Published online 22 October 2022 *Nucleic Acids Research, 2023, Vol. 51, Database issue* ***D1457–D1464***
https://doi.org/10.1093/nar/gkac921

CuGenDBv2: an updated database for cucurbit genomics

Jingyin Yu[1], Shan Wu[1], Honghe Sun[1,2], Xin Wang[3], Xuemei Tang[1], Shaogui Guo[4], Zhonghua Zhang[5], Sanwen Huang[6], Yong Xu[4], Yiqun Weng[7,8], Michael Mazourek[9], Cecilia McGregor[10], Susanne S. Renner[11,12], Sandra Branham[13], Chandrasekar Kousik[14], W. Patrick Wechter[14], Amnon Levi[14], Rebecca Grumet[15], Yi Zheng[16,17] and Zhangjun Fei[1,18,*]

[1]Boyce Thompson Institute, Ithaca, NY 14853, USA, [2]Plant Biology Section, School of Integrative Plant Science, Cornell University, Ithaca, NY 14853, USA, [3]College of Horticulture and Forestry Sciences, Huazhong Agricultural University, Wuhan 430070, China, [4]National Watermelon and Melon Improvement Center, Beijing Academy of Agricultural and Forestry Sciences, Key Laboratory of Biology and Genetic Improvement of Horticultural Crops (North China), Beijing Key Laboratory of Vegetable Germplasm Improvement, Beijing 100097, China, [5]Engineering Laboratory of Genetic Improvement of Horticultural Crops of Shandong Province, College of Horticulture, Qingdao Agricultural University, Qingdao 266109, China, [6]Genome Analysis Laboratory of the Ministry of Agriculture, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, Guangdong 518124, China, [7]U.S. Department of Agriculture-Agricultural Research Service, Vegetable Crops Research Unit, Madison, WI 53706, USA, [8]Department of Horticulture, University of Wisconsin, Madison, WI 53706, USA, [9]Plant Breeding and Genetics Section, School of Integrative Plant Science, Cornell University, Ithaca, NY 14853, USA, [10]Department of Horticulture, University of Georgia, Athens, GA 30602, USA, [11]Faculty of Biology, Systematic Botany and Mycology, University of Munich (LMU), 80638 Munich, Germany, [12]Department of Biology, Washington University, Saint Louis, MO 63130, USA, [13]Coastal Research and Educational Center, Clemson University, Charleston, SC 29414, USA, [14]U.S. Department of Agriculture-Agricultural Research Service, U.S. Vegetable Laboratory, 2700 Savannah Highway, Charleston, SC 29414, USA, [15]Department of Horticulture, Michigan State University, East Lansing, MI 48824, USA, [16]Beijing Key Laboratory for Agricultural Application and New Technique, College of Plant Science and Technology, Beijing University of Agriculture, Beijing 102206, China, [17]Bioinformatics Center, Beijing University of Agriculture, Beijing 102206, China and [18]U.S. Department of Agriculture-Agricultural Research Service, Robert W. Holley Center for Agriculture and Health, Ithaca, NY 14853, USA

Received September 11, 2022; Revised October 03, 2022; Editorial Decision October 04, 2022; Accepted October 06, 2022

ABSTRACT

The Cucurbitaceae (cucurbit) family consists of about 1,000 species in 95 genera, including many economically important and popular fruit and vegetable crops. During the past several years, reference genomes have been generated for >20 cucurbit species, and variome and transcriptome profiling data have been rapidly accumulated for cucurbits. To efficiently mine, analyze and disseminate these large-scale datasets, we have developed an updated version of Cucurbit Genomics Database. The updated database, CuGenDBv2 (http://cucurbitgenomics.org/v2), currently hosts 34 reference genomes from 27 cucurbit species/subspecies belonging to 10 different genera. Protein-coding genes from these genomes have been comprehensively annotated by comparing their protein sequences to various public protein and domain databases. A novel 'Genotype' module has been implemented to facilitate mining and analysis of the functionally annotated variome data including SNPs and small indels from large-scale genome sequencing projects. An updated 'Expression' module has been developed to provide a comprehensive gene expression atlas for cucurbits. Furthermore, synteny blocks between any two and within each of the 34 genomes, representing a total of 595 pair-wise genome comparisons, have been identified and can be explored and visualized in the database.

*To whom correspondence should be addressed. Tel: +1 607 2543234; Fax: +1 607 2541242; Email: zf25@cornell.edu

INTRODUCTION

The Cucurbitaceae (cucurbit) family consists of about 1000 species in 95 genera, mainly grown in tropical, subtropical and temperate regions around the world (1,2). The family includes numerous important fruit and vegetable crops with high nutrition and flavor values such as cucumber, melon, watermelon, squash, pumpkin etc. In addition, some cucurbits can also be used as containers, musical instruments and sources of oils, and serve as ornaments for festivals, medicines for disorder treatment, as well as model systems for the study of sex determination (3–5). Due to their importance, abundant genetic and genomic resources have been developed for various cucurbit plants during the past 15 years or so, with cucumber representing the first fruit or vegetable crop that had a genome sequence, which was released in 2009 (6).

We have developed the Cucurbit Genomics Database (CuGenDB), which serves as a central portal for cucurbit comparative and functional genomics (7). Since the release of CuGenDB in 2019, thanks to the rapid advances in sequencing technologies, novel or improved reference genomes have been generated for a number of cucurbit species and variety groups. In addition, gene expression profiling data generated using RNA sequencing (RNA-Seq) have been rapidly accumulated for cucurbit species, which have provided broad insights into molecular mechanisms underlying biotic and abiotic stresses, and plant growth and development. Furthermore, high-resolution genomic variants including single nucleotide polymorphisms (SNPs) and small insertions and deletions (indels) have been generated for various cucurbit populations, which have helped to understand the genetic diversity, origin, and domestication of these cucurbit crops, as well as genetic bases of key cucurbit agronomic traits. A platform for efficient distribution, mining and analysis of these newly generated genomic data would benefit the plant research and breeding community. Therefore, an updated CuGenDB to embrace these data and additional associated data mining functions is urgently needed.

To this end, we have developed an updated version of CuGenDB, CuGenDBv2 (http://cucurbitgenomics.org/v2), mainly using Tripal v3.0, which, compared to Tripal v2.0, has substantially improved the efficiency of genomic data loading into the backend PostgreSQL database tables (several hours using Tripal v3.0 versus several weeks per genome using Tripal v2.0) (8). The web interfaces in CuGenDBv2 have been built using the legacy functionalities in Tripal v2.0 (9). In this way, genomic data can be loaded into the PostgreSQL database quickly and the web interfaces customized efficiently. Some modules/functions in CuGenDBv2, such as 'Expression', 'Genotype' and 'Synteny Viewer', have been implemented using Perl/CGI combined with the backend MySQL database.

DATABASE CONTENTS

Cucurbit genome assemblies and syntenies

A total of 34 genome assemblies are currently available in CuGenDBv2. These assemblies are from 27 different species or subspecies that belong to 10 genera in the cucurbit family: *Cucumis*, *Citrullus*, *Cucurbita*, *Luffa*, *Momordica*, *Lagenaria*, *Benincasa*, *Sechium*, *Siraitia* and *Trichosanthes* (Table 1). Ten genome assemblies from the genus *Cucumis* are available in CuGenDBv2, including five from cucumber and six from melon. For cucumber, three are from cultivated species *C. sativus* var. *sativus* (Chinese Long, Gy14 and B10), one from the wild progenitor *C. sativus* var. *hardwickii* (PI 183967), and one from the distant wild relative *C. hystrix* (10–13). For melon, five are from different subspecies or variety groups of cultivated species *C. melo* (DHL92, ssp. *agrestis* IVF77, var. *inodorus* Payzawat, var. *reticulatus* Harukei-3 and var. *cantalupensis Charmono*) and one from the wild melon *C. metuliferus* (14–18). For the genus *Citrullus*, six genome assemblies are available, including two from the cultivated watermelon *C. lanatus* ssp. *vulgaris*, an East Asia ecotype 97103 (19) and an America ecotype Charleston Gray (20), one from the possible direct wild progenitor *C. lanatus* ssp. *cordophanus* (21) and one each from the three wild *Citrullus* species, *C. mucosospermus* (USVL531-MDR), *C. amarus* (USVL246-FR2) and *C. colocynthis* (PI 537277). For the genus *Cucurbita*, five genome assemblies are available, with one each from the four cultivated species, *C. maxima* (Rimu), *C. moschata* (Rifu), *C. pepo* (MU-CU-16), *C. argyrosperma* ssp. *argyrosperma* (SMH-JMG-627), and one wild relative, *C. argyrosperma* subsp. *sororia* (22–24). For *Luffa*, two genome assemblies from the cultivated sponge gourd species *L. cylindrica* including that of cultivar P93075, and one from another cultivated species *L. acutangula* (AG-4) are available (25–27). For *Momordica*, two genome assemblies from the cultivated bitter gourd species *M. charantia* (Dali-11 and OHB3-1) and one from a small-fruited wild line, TR, are available (28,29). For *Lagenaria*, one genome assembly from the food-type bottle gourd *L. siceraria* (Hangzhou gourd) and another from the rootstock-type bottle gourd (USVL1VR-Ls) are available (30,31). For the genera *Benincasa*, *Sechium*, *Siraitia*, and *Trichosanthes*, one genome assembly is available for each of the cultivated species in the four genera: *Benincasa hispida* wax gourd (B227), *Sechium edule* chayote, *Siraitia grosvenorii* monk fruit (Qingpiguo) and *Trichosanthes anguina* snake gourd (32–35). Among the 34 genome assemblies currently in CuGenDBv2, four have not been published in literature and are first released in the database, including the cucumber Gy14 genome and genomes of three wild watermelons (*C. mucosospermus* USVL531-MDR, *C. amarus* USVL246-FR2 and *C. colocynthis* PI 537277).

Genomic synteny blocks and syntenic gene pairs have been identified between any two and within each of the 34 cucurbit genome assemblies, representing a total of 595 pair-wise genome comparisons. Protein sequences from the genomes were first compared against each other (between two genomes) or against themselves (within each genome) using DIAMOND BLASTP (36) with an E-value cutoff of 1e–10 and a maximum of five alignments. The BLASTP results were then fed to MCScanX (37) to identify synteny blocks with default parameters. In total, 391 379 synteny blocks and 12 130 719 syntenic gene pairs, with an average of 31 gene pairs per synteny block, have been identified for the 34 cucurbit genomes, and are stored in MySQL database tables in CuGenDBv2.

Table 1. Cucurbit genome assemblies available in CuGenDBv2

Common name	Latin name	Accession	Version	No. genes	Source
Cucumber	*Cucumis sativus* var. *sativus*	Chinese Long	v3	24 317	(10)
	Cucumis sativus var. *sativus*	Gy14	v2.1	22 626	-
	Cucumis sativus var. *sativus*	B10	v3	16 104	(11)
	Cucumis sativus var. *hardwickii*	PI 183967	v1	23 667	(12)
	Cucumis hystrix	–	v1	23 864	(13)
Watermelon	*Citrullus lanatus* subsp. *vulgaris*	97103	v2.5	21 917	(19)
	Citrullus lanatus subsp. *vulgaris*	Charleston Gray	v2.5	22 764	(20)
	Citrullus lanatus subsp. *cordophanus*	cordophanus	v2	21 676	(21)
	Citrullus mucosospermus	USVL531-MDR	v1	22 377	-
	Citrullus amarus	USVL246-FR2	v1	22 028	-
	Citrullus colocynthis	PI 537277	v1	22 723	-
Melon	*Cucumis melo*	DHL92	v4.0	28 299	(14)
	Cucumis melo var. *inodorus*	Payzawat	v1	22 924	(16)
	Cucumis melo ssp. *agrestis*	IVF77	v1	27 073	(15)
	Cucumis melo var. *reticulatus*	Harukei-3	v1.41	33 829	(17)
	Cucumis melo var *cantalupensis*	Charmono	v1.1	31 348	(18)
	Cucumis metuliferus	PI 482460	v1	29 214	(15)
Cucurbita	*Cucurbita maxima*	Rimu	v1.1	32 076	(22)
	Cucurbita moschata	Rifu	v1	32 205	(22)
	Cucurbita pepo subsp. *pepo*	MU-CU-16	v1	27 868	(23)
	Cucurbita argyrosperma subsp. *argyrosperma*	SMH-JMG-627	v2	27 998	(24)
	Cucurbita argyrosperma subsp. *sororia*	–	v1	30 592	(24)
Bitter gourd	*Momordica charantia*	OHB3-1	v2	41 016	(28)
	Momordica charantia	TR	v1	28 827	(29)
	Momordica charantia	Dali-11	v1	26 427	(29)
Bottle gourd	*Lagenaria siceraria*	USVL1VR-Ls	v1	22 472	(31)
	Lagenaria siceraria	Hangzhou Gourd	v1	23 510	(30)
Sponge gourd	*Luffa cylindrica*	–	v1	31 661	(25)
	Luffa cylindrica	P93075	v1	27 147	(27)
	Luffa acutangula	AG-4	v1	42 211	(26)
Wax gourd	*Benincasa hispida*	B227	v1	27 467	(32)
Chayote	*Sechium edule*	–	v1	28 237	(33)
Monk fruit	*Siraitia grosvenorii*	Qingpiguo	v1	30 565	(34)
Snake gourd	*Trichosanthes anguina*	–	v1	22 874	(35)

Cucurbit genes and annotations

A total of 919 903 protein-coding genes predicted from the 34 cucurbit genome assemblies have been comprehensively annotated using various public protein and domain databases. Protein sequences of the protein-coding genes were compared against the GenBank non-redundant (nr) (38), UniProt (SwissProt/TrEMBL) (39) and Arabidopsis (TAIR10) protein databases (40) using DIAMOND BLASTP (36) with parameters '–more-sensitive –masking 0 –evalue 1e-4'. The conserved domains or motifs in the protein-coding genes were identified by searching their protein sequences against the 16 member databases in InterPro (41) using InterProScan (42). Gene ontology (GO) (43) terms were assigned to each protein-coding gene with the BLAST2GO program (44) using the DIAMOND BLASTP results against the nr database and the results from InterProScan. The human-readable functional description of each protein-coding gene was derived from the BLASTP results against the SwissProt/TrEMBL and TAIR10 protein databases using the AHRD program (https://github.com/groupschoof/AHRD). The Pathway Tools software (45) was used to predict metabolic pathways from protein-coding genes in each of the 34 cucurbit genomes. All these analysis results were uploaded into the PostgreSQL database tables organized by the Chado schema (46) through the data loader function implemented in Tripal v3.0.

Cucurbit genome variants

During the past several years, high-density genomic variants such as SNPs and small indels have been identified for cucurbit species through large-scale genome resequencing or genotyping-by-sequencing (GBS). SNPs derived from GBS data of 1365 watermelon (25 308 SNPs) (20), 2083 melon (32 268) (47), 1234 cucumber (18 842) (48), 830 *Cucurbita pepo* (47 544), 372 *C. maxima* (5600) and 314 *C. moschata* (46 924) accessions are currently available in CuGenDBv2. SNPs were called from the GBS reads using the TASSEL 5.0 GBS Discovery Pipeline (49) and only biallelic SNPs with minor allele frequency >0.01 were retained. Recently, based on the GBS SNP data, we have constructed a cucumber core collection comprising 388 accessions and performed genome resequencing (~30×) for these accessions. For watermelon, we previously reported the genome resequencing of 414 accessions (19). We recently generated additional genome sequencing data for 201 accessions, mainly from the wild progenitor and relatives including *C. lanatus* ssp. *cordophanus*, *C. mucosospermus*, *C. amarus* and *C. colocynthis*. After integrating the two datasets, we obtained a genome-sequenced panel of 547 distinct accessions. These genome resequencing data were first processed to remove adaptor and low-quality sequences using Trimmomatic (50), and the cleaned reads were aligned to the representative cucumber (Gy14 v2.1) and watermelon (97103 v2.5) reference genomes, respectively, with BWA-

MEM (51). SNPs and small indels were then called using the Sentieon software package (https://www.sentieon.com/), and same as GBS SNPs, only biallelic SNPs and small indels with minor allele frequency >0.01 were kept. A total of 2 513 882 SNPs and 490 882 small indels were identified for the cucumber core collection, and 13 256 154 SNPs and 2 277 760 small indels for the watermelon resequencing panel. All the SNPs and small indels in CuGenDBv2 were functionally annotated by predicting their effects on protein-coding genes using SnpEff (52). SNPs and small indels are stored in the indexed VCF files in which variants can be quickly explored with BCFtools (53). The metadata associated with SNP and small indel variants such as sample accession information are stored in MySQL database tables of CuGenDBv2.

Cucurbit gene expression profiles

All raw RNA-Seq data (fastq files) from cucurbit species for which reference genomes are available in CuGenDBv2 have been downloaded from NCBI Sequence Read Archive (SRA), as well as the associated project and sample metadata. The metadata were manually curated by checking the publications describing the data (if available), and one brief and informative description for each sample was derived. RNA-Seq data with ambiguous sample information were not included in CuGenDBv2. Raw RNA-Seq reads were first processed to remove adaptor and low-quality sequences using Trimmomatic (50) and polyA/T tails using PRINSEQ++ (54). The processed reads were then aligned to the rRNA database (55) to remove possible contaminating rRNA reads. The final cleaned reads were aligned to the corresponding reference genomes in CuGenBDv2 using HISAT2 (56). Following alignments, raw counts for each protein-coding gene were calculated and then normalized to fragments per kilobase of transcript per million mapped fragments (FPKM). Currently, a total of 221 projects, 1513 distinct samples and 3560 runs (or libraries) are available in CuGenDBv2 (Table 2). The read processing and alignment statistics, raw counts and expression values (FPKM) for each project are available from the CuGenDBv2 download site. The expression value data and the associated project and sample metadata are stored in MySQL database tables of CuGenDBv2.

DATABASE FUNCTIONS

Gene interface

Same as in CuGenDBv1, CuGenDBv2 also provides the basic search functions such as search by gene ID or key words, and the batch query function. However, for easier navigation of gene features, the gene page has been redesigned and a navigation bar has been added for different sections related to gene features including 'Overview', 'Sequences', 'Homology', 'InterPro', 'Relationship' and 'GO annotation' (Figure 1A). Besides basic gene features, the 'Overview' section also provides links of 'RNA-Seq Expression' and 'Synteny' for each protein-coding gene. The 'RNA-Seq Expression' link displays the expression profiles of the gene of interest in various RNA-Seq projects archived in the database (Figure 1B). The 'Synteny' link displays the orthologous and paralogous genes of the gene of interest in different synteny blocks, and the list of synteny blocks that cover the gene (Figure 1C).

Genotype module

A 'Genotype' module has been newly implemented in CuGenDBv2 that provides a suite of functions to extract/download variants including SNPs and small indels from large-scale population genome sequencing projects. In this module four variant retrieval/download functions are available: (i) variant retrieval within a gene of interest; (ii) variant retrieval within a specific genomic region; (iii) variant retrieval at a specific genomic position for a list of accessions or all accessions in the project; (iv) download of variant data within a specific genome region for a list of accessions or all accessions in the project (Figure 2A).

Within the genome region of a gene of interest, or a specific genomic region defined by the user, the functions return a list of variants with their genomic positions, the reference and alternate alleles, and their annotations (effects on genes) (Figure 2B). For each specific variant (or variant at a specific genomic position), the interface displays the basic information of the variant, the flanking sequences (500 bp up- and downstream) with or without other variants in the flanking sequences being provided. Moreover, the allele frequencies in different groups from a list of samples or all samples in the project are shown as a bar chart (Figure 2C). Each bar in the chart is linked to a page that displays the genotype information of the variant in individual accessions from the corresponding group (Figure 2D).

Expression module

To provide a complete cucurbit gene expression atlas, the 'Expression' module in CuGenDBv2 has been redesigned. Under this module, the expression profile data of a specific gene can be easily and directly accessed. As described above, expression profiles in different RNA-Seq projects of a gene of interest can be accessed directly through the 'RNA-Seq Expression' link provided in the gene page. In addition, the main navigation bar of CuGenDBv2 contains the 'Expression' menu, which provides a query interface that also returns the expression profiles (FPKM values) of the queried gene in all corresponding projects/samples. Furthermore, with this redesigned 'Expression' module, newly available RNA-Seq expression data can be easily added and displayed in the database.

Other updated tools

All the other data search, mining and analysis tools in CuGenDBv1 have been preserved in CuGenDBv2. The basic search, analysis and visualization tools, including 'Search', 'BLAST', 'JBrowse', 'Batch Query', 'Synteny Viewer' and 'CucurbitCyc', have been kept with the same functionalities while the backend datasets have been updated with the 34 cucurbit genome assemblies currently available in the database. Specially, synteny blocks between any two and within each of the 34 genome assemblies can be visualized in the database. Functions in the 'Synteny Viewer' module

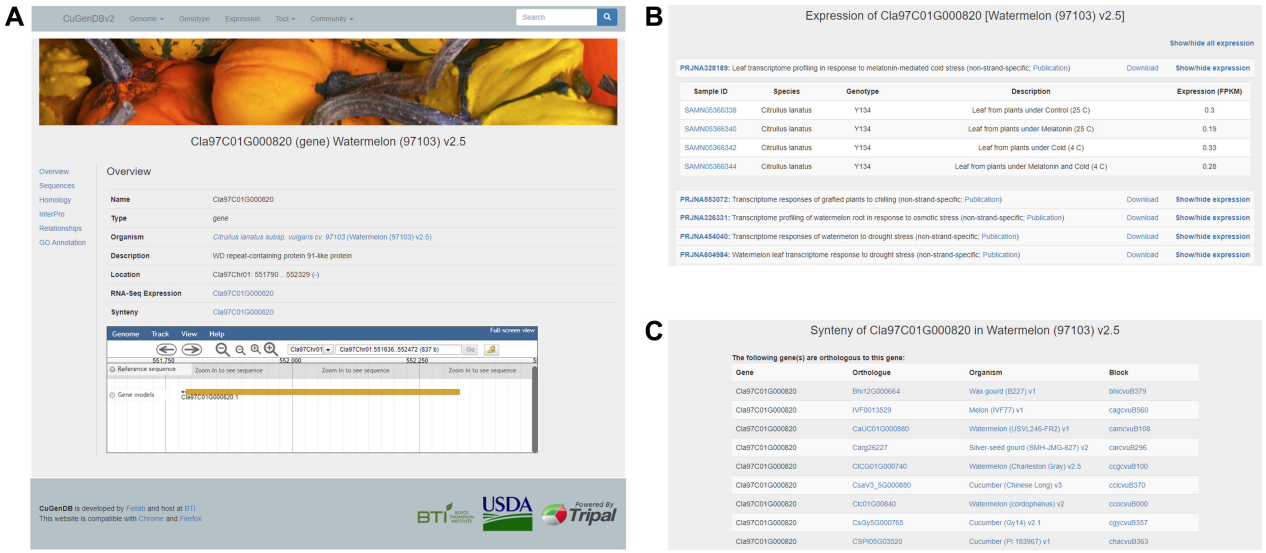

Figure 1. Gene interface in CuGenDBv2. (**A**) Screenshot of the gene feature page showing the navigation menus on the left and gene overview on the right. (**B**) Page showing expression profiles of the queried gene or the gene of interest in different RNA-Seq projects. (**C**) Page showing the list of syntenic orthologous genes of a gene of interest.

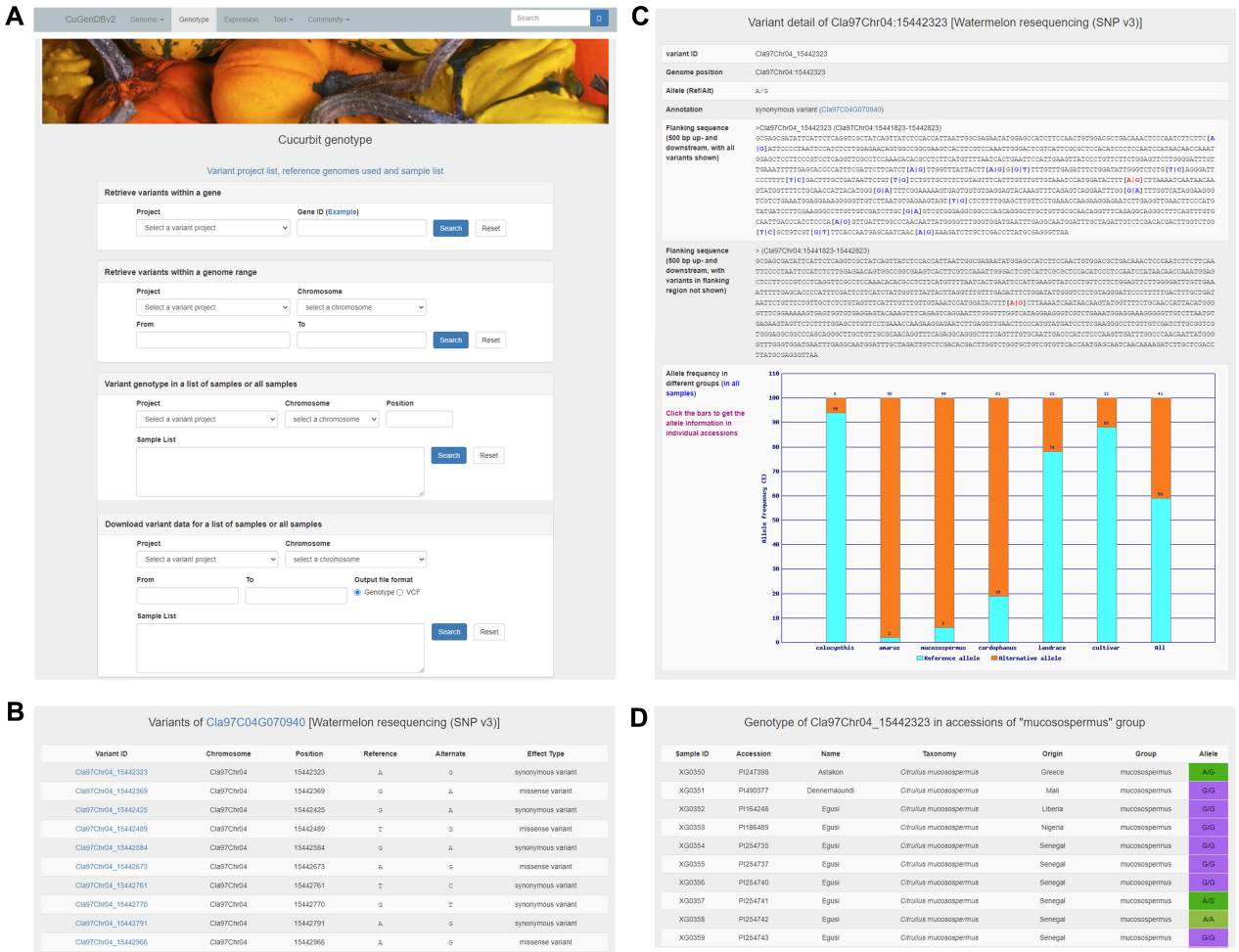

Figure 2. Genotype module in CuGenDBv2. (**A**) Screenshot of the search interfaces of the genotype module. (**B**) Result page showing the list of variants within the genomic region of the queried gene. (**C**) Detailed page of a specific variant. The page shows the basic information of the variant, its genomic flanking sequences, and its allele frequencies in different population groups. (**D**) Detailed genotype information of a specific variant in individual accessions.

Table 2. Summary of RNA-Seq gene expression data in CuGenDBv2

Species	No. projects	No. samples	No. runs	Reference genomes used
Cucumber (*Cucumis sativus*)	88	523	1315	Chinese Long v3, Gy14 v2.1, B10 v3, PI183967 v1
Cucumber (*Cucumis hystrix*)	2	9	17	hystrix v1
Watermelon (*Citrullus* spp.)	49	293	769	97103 v2.5, Charleston Gray v2.5, USVL246-FR2 v1, PI 537277 v1, cordophanus v2, USVL531-MDR v1
Melon (*Cucumis melo*)	41	362	705	DHL92 v4, Payzawat v1, IVF77 v1, Harukei-3 v1.41, Charmono v1.1
Melon (*Cucumis metuliferus*)	1	4	4	PI 482460 v1
Cucurbita pepo	11	94	203	MU-CU-16 v4.1
Cucurbita moschata	15	62	158	Rifu v1
Cucurbita maxima	7	24	64	Rimu v1.1
Cucurbita argyrosperma	2	10	10	SMH-JMG-627 v2, sororia v1
Bottle gourd (*Lagenaria siceraria*)	6	40	86	Hangzhou Gourd v1, USVL1VR-Ls v1
Bitter gourd (*Momordica charantia*)	4	30	75	OHB3-1 v2, TR v1, Dali-11 v1
Wax gourd (*Benincasa hispida*)	5	28	74	B227 v1
Snake gourd (*Trichosanthes anguina*)	1	7	16	Snake gourd v1
Monk fruit (*Siraitia grosvenorii*)	1	6	15	Qingpiguo v1
Sponge gourd (*Luffa* spp.)	8	21	49	cylindrica v1, AG-4 v1, P93075 v1
Chayote (*Sechium edule*)	0	0	0	Chayote v1
Total	225	1513	3560	-

have been re-implemented using Perl/CGI with the newly added syntenic genomic data stored in MySQL database tables, which has substantially improved the performance (mainly speed) of these functions. Other data mining and analysis tools, including 'Pathway enrichment', 'GO enrichment', and 'Gene classification', follow the previous designs in CuGenDBv1 with the newly analyzed results from the Pathway Tools and BLAST2GO for the 34 cucurbit genome assemblies.

CONCLUSIONS AND FUTURE PERSPECTIVES

The CuGenDBv2 currently contains 34 genome assemblies with comprehensive gene functional annotations, from 27 different species/subspecies belonging to 10 cucurbit genera. Compared with CuGenDBv1, a new 'Genotype' module has been developed in CuGenDBv2, which helps mining genomic variants including functionally annotated SNPs and small indels identified from large-scale genome sequencing projects with user-friendly interfaces. RNA-Seq raw reads have been downloaded from NCBI SRA for all cucurbit species for which reference genomes are available in CuGenDBv2 and processed to derive gene expression values. The 'Expression' module has been redesigned and re-implemented, which provides a complete gene expression atlas for cucurbit species. In addition, CuGenDBv2 includes a huge amount of genomic syntenic information derived from the comparisons of the 34 genomes, and the 'Synteny Viewer' have been re-implemented in CuGenDBv2 to improve its performance in handling this type of massive datasets.

CuGenDBv2 will be updated regularly when new genomic datasets are available. New genome assemblies will be included in the database if the assemblies are from species or subspecies that are not covered by the existing genomes in CuGenDBv2 or have substantially higher quality than existing genome assemblies from the same species or subspecies. Variant data can be easily added to the database if the sample metadata are sufficiently clear. Therefore, we will add genome variant data once they are available. RNA-Seq data will be collected from NCBI SRA, processed to derive expression values and included in the database every six months. Furthermore, large-scale phenotypic data are being generated for various cucurbit populations. Functions to mine, analyze and visualize these data and to associate phenotype and genotype data will be implemented in the database.

DATA AVAILABILITY

All data host in CuGenDBv2 are freely available at (http://cucurbitgenomics.org/v2/).

FUNDING

USDA National Institute of Food and Agriculture Specialty Crop Research Initiative [2015-51181-24285, 2020-51181-32139]. Funding for open access charge: USDA National Institute of Food and Agriculture Specialty Crop Research Initiative [2020-51181-32139].
Conflict of interest statement. None declared.

REFERENCES

1. Schaefer,H., Heibl,C. and Renner,S.S. (2012) Gourds afloat: a dated phylogeny reveals an asian origin of the gourd family (Cucurbitaceae) and numerous oversea dispersal events. *Proc. R. Soc. B Biol. Sci.*, **276**, 843–851.
2. Christenhusz,M.J.M. and Byng,J.W. (2016) The number of known plants species in the world and its annual increase. *Phytotaxa*, **261**, 201–217.
3. Wu,P., Tung,C., Lee,C. and Liao,C. (2019) Genomic prediction of pumpkin hybrid performance. *Plant Genome*, **12**, 180082.
4. Abdel-Salam,I.M., Awadein,N.E.S. and Ashour,M. (2019) Cytotoxicity of *Luffacylindrica* (L.) M.Roem. extract against circulating cancer stem cells in hepatocellular carcinoma. *J. Ethnopharmacol.*, **229**, 89–96.
5. Tanurdzic,M. and Banks,J.A. (2004) Sex-determining mechanisms in land plants. *Plant Cell*, **16**, S61–S71.
6. Huang,S., Li,R., Zhang,Z., Li,L., Gu,X., Fan,W., Lucas,W.J., Wang,X., Xie,B., Ni,P. *et al.* (2009) The genome of the cucumber, *Cucumissativus* L. *Nat. Genet.*, **41**, 1275–1281.

7. Zheng,Y., Wu,S., Bai,Y., Sun,H., Jiao,C., Guo,S., Zhao,K., Blanca,J., Zhang,Z., Huang,S. *et al.* (2019) Cucurbit genomics database (CuGenDB): a central portal for comparative and functional genomics of cucurbit crops. *Nucleic Acids Res.*, **47**, D1128–D1136.
8. Spoor,S., Cheng,C.H., Sanderson,L.A., Condon,B., Almsaeed,A., Chen,M., Bretaudeau,A., Rasche,H., Jung,S., Main,D. *et al.* (2019) Tripal v3: an ontology-based toolkit for construction of FAIR biological community databases. *Database*, **2019**, baz077.
9. Ficklin,S.P., Sanderson,L.A., Cheng,C.H., Staton,M.E., Lee,T., Cho,I.H., Jung,S., Bett,K.E. and Main,D. (2011) Tripal: a construction toolkit for online genome databases. *Database*, **2011**, bar044.
10. Li,Q., Li,H., Huang,W., Xu,Y., Zhou,Q., Wang,S., Ruan,J., Huang,S. and Zhang,Z. (2019) A chromosome-scale genome assembly of cucumber (*Cucumis sativus* l.). *GigaScience*, **8**, giz072.
11. Osipowski,P., Pawełkowicz,M., Wojcieszek,M., Skarzyńska,A., Przybecki,Z. and Pląder,W. (2020) A high-quality cucumber genome assembly enhances computational comparative genomics. *Mol. Genet. Genomics*, **295**, 177–193.
12. Qi,J., Liu,X., Shen,D., Miao,H., Xie,B., Li,X., Zeng,P., Wang,S., Shang,Y., Gu,X. *et al.* (2013) A genomic variation map provides insights into the genetic basis of cucumber domestication and diversity. *Nat. Genet.*, **45**, 1510–1515.
13. Qin,X., Zhang,Z., Lou,Q., Xia,L., Li,J., Li,M., Zhou,J., Zhao,X., Xu,Y., Li,Q. *et al.* (2021) Chromosome-scale genome assembly of *Cucumishystrix* - a wild species interspecifically cross-compatible with cultivated cucumber. *Hortic. Res.*, **8**, 40.
14. Castanera,R., Ruggieri,V., Pujol,M., Garcia-Mas,J. and Casacuberta,J.M. (2020) An improved melon reference genome with single-molecule sequencing uncovers a recent burst of transposable elements with potential impact on genes. *Front. Plant Sci.*, **10**, 1815.
15. Ling,J., Xie,X., Gu,X., Zhao,J., Ping,X., Li,Y., Yang,Y., Mao,Z. and Xie,B. (2021) High-quality chromosome-level genomes of *Cucumismetuliferus* and *Cucumismelo* provide insight into *Cucumis* genome evolution. *Plant J.*, **107**, 136–148.
16. Zhang,H., Li,X., Yu,H., Zhang,Y., Li,M., Wang,H., Wang,D., Wang,H., Fu,Q., Liu,M. *et al.* (2019) A high-huality melon genome assembly provides insights into genetic basis of fruit trait improvement. *iScience*, **22**, 16–27.
17. Yano,R., Ariizumi,T., Nonaka,S., Kawazu,Y., Zhong,S., Mueller,L., Giovannoni,J.J., Rose,J.K.C. and Ezura,H. (2020) Comparative genomics of muskmelon reveals a potential role for retrotransposons in the modification of gene expression. *Commun. Biol.*, **3**, 432.
18. Pichot,C., Djari,A., Tran,J., Verdenaud,M., Marande,W., Huneau,C., Gautier,V., Latrasse,D., Arribat,S., Sommard,V. *et al.* (2022) Cantaloupe melon genome reveals 3D chromatin features and structural relationship with the ancestral Cucurbitaceae karyotype. *iScience*, **25**, 103696.
19. Guo,S., Zhao,S., Sun,H., Wang,X., Wu,S., Lin,T., Ren,Y., Gao,L., Deng,Y., Zhang,J. *et al.* (2019) Resequencing of 414 cultivated and wild watermelon accessions identifies selection for fruit quality traits. *Nat. Genet.*, **51**, 1616–1623.
20. Wu,S., Wang,X., Reddy,U., Sun,H., Bao,K., Gao,L., Mao,L., Patel,T., Ortiz,C., Abburi,V.L. *et al.* (2019) Genome of 'Charleston gray', the principal American watermelon cultivar, and genetic characterization of 1,365 accessions in the U.S. national plant germplasm system watermelon collection. *Plant Biotechnol. J.*, **17**, 2246–2258.
21. Renner,S.S., Wu,S., Pérez-Escobar,O.A., Silber,M.V., Fei,Z. and Chomicki,G. (2021) A chromosome-level genome of a Kordofan melon illuminates the origin of domesticated watermelons. *Proc. Natl. Acad. Sci. U.S.A.*, **118**, e2101486118.
22. Sun,H., Wu,S., Zhang,G., Jiao,C., Guo,S., Ren,Y., Zhang,J., Zhang,H., Gong,G., Jia,Z. *et al.* (2017) Karyotype stability and unbiased fractionation in the paleo-allotetraploid *Cucurbita* genomes. *Mol. Plant*, **10**, 1293–1306.
23. Montero-Pau,J., Blanca,J., Bombarely,A., Ziarsolo,P., Esteras,C., Martí-Gómez,C., Ferriol,M., Gómez,P., Jamilena,M., Mueller,L. *et al.* (2018) De novo assembly of the zucchini genome reveals a whole-genome duplication associated with the origin of the C*ucurbita* genus. *Plant Biotechnol. J.*, **16**, 1161–1171.
24. Barrera-Redondo,J., Sánchez-de la Vega,G., Aguirre-Liguori,J.A., Castellanos-Morales,G., Gutiérrez-Guerrero,Y.T., Aguirre-Dugua,X., Aguirre-Planter,E., Tenaillon,M.I., Lira-Saade,R. and Eguiarte,L.E. (2021) The domestication of C*ucurbitaargyrosperma* as revealed by the genome of its wild relative. *Hortic. Res.*, **8**, 109.
25. Zhang,T., Ren,X., Zhang,Z., Ming,Y., Yang,Z., Hu,J., Li,S., Wang,Y., Sun,S., Sun,K. *et al.* (2020) Long-read sequencing and de novo assembly of the *Luffacylindrica* (L.) Roem. genome. *Mol. Ecol. Resour.*, **20**, 511–519.
26. Pootakham,W., Sonthirod,C., Naktang,C., Nawae,W., Yoocha,T., Kongkachana,W., Sangsrakru,D., Jomchai,N., U-thoomporn,S., Sheedy,J.R. *et al.* (2021) De novo assemblies of *Luffaacutangula* and *Luffacylindrica* genomes reveal an expansion associated with substantial accumulation of transposable elements. *Mol. Ecol. Resour.*, **21**, 212–225.
27. Wu,H., Zhao,G., Gong,H., Li,J., Luo,C., He,X., Luo,S., Zheng,X., Liu,X., Guo,J. *et al.* (2020) A high-quality sponge gourd (*Luffa cylindrica*) genome. *Hortic. Res.*, **7**, 128.
28. Matsumura,H., Hsiao,M.C., Lin,Y.P., Toyoda,A., Taniai,N., Tarora,K., Urasaki,N., Anand,S.S., Dhillon,N.P.S., Schafleitner,R. *et al.* (2020) Long-read bitter gourd (*Momordica charantia*) genome and the genomic architecture of nonclassic domestication. *Proc. Natl. Acad. Sci. U.S.A.*, **117**, 14543–14551.
29. Cui,J., Yang,Y., Luo,S., Wang,L., Huang,R., Wen,Q., Han,X., Miao,N., Cheng,J., Liu,Z. *et al.* (2020) Whole-genome sequencing provides insights into the genetic diversity and domestication of bitter gourd (*Momordica* spp.). *Hortic. Res.*, **7**, 85.
30. Xu,P., Wang,Y., Sun,F., Wu,R., Du,H., Wang,Y., Jiang,L., Wu,X., Wu,X., Yang,L. *et al.* (2021) Long-read genome assembly and genetic architecture of fruit shape in the bottle gourd. *Plant J.*, **107**, 956–968.
31. Wu,S., Shamimuzzaman,M., Sun,H., Salse,J., Sui,X., Wilder,A., Wu,Z., Levi,A., Xu,Y., Ling,K.S. *et al.* (2017) The bottle gourd genome provides insights into Cucurbitaceae evolution and facilitates mapping of a papaya ring-spot virus resistance locus. *Plant J.*, **92**, 963–975.
32. Xie,D., Xu,Y., Wang,J., Liu,W., Zhou,Q., Luo,S., Huang,W., He,X., Li,Q., Peng,Q. *et al.* (2019) The wax gourd genomes offer insights into the genetic diversity and ancestral cucurbit karyotype. *Nat. Commun.*, **10**, 5158.
33. Fu,A., Wang,Q., Mu,J., Ma,L., Wen,C., Zhao,X., Gao,L., Li,J., Shi,K., Wang,Y. *et al.* (2021) Combined genomic, transcriptomic, and metabolomic analyses provide insights into chayote (*Sechium edule*) evolution and fruit development. *Hortic. Res.*, **8**, 35.
34. Xia,M., Han,X., He,H., Yu,R., Zhen,G., Jia,X., Cheng,B. and Deng,X.W. (2018) Improved de novo genome assembly and analysis of the Chinese cucurbit *Siraitiagrosvenorii*, also known as monk fruit or luo-han-guo. *GigaScience*, **7**, giy067.
35. Ma,L., Wang,Q., Mu,J., Fu,A., Wen,C., Zhao,X., Gao,L., Li,J., Shi,K., Wang,Y. *et al.* (2020) The genome and transcriptome analysis of snake gourd provide insights into its evolution and fruit development and ripening. *Hortic. Res.*, **7**, 199.
36. Buchfink,B., Reuter,K. and Drost,H.G. (2021) Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods*, **18**, 366–368.
37. Wang,Y., Tang,H., Debarry,J.D., Tan,X., Li,J., Wang,X., Lee,T.H., Jin,H., Marler,B., Guo,H. *et al.* (2012) MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.*, **40**, e49.
38. Sayers,E.W., Cavanaugh,M., Clark,K., Pruitt,K.D., Schoch,C.L., Sherry,S.T. and Karsch-Mizrachi,I. (2021) GenBank. *Nucleic Acids Res.*, **49**, D92–D96.
39. Bateman,A., Martin,M.J., Orchard,S., Magrane,M., Agivetova,R., Ahmad,S., Alpi,E., Bowler-Barnett,E.H., Britto,R., Bursteinas,B. *et al.* (2021) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.*, **49**, D480–D489.
40. Lamesch,P., Berardini,T.Z., Li,D., Swarbreck,D., Wilks,C., Sasidharan,R., Muller,R., Dreher,K., Alexander,D.L., Garcia-Hernandez,M. *et al.* (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.*, **40**, 1202–1210.
41. Blum,M., Chang,H.Y., Chuguransky,S., Grego,T., Kandasaamy,S., Mitchell,A., Nuka,G., Paysan-Lafosse,T., Qureshi,M., Raj,S. *et al.* (2021) The Interpro protein families and domains database: 20 years on. *Nucleic Acids Res.*, **49**, D344–D354.
42. Jones,P., Binns,D., Chang,H.Y., Fraser,M., Li,W., McAnulla,C., McWilliam,H., Maslen,J., Mitchell,A., Nuka,G. *et al.* (2014)

InterProScan 5: Genome-scale protein function classification. *Bioinformatics*, **30**, 1236–1240.
43. Carbon,S., Douglass,E., Good,B.M., Unni,D.R., Harris,N.L., Mungall,C.J., Basu,S., Chisholm,R.L., Dodson,R.J., Hartline,E. *et al.* (2021) The gene ontology resource: enriching a GOld mine. *Nucleic Acids Res.*, **49**, D325–D334.
44. Conesa,A., Götz,S., García-Gómez,J.M., Terol,J., Talón,M. and Robles,M. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674–3676.
45. Karp,P.D., Paley,S.M., Krummenacker,M., Latendresse,M., Dale,J.M., Lee,T.J., Kaipa,P., Gilham,F., Spaulding,A., Popescu,L. *et al.* (2010) Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. *Brief. Bioinform.*, **11**, 40–79.
46. Jung,S., Lee,T., Ficklin,S., Yu,J., Cheng,C.H. and Main,D. (2016) Chado use case: storing genomic, genetic and breeding data of Rosaceae and Gossypium crops in Chado. *Database*, **2016**, baw010.
47. Wang,X., Ando,K., Wu,S., Reddy,U.K., Tamang,P., Bao,K., Hammar,S.A., Grumet,R., McCreight,J.D. and Fei,Z. (2021) Genetic characterization of melon accessions in the U.S. National Plant Germplasm System and construction of a melon core collection. *Mol. Hortic.*, **1**, 11.
48. Wang,X., Bao,K., Reddy,U.K., Bai,Y., Hammar,S.A., Jiao,C., Wehner,T.C., Ramírez-Madera,A.O., Weng,Y., Grumet,R. *et al.* (2018) The USDA cucumber (*Cucumis sativus* L.) collection: genetic diversity, population structure, genome-wide association studies, and core collection development. *Hortic. Res.*, **5**, 64.
49. Glaubitz,J.C., Casstevens,T.M., Lu,F., Harriman,J., Elshire,R.J., Sun,Q. and Buckler,E.S. (2014) TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. *PLoS One*, **9**, e90346.
50. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
51. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, **25**, 1754–1760.
52. Cingolani,P., Platts,A., Wang,L.L., Coon,M., Nguyen,T., Wang,L., Land,S.J., Lu,X. and Ruden,D.M. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. *Fly.*, **6**, 80–92.
53. Danecek,P., Bonfield,J.K., Liddle,J., Marshall,J., Ohan,V., Pollard,M.O., Whitwham,A., Keane,T., McCarthy,S.A., Davies,R.M. *et al.* (2021) Twelve years of SAMtools and BCFtools. *GigaScience*, **10**, giab008.
54. Cantu,V.A., Sadural,J. and Edwards,R. (2019) PRINSEQ++, a multi-threaded tool for fast and efficient quality control and preprocessing of sequencing datasets. *PeerJ Prepr.*, **7**, e27553v1.
55. Quast,C., Pruesse,E., Yilmaz,P., Gerken,J., Schweer,T., Yarza,P., Peplies,J. and Glöckner,F.O. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.*, **41**, D590–D596.
56. Kim,D., Paggi,J.M., Park,C., Bennett,C. and Salzberg,S.L. (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.*, **37**, 907–915.