www.nature.com/scientificdata

scientific data

OPEN

DATA DESCRIPTOR

Comprehensive curation and validation of genomic datasets for chestnut

Jialu Fan[1,2,6], Yu Zhang[1,3,6], Xinghua Nie[1,3], Yang Liu[1,3], Shangxiao Wei[1,2], Haixu Peng[1,2], Hanlei Li[1,2], Mingjun Zhang[1,2], Lu Ning[4], Sen Wang[1,2,5], Ling Qin[1,3], Yi Zheng[1,2,5] ✉ & Yu Xing[1,3] ✉

The Chinese chestnut (*Castanea* mollissima) stands out as a plant with significant ecological and economic value, excellent nutritional quality and natural resistance to pests and diseases. Recent strides in high-throughput techniques have enabled the continuous accumulation of genomic data on chestnuts, presenting a promising future for genetic research and advancing traits in this species. To facilitate the accessibility and utility of this data, we have curated and analyzed a collection of genomic datasets for eight *Castanea* species, including functional annotations, 213 RNA-Seq samples, and 330 resequencing samples. These datasets are publicly available on Figshare and are also available through other platforms such as GEO and EVA, providing a valuable resource for researchers studying *Castanea* genetics, functional genomics, and evolutionary biology. Furthermore, the datasets are integrated into the *Castanea* Genome Database (CGD, http://castaneadb.net), which serves as a complementary platform, offering advanced data mining and analysis tools, including BLAST, Batch Query, GO/KEGG Enrichment Analysis, and Synteny Viewer, to enhance the usability of the curated datasets.

Background & Summary

Castanea is a genus of plants in the family Fagaceae that has significant ecological and economic value[1], including excellent nutritional quality[2]. The genus comprises seven species, including the Chinese chestnut (*Castanea mollissima*), Seguin chestnut (*Castanea seguinii*), Henry chestnut (*Castanea henryi*), and Japanese chestnut (*Castanea crenata*) in East Asia, the American chestnut (*Castanea dentata*) and Chinkapin (*Castanea pumila*) are present in North America, while European chestnut (*Castanea sativa*) is grown in Europe. Chestnuts play an important role in nut production and forest ecosystem services[3]. Chinese chestnuts are widely cultivated in 26 provinces of China[4], with a nut yield of 1,562,685 tons, accounting for 74% of the global chestnut production, ranking first in the world in 2022[5].

Recent advancements in high-throughput sequencing technologies have enabled the continuous accumulation of genomic data for *Castanea* species. However, existing resources are relatively scattered and lack integration, limiting researchers' ability to conduct in-depth studies on *Castanea* genetics, functional genomics, and evolutionary biology.

To address this gap, we have gathered and analyzed a comprehensive collection of genomic datasets for *Castanea*. The datasets include genomic information from eight *Castanea* species, 213 RNA-Seq samples, and 330 resequencing samples and are publicly available on figshare. This curation and validation process ensures the reliability and utility of the data for researchers.

Additionally, the CGD serves as a complementary platform to enhance the usability of these datasets. The CGD provides a user-friendly interface and a suite of advanced data mining and analysis tools, including BLAST, Batch Query, GO/KEGG Enrichment Analysis, and Synteny Viewer. These tools facilitate the exploration and analysis of the curated datasets, enabling researchers to investigate deeper into *Castanea* genetics and functional genomics.

[1]Beijing Key Laboratory for Agriculture Application and New Technique, College of Plant Science and Technology, Beijing University of Agriculture, Beijing, 102206, China. [2]Bioinformatics Center, Beijing University of Agriculture, Beijing, 102206, China. [3]Beijing Advanced Innovation Center for Tree Breeding by Molecular Design, Beijing, 102206, China. [4]Library, Beijing University of Agriculture, Beijing, 102206, China. [5]Ancient Tree Health and Culture Engineering Technology Research Center, National Forestry and Grassland Administration, Beijing, China. [6]These authors contributed equally: Jialu Fan, Yu Zhang. ✉e-mail: yz@moilab.net; xingyu@bua.edu.cn

Genome	Total length (Mb)	N50 length (Mb)	Data records
C. mollissima cv. HBY-2 (Xing *et al.*[6], Liu *et al.*[7])	738.87	NA	PRJNA527178
C. mollissima cv.Sun (Sun *et al.*[9])	773.99	5.88	PRJNA540917
C. mollissima cv. 'Vanuxem' (Staton *et al.*[10])	725.20	0.14	PRJNA46687
C. mollissima cv. N11-1 (Wang *et al.*[8])	688.93	2.83	PRJNA559042
C. crenata cv. Ginyose (Shirasawa *et al.*[12])	721.20	1.6	DRA012289
C. mollissima var. drought-resistant (H7) (Hu *et al.*[11])	678.90	3.39	PRJNA769510
C. mollissima var. early-maturing (ZS) (Hu *et al.*[11])	790.99	2.17	
C. mollissima var. easy-pruning (YH) (Hu *et al.*[11])	671.99	3.65	

Table 1. List of collected *Castanea* genomes.

We envision the CGD as an essential resource for advancing functional genomic research and understanding the evolutionary relationships within the *Castanea* genus. By making these datasets and tools readily accessible, we aim to facilitate collaborative research and drive innovation in the field.

Methods

Data collection.

The CGD contains eight chestnut genomes, including seven from *C. mollissima* ('HBY-2'[6,7], 'N11-1'[8], 'Sun'[9], 'Vanuxem'[10], 'drought-resistant'(H7)[11], 'early-maturing'(ZS)[11] and 'easy-pruning'(YH)[11]) and one from *C. crenata*[12] (Table 1). Genomic information for chestnut, including the genome sequences, mRNA, gene structure annotations in general feature format (GFF), coding (CDS) and protein sequences of protein-coding genes, were obtained based on previously published articles. We collected the genome of 'Vanuxem' from the Hardwood Genomics website (now relocated to the TreeGenes Database, https://treegenesdb.org/), which includes several transcriptome libraries from trees affected by chestnut blight disease[10,13]. The TreeGenes Database also contains data for eight chestnut species. However, it is not specific to chestnut, it includes data for multiple tree species. Noted that the C. mollissima cv. HBY-2 we previously generated using Pacific Biosciences single-molecule sequencing technology[6]. We performed Hi-C analysis on the Chinese chestnut genome HBY-2 to enhance the genome sequence assembly and contiguity[7]. Therefore, this version exhibits superior genome contiguity and sequence quality compared to the previous version. Besides, RNA-Seq data from various samples, which are at different tissues, developmental stages, as well as cultivars (Supplementary Table 1) were downloaded from the NCBI SRA databese. And resequencing datasets from 330 accessions were downloaded from the NCBI database[7,9].

Data processing.

Gene functional annotation. In our previous study, we developed a standard pipeline for comprehensively annotating predicted protein-coding genes[14]. In brief, the protein sequences of the predicted genes were analyzed against the UniPort (Swiss-Port and TrEMBL), and NCBI nonredundant (nr), and Arabidopsis protein (TAIR) databases using DIAMOND[15] with an E-value cutoff of 1e-4. Furthermore, all the protein sequences were compared against the InterPro database to identify functional domains by InterProScan[16]. For the purpose of conducting functional enrichment analyses and generating GO and KEGG pathway annotations, protein sequences were aligned with the EggNOG database using eggnog-mapper[17]. The GO terms assigned to genes/transcripts based on the eggnog-mapper results were transformed into the GO Annotation File (GAF) format. Within the eggnog-mapper outcomes, certain KEGG pathways unrelated to plants were discarded. The iTAK program was employed to identify transcription factors (TFs), transcriptional regulators (TRs), and protein kinases(PKs) from the predicted protein-coding genes and to classify them into different families[18].

After the above analysis, we obtained files containing homologous genes identified by BLAST, protein functional domains, AHRD-based functional descriptions, and GO/KEGG annotations, which have been shared in our database and on Figshare.

RNA-Seq analysis. A specific pipeline was utilized to procedure and analyze hundreds of RNA-Seq datasets, which were downloaded from the NCBI SRA Database. Firstly, raw RNA-Seq reads were processed using the FastQC software (v0.11.9)[19] to evaluate the quality of reads, then treated to remove adaptor and low-quality sequences using Trimmomatic[20], Trimmed reads shorter than 80% of their initial length were removed, and then the remaining cleaned reads were aligned to the reference genome (HBY-2) using the STAR (version 2.7.10b)[21]. Finally, read counts for each gene were calculated based on the alignments and normalized to fragments per kilobase of transcript per million mapped fragments (FPKM) values[22].

Following the analysis, we generated the raw and normalized expression matrices and made them available both in our database and on Figshare.

Variant identification. To identify variants, a pipeline developed by Sentieon Inc.[23] was used with default parameters. First, quality evaluation utilizing the FastQC software (v0.11.9)[19], and resequencing reads were processed to remove adapter and low-quality sequences using Trimmomatic[20]. Then, the cleaned reads were aligned to the reference genomes (HBY-2) utilizing the BWA-MEM algorithm with default parameters[24]. Following alignment, duplicated reads were removed using the 'LocusCollector' and 'Dedup' algorithms of Sentieon. Finally, variants were called using the 'Haplotyper' algorithms of Sentieon software (https://www.sentieon.com/).

	C (S, D)	F	M
C. crenata cv.Ginyose	98.0% (94.2%, 3.8%)	1.1%	0.9%
C. mollissima cv. HBY-2	97.6% (91.3%, 6.3%)	1.1%	1.3%
C. mollissima cv. N11-1	94.0% (89.6%, 4.4%)	1.5%	4.5%
C. mollissima cv. Sun	98.5% (89.8%, 8.7%)	0.9%	0.6%
C. mollissima cv. 'Vanuxem'	95.6% (90.5%, 5.1%)	2.9%	1.5%
C. mollissima var. drought-resistant (H7)	94.3% (91.4%, 2.9%)	1.6%	4.1%
C. mollissima var. early-maturing (ZS)	98.5% (90.8%, 7.7%)	0.9%	0.6%
C. mollissima var. easy-pruning (YH)	92.6% (89.6%, 3.0%)	1.2%	6.2%

Table 2. BUSCO assessment of the completeness of genomes. (C: Complete BUSOs, S: Complete Single-Copy BUSCOs, D: Complete Dupicated BUSCOs, F: Fragemented BUSCOs, M: Missing BUSCOs).

As a result of the analysis, we produced the vcf files and have shared them through our database and Figshare.

Data Records

All functional annotations, expression profiles, and called variant data have been uploaded to Figshare (https://figshare.com/s/d8505ab07724a111b1f3)[25], where researchers can access the required files. Additionally, the expression profiles are available on the Gene Expression Omnibus (GEO) under accession numbers GSE284510[26], GSE284516[27], GSE284517[28], and GSE284518[29]. The variant data are accessible through the European Variation Archive (EVA)[30]. On Figshare, we have organized the data into three main categories: genomic data, gene expression data, and variant data.

The genomic data category includes comprehensive functional annotations for all collected genes. These annotations cover homologous genes identified by BLAST, protein functional domains, AHRD-based functional descriptions, and GO/KEGG annotations. The files are stored in a compressed archive named "genome_anno.tgz," with subfolders named according to the respective genomes. Additionally, the genome files for HBY-2 have also been uploaded in this section. The gene expression data category consists of both the raw and normalized expression matrices. These files are stored in a compressed archive named "gene_expression.zip." The variant data category includes VCF files for variant data, all starting with "variants_data." It encompasses 330 samples from the GBS (Genotyping-by-Sequencing) and resequencing projects.

Additionally, all genomic sequences, functional annotations, expression profiles, and called variant data can be downloaded from the CGD under the "Tools - > Download" module. The sample information for the hierarchical clustering heatmap can be found in Supplementary Table 1. Notably, the updated data for HBY-2 has also been shared on Figshare (https://figshare.com/articles/dataset/Wild_Chinese_chestnut_genome_V2_/28098758).

By organizing the data in this structured manner, we aim to facilitate easy access and usability for researchers interested in chestnut genomics and related studies.

Technical Validation

The integrity of genomes.

We evaluated genome integrity using BUSCO (version 3)[31], the BUSCO analysis is based on conserved orthologous genes among species. In this study, simultaneous analysis of the genomes of eight varieties was conducted, and the proportions of identified complete genes in the BUSCO database (with a total of 1,614 genes) were 98.0%, 97.6%, 94.0%, 98.5%, 95.6%, 94.3%, 98.5%, and 92.6%, respectively (Table 2), the results indicate that the genomes integrity are satisfactory.

RNA-Seq data.

Among the 213 samples, 91.81% of the clean reads were mapped to the reference genome (HBY-2). To ensure the accurate reflection of expression across diverse tissues and developmental stages, hierarchical clustering analysis (Fig. 1) based on FPKMs was conducted using R, derived from various RNA-seq items of chestnut (Supplementary Table 1). The heatmap is divided into four color blocks from top to bottom: the first block represents seed kernel and embryo at relatively early stages, the second block represents leaves, buds, galled leaves, etc., at similar stages, the third block indicates seed kernel and embryo at relatively later stages, and the fourth block includes somatic embryo, embryo, root, and callus at similar stages. Among the 213 samples shown in the heatmap, nine are not clustered within the expected blocks (marked with red horizontal lines in the figure). Overall, similarly developmental stages originating from the same tissue exhibited cohesive clustering patterns. For example, the seed kernel is similar to the tissue of the embryo, and at similar developmental stages, they exhibit cohesive clustering patterns. The same reasoning applies to galled leaves, leaf, insect galls, and some of the tissue of the bud, which indicates the reliability of the data we collected.

Resequencing data.

To verify the accuracy of Sentieon in detecting varinsts, several samples were randomly selected from the resequencing data and GATK pipeline was used to identify SNPs on them. Briefly, the cleaned reads were aligned to the reference genomes (HBY-2) using the BWA-MEM algorithm with default parameters[24]. Next, the aligned reads were processed to remove duplicated reads using the MarkDuplicates algorithm from the Picard. After mapping, 93.81% of clean reads were mapped to the reference genome (HBY-2). The variant were called using the 'HaplotypeCaller' algorithms of GATK3.8, cross-validated with Sentieon, yielding results of 12,631,432 and 12,314,460 respectively (Table 3). The concordance rate of common SNPs identified by Sentieon and GATK3.8 software packages stands at 99.943% and 97.435%, respectively, indicate a high level of consistency in the results.

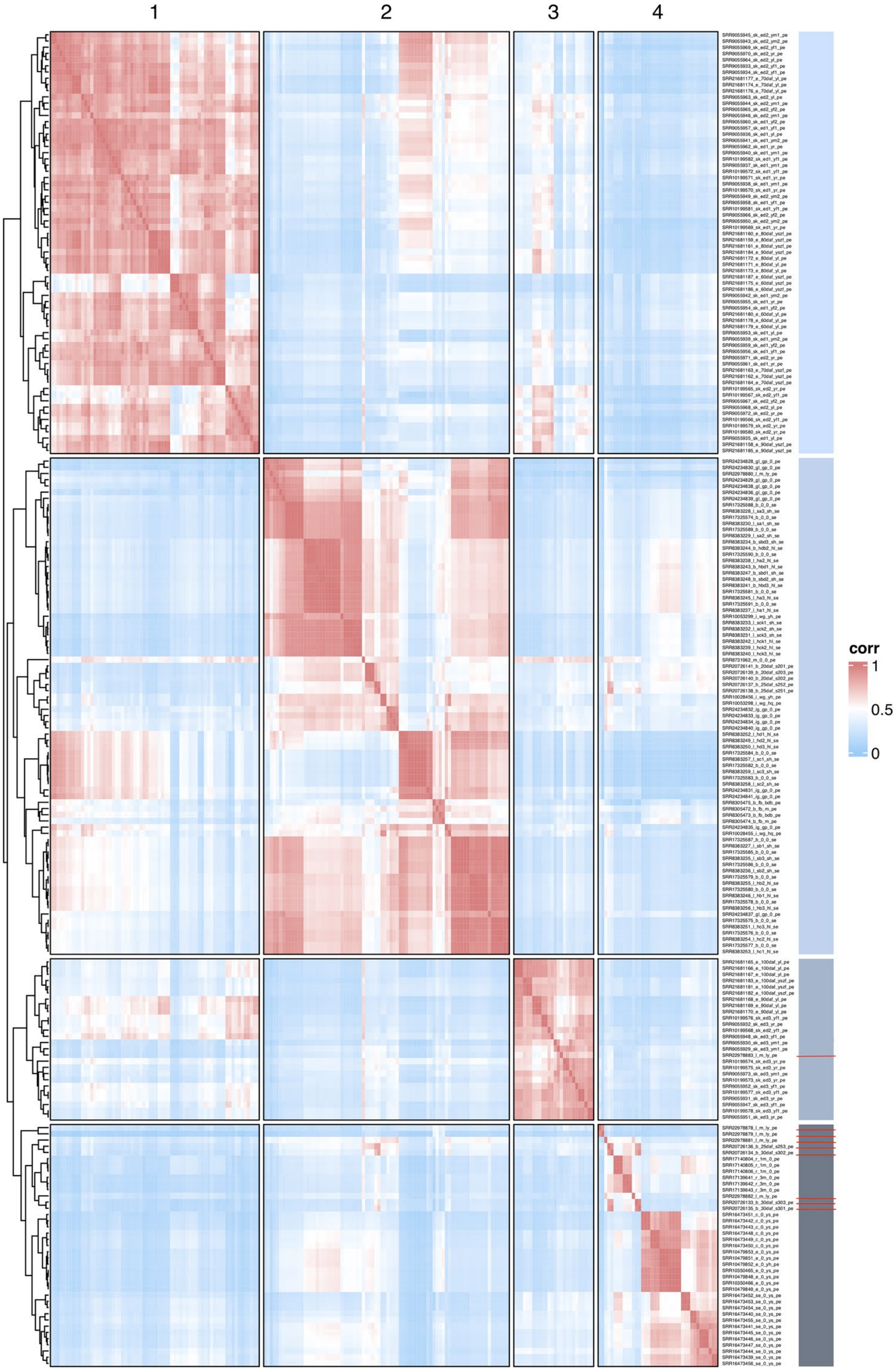


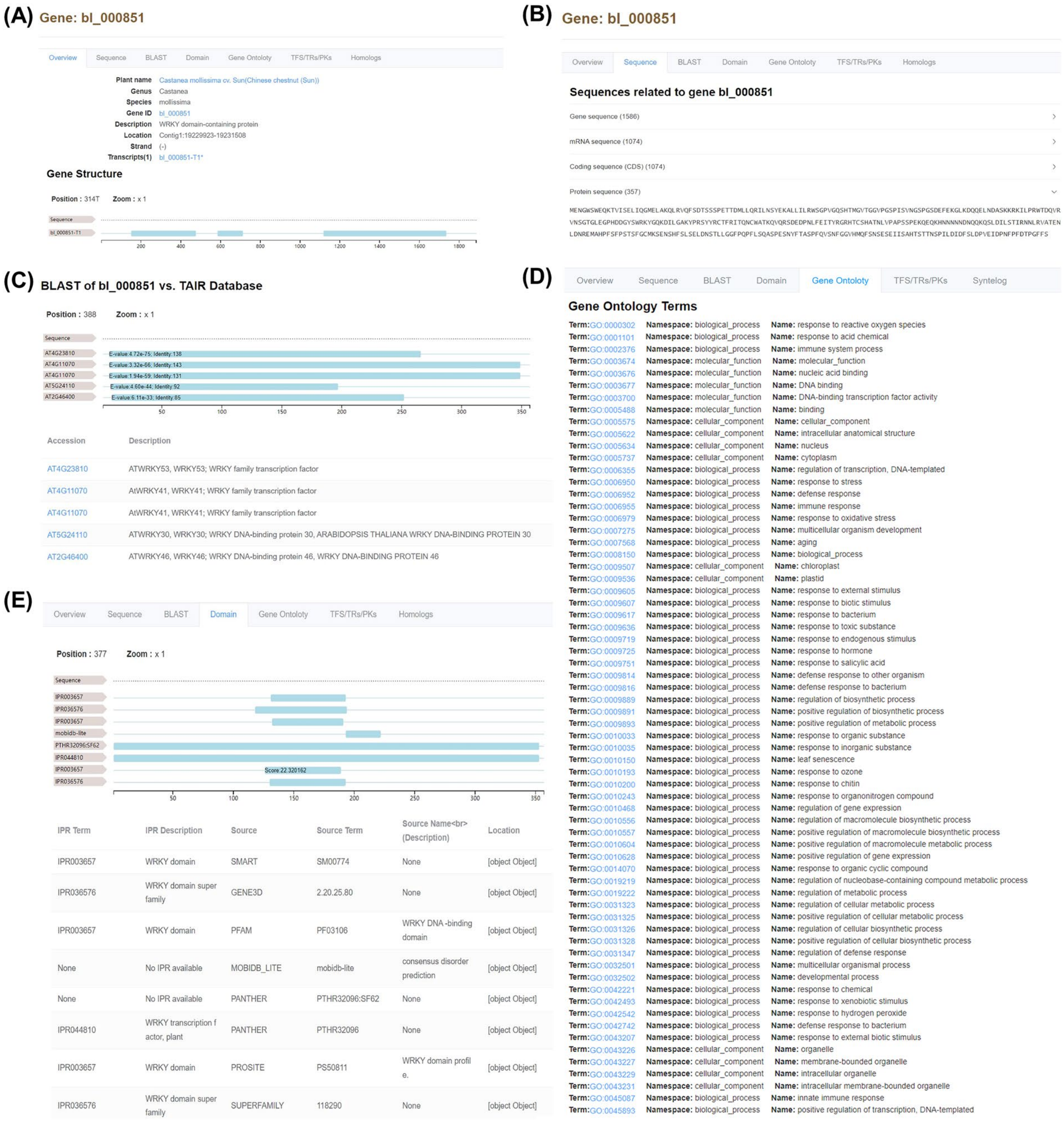
Fig. 1 The heatmap of hierarchical clustering in diferent RNA-seq samples based on FPKM. The four color blocks from top to bottom respectively represent: (1) seed kernel and embryo at relatively early stages, (2) leaves at similar stages, buds, galled leaves, etc., (3) seed kernel and embryo at relatively later stages, (4) somatic embryo at similar stages, embryo, root and callus, etc. Note: There is one cluster of leaf tissue in the third color block, and four clusters of leaf tissue and four clusters of bud tissue in the fourth color block, all marked with red lines on the blocks.

Usage Notes

The CGD serves as a complementary platform to primary datasets hosted on Figshare and other repositories, enhancing data accessibility and utility. The following sections provide a concise overview of how the CGD enables efficient analysis and exploration of the aforementioned datasets.

	Sentieon	GATK
Total	12,314,460	12,631,432
Common	12,307,411 (99.943%)	12,307,411 (97.435%)
Specific	234	317,206
Non-matching overlapping	6815	6815

Table 3. Statistics of SNPs identified by Sentieon and GATK.



Gene: bl_000851

Plant name Castanea mollissima cv. Sun(Chinese chestnut (Sun))
Genus Castanea
Species mollissima
Gene ID bl_000851
Description WRKY domain-containing protein
Location Contig1:19229923-19231508
Strand (-)
Transcripts(1) bl_000851-T1*

Gene Structure

Position : 314T Zoom : x 1

Gene: bl_000851

Sequences related to gene bl_000851

Gene sequence (1586)

mRNA sequence (1074)

Coding sequence (CDS) (1074)

Protein sequence (357)

```
MENGWSWEQKTVISELIQGMELAKQLRVQFSDTSSSPETTDMLLQRILNSYEKALLILRWSGPVGQSHTMGVTGGVPGSPISVNGSPGSDEFEKGLKDQQELNDASKKRKILPRWTDQVR
VNSGTGLEGPHDDGYSWRKYGQKDILGAKYPRSYYRCTFRITQNCWATKQVQRSDEDPNLFEITYRGRHTCSHATNLVPAPSSPEKQEQKHNNNNNDNQQKQSLDILSTIRNNLRVATEN
LDNREMAHPFSFPSTSFGCMKSENSHFSLSELDNSTLLGGFPQPFLSQASPESNYFTASPFQVSNFGGVHMQFSNSESEIISAHTSTTNSPILDIDFSLDPVEIDPNFPFDTPGFFS
```

BLAST of bl_000851 vs. TAIR Database

Position : 388 Zoom : x 1

Accession	Description
AT4G23810	ATWRKY53, WRKY53; WRKY family transcription factor
AT4G11070	AtWRKY41, WRKY41; WRKY family transcription factor
AT4G11070	AtWRKY41, WRKY41; WRKY family transcription factor
AT5G24110	ATWRKY30, WRKY30; WRKY DNA-binding protein 30, ARABIDOPSIS THALIANA WRKY DNA-BINDING PROTEIN 30
AT2G46400	ATWRKY46, WRKY46; WRKY DNA-binding protein 46, WRKY DNA-BINDING PROTEIN 46

Gene Ontology Terms

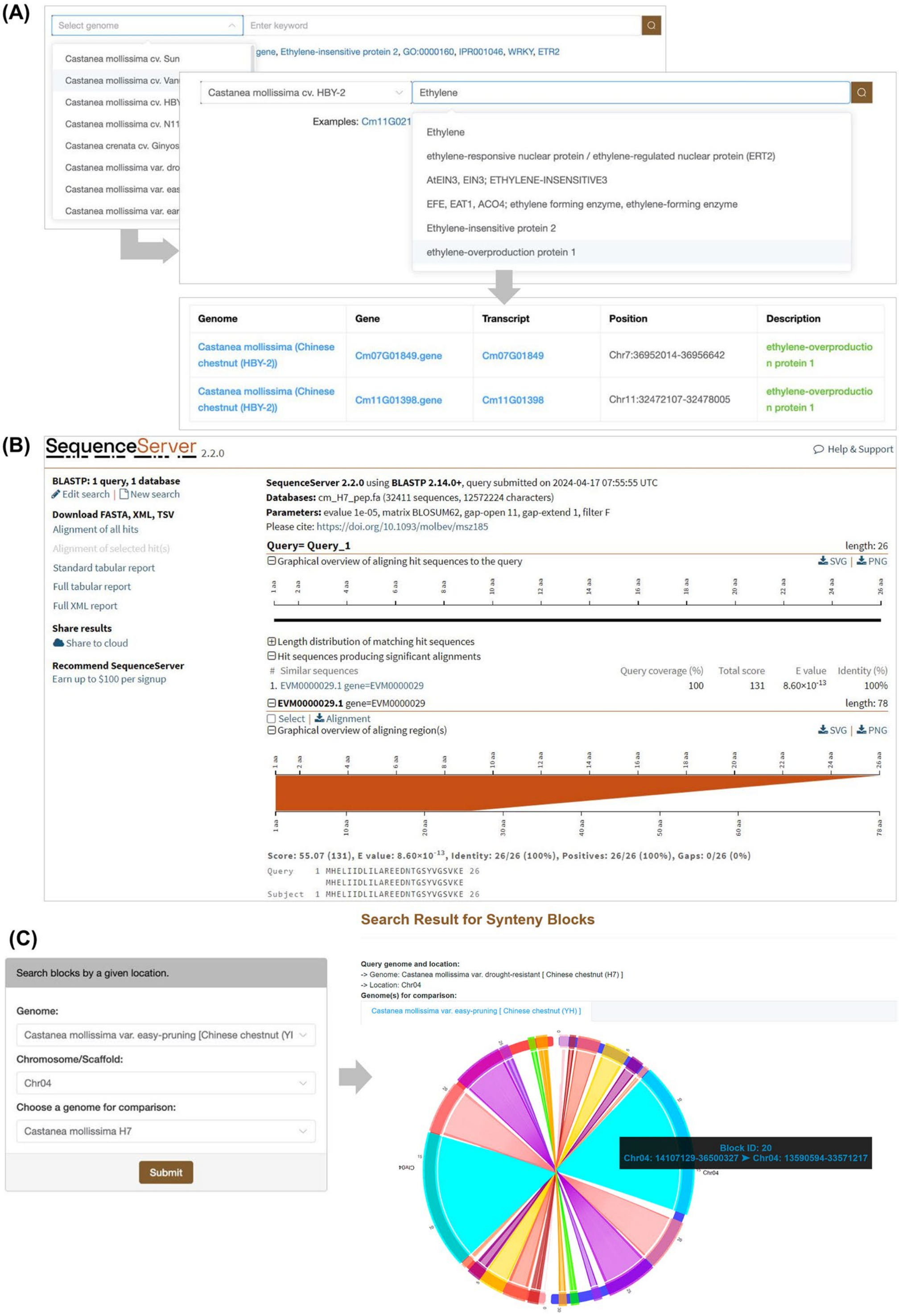
Term:GO:0000302 Namespace: biological_process Name: response to reactive oxygen species
Term:GO:0001101 Namespace: biological_process Name: response to acid chemical
Term:GO:0002376 Namespace: biological_process Name: immune system process
Term:GO:0003674 Namespace: molecular_function Name: molecular_function
Term:GO:0003676 Namespace: molecular_function Name: nucleic acid binding
Term:GO:0003677 Namespace: molecular_function Name: DNA binding
Term:GO:0003700 Namespace: molecular_function Name: DNA-binding transcription factor activity
Term:GO:0005488 Namespace: molecular_function Name: binding
Term:GO:0005575 Namespace: cellular_component Name: cellular_component
Term:GO:0005622 Namespace: cellular_component Name: intracellular anatomical structure
Term:GO:0005634 Namespace: cellular_component Name: nucleus
Term:GO:0005737 Namespace: cellular_component Name: cytoplasm
Term:GO:0006355 Namespace: biological_process Name: regulation of transcription, DNA-templated
Term:GO:0006950 Namespace: biological_process Name: response to stress
Term:GO:0006952 Namespace: biological_process Name: defense response
Term:GO:0006955 Namespace: biological_process Name: immune response
Term:GO:0006979 Namespace: biological_process Name: response to oxidative stress
Term:GO:0007275 Namespace: biological_process Name: multicellular organism development
Term:GO:0007568 Namespace: biological_process Name: aging
Term:GO:0008150 Namespace: biological_process Name: biological_process
Term:GO:0009507 Namespace: cellular_component Name: chloroplast
Term:GO:0009536 Namespace: cellular_component Name: plastid
Term:GO:0009605 Namespace: biological_process Name: response to external stimulus
Term:GO:0009607 Namespace: biological_process Name: response to biotic stimulus
Term:GO:0009617 Namespace: biological_process Name: response to bacterium
Term:GO:0009636 Namespace: biological_process Name: response to toxic substance
Term:GO:0009719 Namespace: biological_process Name: response to endogenous stimulus
Term:GO:0009725 Namespace: biological_process Name: response to hormone
Term:GO:0009751 Namespace: biological_process Name: response to salicylic acid
Term:GO:0009814 Namespace: biological_process Name: defense response to other organism
Term:GO:0009816 Namespace: biological_process Name: defense response to bacterium
Term:GO:0009889 Namespace: biological_process Name: regulation of biosynthetic process
Term:GO:0009891 Namespace: biological_process Name: positive regulation of biosynthetic process
Term:GO:0009893 Namespace: biological_process Name: positive regulation of metabolic process
Term:GO:0010033 Namespace: biological_process Name: response to organic substance
Term:GO:0010035 Namespace: biological_process Name: response to inorganic substance
Term:GO:0010150 Namespace: biological_process Name: leaf senescence
Term:GO:0010193 Namespace: biological_process Name: response to ozone
Term:GO:0010200 Namespace: biological_process Name: response to chitin
Term:GO:0010243 Namespace: biological_process Name: response to organonitrogen compound
Term:GO:0010468 Namespace: biological_process Name: regulation of gene expression
Term:GO:0010556 Namespace: biological_process Name: regulation of macromolecule biosynthetic process
Term:GO:0010557 Namespace: biological_process Name: positive regulation of macromolecule biosynthetic process
Term:GO:0010604 Namespace: biological_process Name: positive regulation of macromolecule metabolic process
Term:GO:0010628 Namespace: biological_process Name: positive regulation of gene expression
Term:GO:0014070 Namespace: biological_process Name: response to organic cyclic compound
Term:GO:0019219 Namespace: biological_process Name: regulation of nucleobase-containing compound metabolic process
Term:GO:0019222 Namespace: biological_process Name: regulation of metabolic process
Term:GO:0031323 Namespace: biological_process Name: regulation of cellular metabolic process
Term:GO:0031325 Namespace: biological_process Name: positive regulation of cellular metabolic process
Term:GO:0031326 Namespace: biological_process Name: regulation of cellular biosynthetic process
Term:GO:0031328 Namespace: biological_process Name: positive regulation of cellular biosynthetic process
Term:GO:0031347 Namespace: biological_process Name: regulation of defense response
Term:GO:0032501 Namespace: biological_process Name: multicellular organismal process
Term:GO:0032502 Namespace: biological_process Name: developmental process
Term:GO:0042221 Namespace: biological_process Name: response to chemical
Term:GO:0042493 Namespace: biological_process Name: response to xenobiotic stimulus
Term:GO:0042542 Namespace: biological_process Name: response to hydrogen peroxide
Term:GO:0042742 Namespace: biological_process Name: defense response to bacterium
Term:GO:0043207 Namespace: biological_process Name: response to external biotic stimulus
Term:GO:0043226 Namespace: cellular_component Name: organelle
Term:GO:0043227 Namespace: cellular_component Name: membrane-bounded organelle
Term:GO:0043229 Namespace: cellular_component Name: intracellular organelle
Term:GO:0043231 Namespace: cellular_component Name: intracellular membrane-bounded organelle
Term:GO:0045087 Namespace: biological_process Name: innate immune response
Term:GO:0045893 Namespace: biological_process Name: positive regulation of transcription, DNA-templated

Position : 377 Zoom : x 1

IPR Term	IPR Description	Source	Source Term	Source Name (Description)	Location
IPR003657	WRKY domain	SMART	SM00774	None	[object Object]
IPR036576	WRKY domain super family	GENE3D	2.20.25.80	None	[object Object]
IPR003657	WRKY domain	PFAM	PF03106	WRKY DNA -binding domain	[object Object]
None	No IPR available	MOBIDB_LITE	mobidb-lite	consensus disorder prediction	[object Object]
None	No IPR available	PANTHER	PTHR32096:SF62	None	[object Object]
IPR044810	WRKY transcription f actor, plant	PANTHER	PTHR32096	None	[object Object]
IPR003657	WRKY domain	PROSITE	PS50811	WRKY domain profil e.	[object Object]
IPR036576	WRKY domain super family	SUPERFAMILY	118290	None	[object Object]

Fig. 2 Gene feature page in Castanea Genome Database. (**A**) Screenshot of the gene page including basic information and gene structure. (**B**) Screenshot of the gene page containing gene, mRNA, CDS, and protein sequences. (**C**) Screenshot of the homolog genes and sequence alignments from the BLAST results. (**D**) Screenshot of the GO terms assigned to the gene. (**E**) Screenshot of the functional domains predicted from the protein sequence of the gene.

Each gene feature page in the CGD offers basic gene details, an interactive genome browser, and access to mRNA, protein, and genomic sequences, along with homologous genes, functional annotations, protein domains, and syntenic blocks (Fig. 2). To facilitate dataset storage, browsing, and querying, we have


(A)
Select genome
Enter keyword
gene, Ethylene-insensitive protein 2, GO:0000160, IPR001046, WRKY, ETR2
Castanea mollissima cv. Sun
Castanea mollissima cv. Van
Castanea mollissima cv. HBY
Castanea mollissima cv. N11
Castanea crenata cv. Ginyos
Castanea mollissima var. dro
Castanea mollissima var. eas
Castanea mollissima var. ear
Castanea mollissima cv. HBY-2
Ethylene
Examples: Cm11G021
Ethylene
ethylene-responsive nuclear protein / ethylene-regulated nuclear protein (ERT2)
AtEIN3, EIN3; ETHYLENE-INSENSITIVE3
EFE, EAT1, ACO4; ethylene forming enzyme, ethylene-forming enzyme
Ethylene-insensitive protein 2
ethylene-overproduction protein 1

Genome	Gene	Transcript	Position	Description
Castanea mollissima (Chinese chestnut (HBY-2))	Cm07G01849.gene	Cm07G01849	Chr7:36952014-36956642	ethylene-overproduction protein 1
Castanea mollissima (Chinese chestnut (HBY-2))	Cm11G01398.gene	Cm11G01398	Chr11:32472107-32478005	ethylene-overproduction protein 1

(B)
SequenceServer 2.2.0
Help & Support
BLASTP: 1 query, 1 database
Edit search | New search
Download FASTA, XML, TSV
Alignment of all hits
Alignment of selected hit(s)
Standard tabular report
Full tabular report
Full XML report
Share results
Share to cloud
Recommend SequenceServer
Earn up to $100 per signup

SequenceServer 2.2.0 using **BLASTP 2.14.0+**, query submitted on 2024-04-17 07:55:55 UTC
Databases: cm_H7_pep.fa (32411 sequences, 12572224 characters)
Parameters: evalue 1e-05, matrix BLOSUM62, gap-open 11, gap-extend 1, filter F
Please cite: https://doi.org/10.1093/molbev/msz185

Query= Query_1 length: 26

⊟ Graphical overview of aligning hit sequences to the query
SVG | PNG
1 aa
2 aa
4 aa
6 aa
8 aa
10 aa
12 aa
14 aa
16 aa
18 aa
20 aa
22 aa
24 aa
26 aa
⊞ Length distribution of matching hit sequences
⊟ Hit sequences producing significant alignments

#	Similar sequences	Query coverage (%)	Total score	E value	Identity (%)
1.	EVM0000029.1 gene=EVM0000029	100	131	8.60×10^{-13}	100%

⊟ **EVM0000029.1** gene=EVM0000029 length: 78

```
Score: 55.07 (131), E value: 8.60×10^-13, Identity: 26/26 (100%), Positives: 26/26 (100%), Gaps: 0/26 (0%)
Query    1 MHELIIDLILAREEDNTGSYVGSVKE 26
           MHELIIDLILAREEDNTGSYVGSVKE
Subject  1 MHELIIDLILAREEDNTGSYVGSVKE 26
```

(C)
Search blocks by a given location.
Submit
Search Result for Synteny Blocks
Chr04
Chr04
Block ID: 20
Chr04: 14107129-36500327 ➤ Chr04: 13590594-33571217
0
5
10
15
20
25
30

Genome: Castanea mollissima var. easy-pruning [Chinese chestnut (YI
Chromosome/Scaffold: Chr04
Choose a genome for comparison: Castanea mollissima H7

Query genome and location:
-> Genome: Castanea mollissima var. drought-resistant [Chinese chestnut (H7)]
-> Location: Chr04
Genome(s) for comparison:
Castanea mollissima var. easy-pruning [Chinese chestnut (YH)]

Fig. 3 Query interfaces of data mining tools in Castanea Genome Database. (**A**) Screenshot of the 'Search' interface and result page. (**B**) Screenshot of the 'BLAST' result page. (**C**) Screenshot of 'Synteny Viewer' result page.

implemented a versatile gene search tool that supports flexible queries based on selected genomes from a drop-down menu. Users can quickly locate specific genes of interest by entering keywords such as gene names, functional descriptions, transcription factor or protein kinase family names, or GO/Pfam terms. The search interface is user-friendly, featuring auto-suggestions to streamline the selection process (Fig. 3A).

The CGD also offers a suite of advanced data mining and analysis tools to maximize the utility of our datasets for comparative genomics, gene function discovery, and molecular breeding. For enrichment analysis, BLAST-indexed databases are organized into nucleotide and protein categories. Nucleotide databases include indexes for genomic, mRNA, and CDS sequences, while protein databases cover all available protein sequence indexes (Fig. 3B).

Furthermore, we have introduced the Synteny Viewer, a tool designed to simplify the identification and visualization of homologs within specific regions of the Castanea genome (Fig. 3C).

For more information, please visit the CGD at http://castaneadb.net. Detailed descriptions of the site's features and functionalities are available directly on the platform.

Code availability

The gene functional annotation pipeline is available on GitHub (https://github.com/fjl23/CGD_annotation-pipeline). All software and pipelines were executed according to the manual and protocols of the published bioinformatics tools. The version and code/parameters of software have been detailed and described in Methods.

Received: 13 May 2024; Accepted: 9 May 2025;
Published online: 24 May 2025

References

1. Bounous, G. The chestnut: a multipurpose resource for the new millennium. *III International Chestnut Congress* **693**, 33–40 (2005).
2. Li, Q. *et al*. Effect of cooking methods on nutritional quality and volatile compounds of Chinese chestnut (Castanea mollissima Blume). *Food Chemistry* **201**, 80–86 (2016).
3. Ji, F. *et al*. Construction of a SNP-Based High-Density Genetic Map Using Genotyping by Sequencing (GBS) and QTL Analysis of Nut Traits in Chinese Chestnut (Castanea mollissima Blume). *Frontiers in plant science* **9**, 816 (2018).
4. Zhang, Y., Liu, L., Liang, W. & Zhang, Y. China Fruit Monograph: Chinese Chestnut and Chinese Hazelnut Volume. (China Forestry Press, 2005).
5. Food and Agriculture Organization of the United Nations. *FAOSTAT statistics database*. https://www.fao.org/faostat/en/#data/QCL (2018).
6. Xing, Y. *et al*. Hybrid de novo genome assembly of Chinese chestnut (Castanea mollissima). *GigaScience* **8** (2019).
7. Liu, Y. *et al*. Beta-amylase and phosphatidic acid involved in recalcitrant seed germination of Chinese chestnut. *Frontiers in plant science* **13**, 828270 (2022).
8. Wang, J. *et al*. Construction of Pseudomolecules for the Chinese Chestnut (Castanea mollissima) Genome. *G3 (Bethesda, Md.)* **10**, 3565–3574 (2020).
9. Sun, Y., Lu, Z., Zhu, X. & Ma, H. Genomic basis of homoploid hybrid speciation within chestnut trees. *Nature communications* **11**, 3375 (2020).
10. Staton, M. *et al*. A reference genome assembly and adaptive trait analysis of Castanea mollissima 'Vanuxem,'a source of resistance to chestnut blight in restoration breeding. *Tree Genetics & Genomes* **16**, 1–23 (2020).
11. Hu, G. *et al*. Pan-genome analysis of three main Chinese chestnut varieties. *Frontiers in plant science* **13**, 916550 (2022).
12. Shirasawa, K. *et al*. Chromosome-level genome assembly of Japanese chestnut (Castanea crenata Sieb. et Zucc.) reveals conserved chromosomal segments in woody rosids. *DNA Research* **28** (2021).
13. Staton, M. *et al*. Substantial genome synteny preservation among woody angiosperm species: comparative genomics of Chinese chestnut (Castanea mollissima) and plant reference genomes. *BMC genomics* **16**, 744 (2015).
14. Zheng, Y. *et al*. Cucurbit Genomics Database (CuGenDB): a central portal for comparative and functional genomics of cucurbit crops. *Nucleic acids research* **47**, D1128–d1136 (2019).
15. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nature methods* **12**, 59–60 (2015).
16. Mitchell, A. L. *et al*. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic acids research* **47**, D351–d360 (2019).
17. Huerta-Cepas, J. *et al*. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic acids research* **47**, D309–d314 (2019).
18. Zheng, Y. *et al*. iTAK: A Program for Genome-wide Prediction and Classification of Plant Transcription Factors, Transcriptional Regulators, and Protein Kinases. *Molecular plant* **9**, 1667–1670 (2016).
19. Brown, J., Pirrung, M. & McCue, L. A. FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool. *Bioinformatics (Oxford, England)* **33**, 3137–3139 (2017).
20. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)* **30**, 2114–2120 (2014).
21. Dobin, A. *et al*. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)* **29**, 15–21 (2013).
22. Florea, L., Song, L. & Salzberg, S. L. Thousands of exon skipping events differentiate among splicing patterns in sixteen human tissues. *F1000Research* **2**, 188 (2013).
23. Freed, D., Aldana, R., Weber, J. A. & Edwards, J. S. The Sentieon Genomics Tools–A fast and accurate solution to variant calling from next-generation sequence data. *bioRxiv*, 115717 (2017).
24. Heng, L. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://arxiv.org/abs/1303.3997 (2013).
25. Zheng, Y. *Castanea* Genome Database (CGD): a comprehensive database of chestnut genome. *figshare* https://doi.org/10.6084/m9.figshare.27060067 (2024).
26. *NCBI Gene Expression Omnibus (GEO)* https://identifiers.org/geo:GSE284510 (2024).
27. *NCBI Gene Expression Omnibus (GEO)* https://identifiers.org/geo:GSE284516 (2024).
28. *NCBI Gene Expression Omnibus (GEO)* https://identifiers.org/geo:GSE284517 (2024).
29. *NCBI Gene Expression Omnibus (GEO)* https://identifiers.org/geo:GSE284518 (2024).
30. *European Variation Archive (EVA)* https://identifiers.org/ena.embl:ERP170716 (2025).
31. Waterhouse, R. M. *et al*. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Molecular biology and evolution* **35**, 543–548 (2018).

Acknowledgements

This work was supported by grants from the National Key Research & Development Program of China (2018YFD1000605) and 2020 Connotation Development Quota Project of Beijing University of Agriculture (5066516001/003) to Y.Xing; as well as Beijing University of Agriculture Start-up fund and Enhancing Technology Innovation Capability Project of Beijing University of Agriculture (5076016192/054) to Y.Zheng.

Author contributions

Y.Xing and Y.Zheng designed and managed the project; Y.Zhang, S.Wei and J.Fan constructed the database; J.Fan, Y.Zhang, X.Nie, Y.Liu, H.Peng, H.Li, M.Zhang, L.Ning and S.Wang collected and analyzed the data; J.Fan, Y.Zheng and Y.Xing wrote and revised the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-025-05162-x.

Correspondence and requests for materials should be addressed to Y.Z. or Y.X.

Reprints and permissions information is available at www.nature.com/reprints.

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