Journal of Wood Science

ORIGINAL ARTICLE Open Access

SugiExDB: data repository for *Cryptomeria japonica* RNA-sequencing read counts in TogoDB

Naoki Takata[1*†], Ryosuke Sato[1†], Soichiro Nagano[2] and Saneyoshi Ueno[3]

Abstract

Transcriptome analysis, particularly RNA-sequencing, is an important tool for predicting gene function and investigating the physiology and phenology of organisms. We conducted RNA-sequencing analyses on 35 samples of *Cryptomeria japonica* collected from 11 specific tissues of 3-year-old trees during the growing season. In addition to these datasets, we retrieved 118 RNA-sequencing datasets from public databases and analyzed gene expression patterns by mapping them to the *C. japonica* genome sequence. By integrating the read counts and counts per million files from all 153 RNA-sequencing datasets, we constructed a comprehensive repository entitled "SugiExDB: Data repository for *Cryptomeria japonica* RNA-sequencing read counts" under TogoDB platform (http://togodb.org/db/index_rnaseq_cj). This repository enables users to see expression patterns for each gene and to download all datasets for further analyses.

Keywords RNA-sequencing, Data storage, *Cryptomeria japonica*

Introduction

Cryptomeria japonica is a coniferous tree with significant potential for research in molecular biology and new breeding technologies. Widely cultivated in Japan and parts of southern China, *C. japonica* accounts for 44% of Japan's forests and is a major species in silviculture. Tree breeders have developed numerous *C. japonica* cultivars, including low- or no-pollen varieties and fast-growing trees, which have been planted in artificial forests for several decades. One notable biological trait of *C. japonica* is its ability to be crossbred at just one or two years old, whereas many tree species require several years to flower. Additionally, advancements in generating transgenic and gene-edited *C. japonica* have greatly facilitated breeding programs in Japan.

C. japonica has 11 chromosomes ($n = 11$) with an estimated genome size of approximately 11 Gb, which is about three times larger than the human genome (3.1 Gb). The genome sequencing of *C. japonica* were completed for two genotypes in recent years [1, 2]. Using HiFi PacBio long-read sequencing, the genome assemblies of the two genotypes yielded 2,650 contigs with an N50 value of 12.0 Mb [1] and 2,740 contigs with an N50 value of 8.3 Mb [2], covering more than 93% of the estimated genome size. Fujino et al. [1] predicted protein-coding genes in these assemblies, providing two gene catalogues: the"Permissive gene set"with 152,527 genes and the"Standard gene set"with 55,246 genes. These predicted gene sets are based on RNA-sequencing data

†Naoki Takata and Ryosuke Sato have contributed equally.

*Correspondence:
Naoki Takata
naokitakata@affrc.go.jp
[1] Forest Bio-Research Center, Forestry and Forest Products Research Institute, Forest Research and Management Organization, 3809-1 Ishi, Juo, Hitachi, Ibaraki 319-1301, Japan
[2] Forest Tree Breeding Center, Forestry and Forest Products Research Institute, Forest Research and Management Organization, 3809-1 Ishi, Juo, Hitachi, Ibaraki 319-1301, Japan
[3] Department of Forest Molecular Genetics and Biotechnology, Forestry and Forest Products Research Institute, Forest Research and Management Organization, 1 Matsunosato, Tsukuba, Ibaraki 305-8687, Japan

Springer Open

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

and homology with protein databases. The availability of common *C. japonica* transcript IDs (e.g., SUGI_0000010) in the genome database marks an important milestone, allowing different research groups to use the same gene sets for transcriptome and proteome analyses, thereby enabling data comparison across studies.

Transcriptome analysis is commonly used to study the physiology and phenology of organisms. In the past decade, RNA-sequencing using massive parallel sequencing technology allows us to analyze transcriptome for even non-model organisms. All RNA-sequencing data are required to be deposited in public databases managed by the International Nucleotide Sequence Database Collaboration (INSDC), which includes the National Institute of Genetics (NIG), the National Center for Biotechnology Information (NCBI), and the European Molecular Biology Laboratory (EMBL), prior to the publication of scientific papers, in accordance with the policies of the respective journals. Researchers typically upload sequencing read files, such as FASTQ files, to these public databases, allowing other researchers interested in transcriptome data to download and reanalyze them. Reanalyzing transcriptome data in a comprehensive manner is a crucial tool for predicting gene function and assessing the specific physiological conditions of an organism.

In this study, we sampled several tissues from *C. japonica* to investigate tissue-specific gene expression patterns and conducted RNA-sequencing analyses using the gene sets published by Fujino et al. [1]. Data sets of tissue-specific expression greatly contribute to elucidate gene function. In addition to our own transcriptome data, we reanalyzed previously deposited transcriptome datasets using the same *C. japonica* gene set. By integrating all available transcriptome data, we created a comprehensive data repository of *C. japonica* RNA-sequencing read counts under the TogoDB platform. This repository allows users to download read count files and visualize gene expression patterns in dot plots.

Experimental

Plant materials

Plantlets of *Cryptomeria japonica* (plant strain #13–8–12) were propagated in vitro using a culture bottle containing half-strength WPF-f medium [3]. The plantlets were transferred to pots containing soil mixture consisting of peat moss-based soil (Jiffy Mix; Sakata Seed, Kanagawa, Japan), akadama soil (granules of volcanic ash soil; Nagahama Shouten, Tochigi, Japan) and vermiculite (Asahi Kogyo, Okayama, Japan) in a 1:1:1 ratio. These plants were subsequently cultivated in a special netted greenhouse under natural photoperiod and ambient temperature conditions. For summer growth, the trees were fertilized with Maruyama No.1 (solid fertilizer; Nihon Ringyo Hiryo, Tokyo, Japan), while during winter they were maintained by regular watering without fertilization.

RNA extraction and RNA-sequencing

We collected samples during the growing season from three independent *C. japonica* plants grown in the special netted greenhouse for three years. Tissue samples were taken from various plant parts, including the shoot tip, leaves from one-year-old stems, xylem and bark from both the upper and lower sections of one-year-old stems, phloem fibers from the lower section of one-year-old stems, xylem, phloem fibers, and bark from 2-year-old stems, and roots. Those samples were frozen with enough liquid nitrogen and stored at − 80 ℃ until RNA extraction. The frozen samples were ground into a fine powder using a MM300 TissueLyser Mill Mixer (Retsch, Haan, Germany) and a ShakeMaster (Bio Medical Science, Tokyo, Japan). Total RNA was extracted using a Maxwell RSC Plant RNA Kit (Promega, Madison, WI, USA). RNA-sequencing libraries were constructed using a NEBNext Ultra II Directional RNA Library Prep Kit for Illumina (New England Biolabs, Ipswich, MA, USA). Sequencing of the RNA-sequencing libraries was performed on a DNBSEQ-T7 platform (MGI Tech, Shenzhen, China), yielding 44.9–104.9 million 150 bp paired-end reads per sample (Table S1). The sequence reads indicate the representative transcripts present in the tissue samples. Library preparation and sequencing were conducted by Novogene. (Beijing, China). The resulting sequence data have been deposited in the DNA Data Bank of Japan (DDBJ) Sequence Read Archive under accession number PRJDB18881.

Data collection

In addition to our RNA-sequencing data, we collected *C. japonica* RNA-sequencing data from the BioProject database of NCBI. We also collected transcriptome data for *Cryptomeria fortunei*, as *C. fortunei* is synonymous with *C. japonica* var. *sinensis*, which is the only accepted scientific name according to The Plant List [4].

Data analyses

Raw reads of RNA-sequencing data were trimmed using Trimmomatic (ver. 0.39) [5] (Table S1). The following parameters were used for trimming: ILLUMINACLIP:adapter.fa:2:30:10, LEADING:20, TRAILING:20, SLIDINGWINDOW:4:15, and MINLEN:36. adapter.fa contained the 5′ and 3′ adapter sequences AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT and GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCTCGTATGCCGTCTTCTGCTTG, respectively.

The trimmed paired reads were aligned to the reference genome of *C. japonica* SUGI_1 (https://forestgen.ffpri.go.jp/en/info_sugi1.html) using HISAT2 (ver. 2.1.0) [6] with the following options: –no-discordant and –no-mixed (Table S1). The obtained Sequence Alignment Map was converted into a Binary Alignment Map using SAMtools (ver. 1.15) [7]. Reads were counted using featureCounts (ver. 2.0.2) [8] with the following options: -p -t gene -g ID. The GFF3 file was downloaded from https://forestgen.ffpri.go.jp/SUGI_1/SUGI_1.pmsv.gene.gff3.gz and used for the analysis. The raw counts of all samples were merged into one file and processed using R (ver. 4.2.1) [9]. The read counts were normalized using the variance stabilizing transformation (VST) in DESeq [10]. The 400 genes with the highest standard deviation (SD) across samples (Table S2) were selected for hierarchical clustering. Heatmaps were generated with the R function heatmap.2, available in the gplots library [11], and clustering was performed using Ward's method.

Results and discussion

Collection of RNA-sequencing data of *C. japonica* plants

To examine tissue-specific gene expression patterns, we conducted RNA-sequencing for 35 total RNA samples extracted from 11 specific tissues, including shoot tips, leaves from one-year-old stems, xylem and bark from both the upper and lower sections of one-year-old stems, phloem fibers from one-year-old lower stems, xylem and phloem fibers from 2-year-old stems, bark from 2-year-old stems, and roots (Table 1). In our RNA-sequencing analyses, RNA sample quality had no significant impact on raw read count throughput, quality trimming, and mapping rate (Table S1). We also searched *C. japonica* RNA-sequencing data in public databases and found 11 BioProject data sets that have been deposited in the BioProject database of NCBI (January, 30 th, 2024). From these, we retrieved 118 RNA-sequencing data from eight BioProjects; (accession IDs; PRJDB13625, PRJDB6436, PRJDB8803, PRJDB10268, PRJNA644276, PRJNA697258, PRJNA793065, and PRJDB12272) [1, 12–18] (Table 1). The remaining three BioProjects were excluded from our analysis: two contained RNA-sequencing data for mixed tissue samples, and the third focused on small RNA. In total, 153 RNA-sequencing datasets were collected and mapped to the *C. japonica* genome sequence (SUGI_1) [1]. The raw read counts were processed and normalized to counts per million (CPM) for all datasets. Cluster analysis of genome-wide gene expression revealed that the 153 datasets could be grouped into six clusters, largely based on the tissue types sampled (Fig. 1). This result suggests that, despite being generated by different research groups, each RNA-sequencing dataset preserved the major gene expression profiles characteristic of the respective tissues.

Construction of the data storage of RNA-sequencing read counts in TogoDB

To store RNA-sequencing data, we selected TogoDB, an online storage server provided by the Database Center for Life Science (DBCLS), part of the Research

Table 1 RNA-sequencing datasets stored in "The data storage of *Cryptomeria japonica* RNA-sequencing read counts"

Accession ID	Tissues	Conditions	References
PRJDB18881	Shoot tip, leaves on one-year-old stem, xylem in one-year-old upper stem, bark in one-year-old upper stem, xylem in one-year-old lower stem, bark in one-year-old lower stem, phloem fibers in one-year-old lower stem, xylem in 2-year-old stem, phloem fibers in 2-year-old stem, bark in 2-year-old stem, roots	–	This study
PRJDB13625	Roots from 2-year-old seedling, female flower buds, cone, seeds in cones	–	[1]
PRJDB6436	Male strobili	–	[15]
PRJDB8803	Needles of seedlings	24 h before heat shock, 1 h after heat shock, 24 h after heat shock	[11]
PRJDB10268	Embryogenic cells, non-embryogenic cells	5 weeks after maturation, 7 weeks after maturation, 9 weeks after maturation	[12]
PRJNA644276	Vascular cambium from 60-year-old tree	–	[13]
PRJNA697258	Seasonal needles of the secondary lateral branches of 5-year-old trees	–	[14]
PRJNA793065	Seasonal needles of the secondary lateral branches of 5-year-old trees	–	[16]
PRJDB12272	Shoot from 2-year-old seedling	Elevated (800 ppm) and lowered (200 ppm) CO2 concentration	[17]

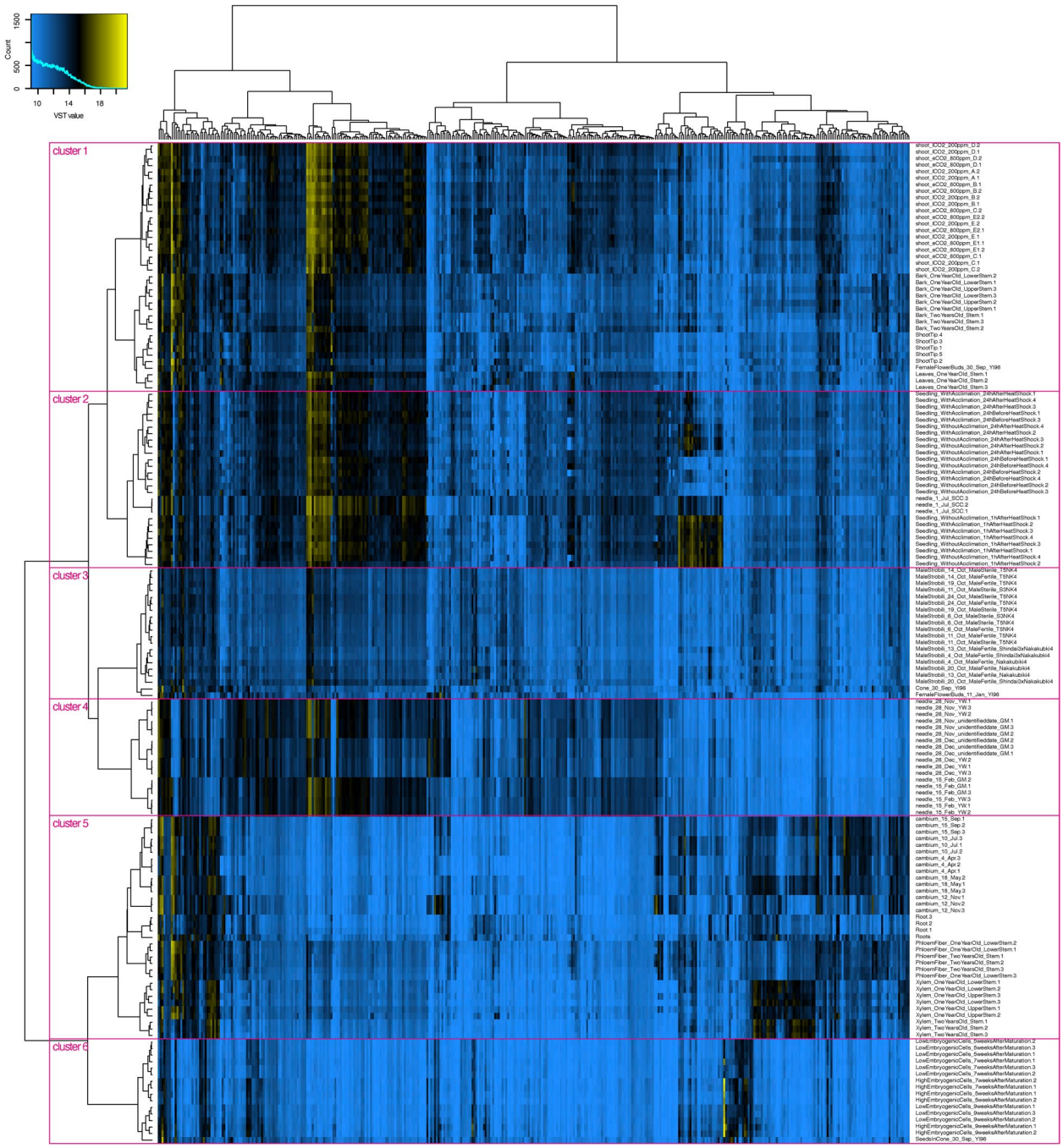


Fig. 1 Hierarchical clustering and heatmap of 153 RNA-sequencing datasets using the top 400 genes with the SD values. Genes with higher expression levels are represented in yellow, while those with lower expression levels are depicted in blue. The 153 datasets were grouped into six clusters, largely based on the tissue types sampled. cluster 1, bark, shoot, younger needles; cluster 2, seedlings and needles in growing season; cluster 3, male strobili; cluster 4, needles in dormant season; cluster 5, xylem, phloem fibers, and roots; cluster 6, embryogenic cells

Organization of Information and Systems (ROIS) in Japan. TogoDB is a simple, user-friendly, and updatable database system that allows data uploads through an intuitive web interface. This enables easy data addition and updates in just a few steps. In this study, all read count files and CPM files were uploaded to TogoDB and indexed under the sample list entitled "SugiExDB: Data repository for *Cryptomeria japonica* RNA-sequencing read counts" (http://togodb.org/db/index_rnaseq_cj) (Fig. 2). Users can access and download all

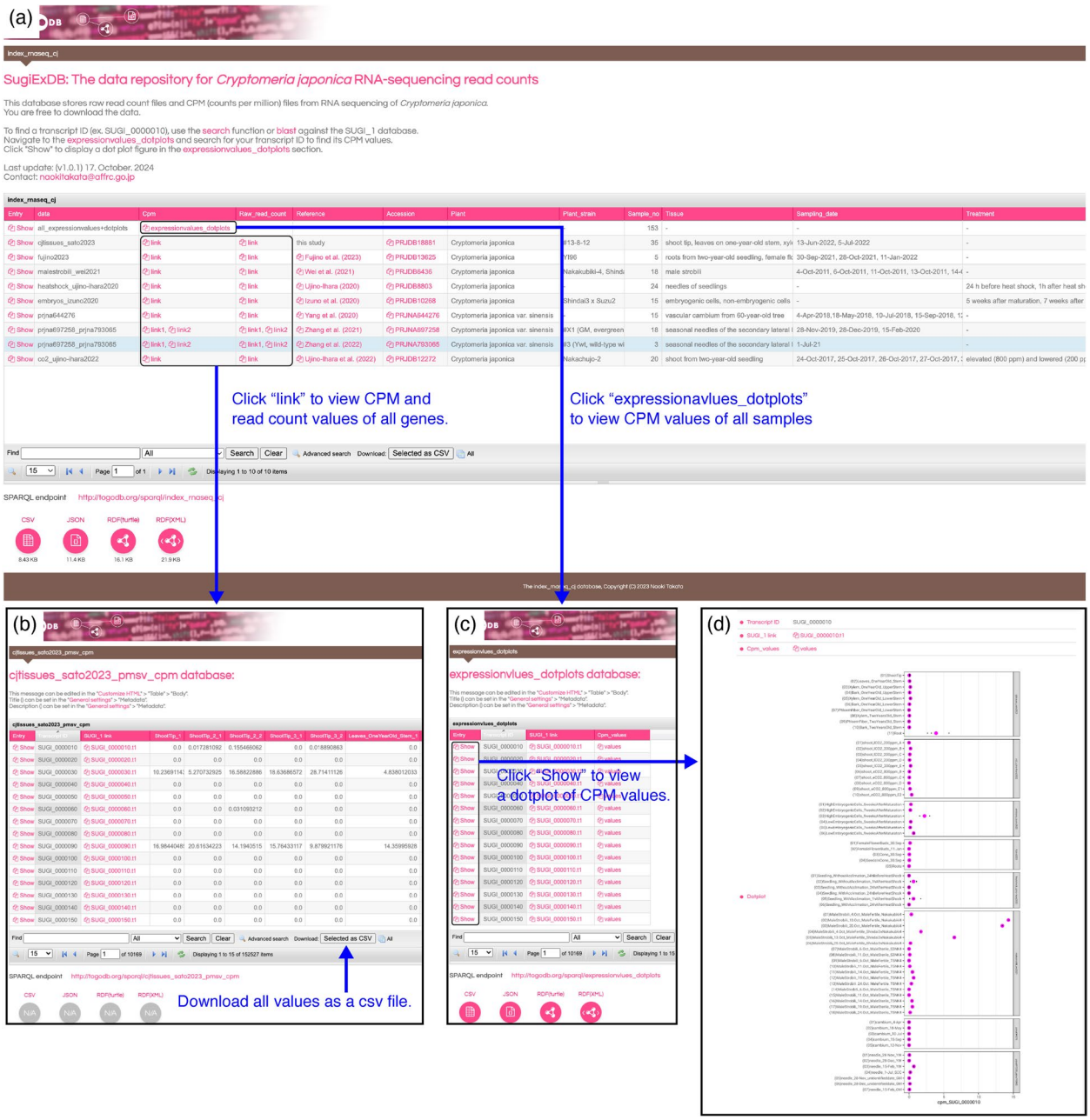


Fig. 2 The data storage of *Cryptomeria japonica* RNA-sequencing read counts. **a** A comprehensive index of the data storage. **b** An example list of CPM values from the dataset"cjtissues_sato2023". Users can download the complete CPM values for all genes as a CSV file. **c** A summary of CPM expression values accompanied by dot plots for all genes. **d** A specific dot plot illustrating the expression of a single gene

the data as CSV files via this storage. To summarize the gene expression patterns for each gene, we created an additional dataset, "all_expressionvalues +dotplots", which contains the CPM values for all samples, along with dot plots visualizing gene expression patterns. Users can retrieve expression values for all samples and view the expression patterns for specific transcript IDs using the dot plot feature. For example, transcript ID SUGI_0000010 shows higher expression levels in roots and male strobili (Fig. 2). Transcript IDs (e.g., SUGI_0000010) are referenced from the *C. japonica* genome database available on ForestGen (https://forestgen.ffpri.go.jp/en/index.html). In summary, we developed a simple RNA-sequencing data storage system, which is regularly updated and provides downloadable read count and CPM data.

Conclusion

We have made the data storage system, "SugiExDB: Data repository for *Cryptomeria japonica* RNA-sequencing read counts" available on TogoDB. Users can freely download the data and perform various analyses, such as co-expression network analysis. We plan to update the storage as new data become available through the BioProject database on NCBI. Users who wish to submit their own RNA-sequencing data are encouraged to contact us for inclusion in the storage.

Abbreviations

INSDC	International Nucleotide Sequence Database Collaboration
NIG	National Institute of Genetics
NCBI	National Center for Biotechnology Information
EMBL	European Molecular Biology Laboratory
DDBJ	DNA Data Bank of Japan
VST	Variance stabilizing transformation
SD	Standard deviation
CPM	Counts per million
DBCLS	Database Center for Life Science
ROIS	Research Organization of Information and Systems

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s10086-025-02205-0.

Additional file 1: Table S1. Quality controls, raw read counts, quality trimmings, and mapping rates of 35 RNA samples that we conducted RNA-sequencing.

Additional file 2: Table S2. 400 gene sets used in hierarchical clustering and heatmap (Figure 1).

Acknowledgements

We are grateful to Ms. Shiho Kamikabeya and Ms. Kazuko Kato (Forest Bio-Research Center, Forestry and Forest Products Research Institute, Forest Research and Management Organization) for skillful technical support. Computations were partly performed on the supercomputer of AFFRIT, MAFF, Japan.

Author contributions

N.T. and R.S. designed the study. N.T. wrote the initial draft of the manuscript. N.T., R.S., and S.U. contributed to data collection. N.T., R.S., S.U., and S.N. contributed to data analysis and interpretation of the data.

Funding

This work was supported in part by the MAFF commissioned project study on "Development of efficient breeding technique aiming at forestry trees with superior carbon storage capacity" (Grant Number JPJ009841) and JSPS KAKENHI (Grant Numbers 22H02412 to N.T.).

Availability of data and materials

The datasets analyzed during the current study are available in "SugiExDB: Data repository for *Cryptomeria japonica* RNA-sequencing read counts" (http://togodb.org/db/index_rnaseq_cj).

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 24 December 2024 Accepted: 12 May 2025
Published online: 12 June 2025

References

1. Fujino T, Yamaguchi K, Yokoyama T, Hamanaka T, Harazono Y, Kamada H, Kobayashi W, Ujino-Ihara T, Uchiyama K, Matsumoto A, Izuno A, Tsumura Y, Toyoda A, Shigenobu S, Moriguchi Y, Ueno S, Kasahara M (2024) A chromosome-level genome assembly of a model conifer plant, the Japanese cedar, *Cryptomeria japonica* D. Don. BMC Genomics 25:1039
2. Shirasawa K, Mishima K, Hirakawa H, Hirao T, Tsubomura M, Nagano S, Iki T, Isobe S, Takahashi M (2024) Haplotype-resolved de novo genome assemblies of four coniferous tree species. J Forest Res 29:151–157
3. Konagaya K, Nanasato Y, Taniguchi T (2020) A protocol for *Agrobacterium*-mediated transformation of Japanese cedar, Sugi (*Cryptomeria japonica* D. Don) using embryogenic tissue explants. Plant Biotechnol 37:147–156
4. The Plant List (2024) http://www.theplantlist.org/tpl1.1/search?q=cryptomeria. Accessed 25 Sep 2024.
5. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120
6. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol 37:907–915
7. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) 1000 Genome project data processing subgroup. The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079
8. Liao Y, Smyth GK, Shi W (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30:923–930
9. R Core Team (2024) R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, Vienna, Austria, 2020). https://www.r-project.org/. Accessed 25 Sep 2024.
10. Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15:550
11. Warnes G, Bolker B, Bonebakker L, Gentleman R, Huber W, Liaw A, Lumley T, Maechler M, Magnusson A, Moeller S, Schwartz M, Venables B, Galili T (2024) gplots: Various R Programming Tools for Plotting Data. R package version 3.2.0, https://talgalili.github.io/gplots/, https://github.com/talgalili/gplots.
12. Ujino-Ihara T (2020) Transcriptome analysis of heat stressed seedlings with or without pre-heat treatment in *Cryptomeria japonica*. Mol Genet Genomics 295:1163–1172
13. Izuno A, Maruyama TE, Ueno S, Ujino-Ihara T, Moriguchi Y (2020) Genotype and transcriptome effects on somatic embryogenesis in *Cryptomeria japonica*. PLoS ONE 15:e0244634
14. Yang J, Guo Z, Zhang Y, Mo J, Cui J, Hu H, He Y, Xu J (2020) Transcriptomic profiling of *Cryptomeria fortunei* Hooibrenk vascular cambium identifies candidate genes involved in phenylpropanoid metabolism. Forests 11:766
15. Zhang Y, Cui J, Hu H, Xue J, Yang J, Xu J (2021) Integrated four comparative-omics reveals the mechanism of the terpenoid biosynthesis in two different overwintering *Cryptomeria fortunei* phenotypes. Front Plant Sci 12:740755
16. Wei FJ, Ueno S, Ujino-Ihara T, Saito M, Tsumura Y, Higuchi Y, Hirayama S, Iwai J, Hakamata T, Moriguchi Y (2021) Construction of a reference transcriptome for the analysis of male sterility in sugi (*Cryptomeria japonica* D. Don) focusing on MALE STERILITY 1 (MS1). PLoS ONE 16:e0247180
17. Zhang Y, Yang L, Yang J, Hu H, Wei G, Cui J, Xu J (2022) Transcriptome and metabolome analyses reveal differences in terpenoid and flavonoid biosynthesis in *Cryptomeria fortunei* needles across different seasons. Front Plant Sci 13:862746
18. Ujino-Ihara T, Tobita H, Miyazawa S (2022) Changes in the *Cryptomeria japonica* shoot transcriptome after short-term treatments with different concentrations of CO_2. Bull FFPRI 21:207–216

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.