Sinbase: An Integrated Database to Study Genomics, Genetics and Comparative Genomics in Sesamum indicum

Linhai Wang1,2, Jingyin Yu1,2, Donghua Li1 and Xiurong Zhang1,*

1Oil Crops Research Institute of the Chinese Academy of Agricultural Sciences, Key Laboratory of Biology and Genetic Improvement of Oil Crops, Ministry of Agriculture, PR China
2These authors contributed equally to this work.
*Corresponding author: E-mail, zhangxr@oilcrops.cn; Fax, +86-27-86811836.
(Received August 20, 2014; Accepted November 5, 2014)

Sesame (Sesamum indicum L.) is an ancient and important oilseed crop grown widely in tropical and subtropical areas. It belongs to the gigantic order Lamiales, which includes many well-known or economically important species, such as olive (Olea europaea), leonurus (Leonurus japonicus) and lavender (Lavandula spica), many of which have important pharmacological properties. Despite their importance, genetic and genomic analyses on these species have been insufficient due to a lack of reference genome information. The now available S. indicum genome will provide an unprecedented opportunity for studying both S. indicum genetic traits and comparative genomics. To deliver S. indicum genomic information to the worldwide research community, we designed Sinbase, a web-based database with comprehensive sesame genomic, genetic and comparative genomic information. Sinbase includes sequences of assembled sesame pseudomolecular chromosomes, protein-coding genes (27,148), transposable elements (372,167) and non-coding RNAs (1,748). In particular, Sinbase provides unique and valuable information on colinear regions with various plant genomes, including Arabidopsis thaliana, Glycine max, Vitis vinifera and Solanum lycopersicum. Sinbase also provides a useful search function and data mining tools, including a keyword search and local BLAST service. Sinbase will be updated regularly with new features, improvements to genome annotation and new genomic sequences, and is freely accessible at http://ocri-genomics.org/Sinbase/.

Keywords: Comparative genomics • Database • Genetics • Genomics • Sesamum indicum.

Abbreviations: InDel, insertion and deletion; SNP, single nucleotide polymorphism; SSR, simple sequence repeat.

Introduction

Sesame (Sesamum indicum L.), an ancient and important oilseed crop, has earned the poetic label ‘queen of oilseeds’ due to the high oil yield and quality of the seed. The higher oil content and fewer lipid-related genes in the small and diploid genome of sesame make it an invaluable plant model for studying oil biosynthesis (Wang et al. 2014). In addition, sesame oil has an excellent composition: >80% is made up of reputable unsaturated fatty acids (mainly oleic acid and linoleic acid, 1:1) (Wei et al. 2013), yet it has an outstanding shelf life due to the presence of antioxidative furofuran lignans (Budowski et al. 1950). The major sesame lignan, sesamin, has been a focus of interest of pharmacologists and was demonstrated to possess many potent pharmacological properties such as decreasing blood lipids (Hirata et al. 1996) and lowering cholesterol levels (Chen et al. 2005). Sesame (2n = 2x = 26) is a member of the family Pedaliaceae and order Lamiales (Angiosperm Phylogeny Group 2009). This order includes many well-known or economically important species, such as olive (Olea europaea), leonurus (Leonurus japonicus), lavender (Lavandula spica) and basil (Ocimum basilicum). Genetic and genomic studies have been lacking in most of these species. Recently, with high-throughput sequencing technology becoming routine, a relatively advanced study in sesame molecular biology was performed. Several simple sequence repeat (SSR) markers have been developed in sesame (Spandana et al. 2012, Wang et al. 2012a, Zhang et al. 2012). The loci associated with the indehiscent capsule trait, definite growth habit and seed coat color have been detected (Uzun et al. 2003, Uzun and Çağrın 2009, Zhang et al. 2013), and the expression levels of sesame genes have been explored using Sanger and high-throughput DNA sequencing technologies (Suh et al. 2003, Wei et al. 2011, Wang et al. 2012b). The phylogenetic position of sesame has been determined using chloroplast genomic data, which indicated the core lineage of Sesamum in Lamiales (Yi and Kim 2012). More specifically, the completed de novo assembly of the genomic sequence of the cultivar ‘Zhongzhi No. 13’ provided new insights into sesame. Sesame was estimated to have diverged from the the Lamiales species Utricularia gibba approximately 98 million years ago (68.6–145.2 MYA) and from the tomato–potato lineage approximately 125 MYA (89.8–185.8 MYA). Its recent whole-genome duplication event was inferred to have occurred independently in the parallel period of the triplication event in the tomato–potato lineage (Wang et al. 2014).

Until now, few public resources have existed for sesame genome data with the exception of the sesame genome project website (http://www.sesamegenome.org/) and the National Center for Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov/) (Pruitt et al. 2007). However, the sesame genome project website did not supply detailed sesame...
genomic or genetic data sets at the time of writing. With respect to sesame, the NCBI presents only partial and less than systematic data, such as genome survey sequences (GSSs), expressed sequence tag (EST) sequences and transcriptome data in Gene Expression Omnibus (GEO) data sets (Barrett et al. 2009). To visualize systematically and understand sesame genomics and genetics, we developed a comprehensive web-based database, referred to as Sinbase (http://ocri-genomics.org/Sinbase), which includes genomic, genetic and comparative genomic information. This user-friendly database is the primary genomic resource for *Sesamum*, which will serve as an infrastructure for researchers to study the molecular function of genomic components, comparative genomics and evolution in related Lamiales species. In this study, we also present an overview of the major sections of Sinbase and some of the tools available for data mining and database searches (Fig. 1).

**Database Implementation**

We developed an integrated database of genomic, genetic and comparative genomic information for the sesame genome. It is currently hosted on a CentOS 5.4 Linux operating system with the Apache Web server. The user-friendly interface was developed using Perl/CGI, Python and JavaScript. Graphical compositions of pseudomolecular chromosomes, distribution of genetic markers on pseudomolecular chromosomes, and gene structures were generated using the SVG software package and Perl GD module from the Comprehensive Perl Archive Network (CPAN) website (Hietaniemi 2008). Basic sesame genome data sets were analyzed using in-house Perl scripts, and all data sets related to the sesame genome were stored in a MySQL rational database management system. GBrowse 2.0 and GBrowse_syn software packages were adapted for visualization of the distribution of genomic components on pseudomolecular chromosomes and relationships colinear to the sesame genome between species (Ashburner et al. 2000, Donlin 2009, McKay et al. 2010).

**Database Contents and Functions**

Currently, Sinbase has collected 16 genetic linkage groups (pseudomolecular chromosomes) of sesame anchored by 406 genetic markers including single nucleotide polymorphisms (SNPs), SSRs and insertion/deletion (InDel) genetic markers, as well as 16,296 scaffolds. Based on these genomic data sets, Sinbase includes all genomic components generated by open-resource software with stringent parameters comprising 27,148 predicted protein-coding genes, 372,167 transposable elements and 1,748 non-coding RNAs (Table 1). With regard to sesame genetics, Sinbase supplies not only the genetic distribution of the 406 DNA markers on pseudomolecular chromosomes, but also their basic information. Moreover, Sinbase offers a comprehensive analysis of comparative genomics, which are displayed using Circos software (Krzywinski et al. 2009) between *S. indicum* and *Arabidopsis thaliana* (Huala et al. 2001), *Glycine max* (Schmutz et al. 2010), *Vitis vinifera* (Jaillon et al. 2007) and *Solanum lycopersicum* (Tomato Genome Consortium 2012) using genomic data from freely available databases.

**Genomic component**

Three typical sesame genomic components, including putative genes (27,148), transposable elements (372,167) and non-coding RNA (1,748), were collected for Sinbase. For each putative gene, Sinbase supplies the functional description, genome analysis and sequences for the sesame-predicted gene. Repetitive elements in sesame are identified employing four methods, i.e. de novo RepeatMasker (ab initio prediction repeats), Proteinmask (prediction by Repeat Protein Mask), RepeatMasker (prediction by RepeatMasker; Repbase is the...
To illustrate the evolutionary relationship and species divergence between sesame and other remarkable species or its relatives, such as *A. thaliana*, *G. max*, *S. lycopersicum* and *V. vinifera*, at both the inter- and intragenomic level, whole-genome comparative analyses were performed, and the colinear regions between sesame and other species were generated using MCscan software (Tang et al. 2008). Sinbase collected 1,058 colinear blocks, including 21,506 orthologous gene pairs among *A. thaliana*, *G. max*, *S. lycopersicum* and *V. vinifera* genomes compared with the *S. indicum* genome (Supplementary Table S1). Using MCscan software, orthologous gene pairs were detected based on an all-against-all BLAST search (Altschul et al. 1997) using protein sequences with a strict parameter: E-value cut-off \( \leq 1e-05 \) between species. The colinear regions were scanned with the following parameters, \( e = 1e-20 \), \( u = 1 \) and \( s = 5 \), which required a minimum of five consecutive orthologous gene pairs in colinear regions. Ultimately, 1,058 colinear regions were detected including 353 in *S. indicum* compared with *A. thaliana*, 377 in *S. indicum* compared with *G. max*, 358 in *S. indicum* compared with *V. vinifera*, and 420 in *S. indicum* compared with *S. lycopersicum* (Fig. 3).

**Utility of Sinbase**

Sinbase is the first comprehensive and integrated resource platform for sesame genomics, genetics and comparative genomics, and provides a user-friendly interface for easy access to sesame genome information. Two critical functional modules were designed in Sinbase: ‘Browse’ and ‘Search’ modules. From the ‘Browse’ module, users can browse sesame phenotypes, sesame genome components, genetic linkage groups and colinear relationships among species in comparison with the sesame genome, as well as the distribution of genome components on sesame pseudomolecular chromosomes shown by GBrowse (Donlin 2009). According to different sections in this module, internal links were equipped to help users maximize the available information. The ‘Search’ module was developed to help users extract sesame genome information, including genome components and corresponding sequences, orthologous genes of sesame in corresponding species, and colinear regions between sesame and corresponding species on pseudomolecular chromosomes. The sequence similarity search function using a standard NCBI BLAST program (Altschul et al. 1997) was included in this module for the user to query nucleotide or protein sequences against a database containing whole-genome data sets of *S. indicum*, *A. thaliana*, *G. max*, *V. vinifera* and *S. lycopersicum*. Data from the sesame genome collected in Sinbase are freely available in bulk.

**Browse**

Users can browse information on five categories: phenotypes, genome components, genetic linkage map, comparative genomics and Genome Browse. The ‘phenotypes’ tab was designed to demonstrate the phenotype variation of sesame, and help users in other fields obtain perceptual knowledge about sesame quickly. It collected a variety of sesame with typical phenotypes, including those with indefinite or definite growth habits, unculm or branching styles, single or multiple flowers per axil, single or multiple capsules per axil, two or more carpels in a capsule, an opened or closed corolla, a normal or dwarf plant and those with different flower color, capsule color, seed coat

**Comparative genomics**

To illustrate the evolutionary relationship and species divergence between sesame and other remarkable species or its relatives, such as *A. thaliana*, *G. max*, *S. lycopersicum* and *V. vinifera*, at both the inter- and intragenomic level, whole-genome comparative analyses were performed, and the colinear regions

---

### Table 1. Statistics of *S. indicum* genome components in Sinbase

<table>
<thead>
<tr>
<th>Categories</th>
<th>Putative genes</th>
<th>Repeat elements</th>
<th>Non-coding RNAs</th>
<th>Total genome components</th>
</tr>
</thead>
<tbody>
<tr>
<td>LG01</td>
<td>2,024</td>
<td>18,719</td>
<td>89</td>
<td>20,832</td>
</tr>
<tr>
<td>LG02</td>
<td>1,934</td>
<td>18,852</td>
<td>77</td>
<td>20,683</td>
</tr>
<tr>
<td>LG03</td>
<td>2,709</td>
<td>35,743</td>
<td>122</td>
<td>38,576</td>
</tr>
<tr>
<td>LG04</td>
<td>1,679</td>
<td>19,264</td>
<td>59</td>
<td>21,002</td>
</tr>
<tr>
<td>LG05</td>
<td>1,679</td>
<td>34,056</td>
<td>93</td>
<td>35,828</td>
</tr>
<tr>
<td>LG06</td>
<td>2,745</td>
<td>36,526</td>
<td>123</td>
<td>39,394</td>
</tr>
<tr>
<td>LG07</td>
<td>1,260</td>
<td>14,310</td>
<td>162</td>
<td>21,772</td>
</tr>
<tr>
<td>LG08</td>
<td>2,535</td>
<td>18,712</td>
<td>143</td>
<td>21,390</td>
</tr>
<tr>
<td>LG09</td>
<td>1,300</td>
<td>12,814</td>
<td>69</td>
<td>14,183</td>
</tr>
<tr>
<td>LG10</td>
<td>1,792</td>
<td>24,803</td>
<td>127</td>
<td>26,722</td>
</tr>
<tr>
<td>LG11</td>
<td>1,617</td>
<td>17,892</td>
<td>81</td>
<td>19,590</td>
</tr>
<tr>
<td>LG12</td>
<td>1,017</td>
<td>4,066</td>
<td>46</td>
<td>5,129</td>
</tr>
<tr>
<td>LG13</td>
<td>522</td>
<td>7,769</td>
<td>36</td>
<td>8,527</td>
</tr>
<tr>
<td>LG14</td>
<td>524</td>
<td>3,178</td>
<td>33</td>
<td>3,735</td>
</tr>
<tr>
<td>LG15</td>
<td>1,187</td>
<td>15,229</td>
<td>112</td>
<td>16,528</td>
</tr>
<tr>
<td>LG16</td>
<td>372</td>
<td>7,033</td>
<td>25</td>
<td>7,430</td>
</tr>
<tr>
<td>Unknown</td>
<td>2,252</td>
<td>83,201</td>
<td>351</td>
<td>85,804</td>
</tr>
</tbody>
</table>
color, capsule size, seed size, and so on. For further information, please refer to the file 'Descriptors for sesame', which is freely accessible at http://www.bioversityinternational.org/e-library/publications/detail/descriptors-for-sesame-emsesamum-em-spp/. For the ‘genome components’ navigation tab, a comprehensive annotation was compiled for three key components of sesame genomics. Putative genes are classified by basic information, genome analysis and sequences, as well as annotation by Gene Ontology (GO; Ashburner et al. 2000), InterPro (Quevillon et al. 2005), Swiss-Prot, TrEMBL (O’Donovan et al. 2002) and the Kyoto Encyclopedia of Genes and Genomes (KEGG; Kanehisa et al. 2008). Each repeat element class is organized into different pseudomolecular chromosomes and is shown as a list in table format. For every repeat element entry, Sinbase offers basic information, including methods on how to obtain this entry, sequence of the entry and type that the entry belongs to, as well as parameters to generate the entry. Similar to repeat element entry, basic information on non-coding RNAs is also supplied. Sinbase also supplies graphics regarding the distribution on genetic linkage groups of sesame genetic markers, which allows users to understand them intuitively. According to what the genetic linkage groups exhibit, users can retrieve the genetic distances and order of those markers, as well as the corresponding scaffolds anchored on pseudomolecular chromosomes. Furthermore, users can obtain basic information, including marker type, physical location on pseudomolecular chromosomes, and flanking sequences, for each sesame genetic marker, which could be employed by genetic breeders and researchers to further their knowledge underlying special traits and phenotypes. For comparative genomics, the relationships between S. indicum and A. thaliana, G. max, V. vinifera and S. lycopersicum genomes are curated in this section, and their colinear regions are displayed using Circos software (Krzywinski et al. 2009). Statistical analyses of gene pairs between colinear regions among different species are scatterplotted by the R programming language. GBrowse_syn is used to display every colinear region between two species to offer users a dynamic and intuitive description (Donlin 2009). Moreover, sequence identities between orthologous gene pairs in colinear regions are displayed in gene pair lists. GBrowse was employed to show the distribution of genome components of the sesame genome on pseudomolecular chromosomes (McKay et al. 2010).

**Search**

Multiple-function search modules were designed to meet a variety of needs for researchers to extract interesting information on the sesame genome in the plant science community. Sinbase search functions refer to three parts according to
sesame genome resources deposited in the database. First, the genomic component search accounts for a large proportion of search functions, which can give users a multichannel search function to obtain sesame genomic component information. This function can be used to access annotations and sequences of putative genes, transposable elements and non-coding RNAs. Secondly, the orthologous genes search panel is an important resource to obtain sesame orthologs in other species. This function is useful for researchers who want to investigate homolog cloning, homolog comparison, and evolution at the gene level. Finally, the colinear regions search function provides homologous regions from the sesame genome to other species. Based on whole pseudomolecular chromosomes, colinear blocks can be visualized as a comparative chromosomal graphic, and lists of colinear regions are shown along with their chromosomal position. This function can aid researchers in their understanding of the colinear and evolutionary relationships between the sesame genome and corresponding species efficiently.

**BLAST**

The similarity sequence search function, which embeds a standard NCBI BLAST software package (Altschul et al. 1997), will assist users in retrieving homologous genome components or regions in Sinbase. Users can upload a sequence file or directly paste the nucleic acid or amino acid sequence(s) into a box to search against all available databases by choosing the correct BLAST program (depending on the sequence type). Thus, this function aids users in extracting annotations of query sequences using the sesame genome data compiled in Sinbase by quick match. By inputting a nucleic acid or amino acid sequence, the similarity sequence search function can also allow users to find the location of query sequences on scaffolds, genetic linkage groups or pseudomolecular chromosomes.

**Discussion**

Few public resources exist on sesame genome information, apart from a sesame genome project website and certain transcriptomic data and genome survey sequences from the NCBI (Pruitt et al. 2007). Sinbase is the first practical and integrated database that focuses on *S. indicum* genomics, genetics and comparative genomics with its relatives and other important species in the plant community. The deposited genomic sequences and relatively accurate genomic component annotations will be convenient for users to study functional genomics and comparative genomics of sesame simultaneously. Compared with another public resource platform for *S. indicum* genomic data (http://www.sesamegenome.org/), Sinbase supplies more comprehensive genomic data sets and detailed...
Conclusions and Perspective

To facilitate functional and comparative genomic studies in sesame, we developed a comprehensive database of the *S. indicum* genome, which includes genomic, genetic and comparative genomic data from sesame. Genomic sequences, genetic linkage groups and colinear regions have been compiled into an integrated sesame genome database. Sinbase consists of two primary functions: a 'browse' section to show data collected in the database and multifunction search systems to display search results. We hope that Sinbase (a comprehensive data resource) will contribute meaningfully not only to functional genomics research, but also to evolutionary biology and molecular breeding.

Supplementary data

Supplementary data are available at PCP online.

Funding

This work was supported by the National Natural Science Foundation of China [No.31101182 and No.31371665]; the National Basic Research Program of China [973 Program, No. 2011CB109304]; and the National Crop Germplasm Resources Infrastructure [2013-014].

Disclosures

The authors have no conflicts of interest to declare.

References


