BIOINFORMATICS ORIGINAL PAPER
Vol. 29 no. 9 2013, pages 1174–1181
doi:10.1093/bioinformatics/btt131

Systems biology
Advance Access publication March 16, 2013

Identification of cancer fusion drivers using network fusion centrality

Chia-Chin Wu[1,2,*], Kalpana Kannan[3], Steven Lin[4], Laising Yen[3] and Aleksandar Milosavljevic[2]

[1]Department of Genomic Medicine, UT MD Anderson Cancer Center, [2]Bioinformatics Research Laboratory, [3]Department of Molecular and Human Genetics, [4]Department of Pathology and Immunology, Baylor College of Medicine and [5]Department of Radiation Oncology, UT MD Anderson Cancer Center, Houston, TX 77030, USA

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: Gene fusions are being discovered at an increasing rate using massively parallel sequencing technologies. Prioritization of cancer fusion drivers for validation cannot be performed using traditional single-gene based methods because fusions involve portions of two partner genes. To address this problem, we propose a novel network analysis method called fusion centrality that is specifically tailored for prioritizing gene fusions. We first propose a domain-based fusion model built on the theory of exon/domain shuffling. The model leads to a hypothesis that a fusion is more likely to be an oncogenic driver if its partner genes act like hubs in a network because the fusion mutation can deregulate normal functions of many other genes and their pathways. The hypothesis is supported by the observation that for most known cancer fusion genes, at least one of the fusion partners appears to be a hub in a network, and even for many fusions both partners appear to be hubs. Based on this model, we construct fusion centrality, a multi-gene-based network metric, and use it to score fusion drivers. We show that the fusion centrality outperforms other single gene-based methods. Specifically, the method successfully predicts most of 38 newly discovered fusions that had validated oncogenic importance. To our best knowledge, this is the first network-based approach for identifying fusion drivers.

Availability: Matlab code implementing the fusion centrality method is available upon request from the corresponding authors.

Contact: perwu777@gmail.com

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

Received on October 30, 2012; revised on February 13, 2013; accepted on March 11, 2013

1 INTRODUCTION

A fusion gene that results from chromosome rearrangements is a hybrid gene formed from distinct genes, either in the same chromosome or in different chromosomes. Many recurrent gene fusions, such as BCR-ABL in chronic myelogenous leukemia (Rowley, 1973), EWS-FIL1 in Ewing's Sarcoma (Aurias, 1984) and TMPRSS2-ERG in prostate cancer (Tomlins *et al.*, 2005), play important roles in cancer progression. Gene fusions can lead to oncogenesis through one of several alternative mechanisms: (i) a strong promoter of one gene is fused to and hence controls the open reading frame of a second proto-oncogene, resulting in the deregulation of its oncogenic function; (ii) the fused gene encodes a fusion protein that leads to new oncogenic function; and (iii) fusion results in the truncation or loss of function of a tumor suppressor gene. In addition, fused transcripts (chimeric RNAs), generated by trans-splicing or read-through events (Kannan *et al.*, 2011) and unrelated to chromosomal rearrangement, may also lead to similar, if not identical, biological consequences for cancer progression.

Chromosome banding analysis and fluorescence *in situ* hybridization have long been used to detect fusion genes. The advent of massively parallel sequencing greatly accelerated the rate of discovery of novel fusion genes and chimeric RNAs (Kannan *et al.*, 2011; Palanisamy *et al.*, 2010; Robinson *et al.*, 2011). A number of bioinformatics methods have recently been developed for identifying gene fusions from sequencing data (Kim and Salzberg 2011; McPherson *et al.*, 2011a, b; Sboner *et al.*, 2010). However, recent studies indicate that the majority of these fusions are likely to constitute passenger events in cancer progression (Stephens *et al.*, 2009). Up to now, only a few identified fusions have been shown to have oncogenic effects, in part due to the labor-intensive nature of required validation assays. Thus, a computational method that allows prioritization of potential fusion drivers would greatly facilitate the discovery of important fusions.

One well established approach to discover driver point mutations is to determine whether its observed mutation rate is significantly higher than the background mutation rate (Ding *et al.*, 2008; Jones *et al.*, 2008). However, this approach is not suitable for the discovery of fusion drivers that involve two genes. Besides, because fusions are rare, background rate for fusion mutation would be hard to be established. Alternatively, network analysis has been successfully used to identify rare cancer driver genes with point mutations without estimating mutation rate (Torkamani and Schork 2009; Wu *et al.*, 2012). These network methods estimate the tendency of individual mutated genes to result in oncogenesis in molecular networks. Prioritization of fusion genes raises a new challenge for the existing single gene-based approaches because fusions result in hybrid genes formed from portions of two genes. Only few attempts have so far been made to study network properties of fusion genes. One study (Höglund *et al.*, 2006) constructed a fusion network based on

*To whom correspondence should be addressed.

all partnerships of known fusion gene pairs. In this analysis, nodes were fusion partner genes and edges represented partnerships between fusion partner genes. The constructed gene fusion network showed a scale-free topology. However, this analysis did not address the problem of prioritizing fusion drivers. Another study developed a computational method to nominate fusion drivers by assessing the association of their partner genes with the identified biological concepts (i.e. gene sets with specific biological functions) (Wang *et al.*, 2009). This method, similar to the aforementioned single gene-based approaches, only accounted for the impact of a single gene. A recent study (Vandin *et al.*, 2011) proposed a multiple gene-based approach to identify groups of proteins that are potential cancer drivers. However, this approach is not specifically developed for identifying fusion drivers.

To address the problem, we here propose fusion centrality method for prioritizing fusion drivers in network context. We first test whether gene fusions are associated with exon/domain shuffling by enrichment analysis of protein domain co-occurrence and protein domain–domain interactions between protein domains of known cancer fusion partner genes. We then proposed a domain-based fusion model, assuming that the potential impact level of a fusion in a network would be associated with all the protein domains or biological interactions of its two partner genes. The model leads to the hypothesis that a fusion is more likely to be an oncogenic driver if its corresponding partner genes are involved in extensive biological interactions (i.e. hub genes in a network). We then test this hypothesis using a network centrality analysis on known cancer fusion partner genes, and then led to the development of fusion centrality, a multiple gene-based network metric, to prioritize driver fusion genes from passengers. We evaluated the performance of the fusion centrality model by a comparison with other single gene-based methods. Finally, we applied the fusion centrality approach to 38 newly discovered cancer fusions.

2 METHODS

2.1 Data sources

A total of 289 known fusion gene pairs and 109 genes with point mutations that have been proven to be causally linked to oncogenesis were downloaded from the Cancer Gene Census database (Futreal *et al.*, 2004) for analysis. We also downloaded a curated human gene network from Pathway Commons (Cerami *et al.*, 2010), consisting of 11 570 genes and >1 000 000 molecular interactions, for network centrality analysis of these fusion partner genes and point mutation genes.

Two types of protein domain data in humans are used for the analysis of domain shuffling in gene fusion. One is protein domain co-occurrence relationships. The UCSC Genome Browser (Fujita *et al.*, 2011) provides data on Pfam domain families for protein-coding genes. Any two domains are considered to have a co-occurrence relationship if they co-exist in at least one protein. In all, we identified 5135 domain co-occurrence relationships among 4412 Pfam domains involving 12 864 genes. The other is a set of 26 219 curated domain–domain interactions among 5410 Pfam domains involving 14 936 genes (Yellaboina *et al.*, 2011).

2.2 Fusion centrality

Centrality measures that are normally associated with one node in a network have been used to determine hub genes in a network (Section S1 of the Supplementary Material). These measures have been used to identify genes related to cancer and other diseases (Barabási *et al.*, 2011). However, these methods are not able to measure network centrality of a fusion. In this study, we hypothesized that the importance of a fusion in a network would be associated with all biological interactions of its partner genes (Section 3.2). Based on this hypothesis, we quantify the importance of a fusion based on the new concept of fusion centrality. A new node in a network, representing fusion between two existing nodes (i.e. the two partner genes in a fusion), inherits all the linkages of the two nodes when it is first created, but those replicated linkages with other nodes are ignored. We then postulate that the centrality of the new node represents the importance of a fusion between two nodes in a network. Thus, it is intuitive to consider the definition of degree centrality of a fusion in terms of total linkages (by ignoring those replicated linkages with other nodes) of its two partner nodes in a network. More precisely, the fusion degree centrality of a fusion i ($\mathrm{FCD_i}$) between partner genes p and q in a binary network is defined as:

$$FCD_i = \sum_{j=1}^{n} \max\left(a_{p,j},\, a_{q,j}\right) \tag{1}$$

where n is the total number of genes in the network, and $a_{p,j}$ ($a_{q,j}$) represents the binary linkage between gene $p(q)$ and gene j in a binary network.

In a weighted network (e.g. the constructed gene network in Section 3.3), where the linkages among nodes have weights assigned to them, similarly, we define the fusion centrality of a fusion i (FCW_i) between genes p and q in terms of the total weight of the connections of gene p and q:

$$FCW_i = \frac{1}{n}\sum_{j=1}^{n} \max\left(w_{p,j},\, w_{q,j}\right) \tag{2}$$

where $w_{p,j}$ ($w_{q,j}$) represent the weighted linkage between fusion partner gene p (q) and gene j in a weighted network.

3 RESULTS

3.1 Domain-based fusion model

Based on exon shuffling theory (Gilbert, 1978), rearrangement of exons would result in novel genes that encode different proteins with different functional domains. Several studies have provided supporting evidence of a strong correlation between protein domains and exon organization (Kaessmann, 2002; Liu and Grigoriev, 2004). We hypothesized that fusion mutations in cancer are associated with exon and domain shuffling. Because most biological processes are ultimately executed by proteins, we focused on domain shuffling for in-frame cancer fusion. To test our hypothesis, we analyzed the enrichment of protein domain co-occurrence relationships and domain–domain interactions between 289 known cancer fusion gene pairs (Section 2.1).

Any two domains are considered to have a co-occurrence relationship if they co-constitute at least one protein. These relationships can represent complex domain combinations in the human proteome. A set of 5135 known protein domain co-occurrence relationships from the protein domains of 12 864 genes (Section 2.1) was used for the analysis. We found that 32.21% of the 289 cancer fusion gene pairs had known co-occurrence relationships between their domains. In contrast, only 6.66% of all the pairs from the 12 864 genes (the background ratio) were found to have co-occurrence relationships between their domains. Thus, a highly significant over-representation of domain co-occurrence was observed among the protein domains of the known cancer fusion partner genes. This finding implies that cancer fusions preferentially involve genes whose protein domains constitute functional proteins.

Several systematic studies have reported the effect of protein domain–domain interactions on protein folding kinetics and stability (Bhaskara and Srinivasan, 2011; Itoh and Sasai, 2008; Osváth *et al.*, 2005). Thus, domain–domain interactions among fusion partner genes may play roles in the synthesis of new fusion proteins. To explore this possibility, we used a set of 5410 curated domain–domain interactions from the protein domains of 14 936 genes (Section 2.1) to investigate whether there was enrichment of domain–domain interactions between the partner genes of the 289 cancer fusions. We found that 48.60% of the 289 cancer fusion gene pairs had such interactions, whereas only 13.33% of all the pairs among the 14 936 genes (the background ratio) did so. This indicates highly significant over-representation of domain–domain interactions among the protein domains of the known cancer fusion partner genes. From this, we conclude that fusions preferentially involve genes whose protein domains interact with each other.

Our results above implied that in-frame cancer fusions would result in protein domain shuffling. Several recent studies also support this view. Two studies have modeled domain rearrangement from the point of view in evolution, concluding that gene fusion is a major mechanism of multi-domain protein formation in bacteria (Pasek *et al.*, 2006) and animals (Buljan *et al.*, 2010). In addition, a very recent work analyzed 7424 chimeric RNAs from human tissues and found that these chimeras contain complete protein domains from their original genes significantly more than in random datasets (Frenkel-Morgenstern and Valencia, 2012). Accordingly, we proposed a domain-based fusion model (Fig. 1). Proteins A and B contains n and q domains, respectively, and denoted $A_1 \sim A_n$ and $B_1 \sim B_q$. The domains $A_{m+1} \sim A_n$ and $B_{p+1} \sim B_q$, which do not appear in the new fusion protein, were considered deletions, while the remaining domains $A_1 \sim A_m$ and $B_1 \sim B_p$ that constituted the new fusion protein, were considered insertions for each other. Thus, the perturbation impact of a fusion mutation between two genes in a network would be associated with all of their protein domains. Also, domains have been considered as basic functional units for biological interactions. We thus further reasoned that the potential impact of a fusion mutation in a biological network would be associated with all biological interactions of its two partner genes. Therefore, a fusion is more likely to be an oncogenic driver if its two partner genes act like hubs in a network because the fusion mutation can deregulate normal functions of many other genes and their associated pathways. This hypothesis will be tested in Section 3.2.

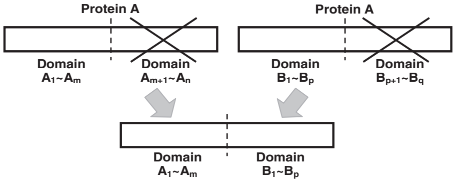

Fig. 1. Domain-based fusion model. Proteins A and B contain n and q domains, denoted $A_1 \sim A_n$ and $B_1 \sim B_q$ respectively. Dotted lines represent breakpoints and the fusion junction point of the fusion

3.2 Centrality analysis of fusion partner genes

To test whether a fusion is more likely to be an oncogenic driver if its two partner genes act like hubs in a network, we performed centrality analysis for the 289 known cancer fusion gene pairs in a curated gene network (Section 2.1). In the analysis, we used two single-gene based network centrality metrics, total degree and betweenness centrality (Supplementary Material Section S1).

We first compared the centrality distribution of all the cancer fusion partner genes with those of the 109 point mutation genes and all other genes in the curated gene network. The centrality analysis (Fig. 2), showed us that fusion partner genes and point mutation genes both tended to have much higher total degree centrality ($P = 1.44\text{e-}022$ and $P = 1.82\text{e-}025$, respectively, by the Wilcoxon signed-rank test) and higher betweenness centrality ($P = 2.21\text{e-}015$ and $P = 3.58\text{e-}019$, respectively) than all other genes in the network. We also observed the same result in the centrality analysis of a protein network (Supplementary Fig. S1). These results confirmed that cancer genes tend to be hubs in biological networks, as suggested in previous studies (Barabási *et al.*, 2011; Cui *et al.*, 2007).

However, we also found that point mutation genes had both a higher total degree centrality ($P = 1.83\text{e-}004$) and betweenness centrality ($P = 1.59\text{e-}005$) than fusion partner genes in the curated gene network (the same result was also observed in the curated protein network). We reasoned that this was due to the fact that in some fusion genes, only one of their partner genes serves as a hub gene. To test this, we compared the centrality of the two partner genes in a fusion separately. We separated the partner genes of the 289 cancer fusions into two groups, according to their total degree centrality in the network. A gene with higher total degree centrality than its fusion partner was included in the Fusion Gene1s group, whereas its fusion partner was included in the Fusion Gene2s group. The results of centrality analysis in the curated gene network are shown in Figure 3. The Fusion Gene1s group tended to have higher total degree centrality ($P = 0.0056$) and higher betweenness centrality ($P = 0.0356$) than the point mutation genes in the network. The Fusion Gene2s group tended to have a much smaller total degree centrality ($P = 2.48\text{e-}013$) and betweenness centrality ($P = 1.66\text{e-}011$) than the point mutation genes in the curated gene network. The same result was also observed in the curated

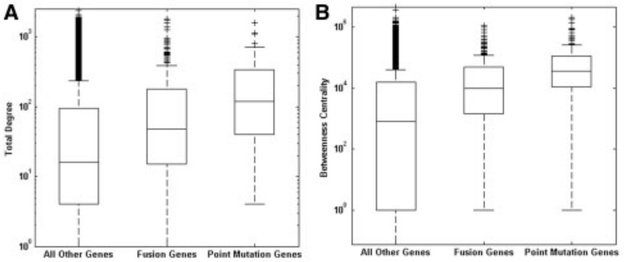

Fig. 2. Centrality distribution comparison of all fusion partner genes with point mutation genes and all other genes

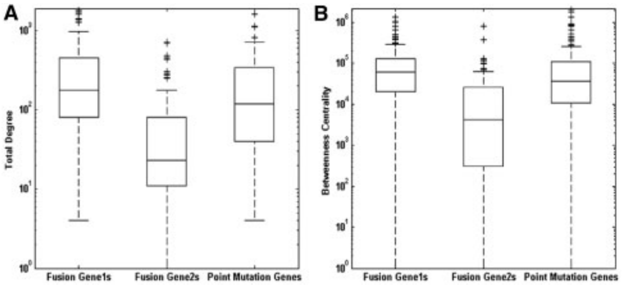

Fig. 3. Centrality distribution comparison of two fusion partner gene groups with point mutation genes

protein network (Supplementary Fig. S2). These results suggest that a fusion mutation may play an important role in oncogenesis if at least one of its two partner genes is a hub gene in a biological network. For example, a fusion is generated by a proto-oncogene and the promoter of another gene. The promoter gene may not hold high centrality, but it can activate the proto-oncogene to play an important role in cancer progression.

We further evaluated the 2D centrality distribution of the 289 fusion gene pairs using contour plots, on which the *x*- and *y*-axes, respectively, represented the centrality values of the Fusion Gene1s and Fusion Gene2s groups. Because the centrality distribution of nodes in biological networks follows the power law (Barabási and Oltvai, 2004), the density contour matrix of the centrality values was generated in log–log scale. The centrality distribution of a set of 1 000 000 randomly selected gene pairs, separated into two groups based on total degree centrality, was used to represent the background distribution. The distributions of total degree and betweenness centrality, presented in Figure 4, clearly show that cancer fusion partner genes had a significantly different centrality distribution than that of the set of randomly selected gene pairs. Again, in most of the cancer fusion gene pairs, we found that at least one partner gene was a hub, and that many cancer fusions even originated from two hub genes. We also found the same result in the curated protein network (Supplementary Figs S3 and S4).

These results support our hypothesis that a fusion is more likely to be an oncogenic driver if its two fusion partner genes are more central in a network. Several previous studies (Palanisamy *et al.*, 2010; Robinson *et al.*, 2011; Wang *et al.*, 2009) selected fusion driver candidates solely based on the criterion that one of their two partner genes is cancer-associated. Because most of cancer genes act like hub genes, the results presented above are consistent with their methods. However, these single-gene based methods may have missed some fusion drivers, as we will explain in Section 3.3.

3.3 Performance comparison of fusion centrality with other methods

Based on our proposed hypothesis and the observations aforementioned, we propose a fusion centrality metric to quantify the relative importance of a fusion in a biological network (as described in Section 2.2). Equation (1) can be used to estimate the relative importance of fusions and identify fusion drivers in a binary network, such as the curated gene networks used in the section 3.2. Because of the limited coverage of the curated gene network, which is only composed of ~25% of the genes in the genome, we were not able to evaluate the importance of many fusions using it. Therefore, we mapped partner genes of fusions to a predicted gene network that was constructed by integrating 17 heterogeneous genomic and proteomic datasets (Wu *et al.*, 2010) for the primary purpose of calculating fusion centrality. The nodes in the constructed network represented all genes of the human genome, and the functional association between any pair of them was weighted by a linkage probability that can reveal the tendency of genes to operate in the same or related pathway. This network not only contained direct molecular interaction information but also broader gene-gene functional relationships within pathways. Thus, the constructed network provided more coverage for prediction. Because the constructed gene network is a weighted network, we applied Equation (2) to quantify centrality of fusion mutations.

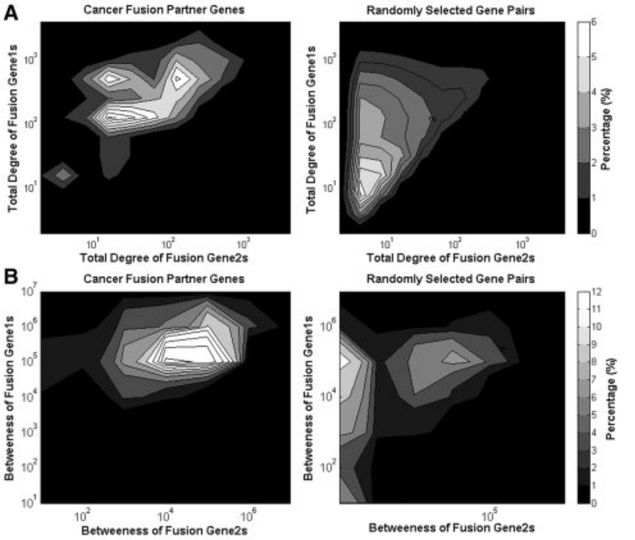

Fig. 4. 2D distribution of degree centrality (**A**) and betweenness centrality (**B**) of two partner gene groups

Next, we compared the performance of the proposed fusion centrality with other single gene-based methods using receiver operating characteristic (ROC) curves. As a positive benchmark set, we used the 289 partner gene pairs of known cancer fusions, and as a negative benchmark, we used a set composed of 1 000 000 randomly selected human gene pairs. In the first method, denoted M1, the proposed fusion centrality is applied to quantify importance of potential fusions generated by these benchmark gene pairs in the constructed gene network. The second method, denoted M2, is used to mimic the situation of those previous studies (Palanisamy *et al.*, 2010; Robinson *et al.*, 2011) that typically selected fusion drivers, based solely on the criterion that one of the two partner genes was a cancer-associated gene (i.e. most of cancer genes act like hub genes). We calculated the single gene-based degree centrality (Barrat *et al.*, 2004) for each partner gene in a benchmark gene pair and used the maximum value of their centrality to represent the importance of a potential fusion generated by the benchmark gene pair.

We also applied the ConSig method (Wang *et al.*, 2009), yet another single-gene based network method, denoted M3. ConSig score was calculated for each gene by considering its association with the identified biologically important concepts. A fusion is nominated as a driver if one of its partner genes has a high ConSig score. Therefore, in this method, we used the maximum value from the ConSig scores of a benchmark gene pair to represent the importance of a fusion generated by the pair. ConSig scores of all human genes were downloaded from (http://consig.cagenome.org/).

The ROC curves and area under curve values for all methods are shown in Figure 5. A higher area under curve value in a method meant that fusions generated by the positive gene pairs tended to have higher fusion centrality than those generated by the negative gene pairs. These results indicate that M1 outperformed the other two methods. As comparing M1 and M2, it seems that M2 has a comparable performance as M1. This indicates single gene-based method can provide moderately good prediction owing to the fact that at least one of partner genes of cancer fusion is a hub gene. However, a high true-positive rate is preferred to the fixed lower false-positive rate in a screening. In the partial ROC area with high true-positive rate and low FDR, we found that M1 significantly outperformed M2. Nevertheless, this result further confirms our hypothesis that the importance of a fusion in a network should be evaluated by considering both of its partner genes. Moreover, M2 performed better than M3. Both M2 and M3 are single-gene based models but were generated using data of the constructed gene network and ConSig scores, respectively. This implies that the constructed network (Wu *et al.*, 2010) contained more critical pathway information than the concept signature method (Wang *et al.*, 2009).

3.4 Application of newly discovered fusion drivers

To further evaluate the performance of our fusion centrality method, we applied Equation (2) to newly discovered fusions. We selected 38 such fusions, 33 of which had validated oncogenetic importance and five of which were highly recurrent in cancer samples. To identify potential fusion drivers, we first set a threshold value on fusion centrality using the set of 1 000 000 randomly selected gene pairs as a control. We found that the fusion centralities of the top 7% of these 1 000 000 pairs were >0.370, whereas the fusion centralities of 73% of the 289 known cancer fusions were >0.370 (Table 1). We therefore set the threshold value to 0.370 for identifying potential fusion drivers in the proposed fusion centrality method. Table 2 only lists the fusion centrality values of 10 of these 38 new fusions (the full information of the 38 fusions is shown in Supplementary Table S1). Of these, 34 fusions had a weighed fusion centrality value larger than or very closer to the threshold value. In all, 29 of these 34 fusions have been confirmed to have oncogenic importance experimentally. The other four of these new fusions, JAZF1-SUZ12, CRTC1-MAML1, R3HDM2-NFE2 and ALG5-PIGU, had fusion centrality values that were much smaller than the threshold value. Among them, JAZF1-SUZ12 and CRTC1-MAML1 have been associated with lower grade cancer subtypes (Lee *et al.*, 2012). We thus suspect that R3HDM2-NFE2 and ALG5-PIGU may also be associated with lower grade cancer subtypes. These results indicate the ability of the proposed fusion centrality method to identify fusion drivers.

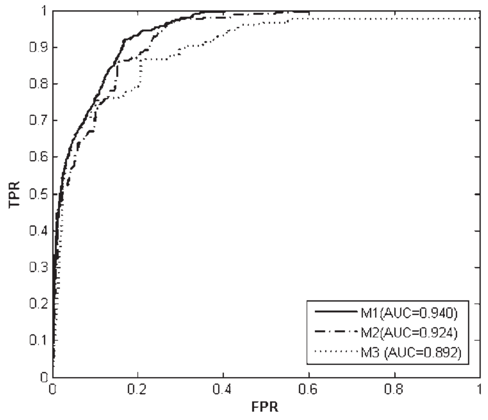

Fig. 5. Prediction performance comparison of fusion centrality (M1) and other methods using ROC curve

Notably, our fusion centrality approach also predicted several fusions with partner genes whose biological roles in oncogenesis are not well-known. For example, little is known about biological roles of microtubule-associated serine/threonine kinase (MAST) proteins, and their somatic alterations just recently were discovered in cancer (Robinson *et al.*, 2011). Expression of these MAST fusions has shown a proliferative effect, both *in vitro* and *in vivo* (Robinson *et al.*, 2011). Our approach

Table 1. Distribution of Fusion centralities of 289 known cancer fusions and 1 000 000 randomly selected gene pairs

Fusion centrality	289 Known cancer fusions	Randomly selected gene pairs
>0.380	48%	1%
>0.370	73%	7%
>0.360	95%	19%

Table 2. Fusion centralities of 10 new fusion drivers

Reference	Fusions	Cancer type	Fusion centrality
Robinson *et al.*, 2011	MAST2-ARID1A	BC	0.386
	MAST2-GPBP1L1		0.383
	MAST1-TADA2L		0.38
	MAST1-NFIX		0.38
Edgren *et al.*, 2011	VAPB -IKZF3	BC	0.376
Lee *et al.*, 2012	YWHAE -FAM22A	HGESS	0.375
	YWHAE -FAM22B		0.368
Lee *et al.*, 2012	JAZF1-SUZ12	LGESS	0.353
Singh *et al.*, 2012	FGFR1-TACC1	GBM	0.385
	FGFR3-TACC3		0.381

BC, breast cancer; GBM, glioblastoma; HGESS, high-grade endometrial stromal sarcoma; LGESS, low-grade endometrial stromal sarcoma.

identified all MAST fusions as driver fusions (i.e. all MAST fusions with >0.38 fusion centralities). Another is the fusion of vesicle-associated membrane protein (VAPB) with IKZF3 (IKAROS family zinc finger 3) in breast cancer, whose fusion centrality in our analysis was 0.376. RNA interference-mediated knock-down of the VAPB-IKZF3 fusion gene revealed its importance to the growth and survival of breast cancer cells (Edgren *et al.*, 2011).

In addition, Lee *et al.* (2012) identified fusions between 14-3-3ε (YWHAE) and either of two nearly identical FAM22 family members (FAM22A and FAM22B) in endometrial stromal sarcoma. In contrast to classic endometrial stromal sarcoma that harbors JAZF1-SUZ12 genetic fusions, tumor cells with YWHAE–FAM22 fusions displayed high-grade histologic features, including larger and more irregular nuclei, increased mitotic activity, a distinct gene-expression profile, and a more aggressive clinical course (Lee *et al.*, 2012). Interestingly, fusions YWHAE–FAM22A and YWHAE–FAM22B have fusion centralities of 0.375 and 0.368, respectively, in our predictions, whereas JAZF1–SUZ12 has a much lower fusion centrality, 0.353. This result suggests that our approach can predict the relative importance of fusion mutations well.

4 DISCUSSION

In this work, we have proposed the concept of fusion degree centrality, which measures the local impact of a fusion gene through its direct connections in a network. The fusion degree centrality, however, is somewhat biased toward those genes that were studied more in literature. Therefore, false negatives could be an issue because some of the genes important for cancer progression may not be predicted by the method because little is known about their function, and there is little connectivity between these genes and others in the network. For example, some bottleneck hub genes (Yu *et al.*, 2007) with only a few direct connections to other nodes, but that act as key connectors in the network. This problem can be addressed by developing of other kinds of fusion centrality metrics that are able to measure the global network impact of a fusion, such as fusion betweenness centrality or fusion eigenvector centrality, using the same concept we proposed here. For example, after creating a fused node of two nodes, the fusion betweenness centrality of a potential fusion generated by these two nodes in a network is represented by the betweenness centrality of the fused node, which can be calculated using several existing algorithms. However, in these algorithms, the calculations of the shortest paths between all pairs of nodes in a network are time consuming and have to occur once for each fused node when applying the algorithms to multiple fusions because a fusion between two nodes in a network may shorten paths between other node pairs. Therefore, faster algorithms are needed to efficiently calculate these fusion centrality measurements for a large number of fusions detected in next-generation sequencing.

The locations of the fusion or breakpoints and the assessment of the directionality to the gene components allowed us to categorize different types of fusions, such as in-frame fusions and out-of-frame fusions. Although the proposed domain-based fusion model was developed based on in-frame fusions, the proposed fusion centrality method can also be applied to out-of-frame fusions because an out-of-frame fusion mutation can be considered the deletion of its two partner genes. More specifically, if the reading frame is not retained across the fusion junction, it will likely lead to the production of a premature stop codon and degradation of the transcript. An out-of-frame fusion can silence one allele of its two partner genes and be associated with haploinsufficiency or loss of heterozygosity. Therefore, the potential impact of an out-of-frame fusion in a network would also be associated with all biological interactions of its two partner genes if the other alleles of the partner genes are lost as well.

A number of studies have identified that different fusion mutations can be generated when the partner genes are identical but the breakpoints different (Baxter *et al.*, 2002; Hernández *et al.*, 2002; Robinson *et al.*, 2011). For example, genomic analysis of three TFG-ALK rearrangements in anaplastic large cell lymphomas revealed that the TFG breakpoints occurred at introns 3, 4 and 5, respectively, whereas the ALK breakpoints always occurred in the same intron (Hernández *et al.*, 2002). Based on our proposed approach, we suggest that all three fusions would have the same level of potential impact in a network, no matter where their breakpoints are located. Interestingly, the study of Hernández *et al.* (2002) revealed that TFG may use a variety of intronic breakpoints in ALK rearrangements, generating fusion proteins with different molecular weights but similar transforming potential. This finding seems to support our view. However, we need to emphasize that our fusion centrality method is used to roughly estimate the tendency of a fusion mutation to be an oncogenic driver in a network, not to estimate its real impact on cellular function.

A critical issue in detecting fusions from sequencing data is the high false-detection rate. False detections may be caused by experimental artifacts, genome assembly errors or the mis-mapping of end sequences (Bashir *et al.*, 2008; Shah *et al.*, 2009; Wang *et al.*, 2009). To date, most of the bioinformatics methods developed for fusion discovery in sequencing data use multiple steps of false-positive filtering (Kim and Salzberg, 2011; McPherson *et al.*, 2011a, b; Sboner *et al.*, 2010). Although applying these filters may reduce the false-positive rate, the false-negative rate would increase, and consequently, some fusion drivers would not be detected (McPherson *et al.*, 2011a). Our fusion centrality method is an alternative way to reduce the false-negative rate when detecting fusion drivers. For example, potential fusion drivers that fail to pass filters but have high predicted fusion centrality may be kept for further investigation.

Although the gene functional network used in this work was constructed without accounting for tissue specificity, the results presented in this work suggest that a single gene functional network can be used to identify fusion mutation drivers in diverse types of cancer. Several studies also successfully applied a single gene functional network to predicted tissue-specific disease genes and phenotypic effects of gene perturbation (Lee *et al.*, 2008; Torkamani and Schork, 2009; Wu *et al.*, 2012). Besides, fusion drivers of a specific type of cancer may also be identified by associating their partner genes with known cancer genes or pathways in the cancer type. For example, by extending the concepts in Equations (1) and (2), the importance of a fusion in a network can be calculated as the sum of the direct functional associations of its two partner genes with those known cancer genes or pathways.

We also applied the centrality analysis to over 1000 cancer chimeric transcripts and fusion genes collected in ChiTaRS database (Frenkel-Morgenstern *et al.*, 2013), and conclude the same result as that of using the original 289 fusion gene pairs set (Supplementary Material Section S4). Besides, we also used thirty-two highly recurrent chimeric RNAs that were detected in 20 human prostate cancer samples and 10 matched benign prostate tissues (Kannan *et al.*, 2011) to validate the performance of our fusion centrality approach. Many of these chimerics RNA also appeared in matched benign tissues owing to a 'field effect' within the histologically normal epithelium. They non-parametric Kolmogorov–Smirnov test was used to determine which chimeric RNAs are significantly enriched in cancer tissues. We hypothesized that some of enriched chimeric RNAs would be fusion drivers and thus have high fusion centrality in our prediction. Indeed, we found most of chimeric RNAs that are significantly enriched in human prostate cancer ($P < 0.05$) tended to have higher fusion centrality than those insignificantly enriched chimeric RNAs. These results are detailed in the Section S5 of the Supplementary Material.

Finally, it is important to note that the fusion centrality method was constructed based on the available data of biological networks that are far from complete. Genes that have not been characterized to date but may be important in cancer progression will not be identified by our fusion centrality approach because little of their functions are known. This limitation applies to all network-based approaches (Barabási *et al.*, 2011), but our method will be improved overtime, as more data are generated and are integrated to construct a more complete gene network in the future. Nevertheless, our results suggest that the method of fusion centrality is an efficient method for identifying novel fusion drivers.

5 CONCLUSION

In summary, all results in this work indicate that the fusion centrality method is a systematic and effective method for prioritizing fusion drivers from hundreds or thousands of fusion candidates identified in diverse cancer types. To our best knowledge, this is the first network-based approaches to nominate fusion drivers.

Funding: This research was supported in part by a grant from The Weathervane Foundation to A.M. L.Y. is supported by DOD W81XWH-10-10327, OCRF PPD/BCM/01.12, NIH R01EB013584. K.K. is supported by PCF Young Investigator award.

Conflict of Interest: none declared.

REFERENCES

Aurias,A. *et al.* (1984) Translocation involving chromosome 22 in Ewing's sarcoma: a cytogenetic study of four fresh tumors. *Cancer Genet. Cytogenet.*, **12**, 21–25.

Barabási,A.L. and Oltvai,Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.*, **5**, 101–113.

Barabási,A.L. *et al.* (2011) Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.*, **12**, 56–68.

Barrat,A. *et al.* (2004) The architecture of complex weighted networks. *Proc. Natl Acad. Sci. USA*, **101**, 3747–3752.

Bashir,A. *et al.* (2008) Evaluation of paired-end sequencing strategies for detection of genome rearrangements in cancer. *PLoS Comput. Biol.*, **4**, e1000051.

Baxter,E.J. *et al.* (2002) The t(4;22)(q12;q11) in atypical chronic myeloid leukaemia fuses BCR to PDGFRA. *Hum. Mol. Genet.*, **11**, 1391–1397.

Bhaskara,R.M. and Srinivasan,N. (2011) Stability of domain structures in multi-domain proteins. *Sci. Rep.*, **1**, 40.

Buljan,M. *et al.* (2010) Quantifying the mechanisms of domain gain in animal proteins. *Genome Biol.*, **11**, R74.

Cerami,E.G. *et al.* (2010) Pathway commons, a web resource for biological pathway data. *Nucleic Acids Res.*, **39**, D685–D690.

Cui,Q. *et al.* (2007) A map of human cancer signaling. *Mol. Syst. Biol.*, **3**, 152.

Ding,L. *et al.* (2008) Somatic mutations affect key pathways in lung adenocarcinoma. *Nature*, **455**, 1069–1075.

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

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.* (2013) ChiTaRS: a database of human, mouse and fruit fly chimeric transcripts and RNA-sequencing data. *Nucleic Acids Res.*, **41**, D142–D151.

Fujita,P.A. *et al.* (2011) The UCSC Genome Browser database: update 2011. *Nucleic Acids Res.*, **39**, D876–D882.

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

Gilbert,W. (1978) Why genes in pieces? *Nature*, **271**, 501.

Hernández,L. *et al.* (2002) Diversity of genomic breakpoints in TFG-ALK translocations in anaplastic large cell lymphomas: identification of a new TFG-ALK(XL) chimeric gene with transforming activity. *Am. J. Pathol.*, **160**, 1487–1494.

Höglund,M. *et al.* (2006) A gene fusion network in human neoplasia. *Oncogene*, **25**, 2674–2678.

Itoh,K. and Sasai,M. (2008) Cooperativity, connectivity, and folding pathways of multidomain proteins. *Proc. Natl Acad. Sci. USA*, **105**, 13865–13870.

Jones,S. *et al.* (2008) Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science*, **321**, 1801–1806.

Kaessmann,H. *et al.* (2002) Signatures of domain shuffling in the human genome. *Genome Res.*, **12**, 1642–1650.

Kannan,K. *et al.* (2011) Recurrent chimeric RNAs enriched in human prostate cancer identified by deep sequencing. *Proc. Natl Acad. Sci. USA*, **108**, 9172–9177.

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

Lee,C.H. *et al.* (2012) 14-3-3 fusion oncogenes in high-grade endometrial stromal sarcoma. *Proc. Natl Acad. Sci. USA*, **109**, 929–934.

Lee,I. *et al.* (2008) A single gene network accurately predicts phenotypic effects of gene perturbation in *Caenorhabditis elegans*. *Nat. Genet.*, **40**, 181–188.

Liu,M. and Grigoriev,A. (2004) Protein domains correlate strongly with exons in multiple eukaryotic genomes–evidence of exon shuffling? *Trends Genet.*, **20**, 399–403.

McPherson,A. *et al.* (2011a) Comrad: detection of expressed rearrangements by integrated analysis of RNA-Seq and low coverage genome sequence data. *Bioinformatics*, **27**, 1481–1488.

McPherson,A. *et al.* (2011b) deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. *PLoS Comput. Biol.*, **7**, e1001138.

Osváth,S. *et al.* (2005) Asymmetric effect of domain interactions on the kinetics of folding in yeast phosphoglycerate kinase. *Protein Sci.*, **14**, 1609–1616.

Palanisamy,N. *et al.* (2010) Rearrangements of the RAF kinase pathway in prostate cancer, gastric cancer and melanoma. *Nat. Med.*, **16**, 793–798.

Pasek,S. *et al.* (2006) Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins. *Bioinformatics*, **22**, 1418–1423.

Robinson,D.R. *et al.* (2011) Functionally recurrent rearrangements of the MAST kinase and Notch gene families in breast cancer. *Nat. Med.*, **17**, 1646–1651.

Rowley,J.D. (1973) Letter: a new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. *Nature*, **243**, 290–293.

Sboner,A. *et al.* (2010) FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data. *Genome Biol.*, **11**, R104.

Shah,S.P. *et al.* (2009) Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature*, **461**, 809–813.

Singh,D. *et al.* (2012) Transforming fusions of FGFR and TACC genes in human glioblastoma. *Science*, **337**, 1231–1235.

Stephens,P.J. *et al*. (2009) Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature*, **462**, 1005–1010.

Tomlins,S.A. *et al*. (2005) Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science*, **310**, 644–648.

Torkamani,A. and Schork,N.J. (2009) Identification of rare cancer driver mutations by network reconstruction. *Genome Res*., **19**, 1570–1578.

Vandin,F. *et al*. (2011) Algorithms for detecting significantly mutated pathways in cancer. *J*. *Comput*. *Biol*., **18**, 507–522.

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.

Wu,C.C. *et al*. (2010) Prediction of human functional genetic networks from heterogeneous data using RVM-based ensemble learning. *Bioinformatics*, **26**, 807–813.

Wu,C.C. *et al*. (2012) TARGETgene: a tool for identification of potential therapeutic targets in cancer. *PLoS One*, **7**, e43305.

Yellaboina,S. *et al*. (2011) DOMINE: a comprehensive collection of known and predicted domain-domain interactions. *Nucleic Acids Res*., **39**, D730–D735.

Yu,H. *et al*. (2007) The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput*. *Biol*., **3**, e59.