IIIDB: a database for isoform-isoform interactions and isoform network modules

Yu-Ting Tseng1, Wenyuan Li2, Ching-Hsien Chen3, Shihua Zhang4, Jeremy JW Chen5,6,7, Xianghong Jasmine Zhou2, Chun-Chi Liu1,5,7

From The Thirteenth Asia Pacific Bioinformatics Conference (APBC 2015)
HsinChu, Taiwan. 21-23 January 2015

Abstract

Background: Protein-protein interactions (PPIs) are key to understanding diverse cellular processes and disease mechanisms. However, current PPI databases only provide low-resolution knowledge of PPIs, in the sense that “proteins” of currently known PPIs generally refer to “genes.” It is known that alternative splicing often impacts PPI by either directly affecting protein interacting domains, or by indirectly impacting other domains, which, in turn, impacts the PPI binding. Thus, proteins translated from different isoforms of the same gene can have different interaction partners.

Results: Due to the limitations of current experimental capacities, little data is available for PPIs at the resolution of isoforms, although such high-resolution data is crucial to map pathways and to understand protein functions. In fact, alternative splicing can often change the internal structure of a pathway by rearranging specific PPIs. To fill the gap, we systematically predicted genome-wide isoform-isoform interactions (IIIs) using RNA-seq datasets, domain-domain interaction and PPIs. Furthermore, we constructed an III database (IIIDB) that is a resource for studying PPIs at isoform resolution. To discover functional modules in the III network, we performed III network clustering, and then obtained 1025 isoform modules. To evaluate the module functionality, we performed the GO/pathway enrichment analysis for each isoform module.

Conclusions: The IIIDB provides predictions of human protein-protein interactions at the high resolution of transcript isoforms that can facilitate detailed understanding of protein functions and biological pathways. The web interface allows users to search for IIIs or III network modules. The IIIDB is freely available at http://syslab.nchu.edu.tw/IIIDB.

Background

Protein-protein interactions (PPIs) perform and regulate fundamental cellular processes. As a consequence, identifying interacting partners for a protein is essential to understand its functions. In recent years, remarkable progress has been made in the annotation of all functional interactions among proteins in the cell. However, in both experimentally derived and computationally predicted protein-protein interactions, a “protein” generally refers to “all isoforms of the respective gene.” Yet, it is known that alternative splicing often impacts PPI by either directly affecting protein interacting domains, or by indirectly impacting other domains, which, in turn, impact the PPI binding [1]. That is, alternative splicing can modulate the PPIs by altering the protein structures and the domain compositions, leading to the gain or loss of specific molecular interactions that could be key links of pathways (reviewed in reference [2]). It is very likely that different isoforms of the same protein interact with different proteins, thus exerting different functional roles. For example, the protein BCL2L1 is alternatively spliced into two isoforms: Bcl-xL (long form) and Bcl-xS (short form) [3], in which Bcl-xL inhibits apoptosis whereas Bcl-xS promotes apoptosis [4]. Vogler et al. reported that
the interaction of Bcl-xL and BAK1 in platelets ensures platelet survival [5]. Therefore, comprehensively identifying protein-protein interactions at the isoform level is important to systematically dissect cellular roles of proteins, to elucidate the exact composition of protein complexes, and to gain insights into metabolic pathways and a wide range of direct and indirect regulatory interactions.

Thus far, a series of studies have systematically predicted PPIs [6-10] and established PPI databases, e.g., OPID [11], POINT [12], STRING [10] and PIPs [7]. With the exception that the IntAct database [13] contains 116 human PPIs with isoform specification, currently, none of those PPI databases has isoform-level PPI data. This is a huge knowledge gap yet to be filled. The rapid accumulation of RNA-seq data provides unprecedented opportunities to study the structures and topological dynamics of PPI networks at the isoform resolution. RNA-seq data provides two unique informative sources for Isoform-Isoform Interaction (III) reconstruction: the absence or presence of specific isoforms under specific conditions, and the co-expression of two isoforms that may contribute to their interaction propensity. In this study, we seize this opportunity to comprehensively predict the possible interactions between splicing isoforms by integrating a series of RNA-seq data with domain-domain interaction data and PPI database. The resulting III network presents a high-resolution map of PPIs, which could be invaluable in studying biological processes and understanding cellular functions.

In this report, we described a database, IIIDB, for accessing and managing predicted human IIIs. In the IIIDB, users can differentiate access high-confidence and low-confidence predictions of human IIIs (see detailed description in Result section), and then see the full evidence values for each predicted III. Figure 1 shows the IIIDB web interface screen-shots of the III search and isoform module search function in the IIIDB. Users can

Figure 1 The screenshots of the isoform interaction and module search function in the IIIDB. (A) Users can upload their own gene expression data for III prediction. (B) IIIDB shows top 100 IIIs and uses these interactions to construct an III network. Users can download complete prediction result as a file. (C) In interaction search function, users can input a gene symbol or gene ID to search associated IIIs and isoform modules. The interaction search section provides interface on searching high-confidence (score > 2.575) or low-confidence (score > 1.692) prediction. The resulting III table for interaction search shows full evidence values including Pearson correlations for 19 RNA-seq datasets and domain interaction score. (D) The resulting page for module search includes the user friendly network graph and pathway/GO enrichment analysis result.
upload their own gene expression data for III prediction, and then the users can download the predicted result. The searching function has three major parts: high-confidence interaction prediction search, low-confidence interaction prediction search, and isoform module search. The IIIDB provides auto-complete function in all search functions. Users can input a gene symbol or gene ID in the auto-complete field which provides an interface to quickly find and select matched values.

The IIIDB allows users to easily search IIIs and isoform modules, and then provides the evidence that led to each III prediction. To visualize the interactions with the isoforms of the input gene, we integrated CytoscapeWeb [14] to generate the interactive web-based III network (Figure 1B). Interestingly, the different isoforms within the same gene can be involved with different isoform modules that may open a new door to study differential functionality of isoforms of the gene. The IIIDB also provided GO/pathway enrichment analysis results for each isoform module, which helps biologists to study the biological insights of network modules at isoform level.

**Results**

To obtain the isoform annotations in the human genome, we used the NCBI Reference Sequences (RefSeq) mRNAs as transcriptome annotation [15]. Figure 2 shows the framework of III prediction. We employed the logistic regression approach with 19 RNA-seq datasets (Table 1) and the domain-domain interaction database to infer IIIs. To confirm with PPI, given an III prediction I₁ and I₂, we only keep this prediction if the gene symbols of I₁ and I₂ have PPI in IntAct database.

**High-confidence and low-confidence prediction of IIIs**

In the IIIDB, we provided two III prediction sets using the logistic regression model: (a) high-confidence prediction: logit score > 2.575 (precision 60% and recall 15%), it resulted in 4,476 IIIs; and (b) low-confidence prediction: logit score > 1.692 (precision 40% and recall 68%). In addition, given a known PPI, the isoform pair with the best logit score among this PPI will be selected as low-confidence prediction. Thus, a PPI has at least one isoform-isoform interaction. It resulted in 54,605 IIIs (Figure 3).

**Isoform module discovery**

To discover functional modules in the III network, we applied MODES network clustering method [16] on low-confidence III network to discover isoform modules with the given parameters (minimum module size 3, maximum module size 30, and density cutoff 0.7). An important feature of MODES is that it can discover overlapping dense isoform modules which allows one isoform to belong to multiple modules. We obtained 1025 modules with size of 5.08 isoforms on average. To provide functional annotation and evaluate the module functional enrichment, we performed the enrichment analyses with GO [17] and KEGG pathway [18] databases. These databases are protein-level annotation which provides approximately functional annotation of isoforms.

To evaluate the significance, we randomly generated the same number and the same size of MODES isoform modules (i.e. 1025 randomized isoform modules). Table 2 shows the results of enrichment analyses for MODES and randomization modules for comparison. The MODES isoform modules have significantly higher functional enrichment rate than those of random cases, showing strong biological relevance of the predicted modules. The MODES isoform modules were further used to build the isoform module database in the IIIDB, and all isoform modules were listed in Additional file 1. We also stored the GO and KEGG enrichment results.
for all isoform modules in the IIIIDB to provide potential functional annotation. In the IIIIDB web interface, Figure 1C shows the result page of the isoform module search including the network graph and the enrichment analysis results. In the module result page, users can click any gene symbol to iteratively search isoform modules, and sort the module result table by clicking the column header.

**Integrating diverse data sources for III prediction**
Currently, most PPI databases do not provide information at the level of isoforms, which thus presents challenges for constructing a gold standard positive set (GSP) for the III prediction. Fortunately, in the June 2013 version of the IntAct database http://www.ebi.ac.uk/intact/ [13], we identified 116 human PPIs with isoform specification (out of the total 43,508 distinct human PPIs). For example, IntAct has III between P29590-5 and P03243-1, which correspond to the 5th isoform of the protein P29590 and the 1st isoform of P03243. In addition, to obtain more IIIs for the GSP, we applied the following rule: given a PPI between protein P1 and P2, if both P1 and P2 only have single isoforms, we take it as the GSP. It resulted in 11,356 IIIs in the GSP set.

GSP covered 5,503 RefSeq IDs, and we used these RefSeq IDs to construct gold standard negative set (GSN). The GSN was defined as isoform pairs in which one isoform was assigned to the plasma membrane cellular component, and the other was assigned to the nuclear cellular component by the isoform-specific sub-cellular localization, in which we performed sequence-based predictions using the CELLO (subCELlular LOcalization predictor) [19]. To obtain the accurate isoform-specific annotation, we only used the cellular localization prediction results that consist of UniProt GO annotations.

### Table 1 19 RNA-seq datasets from SRA

| ID   | SRA ID       | # Exp | Title                                                                 |
|------|--------------|-------|-----------------------------------------------------------------------|
| d1   | ERP000546    | 48    | Illumina bodyMap2 transcriptome                                       |
| d2   | SRP005169    | 41    | Widespread splicing changes in human brain development and aging       |
| d3   | SRP005408    | 31    | Gene expression profile in postmortem hippocampus using RNAseq for addicted human samples |
| d4   | SRP010280    | 31    | Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins |
| d5   | SRP002638    | 30    | Comparative transcriptomic analysis of prostate cancer and matched normal tissue using RNA-seq |
| d6   | ERP000550    | 29    | Complete transcriptomic landscape of prostate cancer in the Chinese population using RNA-seq |
| d7   | SRP005242    | 21    | A Comparison of Single Molecule and Amplification Based Sequencing of Cancer Transcriptomes: RNA-Seq Comparison |
| d8   | SRP002079    | 20    | GSE20301: Dynamic transcriptomes during neural differentiation of human embryonic stem cells |
| d9   | ERP000992    | 18    | The effect of estrogen and progesterone and their antagonists in Ishikawa cell line compared to MCF7 and T47D cells |
| d10  | SRP007272    | 16    | Alternative Isoform Regulation in Human Tissue Transcriptomes          |
| d11  | SRP007338    | 16    | GSE30017: Widespread regulated alternative splicing of single codons accelerates proteome evolution |
| d12  | SRP010166    | 16    | GSE34914: Deep Sequence Analysis of non-small cell lung cancer: Integrated analysis of gene expression, alternative splicing, and single nucleotide variations in lung adenocarcinomas with and without oncogenic KRAS mutations |
| d13  | ERP000710    | 12    | Transcriptome profiling of ovarian cancer cell lines                  |
| d14  | SRP005411    | 11    | RNA-Seq Quantification of the Complete Transcriptome of Genes Expressed in the Small Airway Epithelium of Nonsmokers and Smokers |
| d15  | SRP006731    | 11    | GSE29155: RNA-Seq analysis of prostate cancer cell lines using Next Generation Sequencing |
| d16  | SRP013224    | 11    | GSE38006: Next-generation sequencing reveals HIV-1-mediated suppression of T cell activation and RNA processing and the regulation of non-coding RNA expression in a CD4+ T cell line |
| d17  | ERP000418    | 10    | Gene expression profiles between normal and breast tumor genomes       |
| d18  | ERP000573    | 10    | RNA and chromatin structure                                           |
| d19  | SRP010483    | 10    | GSE35296: The human pancreatic islet transcriptome: impact of pro-inflammatory cytokines |

**Figure 3** Precision and recall curve for logistic regression model. At recall 15%, logistic regression model achieve 60% precision (High-confidence prediction); at recall 68%, logistic regression model achieve 40% precision (Low-confidence prediction).
It resulted in 36 RefSeq IDs for plasma membrane cellular component and 739 RefSeq IDs for nuclear cellular component. In addition, isoforms that are assigned to both the plasma membrane and the nuclear cellular component are excluded in GSN.

To calculate precision and recall, we used timestamp to divide GSP into training and test GSP sets, in which if an interaction with timestamp after 1st Jan 2012, it will be assigned to test GSP set (10,408 IIIs); otherwise, it will be assigned to training GSP set (948 IIIs). When the GSP is decided, we used the RefSeq IDs covered in GSP to build GSN. Figure 3 shows the precision and recall curve for the logistic regression model.

Case studies

To demonstrate the biological importance of the IIIDB, we searched for isoform-associated reports in the literature. Although isoform-specific protein function studies are very rare, we found isoform-specific biological evidences with BCL2L1, which validated our III prediction of BCL2L1. In addition, we also found diverse biological functions with Ras association domain family in our isoform modules.

**BCL2L1**

BCL2L1 has two isoforms (Table 3), in which BCL2L1 isoform 1 is called Bcl-xL (long form) and BCL2L1 isoform 2 is called Bcl-xS (short form) [3]. These isoforms play important roles in apoptosis as follows: Bcl-xL inhibits apoptosis whereas Bcl-xS promotes apoptosis [4]. In our high-confidence prediction, BCL2L1 isoform 1 (Bcl-xL) interacts with BAK1, BAX and NLRP1 to inhibit apoptosis, but BCL2L1 isoform 2 (Bcl-xS) doesn’t (Table 3). In previous reports, the fluorescence anisotropy, analytical ultracentrifugation, and NMR assays confirmed a direct interaction between Bcl-xL and BAK1 [5,20]. Vogler et al. also reported that the interaction of Bcl-xL and BAK1 in platelets ensures cell survival [5]. Edlich et al. reported that an interaction between Bcl-xL and BAX not only inhibits BAX activity but also maintains BAX in the cytosol [21]. On the other hand, Chang et al. demonstrated that Bcl-xL interacted with endogenous BAX in 293 cells. However, no significant amount of BAX was detectable in the Bcl-xS immunoprecipitation [21], suggesting that Bcl-xS does not interact with BAX. In addition, Bruey et al. reported that Bcl-xL interacts with NALP1 to suppress apoptosis [22]. Thus, these previous biological studies validated our III prediction of BCL2L1.

**RASSF1**

RASSF1 is RAS association domain family member 1. RASSF1 has four isoforms, of which two isoforms belong to isoform modules (Mod-162 and Mod-384). The Mod-162 consists of RASSF1 isoform A, RASSF5 and STK4, whereas the Mod-384 consists of RASSF1 isoform C, RASSF3, RASSF4, SAV1, STK3 and STK4 (Figure 4). Interestingly, several studies reported that RASSF5 may interact with RASSF1 isoform A to suppress tumors [23-25], suggesting that the Mod-162 is validated. On the other hand, RASSF1 isoform C may play a completely different role as an oncogene in high-grade tumors [26]. Although the function of RASSF1 isoform C is still not clear, the RASSF1 isoform A and C should have distinct functions. The IIIDB assigned the RASSF1 isoform A and C into the Mod-162 and the Mod-384, respectively, suggesting a new hypothesis of the isoform-level modules.

**Methods**

**Isoform coexpression network construction**

To construct the Isoform coexpression networks, we selected 19 RNA-seq datasets with at least 10 experiments from Sequence Read Archive (SRA) database (Table 1) [27]. We performed the eXpress [28] with Bowtie2 aligner [29] to obtain isoform expression values. NCBI Reference Sequences (RefSeq) mRNAs with protein sequences was used as transcriptome annotation, which included 31,454 RefSeq IDs (Jan 2013 version) [15].

**Logistic regression model**

The logistic regression model has been applied to PPI prediction [30,31], as it is suitable to describe the relationship between a binary response variable and a set of

---

**Table 2 The enrichment rate of isoform modules based on GO and pathway enrichments**

| Isoform modules | # modules | % modules enriched with GOa | % modules enriched with pathwayb |
|-----------------|-----------|----------------------------|---------------------------------|
| MODES modules   | 1025      | 88.7%                      | 36.1%                           |
| Randomization   | 1025      | 49.7%                      | 10.1%                           |

a The modules enriched with GO term (P-value < 0.001).

b The modules enriched with pathway (P-value < 0.01).

**Table 3 BCL2L1 isoform interaction partners.**

| Isoform         | RefSeq ID   | mRNA length | Protein length | UniProt ID  | Interaction partner (high confidence) |
|-----------------|-------------|-------------|----------------|-------------|--------------------------------------|
| BCL2L1 isoform 1 (Bcl-xL) | NM_138578   | 2575        | 233            | Q07817-1    | BAK1, BAX, NLRP1                      |
| BCL2L1 isoform 2 (Bcl-xS)  | NM_001191   | 2386        | 170            | Q07817-2    | BCL2                                 |
We used logistic regression approach to build the prediction model as follows:

\[
\text{logit}(y_{ij}) = \alpha_0 + \alpha_1 E_{1ij} + \alpha_2 E_{2ij} + \cdots + \alpha_{19} E_{19ij} + \alpha_{20} \text{DDI}_{ij}
\]

where \(\alpha_0, \alpha_1, \ldots, \alpha_{20}\) are regression coefficients and \(y_{ij}\) is the probability of the isoform interaction between isoform \(i\) and isoform \(j\). DDI, \(E_1\), \(E_2\), \ldots, \(E_{19}\) are described as follows:

(a) Domain-domain interaction (DDI) score: For all combinations of human isoform pairs, if a isoform pair has domain-domain interaction in the DOMINE database [32,33], then we assign the DDI score by the confidence level in the DOMINE database as follows: high-confidence prediction: 3, medium-confidence prediction: 2, and low-confidence prediction: 1. If the isoform pair has several DDI scores, we take the highest score.

(b) 19 RNA-seq datasets (\(E_1, E_2, \ldots, E_{19}\)): the absolute values of Pearson correlations for all isoform pairs derived from RNA-seq datasets. Since each RNA-seq dataset may have different quality and data type, we used the logistic regression model to integrate 19 RNA-seq datasets, and then the coefficients of datasets reflected quality of the RNA-seq datasets.

Additional material

Additional file 1: The isoform modules. There are 1025 isoform modules with gene symbols and RefSeq IDs.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

The National Institutes of Health (NIH) [NHLBI MapGen U01HL108634 and NIGMS R01GM105431] and the National Science Foundation [0747475 to X.J.Z.]; the Taiwan Ministry of Science and Technology grant [MOST 103-2320-B-005-004] and the Ministry of Education, Taiwan, R.O.C. under the ATU plan (to C.C.L.). This article has been published as part of BMC Genomics Volume 16 Supplement 2, 2015. Selected articles from the Thirteenth Asia Pacific Bioinformatics Conference (APBC 2015): Genomics. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcgenomics/supplements/16/S2.

Authors’ details

1Institute of Genomics and Bioinformatics, National Chung Hsing University, Taiwan. 2Molecular and Computational Biology, University of Southern California, USA. 3Department of Internal Medicine, Division of Pulmonary and Critical Care Medicine and Center for Comparative Respiratory Biology and Medicine, University of California Davis, Davis, California, USA. 4National Center for Mathematics and Interdisciplinary Sciences, Academy of Mathematics and Systems Science, CAS, Beijing, China. 5Institute of Biomedical Sciences, National Chung Hsing University, Taiwan. 6Institute of Molecular Biology, National Chung Hsing University, Taiwan. 7Agricultural Biotechnology Center, National Chung Hsing University, Taiwan.

Published: 21 January 2015

References

1. Ellis JD, Barrios-Rodiles M, Colak R, Irimia M, Kim T, Calarco JA, Wang X, Pan Q, O’Hanlon D, Kim PM, et al. Tissue-specific alternative splicing remodels protein-protein interaction networks. Molecular cell 2012, 46(4):884-892.
2. Leeman JR, Gilmore TD: Alternative splicing in the NF-kappaB signaling pathway. Gene 2008, 423(2):97-107.
3. Perumalsamy A, Fernandes R, Lai I, Detmar J, Varmuza S, Casper RF, Jurisicova A. Developmental consequences of alternative Bcl-x splicing during preimplantation embryo development. FEBS J 2010, 277(15):219-233.

4. Rajan P, Elliott DJ, Robson CN, Leung HY. Alternative splicing and biological heterogeneity in prostate cancer. Nat Rev Urol 2009, 6(8):454-460.

5. Vogler M, Hamali HA, Sun XM, Bampton ET, Dinsdale D, Snowden RT, Dyer MJ, Goodall AH, Cohen GM. BCL2/BCL-XL inhibition induces apoptosis, disrupts cellular calcium homeostasis, and prevents platelet activation. Blood 2011, 117(26):7145-7154.

6. Cho YR, Shi L, Ramanathan M, Zhang A. A probabilistic framework to predict protein function from interaction data integrated with semantic knowledge. BMC Bioinformatics 2008, 9:382.

7. McDowell MD, Scott MS, Barton GJ. PIPs: human protein-protein interaction prediction database. Nucleic Acids Res 2009, 37 Database: D651-656.

8. Rhodes DR, Tomlins SA, Varambally S, Mahavuso V, Barrette T, Kalyana-Sundaram S, Ghosh D, Pande A, Chinnaiyan AM. Probabilistic model of the human protein-protein interaction network. Nat Biotechnol 2005, 23(8):951-959.

9. Scott MS, Barton GJ. Probabilistic prediction and ranking of human protein-protein interactions. BMC Bioinformatics 2007, 8:239.

10. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P, et al. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res 2011, 39 Database: D651-656.

11. Brown KR, Jurisica I. Online predicted human interaction database. Bioinformatics 2005, 21(9):206-208.

12. Huang TW, Tien AC, Huang WS, Lee YC, Peng CL, Tseng HH, Kao CY, Huang CY. POINT: a database for the prediction of protein-protein interactions based on the orthologous interactome. Bioinformatics 2004, 20(17):2732-2736.

13. Aranda B, Achuthan P, Alam-Faruque Y, Armean I, Bridge A, Derow C, Feuermann M, Ghanbarian AT, Kermarrec M, Khadake J, et al. The IntAct molecular interaction database in 2010. Nucleic Acids Res 2010, 38 Database: DS25-S31.

14. Lopes CT, Franz M, Kazi F, Donaldson SL, Morris Q, Bader GD. Cytoscape Web: an interactive web-based network browser. Bioinformatics 2010, 26(18):2347-2348.

15. Pruitt KD, Tatusova T, Brown GR, Maglott DR. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res 2012, 40 Database: D130-135.

16. Hu H, Yan X, Huang Y, Han J, Zhou XJ. Mining coherent dense subgraphs across massive biological networks for functional discovery. Bioinformatics 2005, 21(Suppl 1):i213-i221.

17. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000, 25(1):25-29.

18. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000, 28(1):27-30.

19. Yu CS, Chen YC, Li LH, Hwang JK. Prediction of protein subcellular localization. Proteins 2006, 64(3):641-651.

20. Sat B, Freund SM, Fersht AR. Comparative biophysical characterization of p33 with the pro-apoptotic BAK and the anti-apoptotic BCL-xl. J Biol Chem 2007, 282(40):29919-29920.

21. Edlich F, Banerjee S, Suzuki M, Celsand MM, Arnould D, Wang C, Neutzner A, Tjandra N, Youle RJ. The RASSF proteins in cancer: from epigenetic silencing to functional characterization. Biochim Biophys Acta 2009, 1796(2):114-28.

22. Pelosi G, Fumagalli C, Trubia M, Sonzogni A, Rekhtman N, Maisonnoue P, Galetta D, Spaggiari L, Veronesi G, Scapar A, et al. Dual role of RASSF1 as a tumor suppressor and an oncogene in neuroendocrine tumors of the lung. Anticancer Res 2010, 30(10):4269-4281.

23. Kodama Y, Shumway M, Leinonen R. International Nucleotide Sequence Database C. The Sequence Read Archive: explosive growth of sequencing data. Nucleic Acids Res 2012, 40 Database: D54-56.

24. Roberts A, Pachter L. Streaming fragment assignment for real-time analysis of sequencing experiments. Nat Methods 2013, 10(1):71-73.

25. Liu Y, Schmidt B. Long read alignment based on maximal exact match seeds. Bioinformatics 2012, 28(18):i318-i324.

26. Qi Y, Klein-Seetharaman J, Bar-Joseph Z. A mixture of feature experts approach for protein-protein interaction prediction. BMC Bioinformatics 2007, 8(Suppl 10):56.

27. Spinkzik E, Altuvia Y, Margalit H. Characterization and prediction of protein-protein interactions within and between complexes. Proc Natl Acad Sci USA 2006, 103(40):14718-14723.

28. Raghavachari B, Taneeen A, Przytycka TM, Jothi R. DOMINE: a database of protein domain interactions. Nucleic Acids Res 2008, 36 Database: D656-661.

29. Yellaboina S, Taneeen A, Zaykin DV, Raghavachari B, Jothi R. DOMINE: a comprehensive collection of known and predicted domain-domain interactions. Nucleic Acids Res 2011, 39 Database: D730-735.