

# SCIENTIFIC DATA

## OPEN Data Descriptor: Novel sequences, structural variations and gene presence variations of Asian cultivated rice

Received: 22 June 2017  
 Accepted: 15 September 2017  
 Published: 2 May 2018

Zhiqiang Hu<sup>1,2,\*</sup>, Wensheng Wang<sup>1,\*</sup>, Zhichao Wu<sup>1,\*</sup>, Chen Sun<sup>2</sup>, Min Li<sup>1,5</sup>, Jinyuan Lu<sup>2</sup>, Binying Fu<sup>1,4</sup>, Jianxin Shi<sup>2</sup>, Jianlong Xu<sup>1,4</sup>, Jue Ruan<sup>3</sup>, Chaochun Wei<sup>2</sup> & Zhikang Li<sup>1,4</sup>

Genomic diversity within a species genome is the genetic basis of its phenotypic diversity essential for its adaptation to environments. The big picture of the total genetic diversity within Asian cultivated rice has been uncovered since the sequencing of 3,000 rice genomes, including the SNP data publicly available in the SNP-Seek database. Here we report other aspects of the genetic diversity, including rice sequences assembled from over 3,000 accessions but absent in the Nipponbare reference genome, structural variations (SVs) and gene presence/absence variations (PAVs) in 453 accessions with sequencing depth over 20x. Using either SVs or gene PAVs, we were able to reconstruct the population structure of *O. sativa*, which was consistent with previous result based on SNPs. Moreover, we demonstrated the usefulness of the new data sets by successfully detecting the strong association of the "Green Revolution gene", *sd1*, with plant height. Our data provide a more comprehensive view of the genetic diversity within rice, as well as additional genomic resources for research in rice breeding and plant biology.

Design Type(s)	sequence analysis objective
Measurement Type(s)	genetic sequence variation analysis
Technology Type(s)	bioinformatics analysis
Factor Type(s)	
Sample Characteristic(s)	Oryza sativa

<sup>1</sup>Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing 100081, China. <sup>2</sup>School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China. <sup>3</sup>Agricultural Genomics Institute, Chinese Academy of Agricultural Sciences, Shenzhen 518120, China. <sup>4</sup>Shenzhen Institute for Innovative Breeding, Chinese Academy of Agricultural Sciences, Shenzhen 518120, China. <sup>5</sup>Anhui Agricultural University, Hefei 230036, China. \*These authors contributed equally to this work. Correspondence and requests for materials should be addressed to C.W. (email: ccwei@sjtu.edu.cn) or to Z.L. (email: lizhikang@caas.cn).

## Background & Summary

With the continued reduction of croplands, we are facing a great challenge of feeding the fast growing world population. Asian cultivated rice (*Oryza sativa* L.) is grown worldwide and is the staple food for half the world population. The rich resources of the global rice germplasms are expected to provide a sustainable solution in future rice improvement. However, the total genomic diversity within the primary gene pool of *O. sativa* remains poorly explored. In order to reveal the comprehensive genomic diversity, ‘The 3,000 rice genomes project (3k RG)’<sup>1,2</sup> was initiated, in which 3,024 accessions representing the most diversity of the cultivated rice population were selected from 89 countries. The 3,024 rice genomes were sequenced using HiSeq 2000 platform with 500 bp DNA fragment libraries and a total of 17 trillion bases were generated. The sequencing data of the 3,024 rice accessions were freely available at GigaDB<sup>1</sup> (Data Citation 1), NCBI SRA (Data Citation 2), Amazon cloud (<https://aws.amazon.com/cn/public-data-sets/3000-rice-genome/>) and Aliyun cloud (<http://ricecloud.org/>). We have previously reported the single nucleotide polymorphisms (SNP) data derived from these sequencing data<sup>3</sup>, which is very useful for QTL mapping by genome-wide association studies.

Structural variations (SVs) and gene presence/absence variations (PAVs) represent additional dimensions of the total genetic diversity within a species and remain largely unknown in almost all eukaryotes. In this descriptor, we reported the SV data and gene PAV data of *O. sativa*, together with the novel sequences absent in the widely used Nipponbare reference genome IRGSP-1.0, as key results of the in-depth analyses of the sequencing data of the 3k RG<sup>4</sup>.

In this data descriptor, we present the rice SV data obtained by calling against the Nipponbare reference genome using novoBreak<sup>5</sup> because it had the lowest false positive rate when compared with results from several tools such as BreakDancer<sup>6</sup> and Delly<sup>7</sup>. This SV data set contains a total number of 93,683 SVs (deletions, inversions and duplications of >100 bps and < 1 Mbps, and translocations), or an average of 12,178 SVs per genome detected in the 453 high-quality genomes with sequencing depths  $\geq 20\times$  and mapping depths  $\geq 15\times$ . We also present the PAV data sets of 48,098 full-length protein-coding genes (35,633 Nipponbare reference genes and 12,465 novel genes) and 23,876 gene families in the 453 rice accessions, which were obtained from a ‘map-to-pan’ pipeline<sup>8</sup>. These data not only provide a more comprehensive understanding of the genomic diversity within *O. sativa*, but also provide additional genomic markers for genome-wide association studies of rice.

## Methods

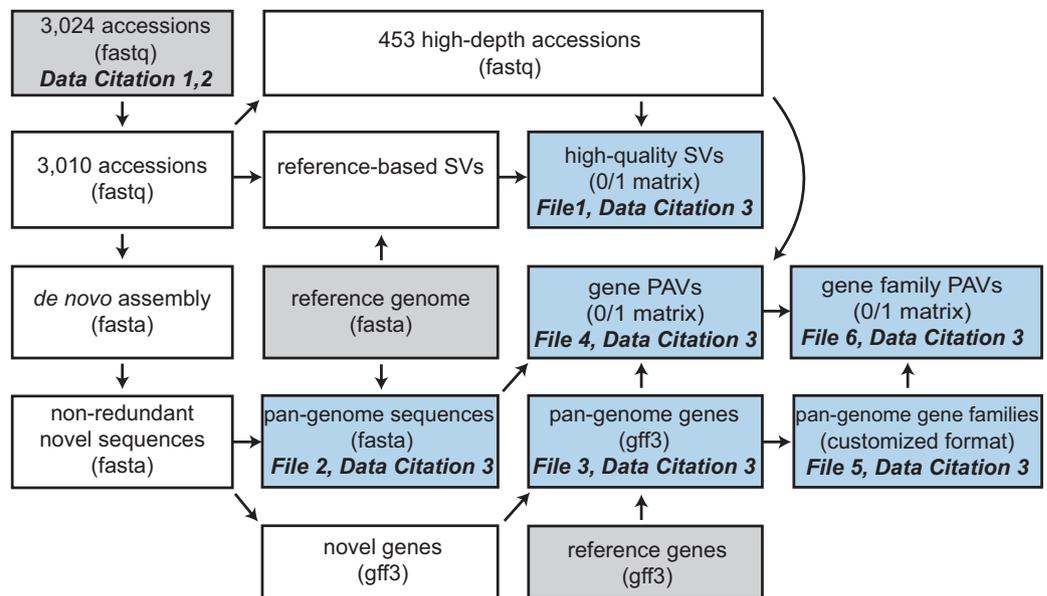
In this descriptor, we provided the key outputs of the in-depth analyses in the 3k RG (Fig. 1), including (1) SVs in the 453 high sequencing-depth accessions (File 1, Data Citation 3); (2) the pan-genome sequences of *O. sativa* (File 2, Data Citation 3) and its protein-coding gene annotation (File 3, Data Citation 3); (3) a matrix of gene PAVs of the 453 accessions (File 4, Data Citation 3); and (4) gene family annotation (File 5, Data Citation 3) and a matrix of gene family PAVs of the 453 accessions (File 6, Data Citation 3). The methods described below are expanded versions of descriptions in our related work<sup>4</sup>.

The sequencing data of 3,024 accessions of 3k RG were available from various sources (Data Citation 1 and Data Citation 2). To begin with, 14 accessions were removed from the 3,024 sequenced accessions. Accessions CX400, CX401, CX402, IRIS\_313-11415 and IRIS\_313-10729 belong to African cultivated rice (*Oryza glaberrima* L.). Accessions IRIS\_313-8502, IRIS\_313-9233, IRIS\_313-8444, IRIS\_313-10057, IRIS\_313-9184, B014, IRIS\_313-9404 were removed due to significant contamination, and accession B101 was removed due to a very small estimated genome size. Accession IRIS\_313-8921 was removed due to its extremely low sequencing depth ( $0.3\times$ ).

The sequencing depths of the remaining 3,010 genomes ranged from 4.2x to 63.8x. Clean reads of each accession were mapped to the Nipponbare reference genome (IRGSP-1.0) using ‘bwa mem’ with default parameters<sup>9</sup>. The mapping depth was calculated as total mapped bases divided by the size of IRGSP-1.0 genome. Only the 453 rice accessions with sequencing depth  $\geq 20$  and mapping rate  $\geq 15$  were selected for reporting final SVs and PAVs.

## Structural variations

novoBreak<sup>5</sup> (<https://sourceforge.net/projects/novobreak/?source=navbar>) was used for SV calling against the Nipponbare genome. The comparison of performances of three software, novoBreak<sup>5</sup>, BreakDancer<sup>6</sup> and Delly<sup>7</sup> are available in our related work<sup>4</sup>. novoBreak was first run on each of the 3,010 accessions independently with default parameters. We detected deletions, inversions and duplications with sizes between 100 bps and 1Mbps, and translocations (interchromosomal breakpoints). SVs which could be inferred by no less than 3 kmer/reads were further selected with the following filter conditions: 1) more than 4 supporting split reads; or 2) no less than 3 discordant read pairs for both breakpoint. To identify SVs in different samples, all SVs passed the filter condition in the 453 accessions were pooled together to justify coordinated variants. Any two adjacent SVs were identified as a single SV if their start and end positions varied no more than 1 kb, and their overlapping region is over half of the total size. The presence/absence matrix of SVs in each accession was built with these criteria. To minimize the false positive rates, SVs detected in less than 6 accessions or more than 80% of the 3,010 accessions were removed. The SVs in the 453 high-depth accessions were then extracted as a more confident subset (File 1, Data Citation 3).

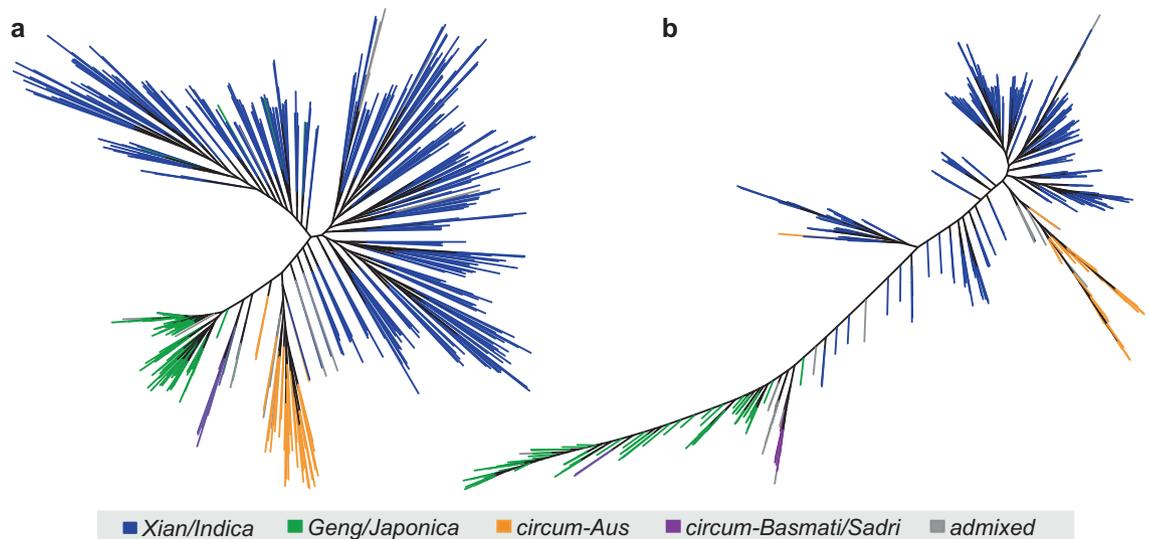


**Figure 1.** A diagram of data generation and data dependencies. Raw inputs are displayed in grey boxes and important outputs are shown in blue boxes. Data formats are shown in the brackets. The analysis procedure includes: 1) removal of 14 accessions; 2) selection of 453 high-depth accessions; 3) SV calling for 3,010 accessions against the Nipponbare reference genome and extraction of high-quality SVs for the 453 accessions; 4) construction of the pan-genome sequences by combining the Nipponbare reference genome and non-redundant novel sequences; and 5) determination of presence/absence variation of genes and gene families in the 453 accessions with the “map-to-pan” strategy. Detailed methods are described in the Methods section.

### Novel sequences

Novel sequences absent in the reference genome were derived from the *de novo* assemblies of the 3,010 rice accessions. To achieve high-quality assemblies, we utilized a method with iterative use of SOAPdenovo r240<sup>10</sup> to select the best Kmer parameter for each accession. In detail, we first determined an initial Kmer ( $K_{init}$ ) for rice accession  $R$  based on a linear model  $K_{init} = \text{ROUND}(a * \text{Dep}_R + b)$ , where  $\text{Dep}_R$  is the sequencing depth of  $R$ , and  $a$  and  $b$  are the parameters of the linear model. This linear model was trained on 50 randomly selected rice accessions and the initial Kmer was set as ‘ $K = 2 * \text{int}(0.38 * (\text{sequencing depth}) + 10) + 1$ ’ ( $R^2 = 0.68$ ). After  $K_{init}$  was determined, we ran SOAPdenovo 3 times with stepped Kmers ( $K_{mer_{low}} = K_{mer_{init}} - 2$ ,  $K_{mer_{mid}} = K_{mer_{init}}$  and  $K_{mer_{high}} = K_{mer_{init}} + 2$ ) using command line “SOAPdenovo-63mer(or SOAPdenovo-127mer) all -s configure\_file (average insertion length set as 460 in the configure file) -o output\_directory -K Kmer -R -F -p 8”; Gapcloser tool within SOAPdenovo package was then used on the SOAPdenovo raw scaffolds with command “GapCloser -b configure\_file (the same as the configure file used in running SOAPdenovo) -a soap\_raw.scaf -o output\_gc.scaf -t 8”; Gap closed scaffolds were then broken down to contigs by cutting scaffolds at sites with >10 consecutive Ns. The  $N50$  values of the gap-closed contigs were calculated for  $K_{mer_{low}}$ ,  $K_{mer_{mid}}$  and  $K_{mer_{high}}$  respectively. If  $N50$  of  $K_{mer_{mid}}$  was the highest, assembly at  $K_{mer_{mid}}$  was selected as the final result. If  $N50$  of  $K_{mer_{high}}$  was the highest, an additional run with  $K_{mer} = K_{mer_{high}} + 2$  was needed and we compared  $N50$  of  $K_{mer_{mid}}$ ,  $K_{mer_{high}}$  and  $K_{mer_{high}} + 2$ ; this iteration continued until  $N50$  of the median Kmer became the highest. If  $N50$  of  $K_{mer_{low}}$  was the highest, Kmer went down and similar iteration was carried out to determine the best Kmer. In order to control the false positives, though GapCloser was involved in the assembly pipeline and provided larger  $N50$ , only raw contig assemblies from SOAPdenovo were used for detection of novel sequences.

Assemblies were assessed by comparing to the Nipponbare reference genome using QUAST v2.3<sup>11</sup>, and those unaligned contigs were then selected. All the unaligned contigs with length >500 bp from each rice accession were merged into a single sequence set. Next, CD-HIT v4.6.123<sup>12</sup> was used to remove redundant sequences at an identity cut-off of 90% with command “cd-hit-est -i input.fa -o output.fa -c 0.9 -T 16 -M 50000”. This process was carried out for 3 times to ensure redundancies were removed. Next, various contaminants including archaea, bacteria, virus, fungi and metazoa were removed. In practice, the non-redundant sequences were aligned to the NT database (07/26/2014) with NCBI-blast v2.2.28+ with command “blastn -db nt.fa -task megablast -query input.fa -num\_threads 16 -evaluate 1e-5 -out o -outfmt '\%6 qseqid sseqid qlen length qstart qend sstart send pident evaluel' -best\_hit\_overhang 0.25 -perc\_identity 0.5 -max\_target\_seqs 10”. Contigs whose best alignments (E-value as 1<sup>st</sup> keyword and identity as 2<sup>nd</sup> keyword) were not in Viridiplantae (annotation information



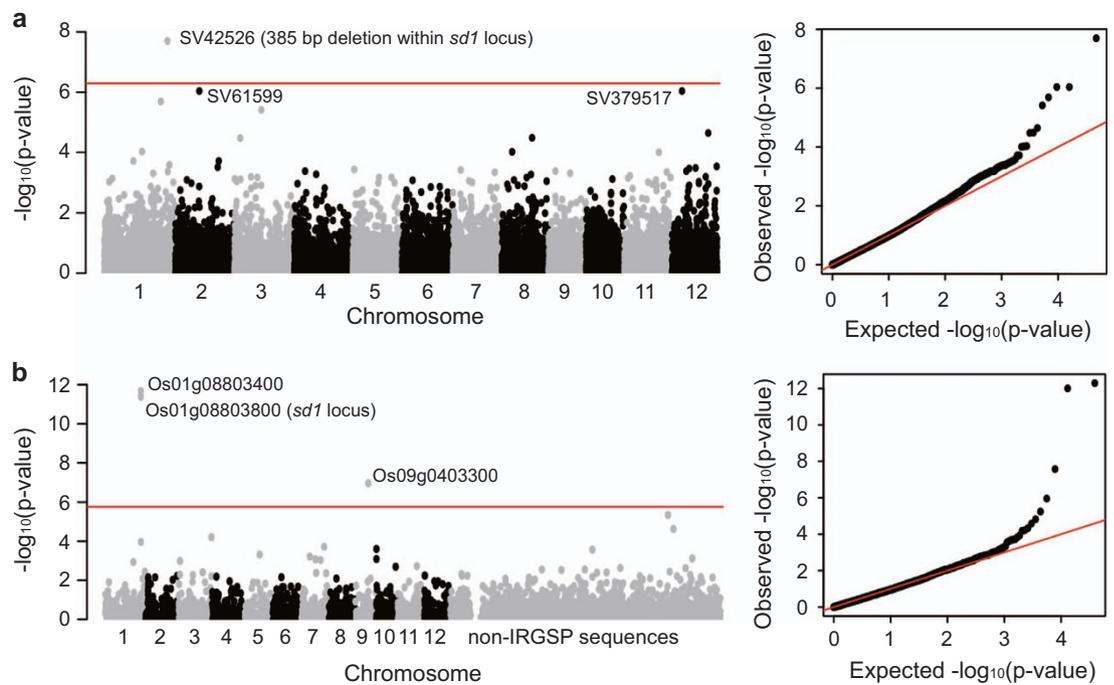
**Figure 2.** Phylogenetic analyses of the 453 rice accessions based on SVs and gene PAVs. (a) the phylogenetic tree constructed from SVs of the 453 accessions; (b) the phylogenetic tree constructed from gene PAVs of the 453 accessions. Different Colors represent the grouping information derived from SNP analyses. Both SVs and PAVs clearly classified the 453 rice accessions into *Xian/Indica*, *Geng/Japonica*, *circum-Aus* and *circum-Basmati/Sadri* groups.

downloaded from NCBI taxonomy database) were considered as contaminants and were filtered out. The remaining contigs formed the non-redundant novel sequence dataset (identity < 0.9 in comparison to IRGSP genome and identity < 0.9 among the novel sequences themselves). In order to be browsed in a genome browser, the novel contigs were concatenated with 100 consecutive Ns as the delimiters. The reference sequences and the novel sequences were combined to form the rice pan-genome sequences (File 2, Data Citation 3).

### Gene presence/absence variations

Both the Nipponbare reference genes and non-Nipponbare novel genes were included in the PAV analyses. The gene annotation of IRGSP-1.0 genome was downloaded from the Rice Annotation Project (RAP)<sup>13</sup>. In order to be consistent with gene annotation for non-IRGSP sequences, also to simplify the subsequent analyses, only those transcripts with the longest open reading frame (ORF) were selected as representatives. Protein-coding genes on the novel sequences were predicted using MAKER2<sup>14</sup> (a gene prediction pipeline combining *ab initio* predictions, expression evidence and protein homologies). In detail, low-complexity repeats were first masked. Two *ab initio* predictors, snap<sup>15</sup> and AUGUSTUS<sup>16</sup>, were utilized to predict gene models with their default parameters for rice. All rice ESTs were downloaded from Genbank (12/15/2014) and were aligned to the novel sequences with NCBI-blastn. All rice proteins were downloaded from NCBI (12/15/2014) and were aligned to the novel sequences with NCBI-blastx. To get more informative alignments, Exonerate<sup>17</sup> was used to realign each sequence identified by blast around splice sites. EVIDENCEModeler<sup>18</sup> was used to combine and refine the *ab initio* predictions with RNA and protein evidences. Incomplete gene models were further removed. Since we used a method based on representative sequences to build the novel sequences, there might be local similarities between novel sequences and the reference genome, as well as among novel sequences themselves. Therefore, there might also be genes predicted in the novel sequences similar to genes in the reference genome and also novel genes similar to each other. In order to remove these genes, we clustered all genes at global protein identity of 95%, and remove those novel genes who are not the representative of the group. The non-redundant novel genes together with the reference genes formed the rice pan-genome (File 3, Data Citation 3).

We utilized a “map-to-pan” strategy<sup>8</sup> to determine gene presence/absence based on the observation that more than 98% of the genome can be covered by short read mapping at sequencing depth  $\geq 20$ . The presence/absence of each gene was determined by the read coverage of the gene body and its coding region. In practice, we first mapped reads of each accession to the pan-genome sequences with “bwa mem” and a gene was considered as present if it had 1) coding coverage > 0.95 (covered bases in the ORF divided by the ORF length) and 2) gene body coverage > 0.85 (the number of covered bases in the gene body divided by the gene length). The presence/absence information of each gene from the 453 high sequencing-depth accessions were then merged together as a 0/1 matrix (File 4, Data Citation 3).



**Figure 3. Genome-wide association studies of plant height based on both SVs and gene PAVs.** (a) genome-wide association study based on SVs including deletions, duplications and inversions; (b) genome-wide association study based on PAVs of distributed genes. The left panels show the Manhattan plots and the right panels show the quantile-quantile plots. The thresholds of significant  $P$ -values were  $5.1e-7$  and  $4.2e-7$ , respectively (with bonferroni correction, significant level at 0.01).

### Gene family presence/absence variations

Similar to a recent pan-genome study on soybean<sup>19</sup>, the genes were clustered to gene families with OrthoMCL v2.0.9<sup>20</sup> at protein similarity cut-off of 0.5. In detail, all coding sequences of genes on the pan-genome sequences were first extracted and then translated into protein sequences. All protein sequences were compared by using all-by-all blastp (Evalue set as  $1e-5$ ). OrthoMCL was used to process the blastp output and cluster genes to gene families (File 5, Data Citation 3).

Gene family presence/absence (File 6, Data Citation 3) was further determined with a straightforward rule based on gene presence/absence. If at least one gene of a gene family presented in the rice accession, this gene family was considered to be present in this rice accession.

### Code availability

All codes for detection of the novel sequences and determination of gene PAVs were implemented in EUPAN package<sup>8</sup> available at <http://cgm.sjtu.edu.cn/eupan/>. Software and their used versions were described as above.

### Data Records

The raw data of ‘The 3,000 rice genome project’ (3k RG) were deposited in GigaDB (Data Citation 1) and NCBI’s Short Read Archive (SRA) under accession number PRJEB6180 (Data Citation 2). Data records presented in this descriptor are available online from Figshare. The presence/absence matrix of SVs including deletion, duplication, inversion and translocations (File 1, Data Citation 3) contains both SV annotations and their presence/absence in 453 high-depth rice accessions. The pan-genome sequences (File 2, Data Citation 3) include both the IRGSP-1.0 sequences and novel sequences. The protein-coding gene annotation (File 3, Data Citation 3) is in GFF3 format. The gene family annotation file (File 5, Data Citation 3) is formatted as a gene family id followed by all its gene members in a line. The gene PAV matrix file (File 4, Data Citation 3) and gene family PAV matrix file (File 6, Data Citation 3) are formatted as a matrix with values of either 0 (absence) or 1 (presence). Each line in the matrix represents a gene/gene family and each column represents a rice accession.

### Technical Validation

Based on the SNP data, the 3,010 rice accessions were clearly divided into *Xian-Indica*, *Geng-Japonica*, circum-Aus and circum-Basmati groups, together with some admixed accessions<sup>4</sup>. Here we reconstructed the phylogenetic relationship of the 453 rice accessions based on SVs and gene PAVs with ‘pars’ tool of

the phylip package (<http://evolution.genetics.washington.edu/phylip.html>). As a result, both the SVs and the gene PAVs showed consistent phylogenies (Fig. 2a and b), suggesting the validity of these data.

In order to demonstrate the usefulness of these data, we carried out genome-wide association studies of plant height, an agronomically important phenotype, based on both the SVs and the gene PAVs. For 323 out of the 453 rice accessions, the plant height data were collected. SVs including deletions, duplications and inversions were utilized for the analysis. Efficient mixed model analyses (EMMA) by EMMAX software<sup>21</sup> were used for the association studies. In detail, BN kinship matrixes were calculated to measure the genetic similarity between accessions based on the genome-wide SV data and gene PAV data, respectively. As a result, GWAS based on both SVs (Fig. 3a) and gene PAVs (Fig. 3b) successfully detected the *sd1* locus, the well-known “Green Revolution gene”, whose loss of function is associated with the greatly reduced plant heights<sup>22</sup>. Moreover, we also detected several other candidate variations significantly associated with plant heights.

This paper released two important types of variations beyond SNPs for rice. This is also the first release of gene presence/absence variations among hundreds of individuals for a higher eukaryote. We demonstrated the potential of these variations to explain phenotype variations.

## References

1. The 3k RGP. The 3,000 rice genomes project. *GigaScience* **3**, 1–6 (2014).
2. Li, J., Wang, J. & Zeigler, R. S. The 3,000 rice genomes project: new opportunities and challenges for future rice research. *GigaScience* **3**, 8 (2014).
3. Alexandrov, N. *et al.* SNP-Seek database of SNPs derived from 3000 rice genomes. *Nucleic Acids Res.* **43**, D1023 (2015).
4. Wang, W. *et al.* Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature*, <https://doi.org/10.1038/s41586-018-0063-9> (2018).
5. Chong, Z. *et al.* novoBreak: local assembly for breakpoint detection in cancer genomes. *Nat. Methods* **14**, 65–67 (2017).
6. Chen, K. *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods* **6**, 677–681 (2009).
7. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).
8. Hu, Z. *et al.* EUPAN enables pan-genome studies of a large number of eukaryotic genomes. *Bioinformatics*, *btx170* (2017).
9. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
10. Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* **1**, 18 (2012).
11. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
12. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
13. Sakai, H. *et al.* Rice Annotation Project Database (RAP-DB): An Integrative and Interactive Database for Rice Genomics. *Plant Cell Physiol.* **54**, e6 (2013).
14. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011).
15. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
16. Stanke, M., Steinkamp, R., Waack, S. & Morgenstern, B. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res.* **32**, W309–W312 (2004).
17. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
18. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
19. Li, Y. *et al.* De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat. Biotechnol.* **32**, 1045–1052 (2014).
20. Li, L., Stoeckert, C. J. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003).
21. Zhou, X. & Stephens, M. Genome-wide Efficient Mixed Model Analysis for Association Studies. *Nat. Genet.* **44**, 821 (2012).
22. Sasaki, A. *et al.* Green revolution: a mutant gibberellin-synthesis gene in rice. *Nature* **416**, 701–702 (2002).

## Data Citations

1. The 3,000 Rice Genome Project. *GigaScience Database* <http://dx.doi.org/10.5524/200001> (2014).
2. BGI. *NCBI BioProject* PRJEB6180 (2014).
3. Hu, Z. *et al.* *Figshare* <https://doi.org/10.6084/m9.figshare.c.3876022.v1> (2017).

## Acknowledgements

This work was supported by the National Key R&D Program (2016YFD0100904), the grants to Z.L. from the Bill & Melinda Gates Foundation Project (OPPGD1393 and OPP1130530), the CAAS Innovative Team Award, the National High Technology Research and the International S&T Cooperation Program of China (2012DFB32280), 863 Program (2014AA10A601 and 2014AA10A602-4), the Shenzhen Peacock Plan to Z.L. by Shenzhen Municipal Government of China, the National Key Technology Support Program (2015BAD01B02), Fundamental Research Funds for Central Non-Profit of Institute of Crop Sciences, Chinese Academy of Agricultural Sciences (Y2017CG21) and National Natural Science Foundation of China (31501291). We thank the High Performance Computing Center at Shanghai Jiao Tong University and Computing Center at Agricultural Genomics Institute at Shenzhen of CAAS for the computing. We thank Zechen Chong from M.D. Anderson Cancer Center for the help in running novoBreak.

### Author Contributions

Z.L. designed and conceived research; Z.H., W.W., C.S., J.S., J.L. and C.W. performed the novel sequences and pan-genome analyses; Z.W., M.L., W.W., J.R., B.F. and Z.L. performed the structural variation analyses; H.Z., W.W., Z.W., C.W. and Z.L. wrote the manuscript. All authors read and approved the final manuscript.

### Additional information

**Competing interests:** The authors declare no competing interests

**How to cite this article:** Hu Z. *et al.* Novel sequences, structural variations and gene presence variations of Asian cultivated rice. *Sci. Data* 5:180079 doi: 10.1038/sdata.2018.79 (2018).

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



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

The Creative Commons Public Domain Dedication waiver <http://creativecommons.org/publicdomain/zero/1.0/> applies to the metadata files made available in this article.

© The Author(s) 2018