**funRiceGenes dataset for comprehensive understanding and application of rice functional genes**

---Manuscript Draft---

| Manuscript Number:          | GIGA-D-17-00154R1 |
|----------------------------|-------------------|
| Full Title:                | funRiceGenes dataset for comprehensive understanding and application of rice functional genes |
| Article Type:              | Technical Note    |
| Funding Information:       |                   |
|                            | National Key Research and Development Program of China (2016YFD0100903) Not applicable |
|                            | National Natural Science Foundation of China (31771873) Dr. Yidan Ouyang |
|                            | National Natural Science Foundation of China (31371599) Dr. Yidan Ouyang |
|                            | Outstanding Young Talents Program Dr. Yidan Ouyang |

**Abstract:**

Background: As a main staple food, rice is also a model plant for functional genomic studies of monocots. Decoding of every DNA element of the rice genome is essential for genetic improvement of rice to address the increasing food demands. The past 15 years have witnessed extraordinary advances in rice functional genomic studies. Systematic characterization and proper deposition of every rice gene are vital for both functional studies and crop genetic improvement.

Findings: We built a comprehensive and accurate dataset of ~2,800 functionally characterized rice genes and ~5,000 members of different gene families, by integrating data from available database and reviewing of every publication of rice functional genomic studies. The dataset accounts for 19.2% of the 39,045 annotated protein-coding rice genes, which provides the most exhaustive archive for investigating the functions of rice genes. We also constructed 214 gene interaction networks based on 1,841 connections between 1,310 genes. The largest network with 762 genes indicated that pleiotropic genes linked different biological pathways. Increasing degree of conservation of the flowering pathway was observed among closer related plants, implying substantial value of rice genes for future dissection of flowering regulation in other crops. All data are deposited in the funRiceGenes database (https://funricegenes.github.io/). Functionality for advanced search and continuous updating of the database are provided by a Shiny application (http://funricegenes.ncpgr.cn/).

Conclusions: The funRiceGenes dataset would enable further exploring of the crosslink between gene functions and natural variations in rice, which can also facilitate breeding design to improve target agronomic traits of rice.

**Corresponding Author:**

Yidan Ouyang  
CHINA

**First Author:**

Wen Yao

**Order of Authors:**

Wen Yao  
Guangwei Li
Dear Editor:

We are now submitting our revised manuscript, entitled funRiceGenes dataset, for comprehensive understanding and application of rice functional genes, to your consideration for publication in GigaScience. This manuscript was assigned GIGA-D-17-00154 in the previous submission. We are grateful to the editor’s and reviewers’ comments and suggestions. We have done a thorough revision of the manuscript to address the editor’s and reviewers’ concerns. We have rewritten many parts of the manuscript to improve the writing of the manuscript. In pages that follow, point-by-point responses to the comments and suggestions by the editor and reviewers are provided. We will greatly appreciate your attention to this submission.

Sincerely,

Wen Yao, Ph. D.
National Key Laboratory of Crop Genetic Improvement
Huazhong Agricultural University
Wuhan 430070, China

Yidan Ouyang, Ph. D.
Professor, College of Life Science and Technology
National Key Laboratory of Crop Genetic Improvement
Huazhong Agricultural University
Wuhan 430070, China

EDITOR COMMENTS

I agree with reviewer 2 that such a database needs to be up-to-date if we understand correctly, data incorporation was up to 2014 only, and I agree with reviewer 2 that it should be up-to-date.
Response: Thanks for the suggestion and we agree with this. Actually, our database is almost always up-to-date. Data collection from different sources until 2014 mentioned in the manuscript were performed for initial construction of the database. Since then, this database was updated by tracking publications from PubMed and new records from the China Rice Data Center and Oryzabase database using a Shiny application. All the updated records are available at https://funricegenes.github.io/news/. The latest update was performed on Sep 20th, 2017.

In addition, it is not quite clear in how far this new dataset is an advance over existing resources - this needs to be discussed and explained in detail. As reviewer 2 says, “clear examples where functional descriptions were improved by the authors’ effort need to be provided.”
Response: Many thanks for your valuable suggestions. We discussed the advance of the funRiceGenes database in the first paragraph of the Discussion section of the revised manuscript.

In this study, we built a comprehensive and accurate database of functionally characterized rice genes, funRiceGenes, which provides a valuable resource for rice functional genomic studies. funRiceGenes was constructed by integrating data from PubMed, Oryzabase, and China Rice Data Center, and was updated every two weeks using a Shiny application. For each gene in the funRiceGenes database, the gene symbol, the genomic locus in the reference genome and the published papers on this gene were identified. Compared with Textpresso for Oryza sativa (http://map.lab.nig.ac.jp:8095/textpresso/index.html), which is a comprehensive collection of literatures on rice, we further built the associations between genomic locus or symbol of genes and literatures. Based on the literatures identified for each gene, we summarized the brief functions of each gene and constructed interaction networks for all genes. The evidences supporting the functions of all collected genes and the interaction networks are unique to the funRiceGenes database. In addition, user-friendly query interface and tidy data for downloading are provided in the funRiceGenes database.
An interesting feature of your submission are the automatic updates to the database. Please elaborate on this feature, and how the updates are implemented in practice, as it seems to be a useful functionality that may convince the reviewers regarding the merits of your manuscript.

Response: Many thanks for your valuable suggestions. We have given an in-depth description on the automatic updates to the database (from page 5 line 17-25 to page 6 line 1-4 of the revised manuscript). The process for implementation of the updates using the Shiny application was described in the help manual (https://funricegenes.github.io/help.pdf).

New genes were added to this database using the Shiny application, based on daily email alert of the searching results from the PubMed database with the keyword rice (rice[Title] OR rice[Title/Abstract]) (https://funricegenes.github.io/help.pdf). For all the PubMed records in the email alert, we identified ones on functionally characterized rice genes. We then went over the full publication of each record and identified the gene symbol and gene model in the reference genome. After inputting the gene symbol, the gene model in the reference genome and the PubMed identifier, the Shiny application will fetch the corresponding publication record from PubMed and extract key information automatically. We also kept track of new records in the database of Oryzabase and China Rice Data Center, which were then added to our database using the Shiny application. Since 13 Feb 2014, funRiceGenes was updated every two weeks using the Shiny application. All the updated records are available at https://funricegenes.github.io/news/.

Regarding the article type, in case of acceptance, we feel the manuscript would be suitable as a "Data Note" (https://academic.oup.com/gigascience/pages/data_note), or maybe also as a "Technical Note" - we can discuss this further when you submit a revised manuscript.

Response: Many thanks for your suggestion. We would like to change our manuscript as a "Technical Note".

REVIEWER COMMENTS
Reviewer: 1
The manuscript provides an integration of publicly available information on rice gene functions and associated attributes from heterogenous sources, in order to make the information available for biological interpretation. A number of search tools have been developed or applied to derive associations between heterogeneous data subjects. These associations have also been used to derive networks of functional associations from literature that can provide a basis for further searches.

The interactive search page with a Shiny application for updating was tested with a number of genes of interest, and they made links between loci numbers and new publications, providing a potential gene function from available literature. I see that as a very good tool to test data and hypotheses in a research. Although the interactive page is a bit slow, and might be even more with more traffic from searches, it is user friendly and would be an asset for researchers doing GWAS or gene function identification. The utility for gene function information goes beyond Gramene and RAPdb, but will only be able to remain so if the planned automatic updates to the database remain functional.

Response: Many thanks for the positive comments. We updated the funRiceGenes database every two weeks since its initial construction in 2014. Since 2014, this database was updated using a Shiny application by tracking publications from PubMed and new records in the China Rice Data Center and Oryzabase databases. All updated records are available at https://funricegenes.github.io/news/, with the latest update performed on Sep 20th, 2017. We will keep updating of the funRiceGenes database in future.

The speed of the interactive page is probably restricted by the internet speed in our university. However, the Shiny application can be downloaded and deployed on local computer, which can be then accessed without speed limit. Please check the help manual (https://funricegenes.github.io/help.pdf) for downloading and deploying of the Shiny application on local computer.

Since the Nipponbare genome basis and annotation is used, is there a potential to survey overlapping genomic intervals from the indica genome sequences and make predictions of intervening syntenic genes?
Response: Thanks for your valuable suggestion. We provide functions allowing conversion between indica and japonica syntenic gene IDs in the IDConversion menu of the updated Shiny application (http://funricegenes.ncpgr.cn/), based on synteny analysis between Nipponbare genome and two high-quality indica reference genomes reported in Zhang et al. 2016, PNAS (http://www.pnas.org/content/113/35/E5163.full). In the conversion result, we provide links to the RIGW database (http://rice.hzau.edu.cn/), which contains the detailed information for the indica genes. In the RIGW database, syntenic alignments between the Nipponbare and two indica genomes are provided (http://rice.hzau.edu.cn/cgi-bin/gb2/gbrowse_syn/3rice_syn/).

Is the search scalable to use larger datasets or gene lists rather than individual genes to derive hypotheses from experimental data, eg what would be the pathways affected from mutation of a specific candidate gene, when no experimental data is available? Or, could one predict candidate genes that might perturb/affec specific biological process. The availability of other network-based predictive methods and integration into funRiceGenes would be able to provide further tools for experimenters.

Response: Thanks for your valuable suggestions. We provided batch query functions allowing search of the funRiceGenes database with gene lists in the Download menu of the updated Shiny application (http://funricegenes.ncpgr.cn/). We also integrated the data from the RiceNet V2 database into funRiceGenes, which provides genome-scale probabilistic functional gene networks of O. sativa (RiceNet v2: an improved network prioritization server for rice genes, Nucl. Acids Res, 2015, 43:W122-7).

The funRiceGenes application on publications has similarities to the Texptresso application for many model systems from Arabidopsis (http://www.textpresso.org/arabidopsis/) to mouse and also initiated for Oryza sativa (http://map.lab.nig.ac.jp:8095/textpresso/index.html). This rice functional genomics application funRiceGenes should be shown how it distinguishes from the textpresso tool with differences outlined in the manuscript.

Response: Many thanks for your valuable suggestions. Textpresso provides an archive of biological literature allowing information extracting by keywords. Only if the symbol of a gene is present in the title and/or the abstract of published papers, matched results will be shown. In addition to information extracting by keywords, the funRiceGenes database allows searching by gene symbol and genomic locus from either MSU or RAPdb (e.g., LOC_Os07g15770 or Os05g0158500), as the funRiceGenes database builds the associations between genomic locus of a gene and related published papers. Besides, funRiceGenes also lists all the genes related to a specified publication, which provides another option for information retrieving. We discussed this in the first paragraph of the Discussion section in the revised manuscript.

Reviewer: 2

The authors have created a new database, funRiceGenes, which contains functional information of rice genes and some other related data. The data were first collected from other databases and manually curated. The database is possibly useful, but I have some serious concerns as follows: Oryzabase, which was created in 2000 and is still actively maintained, harbors a large amount of literature information. https://shigen.nig.ac.jp/rice/oryzabase/about/oryzabase

Though the data of Oryzabase are all curated, the authors seemed to re-curate them, and I don't understand why this was needed and what really had to be done.

Response: A number of genes archived in Oryzabase are merely members of gene families identified by bioinformatics analysis. We need to separate them from genes functionally characterized by experiments. In addition, Oryzabase also contains quantitative trait loci (QTL) associated with agronomic traits and assigns gene symbols to these QTL (https://shigen.nig.ac.jp/rice/oryzabase/gene/advanced/list). However, the casual gene of these QTL has not been identified yet. Thus these “genes” should be distinguished from genes functionally characterized by experiments. In addition, we re-curated all the data collected from the China Rice Data Center and the Oryzabase database as a double-check to make sure all the information in our database is correct. And we did find some error information in the two databases.

While the database of the Michigan State Univ is virtually abandoned without new updates since 2013, Oryzabase and RAP-DB have been releasing newly curated hundreds or thousands of data every year. The authors’ data that were “collected until 13 Feb 2014” (page 5) are very old and my feeling is that the researchers should need
Response: We updated the funRiceGenes database every two weeks since its initial construction in 2014. Since 2014, this database was updated using a Shiny application by tracking publications from PubMed and new records in the China Rice Data Center and Oryzabase databases. All updated records are available at https://funricegenes.github.io/news/, with the latest update performed on Sep 20th, 2017. We will keep updating of the funRiceGenes database in future.

First of all, the authors should mention that there are other efforts of extensive data curation of the rice genes. And, the authors should clearly state what are new and different from Oryzabase and RAP-DB in their database. Some clear example where functional descriptions were improved by the authors’ effort should be shown.
Response: Many thanks for your valuable suggestions. We discussed the features of funRiceGenes and difference of this database from Oryzabase and RAP-DB in the first paragraph of the Discussion section of the revised manuscript. We also clearly indicated the efforts of data curation from other database in the Background (page 3 line 11-16) and Result section (page 4 line 23-25).

Compared with Oryzabase and RAPdb, funRiceGenes has the following improvements:
1. The symbols of genes collected in funRiceGenes are much more accurate.
2. We separated member of gene families from functional characterized rice genes in the funRiceGenes database. A number of genes archived in Oryzabase and RAPdb are merely member of reported rice gene families identified by bioinformatics analysis rather than genes functionally characterized by experiments.
3. Some of the "genes" archived in Oryzabase are uncloned QTL rather than functionally characterized genes. The casual gene for the QTL has not been identified. We filtered these "genes" when we built the funRiceGenes database.
4. User-friendly query interface and tidy data for downloading are provided in the funRiceGenes database.

funRiceGenes also provides several additional functions:
1. Brief descriptions of the functions of collected genes and the supporting evidences are provided in the funRiceGenes database.
2. The interactions between different genes and the supporting evidences are provided in the funRiceGenes database.
3. Live update of the database every two weeks.

As I noted, the data of MSU are somewhat obsolete, but the authors’ analyses depended heavily upon such data. Isn't it necessary to use up-to-date information?
Response: We totally agree with you that it is necessary to use up-to-date information. Thus we use not only the data from MSU, but also up-to-date data from other database.

We used the information of orthologous groups of seven plants from MSU, which is absent from the database of RAPdb and Oryzabase. We also used the data of gene families from both MSU and Oryzabase to build our database, to make a more comprehensive collection.

In addition, both MSU and RAPdb provide annotation of the Nipponbare reference genome, which are extensively used by a wide range of researchers. For each gene in the funRiceGenes database, both the MSU and the RAPdb genomic locus are provided for convenience of researchers.

**Additional Information:**

| Question                                                        | Response   |
|-----------------------------------------------------------------|------------|
| Are you submitting this manuscript to a special series or article collection? | No         |
| **Experimental design and statistics**                         | Yes        |
| Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available |            |
Have you included all the information requested in your manuscript?

| Resources                                      | Yes |
|-----------------------------------------------|-----|
| A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible. |
| Have you included the information requested as detailed in our Minimum Standards Reporting Checklist? |

| Availability of data and materials            | Yes |
|-----------------------------------------------|-----|
| All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in publicly available repositories (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript. |
| Have you have met the above requirement as detailed in our Minimum Standards Reporting Checklist? |
funRiceGenes dataset for comprehensive understanding and application of rice functional genes

Wen Yao*, Guangwei Li, Yiming Yu, Yidan Ouyang*

National Key Laboratory of Crop Genetic Improvement, National Center of Plant Gene Research, Huazhong Agricultural University, Wuhan 430070, China

*Corresponding author: Wen Yao, ywhzau@gmail.com; Yidan Ouyang, diana1983941@mail.hzau.edu.cn

Abstract

Background: As a main staple food, rice is also a model plant for functional genomic studies of monocots. Decoding of every DNA element of the rice genome is essential for genetic improvement of rice to address the increasing food demands. The past 15 years have witnessed extraordinary advances in rice functional genomic studies. Systematic characterization and proper deposition of every rice gene are vital for both functional studies and crop genetic improvement.

Findings: We built a comprehensive and accurate dataset of ~2,800 functionally characterized rice genes and ~5,000 members of different gene families, by integrating data from available database and reviewing of every publication of rice functional genomic studies. The dataset accounts for 19.2% of the 39,045 annotated protein-coding rice genes, which provides the most exhaustive archive for investigating the functions of rice genes. We also constructed 214 gene interaction networks based on 1,841 connections between 1,310 genes. The largest network with 762 genes indicated that pleiotropic genes linked different biological pathways. Increasing degree of conservation of the flowering pathway was observed among
closer related plants, implying substantial value of rice genes for future dissection of flowering regulation in other crops. All data are deposited in the funRiceGenes database (https://funricegenes.github.io/). Functionality for advanced search and continuous updating of the database are provided by a Shiny application (http://funricegenes.ncpgr.cn/).

**Conclusions:** The funRiceGenes dataset would enable further exploring of the crosslink between gene functions and natural variations in rice, which can also facilitate breeding design to improve target agronomic traits of rice.

**Keywords:** *Oryza sativa* (rice), functional genomics, interaction network, genetic improvement

**Background**

Rice is a main staple food that feeds half of the world population. Improvement of the yield and the resistance to multiple biotic and abiotic stresses of rice is an essential strategy to cope with the increasing world population and the diminishing arable land. Decoding of the genetic reservoirs of rice is the basis for rice phenotype improvement.

Functional genomic studies in model organisms have made great contributions to the studies of a wide range of other species [1]. In the last decade, the functions of a number of rice genes were explored with the availability of the genome sequence of *Oryza sativa* L. ssp. *japonica* cv. Nipponbare [2]. Genes controlling important agronomic traits, including grain yield [3, 4], blast [5] and blight [6, 7] disease resistance, insect resistance [8], and abiotic stress resistance [9, 10], were functionally characterized. Some of these genes were utilized in rice breeding directly based on
marker-assisted strategy and CRISPR [11-13]. Moreover, the putative homologs of some rice genes were investigated in other crops such as wheat [14-17], barley [18] and maize [19]. As rice is an ideal model of the grass family, characterization of rice genes would greatly facilitate genomic studies and molecular breeding in other crops.

Abundant information on functionally characterized genes of Arabidopsis is archived in The Arabidopsis Information Resource (TAIR) [20], while a list of functionally characterized maize genes are integrated in the maizeGDB database (http://maizegdb.org/web_newgene.php?window=alltime), which greatly promoted the functional genomics studies in plants. Detailed information on Drosophila genes stored in the FlyBase database (http://flybase.org) is of great value to the studies in Drosophila and human [21]. The rice genome annotation project maintained by the Michigan State University of the USA [22] (http://rice.plantbiology.msu.edu/) and Rice Annotation Project Database (RAP-DB) [23] (http://rapdb.dna.affrc.go.jp/) greatly promoted the progress of rice functional genomics. Although a number of curated rice genes are collected in RAP-DB and Oryzabase (http://www.shigen.nig.ac.jp/rice/oryzabase/download/gene), not all the functionally characterized rice genes are properly deposited in existing databases. In the long term, the functions of all rice genes will be decoded [24]. As a result, comprehensive archive of all functionally characterized rice genes involved in diverse pathways with live updating is urgently in demand.

In this study, we constructed a comprehensive database of rice functional genes up to date, which includes ~2,800 cloned rice genes and ~5,000 members of different gene families. Interaction networks comprising 1,310 functionally characterized rice genes were constructed, which revealed the complex regulation and crosstalk of different biological pathways. We also developed a Shiny application allowing easily
addition of newly reported rice genes. As far as we are concerned, this is the most comprehensive and accurate database of functionally characterized rice genes with continuous updating.

Results

Collection of functionally characterized rice genes

A database (http://www.ricedata.cn/gene) maintained by the China Rice Data Center collects information on thousands of cloned rice genes in Chinese. Information on these genes was downloaded using in-house R scripts, including the gene symbol, the publications, the corresponding gene model in the Nipponbare reference genome, and a brief summary of the corresponding gene. The abstract, the author affiliation, and the full text of each publication were subsequently extracted from the PubMed database. Next, we manually curated the dataset based on the full text of each publication, and obtained 1,297 functionally characterized rice genes.

We further downloaded 29,982 publication records by querying the PubMed database with the keyword rice ((rice>Title] OR rice>Title/Abstract]), data until 13 Feb 2014). All the records were grouped by the published journal. After removing of the records involved in the China Rice Data Center and ones irrelevant to rice functional genomics, the full texts of the remaining publications were downloaded and reviewed, which identified additional 441 functionally characterized rice genes. Information on each gene, including the GenBank accession number and the corresponding gene model in the Nipponbare genome was extracted.

As an integrated rice science database, the Oryzabase (http://www.shigen.nig.ac.jp/rice/oryzabase/download/gene) also provides information on a portion of functionally characterized rice genes with manual curation.
We downloaded 10,140 records comprising a list of genes from this database
(http://www.shigen.nig.ac.jp/rice/oryazabase/gene/download;jsessionid=52FB01A7F53441CF4F823AA1ED71DE0?classtag=GENE_EN_LIST), and 5,531 records with
assigned Nipponbare genomic locus were retained. After removing of redundant
records in datasets obtained from the other two approaches, 469 functionally
characterized genes excluding members of gene families were retrieved. All the
information on the 469 genes was manually curated based on the review of research
publications. Finally, 2,207 functionally characterized rice genes were collected till 13
Feb 2014.

We further collected ~3,600 members of various gene families by integrating data
from the database of Rice Genome Annotation Project
(http://rice.plantbiology.msu.edu/annotation_community_families.shtml), the
Oryzabase database and research publications. All the data were deposited in the
funRiceGenes database (https://funricegenes.github.io/).

A Shiny application (http://funricegenes.ncpgr.cn/) was then developed to
facilitate utilization of this dataset, which also enabled easy addition of newly
reported genes to the database. New genes were added to this database using the
Shiny application, based on daily email alert of the searching results from the PubMed
database with the keyword rice (rice[Title] OR rice[Title/Abstract])
(https://funricegenes.github.io/help.pdf). For all the PubMed records in the email alert,
we identified ones on functionally characterized rice genes. We then went over the
full publication of each record and identified the gene symbol and gene model in the
reference genome. After inputting the gene symbol, the gene model in the reference
genome and the PubMed identifier, the Shiny application will fetch the corresponding
publication record from PubMed and extract key information automatically. We also
kept track of new records in the database of Oryzabase and China Rice Data Center, which were then added to our database using the Shiny application. Since 13 Feb 2014, funRiceGenes was updated every two weeks using the Shiny application. All the updated records are available at https://funricegenes.github.io/news/. Till 23 Feb 2017, ~2,800 functionally characterized genes and ~5,000 gene family members were archived in the funRiceGenes database, which accounted for 19.2% of the 39,045 annotated protein-coding rice genes (Supplementary Table S1, Supplementary Table S2) [22].

**Overview of the dataset regarding functionally characterized rice genes**

Rice functional genomic studies got rapid development after the public availability of the Nipponbare reference genome (Supplementary Figure S1). In total, about 3,553 publications with respect to ~2,800 functionally characterized genes were collected (Supplementary Table S3). These publications came from more than 215 journals, 31.0% of which were published in *The Plant Journal, Plant Physiology, Plant Molecular Biology, The Plant Cell, Molecular Plant, and New Phytologist* (Supplementary Table S3). Among all published papers, four words, rice, gene, protein, and expression, were observed with the highest frequencies in titles, while the words including rice, gene, expression, protein, plant, mutant, and stress were found with the highest frequencies in the abstract (Supplementary Figure S2, Supplementary Figure S3). More than 1,800 affiliations from all over the world contributed to rice functional genomic studies (Supplementary Table S4), and scientists from China, Japan, Korea, USA and India accounted for the majority of the progress (Supplementary Figure S4).

Genomic positions were determined for more than 98.1% of all functionally characterized rice genes based on the corresponding gene models of the Nipponbare...
reference genome (Supplementary Table S1, Figure 1). Twenty-five genes were absent from or showed substantial sequence divergence relative to the Nipponbare reference genome, and their genomic positions were determined based on the reference genome sequences of indica varieties Zhenshan 97 and Minghui 63 [25]. The remaining 24 genes were unable to be located in the genome, which was likely due to the sequence divergence between different rice germplasms.

A number of genes were investigated simultaneously by distinct research groups based on various rice accessions, mutants or phenotypic traits. As a result, 637 genes were assigned more than one symbol (Supplementary Table S1). In contrast, the same symbols were sometimes assigned to different genes due to the lack of communication (Supplementary Table S5).

Based on the concurrence of gene symbols and keywords regarding phenotype description or biological process in the same sentence of an abstract or a title in literatures, the functions of corresponding genes were summarized with manual curation. A total of 441 keywords were investigated, which generated 21,872 records for 1,952 genes (Supplementary Table S6). Among all 441 keywords, yield and grain yield were found in 311 records for 115 genes, while grain width, grain length, grain weight and grain size were detected in 139 records for 53 genes. Among all 77 genes retrieved with heading date or flowering time, 13 were also associated with yield or grain yield. Likewise, seven genes involved in iron utilization, phosphate uptake and sugar transporting were related to grain yield. We also found that 335 genes were involved in different stress signaling pathways, while 139 genes were related to rice diseases, including blast, bacterial blight, and sheath blight.

Progress in rice functional genomics benefited from the development of various technologies and the available of diverse genomic and genetic resources. We found
that homolog information was the most frequently used resource in rice functional genomics studies, and RT-PCR was the most commonly used technique to analyze gene expression level (Figure 2). Overexpression or RNAi were frequently used to disturb gene expression, which contributed to the dissection of the association between gene expression and phenotype variation. Creation of mutants using T-DNA and Tos17 insertions contributed significantly to rice gene cloning, while GWAS and CRISPR became new strategies to dissect the functions of rice genes in recent years [26, 27].

Interaction networks of functionally characterized rice genes

Physical and genetic interactions between different rice genes were frequently reported. However, a global view of the interaction networks for all functionally characterized rice genes remains to be elaborated. We constructed interaction networks of functionally characterized genes based on the concurrence of the symbols of two or more genes in the same sentence of an abstract or a title of research publications using in-house R script with manual curation. A sentence, in which two or more genes were observed, was regarded as an evidence supporting the connection between these genes. In total, 1,841 connections supported by 4,046 evidences were detected, which comprised 1,310 genes constituting 214 interaction networks (Supplementary Table S7).

The largest network was composed of 762 genes including ones associated with flowering, phosphate uptake and homeostasis, iron uptake, stress signaling, blight disease resistance, meiosis, BR and GA signaling, grain weight, and endosperm development (Figure 3). Genes related to the same trait were clustered together, indicating the trustworthy of this approach. The enormous size of this network was mainly caused by pleiotropic genes involved in different biological pathways. For
example, Ghd8 was responsible for grain number, plant height and heading date [28].

Ghd8 connected to genes controlling heading date including Ehd1 [29], Hd16 [30], and RFT1 [29], and genes controlling tillering including MOC1 [31], which was further connected with MIP1, a gene regulating tillering and plant height [32]. The other 213 interaction networks were made up of 548 rice genes, 88% of which contained only two or three genes (Supplementary Figure S5). The second largest network contained 14 genes involved in glutamine metabolism, including OsAMT1;3, GAD3, and GAT1 [33, 34]. Genes in terms of small RNA biogenesis including OsDCL3a, OsDCL1 and OsHEN1 were observed in a 10-gene network [35-37] (Supplementary Figure S5).

We further constructed an interaction network using 77 genes involved in flowering regulation (Figure 4). Based on the orthologous groups among seven plants provided by the Rice Genome Annotation Project (http://rice.plantbiology.msu.edu/annotation_pseudo_apk.shtml), we found that 40 of the 77 genes had orthologous genes in sorghum, maize, Brachypodium, Arabidopsis, poplar and grapevine, and orthologous genes were also identified for another 20 rice genes in sorghum, maize and Brachypodium (Figure 4, Supplementary Table S8).

Only seven genes, RFT1, Ehd4, Hd6, OsCO3, ROC4, Se14 and OsPIL15, were unique to rice. These results demonstrated the increasing degree of conservation of the flowering pathway among plants with closer phylogenetic relationships, implying substantial value of knowledge on functionally characterized rice genes to future dissection of flowering time regulation in other crops.

Discussion

In this study, we built a comprehensive and accurate database of functionally
characterized rice genes, funRiceGenes, which provides a valuable resource for rice functional genomic studies. funRiceGenes was constructed by integrating data from PubMed, Oryzabase, and China Rice Data Center, and was updated every two weeks using a Shiny application. For each gene in the funRiceGenes database, the gene symbol, the genomic locus in the reference genome and the published papers on this gene were identified. Compared with Textpresso for *Oryza sativa* (http://map.lab.nig.ac.jp:8095/textpresso/index.html), which is a comprehensive collection of literatures on rice, we further built the associations between genomic locus or symbol of genes and literatures [38]. Based on the literatures identified for each gene, we summarized the brief functions of each gene and constructed interaction networks for all genes. The evidences supporting the functions of all collected genes and the interaction networks are unique to the funRiceGenes database. In addition, user-friendly query interface and tidy data for downloading are provided in the funRiceGenes database.

Along with the sequence and phenotype data of thousands of rice accessions reported in recent years, the affluent information of rice genes in our database would enable further exploring of the crosslink between gene functions and natural variations. We found that a cloned rice gene *OsSGL* (LOC_Os02g04130, chr02:1799733-1800811), which regulated grain weight in rice, was ~70 kb away from a GWAS peak (chr02:1871732) in terms of grain weight [39, 40]. Likewise, another gene *OsPPKL3* (LOC_Os12g42310, chr12:26273157-26282197), which regulated grain length, is ~90 kb away from a GWAS peak (chr12:26182880) associated with grain length [41, 42]. The functions of *OsSGL* and *OsPPKL3* were characterized by transgenic studies and the natural variations of the two genes are yet to be dissected.
Our database is also beneficial to the interpretation of the large scale DNA, mRNA and other sequencing dataset in rice. Analyses of these data usually identify differentially expressed genes, gene co-expression networks, differentially methylated regions and ChIP-seq peaks, etc. The detailed information concerning several thousands of rice genes archived in this database would be helpful for illustration of these results [43]. Batch query functions are provided, allowing search of this database with multiple genes belonging to a pathway/biological process or defined gene set. Our work in rice would facilitate functional genomic studies of other crops including wheat, sorghum, and maize.

Pyramiding and editing of functionally characterized rice genes regulating important agronomic traits by molecular marker assisted selection and CRISPR are two promising approaches used to breed new rice varieties in recent years [44-46]. Thus, this database would play important roles in future rice breeding. For a specific agronomic trait, all related genes could be retrieved from this database conveniently for further manipulation (https://funricegenes.github.io/tags/#blight disease). For any of these genes, all relevant publications and a brief summary are available in this database (https://funricegenes.github.io/xa21/). The sequences of different alleles reported are also archived in this database. These resources would greatly facilitate breeding design to improve target agronomic traits by pyramiding of elite alleles or knocking out deleterious alleles. In addition, the effect of one gene might be enhanced or masked by other genes [47]. Thus, the gene interaction networks provided in this database could also be taken into account when making breeding designs.

Materials and Methods
Geocoding of author affiliations
The latitudes and longitudes of all the author affiliations were obtained using the application interface provided by the DATASCIENCETOOLKIT website (http://www.datasciencetoolkit.org/) with in-house R scripts. For author affiliations failed to be geocoded at high resolutions, we further used the Mapeasy website (http://www.mapeasy.com/adress-to-gps-coordinates.php) to find the accurate latitudes and longitudes. The R package ggmap was used to demonstrate the positions of all affiliations on the world map [48].

Extraction of information from PDF files

The occurrence of keywords, including map-based cloning, positional cloning, accession number, accession No., northern blot, northern analysis, northern hybridization and the regular expression "os[0-1][0-9]g[0-9]+.*", in PDF files were inspected utilizing the R tm [49] package.

Construction of interaction networks

The R package igraph [50] was used to build the interaction networks based on all the connection information between genes. The networks were then exported in data format suitable for Cytoscape, which was used to visualize the network [51].

Additional files

Additional file 1: Table S1: A comprehensive list of functionally characterized rice genes.

Additional file 2: Table S2: List of rice gene families.

Additional file 3: Table S3: Publications on functionally characterized rice genes.

Additional file 4: Table S4: The geocoding results of author affiliations.

Additional file 5: Table S5: Genes with different functions that were assigned the same symbols.
Additional file 6: Table S6: Concurrence of the gene symbols and the keywords regarding phenotype description or biological process in the same sentence of abstracts or titles of literatures.

Additional file 7: Table S7: Concurrence of the symbols of two or more genes in the same sentence of abstracts or titles of research publications.

Additional file 8: Table S8: Orthologs of genes regulating heading date in rice.

Additional file 9: Figure S1. Number of papers on rice functional genomic studies published in each year.

Additional file 10: Figure S2: Word cloud analysis of the titles of all the publications on rice functional genomic studies.

Additional file 11: Figure S3: Word cloud analysis of the abstracts of all the publications on rice functional genomic studies.

Additional file 12: Figure S4: Global distribution of affiliations contributed to rice functional genomics studies. All the affiliations are marked on the world map as blue circles based on their longitudes and latitudes. The size of the circle represents the number of publications conducted by each affiliation. Data after 18 Jun 2015 are not shown.

Additional file 13: Figure S5: Gene interaction networks constructed based on the concurrence of two or more genes in the same sentence of abstracts or titles of publications. Each white node represents a gene while each green edge indicates a connection between two genes.

Conflicts of interest

The authors declare that they have no competing interests.
Authors’ Contributions

W.Y. conceived and designed the experiments; W.Y., G.L., Y.Y. and Y.O. analyzed the data; W.Y. and Y.O. wrote the paper.

Acknowledgements

Not applicable.

Funding

This research was supported by grants from the National Key Research and Development Program of China (2016YFD0100903), the National Natural Science Foundation of China (31771873 and 31371599), and the Outstanding Young Talents Program.

Reference

1. Fontana L and Partridge L. Promoting health and longevity through diet: from model organisms to humans. Cell. 2015;1:106-18.

2. Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, et al. A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). Science. 2002;5565:92-100.

3. Wang J, Yu H, Xiong G, Lu Z, Jiao Y, Meng X, et al. Tissue-specific ubiquitination by IPA1 INTERACTING PROTEIN 1 modulates IPA1 protein levels to regulate plant architecture in rice. The Plant Cell. 2017;4:697-707.
4. Fan C, Xing Y, Mao H, Lu T, Han B, Xu C, et al. *GS3*, a major QTL for grain length and weight and minor QTL for grain width and thickness in rice, encodes a putative transmembrane protein. Theoretical and Applied Genetics. 2006;6:1164-71.

5. Deng Y, Zhai K, Xie Z, Yang D, Zhu X, Liu J, et al. Epigenetic regulation of antagonistic receptors confers rice blast resistance with yield balance. Science. 2017;6328:962-5.

6. Gu K, Yang B, Tian D, Wu L, Wang D, Sreekala C, et al. *R* gene expression induced by a type-III effector triggers disease resistance in rice. Nature. 2005;7045:1122-5.

7. Hu K, Cao J, Zhang J, Xia F, Ke Y, Zhang H, et al. Improvement of multiple agronomic traits by a disease resistance gene via cell wall reinforcement. Nature Plants. 2017:17009.

8. Zhao Y, Huang J, Wang Z, Jing S, Wang Y, Ouyang Y, et al. Allelic diversity in an NLR gene *BPH9* enables rice to combat planthopper variation. Proceedings of the National Academy of Sciences. 2016;45:12850-5.

9. Xu K, Xu X, Fukao T, Canlas P, Maghirang-Rodriguez R, Heuer S, et al. *Sub1A* is an ethylene-response-factor-like gene that confers submergence tolerance to rice. Nature. 2006;7103:705-8.

10. Tan J, Tan Z, Wu F, Sheng P, Heng Y, Wang X, et al. A novel chloroplast-localized pentatricopeptide repeat protein involved in splicing affects chloroplast development and abiotic stress response in rice. Molecular
1. Jiang H, Feng Y, Bao L, Li X, Gao G, Zhang Q, et al. Improving blast resistance of Jin 23B and its hybrid rice by marker-assisted gene pyramiding. Molecular Breeding. 2012;4:1679-88.

2. Wang S, Wu K, Yuan Q, Liu X, Liu Z, Lin X, et al. Control of grain size, shape and quality by OsSPL16 in rice. Nature Genetics. 2012;8:950-4.

3. Shan Q, Zhang Y, Chen K, Zhang K and Gao C. Creation of fragrant rice by targeted knockout of the OsBADH2 gene using TALEN technology. Plant Biotechnology Journal. 2015;6:791-800.

4. Bednarek J, Boulaflous A, Girousse C, Ravel C, Tassy C, Barret P, et al. Down-regulation of the TaGW2 gene by RNA interference results in decreased grain size and weight in wheat. Journal of Experimental Botany. 2012;16:5945-55.

5. Liu Y-N, Xia X-C and He Z-H. Characterization of Dense and Erect Panicle 1 gene (TaDep1) located on common wheat group 5 chromosomes and development of allele-specific markers. Acta Agronomica Sinica. 2013;4:589-98.

6. Nemoto Y, Kisaka M, Fuse T, Yano M and Ogihara Y. Characterization and functional analysis of three wheat genes with homology to the CONSTANS flowering time gene in transgenic rice. The Plant Journal. 2003;1:82-93.

7. Nakamura S, Abe F, Kawahigashi H, Nakazono K, Tagiri A, Matsumoto T, et al. A wheat homolog of MOTHER OF FT AND TFL1 acts in the regulation of...
1. Germination. The Plant Cell. 2011;9:3215-29.

2. Comadran J, Kilian B, Russell J, Ramsay L, Stein N, Ganal M, et al. Natural variation in a homolog of *Antirrhinum CENTRORADIALIS* contributed to spring growth habit and environmental adaptation in cultivated barley. Nature Genetics. 2012;12:1388-92.

3. Yang Q, Li Z, Li W, Ku L, Wang C, Ye J, et al. CACTA-like transposable element in *ZmCCT* attenuated photoperiod sensitivity and accelerated the postdomestication spread of maize. Proceedings of the National Academy of Sciences. 2013;42:16969-74.

4. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, et al. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic acids research. 2012;Database issue:D1202-10.

5. Gramates LS, Marygold SJ, Santos Gd, Urbano J-M, Antonazzo G, Matthews BB, et al. FlyBase at 25: looking to the future. Nucleic acids research. 2017;D1:D663-D71.

6. Kawahara Y, de la Bastide M, Hamilton J, Kanamori H, McCombie W, Ouyang S, et al. Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. Rice. 2013;1:1-10.

7. Sakai H, Lee SS, Tanaka T, Numa H, Kim J, Kawahara Y, et al. Rice Annotation Project Database (RAP-DB): an integrative and interactive database for rice genomics. Plant and Cell Physiology. 2013;2:e6.
24. Zhang Q, Li J, Xue Y, Han B and Deng XW. Rice 2020: A call for an international coordinated effort in rice functional genomics. Molecular plant. 2008;5:715-9.

25. Zhang J, Chen L-L, Xing F, Kudrna DA, Yao W, Copetti D, et al. Extensive sequence divergence between the reference genomes of two elite *indica* rice varieties Zhenshan 97 and Minghui 63. Proceedings of the National Academy of Sciences. 2016;35:E5163-71.

26. Si L, Chen J, Huang X, Gong H, Luo J, Hou Q, et al. *OsSPL13* controls grain size in cultivated rice. Nature Genetics. 2016;4:447-56.

27. Yamauchi T, Yoshioka M, Fukazawa A, Mori H, Nishizawa NK, Tsutsumi N, et al. An NADPH oxidase RBOH functions in rice roots during Lysigenous Aerenchyma formation under oxygen-deficient conditions. The Plant Cell. 2017;4:775-90.

28. Yan WH, Wang P, ChenHX, Zhou HJ, Li QP, Wang CR, et al. A major QTL, *Ghd8*, plays pleiotropic roles in regulating grain productivity, plant height, and heading date in rice. Molecular plant. 2011;2:319-30.

29. Dai X, Ding Y, Tan L, Fu Y, Liu F, Zhu Z, et al. *LHD1*, an allele of *DTH8/Ghd8*, controls late heading date in common wild rice (*Oryza rufipogon*). Journal of Integrative Plant Biology. 2012;10:790-9.

30. Hori K, Ogiso-Tanaka E, Matsubara K, Yamanouchi U, Ebana K and Yano M. *Hd16*, a gene for casein kinase I, is involved in the control of rice flowering time by modulating the day-length response. The Plant Journal. 2013;1:36-46.
31. Li X, Qian Q, Fu Z, Wang Y, Xiong G, Zeng D, et al. Control of tillering in rice. Nature. 2003;6932:618-21.

32. Sun F, Zhang W, Xiong G, Yan M, Qian Q, Li J, et al. Identification and functional analysis of the MOC1 interacting protein 1. Journal of Genetics and Genomics. 2010;1:69-77.

33. Yang S, Hao D, Cong Y, Jin M and Su Y. The rice OsAMT1;1 is a proton-independent feedback regulated ammonium transporter. Plant Cell Reports. 2015;2:321-30.

34. El-kereamy A, Bi Y-M, Ranathunge K, Beatty PH, Good AG and Rothstein SJ. The rice R2R3-MYB transcription factor OsMYB55 is involved in the tolerance to high temperature and modulates amino acid metabolism. PLOS ONE. 2012;12:e52030.

35. Wei L, Gu L, Song X, Cui X, Lu Z, Zhou M, et al. Dicer-like 3 produces transposable element-associated 24-nt siRNAs that control agricultural traits in rice. Proceedings of the National Academy of Sciences. 2014;10:3877-82.

36. Liu B, Li P, Li X, Liu C, Cao S, Chu C, et al. Loss of function of OsDCL1 affects microRNA accumulation and causes developmental defects in rice. Plant Physiology. 2005;1:296-305.

37. Abe M, Yoshikawa T, Nosaka M, Sakakibara H, Sato Y, Nagato Y, et al. WAVY LEAF1, an ortholog of Arabidopsis HEN1, regulates shoot development by maintaining microRNA and trans-acting small interfering RNA accumulation in rice. Plant Physiology. 2010;3:1335-46.
38. Müller H-M, Kenny EE and Sternberg PW. Textpresso: an ontology-based information retrieval and extraction system for biological literature. PLOS Biology. 2004;11:e309.

39. Wang M, Lu X, Xu G, Yin X, Cui Y, Huang L, et al. OsSGL, a novel pleiotropic stress-related gene enhances grain length and yield in rice. Scientific Reports. 2016;38157.

40. Yang W, Guo Z, Huang C, Duan L, Chen G, Jiang N, et al. Combining high-throughput phenotyping and genome-wide association studies to reveal natural genetic variation in rice. Nature Communications. 2014;5087.

41. Zhang X, Wang J, Huang J, Lan H, Wang C, Yin C, et al. Rare allele of OsPPKL1 associated with grain length causes extra-large grain and a significant yield increase in rice. Proceedings of the National Academy of Sciences. 2012;52:21534-9.

42. McCouch SR, Wright MH, Tung C-W, Maron LG, McNally KL, Fitzgerald M, et al. Open access resources for genome-wide association mapping in rice. Nature Communications. 2016;10532.

43. Zong W, Tang N, Yang J, Peng L, Ma S, Xu Y, et al. Feedback regulation of ABA signaling and biosynthesis by a bZIP transcription factor targets drought resistance related genes. Plant Physiology. 2016;4:2810-25.

44. Collard BC and Mackill DJ. Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Philosophical Transactions of the Royal Society B: Biological Sciences. 2008;1491:557-72.
45. Zeng D, Tian Z, Rao Y, Dong G, Yang Y, Huang L, et al. Rational design of high-yield and superior-quality rice. Nature Plants. 2017:17031.

46. Zhou H, He M, Li J, Chen L, Huang Z, Zheng S, et al. Development of commercial thermo-sensitive genic male sterile rice accelerates hybrid rice breeding using the CRISPR/Cas9-mediated TMS5 editing system. Scientific Reports. 2016:37395.

47. Gao X, Zhang X, Lan H, Huang J, Wang J and Zhang H. The additive effects of GS3 and qGL3 on rice grain length regulation revealed by genetic and transcriptome comparisons. BMC Plant Biology. 2015;1:156.

48. Kahle D and Wickham H. ggmap: Spatial visualization with ggplot2. The R Journal. 2013;1:144-61.

49. Meyer D, Hornik K and Feinerer I. Text mining infrastructure in R. Journal of Statistical Software. 2008;5:1-54.

50. Csardi G and Nepusz T. The igraph software package for complex network research. InterJournal. 2006:1695.

51. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Research. 2003;11:2498-504.

---

**Figure legends**

**Figure 1.** Chromosome distribution of representative functionally characterized rice genes.
The chromosomes are represented as vertical rectangles and each horizontal line denotes the position of a functionally characterized rice gene. Symbols of all genes are labeled. A total of 930 representative genes are shown.

Figure 2. Usage of various biotechniques in rice functional genomics studies.

The y-axis indicates the number of publications using a specific biotechnique. Data after 18 Jun 2015 are not shown.

Figure 3. The gene interaction network comprising 762 genes.

Each white node represents a functionally characterized rice gene and gene symbols are marked beside the node. Each green edge indicates a connection between two genes. Genes involved in the same biological pathways are indicated.

Figure 4. Interaction network of genes regulating flowering in rice and the orthologs of these genes in other plants.

Each node represents a functionally characterized rice gene. Each edge indicates a connection between two genes. Genes with different number of orthologs are indicated with different color and shape. “Rice + (Maize | Poplar)” indicates “Rice and Maize” or “Rice and Poplar”. Detailed information is shown in Supplementary Table S8.
Rice

Rice + (Maize | Poplar | Brachypodium)

Rice + (Sorghum + Maize) | (Sorghum + Brachypodium) | (Maize + Brachypodium)

Rice + Maize + Sorghum + Brachypodium + (Arabidopsis | Poplar | Grapevine)

Rice + Maize + Brachypodium + Poplar + Grapevine + Arabidopsis

Rice + Maize + Sorghum+ Brachypodium + Poplar + Grapevine + Arabidopsis
Click here to access/download

**Supplementary Material**

Supplementary_Table_S1.xlsx
Click here to access/download

**Supplementary Material**

Supplementary_Table_S2.xlsx
Click here to access/download
Supplementary Material
Supplementary_Table_S3.xlsx
Click here to access/download

**Supplementary Material**

Supplementary_Table_S4.xlsx
Click here to access/download

**Supplementary Material**

Supplementary_Table_S5.xlsx
Click here to access/download
Supplementary Material
Supplementary_Table_S6.xlsx
Click here to access/download

Supplementary Material
Supplementary_Table_S8.xlsx
Click here to access/download
**Supplementary Material**
Supplementary_Figure_S1.tif
Click here to access/download
Supplementary Material
Supplementary_Figure_S2.tif
Click here to access/download
Supplementary Material
Supplementary_Figure_S3.tif
Click here to access/download
Supplementary Material
Response to Decision Letter.docx
