Association Rules Identification in Nervous System Tumors Mutation Data

S. P. B. M. Senadheera[1] ,A. R. Weerasinghe[2] and C. R. Wijesinghe[3]

[1,2,3]University of Colombo School of Computing- Sri Lanka
[1]madhubhaniesenadheera@gmail.com, [2]arw@ucsc.cmb.ac.lk, [3]crw@ucsc.cmb.ac.lk

Abstract. Cancer is a genetic disease which begins by accumulating mutations in an organ. It is important to understand relationship between mutated genes in cancers in order to predict the oncogenic patterns. Objective of the following research is to identify association rules in Human Nervous system cancers. We have analysed fifteen cancer single nucleotide polymorphisms (SNPs) mutation datasets in order to generate the association rules. Initially, we have performed data purification and data enrichment via using Variant Effect Prediction tool and Ensemble Genome conversion tool. According to the Central Dogma protein mutating SNPs will have the highest influence for cancers. In the analysis we considered the protein mutated genes for the association rule identification. We have generated 23 gene level association rules. We have validated the association rules with biological evidence. According the results, ATRX,TP53>IDH1, TP53,TTN>IDH1 and SMARCA4,TP53> IDH1 association rules have biological evidence that they work together inside a living cell.

Keywords: Cancer, Association rules, Mutation, Human nervous system

1. Introduction

Cancer is a disease in which abnormal cells acquire the ability to divide without control and invade other tissues. According to the central dogma [1], genomic mutations will affect the protein function and structure[2]. Therefore, proteins affected by the mutations will provide more insight into cancer genesis[3]. Single nucleotide polymorphism (SNP) is the most common mutation type in genomic mutation datasets. Mutation data analysis requires a sequential process in order to map different data to their genomic and proteomic positions.

Furthermore, some mutated genes will trigger mutations in other genes[4]. Cancers occur due to accumulating such mutations in the genome. Importance of studying the patterns in the mutations, is to understand the mutation/gene/protein level combinations in cancers. Following study focused on understanding the human nervous system cancer gene level patterns.

Identifying association rule is a machine learning technique to understand significant relationships between features in a dataset[5][6]. This technique will perform better with large datasets. Due to the above reason this study considered fifteen types of human nervous system cancers. We have considered relationship identification technique (association rules) which evaluate the interestingness of gene combinations.

2. Objectives

Main objective of the study is to identify the human nervous system cancer related gene combination which has relationship between each other in biological context. In order to achieve the main objective there are three sub objectives. First objective is to get all datasets into common platform. Second objective is to identify significant mutation positions in the human nervous system cancer datasets. Generating workflow for mutation analysis is another parallel objective of this research. Final objective of the research is to identify and validate the gene combinations in human nervous system cancers which have biological relatedness evidence.

3. Methodology

All datasets were downloaded from freely available cBioPortal (July 2018 version v1.13.2[7]). Table 1 shows the dataset details used for the analysis. These datasets annotated with old genomic version (GRCh37) in genomic level and protein level. We have used 15 cancer projects, 11 from central nervous system and 4 from the peripheral nervous system. Each donor/patient is different however there are similar mutations in different projects.

Initial step of the analysis is data purification. Data were generated in Genome Reference Consortium Human Build 37 (GRCh37). We attempted to predict variant effect prediction with the old version. However, we found several errors while predicting the effect. Therefore, first step of the analysis begins with data conversion to the latest version. There were two methods for data conversion. First method was to convert old genome position to new protein positions. Second method was converting old protein position to new protein position.

We considered both conversion methods and analyse the advantages and disadvantages of the two processes. We have used VEP[8] genomic reference conversion tool for genomic position conversion [9] BISQUE [10] and UNIPROT tool [11] for protein position conversion. REST API in Ensemble database was used for the Ensemble ID conversion. We have considered the Ensemble IDs and UNIPROT Ids as protein identifiers. However, many tools prefer the UNIPROT IDs for data annotation. We considered both methods and selected the most effective methods which gives minimum errors in data conversion.

We filtered single nucleotide polymorphisms for further analysis. In order to get the mutation impact Ensemble Variant Effect Prediction Tool [8] was used. Data matrix were generated with R package. Furthermore,

we used Reactome pathway tool [12] to map biological functions for each mutated proteins. All these methods required batch processing hence all datasets have large mutation list to annotate.

In the analysis, we have identified significant changes in genomic positions after genome build conversion and we will discuss them under results and discussion section. The table 1 shows projects we considered and their respective single nucleotide polymorphism percentage. According to the analysis, the majority of reported mutations in all datasets are SNPs. Furthermore, we generated a large dataset which consisted with 15 different tumor projects and 100141 mutation entries. We filtered all protein effecting and genomic changes for association rule generation.

In the association rule identification we considered mutational level association rules, gene level association rules and protein level association rules by using arules algorithm[5]. In mutational level no association rules we generated since mutational diversification was high. Furthermore, protein level association rules not generated because of protein isoform impact. Gene level association rules were generated and these rules were validated with biological evidences. To the above mention task we have used REACTOME[13] tool, PSICQUIC REST API and GENEMANIA tool[14]. These biological evidences we considered based on physical interactions, co-expression interactions, co-localization interactions, biological pathways interactions, predicted interactions, genetic interactions and shared protein domain interactions. We have visualize the interactions among genes in association rules via using REACTOME tool, GENEMANIA tool and R network analysis package. Finally, we have generated reusable workflow for association rule identification for genomic mutation. This method can be used to analyse other types of cancer genomic mutation.

TABLE 3: DATASET DETAILS

Type	Dataset	Mutation	Patients	SNP	SNP/Mutation %
Peripheral nervous system	Mpnst_mskcc	4582	15	3767	82.21
	Nbl_amc_2012	568	73	506	89.08
	Nbl_ucologne_2015	920	56	818	88.91
	Nbl_target_2018_pub	1035	372	1035	100
Central Nervous system	Pcpg_tcga	4662	184	3823	82
	Past_dkfz_heidelberg_2013	236	78	217	91.95
	Lgg_tcga_pan_can_atlas_2018	39299	510	37481	95.37
	Mbl_sickkids_2016	4779	44	4581	95.86
	Mbl_icgc	1059	114	1018	96.13
	mbl_pcgp	558	37	536	96.06
	Mbl_broad_2012	1808	92	1696	93.81
	Odg_msk_2017	281	22	229	81.49
	Gbm_tcga	22073	290	20949	94.91
	Lgg_tcga	9885	286	9228	93.35
	Lgg_ucsf_2014	15155	61	14804	97.68

4. Results and Discussion

4.1. Basic Analysis and Statistics

Dataset consisted with 15 projects and 4 of the projects comes under peripheral tumours and 11 are from central nervous system cancers. We merged all mutation projects into single project for the analysis. There are 106900 variants and 2234 donors in the merged full dataset. Dataset was annotated in GRCh37 (old genome version).

4.2. Genome Build Conversion

Initial data file have old version genomic position and their respective protein mutation positions. We performed basic analysis in order to understand the version compatibility of datasets. We have tested two methods to identify protein position changes. Figure 1 shows the process of analysis.

In the method 1, we have tried to convert protein positions directly. In the analysis we have faced three major issues. Some protein positions do not provide genomic positions. Due to the above reason, we could not find entries for protein such as Q9BX63 position 990. Secondly, there were protein position mapped with

different chromosomes genomic positions (eg: P62805-1,6,12 chromosomes and P84243-1, 17 chromosomes). According to the analysis, we have found out these proteins synthesised in different chromosome locations and they are histone proteins. Third problem was filtering SNP mutations from protein changes. Protein positions such as Q9HC77:1235 are occurred due to intron regions deletion or insertions. Therefore, they are not single nucleotide polymorphisms in genomic level. To solve the above issues, we used the method 2 which had an intermediate steps. Even though, method 2 has more steps than method 1, it will reduce the data loss during data annotation and enrichment.

Initially we filtered the SNP mutations for all datasets and converted to the new genome build. Some genome positions in 37 build are not available in the 38 genome build. This may be due to the assembly errors between the versions. Data loss in the conversion is given below (Table 2). However, it is concluded that from the above mentioned methods, Method 2 is more effective. Therefore, we integrated the method 2 into the workflow.

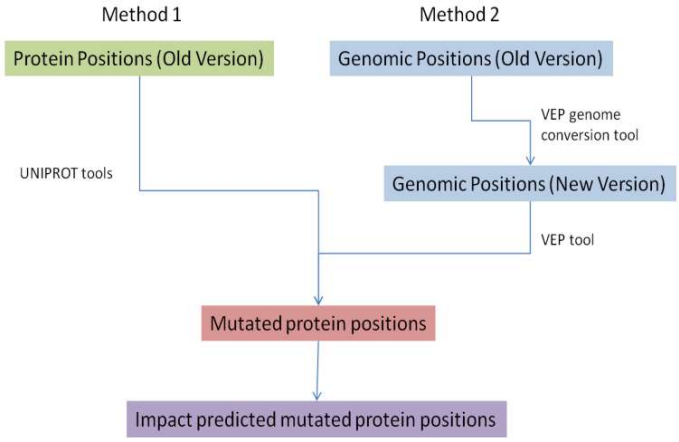

Fig 14 Data conversion workflow of the research

TABLE 4 DATA LOSS IN GENOMIC BUILD VERSION CONVERSION

Project	Input	Output	Data loss %
Mpnst_mskcc	3427	3301	3.676685
Nbl_amc_2012	502	501	0.199203
Nbl_ucologne_2015	816	812	0.490196
Nbl_target_2018_pub	973	944	2.980473
Pcpg_tcga	3371	3332	1.156927
Past_dkfz_heidelberg_2013	211	211	0
Lgg_tcga_pan_can_atlas_2018	36710	36572	0.375919
Mbl_sickkids_2016	4553	4542	0.241599
Mbl_icgc	1001	998	0.2997
Mbl_pcgp	532	530	0.37594
Mbl_broad_2012	1687	1684	0.17783
Odg_msk_2017	173	173	0
Gbm_tcga	20407	20332	0.367521
Lgg_tcga	8850	8818	0.361582
Lgg_ucsf_2014	14003	13961	0.299936

4.3. Mutation Enrichment

After the data conversion step, genome mapped dataset was used for the mutation impact annotation. We have used Variant Effect Prediction tool (VEP) [15] for the following process. This process acquired the longest time in the workflow. We compared web portal and REST API for the analysis. Web portal has taken more time compared to REST API given by VEP tool. Figure 2 shows the visual representation of a project considered in the analysis. Similar to this project, missense variants are the highest mutation type in all projects.

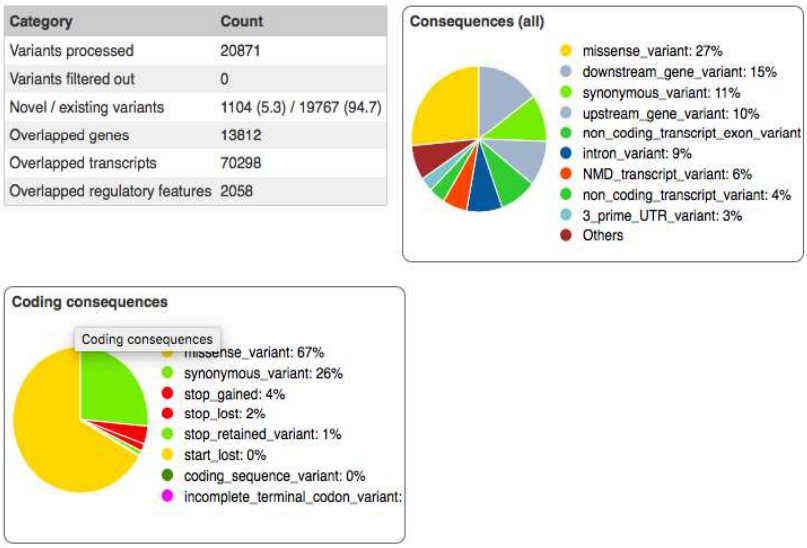

Fig 15 Variant Effect Prediction tool output for brain gbm tcga dataset

From the VEP tools we can annotate the nomenclature, transcript, gene name, protein id, protein change impact, literature regarding the mutation, COSMIC id and exon number. One genomic mutation may alter different transcript and different protein isoforms. Due to the above reason, each genomic position should be annotate with its' respective transcripts and protein isoform ids.

VEP tools provide each of the affected transcript (ENST). We have identified 1106362 mutations which alter protein sequences however these entries are not unique since one mutation position can be altered in different patients. In order to filter the effect of the mutations we considered the protein alteration, miner allele frequency, related publications and predicted protein impact score. For predicting the protein impact we considered the SIFT and Polyphen scores. For these annotation we have used different databases such as DBSNP[16], Clinvar[17], UNIPROT[18]. These features we selected based on ACMG guidelines for mutation impact predictions [19]. We filtered protein altering mutation after considering above information annotated to each mutation.

4.4. Association Rules

We analysed mutational level, gene level and protein level association rules from the dataset we filtered. In the mutational level, diversity of genomic position was high. Figure 3 shows the heatmap drawn for the mutation to patient matrix. According to the results, there is no significant similarity among patients based on genomic level. There are no clear clusters neither a hierarchy in genomic level. Thus, there are no association rules in the genomic mutation level. There are different isoforms for a single protein. Due to that reason there are redundancy in association rules in protein level.

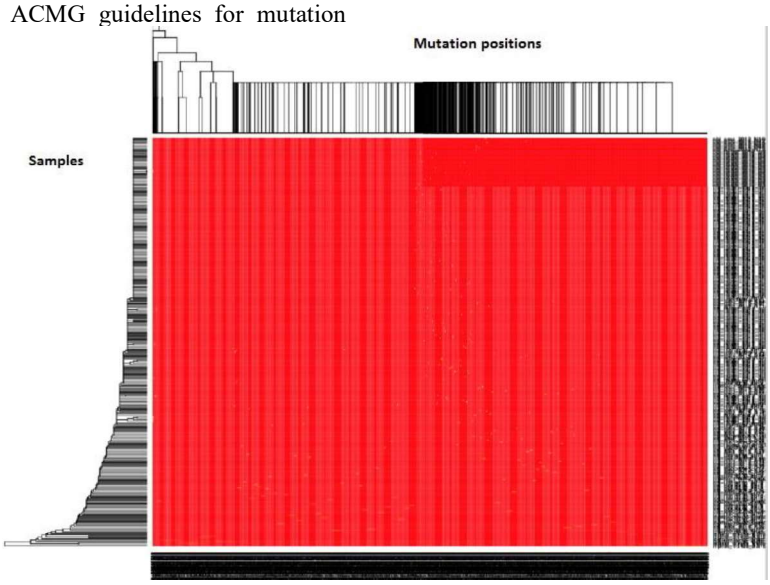

Fig 16 Mutation Heatmap for genomic level mutation vs patients

We have generated 23 association rules for gene level. Table 3 shows the results of association rules and their confidence level. According to the results, Highest confident rules is ATRX, TP53> IDH1. If there are mutations in ATRX and TP53 genes in a patient's cancer there is a 94.78% confidence that IDH1 also mutated in the patient cancer cells. Confidence is an indication of how often the rule has been found to be true in the dataset. However, Support for this is 5.56% which is low. Support is an indication of how frequently the gene set appears in the dataset. According to the results, even though confidence is high diversity of the gene combinations are high.

Furthermore, According to the statistics in genomic SNP mutation IDH1 and IDH2 gene are altered significantly among patients. When investigating the insight of Isocitrate dehydrogenases (IDH1) and cancer, there are several biological proven evidence that mutant IDH1 R132 and IDH2 R172 are used and targeting therapy [20]. Therefore, IDH1 mutation can be considered as an advantage in cancer treatment. According to the results we can see association rules which has IDH1 can be treated as therapy target association rules. On the other hand, most of the rules have IDH1 alteration. This mutation may misguide the association rule results since it always have the high frequency among patients.

TABLE 5 ASSOCIATION RULES FOR GENES

No.	Left hand side of the rule	rhs	Support	Confidence	Lift	Count
(1)	ATRX,TP53	IDH1	0.055584	0.947826	2.614187	109
(2)	SMARCA4,TP53	IDH1	0.012239	0.923077	2.545927	24
(3)	CIC	IDH1	0.048445	0.913462	2.519407	95
(4)	FUBP1	IDH1	0.010709	0.913044	2.518254	21
(5)	APOB,TP53	IDH1	0.010199	0.909091	2.507352	20
(6)	ATRX	IDH1	0.065273	0.882759	2.434725	128
(7)	ATRX,IDH1	TP53	0.055584	0.851563	3.407988	109
(8)	NOTCH1	IDH1	0.018358	0.837209	2.309096	36
(9)	ATRX,TTN	IDH1	0.010199	0.833333	2.298406	20
(10)	TP53	IDH1	0.198878	0.795918	2.195212	390
(11)	ATRX	TP53	0.058644	0.793103	3.174032	115
(12)	ARID1A	IDH1	0.012239	0.774194	2.135293	24
(13)	SMARCA4	IDH1	0.020398	0.727273	2.005882	40
(14)	MUC16,TP53	IDH1	0.016318	0.727273	2.005882	32
(15)	APOB,IDH1	TP53	0.010199	0.714286	2.858601	20
(16)	TP53,TTN	IDH1	0.023457	0.657143	1.812457	46
(17)	APOB	IDH1	0.014278	0.636364	1.755146	28
(18)	IDH1,MUC16	TP53	0.016318	0.627451	2.511084	32
(19)	IDH1,TTN	TP53	0.023457	0.613333	2.454585	46
(20)	IDH1,SMARCA4	TP53	0.012239	0.6	2.401224	24
(21)	IDH1	TP53	0.198878	0.548523	2.195212	390
(22)	LRP2	TP53	0.011729	0.511111	2.045488	23
(23)	APOB	TP53	0.011219	0.5	2.00102	22

4.5. Validation Association Rules

After generating the association rules we have validated the results with biological evidences. We have validated each association rules by using PSICQUIC tool (Platform which combines biological interaction evidence databases) which has combination of protein-protein and gene interactions. Since PSICQUIC provides literature evidence regarding the association between two genes, association rules that computationally generated from the research can be proven from biological evidence. Figure 4 shown below is the PSICQUIC output of SMARCA4,TP53> IDH1 association rules. According to the results there is a direct association between TP53 and SMARCA4 however, there are no biological evidence shown to prove between SMARCA4 and IDH1 or TP53 and IDH1. Furthermore, most of the association rules could not validated via this method since all genes in a particular rule will not interact directly. However, in three gene combinations at least two genes interact directly. We have consider the interaction confidence score in order to confirm the interaction. Confidence does not show the

interaction robustness between genes therefore it can only explain how confidence the gene interaction is validated in research domain.

interactionID	confidenceScore	provider	A.name	B.name
biogrid:261857	*NA*	BioGrid	TP53	SMARCA4
biogrid:261858	*NA*	BioGrid	TP53	SMARCA4
biogrid:261862	*NA*	BioGrid	TP53	SMARCA4
biogrid:657418	*NA*	BioGrid	TP53	SMARCA4
biogrid:669727	*NA*	BioGrid	TP53	SMARCA4
intact:EBI-1750529|imex:IM-20034-3	intact-miscore:0.35	IntAct	TP53	SMARCA4
intact:EBI-1750529|imex:IM-20034-3	intact-miscore:0.35	IMEx	TP53	SMARCA4
biogrid:261862	mentha-score:0.814	mentha	TP53	SMARCA4
biogrid:261857	mentha-score:0.814	mentha	TP53	SMARCA4
intact:EBI-1750529	mentha-score:0.814	mentha	TP53	SMARCA4
biogrid:261858	mentha-score:0.814	mentha	TP53	SMARCA4

Fig 17 Screenshot of the RESTAPI output of PSICQUIC

As the next step we consider the 1st neighbour linkage for the gene combination. Human Interactome has published 11,999 proteins and interactions 74,771 as at 31 June 2018. Human Interactome was generated based on literature and predicted protein-protein interactions. We considered the human protein interactome as a network and calculated shortest paths for each gene combinations. According to the results, it has shown that many links between two genes may be not direct interaction but 1st or 2nd neighbour interactions. Figure 5 shown below is the GENEMANIA network visualization for ATRX,TTN> IDH1 association rule. According to the results shown here, IDH1 is having interaction with ATRX and TTN through IDH2 and VIM. However, these interactions should be further investigated since they should be in the same pathway or share common attributes in co-expression level.

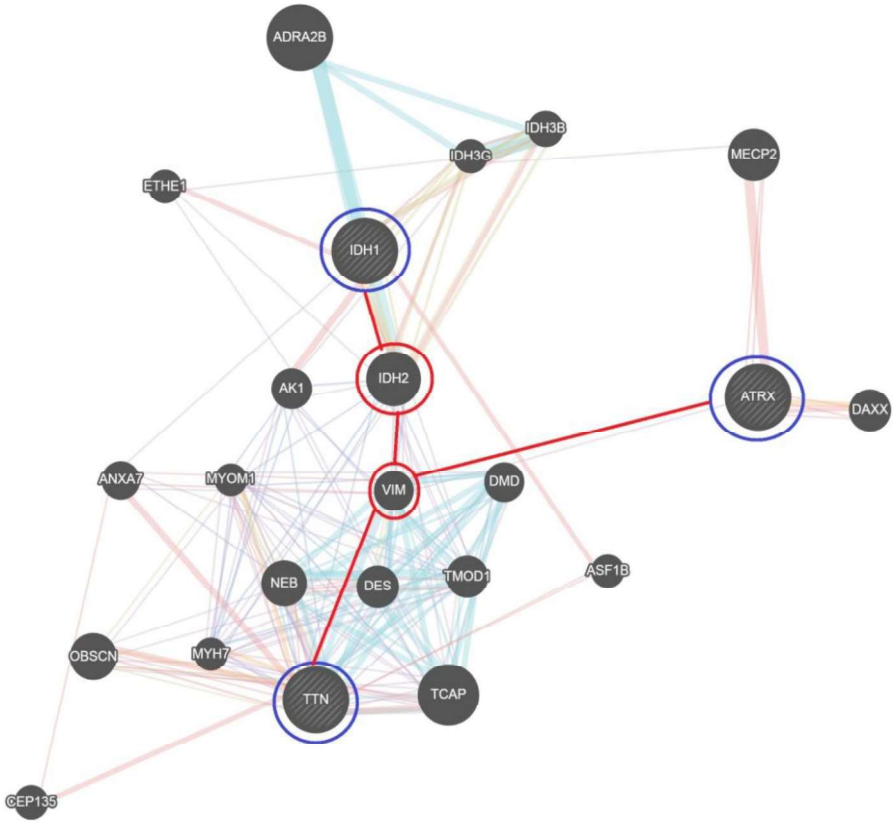

Fig 18 Association rule ATRX,TTN> IDH1 visualization with GENEMANIA tool

We considered the reactome pathways to track the relationship between mutations and biological impact. Initially, we mapped genes to their respective biological pathways. Most of the mutated genes are affecting signal

transduction pathway. Figure 6 shows the affected biological pathways and mutation frequencies.

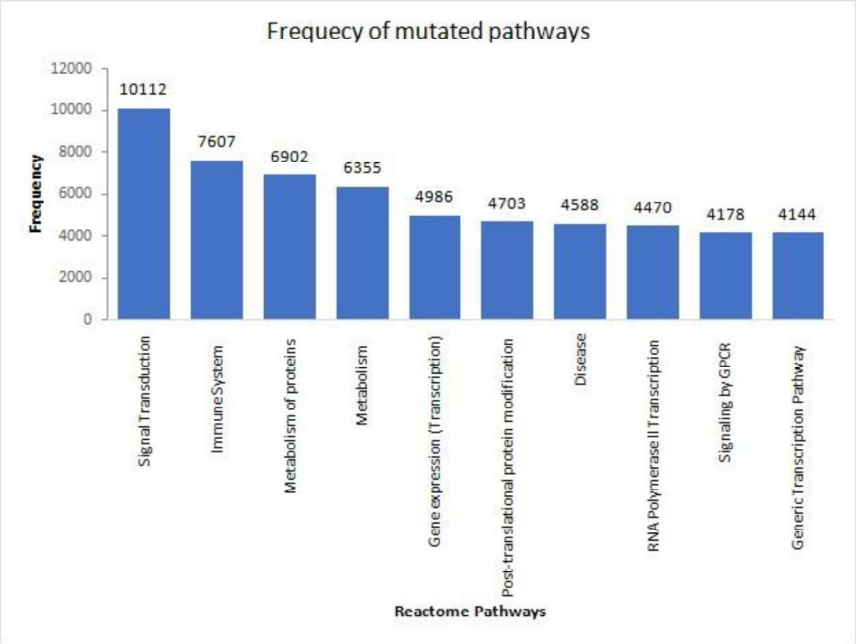

Fig 19 Mutations vs affecting reactome pathways

We visualized the association rules based on reactome pathways. Reactome pathways have a hierarchy which connects main reacotme pathways to sub level reactome pathways. We have designed a panel to visualize the hierarchy of the reactome pathways affected by particular association rule. Figure 7 highlighted the association rule genes and reactome pathways which combine them. According to the results, many association rules have common reactome in higher level however, all the genes in an association rule do not work under one reactome pathway.

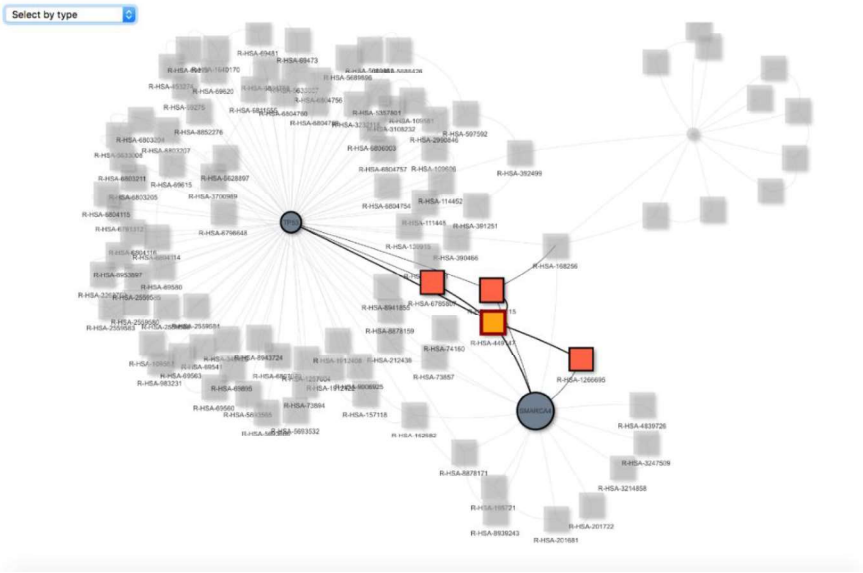

Fig 20 Association rule SMARCA4,TP53> IDH1 visualization with reactome pathway interactions

We have omitted the two gene combinations since they have lower impact as a combination. Moreover, some association rules gene combinations are repetitive if all genes considered as a single group. (eg: SMARCA4,TP53> IDH1 and SMARCA4,IDH1> TP53) These combinations are validated as a single case.

5. Conclusions

From the following study we can conclude SNP effect prediction required combination of tools for effective data annotation. There are main three steps of the mutation effect prediction. Initially all mutations should be converted to the latest genome build. Secondly, mutation should be annotated with genomic data. Finally genomic position which alter the proteins should be annotated with protein data. We have generated a reusable effective workflow for genomic data annotation and data mining. Mainly we proved ATRX,TP53>IDH1, TP53,TTN>IDH1 and SMARCA4,TP53> IDH1 association rules with biological evidences. In the data validation step we have considered direct association and indirect association between genes. Moreover, computationally captured rules were validated with biological evidences via different databases. These validation process was designed as a workflow which can be reusable. For association rule visualization, we have designed a panel which explains the interaction and hierarchy of the reactome pathways. As future work, we need to design more insightful association rules by combining other cancer types.

6. Acknowledgement

We would like to acknowledge the guidance provided by the European Bioinformatics Institute (United Kingdom) for this study. We would like to thank specially the molecular interaction team for providing the information regarding data and tools for the analysis.

References

1. B. Alberts, A. Hnson, J. Ewis, M. Raff, K. Roberts, and P. Alter, *The Cell*, vol. 40, no. 6. Taylor & Francis Group, LLC, an informa business, 270 Madison Avenue, NewYork NY f 0016, USA, and 2 park Square, Milton park, Abingdon, OXl4 4RN, UK., 2001.
2. D. Hanahan and R. a Weinberg, "Hallmarks of cancer: the next generation.," *Cell*, vol. 144, no. 5, pp. 646–74, Mar. 2011.
3. O. N. Ikediobi *et al.*, "Mutation analysis of 24 known cancer genes in the NCI-60 cell line set.," *Mol. Cancer Ther.*, vol. 5, no. 11, pp. 2606–2612, 2006.
4. P. Lecca, N. Casiraghi, and F. Demichelis, "Defining order and timing of mutations during cancer progression: the TO-DAG probabilistic graphical model.," *Front. Genet.*, vol. 6, p. 309, 2015.
5. R. Agrawal and S. Ramakrishnan, "Fast Algorithms for Mining Association Rules," *IBM Almaden Research Center*, 1994. [Online]. Available: http://rakesh.agrawal-family.com/papers/vldb94apriori.pdf. [Accessed: 17-Mar-2016].
6. Y. Xu, M. Zeng, Q. Liu, and X. Wang, "A Genetic Algorithm Based Multilevel Association Rules Mining for Big Datasets," *Math. Probl. Eng.*, vol. 2014, pp. 1–9, Aug. 2014.
7. J. Gao *et al.*, "Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal.," *Sci. Signal.*, vol. 6, no. 269, p. pl1, Apr. 2013.
8. W. McLaren *et al.*, "The Ensembl Variant Effect Predictor," *Genome Biol.*, vol. 17, no. 1, p. 122, Dec. 2016.
9. "Ensembl genome browser 96." [Online]. Available: https://asia.ensembl.org/index.html. [Accessed: 17-May-2019].
10. M. J. Meyer, P. Geske, and H. Yu, "BISQUE: locus- and variant-specific conversion of genomic, transcriptomic and proteomic database identifiers," *Bioinformatics*, vol. 32, no. 10, pp. 1598–1600, May 2016.
11. "UniProt." [Online]. Available: https://www.uniprot.org/. [Accessed: 17-May-2019].
12. D. Croft *et al.*, "Reactome: a database of reactions, pathways and biological processes.," *Nucleic Acids Res.*, vol. 39, no. Database issue, pp. D691-7, Jan. 2011.
13. A. Fabregat *et al.*, "The Reactome Pathway Knowledgebase," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D649–D655, Jan. 2018.
14. D. Warde-Farley *et al.*, "The GeneMANIA prediction server: Biological network integration for gene prioritization and predicting gene function," *Nucleic Acids Res.*, vol. 38, no. SUPPL. 2, pp. 214–220, 2010.
15. W. McLaren *et al.*, "The Ensembl Variant Effect Predictor," *Genome Biol.*, vol. 17, no. 1, p. 122, Dec. 2016.
16. I. A. Adzhubei *et al.*, "A method and server for predicting damaging missense mutations," *Nat. Methods*, vol. 7, no. 4, pp. 248–249, Apr. 2010.
17. M. J. Landrum and B. L. Kattman, "ClinVar at five years: Delivering on the promise," *Hum. Mutat.*, vol. 39, no. 11, pp. 1623–1630, Nov. 2018.
18. "UniProt: a worldwide hub of protein knowledge," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D506–D515, Jan. 2019.
19. M. a. Sukhai *et al.*, "A classification system for clinical relevance of somatic variants identified in molecular profiling of cancer," *Genet. Med.*, no. April, pp. 1–9, 2015.
20. H. Yan *et al.*, "IDH1 and IDH2 mutations in gliomas.," *N. Engl. J. Med.*, vol. 360, no. 8, pp. 765–73, Feb. 2009.