


Investigation and Prediction of Fusion Specific Protein-Protein Interactome Changes

Wu Zhifan

April 4th, 2018



Course name: BCB330
Supervisor: Igor Jurisica

Introduction:

Fusion genes are hybrid genes formed when two previously independent parent genes become juxtaposed (Latysheva and Badu, 2016; Latysheva et al., 2016; Abate et al., 2014). A variety of mechanisms (Figure 1) can result in gene fusion, for example, insertions, deletions, inversions, and translocations.

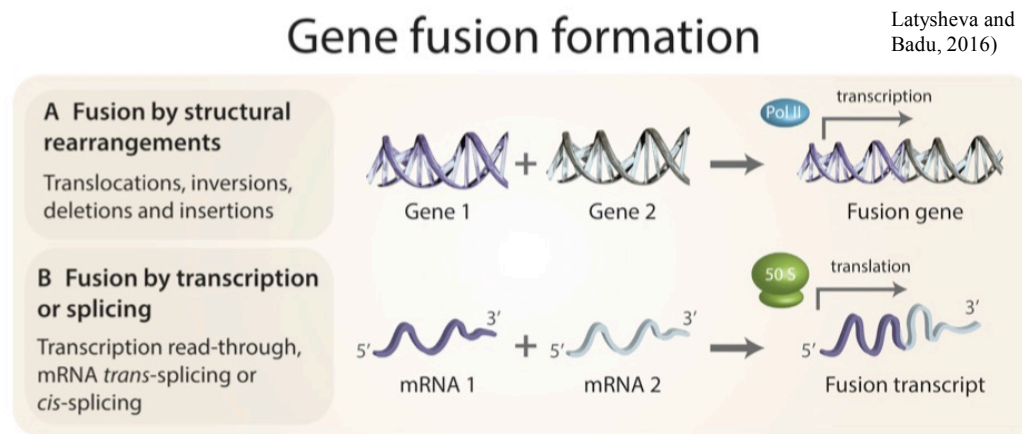


Figure 1. Mechanisms of gene fusion formation. (A) Structural rearrangements of chromosomes, such as translocations, inversions, deletions and insertions, can lead to the formation of gene fusions. These hybrid genes may then be transcribed and translated into fusion transcripts and proteins. (B) Non-structural rearrangement mechanisms, such as transcription read-through of neighboring genes or splicing of mRNA molecules, are increasingly recognized as leading to the formation of a large proportion of gene fusions.

Fusion genes are recognized as oncogenes or common driver mutations in many cancer types including blood/lymph/bone marrow tissue malignancies, solid tumors, lymphomas, leukemias and prostate cancer (especially *TMPRSS2-ERG* fusion) (Latysheva and Badu, 2016; Latysheva et al., 2016; Frenkel-Morgenstern et al., 2017). They are either deregulating one of its partner genes (proto-oncogene fused with a strong promoter), translating to a fusion protein such as tyrosine kinase or introducing a loss of function (e.g. by truncating a tumor suppressor gene) to express the oncogene functionality (Latysheva and Badu, 2016). For example, the BCR:ABL fusion gene (Figure 2), which is the first discovered and most famous fusion genes, is considered to be the principle oncogenic driver in chronic myeloid leukemia (CML). ABL genes are normally on chr9 and encode tyrosine kinase while BCR genes are normally on chr22. When a translocation

happened between chr9 and chr22, a phosphate group from BCR was added to ABL and become to express oncogene functionality (Sun et al., 2013). The discovery of BCR-ABL fusion event leads to the development of a drug (Imatinib/Gleevec) that is “highly specific for inhibiting the fusion kinase, resulting in a breakthrough treatment for a poorly responsive disease” (Frenkel-

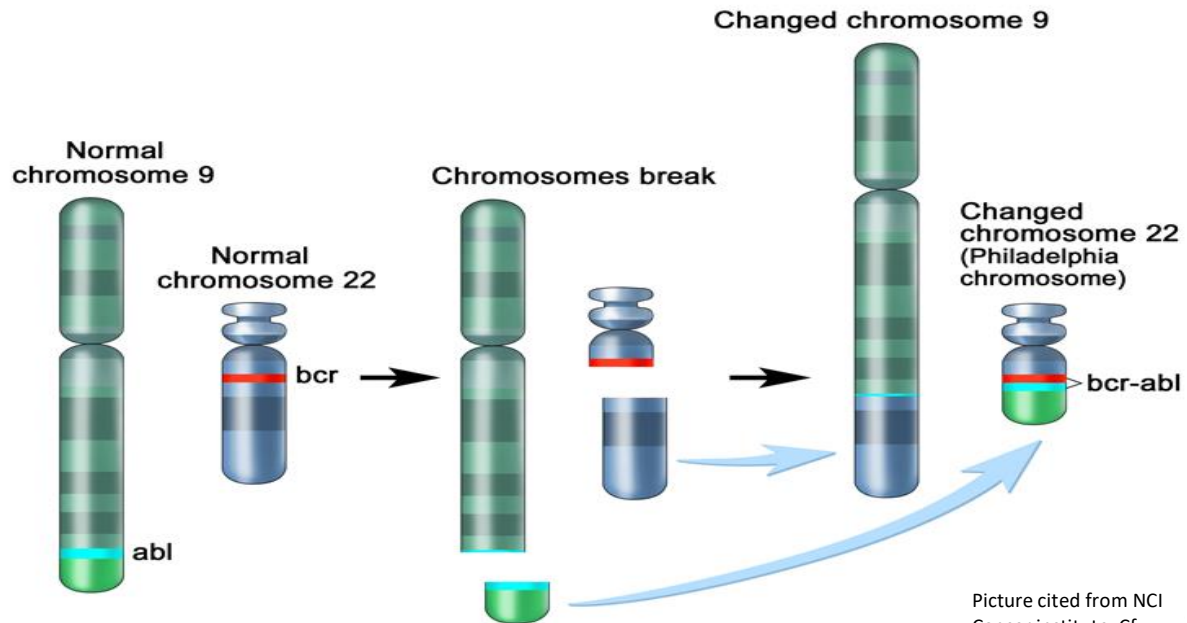


Figure 2. ABL-BCR fusion gene creation process (Cf. <https://www.cancer.gov/about-nci>)

Picture cited from NCI Cancer institute. Cf. <https://www.cancer.gov/about-nci>

Morgenstern et al., 2017). “One estimate states that translocations and gene fusions are responsible for 20% of global cancer morbidity (11), largely due to their central involvement in prostate cancer” (Latysheva and Badu, 2016). Nowadays, due to the advent of next-generation sequencing technology (Mardis, 2008) and advanced bioinformatics theory and algorithms, the DNA/RNA level fusions are vastly well-studied and the related databases are well constructed, organized and up-to-date. Several widely used databases include Mitelman

[<https://cgap.nci.nih.gov/Chromosomes/Mitelman>], ChimerDB

3.0[<http://203.255.191.229:8080/chimerdbv31/mhelp.cdb>], and ChiTaRs

[<http://chitars.md.biu.ac.il/>]. Yet, approximately 10,000 fusion genes have been discovered, the

functions of fusion proteins they encode and the cellular context in which they operate remain relatively poorly understood.

Fusion proteins (comprising peptide of two parental proteins) or chimeras, one possible outcome of fusion genes, are well known as a result of chromosomal translocation in cancerous cells or continuous transcription of neighboring genes or trans/cis-splicing of pre-mRNA (Frenkel-Morgenstern et al., 2017; Saha and Jones, 2005; Latysheva et al., 2016). If fusion transcript is translated (Figure 3), the resulting fusion protein might escape cellular regulation pathways and redirect cellular signaling pathways, thus acting as primary oncogenic drivers (Latysheva et al., 2016). Although some scientist (Yu et al., 2014) were concerned about whether certain putative mRNA fusion event might be artifacts during the sequencing process, much more evidence from clinical experiment reports and research papers indicate the potential of fusion proteins for important biological impact (Latysheva et al., 2016). While fusion genes are well-studied, how fusion proteins influence the interaction networks and how they cause disease stay not well-understood. Saha et al. argued that fusion proteins are related to several diseases including development malignancies and thyroid carcinomas (2005). Latysheva et al. argued about the expression and regulation of fusion proteins: “

1. Parents of fusion proteins occupy central positions in protein interaction networks,
2. Parents are rich in interaction-mediating features, which are often lost via fusion,
3. Fusions preferentially join proteins with no previous connection in protein networks,
4. Fusion proteins escape regulation by losing posttranslational modification sites.” (2016). And such outline is the expected test-purpose for my project. Frenkel-Morgenstern et al. presented a novel method to map the chimeric protein-protein interactions based on domain-domain co-occurrence scores (2017). The advantage of their method is that they predict interactor of both

normal proteins and fusion proteins. Interestingly, Sun et al. performed a fusion protein database, CanProFu, construction based on 6259 reported and predicted gene fusion pairs from ChimerDB 2.0 and Cancer Gene Census (2013). In CanProFu database, the fusion peptides formed by exon-exon linkage of each gene pairing are comprehensively annotated. CanProFu database is flexible to MS/MS-based human cancer experiments.

According to the physical interactions between proteins, we can organize proteins into a network or groups, also known as protein-protein interactions (PPI). PPI induced the discovery of a huge number of functions of otherwise anonymous proteins such as the co-activator or co-repressor (Frenkel-Morgenstern et al., 2017). Nowadays, only ~10% of human PPIs might be understood while around one-third of human proteins have no known interactions (Kotlyar et al., 2015). More importantly, another one-third of human proteins have fewer than five known interactions. Recently, the PPI networks are often analyzed by the computational algorithm or mathematical methods (Machine learning) because of the research bias and limitations among biological assays (Kotlyar et al., 2015). Furthermore, PPIs have been applied to tons of predictions of the normal form of protein interactome changes, but not been yet systematically addressed to the prediction of fusion protein interactome changes. Frenkel-Morgenstern et al. (2017) performed a fusion specific protein preserved interactors identification using domain-domain co-occurrence scores.

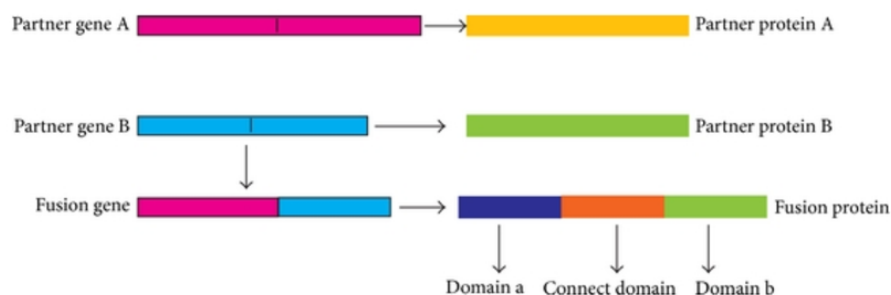


Figure 3. Fusion protein formation. (Cf. https://www.researchgate.net/figure/A-multistep-process-of-the-formation-of-fusion-proteins-A-novel-fusion-gene-is-formed_fig4_281622306)

Since fusion proteins are being recognized as “important diagnostic and prognostic biomarks”,

and the systematical prediction of fusion-specific protein-protein interactomes changes has not yet been well-studied. Here I used fusion gene-pair data downloaded from [http://www.unav.es/genetica/TICdb/] (release 3.3 (August 2013)), protein interactor data from IID database [http://iid.ophid.utoronto.ca/iid/Search_By_Proteins/] (version 2017-04), protein domains data downloaded from FpClass [http://dcv.uhnres.utoronto.ca/FPCLASS/properties/] and fusion transcript data from ChiTaRs-3.1 database [http://chitars.md.biu.ac.il/index.html] to approach the prediction of interactome changes.

Method / Experiment process:

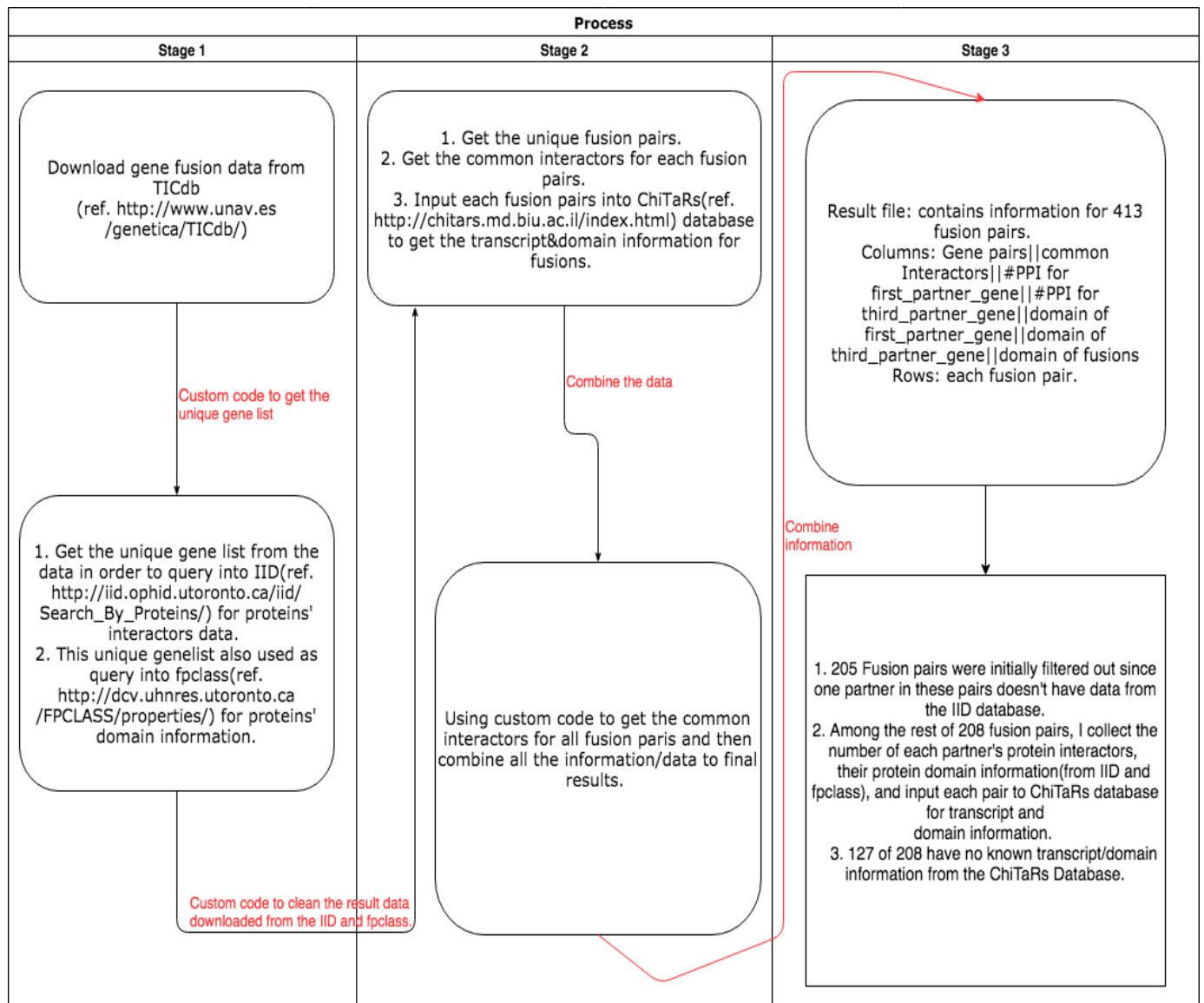


Figure 4. Workflow and methods overview.

I separated the project into 3 stages (Figure 4) and talked about the methods I used below:

Stage 1: Download the data and clean the data.

In this stage, I downloaded the fusion gene data from [http://www.unav.es/genetica/TICdb/] (release 3.3 (August 2013)) since I couldn't find fusion protein data directly. The data file downloaded from TICdb are well-organized in a tab separated

1	5'_PARTNER_GENE	3'_PARTNER_GENE	REFERENCE	FUSION_SEQUENCE
2	ACSL3	ETV1	18594527	ctgtgtcacaccaccttagcctcttgatcgaggaagTGCCTATGATCAGAAGCCACAAGTGGGAATGAGGC
3	ACTB	GLI1	22575261	atgaccagCatgaccacacatcct
4	AFF3	BCL2	18622426	AAAAGGAACGGGAGCTGAGAGcatcacaggaagtagactgatattaac
5	AGTRAP	BRAF	20526349	agtcaagatgcccgaggGccccaattctcaccagtcggtctccttcaaa
6	AHRR	NC0A2	22337624	aactgcatggaaaaccaattactcagcagGaatgatttgtaacagtgcttctcggcctacta
7	AHRR	NC0A2	22337624	agggaacagtagcacagGaaggagcagcagagagagcgcg
8	AHRR	NC0A2	22337624	gacaccgcagccacgcgagtgcaaaGcagccatttggcagttctccagatgacttgct
9	AHRR	NC0A2	22337624	tcttttgccagccaaaacagAgtaaaagccaccaccagtcgtgcg
10	AKAP9	BRAF	15630448	GAACAGgacttg
11	ALK	PTPN3	22334442	aggaggttacacaaggggtgtTgtttcccccggtg
12	ANKRD28	NUP98	17988990	ATTTTGAAGAGGACACCTATTCATGCAGCAGctttgacagatccaaatgcttctgctgcc
13	ARHGAP6	PRCC	8986805	cccagggaggtgctcaagTcagattctgaggaagatga
14	ASPSR1	TFE3	11438465	accacaggactgcagctccccagcagCccgtggaccgggag
15	ASPSR1	TFE3	11438465	cacctgcagcaggcgccggcagcagCccgtggaccgggag

Figure 5. Sample gene fusions data downloaded from TICdb [http://www.unav.es/genetica/TICdb/] (release 3.3 (August 2013)).

format. As shown in Figure 5, there are many duplicated gene fusion pairs with different sequences and this is matched to the TICdb's record they have 1374 fusion sequences and 431 unique genes. The reason for many duplicated fusion pairs is due to different biological assays, and sequencing bias/methods, and also genes in DNA forms always have much more nucleotides in length compared to their corresponding RNA and Protein forms. Since I am only interested in predicting protein-level interactome changes, I filtered out the 432 unique genes' name and 414 unique fusion pairs from TICdb data. Then I used the 432 unique genes as a query into the IID database [http://iid.ophid.utoronto.ca/iid/Search_By_Proteins/] (version 2017-04) with the following parameters:

1. Select Experimental evidence, Orthologous interaction evidence, Computational predictions and Include interactions among partners of query proteins in "Find interaction partners supported by:"

2. Select Search using orthologs of your proteins in “Options for searching across species:”
3. Select Required evidence: gene OR protein expression in “Options for searching across tissues:”
4. Select species Human and select tissues any. Also, in “Information to include in output table:” select Source information: detection methods, PubMed IDs, reporting databases and Tissue information: presence/absence of interactions in selected tissues.

Then clean the results downloaded from the IID database using custom code. The concept behind my code is to:

1. Collecting the interactor pairs by filtering out columns (“Query Symbol” and “Partner Symbol”).
2. Delete the duplicated ones.

And also I use the 432 unique genes as a query to FpClass

[<http://dcv.uhnres.utoronto.ca/FPCLASS/properties/>] to collect protein domains data. The result showing in Figure 6. I only selected the Query ID and Protein domains.

A

```

1  10 IDs could not be mapped to proteins:
2  C15orf21, HELIOS, ERVK-17, HIST1H4I, Ig, MIR142, EWS, NABP1, Immunoglobulin, KIAA1549L
3
4  17 IDs had no data:
5  C15orf55, C6orf204, CEP110, HMG2P46, KIAA1549, MALAT1, TPR,
6  RET, KIT, MIPOL1, MDS2, ERG, MLLT11, CCDC28A, LNP1, C20orf112, SSX2

```

B

Query ID	Protein Domains
ACSL3	AMP-binding_conserved_site, AMP-dependent_synthetase/ligase
ACTB	Actin-like, Actin/actin-like_conserved_site, Actin_conserved_site
AFF3	Transcription_factor_AF4/FMR2
AGTRAP	Angiotensin_II_type_I_receptor-associated
AHRR	Helix-loop-helix_DNA-binding, Helix-loop-helix_DNA-binding_domain, PAS, PAS_fold
AKAP9	Pericentrin/AKAP-450_centrosomal_targeting_domain

Figure 6. A is the result from fpclass that could not be mapped to a protein or had no data. B is the sample of data downloaded from fpclasses.

Stage 2: Find the common interactors and chimeric domain data.

In stage2, since interactors data and domains data ready, I tried to find the common interactors of each fusion pairs and their fusion domain information.

I used a python script to find the common interactors, the idea behind my code is: I started to build a large dictionary to store each gene names as key and their interactors data as value. Then I used this large dictionary to collect each pairs' common interactors and assign to each gene pairs in the final result file. After collecting the common interactors, I started to collect each pair's domain information from fpclass-result and assigned to the result file.

After I finished collect the domain information, I started to query 414 unique fusion pairs into ChiTaRs-3.1 database [<http://chitars.md.biu.ac.il/index.html>] for fusion-specific domains information. The reason I filtered out the unique fusion pairs is only certain regions of fusion DNA would transcript to RNA and translate to protein. Also, the transcript and protein data for fusions are far from completion compared to fusion DNA data. And the result also provided supports to such pre-filtering step.

Stage 3: Clean the result file and analysis the results.

In this stage, I collect some more data into my final result file. I added the number of interactors for each partner in each fusion pair. Then I wrote an R script (Figure 7) and used it to analyze the result. The idea behind this script is it reads in the result file and treated it as a data-frame, thus to get the information for each fusion pair easily.

```
# Example
find_FUdomain("ANKRD28 NUP98")
# [1] "Since this pair don't share interactors, I considered it as not important"
find_FUdomain("ACSL3 ETV1")
# [1] "AMP-binding enzyme, PEA3 subfamily ETS-domain transcription factor N terminal domain and Ets-domain"
find_domain("ACSL3 ETV1")
# [1] "AMP-binding_conserved_site, AMP-dependent_synthetase/ligase"
# [2] "Ets, PEA3-type_ETS-domain_transcription_factor_N-terminal, Winged_helix-turn-helix_transcription_repressor_DNA-binding"
```

Figure 7. Sample test case for the R script. We can directly find the conserved domain of ACSL3 and ETV1 fusions and the lost domains.

Result:

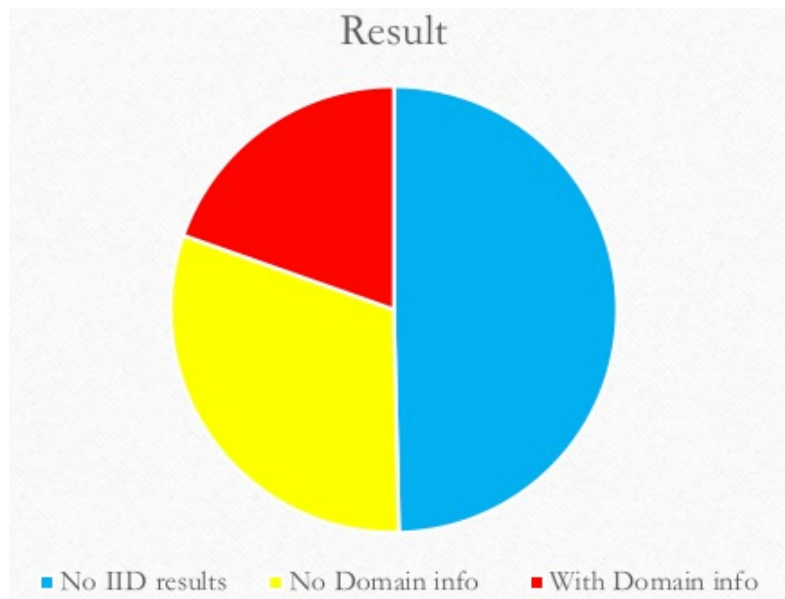


Figure 8. Final results.

I stored all the result into a file (file-name: result.txt). It (Figure 8) contains information including fusion pairs, common interactors, number of protein-protein interactors for first_partner_gene, number of protein-protein interactors for second_partner_gene, domain of first_partner_gene, domain of

second_partner_gene, and domain of fusions. It is separated by "|". Noted that 205 fusion pairs were initially filtered out since either one partner in these pairs doesn't have data from the IID database, which means either one partner has no interactors data. Thus, I considered these pairs are not important for my cases. Among the rest of 208 fusion pairs, I collect the number of each partner's protein interactors, their protein domain information (from IID and fpclass) and input each pair to ChiTras database for fusion domain information. The result shows that 127 in 208 have no known domain information from the ChiTaRs Database.

Discussion:

Nearly half of the fusion pairs were initially ignored since either one partner has no protein interactors data from IID database. I was confident to perform this pre-filter step due to the theory from Latysheva et al. (2016) that "in most known cancer fusion gene pairs, at least one of the fusion partners acts as a hub (i.e. has many interaction partners) in a gene interaction

network (where genes are nodes and edges indicate a regulatory or protein-protein interaction).” Thus, further investigation is required to be performed for these data.

In the remaining fusion pairs, more than half of them don’t have domain information from the ChiTaRs database. This is also consistent since the Chimeric transcript data is still far from completion. And since I only search for the fusion pairs in Human data, this might be also responsible. But in the fusion pairs with domain information, we can find out the domains conserved or lost using my R script easily (Figure 7). For example, in fusion pair “AGTRAP BRAF”, the domains for “AGTRAP” is “Angiotensin_II_type_I_receptor-associated”, the domains for “BRAF” is “Diacylglycerol/phorbol-ester_binding, Protein_kinase-like_domain, Protein_kinase_ATP_binding_site, Protein_kinase_C-like_phorbol_ester/diacylglycerol_binding, Protein_kinase_catalytic_domain, Raf-like_Ras-binding, Serine-threonine/tyrosine-protein_kinase, Serine/threonine-protein_kinase_active_site”, and the domains for fusion is “Angiotensin II, type I receptor-associated protein (AGTRAP) and Protein tyrosine kinase”. We can easily find that Angiotensin domain and tyrosine kinase domain are conserved while other domains are lost after the fusion event.

One problems is that the find_interactors.py python script I used to organize the interactors data has an extremely bad runtime (runs 10 hours in a 16GB-RAM and I7 on MacBook pro 15’ 2017). It might because the read-in data was large, and I used 2 loops inside the code.

Although the data in my final result was not enough to perform a machine learning prediction algorithm, the approach I used was proved to be correct. However, I was only able to find the domain data from ChiTaRs database and more than half of the fusion pairs don’t have a record on it, such result indicates the need for more transcript and protein level fusion data.

Conclusion:

To summarize, I collected the fusion data from DNA level with the available fusion pairs' gene symbol information, and successfully found part of the protein level fusion domains' data. All the clean-data processes are done by custom python script. The final data was put into one result file which was able to be analyzed. By using the custom R script, I can find the domain information about some fusion pairs. Although I am not able to perform the prediction step, I successfully collect the fusion pair names, their number of protein interactors, their protein domain information and, for some I collect the fusion domains information.

Acknowledge:

I would like to thank Dr. Igor Jurisica for supervising this project. Thanks to Chiara Pastrello for the help to find some of the databases I used in this project. Thanks to Max Kotlyar and Dylan Bethune-Waddell for the help in finding papers and related online methods for this project.

Reference:

Abate F., Zairis S., Ficarra E., et al. (2014) Pegasus: a comprehensive annotation and prediction tool for detection of driver gene fusions in cancer. *BMC Syst Biol* 2014;8:97.

Carrara, M., Beccuti, M., Lazzarato, F., Cavallo, F., Cordero, F., Donatelli, S., & Calogero, R. A. (2013). State-of-the-Art Fusion-Finder Algorithms Sensitivity and Specificity. *BioMed Research International*, 2013, 340620. <http://doi.org/10.1155/2013/340620>

Kotlyar M., Rossos AE.M., Jurisica I. Prediction of Protein-Protein interactions (draft).
Latysheva,N.S., Oates,M.E., Maddox,L., Flock,T., Gough,J., Buljan,M., Weatheritt,R.J. and Babu,M.M. (2016) Molecular principles of gene fusion mediated rewiring of protein interaction networks in cancer. *Mol. Cell*, 63, 579–592.

Kotlyar M, Pastrello C, Pivetta F, Lo Sardo A, Cumbaa C, Li H, Naranian T, Niu Y, Ding Z, Vafae F, Broackes-Carter F, Petschnigg J, Mills GB, Jurisicova A, Stagljjar I, Maestro R, Jurisica I (2015) In silico prediction of physical protein interactions and characterization of interactome orphans. *Nat Methods* 12: 79–84

Latysheva,N.S. and Babu,M.M. (2016) Discovering and understanding oncogenic gene fusions through data intensive computational approaches. *Nucleic Acids Res.*, 44, 4487–4503.

Mardis ER. (2008) Next-generation DNA, sequencing methods. *Annu Rev Genomics Hum Genet* 2008;9:387–402

Milana Frenkel-Morgenstern, Alessandro Gorohovski, Somnath Tagore, Vaishnovi Sekar, Miguel Vazquez, Alfonso Valencia. (2017) ChiPPI: a novel method for mapping chimeric protein–protein interactions uncovers selection principles of protein fusion events in cancer, *Nucleic Acids Research*, Volume 45, Issue 12, 7 July 2017, Pages 7094–7105.

National cancer institute (<https://www.cancer.gov/about-nci>). NCI dictionary about cancer term.
Latysheva N.S., Oates M.E., Maddox L., Flock T., Gough J., Buljan M., Weatheritt R.J., Babu M.M. (2016) Molecular principles of gene fusion mediated rewiring of protein interaction networks in cancer. *Mol. Cell.*; 63:579–592.

Okonechnikov K, Imai-Matsushima A, Paul L, Seitz A, Meyer TF, Garcia-Alcalde F (2016) InFusion: Advancing Discovery of Fusion Genes and Chimeric Transcripts from Deep RNA-Sequencing Data. *PLoS ONE* 11(12): e0167417. <https://doi.org/10.1371/journal.pone.0167417>

Sun, H. et al. (2013) Identification of gene fusions from human lung cancer mass spectrometry data. *BMC Genomics* 2013, 14(Suppl 8):S5.

Saha V., Jones LK. (2006) Fusion Proteins and Diseases. In: eLS. John Wiley & Sons Ltd, Chichester.

Youri Hoogstrate, René Böttcher, Saskia Hiltemann, Peter J. van der Spek, Guido Jenster, Andrew P. Stubbs. (2016) FuMa: reporting overlap in RNA-seq detected fusion genes, *Bioinformatics*, Volume 32, Issue 8, 15 April 2016, Pages 1226–1228

Yu, C.-Y., Liu, H.-J., Hung, L.-Y., Kuo, H.-C., and Chuang, T.-J. (2014). Is an observed non-co-linear RNA product spliced in trans, in cis or just in vitro? *Nucleic Acids Res.* 42, 9410–9423.

Kotlyar M, Pastrello C, Pivetta F, Lo

Sardo A, Cumbaa C, Li H, Naranian T, Niu Y, Ding Z, Vafaei F, Broackes-Carter F, Petschnigg J, Mills GB, Jurisicova A, Stagljar I, Maestro R, Jurisica I (2015) In silico prediction of physical protein interactions and characterization of interactome orphans. *Nat Methods* **12**: 79–84