1 GGDB: A Gramineae Genome Alignment Database of Homologous 2 Genes Hierarchically Related to Evolutionary Events

3

4 Qihang Yang¹, Tao Liu¹, Tong Wu¹, Tianyu Lei¹, Yuxian Li¹, Xiyin Wang^{1*}

5

6 ¹Center for Genomics and Bio-computing, North China University of Science and
7 Technology, Tangshan, Hebei 063210, China;

8

9 *To whom correspondence should be addressed. Tel: 86-315-8805592; Fax:
10 86-315-8805592; Email: wangxiyin@vip.sina.com

11

12

13 ABSTRACT

14 Owing to their economic values, Gramineae plants have been
15 preferentially sequenced their genomes. These genomes are often quite
16 complex, e.g., harboring many duplicated genes, which were the main
17 source of genetic innovation and often the results of recurrent
18 polyploidization. Deciphering the complex genome structure and linking
19 duplicated genes to specific polyploidization events are important to
20 understand the biology and evolution of plants. However, the effort has
21 been held back due to its high complexity in analyzing these genomes.
22 Here, by hierarchically relating duplicated genes in colinearity to each
23 polyploidization or speciation event, we analyzed 29 well-assembled and
24 up-to-date Gramineae genome sequences, separated duplicated genes
25 produced by each event, established lists of paralogous and orthologous

genes, and eventually constructed an on-line database, GGDB (<http://www.grassgenome.com/>). Homologous gene lists from each plant and between them can be displayed, searched, and downloaded from the database. Interactive comparison tools were deployed to demonstrate homology among user-selected plants, to draw genome-scale or local alignment figures, phylogenetic trees of genes corrected by exploiting gene colinearity, etc. Using these tools and figures, users can easily observe genome structural changes, and explore the effects of paleo-polyploidy on crop genome structure and function. The GGDB will be a useful platform to improve understanding the genome changes and functional innovation of Gramineae plants.

Keywords: Gramineae; Colinearity; Polyploidy; Homologous gene; Database

Key points:

1. GGDB is the only portal hosting Grameneae colinear homologous genes hierarchically related to evolutionary events, especially polyploidization, which have occurred recursively.
2. Allows systematic analysis of colinear gene relationships and function origination and/or divergence across Grameneae plants.
3. Serving the Grameneae research community, with new genomes, modules, tools, and analysis.

INTRODUCTION

Gramineae is a large group of monocotyledonous flowering plants, which can be divided into more than 620 genera and more than 10000 species, covering 20% of the land area of the earth^[1]. Gramineae contains many important food crops, such as wheat (*Triticum aestivum*), rice (*Oryza sativa*), corn (*Zea mays*), sorghum (*Sorghum bicolor*), and so on^[2-5]. With the tens of Gramineae genomes being sequenced, it provides a solid data basis for in-depth analysis of functional innovation and evolution of Gramineae genomes.

The study of the Gramineae genomes revealed repeated polyploidy events during the evolutionary history^[6-8]. Polyploidy is an abrupt event, which can create a new species with doubled number of chromosomes, produce a large number of repetitive genes^[9-11], trigger large-scale reorganization of biological functions, such as regulatory network re-programming and debugging. Polyploidy leads to genomic instability, and a considerable amount of gene loss may occur^[12; 13]. Gramineae plants could be taken as a good example in that their common ancestor was affected by a tetraploidization ~100 million years ago, followed by the fast originization and divergence of derivative plants^[14-16]. Gramineae crops, such as wheat and maize, were

71 resulted from further polyploidization event(s). Recursive
72 polyploidization and genome reorganization makes their genomes
73 rather complex ^[17-19].

74 Gene colinearity provides precious means to study complex
75 genomic structures. In extant genomes, a considerable number of
76 colinear genes produced by polyploidy have been retained ^[20]. In
77 rice genome, there still remain thousands of colinear genes,
78 resulted from the Gramineae-common tetraploidization ^[21; 22]. The
79 analysis of colinear genes helps identify ancient polyploidy events,
80 deduce the scale and time of their occurrence, and infer gene
81 functional changes ^[23; 24].

82 In that the importance of gene colinearity, the information was
83 often inferred and stored in biological databases, such as JCVI,
84 PGDD, and COGE^[25-27]. However, the gene colinearity
85 information in these existing databases has non-negligible
86 shortcomings. First, colinear genes were not related to specific
87 polyploidy events due to methodological difficulty to analyze these
88 recursive events. Second, only two or three species were involved
89 to infer gene colinearity, hampering the efforts to do genus- or
90 family-scale level evolutionary or functional analysis.

91 Here, by integrating approaches of homologous gene
92 dotplotting to compare genomes, characterizing divergence levels

of colinear gene homologs, and checking complement patterns of chromosome breakages, colinear homologs produced by different polyploidies (or speciation) could be separated, and hierarchically related to each relative event [28; 29]. Then, we implemented the above mentioned hierarchical gene colinearity inference approaches with the Gramineae plants, and thus using the results established the Gramineae Genome Alignment Database GGDB (<http://www.grassgenome.com/>), which contains gene colinearity information to explore the chromosome changes, genomic repatterning, and the actual phylogeny of duplicated genes^[30; 31] at the Gramineae family-scale level.

MATERIALS AND METHODS

We summarize the species information contained in GGDB, the way of data processing, and the composition structure of webpages (Fig.1A-B). The database is constructed by using MySQL to store analysis results^[32-34], such as colinear gene lists, chromosome homology within a genome or between genomes, figures to show homolog between genomes, etc. The website was developed by using HTML and PHP. Data was analyzed by using scripts developed by Python and R^[35-38].

Data sources

In determining whether a species should be included in the GGDB or not, we require its genome to be assembled to the chromosome-level,

115 allowing credible gene colinearity inference. If multiple genome
116 assemblies are available, we use the latest assembly version in the
117 database^[39-41]. Most genome data was downloaded from the NCBI
118 database (<https://www.ncbi.nlm.nih.gov/>) and the JGI database
119 (<https://phytozome-next.jgi.doe.gov/>) (Table 1). Genome data was
120 preprocessed by using home-made Perl scripts: (i) uniform format of gene
121 names was adopted; (ii) redundant information in the genome annotation
122 is removed; (iii) gene location files was extracted, including the
123 information of chromosome numbers, gene IDs, gene locations and
124 orders on chromosomes, etc.

125 **Inferring colinear genes**

126 Sequence similarity alignment software BLAST+ was used to infer
127 putative homologous genes (BLAST E-value $\leq 10^{-5}$ and sequence
128 matching score ≥ 100). Homologous gene dotplots were drawn, in that
129 they are helpful to show and infer homology and structural changes
130 within a genome or between genomes. In a homologous gene dotplot,
131 dots represent homologous gene pairs and are often assigned with
132 different colors, to show sequence divergence levels between compared
133 homologous genes.

134 Colinear genes within and between genomes of Gramineae were
135 inferred by using software ColinearScan^[42], using the above BLAST
136 inferred putative homologs, with BLAST accumulated hit scores ≥ 50 of

137 homologous blocks with colinear gene numbers ≥ 5 , and the statistical
138 significance is set to be $\leq 1 \times 10^{-10}$ estimated by ColinearScan.

139 To distinguish colinear gene blocks produced by different events,
140 polyploidies or speciation, we estimated the synonymous nucleic acid
141 substitution rate at the synonymous sites (Ks), which could be used to
142 measure divergence levels between homologous genes. The Nei-Gojobori
143 method implemented in PAML package was used to estimate the above
144 values^[43]. Actually, homologous blocks produced by different events,
145 especially two polyploidy events, could be mixed together due to having
146 similar Ks values, complicated by the fact that genes evolve at divergent
147 rates. We checked whether homologous blocks shared the chromosome
148 breakage points, which is a hallmark to show them to have been produced
149 by the same polyploidy event. These shared chromosome breakage points
150 could be identified in homologous gene dotplots. When dealing with
151 cross-species gene colienarity, orthologous genes between species often
152 form much better gene collinearity and have much smaller Ks than the
153 outparalogous genes do produced by polyploidy in the common ancestor.
154 Seldom cases need to check chromosome breakage points to distinguish
155 orthologous and outparalogous blocks. Eventually, lists of colinear genes
156 associated with specific polyploidy events or speciation were generated
157 and stored in the MySQL database.

158 **Multi-genome alignment map**

159 Multi-genome alignment maps were constructed by integrating lists of
160 colinear genes between any two species. A reference genome was
161 selected, and then another plant genomes were aligned to it one by one,
162 with colinear genes as markers. Eventually, a table containing aligned
163 colinear genes was produced, and a gene in the reference genome often
164 has no corresponding homolog in another genome or in the duplicated
165 regions, the corresponding cell in the table were filled with dots.

166 **Local colinear alignment**

167 Local colinear alignment was constructed by integrating the list of
168 colinear genes among species. We used a Python script to call the
169 integration package Matplotlib module, built a two-dimensional atlas to
170 show gene homology information. Based on the genome comparison
171 software MCSanX[44] , we developed the "Local colinear alignment"
172 module. Compared to previous databases, by checking chromosome
173 breakage points, GGDB can distinguish subgenomes produced by
174 genome doubling. After receiving the query data and parameters sent
175 from the web interface, the GGDB server queries the colinear gene results
176 in the database and draws the local colinear alignment figure. The final
177 result file is to be packaged and sent to the browser in PDF format.

178 **Gene evolutionary tree**

179 To help explore the evolution of duplicated genes, which is critical in
180 genetic innovation, we used phylogenetic analysis software IQTREE,
181 MUSCLE, and FASTTREE^[45-47], to construct an evolutionary tree using
182 DNA or protein sequences of a set of homologous genes. Actually,
183 previous research found that duplicated genes, especially those produced
184 by polyploidies, could form trees inconsistent to their real evolutionary
185 relationship^[48]. For a set of homologous genes, we can construct the
186 expected tree reflecting the actual relationship of colinear genes,
187 including paralogs produced by specific polyploidies, and orthologs
188 originated from speciation.

189 After accepting the user query parameters (Selected species, reference
190 species, Gene ID) from the browser, the GGDB server queries the
191 database for the homologous colinear genes and writes the gene
192 sequences into a fasta file. The file is used as input for the software
193 MUSCLE to do sequence alignment, and then a tree is built by using
194 FASTTREE^[49; 50]. Default parameters of these software were used. Both
195 the sequence-alignment derived tree and the expected tree are stored in
196 nwk format, convenient for the user download and further editing. The
197 software EChart plug-in is implemented in JavaScript in our interface,
198 and the interaction between the user and the server is realized^[51; 52].

199 **Construction and content**

200 We developed the GGDB database to provide homologous
201 colinear gene information within each of or between Gramineae
202 plants. The database is currently installed on the CentOS operating
203 system. It has a three-tier architecture, namely, the client tier, the
204 middle tier and the database tier. The client layer that users
205 directly access is developed using PHP and JavaScript. In the
206 database layer, GGDB-related data is stored in a MySQL database.
207 The middle tier receives HTTP requests and is processed by a
208 Apacheweb server. In addition, we include different levels of
209 genomic colinearity analysis tools (Fig.2).

210 **Overview of data**

211 At present, GGDB contains the information of colinear genes
212 within a genome and between the genomes of 29 Gramineae plants.
213 A referring outgroup, pineapple, is also included to help explore
214 gene evolution, especially infer real evolutionary relationship of
215 genes. As the original input data, three types of files are used:
216 coding sequence files, protein sequence files, and general feature
217 format (GFF) files containing chromosome sequence annotation
218 data.

219 **Colinearity data**

220 A polyploidy whole-genome duplication (WGD) event common to all

main lineages of Gramineae (cWGD) was dated to 96 million years ago based on putatively neutral DNA substitution rates between duplicated genes^[53]. Some species are further affected by another polyploid (mWGD) event. We analyzed the colinearity of pineapple and 29 species of Gramineae, made 784 comparisons between the two species, and finally generated 784 colinear lists. Then all the colinear lists are summarized into colinear tables with reference to 29 species.

Based on homologous correspondence and colinearity analysis between genomes, we obtained the homologous information of various species of Gramineae (Table 2; Supplemental table 1-4). The homologous regions are detected under the threshold of 5, 10, 20, and 50 colinear genes, respectively. Paralogous and orthologous genes are associated with polyploidy events. For example, There are 20763 homologous genes in maize genome, of which 13134 paralogous genes produced by the mWGD, and 7629 paralogous genes produced by the cWGD (Table 3; Supplemental table 5). Between rice and maize, there are 26947 orthologous genes, and 12600 outparalogous genes produced by the cWGD.

Overview of the interface

On the home page of GGDB, we marked out the geographical originating locations of 29 Gramineae species on a world map^[54]. An interactive evolution tree including the above species is also provided on

the home page. The chart interface on the home page provides interactive view of chromosomes from all species, including the numbers and lengths of chromosomes from each species and the numbers of genes on each chromosome. In addition, we use bar charts and line charts to display the chromosome numbers of each species, which makes it easier for users to compare their differences. These interactive charts can be downloaded.

Species information page

We provide a web page for each Gramineae species, showing basic information about its name (Latin name, common name, Chinese name), picture, classification, profile (geographical distribution, biological characteristics, living habits), genome information (genome size, chromosome information, number of genes, numbers of genes located on chromosomes, number of scaffolds, length of scaffold N50), etc. We provide hyperlinks to sequencing literature for each species. Users are allowed to download DNA sequence files, protein sequence files, and general feature format (GFF) files. These species web pages can shorten the data collection time for researchers, and the format-consistent files provides convenience for following genomics research.

Homologous gene dotplotting

Homologous gene dotplotting module is provided to show chromosome-level homology within a genome or between genomes.

266 A homologous dotplot can be directly derived from the BLAST
267 result (Fig. 3A), which contains relatively full information of genomic
268 homology, as compared to the other dotplots shown below. A color
269 scheme for gene-pair dots is adopted to separate the best-matched,
270 often representing orthologs while comparing different genomes,
271 and secondarily matched, and the other matched homologs.

272 A homologous gene dotplot can be drawn by using information
273 of inferred colinear genes and the Ks values between them (Fig. 3B).
274 A color-scheme representing varied Ks values makes it easy to
275 separate colinear blocks produced by different evolutionary event.
276 Owing to the mWGD, a sorghum chromosome corresponds to two
277 overlapping maize chromosome regions. In the meantime, it
278 matched the other rather smaller homologous regions in maize
279 produced due to the cWGD. Through the comparison of multiple
280 sets of data, it is found that the cWGD of Gramineae can be
281 distinguished from the extra whole-genome duplication in Ks=0.65.
282 According to the Ks value of the regions, the four homologous
283 regions of maize were divided into two groups corresponding to
284 different whole-genome duplication events, and each group of
285 regions was divided in detail according to the continuity of
286 chromosome regions.

287

288 **Ks distribution**

289 We provide a module to show Ks distribution between colinear
290 genes. The peak of Ks between colinear genes in each species can
291 help identify WGD event(s), for example, two Ks peaks produced
292 be the maize colinear genes correspond to the occurrence of two
293 WGD events affecting its evolution. The Ks peaks of colinear
294 genes between genomes correspond to the occurrence time of
295 species differentiation and more ancient WGD event(s), e.g., the
296 two peaks of Ks in sorghum-maize colinear genes correspond to
297 their speciation event and the cWGD, respectively.

298

299 **Tools**

300 This section includes tools for comparative genomics analysis to
301 visualize gene colinearity and phylogeny.

302 **Pairwise gene colinearity**

303 The pairwise gene colinearity module can show gene colinearity at
304 the chromosome level or at the gene level. In the chromosome level
305 module, users select reference species and the other compared species, to
306 find their chromosome-level homology. The colinearity at the
307 chromosome level can help infer whether the species has experienced
308 WGDs and/or distinguish the duplicated genes produced by events. In the
309 gene level module, users submit an interested gene ID, select the

reference species, and the other compared species, to produce an alignment of local regions. This can help find genomic changes, homologous/neighboring gene loss, DNA inversion, etc, related to the interested gene.

Multi-genome colinearity list

The multi-genome colinear list module can produce the colinear lists generated for the species under study. There are two types of lists, with one type showing only orthologous genes between species and the other also including (out)paralogs. Users can select the reference genome and the other genomes to map onto the former to show multiple-genome alignment. In addition, we provide a variety of export formats, including excel, pdf, and csv, and copy and print functions.

Multi-genome colinearity alignment map

The multi-genome colinearity map module is used to produce the alignment map for selected species. Users can choose any of the 29 species as reference species for comparative analysis, and maps can also be in two types, corresponding to colinearity lists shown above. With the module, a joint multi-species genome alignment map can be drawn (Fig. 4B).

Multi-genome local colinearity

The multi-genome local colinearity module provides the function of

332 drawing colinear maps in homologous regions of multiple species (Fig.
333 4C). Users select a reference species, enter an interested gene ID in the
334 colinear list of the reference species, select the species names that needs
335 to be compared to the reference species, and then generate a PDF format
336 figure.

337 **Homologous gene evolution tree**

338 The homologous gene evolution tree module can construct a gene
339 evolutionary tree corrected by gene coinearity information. The user
340 selects the reference species and a gene ID, selects the species to compare
341 to. Two gene trees will be generated at the resulting interface, a tree
342 based on pure sequence alignment, and the other one corrected by gene
343 colinearity(Fig. 4D; Fig. 4E).

344 Actually, we found 46% of maize genes have elevated their
345 evolutionary rates and resulted in weird phylogeny. By retrieving maize
346 genes and their orthologous and colinear genes from sorghum, foxtail
347 millet, rice, and weeping lovegrass (taken as outgroup), we constructed
348 7014 evolutionary trees and found that in 46% (3231) trees the maize
349 genes seemed to have elevated rates (Supplemental fig. 1). In trees with
350 one mWGD paralog, with the other one likely lost or relocated to other
351 genomic regions, 38% (1778) showed elevated rates. In 1453 trees with
352 two mWGD paralogs, 29% have both genes to have elevated rates and 69%
353 to have only one gene to have elevated rates. These findings may be

354 explained by the instability of the maize genome after the mWGD .

355 **Help interface**

356 In the help interface, we provide the researcher with a detailed
 357 GGDB user manual. Users can view detailed parameter
 358 descriptions and instructions for each function in this interface.

359

360

361 DISCUSSION

362 Grameneae plants have been recursively affected by polyploidization,
 363 which makes their genome much complex to decipher^[55; 56]. Owing to this
 364 fact, gene collinearity inference in present databases is not well related to
 365 each polyploidization event, which holds back the research to understand
 366 gene function origination and innovation. Actually, the appearance and
 367 establishment of novel gene functions can often be related the production
 368 and divergence of duplicated genes, the most of which were produced by
 369 ancestral polyploidization^[57; 58]. Here, by separating duplicated genes
 370 produced by different polyploidization, and separating orthologous genes
 371 from outparalogous when comparing different species, we set up the
 372 GGDB to store event-related collinear genes in 29 Grameneae plants and
 373 one outgroup. The related work is intense in that 420 pairwise
 374 comparisons have been done between genomes and eventually
 375 hierarchically constructed multiple comparisons by selecting reference
 376 genomes in each major groups, subfamilies or genus. To separate
 377 homologs produced by different events involved, computational analysis
 378 of their divergence and artificial identification of complement
 379 homologous blocks aroused by chromosome breakages were performed.

380 The present database provided friendly tools for the users to show
 381 pairwise or multiple genome alignment in the global or local levels,

382 and produce lists of homologous genes, related to evolutionary
383 events, which provides opportunities to perform deep study of their
384 evolution and functional innovation. Figures of homologous gene
385 dotplots, alignment of selected genomes, and evolutionary trees of
386 selected genes can be downloaded for further research on genome
387 structural changes, chromosome rearrangements, gene losses, and
388 gene functional evolution.

389 In that different copies of homologous genes have divergent
390 evolutionary rates^[59; 60], we provided a module to correct evolutionary
391 trees constructed purely using sequence alignment, by using information
392 of gene colinearity and shared evolutionary events. Actually, elevated
393 evolutionary rates were observed in considerable percentages of
394 paralogous genes produced by polyploidization^[48; 53], which often resulted
395 in aberrant evolutionary trees that cannot be corrected by selecting
396 methods to construct the trees. The tree correction module provides a
397 means to make a realistic tree for the researchers to manipulate for their
398 further study of gene evolution and functional innovation.

399 In the future, with more and more genome sequences released,
400 we will continue to add new genome data in the GGDB. We also
401 encourage users to submit their new Gramineae sequencing data
402 sets to the GGDB to enrich and improve the database. GGDB will

act as a comparative genomics platform for genomics research of
Gramineae and the other related Monocotyledons.

FIGURES

406
407

Figure 1 Composition structure of GGDB database. A. A phylogenetic tree of
Gramineae species involved in the database; B. the computer languages used to set up
tiers of the database and the functional interfaces.

Figure 2. Logical relationship of modules in the database.

Figure 3. Evolutionary Analysis of homologous correspondence between sorghum and maize genomes. A. Homologous gene dotplot with blocks related to divergent ancestral whole-genome duplication events, the Gramineae-common one (cWGD) and the maize-specific one (mWGD); B. Inferred gene colinearity blocks, colored as the synonymous nucleotide substitution rates (Ks) and related to divergent whole-genome duplication events by colored arrows.

Figure 4. Multi-genome colinearity analysis. A. An expected gene tree of colinear homologs from selected species. Besides a common whole-genome duplication (cWGD), some species have been further affected by extra whole-genome duplications (eWGDs); *Sateria italica* (Sit), *Panicum hallii* (Pha), *Zea mays* (Zma), *Sorghum bicolor* (Sbi), *Miscanthus sinensis* (Msi), and *Coix lachryma* (Cla) are involved; B. Multi-genome alignment at a genome level; C. Multi-genome alignment in local homologous regions with millet as the reference; D. A tree based on pure sequence alignment; E. A corrected tree based on by gene colinearity information.

TABLES

Table 1. 29 plants currently involved in the GGDB

Species name	Common name	Release version	Gene number
<i>Avena atlantica</i>	<i>Avena atlantica</i>	Version 1.0 (Nov 2019)	45724
<i>Avena eriantha</i>	<i>Avena eriantha</i>	Version 1.0 (Nov 2019)	46234
<i>Avena strigosa</i>	Oats	Version 1.0 (May 2021)	39812
<i>Aegilops tauschii</i>	Secundum	Version 1.0 (May 2021)	59569
<i>Brachypodium distachyon</i>	Purple falsebrome	Version 3.0 (Feb 2010)	37797
<i>Brachypodium hybridum</i>	<i>Brachypodium hybridum</i>	Version 1.1 (Jul 2020)	80980
<i>Brachypodium stace</i>	<i>Brachypodium stace</i>	Version 1.1 (Jul 2020)	36332
<i>Cenchrus purpureus</i>	Elephant grass	Version 1.0 (Oct 2020)	63758
<i>Cleistogenes songorica</i>	Awnless cleistogenes	Version 1.0 (Sep 2020)	55318
<i>Coix lachryma-jobi</i>	Coix seed	Version 1.0 (May 2020)	64296
<i>Eragrostis curvula</i>	Weeping lovegrass	Version 1.0 (Jul 2019)	32741
<i>Miscanthus sinensis</i>	<i>Miscanthus</i>	Version 1.0 (Oct 2020)	82046
<i>Oryza rufipogon</i>	<i>Oryza rufipogon</i>	Version 1.0 (Jul 2019)	44059
<i>Oryza sativa</i>	Rice	Version 7.0 (Jan 2007)	48876
<i>Oryza sativa-indica</i>	Rice	Version 1.0 (Jan 2015)	37358
<i>Panicum hallii</i>	Panicum	Version 3.1 (Dec 2018)	37542
<i>Panicum miliaceum</i>	Millet	Version 1.0 (Jan 2019)	85636
<i>Phyllostachys edulis</i>	Moso bamboo	Version 1.0 (Oct 2018)	49085
<i>Panicum virgatum</i>	Switchgrass	Version 1.0 (Jan 2021)	129186
<i>Saccharum spontaneum</i>	Modern sugarcane	Version 1.0 (Oct 2018)	67456
<i>Saccharum spp-R570</i>	Sugarcane	Version 1.0 (Jul 2018)	24341
<i>Secale cereale</i>	Rye	Version 1.0 (Mar 2021)	43928
<i>Setaria italica</i>	Foxtail millet	Version 1.0 (Jun 2021)	35591

<i>Setaria viridis</i>	<i>Setaria viridis</i>	Version 1.0 (Oct 2020)	39114
<i>Sorghum bicolor</i>	Sorghum	Version 1.0 (Nov 2018)	39016
<i>Triticum aestivum</i>	Wheat	Version 1.0 (Jun 2021)	130379
<i>Triticum urartu</i>	Durum wheat	Version 1.0 (Apr 2019)	56809
<i>Zea mays</i>	Corn	Version 1.2 (Feb 2009)	56906
<i>Zea mays-MO17</i>	Corn	Version 1.0 (Apr 2018)	46530

465

466

467

468

469

470

471

472

473

474

475

476

477

478

479

480

481

482

483

484

485

486

487

488

489

490

491

492

493

494

495

Table 2. List of orthologous information of some species of Gramineae

	Bdi	Msi	Osa	Pmi	Pvi	Sbi	Sit	Svi	Tae	Tur	Zma
Bdi	\	30499	21053	30007	37465	19857	20782	21514	52530	13137	25708
Msi	2249	\	27451	47405	60769	34149	29315	32667	72889	17040	41732
Osa	1144	2133	\	32177	38906	21147	21397	22103	49571	12159	26025
Pmi	1397	3279	1236	\	56986	31625	31097	29265	63541	16961	36356
Pvi	2654	5401	2547	3910	\	42055	43185	45356	92787	23668	54055
Sbi	1327	2132	1068	1238	2456	\	17520	21621	47707	11872	29369
Sit	1434	2269	1184	1345	2516	1353	\	8954	47686	11494	27314
Svi	1481	2358	1227	1425	2764	1440	1398	\	48844	12102	28198
Tae	2548	5482	2470	3560	6457	2311	2515	2630	\	54123	40335
Tur	1406	2248	1345	1810	2991	1380	1342	1432	2979	\	17563
Zma	1662	3265	1465	1960	3931	1590	1574	1762	3591	1822	\

Note: Above the diagonal is the number of orthologs genes, and below the diagonal is the number of orthologs blocks.

501

Table 3 Homologous blocks produced by different whole-genome duplication

	events									
	Common whole-genome duplication					Extra whole-genome duplication				
	5	10	20	50	Gene Number	5	10	20	50	Gene Number
Aat	471	275	95	39	5422					
Aer	409	242	101	39	4900					
Ast	571	282	96	46	6112					
Ata	488	412	91	17	5137					
Bdi	547	334	79	27	6809					
Bhy	563	95	2	/	2935	1086	602	273	153	50953
Bst	511	284	71	28	6088					
Cla	780	473	126	32	8016					
Cpu	648	409	241	108	12692	1375	495	149	75	39248
Cso	1539	611	229	104	12363	325	213	165	124	36212
Ecu	349	308	120	26	3800					
Msi	1149	472	136	12	9207	1159	866	455	164	31517
Oru	196	114	28	7	2381					
Osa	473	291	85	34	7891					
Ped	1132	236	22	/	7099	1393	752	211	3	16349
Pha	599	361	92	35	8090					
Pmi	883	403	159	56	9983	364	205	132	92	40148
Pvi	1399	526	156	14	11307	2214	1313	646	259	50424
Sbi	601	334	98	43	7928					
Sce	527	331	62	19	4215					
Sit	677	340	86	33	8239					
Ssp	693	1392	738	355	9403					
Ssp_R	175	113	63	32	2471					
Svi	700	398	104	33	8444					

Tae	3043	957	382	73	16125	3402	1422	691	319	108771
Tur	345	153	13	2	2571					
Zma	760	285	108	26	7629	392	224	131	58	13134

502

Parsed Citations

1. Watson L, Dallwitz M J, Weiller C, et al. The grass genera of the world[J]. 1992, 16(2): 151-152.
Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)
- [2] Sato K, Abe F, Mascher M, et al. Chromosome-scale genome assembly of the transformation-amenable common wheat cultivar 'Fielder'[J]. DNA Res, 2021, 28(3).**
3. Ouyang S, Zhu W, Hamilton J, et al. The TIGR Rice Genome Annotation Resource: improvements and new features[J]. Nucleic Acids Res, 2007, 35(Database issue): D883-7.
Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)
4. Hu Y, Colantonio V, Müller B S F, et al. Genome assembly and population genomic analysis provide insights into the evolution of modern sweet corn[J]. Nat Commun, 2021, 12(1): 1227.
Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)
5. Deschamps S, Zhang Y, Llaca V, et al. A chromosome-scale assembly of the sorghum genome using nanopore sequencing and optical mapping[J]. Nat Commun, 2018, 9(1): 4844.
Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)
6. Guo L, Qiu J, Han Z, et al. A host plant genome (*Zizania latifolia*) after a century-long endophyte infection[J]. Plant J, 2015, 83(4): 600-9.
Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)
7. Tanaka H, Hirakawa H, Kosugi S, et al. Sequencing and comparative analyses of the genomes of zoysiagrasses[J]. DNA Res, 2016, 23(2): 171-80.
Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)
8. Schnable J C, Springer N M, Freeling M. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss[J]. Proc Natl Acad Sci U S A, 2011, 108(10): 4069-74.
Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)
9. Jiao Y, Li J, Tang H, et al. Integrated syntenic and phylogenomic analyses reveal an ancient genome duplication in monocots[J]. Plant Cell, 2014, 26(7): 2792-802.
Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)
10. Soltis P S, Soltis D E. Ancient WGD events as drivers of key innovations in angiosperms[J]. Curr Opin Plant Biol, 2016, 30: 159-65.
Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)
11. Soltis P S, Marchant D B, Van De Peer Y, et al. Polyploidy and genome evolution in plants[J]. Curr Opin Genet Dev, 2015, 35: 119-25.
Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)
12. Zhang Y, Zheng C, Sankoff D. A branching process for homology distribution-based inference of polyploidy, speciation and loss[J]. Algorithms Mol Biol, 2019, 14: 18.
Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)
13. Zhang Y, Zheng C, Sankoff D. Distinguishing successive ancient polyploidy levels based on genome-internal syntenic alignment[J]. BMC Bioinformatics, 2019, 20(Suppl 20): 635.
Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)
14. Tang H, Bowers J E, Wang X, et al. Angiosperm genome comparisons reveal early polyploidy in the monocot lineage[J]. Proc Natl Acad Sci U S A, 2010, 107(1): 472-7.
Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)
15. Paterson A H, Bowers J E, Chapman B A. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics[J]. Proc Natl Acad Sci U S A, 2004, 101(26): 9903-8.
Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)
16. Bowers J E, Chapman B A, Rong J, et al. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events[J]. Nature, 2003, 422(6930): 433-8.
Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)
17. Wang X, Shi X, Hao B, et al. Duplication and DNA segmental loss in the rice genome: implications for diploidization[J]. New Phytol, 2005, 165(3): 937-46.
Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)
18. Wang X, Tang H, Paterson A H. Seventy million years of concerted evolution of a homoeologous chromosome pair, in parallel, in major Poaceae lineages[J]. Plant Cell, 2011, 23(1): 27-37.

- Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)
19. Murat F, Xu J H, Tannier E, et al. Ancestral grass karyotype reconstruction unravels new mechanisms of genome shuffling as a source of plant evolution[J]. *Genome Res*, 2010, 20(11): 1545-57.
Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)
 20. Yuan J, Wang J, Yu J, et al. Alignment of Rutaceae Genomes Reveals Lower Genome Fractionation Level Than Eudicot Genomes Affected by Extra Polyploidization[J]. *Front Plant Sci*, 2019, 10: 986.
Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)
 21. Lovell J T, Macqueen A H, Mamidi S, et al. Genomic mechanisms of climate adaptation in polyploid bioenergy switchgrass[J]. *Nature*, 2021, 590(7846): 438-444.
Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)
 22. Zhang J, Wu F, Yan Q, et al. The genome of *Cleistogenes songorica* provides a blueprint for functional dissection of dimorphic flower differentiation and drought adaptability[J]. *Plant Biotechnol J*, 2021, 19(3): 532-547.
Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)
 23. Wei C, Wang Z, Wang J, et al. Conversion between 100-million-year-old duplicated genes contributes to rice subspecies divergence[J]. *BMC Genomics*, 2021, 22(1): 460.
Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)
 24. Liu C, Wang J, Sun P, et al. Illegitimate Recombination Between Homeologous Genes in Wheat Genome[J]. *Front Plant Sci*, 2020, 11: 1076.
Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)
 - [25] Wang Y P, Tang H B, Debarry J D, et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity[J]. *Nucleic Acids Research*, 2012, 40(7).
 26. Lee T H, Tang H B, Wang X Y, et al. PGDD: a database of gene and genome duplication in plants[J]. *Nucleic Acids Research*, 2013, 41(D1): D1152-D1158.
Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)
 27. Nelson A D L, Haug-Baltzell A K, Davey S, et al. EPIC-CoGe: managing and analyzing genomic data[J]. *Bioinformatics*, 2018, 34(15): 2651-2653.
Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)
 28. Zhuang W J, Chen H, Yang M, et al. The genome of cultivated peanut provides insight into legume karyotypes, polyploid evolution and crop domestication[J]. *Nature Genetics*, 2019, 51(5): 865-+.
Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)
 29. Wang J P, Sun P C, Li Y X, et al. An Overlooked Paleotetraploidization in Cucurbitaceae[J]. *Molecular Biology and Evolution*, 2018, 35(1): 16-26.
Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)
 30. Song X M, Sun P C, Yuan J Q, et al. The celery genome sequence reveals sequential paleo-polyploidizations, karyotype evolution and resistance gene reduction in apiales[J]. *Plant Biotechnology Journal*, 2021, 19(4): 731-744.
Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)
 31. Song X M, Wang J P, Li N, et al. Deciphering the high-quality genome sequence of coriander that causes controversial feelings[J]. *Plant Biotechnology Journal*, 2020, 18(6): 1444-1456.
Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)
 32. Zhu Y, Davis S, Stephens R, et al. GEOmetadb: powerful alternative search engine for the Gene Expression Omnibus[J]. *Bioinformatics*, 2008, 24(23): 2798-800.
Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)
 33. Zhou H, Jin J, Zhang H, et al. IntPath--an integrated pathway gene relationship database for model organisms and important pathogens[J]. *BMC Syst Biol*, 2012, 6 Suppl 2(Suppl 2): S2.
Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)
 34. Zheng W, Mutha N V, Heydari H, et al. NeisseriaBase: a specialised Neisseria genomic resource and analysis platform[J]. *PeerJ*, 2016, 4: e1698.
Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)
 35. Zhang Y, Xu B, Yang Y, et al. CPSS: a computational platform for the analysis of small RNA deep sequencing data[J]. *Bioinformatics*, 2012, 28(14): 1925-7.
Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)
 36. Zhang X, Sun X F, Cao Y, et al. CBD: a biomarker database for colorectal cancer[J]. *Database (Oxford)*, 2018, 2018.
Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)

37. Zhang Q, Yang B, Chen X, et al. Renal Gene Expression Database (RGED): a relational database of gene expression profiles in kidney disease[J]. Database (Oxford), 2014, 2014.
Google Scholar: [Author Only Title Only Author and Title](#)
38. Lee T H, Tang H, Wang X, et al. PGDD: a database of gene and genome duplication in plants[J]. Nucleic Acids Res, 2013, 41(Database issue): D1152-8.
Google Scholar: [Author Only Title Only Author and Title](#)
39. Genome sequencing and analysis of the model grass *Brachypodium distachyon*[J]. Nature, 2010, 463(7282): 763-8.
Google Scholar: [Author Only Title Only Author and Title](#)
40. Zou C, Li L, Miki D, et al. The genome of broomcorn millet[J]. Nat Commun, 2019, 10(1): 436.
Google Scholar: [Author Only Title Only Author and Title](#)
41. Zhang J, Zhang X, Tang H, et al. Allele-defined genome of the autopolyploid sugarcane *Saccharum spontaneum* L.[J]. Nat Genet, 2018, 50(11): 1565-1573.
Google Scholar: [Author Only Title Only Author and Title](#)
42. Wang X, Shi X, Li Z, et al. Statistical inference of chromosomal homology based on gene colinearity and applications to *Arabidopsis* and rice[J]. BMC Bioinformatics, 2006, 7: 447.
Google Scholar: [Author Only Title Only Author and Title](#)
43. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood[J]. Mol Biol Evol, 2007, 24(8): 1586-91.
Google Scholar: [Author Only Title Only Author and Title](#)
44. Wang Y, Tang H, Debarry J D, et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity[J]. Nucleic Acids Res, 2012, 40(7): e49.
Google Scholar: [Author Only Title Only Author and Title](#)
45. Edgar R C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity[J]. BMC Bioinformatics, 2004, 5: 113.
Google Scholar: [Author Only Title Only Author and Title](#)
46. Nguyen L T, Schmidt H A, Von Haeseler A, et al. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies[J]. Mol Biol Evol, 2015, 32(1): 268-74.
Google Scholar: [Author Only Title Only Author and Title](#)
47. Price M N, Dehal P S, Arkin A P. FastTree 2--approximately maximum-likelihood trees for large alignments[J]. PLoS One, 2010, 5(3): e9490.
Google Scholar: [Author Only Title Only Author and Title](#)
48. Meng F, Pan Y, Wang J, et al. Cotton Duplicated Genes Produced by Polyploidy Show Significantly Elevated and Unbalanced Evolutionary Rates, Overwhelmingly Perturbing Gene Tree Topology[J]. Front Genet, 2020, 11: 239.
Google Scholar: [Author Only Title Only Author and Title](#)
49. Liu K, Linder C R, Warnow T. RAXML and FastTree: comparing two methods for large-scale maximum likelihood phylogeny estimation[J]. PLoS One, 2011, 6(11): e27731.
Google Scholar: [Author Only Title Only Author and Title](#)
50. Fouquier J, Rideout J R, Bolyen E, et al. Ghost-tree: creating hybrid-gene phylogenetic trees for diversity analyses[J]. Microbiome, 2016, 4: 11.
Google Scholar: [Author Only Title Only Author and Title](#)
51. Wang R, Perez-Riverol Y, Hermjakob H, et al. Open source libraries and frameworks for biological data visualisation: a guide for developers[J]. Proteomics, 2015, 15(8): 1356-74.
Google Scholar: [Author Only Title Only Author and Title](#)
52. Yachdav G, Wilzbach S, Rauscher B, et al. MSASviewer: interactive JavaScript visualization of multiple sequence alignments[J]. Bioinformatics, 2016, 32(22): 3501-3503.
Google Scholar: [Author Only Title Only Author and Title](#)
53. Wang X, Wang J, Jin D, et al. Genome Alignment Spanning Major Poaceae Lineages Reveals Heterogeneous Evolutionary Rates and Alters Inferred Dates for Key Evolutionary Events[J]. Mol Plant, 2015, 8(6): 885-98.
Google Scholar: [Author Only Title Only Author and Title](#)
54. Zuloaga F O, Salariato D L, Scataglini A. Molecular phylogeny of *Panicum* s. str. (Poaceae, Panicoideae, Paniceae) and insights into its biogeography and evolution[J]. PLoS One, 2018, 13(2): e0191529.
Google Scholar: [Author Only Title Only Author and Title](#)
55. Mckain M R, Tang H, Mcneal J R, et al. A Phylogenomic Assessment of Ancient Polyploidy and Genome Evolution across the Poales[J]. Genome Biol Evol, 2016, 8(4): 1150-64.

Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)

56. Ma P F, Liu Y L, Jin G H, et al. The Pharus latifolius genome bridges the gap of early grass evolution[J]. Plant Cell, 2021, 33(4): 846-864.

Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)

57. Cheng F, Wu J, Cai X, et al. Gene retention, fractionation and subgenome differences in polyploid plants[J]. Nat Plants, 2018, 4(5): 258-268.

Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)

58. Wang X, Gowik U, Tang H, et al. Comparative genomic analysis of C4 photosynthetic pathway evolution in grasses[J]. Genome Biol, 2009, 10(6): R68.

Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)

59. Chalhoub B, Denoeud F, Liu S, et al. Plant genetics. Early allopolyploid evolution in the post-Neolithic Brassica napus oilseed genome[J]. Science, 2014, 345(6199): 950-3.

Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)

60. The tomato genome sequence provides insights into fleshy fruit evolution[J]. Nature, 2012, 485(7400): 635-41.

Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)

AVAILABILITY

The GGDB can be accessed through the web server at <http://www.grassgenome.com/>.