3 Biotech (2024) 14:282
https://doi.org/10.1007/s13205-024-04132-1

ORIGINAL ARTICLE

PFusionDB: a comprehensive database of plant-specific fusion transcripts

Ajay Arya[1] · Simran Arora[1] · Fiza Hamid[1] · Shailesh Kumar[1]

Received: 11 June 2024 / Accepted: 20 October 2024 / Published online: 28 October 2024

Abstract

Fusion transcripts (FTs) are well known cancer biomarkers, relatively understudied in plants. Here, we developed PFusionDB (www.nipgr.ac.in/PFusionDB), a novel plant-specific fusion-transcript database. It is a comprehensive repository of 80,170, 39,108, 83,330, and 11,500 unique fusions detected in 1280, 637, 697, and 181 RNA-Seq samples of *Arabidopsis thaliana*, *Oryza sativa japonica*, *Oryza sativa indica*, and *Cicer arietinum* respectively. Here, a total of 76,599 (*Arabidopsis thaliana*), 35,480 (*Oryza sativa japonica*), 72,099 (*Oryza sativa indica*), and 9524 (*Cicer arietinum*) fusion transcripts are non-recurrent i.e., only found in one sample. Identification of FTs was performed by using a total of five tools viz. EricScript-Plants, STAR-Fusion, TrinityFusion, SQUID, and MapSplice. At PFusionDB, available fundamental details of fusion events includes the information of parental genes, junction sequence, expression levels of fusion transcripts, breakpoint coordinates, strand information, tissue type, treatment information, fusion type, PFusionDB ID, and Sequence Read Archive (SRA) ID. Further, two search modules: 'Simple Search' and 'Advanced Search', along with a 'Browse' option to data download, are present for the ease of users. Three distinct modules viz. 'BLASTN', 'SW Align', and 'Mapping' are also available for efficient query sequence mapping and alignment to FTs. PFusionDB serves as a crucial resource for delving into the intricate world of fusion transcript in plants, providing researchers with a foundation for further exploration and analysis. Database URL: www.nipgr.ac.in/PFusionDB.

Keywords Fusion transcripts · Chimeric RNAs · Fusion tools · Plant database · Genome regulation

Introduction

Fusion transcripts (FTs), where two separate transcripts merge to form a hybrid RNA molecule, have emerged as crucial players in molecular biology (Kumar et al. 2016; Singh et al. 2019). These FTs or chimeric RNAs can be generated by two possible mechanisms: gene-fusion events at the DNA level, and the fusion of two transcripts at the RNA level (Kumar et al. 2016). At the DNA level, gene fusion can originate as a result of deletions, inversions, duplication, transversions, or translocations, which further transcribe to generate FTs (Li et al. 2009). On the other hand, different FTs generated at the RNA level such as transcription-induced chimeras or cis-FTs (Akiva et al. 2006), tandem chimeric RNAs (Greger et al. 2014), transcription-induced gene fusions (TIGF) (Mertens et al. 2015), or cis-SAGes (Annala et al. 2013) through various mechanisms such as intergenic cis-splicing, trans-splicing, and transcriptional read-through processes (Gingeras 2009; Frenkel-Morgenstern et al. 2012; Qin et al. 2015; Dorney et al. 2023). Notably, they can encode novel fusion proteins or act as long non-coding RNAs, which regulate the overall functionality of various organisms (Latysheva and Babu 2016; Jia et al. 2016; Han et al. 2019; Mukherjee et al. 2021).

These fusion events have been originally recognized as valuable biomarkers and therapeutic targets in cancer, where they can act as potent oncogenic drivers (Varley et al. 2014). According to Gao et al., fusion events contribute to 16.5% of human cancers, and are the sole driver in more than 1% of cases (Gao et al. 2018). For instance, the TMPRSS2-ERG fusions in prostate adenocarcinoma (St. John et al. 2012) and BCR-ABL fusion transcript (Tkachuk et al. 1990), serves as

[$] Ajay Arya, Simran Arora, and Fiza Hamid contributed equally to this work.

✉ Shailesh Kumar
shailesh@nipgr.ac.in

[1] Bioinformatics Lab, National Institute of Plant Genome Research, Aruna Asaf Ali Marg, New Delhi 110067, India

well-known examples. Beyond oncology, the presence and impact of FTs are also being explored in other eukaryotic organisms (Rogers et al. 2010; Fu et al. 2010; Chwalenia et al. 2017), including normal and aberrant human cellular processes (Babiceanu et al. 2016; Singh et al. 2020), and plants (Koller et al. 1987; Kawasaki et al. 1999; Chen et al. 2017).

Plant genes frequently undergo fusion events, thereby increasing the complexity of transcriptome (Zhang et al. 2010; Zhou et al. 2022a; Parakkunnel et al. 2022). In rice, about half of the novel genes present on the short arm of chromosome 3 have originated from chimeric events, and 6.8% of the spliced isoforms were discovered as FTs (Zhang et al. 2013). Furthermore, recent advancements in sequencing technologies have improved the identification and study of fusion genes in a variety of plants such as *Arabidopsis* (Zhang et al. 2020), *Oryza* (Zhang et al. 2010; Chen et al. 2017; Zhou and Zhang 2019; Hasan et al. 2022; Zhou et al. 2022b), *Zea mays* (Zhou et al. 2022a), *Solanum lycopersicum* (Chen et al. 1997; Kim et al. 2001) and *Trifolium pratense* L. (Chao et al. 2018). Furthermore, Qiao et al. also conducted an in-silico study to understand the secondary metabolite pathways (biosynthesis of theanine and caffeine) in tea plant (*Camellia sinensis*) and reported 5777 FTs (Qiao et al. 2019). These fusion events, generated from genomic rearrangements or post-transcriptional splicing events, possess the potential to exert profound effects on plant metabolism (Hagel and Facchini 2017), stress responses (Zhou et al. 2022a), gene regulation (Mukherjee et al. 2021), and adaptation mechanisms (Zhou et al. 2022b). Consequently, a critical necessity arises for a centralized repository that systematically compiles, annotates, and analyzes FTs, offering a comprehensive resource for researchers navigating the intricacies of plant transcriptomics.

This study presents a comprehensive database named ‘PFusionDB’ (www.nipgr.ac.in/PFusionDB), exclusively dedicated to plant FTs. Here, we have identified and characterized the FTs in four plants species viz. *Arabidopsis thaliana*, *Oryza sativa japonica*, *Oryza sativa indica*, and *Cicer arietinum*. This database provides comprehensive features of all FTs predicted in the RNA-Seq samples of aforementioned plants using a systematic bioinformatics approach. Apart from basic fusion-transcript information, details like the sequence of FTs, gene ontology of fusion parental genes, tissue, and condition-specific expression of FTs can also be obtained, thus making the database more informative. This database provides various search and browsing options, tools for user-submitted query sequences, and a data download facility. A proper help page is also provided for each module of the PFusionDB for its adequate utilization. It is an invaluable resource for plant research, enabling a deeper understanding of fusion events. It can unravel previously undiscovered regulatory networks and adaptive mechanisms employed by plants in response to environmental cues. The overall architecture and different features of the PFusionDB are illustrated in Fig. 1.

Materials and methods

Data retrieval

All reference genomes, transcriptomes, and annotation files for *Arabidopsis thaliana* (TAIR10.1), *Oryza sativa japonica* (IRGSP-1.0), and *Oryza sativa indica* (ASM465v1) were downloaded from Ensembl Plants (https://plants.ensembl.org/), while those for *Cicer arietinum* (ASM33114v1) were sourced from the genome database (https://www.ncbi.nlm.nih.gov/) of NCBI. In addition, paired-end RNA-Seq samples of non-mutant (Supplementary Table 1) were downloaded from a diverse range of independent experiments across four species using NCBI SRA Toolkit (v2.11.0) (https://www.ncbi.nlm.nih.gov/sra/docs/toolkitsoft/).

Identification of fusion transcripts

To identify FTs in RNA-Seq dataset, we utilized five fusion detection algorithms viz. EricScript-Plants (v0.5.5b) (https://github.com/asherkhb/EricScript-Plants), STAR-Fusion (1.10.0) (Haas et al. 2019), TrinityFusion (v0.3.5) (Haas et al. 2019), SQUID (v1.5) (Ma et al. 2018), and MapSplice (v2.2.1) (Wang et al. 2010). These tools were selected because of their effectiveness in handling plant datasets. Using Ensembl genome files, EricScript-Plants autonomously built its genome database; however, this method did not apply to *Cicer arietinum* because this plant do not have any relevant genome files at Ensembl database. Therefore, to overcome this limitation, we chose to utilize the MapSplice tool specifically for the analysis of *Cicer arietinum*. The initial step involved preprocessing raw sequencing reads to remove low-quality bases and adaptor sequences using the fastp tool (v0.21.0) (Chen et al. 2018). Subsequently, each fusion detection tool was utilized with default parameters to detect FTs in the paired-end RNA-Seq datasets.

PFusionDB web interface development

After collecting and compiling all information on fusion events, the database was constructed on an Apache Hypertext Transfer Protocol (HTTP) Server, and MySQL for data storage, querying, and management. At the same time, HyperText Markup Language (HTML), Cascading Style Sheets (CSS), JavaScript, and Bootstrap were employed to develop a user-friendly web interface. In addition, PHP and PERL scripts were used to create database interfaces and common gateway interfaces.

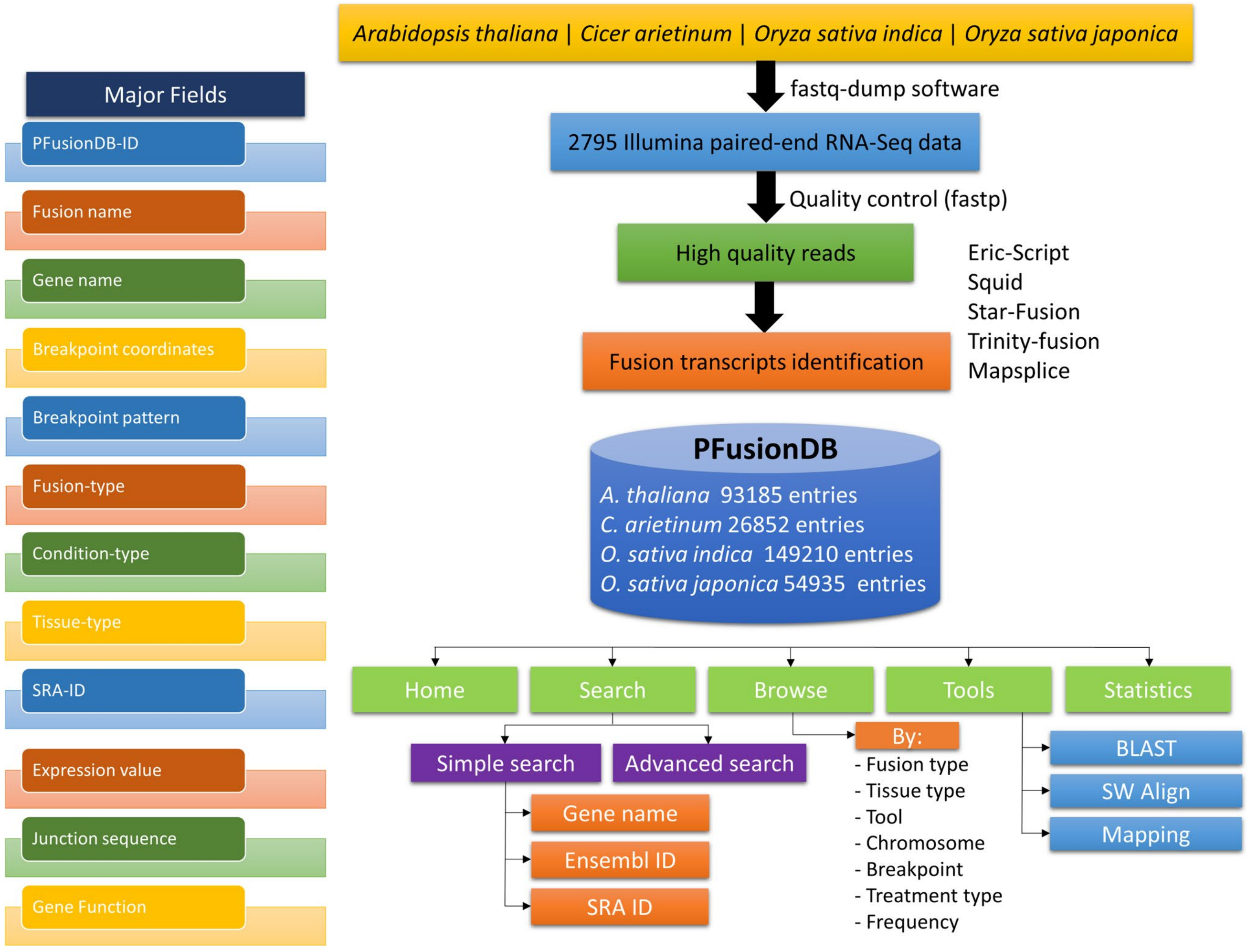

Fig. 1 Schematic representation of methodology for the identification of fusion transcripts in RNA-Seq datasets, along with an overview of various features, and different modules available at PFusionDB database

PFusionDB web interface architecture and features

At PFusionDB, detailed and comprehensive information is incorporated for each fusion-transcript entry. In addition to the basic information including parental genes involved in fusion formation, breakpoint location, source organism, tissue, condition, fusion type, and breakpoint type; users may find additional information like breakpoint sequence, detailed information of parental genes, and specific expression patterns of FTs in different tissues and conditions. PFusionDB web interface is user-friendly, and includes multiple options to easily navigate to the desired page.

Search

The fusion transcript data are systematically organized across different layers to facilitate easy querying and retrieval through both simple and advanced search options. Initially, users are required to choose the species of their interest, which directs them to a basic search page. Here, users can input keywords such as gene name, SRA-ID, and PFusionDB ID to initiate a search. If a user wants to search these keywords for specific conditions like the transcripts generated by a specific tool or transcript present on some chromosome location that can be selected from the drop-down menu beside the search bar.

If user desires a more refined search with additional criteria, the 'Advanced Search' option is available to facilitate a thorough and filtered query. Users can input gene names and SRA IDs and apply filters using conditions such as 'AND' or'OR.' The defined conditions allow for the inclusion or exclusion of selected fields (Tool, Tissue, Treatment, Fusion type, Chromosome, and Breakpoint pattern), tailoring the displayed results accordingly. Within the second layer, which serves as a detailed page, users can obtain a more comprehensive set of information for their chosen fusion transcript by selecting the specific one of interest. In the third layer of the database, there are links established to interconnect fusion events with their associated genes, along with links that offer enhanced insights and biological context.

Browse

For seamless data exploration, the 'Browse' option directs users to a browsing page where they can explore the data based on specific types, including fusion, tissue, tool, chromosome, breakpoint pattern, treatment type, and even highly frequent transcript fusions under different conditions.

Tools

The 'Tools' section within PFusionDB provides users with three unique tools i.e., 'SW Align, 'BLAST', and 'Mapping', each designed to extract valuable insights into FTs (Smith and Waterman 1981; Altschul et al. 1990). The 'SW Align' allows users to align their query sequences with junction sequences of chimeric transcripts stored in the PFusionDB database. It utilizes the 'WATER' utility from the EMBOSS-6.6.0 package, which employs the Smith–Waterman Algorithm (Smith and Waterman 1981). This feature aids users to identify and characterize their specific sequence of interest. The 'Mapping' feature enables users to align fusion-junction sequences from the PFusionDB database with their query gene sequences by utilizing 'BLASTn' module of the BLAST software package. It displays only those PFusionDB sequences that have a 100% match with the user's query sequences.

The 'BLAST' tool in PFusionDB utilizes BLASTn (Chen et al. 2015) to identify regions of similarity between user-provided queries with fusion-junction sequences kept in the database. Users can adjust the E-value for more precise results. Significant alignments link to the corresponding PFusionDB sequence IDs, providing access to detailed information about these sequences.

Other PFusionDB features

In addition to the key features of the database, several additional features have been added to make the database more informative and user-interactive. In the 'Statistics' segment, a visual representation is provided for the total and unique FTs across four distinct species within PFusionDB. Users can examine a bar chart showcasing the distribution of FTs identified in SRA samples, categorized by different tools used for each species. In addition, users can access pie charts illustrating the proportions of identified FTs in diverse tissues and conditions, along with classifications based on their breakpoint pattern and fusion type. The 'Help/Guide' section serves as a valuable reference for users to understand the PFusionDB database and enhance their effective utilization.

Result and discussion

We collected 2795 Illumina paired-end RNA-Seq datasets from NCBI SRA with 1280, 637, 697, and 181 FASTQ files of *Arabidopsis thaliana*, *Oryza* sativa *japonica*, *Oryza sativa indica*, and *Cicer arietinum,* respectively. The number of samples analysed and the output generated by each tool across different plant species are shown in Fig. 2.

A total of 3,24,182 fusion entries are included in our database which represents 93,185, 54,935, 149,210, and 26,852 FTs detected in *Arabidopsis thaliana*, *Oryza* sativa *japonica*, *Oryza sativa indica*, and *Cicer arietinum,* respectively. A total of 76,599, 35,480, 72,099, and 9524 FTs were non-recurrent i.e., only found in one sample of *Arabidopsis thaliana, Oryza sativa japonica, Oryza sativa indica,* and *Cicer arietinum,* respectively.

Fusions were classified based on generation mechanisms e.g., intra-chromosomal (proximal and distal-type) and inter-chromosomal. The graphical representation of their proportion in different species is provided in the Statistics section of webpage of PFusionDB database. We noticed inter-chromosomal FTs are more abundant than intra-chromosomal, across all four species which indicates that fusion formation is not distance-dependent and no bias was observed in fusion formation between closely located genes. The FTs were also classified on the basis of their breakpoint pattern relative to the parental gene structure, whether the junction is present on the exon boundaries, within the exon, or in the UTR. It was observed that most of the FTs are of MM type, followed by EE and UU. A graphical representation of their distribution is available in the Statistics section of PFusionDB.

The tissue-wise and condition-wise distribution of the FTs was also carried out for each species. All the collected samples were categorized based on their tissue origin and condition, and it was observed that most fusions were found in leaf, root, and seedling among all the species. A high number of FTs was observed under various stresses such as biotic and abiotic stresses, drought, heat, and salt, which may suggest their role in stress responses (Fig. 3). Hence, this resource also offers to explore tissue and condition-specific expression of a fusion transcript. The top 10 most frequently occurring FTs in each species along with their corresponding frequencies are shown in Supplementary Table 2.

To illustrate the application of PFusionDB, we have selected a highly recurrent fusion transcript i.e., PsbE_PsbF, in which the parental genes are situated within the chloroplast genome and govern the synthesis of the alpha and beta subunits of the cytochrome b559 (cyt b559) protein. In PFusionDB, this fusion is categorized as an E/E

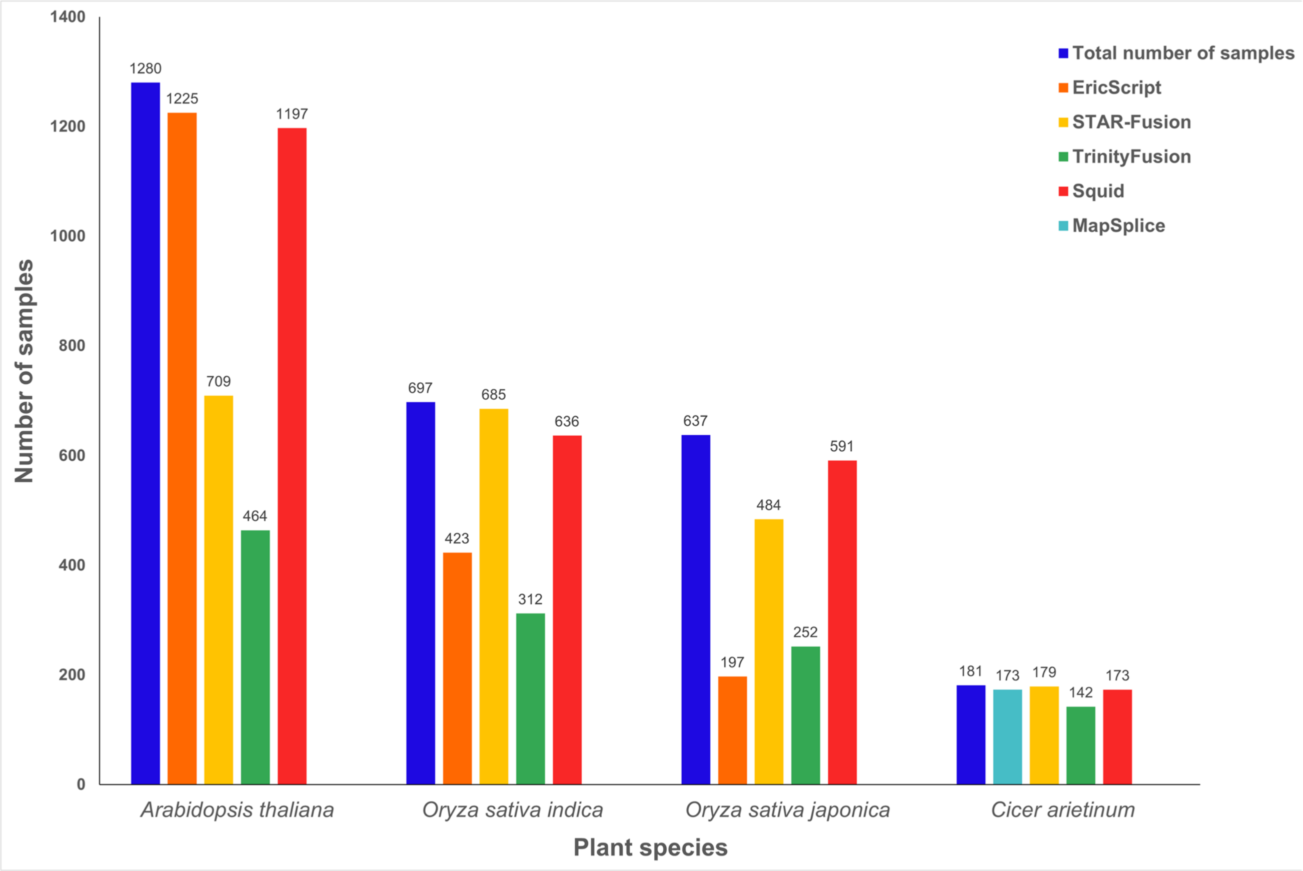

Fig. 2 A bar plot representation of total number of samples selected for the identification of fusion transcripts (blue), and the number of samples on which different tools produced the output in form of fusion transcripts

type, indicating the fusion formation at the exon boundaries of the two genes. As a result, this fusion maintains the open reading frames of the parent genes, suggesting the potential translation into a functional fusion protein. Our analysis revealed the widespread presence of this fusion transcript in over 70 samples, spanning diverse conditions and tissue types. This prevalence underscores the significance of the PsbE_PsbF fusion in various biological contexts. Furthermore, we extended our investigation to search for homologous fusion events in *Oryza sativa japonica*, where homologs of PsbE and PsbF were explored. In this database, we have also reported a fusion event involving the genes Os03g0659233, encodes the alpha subunit, and Os03g0659266, encodes the beta subunit of Cytochrome b559. This discovery adds valuable information to the database, expanding our understanding of similar fusion events in different plant species.

The advancement in sequencing technologies has led to the availability of vast amounts of RNA-Seq data that can be exploited for the investigation of new transcripts such as FTs and further study of the influence of gene-fusion products on plant growth, development, and physiology and it will be helpful for better understanding of genes regulation and genome evolution. Even though a large number of FTs have been identified in various plants over the past few years using RNA-Seq data, a thorough analysis, validation, and functional characterization of these transcripts and their encoded products remain unadressed. This database will be helpful for researchers working in this unexplored domain of transcriptome in plants. Although we have detected the FTs in four plant species by analyzing available RNA-Seq data produced by NGS technologies, experimental validation is further required for their functional characterization.

Conclusion

Recent advancements in sequencing technologies have broadened our understanding of transcriptome diversity beyond alternative splicing. One intriguing aspect of this exploration is the investigation of fusion events, which introduce novel transcripts and peptides without complicating the genome's structure. Recent studies showed the involvement of fusion transcripts in important biological processes

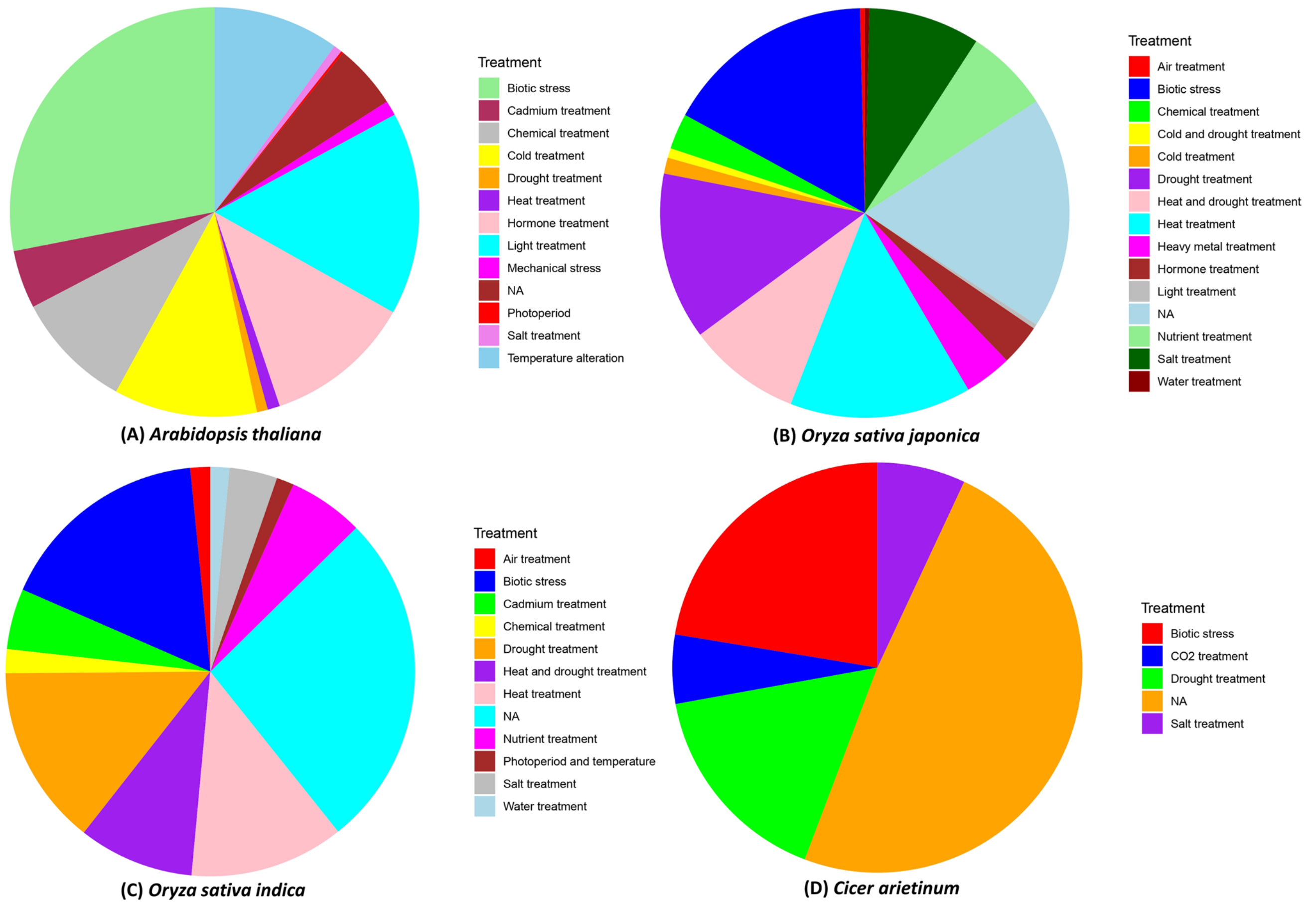

Fig. 3 Each pie chart representing the number of fusion transcripts identified under different biotic and abiotic stress conditions for each plant species. This representation is based on the fusion transcripts identified in RNA-Seq samples across the different conditions for each plant species.

such as metabolism, stress response, and gene regulation. Recognizing the need for tools and resources to support fusion-transcript research in plants, we embarked on a study to develop a comprehensive repository titled PFusionDB. By leveraging RNA-sequencing datasets, we aim to provide researchers with a comprehensive and user-friendly platform for exploring fusion transcripts across multiple plant species. This initiative seeks to empower the research community and facilitate further discoveries in fusion-transcript biology. We emphasize the importance of experimental validation to fully understand the functional roles of fusion transcripts. Our commitment includes regularly updating the database with new fusion-transcript data to ensure its relevance and usefulness. In the future, we anticipate that analyzing additional RNA-Seq datasets from diverse plant species will enhance our understanding of fusion-transcript dynamics across the plant kingdom.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s13205-024-04132-1.

Acknowledgements The authors are thankful to the Department of Biotechnology (DBT)-eLibrary Consortium, India, for providing access to e-resources. SA and FH are thankful to the University Grants Commission (UGC), India, and the Department of Biotechnology (DBT), India, respectively, for research fellowships. The authors acknowledge the Computational Biology & Bioinformatics Facility (CBBF) of the National Institute of Plant Genome Research (NIPGR).

Author contributions Conceptualization: SK. Data curation: SA and FH. Formal analysis: SA and FH. Web-interface development: AA. Funding acquisition, project administration, resources, and supervision: SK. Roles/writing—original draft: AA, SA, and FH; and writing—review and editing: SK.

Funding This research is supported by the BT/PR40146/BTIS/137/4/2020 project grant from the Department of Biotechnology (DBT), Government of India, and EEQ/2019/000231 Science and Engineering Research Board (SERB), Department of Science and Technology, Government of India.

Availability of data and materials All the fusion data generated is available in a database named 'PFusionDB', www.nipgr.ac.in/PFusionDB. The study's scripts and code used for analysis are available in the GitHub repository (https://github.com/skbinfo/PFusion).

Declarations

Conflict of interest The authors have no conflicts of interest to declare.

References

Akiva P, Toporik A, Edelheit S et al (2006) Transcription-mediated gene fusion in the human genome. Genome Res 16:30. https://doi.org/10.1101/GR.4137606

Altschul SF, Gish W, Miller W et al (1990) Basic local alignment search tool. J Mol Biol 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Annala MJ, Parker BC, Zhang W, Nykter M (2013) Fusion genes and their discovery using high throughput sequencing. Cancer Lett 340:192–200. https://doi.org/10.1016/J.CANLET.2013.01.011

Babiceanu M, Qin F, Xie Z et al (2016) Recurrent chimeric fusion RNAs in non-cancer tissues and cells. Nucleic Acids Res 44:2859–2872. https://doi.org/10.1093/NAR/GKW032

Chao Y, Yuan J, Li S et al (2018) Analysis of transcripts and splice isoforms in red clover (Trifolium pratense L.) by single-molecule long-read sequencing. BMC Plant Biol 18:1–12. https://doi.org/10.1186/S12870-018-1534-8/FIGURES/7

Chen JJ, Janssen BJ, Williams A, Sinha N (1997) A gene fusion at a homeobox locus: alterations in leaf shape and implications for morphological evolution. Plant Cell 9:1289. https://doi.org/10.1105/TPC.9.8.1289

Chen Y, Ye W, Zhang Y, Xu Y (2015) High speed BLASTN: an accelerated MegaBLAST search tool. Nucleic Acids Res 43:7762–7768. https://doi.org/10.1093/NAR/GKV784

Chen H, Tang Y, Liu J et al (2017) Emergence of a novel chimeric gene underlying grain number in rice. Genetics 205:993–1002. https://doi.org/10.1534/GENETICS.116.188201/-/DC1

Chen S, Zhou Y, Chen Y, Gu J (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34:i884–i890. https://doi.org/10.1093/BIOINFORMATICS/BTY560

Chwalenia K, Facemire L, Li H (2017) Chimeric RNAs in cancer and normal physiology. Wiley Interdiscip Rev RNA. https://doi.org/10.1002/WRNA.1427

Dorney R, Dhungel BP, Rasko JEJ et al (2023) Recent advances in cancer fusion transcript detection. Brief Bioinform 24:1–12. https://doi.org/10.1093/BIB/BBAC519

Frenkel-Morgenstern M, Lacroix V, Ezkurdia I et al (2012) Chimeras taking shape: potential functions of proteins encoded by chimeric RNA transcripts. Genome Res 22:1231–1242. https://doi.org/10.1101/GR.130062.111

Fu B, Chen M, Zou M et al (2010) The rapid generation of chimerical genes expanding protein diversity in zebrafish. BMC Genomics. https://doi.org/10.1186/1471-2164-11-657

Gao Q, Liang WW, Foltz SM et al (2018) Driver fusions and their implications in the development and treatment of human cancers. Cell Rep 23:227-238.e3. https://doi.org/10.1016/J.CELREP.2018.03.050

Gingeras TR (2009) Implications of chimaeric non-co-linear transcripts. Nature 461:206–211. https://doi.org/10.1038/nature08452

Greger L, Su J, Rung J et al (2014) Tandem RNA chimeras contribute to transcriptome diversity in human population and are associated with intronic genetic variants. PLoS ONE 9:e104567. https://doi.org/10.1371/JOURNAL.PONE.0104567

Haas BJ, Dobin A, Li B et al (2019) Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. Genome Biol 20:1–16. https://doi.org/10.1186/S13059-019-1842-9/FIGURES/4

Hagel JM, Facchini PJ (2017) Tying the knot: occurrence and possible significance of gene fusions in plant metabolism and beyond. J Exp Bot 68:4029–4043. https://doi.org/10.1093/jxb/erx152

Han C, Sun LY, Wang WT et al (2019) Non-coding RNAs in cancers with chromosomal rearrangements: the signatures, causes, functions and implications. J Mol Cell Biol 11:886–898. https://doi.org/10.1093/JMCB/MJZ080

Hasan S, Huang L, Liu Q et al (2022) The long read transcriptome of rice (Oryza sativa ssp. japonica var. Nipponbare) reveals novel transcripts. Rice 15:1–17. https://doi.org/10.1186/S12284-022-00577-1/FIGURES/5

Jia Y, Xie Z, Li H (2016) Intergenically spliced chimeric RNAs in cancer. Trends Cancer 2:475–484. https://doi.org/10.1016/J.TRECAN.2016.07.006

Kawasaki T, Okumura S, Kishimoto N et al (1999) RNA maturation of the rice SPK gene may involve trans-splicing. Plant J 18:625–632. https://doi.org/10.1046/J.1365-313X.1999.00493.X

Kim M, Canio W, Kessler S (2001) Developmental changes due to long-distance movement of a homeobox fusion transcript in tomato. Science 293:287–289. https://doi.org/10.1126/SCIENCE.1059805/SUPPL_FILE/1059805S2_THUMB.GIF

Koller B, Fromm H, Galun E, Edelman M (1987) Evidence for in vivo trans splicing of pre-mRNAs in tobacco chloroplasts. Cell 48:111–119. https://doi.org/10.1016/0092-8674(87)90361-8

Kumar S, Razzaq SK, Vo AD et al (2016) Identifying fusion transcripts using next generation sequencing. Wiley Interdiscip Rev RNA 7:811–823. https://doi.org/10.1002/WRNA.1382

Latysheva NS, Babu MM (2016) Discovering and understanding oncogenic gene fusions through data intensive computational approaches. Nucleic Acids Res 44:4487–4503. https://doi.org/10.1093/NAR/GKW282

Li H, Wang J, Ma X, Sklar J (2009) Gene fusions and RNA trans-splicing in normal and neoplastic human cells. Cell Cycle 8:218–222. https://doi.org/10.4161/CC.8.2.7358

Ma C, Shao M, Kingsford C (2018) SQUID: Transcriptomic structural variation detection from RNA-seq. Genome Biol 19:1–16. https://doi.org/10.1186/S13059-018-1421-5/FIGURES/7

Mertens F, Johansson B, Fioretos T, Mitelman F (2015) The emerging complexity of gene fusions in cancer. Nat Rev Cancer 15:371–381. https://doi.org/10.1038/nrc3947

Mukherjee S, Detroja R, Balamurali D et al (2021) Computational analysis of sense-antisense chimeric transcripts reveals their potential regulatory features and the landscape of expression in human cells. NAR Genom Bioinform. https://doi.org/10.1093/NARGAB/LQAB074

Parakkunnel R, Bhojaraja Naik K, Vanishree G et al (2022) Gene fusions, micro-exons and splice variants define stress signaling by AP2/ERF and WRKY transcription factors in the sesame pangenome. Front Plant Sci 13:1076229. https://doi.org/10.3389/FPLS.2022.1076229/BIBTEX

Qiao D, Yang C, Chen J et al (2019) Comprehensive identification of the full-length transcripts and alternative splicing related to the secondary metabolism pathways in the tea plant (Camellia sinensis). Sci Rep 9:1–13. https://doi.org/10.1038/s41598-019-39286-z

Qin F, Song Z, Babiceanu M et al (2015) Discovery of CTCF-sensitive cis-spliced fusion RNAs between adjacent genes in human prostate cells. PLoS Genet 11:e1005001. https://doi.org/10.1371/JOURNAL.PGEN.1005001

Rogers RL, Bedford T, Lyons AM, Hartl DL (2010) Adaptive impact of the chimeric gene Quetzalcoatl in Drosophila melanogaster. Proc Natl Acad Sci U S A 107:10943–10948. https://doi.org/10.1073/PNAS.1006503107

Singh A, Zahra S, Das D, Kumar S (2019) AtFusionDB: a database of fusion transcripts in Arabidopsis thaliana. Database 2019:135. https://doi.org/10.1093/DATABASE/BAY135

Singh S, Qin F, Kumar S et al (2020) The landscape of chimeric RNAs in non-diseased tissues and cells. Nucleic Acids Res 48:1764–1778. https://doi.org/10.1093/NAR/GKZ1223

Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147:195–197. https://doi.org/10.1016/0022-2836(81)90087-5

St. John J, Powell K, Katie Conley-LaComb M, Chinni SR (2012) TMPRSS2-ERG fusion gene expression in prostate tumor cells and its clinical and biological significance in prostate cancer progression. J Cancer Sci Ther 4:94. https://doi.org/10.4172/1948-5956.1000119

Tkachuk DC, Westbrook CA, Andreeff M et al (1990) Detection of bcr-abl fusion in chronic myelogeneous leukemia by in situ hybridization. Science 250:559–562. https://doi.org/10.1126/SCIENCE.2237408

Varley KE, Gertz J, Roberts BS et al (2014) Recurrent read-through fusion transcripts in breast cancer. Breast Cancer Res Treat 146:287–297. https://doi.org/10.1007/S10549-014-3019-2/FIGURES/4

Wang K, Singh D, Zeng Z et al (2010) MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res 38:e178. https://doi.org/10.1093/NAR/GKQ622

Zhang G, Guo G, Hu X et al (2010) Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. Genome Res 20:646. https://doi.org/10.1101/GR.100677.109

Zhang C, Wang J, Marowsky NC et al (2013) High occurrence of functional new chimeric genes in survey of rice chromosome 3 short arm genome sequences. Genome Biol Evol 5:1038–1048. https://doi.org/10.1093/GBE/EVT071

Zhang S, Li R, Zhang L et al (2020) New insights into Arabidopsis transcriptome complexity revealed by direct sequencing of native RNAs. Nucleic Acids Res 48:7700–7711. https://doi.org/10.1093/NAR/GKAA588

Zhou Y, Zhang C (2019) Evolutionary patterns of chimeric retrogenes in Oryza species. Sci Rep 9:1–12. https://doi.org/10.1038/s41598-019-54085-2

Zhou Y, Lu Q, Zhang J et al (2022) genome-wide profiling of alternative splicing and gene fusion during rice black-streaked dwarf virus stress in maize (Zea mays L.). Genes 13:456. https://doi.org/10.3390/GENES13030456/S1

Zhou Y, Zhang C, Zhang L et al (2022b) Gene fusion as an important mechanism to generate new genes in the genus Oryza. Genome Biol 23:1–23. https://doi.org/10.1186/S13059-022-02696-W/FIGURES/8