Data in Brief 35 (2021) 106780

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

A compiled dataset of molecular pathways associated with fusion genes identified in pediatric cancers

Neetha N. Vellichirammal*, Chittibabu Guda*

Department of Genetics, Cell Biology, and Anatomy, University of Nebraska Medical Center, Omaha, NE 68198, United States

ARTICLE INFO

Article history:
Received 16 November 2020
Revised 15 January 2021
Accepted 18 January 2021
Available online 21 January 2021

Keywords:
Fusion gene
Pediatric cancers
Molecular pathways
Gene networks

ABSTRACT

Fusion genes can serve as actionable biomarkers for diagnosis, prognosis or therapeutic stratification in the clinic. Pathways associated with fusion genes identified in different pediatric cancers are compiled in this article. Fusion genes reported in each cancer were collected using the PubMed search option with the keywords 'fusion transcript', 'fusion gene,' 'chromosomal translocation,' or 'DNA translocation' along with the corresponding pediatric cancer type. Research articles that identified fusion genes using conventional Fluorescence in situ hybridization (FISH) or quantitative real-time polymerase chain reaction (RT-PCR) methods or high-throughput RNA or DNA sequencing were included. The collected fusion gene data were compiled for each cancer and analyzed to identify their functions related to cancer and associated pathways using Ingenuity Pathway Analysis (IPA) and ClueGO software programs. Similarities in associated pathways across different cancers were also analyzed using IPA to identify commonly affected genes and pathways. This value-added and functionally annotated dataset will be an excellent resource for pediatric cancer researchers and clinicians interested in exploring fusion genes in different

DOI of original article: 10.1016/j.canlet.2020.11.015
* Corresponding authors.
E-mail addresses: n.nanothvellichiram@unmc.edu (N.N. Vellichirammal), babu.guda@unmc.edu (C. Guda).
Social media: (N.N. Vellichirammal), (C. Guda)

https://doi.org/10.1016/j.dib.2021.106780

cancers. This article is a companion article to 'Fusion genes as biomarkers in pediatric cancers: A review of the current state and applicability in diagnostics and personalized therapy'[1].

Specifications Table

Subject	Cancer Research
Specific subject area	Fusion genes in pediatric cancer
Type of data	Table Figure
How data were acquired	Electronic searches in PubMed, Google Scholar and Web of Science for studies reporting fusion genes in different pediatric cancers using the search terms' Fusion transcript', 'fusion gene', 'chromosomal translocation', 'DNA translocation' along with each cancer type.
Data format	Analyzed Filtered
Parameters for data collection	All reports that identified fusion genes in selected pediatric cancers were included in this study. Only articles in English were included in this dataset.
Description of data collection	Electronic searches in PubMed, Google Scholar and Web of Science for studies reporting fusion genes in different pediatric cancers using the search terms 'Fusion transcript', 'fusion gene', 'chromosomal translocation', or 'DNA translocation' along with each cancer type. Fusion genes identified in each pediatric cancer were extracted from the reports and compiled as a list.
Data source location	Primary data sources: Provided as a text file in the data repository (File 35, Primary_data_sources.txt) Repository name: Mendeley Data identification number: doi: 10.17632/24nnx5w5bj.1 Direct URL to data: https://data.mendeley.com/datasets/24nnx5w5bj/draft?a=81fcdde3-0e06-4aef-9188-704e42820614
Data accessibility	Representative Figures are provided with the article and the rest of the Figures and Tables are deposited in a data repository. Repository name: Mendeley Data identification number: doi: 10.17632/24nnx5w5bj.1 Direct URL to data: https://data.mendeley.com/datasets/24nnx5w5bj/draft?a=81fcdde3-0e06-4aef-9188-704e42820614
Related research article	N.N. Vellichirammal, N.K. Chaturvedi, S.S. Joshi, D.W. Coulter, C. Guda, Fusion genes as biomarkers in pediatric cancers: A review of the current state and applicability in diagnostics and personalized therapy (In press -Cancer Letters [1])

Value of the Data

- These data are compiled from published reports to date in pediatric cancers. Compiling such information into one accessible resource provides researchers easy access to publicly available data. Moreover, we were able to add value to the data by annotating the associated gene functions, gene interactions, enriched pathways and perform comparative analyses among different pediatric cancers.
- Pediatric cancer researchers investigating fusion genes in different types of cancers and clinicians interested in identifying clinically actionable targets will benefit from this data set.
- This data can be reused as a reference to investigate common and unique genes and pathways across different pediatric cancers and for designing experiments to identify new drug targets.

1. Data Description

The data article represents the compilation of detailed downstream functional analyses of genes participating in fusions associated with various pediatric cancers. Fusion genes participating in pediatric cancer were compiled for each cancer using standard methods for data collection and curation (Fig. 1) and subsequently analyzed to determine their biological functions related to cancer and associated pathways. This analysis performed with Ingenuity Pathway Analysis (IPA) and ClueGO software tools identified significant pathways associated with each cancer. Since the genes participating in fusions may be functionally disrupted, this analysis provides information on the pathways affected due to fusion in each cancer (Fig. 2, 3). IPA was also used for performing gene network analysis to identify significant biological interactions between genes participating in fusions (Fig. 4). The results from this network analysis inform us of the evidence-based interactions among different genes participating in fusions in each cancer. Finally, IPA was also used to identify similarities across pathways among various pediatric cancers to identify commonly affected functions (Fig. 5). The list of fusion genes in pediatric cancers and their implications in diagnosis and treatment protocols are described in the companion research article [1].

2. Supplementary files (deposited in repository)

File 1: IPA top canonical pathways generated from gene lists comprised of genes participating in fusions in T-cell acute lymphoblastic leukemia. IPA uses a knowledge database that is manually curated and comprehensive, representing biological interactions and functional annotations focused on genes, pathways, drugs, and diseases. This method also uses the pathway enrichment analysis as described in Fig. 2. Input for this analysis is the list of unique genes that form fusions in a cancer type. This input gene list is compared to the database. Right-Tailed Fisher's Exact Test is calculated that reflects the likelihood that the association or overlap between the genes and a specific pathway is due to random chance. This excel file represents IPA canonical pathways, their corresponding Fisher's Exact Test P-values, and the genes forming fusions in each pathway. Only P-values $\leq$0.05 multiple testing (Benjamini & Hochberg (BH) method) corrected is represented here. Files 2–8 were prepared using the same methodology.

File 2: IPA top canonical pathways generated from gene lists comprised of genes participating in fusions in B-cell acute lymphoblastic leukemia.

File 3: IPA top canonical pathways generated from gene lists comprised of genes participating in fusions in acute myeloid leukemia.

File 4: IPA top canonical pathways generated from gene lists comprised of genes participating in fusions in Ewing's Sarcoma.

File 5: IPA top canonical pathways generated from gene lists comprised of genes participating in fusions in Osteosarcoma.

File 6: IPA top canonical pathways generated from gene lists comprised of genes participating in fusions in Alveolar rhabdosarcoma

File 7: IPA top canonical pathways generated from gene lists comprised of genes participating in fusions in Medulloblastoma

File 8: IPA top canonical pathways generated from gene lists comprised of genes participating in fusions in Neuroblastoma.

File 9: Disease functions from IPA analysis significantly affected by genes involved in fusions in T-cell acute lymphoblastic leukemia. IPA analysis was performed as described in Fig. 2. Genes involved in fusions known to be associated with diseases are represented in this excel file. P-Value indicates Fishers Exact test P-value. Gene IDs and numbers of genes involved in fusions related to each condition are also shown. Files 10–16 were prepared using the same methodology.

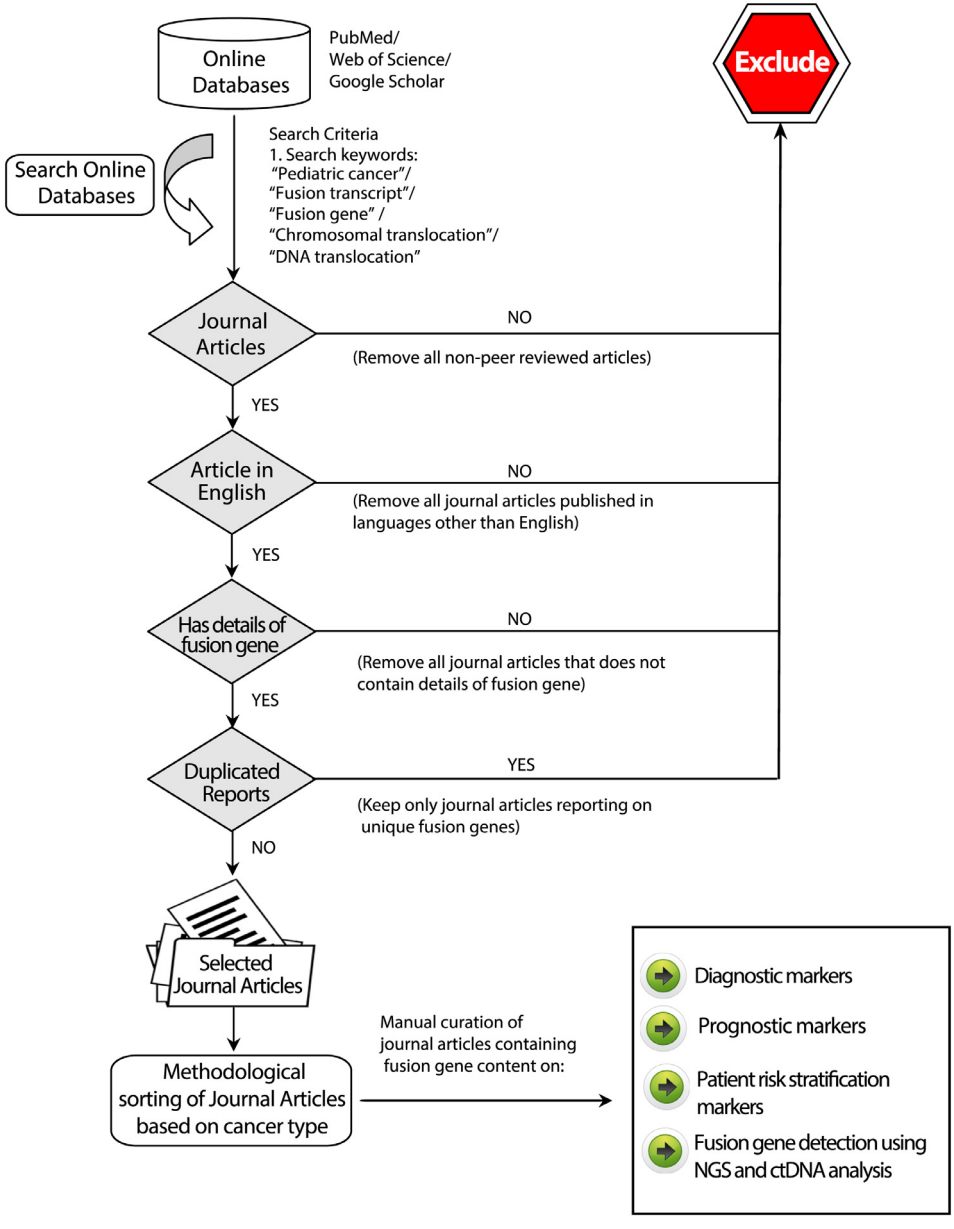

Fig. 1. Data collection and curation protocol used in the review. This figure explains the inclusion and exclusion criteria for selecting reports of fusion genes identified in pediatric cancers. Online database searches were conducted using the keywords 'Fusion transcript', 'fusion gene', 'chromosomal translocation', 'DNA translocation' along with each pediatric cancer type. All non-peer-reviewed articles and articles published in languages other than English were excluded manually. We only included reports that contained unique fusions identified in pediatric cancers and articles containing details of fusion genes, including methods used for fusion identification, cancer type associated with it, associated clinical data, and reports of fusion genes as diagnostic, prognostic, or patient risk stratification markers. We manually sorted these selected journal articles for review based on the information collected and the cancer type associated.

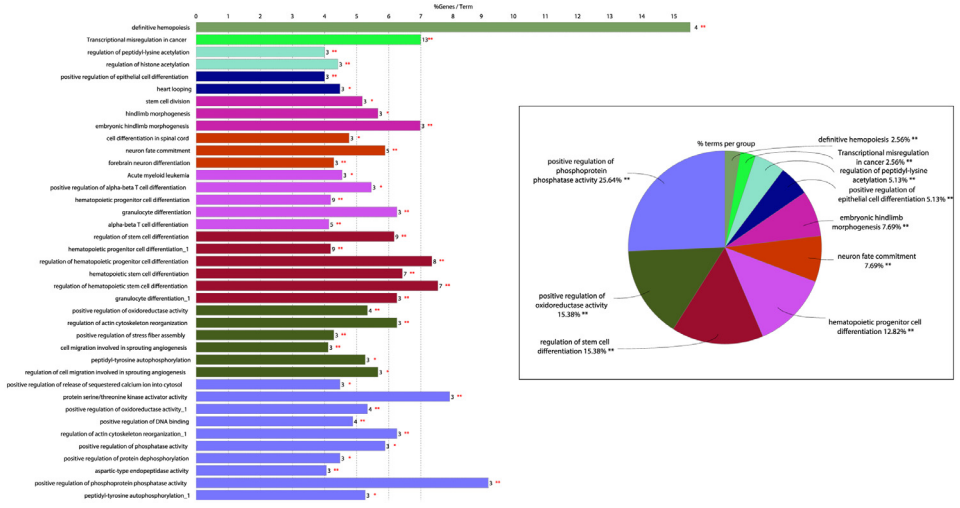

Fig. 2. Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis of genes participating in fusions in T-cell acute lymphoblastic leukemia identified using ClueGO. KEGG is a database resource to link genomic data to higher-order functional information. This database provides information about how genes are networked. Pathway enrichment analysis is performed to extract biologically meaningful results from a gene list. An input gene list is prepared by compiling all genes that participated in fusions in each cancer and then curating the list to represent unique genes. This input gene list is then compared against the KEGG database, and gene hits along with the corresponding pathways are retrieved. The significance of enrichment in each pathway is then calculated using Fisher's Exact Test with multiple corrections. The results of enrichment analysis are represented as a bar graph and pie chart. The bar graph represents KEGG pathway terms associated with target genes that are significantly enriched. The bars represent the percentage of genes associated with the terms, and the bar label represents the number of genes participating in fusions per term. The pie chart represents the overview of functional groups for target genes, ordered based on percentage terms per group.

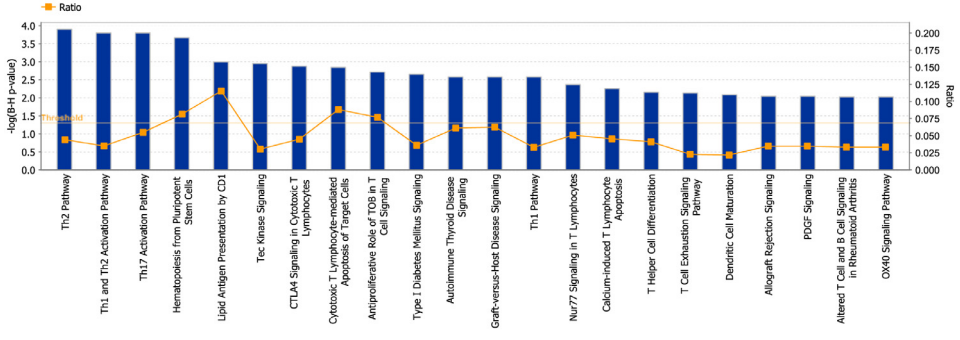

Fig. 3. Top canonical pathways generated from gene lists comprised of genes participating in fusions in T-cell acute lymphoblastic leukemia using Ingenuity Pathway Analysis (IPA) gene ontology analysis. IPA uses a knowledge database that is manually curated and comprehensive, representing biological interactions and functional annotations focused on genes, pathways, drugs, and diseases. This method also uses the pathway enrichment analysis as described in Fig. 2. Input for this analysis is the list of unique genes that form fusions in a cancer type. This input gene list is compared to the database. Right-Tailed Fisher's Exact Test is calculated that reflects the likelihood that the association or overlap between the genes and a specific pathway is due to random chance. The bar-chart represents the significance of gene enrichment for pathways. Only P-values ≤ 0.05 with multiple testing (Benjamini & Hochberg (BH) method) correction is represented here. The ratio represented here is the number of genes involved in fusions compared to the total number of genes in that canonical pathway in IPA. The threshold indicates the minimum significance level (scored as –log [P-value] from Fisher's exact test).

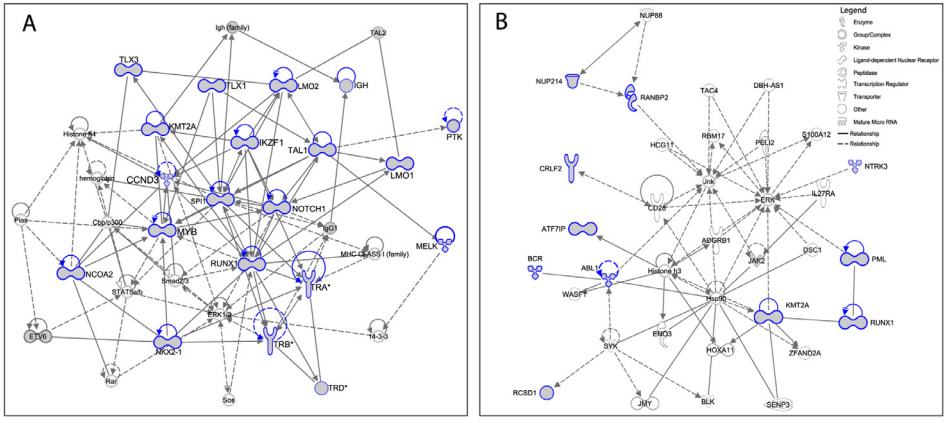

Fig. 4. This network represents the interactions among the top-ranked genes participating in fusions in (A)T-cell acute lymphoblastic leukemia and (B) B-cell acute lymphoblastic leukemia. Biological network analysis was performed using Ingenuity Pathway Analysis (IPA). Input for this analysis is the list of unique genes that form fusions in a cancer type. IPA compares this list to relevant networks, gene regulators, and mechanistic networks based on their connectivity score. Two genes are considered to be connected in a network if a path (edge) is present between them in the network. Arrows indicate gene/protein interactions of molecules (in gray) within genes in the pathway. The shape of the gene represents the molecule/functional class, which are nodes in the network, and the relationship between them is indicated by edges. Genes participating in fusions in each cancer is highlighted in blue in this figure. Biological network analysis was performed using Ingenuity Pathway Analysis (IPA).

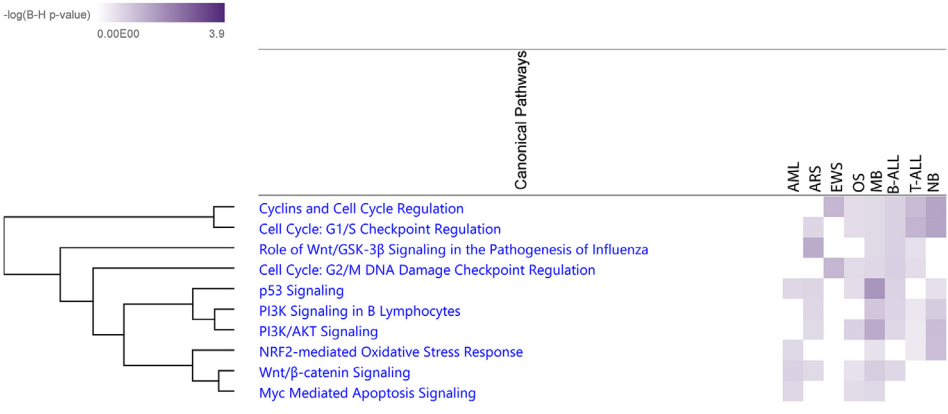

Fig. 5. Comparison of over-represented pathways enriched by genes participating in fusions across different pediatric cancers using Ingenuity Pathway Analysis. In this analysis, IPA compares common pathways over-represented in each of the cancer types studied. Over-representation of IPA canonical pathway annotation terms is represented here as Fisher's Exact Test P-value with Benjamin-Hochberg multiple testing correction. Only canonical pathways associated with nine oncogenic pathways are represented in this figure. Pathways are clustered using the hierarchical clustering method. T-ALL: T-cell acute lymphoblastic leukemia, B-ALL: B-cell acute lymphoblastic leukemia, AML: Acute myeloid leukemia, ARS: Alveolar rhabdosarcoma, EWS: Ewing's sarcoma, NB: neuroblastoma, OS: osteosarcoma, MB: medulloblastoma.

File 10: Disease functions from IPA analysis significantly affected by genes involved in fusions in B-cell acute lymphoblastic leukemia.

File 11: Disease functions from IPA analysis significantly affected by genes involved in fusions in Acute Myeloid Leukemia.

File 12: Disease functions from IPA analysis significantly affected by genes involved in fusions in Ewing's Sarcoma.

File 13: Disease functions from IPA analysis significantly affected by genes involved in fusions in Alveolar rhabdosarcoma.

File 14: Disease functions from IPA analysis significantly affected by genes involved in fusions in Osteosarcoma.

File 15: Disease functions from IPA analysis significantly affected by genes involved in fusions in Medulloblastoma.

File 16: Disease functions from IPA analysis significantly affected by genes involved in fusions in Neuroblastoma.

File 17: Comparison of over-represented pathways enriched by genes participating in fusions across different pediatric cancers using Ingenuity Pathway Analysis. In this analysis, IPA compares common pathways over-represented in each of the cancer types studied. Over-representation of IPA canonical pathway annotation terms is represented here as Fisher test p-value with Benjamin-Hochberg multiple testing correction (scored as –log [P-value] from Fisher's Exact Test).

File 18: This network represents the interactions among the top-ranked genes participating in fusions in (A) Ewing's Sarcoma and (B) Acute Myeloid Leukemia. Biological network analysis was performed using Ingenuity Pathway Analysis (IPA). Input for this analysis is the list of unique genes that form fusions in a cancer type. IPA compares this list to relevant networks, gene regulators, and mechanistic networks based on their connectivity score. Two genes are considered to be connected in a network if a path (edge) is present between them in the network. Arrows indicate protein-protein interactions of molecules (in gray) within genes in the pathway. The shape of the gene represents the molecule/functional class, which are nodes in the network, and the relationship between them is indicated by edges. Genes participating in fusions in each cancer is highlighted in blue in this figure. Legend is as same as in Fig. 2 in the main text. Files 19 and 20 were prepared using the same methodology.

File 19: This network represents the interactions among the top-ranked genes participating in fusions in (A)Osteosarcoma and (B) Alveolar rhabdosarcoma.

File 20: This network represents the interactions among the top-ranked genes participating in fusions in (A) Medulloblastoma and (B) Neuroblastoma.

File 21: Comparison of over-represented pathways enriched by genes participating in fusions across different pediatric cancers using Ingenuity Pathway Analysis. In this analysis, IPA compares common pathways over-represented in each of the cancer types studied. Over-representation of IPA canonical pathway annotation terms is represented here as Fisher test p-value with Benjamin-Hochberg multiple testing correction. Only canonical pathways associated with important cancer signaling are represented in this figure. Pathways are clustered using the hierarchical clustering method. T-ALL: T-cell acute lymphoblastic leukemia, B-ALL: B-cell acute lymphoblastic leukemia, AML: Acute myeloid leukemia, ARS: Alveolar rhabdosarcoma, EWS: Ewing's sarcoma, NB: neuroblastoma, OS: osteosarcoma, MB: medulloblastoma.

File 22: Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis of genes participating in fusions in B-cell acute lymphoblastic leukemia identified using ClueGO. KEGG is a database resource to link genomic data to higher-order functional information. This database provides information about how genes are networked. Pathway enrichment analysis is performed to extract biologically meaningful results from a gene list. An input gene list is prepared by compiling all genes that participated in fusions in each cancer and then curating the list to represent unique genes. This input gene list is then compared against the KEGG database, and gene hits along with the corresponding pathways are retrieved. The significance of enrichment in each pathway is then calculated using Fisher's Exact Test with multiple corrections. The results of enrichment analysis are represented as a bar graph and pie chart. The bar graph represents KEGG pathway terms associated with target genes that are significantly enriched. The bars represent the percentage of genes associated with the terms, and the bar label represents the number of differentially expressed genes per term. The pie chart illustrates the overview of functional groups for target genes, ordered based on percentage terms per group. Files 23–28 were prepared using the same methodology.

File 23: Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis of genes participating in fusions in acute myeloid leukemia identified using ClueGO.

File 24: Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis of genes participating in fusions in Alveolar rhabdosarcoma identified using ClueGO.

File 25: Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis of genes participating in fusions in Ewing's sarcoma identified using ClueGO.

File 26: Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis of genes participating in fusions in neuroblastoma identified using ClueGO.

File 27: Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis of genes participating in fusions in osteosarcoma identified using ClueGO.

File 28: Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis of genes participating in fusions in medulloblastoma identified using ClueGO.

File 29: Top canonical pathways generated from gene lists comprised of genes participating in fusions in B-cell acute lymphoblastic leukemia using Ingenuity Pathway Analysis (IPA) gene ontology analysis. IPA uses a knowledge database that is manually curated and comprehensive, representing biological interactions and functional annotations focused on genes, pathways, drugs, and diseases. This method also uses the pathway enrichment analysis as described in Fig. 2. Input for this analysis is the list of unique genes that form fusions in a cancer type. This input gene list is compared to the database. Right-Tailed Fisher's Exact test is calculated that tests the likelihood that the association or overlap between the genes and a specific pathway is due to random chance. The bar-chart represents the significance of gene enrichment for pathways. Only P-values $\leq$0.05 with multiple testing (Benjamini & Hochberg (BH) method) correction is represented here. The ratio represented here is the number of genes involved in fusions compared to the total number of genes in that canonical pathway in IPA. The threshold indicates the minimum significance level (scored as –log [P-value] from Fisher's Exact Test). Files 30–34 were prepared using the same methodology.

File 30: Top canonical pathways generated from gene lists comprised of genes participating in fusions in Acute myeloid leukemia using Ingenuity Pathway Analysis (IPA) gene ontology analysis.

File 31: Top canonical pathways generated from gene lists comprised of genes participating in fusions in Alveolar rhabdosarcoma using Ingenuity Pathway Analysis (IPA) gene ontology analysis.

File 32: Top canonical pathways generated from gene lists comprised of genes participating in fusions in Ewing's sarcoma using Ingenuity Pathway Analysis (IPA) gene ontology analysis.

File 33: Top canonical pathways generated from gene lists comprised of genes participating in fusions in neuroblastoma using Ingenuity Pathway Analysis (IPA) gene ontology analysis.

File 34: Top canonical pathways generated from gene lists comprised of genes participating in fusions in osteosarcoma using Ingenuity Pathway Analysis (IPA) gene ontology analysis.

File 35: Top canonical pathways generated from gene lists comprised of genes participating in fusions in medulloblastoma using Ingenuity Pathway Analysis (IPA) gene ontology analysis.

File 36: List of primary data sources selected for this review. This text file lists all the peer-reviewed journal articles included in the review that passed the criteria discussed in Fig. 1.

3. Experimental Design, Materials and Methods

3.1. Data collection and curation

Data were collected from studies included in the companion review paper [1]. Relevant articles for this study were selected through a literature search that was carried out till July 2020 through PubMed, Google Scholar, and Web of Science. For each pediatric cancer, we searched terms including 'Fusion gene,' 'Fusion transcript,' 'DNA translocation,' along with the associated cancer type to retrieve published articles related to this field. Eligible articles included studies

that have reported gene fusions in cancer either through traditional methods or through high–throughput sequencing. We included only the reports that were published in the English language. We excluded reports that were not peer-reviewed. Fig. 1 explains in detail the inclusion and exclusion criteria for collecting reports in this review. Previously published review articles containing information on reported fusions in pediatric cancer were also analyzed to identify original reports. For each selected peer-reviewed journal article, we extracted the information pertaining to methods used for fusion identification, cancer type associated with it, clinical data, and reports of fusion genes as diagnostic, prognostic, or patient risk stratification markers. This information was manually extracted from journal reports and compiled for each fusion gene in each pediatric cancer studied for further analysis. The list of fusion genes was extracted from supplementary files in articles that used Next Generation Sequencing methods for identifying fusion genes. We next analyzed the compiled fusion gene list associated with each cancer using pathway enrichment analysis for biological interpretation.

3.2. Pathway and gene network analysis using IPA

The list of genes participating in fusions compiled in each cancer was evaluated further using Ingenuity Pathway Analysis(IPA)[QIAGEN(RedwoodCity, CA)] to identify their functional significance. IPA implements Fisher's exact test with multiple testing correction (Benjamini & Hochberg (BH) method) [2] to identify enriched pathways containing genes of interest. The genes analyzed in IPA using core analysis were followed by comparative analysis among different pediatric cancers. Comparative analysis identified genes participating in fusions that affected common pathways across multiple cancers. Selected canonical oncogene pathways were extracted and represented in Fig. 5. Data from canonical pathways enriched in each cancer were extracted and presented as Tables (deposited into the data repository). Genes participating in the top molecular networks in each cancer were also identified and presented in the Figures. IPA generates networks of genes based on their known interactions from the Ingenuity Pathway Knowledge Base.

3.3. Gene ontology (GO) enrichment analysis using ClueGO

Gene ontology analysis of genes participating in fusions in each cancer was performed using ClueGO software (v) [3], a Cytoscape 3.7.1 plug-in. Significant KEGG pathways in each data set were identified, comparing the ratio of target genes identified in each pathway to the total number of genes within the pathway. The statistical test used to determine the enrichment score for KEGG pathways was based on a right-sided hypergeometric distribution with multiple testing correction (Benjamini & Hochberg (BH) method) [2].

Ethics Statement

Not applicable

CRediT Author Statement

Neetha Nanoth Vellichirammal: Conceptualization, methodology, data curation, writing-original draft preparation; **Chittibabu Guda:** Conceptualization, Writing- Reviewing, and Editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships which have or could be perceived to have influenced the work reported in this article.

Data Availability

Dataset of pathways associated with fusion genes identified in pediatric cancer (Original data) (Mendeley Data)

Acknowledgments

This work has been supported by the National Institutes of Health awards [5P20GM103427, 1P30GM127200, 5P30CA036727] and National Science Foundation's EPSCoR Award [Grant No. OIA-1557417] to CG, and the Fred & Pamela Buffett Cancer Center, which is supported by the National Cancer Institute under award number P30 CA036727, in conjunction with the UNMC/Children's Hospital & Medical Center Child Health Research Institute Pediatric Cancer Research Group to NNV. The authors are grateful to the Bioinformatics and Systems Biology Core at the University of Nebraska Medical Center (UNMC) for providing access to the computational infrastructure. Authors also acknowledge the Holland Computing Center of the University of Nebraska-Lincoln for high-performance computational resources.

References

[1] N.N. Vellichirammal, N.K. Chaturvedi, S.S. Joshi, D.W. Coulter, C. Guda, Fusion genes as biomarkers in pediatric cancers: a review of the current state and applicability in diagnostics and personalized therapy, Cancer Lett. 499 (2021) 24–38, doi:10.1016/j.canlet.2020.11.015.

[2] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, J. R. Stat. Soc. Ser. B. 57 (1995) 289–300, doi:10.1111/j.2517-6161.1995.tb02031.x.

[3] G. Bindea, B. Mlecnik, H. Hackl, P. Charoentong, M. Tosolini, A. Kirilovsky, W.H. Fridman, F. Pagès, Z. Trajanoski, J. Galon, ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks, Bioinformatics 25 (2009) 1091–1093, doi:10.1093/bioinformatics/btp101.