a number of supporting reads greater than 10, Pegasus Driver Score (PDS) greater than 0.8 and recurrence greater than 1 (for the complete list see Additional file 4: Table S1). As shown in the non-public RNA-Seq data, these filtering thresholds provide a good validation rate by RT-PCR. The application of these filters reduces the original list of 9349 candidates down to 13 high confidence fusions, making further functional analysis and validation tractable. Pegasus computes a score greater than 0.8 for both FGFR3-TACC3 and EGFR-SEPT14 gene fusions, which are already reported as driver translocation events in GBM [10,37]. However, since TCGA biological material is not available, we are unable to perform further functional analysis of all predicted driver fusions with experimental procedures. Nonetheless, in order to validate Pegasus performance we compare PDS values with the frequencies reported in both Frattini *et al.* and Brennar *et al.* (see Table 2). Of the 13 high confidence Pegasus predictions, 6 are recurrent in TCGA data suggesting a potential functional driver role in GBM tumorgenesis [39]. Some of the recurrent fusions involve the EGFR gene that is usually amplified in GBM, where it is known to activate STAT3 signaling and is thus a drug target. Particularly interesting is also the BCAN-NTRK1 gene fusion. In fact, NTRK1 is often translocated with different partners in cancers beyond just GBM [37].

Table 2 Pegasus predictions on GBM RNA-Seq data

5′ Gene partner	3′ Gene partner	Pegasus driver score	Recurrence
YEATS4	XRCC6BP1	0.9598	1
EIF4H	GTF2I	0.9440	1
ASH1L	C1orf61	0.9256	1
SEC61G	EGFR	0.9234	4
BCAN	NTRK1	0.9182	1
EGFR	VOPP1	0.9130	2
EGFR	SEPT14	0.9042	6
TDRD3	ESD	0.8959	1
TFG	GPR128	0.8880	4
PPP2R2B	CCT3	0.8699	1
TBC1D14	HTRA3	0.8697	1
FGFR3	TACC3	0.8442	3
LANCL2	SEPT14	0.8428	3

Pegasus top driver scores (PDS > 0.8) predicts 46% of known driver gene fusions in GBM data from TCGA cohort. Recurrence is assessed on the basis of gene fusion frequency reported in Frattini *et al.* and Brennar *et al.*

Functional validation of new recurrent driver fusion in anaplastic large cell lymphoma

Anaplastic large cell lymphoma (ALCL) is a form of peripheral T-cell lymphoma that is often associated with translocations of the ALK gene. In 23 non-public ALCL samples (~450 million properly mapped reads) we detect a total of 5201 candidate fusion transcripts by means of deFuse [14] and ChimeraScan [15]. Beyond the two NPM1-ALK fusion transcripts (PDS = 0.98) that are already reported, Pegasus properly annotates and reveals 16 new biologically relevant fusions in these 23 samples. All 16 candidate driver fusions have been validated with RT-PCR, and 4 gene fusions have successfully undergone functional assays and *in vivo* validation. An example of Pegasus' effectiveness in functional domain analysis lies in the oncogenic role of TRAF1-ALK [40], a novel fusion in ALCL that Pegasus reports as driver.

The TRAF1-ALK fusion has been reported in three cases of ALCL (one in [40] and two in Abate *et al.*, manuscript under review) suggesting a driver role. Pegasus accurately assembles and annotates the in-frame fusion sequence and correctly detects that the ALK protein kinase domain is completely conserved. Various fusions involving the ALK gene have been reported in

literature and the oncogenic effect is generally promoted by ALK signaling. Interestingly, TRAF1 is also known to be involved in both the canonical and non-canonical NFkB pathway. Pegasus correctly annotates that the meprin and TRAF-C homology (MATH) domain is conserved, a domain that ubiquinates the IKK complex, activating NFkB transcription factors. As depicted in Figure 8, Pegasus properly identifies both the presence of an oncogenic protein domain (ALK) and an interacting oncogenic domain (TRAF1 to NFKB activation). Experimental work demonstrates and validates this Pegasus prediction showing the oncogenic effect of TRAF1-ALK *in vivo*, with activation of both ALK and NFkB signaling (Abate *et al*., manuscript under review).

Conclusion

Since the first application of whole transcriptome sequencing to gene fusion discovery in 2009 [8], many new aberrant events have been reported, opening an exciting frontier for molecular understanding of cancer biology and targeted therapies. The unprecedented sensitivity of NGS technology, however, often yields numbers of fusion candidates too large to be experimentally validated. The frontier is made ever more exciting (and challenging) by large consortia such as TCGA and ICGC (International Cancer Genome Consortium) who are making available large sets of RNA-Seq samples spanning the spectrum of human malignancies. The analysis of this data has been revealing the limits of the theory that associates *driver* events with *recurrent* events. In fact, out of 161 RNA-seq GBM samples, the most frequent fusion (EGFR-SEPT14) occurs in only 6 samples, and the highly expressed FGFR3-TACC3 fusion is recurrent in only the 3% of GBM cases. Thus, this data suggest that in order to select relevant driver fusion candidates for biological validation, a functional analysis of the putative gene fusion candidate is necessary.

Here we present Pegasus, an accurate prediction tool for the discovery of new driver gene fusions in cancer studies. The proposed methodology is based on a computational model of the features that make chimeric transcript a driver oncogenic event. The framework provides a common interface for several fusion detection tools and it predicts driver events by properly analyzing the detected gene fusion candidates according to the assembled fused sequence.

The application of ensemble learning techniques reveals the most informative features in discriminating oncogenic gene fusions. The data confirm our intuition that an accurate analysis of fusion transcript sequence is necessary. The reading frame in particular is a dominant features in the discrimination of passenger and driver fusions. Similarly, the molecular characterization of the main reported oncogenic domains accurately increase the sensitivity and the PDS computation.

The problem of computationally assessing the biological and clinical relevance of a gene fusion is still very much an open question. However, some driver prediction tools have been recently proposed. To better determine Pegasus performance and accuracy, we compare our predicted results with Oncofuse. The data confirm that an approximate analysis of the fusion transcript sequence negatively impacts the performance of the algorithm. Using a set of known driver fusions as positive examples and a set of passenger fusions from normal tissue as negative examples, we observe the superior performance of Pegasus in ROC space where its AUC is 0.97.

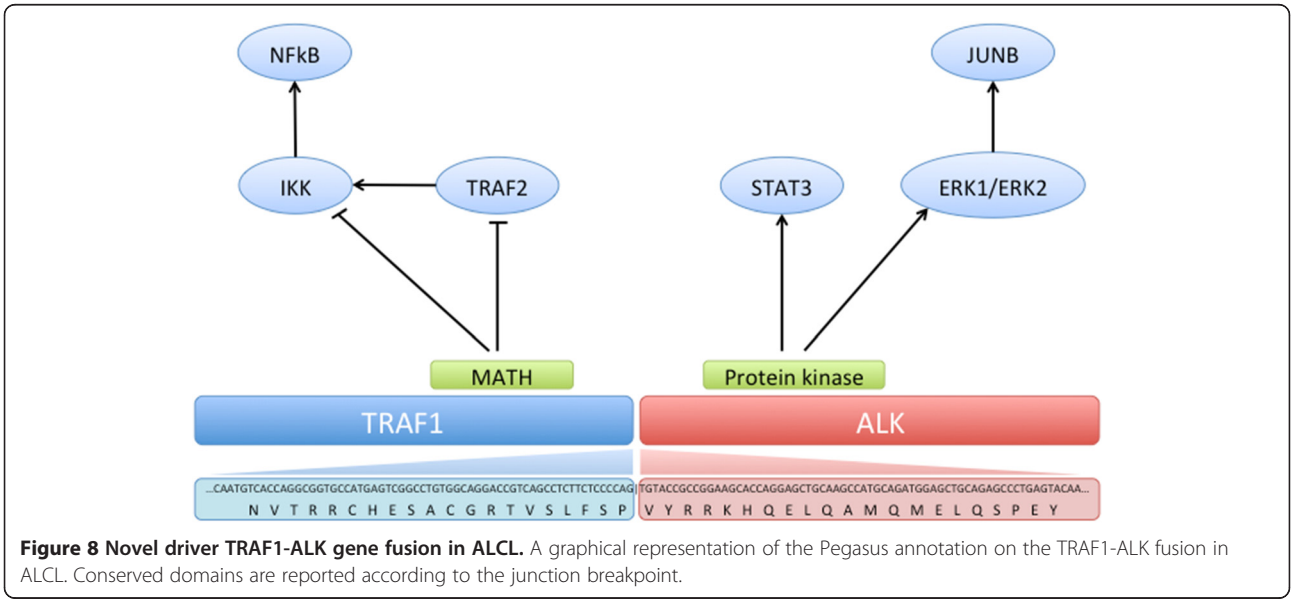

Figure 8 Novel driver TRAF1-ALK gene fusion in ALCL. A graphical representation of the Pegasus annotation on the TRAF1-ALK fusion in ALCL. Conserved domains are reported according to the junction breakpoint.

Moreover, we demonstrate the practical role of the Pegasus framework in computing PDS scores that allow for triaging lists of gene fusion candidates for experimental validation in two actual case studies. In the first, we apply Pegasus to public GBM TCGA data and almost 50% of the detected driver fusions turn out to have been reported as a biologically relevant in independent studies [37,38]. In the second, we compute Pegasus scores for internal ALCL data and we successfully detect novel oncogenic and targetable driver fusions that have undergone complete functional and experimental validation. In this work, we extensively report the driver fusion TRAF1-ALK that has been correctly detected and highly ranked by Pegasus.

Finally, the accuracy of Pegasus in detecting driver fusions in both the curated validation dataset and the real biological cases demonstrates the efficacy of the framework in supporting biological analysis and cancer research. We believe that the Pegasus prediction score, as well as the accurate annotations provided via our feature engineering, will be of great use to other investigators searching for biologically relevant gene fusions in NGS data.

Availability and Requirements

Project name: Pegasus
Project home page: http://sourceforge.net/p/pegasus-fus.
Operating system: UNIX
Programming language: Java, Perl, Python, BASH.

Additional files

Additional file 1: Supplementary details about the training set design and description are provided. Two important points are hereby addressed. First is the issue of why we augment the negative training set with deliberately frame-shifted fusions. The second regards the issue of whether we can quantify the importance of the "in-frame" feature given that we have altered the training set composition.

Additional file 2: Figure S1. Performances of Pegasus compared to Oncofuse on randomly drawn sets of 39 non-oncogenic transcripts.

Additional file 3: Figure S2. RT-PCR on 4 fusions (MAPK1-METTL21B, CAND1-EGFR, VOPP1-IL22, ADCK4-NUMBL) that Pegasus predicted as highly oncogenic in TCGA GBM data.

Additional file 4: Table S1. Complete list of candidates considered as putative driver fusions in TCGA GBM dataset. All fusion transcripts having a number of supporting reads greater than 10, Pegasus Driver Score (PDS) greater than 0.8 and recurrence greater than 1 are reported.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

FA, SZ designed and developed the tool; RR supervised the project; EF, AA, CHW contributed to the architecture design; AI, AL, VF, GI performed biological validations; FA, SZ, RR wrote the manuscript and all the authors improved it with significant revisions. All authors read and approved the final manuscript.

Funding

This project is supported by U54 CA121852 (R.R.), R01 CA179044-01A1 (R.R.), R01 CA164152-01, Stewart Trust Foundation and The Italian Association for Cancer Research (AIRC) Special Program in Clinical Molecular Oncology (5x1000 No. 10007, Milan, Italy), Regione Piemonte (ONCOPROT, CIPE 25/2005), ImmOnc (Innovative approaches to boost the immune responses, Programma Operativo Regionale, Piattaforme Innovative BIO F.E.S.R. 2007/13, Asse 1 'Ricerca e innovazione' della LR 34/2004), and the Oncology Program of Compagnia di San Paolo (Turin, Italy) (G.I.).

Author details

[1]Department of Biomedical Informatics, Columbia University, 1130 St. Nicholas Ave, New York, NY 10032, USA. [2]Center for Computational Biology and Bioinformatics, Columbia University, 1130 St. Nicholas Ave, New York, NY 10032, USA. [3]Department of Control and Computer Engineering, Politecnico di Torino, Torino 10129, Italy. [4]Department of Pathology, Center for Experimental Research and Medical Studies, Laboratory of Functional Genomics, University of Torino, Torino, Italy. [5]Institute for Cancer Genetics, Columbia University Medical Center, New York, New York, USA. [6]Department of Applied Physics and Applied Mathematics, The Fu Foundation School for Engineering and Applied Sciences, Columbia University, 500 W. 120th Street, Mudd 200, MC 4701, New York, New York 10027, USA. [7]Institute for Data Sciences and Engineering, Columbia University, 500 W. 120th Street, Mudd 524, New York, New York 10027, USA.

Received: 6 February 2014 Accepted: 5 August 2014
Published online: 04 September 2014

References

1. Nowell P, Hungerford D: **A minute chromosome in chronic granulocytic leukemia.** *Science* 1960, **132**(3438):1488–1501.
2. Zhao X, Ghaffari S, Lodish H, Malashkevich VN, Kim PS: **Structure of the Bcr-Abl oncoprotein oligomerization domain.** *Nat Struct Biol* 2002, **9**(2):117–120.
3. Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun XW, Varambally S, Cao X, Tchinda J, Kuefer R, Lee C, Montie JE, Shah RB, Pienta KJ, Rubin MA, Chinnaiyan AM: **Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer.** *Science* 2005, **310**(5748):644–648.
4. Merson S, Jhavar S, Flohr P, Edwards S, Foster CS, Eeles R, Martin FL, Phillips DH, Crundwell M, Christmas T, Thompson A, Fisher C, Kovacs G, Cooper CS: **Diversity of TMPRSS2-ERG fusion transcripts in the human prostate.** *Oncogene* 2007, **26**(18):2667–2673.
5. Voena C, Ambrogio C, Piva R, Inghirami G: **The anaplastic lymphoma kinase in the pathogenesis of cancer.** *Nat Rev Cancer* 2008, **8**(1):11–23.
6. Morris SW, Kirstein MN, Valentine MB, Dittmer KG, Shapiro DN, Saltman DL, Look AT: **Fusion of a kinase gene, ALK, to a nucleolar protein gene, NPM, in non-Hodgkin's lymphoma.** *Science* 1994, **263**(5151):1281–1284.
7. Soda M, Choi YL, Enomoto M, Takada S, Yamashita Y, Ishikawa S, Fujiwara S, Watanabe H, Kurashina K, Hatanaka H, Bando M, Ohno S, Ishikawa Y, Aburatani H, Niki T, Sohara Y, Sugiyama Y, Mano H: **Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer.** *Nature* 2007, **448**(7153):561–566.
8. Maher CA, Palanisamy N, Brenner JC, Cao X, Kalyana-Sundaram S, Luo S, Khrebtukova I, Barrette TR, Grasso C, Yu J, Lonigro RJ, Schroth G, Kumar-Sinha C, Chinnaiyan AM: **Chimeric transcript discovery by paired-end transcriptome sequencing.** *Proc Natl Acad Sci U S A* 2009, **106**(30):12353–12358.
9. Steidl C, Shah SP, Woolcock BW, Rui L, Kawahara M, Farinha P, Johnson NA, Zhao Y, Telenius A, Neriah SB, McPherson A, Meissner B, Okoye UC, Diepstra A, van den Berg A, Sun M, Leung G, Jones SJ, Connors JM, Huntsman DG, Savage KJ, Rimsza LM, Horsman DE, Staudt LM, Steidl U, Marra MA, Gascoyne RD: **MHC class II transactivator CIITA is a recurrent gene fusion partner in lymphoid cancers.** *Nature* 2011, **471**(7338):377–381.
10. Singh D, Chan JM, Zoppoli P, Niola F, Sullivan R, Castano A, Liu EM, Reichel J, Porrati P, Pellegatta S, Qiu K, Gao Z, Ceccarelli M, Riccardi R, Brat DJ, Guha A, Aldape K, Golfinos JG, Zagzag D, Mikkelsen T, Finocchiaro G, Lasorella A, Rabadan R, Iavarone A: **Transforming fusions of FGFR and TACC genes in human glioblastoma.** *Science* 2012, **337**(6099):1231–1235.
11. Williams SV, Hurst CD, Knowles MA: **Oncogenic FGFR3 gene fusions in bladder cancer.** *Hum Mol Genet* 2013, **22**(4):795–803.

12. Majewski IJ, Mittempergher L, Davidson NM, Bosma A, Willems SM, Horlings HM, de Rink I, Greger L, Hooijer GK, Peters D, Nederlof PM, Hofland I, de Jong J, Wesseling J, Kluin RJ, Brugman W, Kerkhoven R, Nieboer F, Roepman P, Broeks A, Muley TR, Jassem J, Niklinski J, van Zandwijk N, Brazma A, Oshlack A, van den Heuvel M, Bernards R: **Identification of recurrent FGFR3 fusion genes in lung cancer through kinome-centred RNA sequencing.** *J Pathol* 2013, **230**(3):270–276.
13. Sboner A, Habegger L, Pflueger D, Terry S, Chen DZ, Rozowsky JS, Tewari AK, Kitabayashi N, Moss BJ, Chee MS, Demichelis F, Rubin MA, Gerstein MB: **FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data.** *Genome Biol* 2010, **11**(10):R104.
14. McPherson A, Hormozdiari F, Zayed A, Giuliany R, Ha G, Sun MG, Griffith M, Heravi Moussavi A, Senz J, Melnyk N, Pacheco M, Marra MA, Hirst M, Nielsen TO, Sahinalp SC, Huntsman D, Shah SP: **deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data.** *PLoS Comput Biol* 2011, **7**(5):e1001138.
15. Iyer MK, Chinnaiyan AM, Maher CA: **ChimeraScan: a tool for identifying chimeric transcription in sequencing data.** *Bioinformatics* 2011, **27**(20):2903–2904.
16. Abate F, Acquaviva A, Paciello G, Foti C, Ficarra E, Ferrarini A, Delledonne M, Iacobucci I, Soverini S, Martinelli G, Macii E: **Bellerophontes: an RNA-Seq data analysis framework for chimeric transcripts discovery based on accurate fusion model.** *Bioinformatics* 2012, **28**(16):2114–2121.
17. Trapnell C, Pachter L, Salzberg SL: **TopHat: discovering splice junctions with RNA-Seq.** *Bioinformatics* 2009, **25**(9):1105–1111.
18. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.** *Nat Biotechnol* 2010, **28**(5):511–515.
19. Edgren H, Murumagi A, Kangaspeska S, Nicorici D, Hongisto V, Kleivi K, Rye IH, Nyberg S, Wolf M, Borresen-Dale AL, Kallioniemi O: **Identification of fusion genes in breast cancer by paired-end RNA-sequencing.** *Genome Biol* 2011, **12**(1):R6.
20. Carrara M, Beccuti M, Lazzarato F, Cavallo F, Cordero F, Donatelli S, Calogero RA: **State-of-the-art fusion-finder algorithms sensitivity and specificity.** *Biomed Res Int* 2013, **2013**:340620.
21. Ozsolak F, Milos PM: **RNA sequencing: advances, challenges and opportunities.** *Nat Rev Genet* 2011, **12**(2):87–98.
22. Mitelman F, J.B.a.M.F: **Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer**. In; 2013. Available from: http://cgap.nci.nih.gov/Chromosomes/Mitelman.
23. Novo FJ, de Mendibil IO, Vizmanos JL: **TICdb: a collection of gene-mapped translocation breakpoints in cancer.** *BMC Genomics* 2007, **8**:33.
24. Kim P, Yoon S, Kim N, Lee S, Ko M, Lee H, Kang H, Kim J, Lee S: **ChimerDB 2.0–a knowledgebase for fusion genes updated.** *Nucleic Acids Res* 2010, **38**(Database issue):D81–D85.
25. Wang XS, Prensner JR, Chen GA, Cao Q, Han B, Dhanasekaran SM, Ponnala R, Cao XH, Varambally S, Thomas DG, Giordano TJ, Beer DG, Palanisamy N, Sartor MA, Omenn GS, Chinnaiyan AM: **An integrative approach to reveal driver gene fusions from paired-end sequencing data in cancer.** *Nat Biotechnol* 2009, **27**(11):1005.
26. Wu CC, Kannan K, Lin S, Yen L, Milosavljevic A: **Identification of cancer fusion drivers using network fusion centrality.** *Bioinformatics* 2013, **29**(9):1174–1181.
27. Shugay M, Ortiz De Mendibil I, Vizmanos JL, Novo FJ: **Oncofuse: a computational framework for the prediction of the oncogenic potential of gene fusions.** *Bioinformatics* 2013, **29**(20):2539–2546.
28. Jin Y, Mertens F, Kullendorff CM, Panagopoulos I: **Fusion of the tumor-suppressor gene CHEK2 and the gene for the regulatory subunit B of protein phosphatase 2 PPP2R2A in childhood teratoma.** *Neoplasia* 2006, **8**(5):413–418.
29. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, Gil L, Gordon L, Hendrix M, Hourlier T, Johnson N, Kahari AK, Keefe D, Keenan S, Kinsella R, Komorowska M, Koscielny G, Kulesha E, Larsson P, Longden I, McLaren W, Muffato M, Overduin B, Pignatelli M, Pritchard B, Riat HS, *et al*: **Ensembl 2012.** *Nucleic Acids Res* 2012, **40**(Database issue):D84–D90.
30. UniProt Consortium: **Reorganizing the protein space at the Universal Protein Resource (UniProt).** *Nucleic Acids Res* 2012, **40**(Database issue):D71–D75.
31. Friedman J, Hastie T, Tibshirani R: **Additive Logistic Regression: a Statistical View of Boosting.** *Ann Stat* 2000, **28**(2):337–407.
32. Friedman JH: **Greedy Function Approximation: A Gradient Boosting Machine.** *Ann Stat* 2000, **29**:1189–1232.
33. Hastie T, Tibshirani R, Friedman JH: **The Elements of Statistical Learning.** *Springer Series Stat* 2001.
34. Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E: **Scikit-learn: Machine Learning in Python.** *J Mach Learn Res* 2011, **12**:2825–2830. MIT Press.
35. Ananth Mohan ZC: **Kilian Weinberger Web-Search Ranking with Initialized Gradient Boosted Regression Trees.** *JMLR: Workshop and Conference Proceedings* 2011, **14**:77–89.
36. Breiman L: **Classification and regression trees**. In ; 1984.
37. Frattini V, Trifonov V, Chan JM, Castano A, Lia M, Abate F, Keir ST, Ji AX, Zoppoli P, Niola F, Danussi C, Dolgalev I, Porrati P, Pellegatta S, Heguy A, Gupta G, Pisapia DJ, Canoll P, Bruce JN, McLendon RE, Yan H, Aldape K, Finocchiaro G, Mikkelsen T, Prive GG, Bigner DD, Lasorella A, Rabadan R, Iavarone A: **The integrated landscape of driver genomic alterations in glioblastoma.** *Nat Genet* 2013, **45**(10):1141–1149.
38. Brennan CW, Verhaak RG, McKenna A, Campos B, Noushmehr H, Salama SR, Zheng S, Chakravarty D, Sanborn JZ, Berman SH, Beroukhim R, Bernard B, Wu CJ, Genovese G, Shmulevich I, Barnholtz-Sloan J, Zou L, Vegesna R, Shukla SA, Ciriello G, Yung WK, Zhang W, Sougnez C, Mikkelsen T, Aldape K, Bigner DD, Van Meir EG, Prados M, Sloan A, Black KL, *et al*: **The somatic genomic landscape of glioblastoma.** *Cell* 2013, **155**(2):462–477.
39. Stratton MR, Campbell PJ, Futreal PA: **The cancer genome.** *Nature* 2009, **458**(7239):719–724.
40. Feldman AL, Vasmatzis G, Asmann YW, Davila J, Middha S, Eckloff BW, Johnson SH, Porcher JC, Ansell SM, Caride A: **Novel TRAF1-ALK fusion identified by deep RNA sequencing of anaplastic large cell lymphoma.** *Genes Chromosomes Cancer* 2013, **52**(11):1097–1102.

doi:10.1186/s12918-014-0097-z
Cite this article as: Abate *et al.*: Pegasus: a comprehensive annotation and prediction tool for detection of driver gene fusions in cancer. *BMC Systems Biology* 2014 **8**:97.