Supplemental File for

Long-read sequencing of 111 rice genomes reveals significantly larger pan-genomes

Fan Zhang, Hongzhang Xue, Xiaorui Dong, Min Li, Xiaoming Zheng, Zhikang Li, Jianlong Xu, Wensheng Wang, and Chaochun Wei

Table of Contents

Supplemental Notes

De novo assembly, polishing, scaffolding and evaluation

For 75 newly sequenced OS accessions, each genome size was estimated using KmerGenie v1.7051 (Chikhi and Medvedev 2014) with short reads. The raw nanopore long reads were checked by NanoPlot v1.0.0 (De Coster et al. 2018) and trimmed (≥ Q7, ≥1000bp) by NanoFilt v2.6.0 (De Coster et al. 2018) with parameter "q 7 -l 1000". The trimmed long reads were corrected and assembled using NextDenovo v2.2.0 (https://github.com/Nextomics/NextDenovo) with the parameter "read_cutoff = 1000, seed_cutoff = 20000".

After genome assembling, the contigs were polished with both long and short reads. First, contigs were polished using Racon v1.4.11 (Vaser et al. 2017) with the recommended parameter "-m 8 -x -6 -g -8 -w 500" and Medaka v0.11.5 (https://github.com/nanoporetech/medaka) with the parameter "medaka_consensus -m r941_min_high_g303" with long reads. Next, all short reads were quality-controlled using FastQC v0.11.8 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) and trimmed using Trimmomatic v0.39 (Bolger et al. 2014) with the parameter "LEADING:3 TRAILING:3 MINLEN:36 HEADCROP:10". The contigs were mapped with short reads using Bowtie 2 v2.3.5.1 (Langmead and Salzberg 2012) and polished one round using Pilon v1.23 (Walker et al. 2014) with the default parameters.

For nine accessions sequenced by PacBio and Illumina platforms (H7L1 [Huanghuazhan], H7L26 [CDR22], H7L27 [PsBRC28], H7L28 [PsBRC66], H7L29 [IR64], H7L30 [Teqing], H7L31 [IR50], H7L32 [OM1723] and H7L33 [Phalguna]) in the 13 OS accessions, the long

reads were assembled using FALCON v1.8.7 (Chin et al. 2016). The assembled contigs were then polished using smrtlink v4.0 (https://www.pacb.com/support/software-downloads) with long reads and using Pilon with short reads.

Four of the 13 accessions were sequenced by Nanopore and Illumina platforms (SE-3 [BR 24], SE-19 [Zhong 413], SE-33 [BG 300] and SE-134 [Haonnong]).The long reads were first corrected using Nextdenovo and then assembled using smartdenovo (https://github.com/ruanjue/smartdenovo). The assembled contigs were polished using Pilon for three times with short reads.

For other rice accessions, we directly used assemblies and long reads published in NCBI SRA database in downstream analysis. With the NipRG's guide, contigs misassembled were corrected using "ragtag.py correct" and chromosome-level scaffolds were achieved using "ragtag.py scaffold" in RaGOO RagTag v1.0.0 (Alonge et al. 2019), which invoked MUMmer v3.9.4 (Kurtz et al. 2004) at the mapping step and minimap2 v2.17 (Li 2018) at the checking step. The quality of each genome assembly was evaluated by mapping to the NipRG with at least 90% as the threshold using QUAST v5.0.2 (Mikheenko et al. 2018) identity with the parameter "--min-identity 90". The completeness of each genome assembly was evaluated using BUSCO v5.1.2 (Seppey et al. 2019) with the database embryophyta_odb10 (eukaryota, 2020-09-10).

Pan-genome construction

The “map-to-pan” strategy was used to build the rice pan-genomes by combining NipRG and all novel sequences obtained from 111 cultivated and wild rice accessions / 105 cultivated rice accessions from both TGS and SGS (Supplemental Fig. S3). The sequences and gene annotations of the reference genome are available and high-quality (http://rice.plantbiology.msu.edu), so we focused on novel gene annotations from the obtained novel sequences.

At the step of unaligned sequences filtering, we tried to cut the partially unaligned sequences in order to retain the sequences not similar with reference genome (defined as “unmapped sequence blocks”). A scaffold/contig may contribute more than one unmapped sequence blocks.

At the step of redundancy removing, we chose Gclust instead of cd-hit-est for redundant sequence removing due to its ability to handle very long sequences. The remaining sequences were clustered into non-redundant sequences with identity cutoff of 90% using Gclust v1.0.0 (Li et al. 2019) with the parameter “-minlen 20 -both -nuc -threads 40 -ext 1 -sparse 2 -memiden 90”. Since redundant removing step is important for pan-genome construction, we applied the EUPAN blastCluster steps with identity cutoff of 90% again after the initial sequence redundancy removing again to ensure that we did not over-estimate the size of novel sequences.

In order to remove various contaminants not from Viridiplantae such as archaea, bacteria, viruses, fungi, the remaining sequences were mapped to NT database (18 Jun 2020) using BLAST+ v2.10.1 (Camacho et al. 2009) BLASTN with the parameter “-evalue 1e-5 -

best_hit_overhang 0.25 -perc_identity 0.5 -max_target_seqs 1". The sequences with hits from contaminants were dropped and the rest sequences were defined as candidate novel sequences.

A coverage-based method was used to check mis-assemblies in candidate novel sequences. First, the trimmed short reads and trimmed long reads were mapped to sequences combining NipRG and candidate novel sequences with Bowtie 2 v2.3.5.1 (Langmead and Salzberg 2012) and minimap2 v2.17 (Nanopore reads: map-ont, PacBio reads: map-pb) (Li 2018). Second, the alignment results were sorted with SAMtools v1.9 (Li et al. 2009). For each candidate novel sequence, we calculated the maximum mapped reads coverage in all samples from either short reads or long reads using BEDTools v2.29.2 (Quinlan and Hall 2010) with the parameter "genomecov -bga -split". Finally, the candidate novel sequences with more than 90% mapped region in at least one sample were considered as verified novel sequences.

Although the novel sequences were non-redundant, their borders may cover parts of gene bodies, resulting in predicting incomplete gene structures. To minimize the effect of this issue, we tried to elongate sequences to retain the whole gene body in the gene annotation step, and shortened sequences with no novel genes in elongated regions after the similar gene removing step. We estimated the elongated length = 5,000bp according to 90% quantile length of the predicted genes (~4,600bp). When two elongated sequences were close enough so that they overlap in a specific genomic location, they were merged into one single sequence for gene annotation. The novel representative genes in gene families were kept. We shortened

the sequences with no novel predicted genes in the elongated regions. The final pan-genome was generated by combining NipRG and novel sequences, with MSU7 and novel genes.

Comparison to EUPAN and HUPAN

For the comparison to the previous state-of-the-art approaches, we have applied this method and two of the state-of-the-art methods, EUPAN and HUPAN, to the 63-TGSRG data. EUPAN and HUPAN, were both developed to construct pan-genomes from short-read sequencing data, and they could not finish the whole pan-genome construction process for the long-read sequencing data. EUPAN keeps only the fully unaligned sequences, while HUPAN (an improved method for human pan-genome construction published in Genome biology, 2019) keeps the fully unaligned and partially unaligned sequences. However, at the step of redundancy removing, EUPAN used cd-hit-est as the tool to do the first round redundant sequence removing, and it may fail for long sequence mapping. We chose Gclust instead of cd-hit-est for redundant sequence removing. As mentioned in the cd-hit website (http://weizhong-lab.ucsd.edu/cd-hit/servers.php), the newly developed Gclust can deal with very long sequences. If EUPAN was applied to the 63-TGSRG data, only 1.05 Mb with 71 sequences in total from 40 of the 63 samples would be considered as full unaligned (before the redundancy removing step), the other sequences would be considered as partially unaligned sequences. For HUPAN, similarly but more extremely, it would consider almost all assembled contigs/scaffolds as partially unaligned. Because the assembled contigs/scaffold were very long and we knew some regions could be aligned to the reference genome with

high identify percentage, we considered both results from EUPAN and HUPAN not reasonable. The method introduced in this article tried to cut the partially unaligned sequences into blocks and retain the sequence blocks not similar to the reference genome to ensure the novel sequences remained in the final pan-genome were at least as novel (not similar to the reference genome) as the results from EUPAN or HUPAN for short-read sequencing data.

RNA-seq validation, functional domain and GO annotation of the predicted novel genes

RNA-seq data from 122 public samples (61 rice accessions of 2 tissues) were collected to validate the expressions of genes. All raw reads were quality-controlled using FastQC v0.11.8 and trimmed using Trimmomatic v0.39 with the parameter “LEADING:3 TRAILING:3 MINLEN:36 HEADCROP:10”. The trimmed reads were mapped to all transcripts (including 55,986 MSU7 genes and 19,319 novel genes). Only reads mapped in proper pair (that is, reads of a pair were mapped to the same transcript) were considered using SAMtools v1.9 with the parameters “view -f 2”. The coverage of transcripts was computed from BAM files using BEDTools v2.29.2 with the parameters “genomecov -ibam bamfile -max 1”. The transcripts with more than 95% coverage were considered as with expression evidence. The protein sequences of the predicted novel genes were extracted and input to InterProScan v5.45-80.0 (Jones et al. 2014), a tool integrated CDD-3.17, Coils-2.2.1, Gene3D-4.2.0, Hamap-2020_01, MobiDBLite-2.0, Pfam-33.1, PIRSF-3.10, PRINTS-42.0, ProSitePatterns-2019_11, ProSiteProfiles-2019_11, SFLD-4, SMART-7.1, SUPERFAMILY-1.75 and TIGRFAM-15.0 to predict domains and important sites of their proteins. The GO terms of

proteins were annotated as described in a previous study (Wang et al. 2018). Finally, 75.9% (14,658/19,319) of the predicted novel genes were annotated with at least one GO. GO enrichment analysis was performed using the package clusterProfiler v3.16.1 (Yu et al. 2012) in R v4.0.2. The GO terms with adjust P<0.05 using the Benjamini & Hochberg (BH) method were retained.

Impact of sequencing depths and sequencing platforms on gene PAV detection

In order to assess the possible effects of sequencing depths and sequencing platforms on gene PAVs detection, we compared gene family PAVs of accession IR64 sequenced by both Nanopore (ONT) and PacBio (PB) technologies with different sequencing depths (ont24×: 9.3Gbs; pb157×: 59.7Gbs; and pb85×: 32.2Gbs) and observed high Jaccard Indices (JIs) among them (0.962 ~ 0.977). This indicated that the sequencing platforms and sequencing depths have very limited impact on the gene family PAVs assessment, but higher sequencing depths did detect more gene families (Supplemental Fig. S7A).

Gene family PAV-based classification of rice populations

Based on the gene family PAVs, the 105 OS accessions could be classified into four major populations of XI (XI-1A, XI-1B, XI-2 and XI-3), GJ (GJ-tmp, GJ-sbtrp, and GJ-trp), cA and cB, largely consistent with previous classification of 3K-RG using single nucleotide polymorphisms (SNPs) (Wang et al. 2018) (Fig. 2E,F; Supplemental Fig. S7E,F). The genetic similarity was high between cA and XI (median JI = 0.905), and between GJ and cB (median JI = 0.909),

relatively low between XI/cA and GJ (median JI = 0.886/0.893), while the genetic similarity was high between subpopulations within XI (median JI = 0.923) or between subpopulations within GJ (median JI = 0.933) (Supplemental Fig. S7G,H). We noted three GJ accessions (TG64 [KAUK PAHLING], TG52 [VARIRANGAHY] and TG75 [Annongwangeng B] from GJ-sbtrp, GJ-trp and GJ-tmp), a cA accession (TG11 [JHONA 101]) and a cB accession (TG85 [Karnal Local]) were clustered into XI. An XI accession (TG16 [PHAN PHAE]) from Laos was clustered into population cB (Fig. 2F). The six OR accessions were well differentiated from the OS populations and formed two subpopulations with a genetic similarity of 0.901 (wild12, wild65, wild111 and wild131) and 0.909 (wild219 and wild273). The first cluster contains four accessions (wild12, wild65, wild111 and wild131 collected from Hainan [China], Sri Lanka, Naypyidaw [Myanmar] and Yangon [Myanmar]) and has JI of 0.878, 0.880, 0.896, 0.890 with XI, cA, GJ and cB. The remaining two accessions (wild219 and wild273 collected from Fujian [China] and Guangxi [China]) formed a separate cluster more closely related to population GJ and has JI of 0.874, 0.875, 0.874 and 0.869 with XI, cA, GJ and cB, respectively.

Examples of different PAVs derived from SGS and TGS

Two DNA transposon related genes, *LOC_Os05g27600* and *LOC_Os06g49820*, of Chr5 and Chr6 (Fig. 3G; Supplemental Fig. S8K) were shown in the main manuscript as examples to show short reads of SGS data did not have the complete upstream and downstream sequences of the genes with

repetitive sequences and thus confused PAVs detection or structure variation (SV) detection. These two homozygous deletions were also supported with the SV detection (10.4 kb deletion near *LOC_Os06g49882* with 65 supported reads; 5.7 kb deletion near LOC_Os05g27600 with 77 supported reads).

More examples, such as the gene PAVs for genes *LOC_Os04g01520* (Fig. 3H) and *LOC_Os04g31640* (Supplemental Fig. S8L) were detected only by TGS but rarely detected by SGS.

Comparisons of novel sequences and genes between different rice pan-genomes

OS pan-genomes (63-TGSRG/63-SGSRG) were built from 63 OS accessions selected from 3K-RG (TG1, TG2, TG3, TG4, TG5, TG6, TG7, TG8, TG9, TG10, TG13, TG14, TG15, TG17, TG18, TG19, TG21, TG22, TG24, TG27, TG28, TG29, TG30, TG31, TG32, TG33, TG34, TG42, TG43, TG45, TG46, TG49, TG50, TG51, TG52, TG53, TG55, TG56, TG58, TG59, TG60, TG61, TG62, TG63, TG64, TG65, TG68, TG70, TG75, TG76, TG77, TG78, TG80, TG81, TG82, TG83, TG84, TG85, TG86, TG87, TG88, TG90 and WW8) . These two pangenomes were constructed with the same method reported previously (Wang et al. 2018) for SGS data in three steps including getting unaligned contigs (more than 500bp), removing redundant sequences, and dropping contaminants.

The repeat-masked non-redundant novel sequences in different pan-genomes (111-TGSRG, 63-TGSRG, 63-SGSRG, 3K-RG,105OS-TGSRG and 6OR-TGSRG) were compared. For genome sequence level, they were aligned to each other using BLASTN with the

parameter “-evalue 1e-5 -max_target_seqs 10000”. A query sequence with various identity and length was considered according to different cutoffs. The total length of aligned regions was divided by the total repeat-masked non-redundant novel sequence length to obtain the total mapping length rate, which was used to measure the similarity between the pan-genomes.

We compared novel genes to the genomic sequences, transcripts and proteins from different pan-genomes. For novel genes with multiple transcripts, we chose the longest transcripts as queries and we constructed the pan-genomes as the database. The transcripts were aligned to the pan-genomes using BLASTN with the parameter of “-evalue 1e-5”. The high-scoring segment pairs (HSPs) with identity ≥95% were considered as hits and transcripts with ≥ 95% of regions covered were considered as aligned. At transcriptomic level, we compared transcripts from different pan-genomesusing cd-hit-est-2d in CD-HIT v4.8.1 with sequence identity threshold parameter of “-c 0.95”. At proteomic level, we mapped proteins of the predicted novel genes to all proteins from different rice pan-genomes (3K-RG, 63-TGSRG, 111-TGSRG, 66-RG) using cd-hit-2d in CD-HIT v4.8.1 with two sequence identity thresholds’ parameters of “-c 0.95” or “-n 2 -c 0.5”. The mapping rate is the number of genes mapped to the database divided by the total number of genes.

Supplemental Figures

Supplemental Figure S1

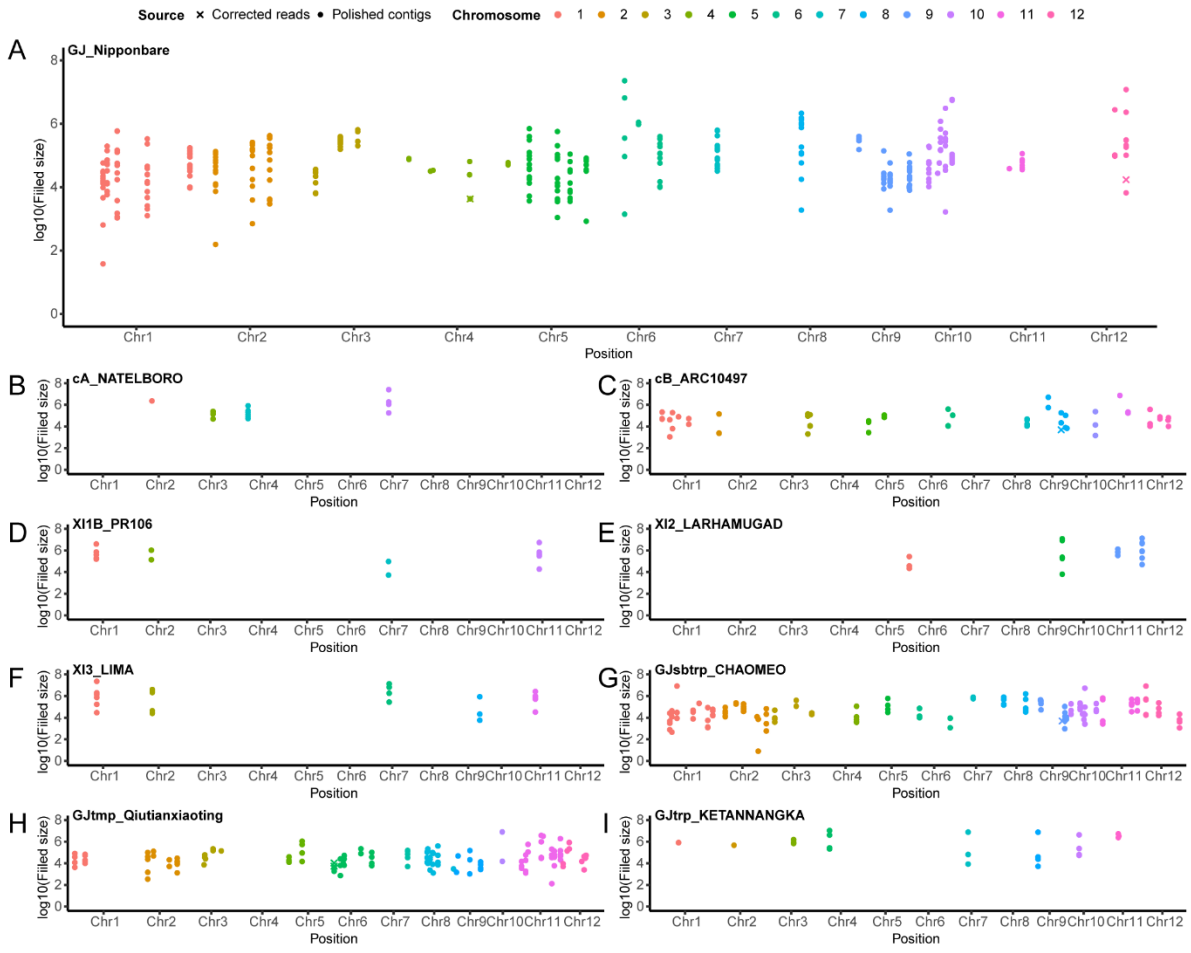

Figure S1. The filled gaps of 9 high-quality reference genomes. (A) The sizes of filed gaps (>=1000bp) in different chromosomes of Nipponbare. (B-I) The sizes of filled gaps in different chromosomes of (B) cA_NATELBORO , (C) cB_ARC10497, (D) XI-1B_PR106, (E) XI-2_LARHAMUGAD, (F) XI-3_LIMA, (G) GJ-sbtrp_CHAOMEO, (H) GJ-tmp_Qiutianxiaoting, and (I) GJ-trp_KETANNANGKA.

Supplemental Figure S2

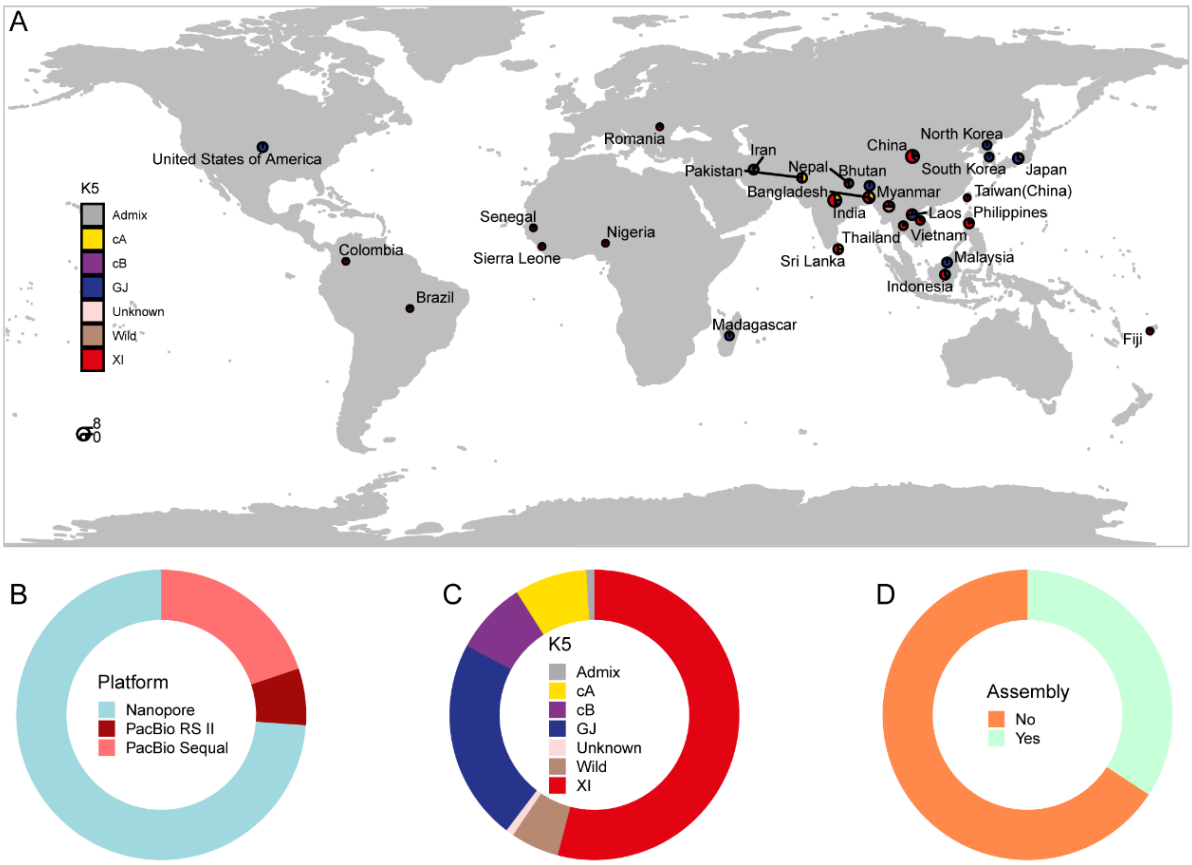

Figure S2. The summary of 111 rice accessions used in this study. (A) The geographical distribution and K5 subpopulation information of 111 rice accessions. (B-D) The sequencing platforms, K5 subpopulation information, data types (assembled genomes downloaded from a public database or newly sequenced genomes) of 111 rice accessions. The color stands for the K5 subpopulations of 111 rice accessions. (cA: Aus, cB: Bas, XI: Xian/Indica, GJ: Geng/Japonica, and Admix: Admixture).

Supplemental Figure S3

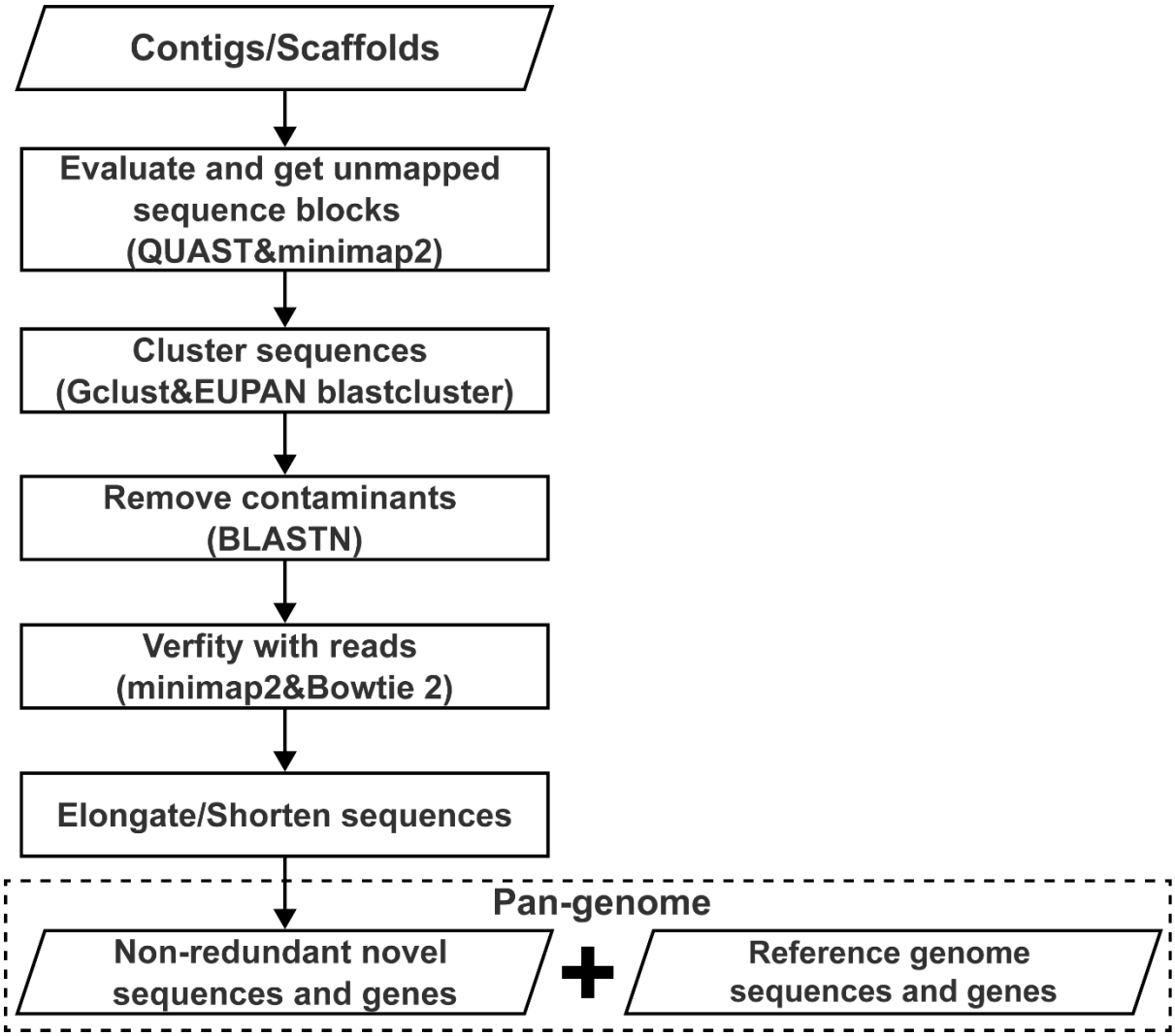

Figure S3. The system diagram of pan-genome construction method for 111 rice accessions. Evaluate and get unmapped sequence blocks: The unmapped sequence blocks more than 500 bp (from QUAST evalution results of contigs/scaffolds) were mapped to NipRG (including mitochondrion and plastid) again using minimap2. Cluster sequences: The unmapped sequences were clustered using Gclust and EUPAN blastcluster to choose the representative ones. Remove contaminants: The unmapped sequence blocks were aligned to NT database using BLASTN and the ones aligned to organisms out of Viridiplantae were removed. Verify with reads: The reads were mapped to the retained unmapped sequence blocks and remove low-coverage ones. Elongate/Shorten sequences: The unmapped sequences will be elongated if the new genes are predicted overlap the bounds of unmapped sequence blocks. The final pan-genome combines non-redundant novel sequences and genes with reference genome sequences and genes.

Supplemental Figure S4

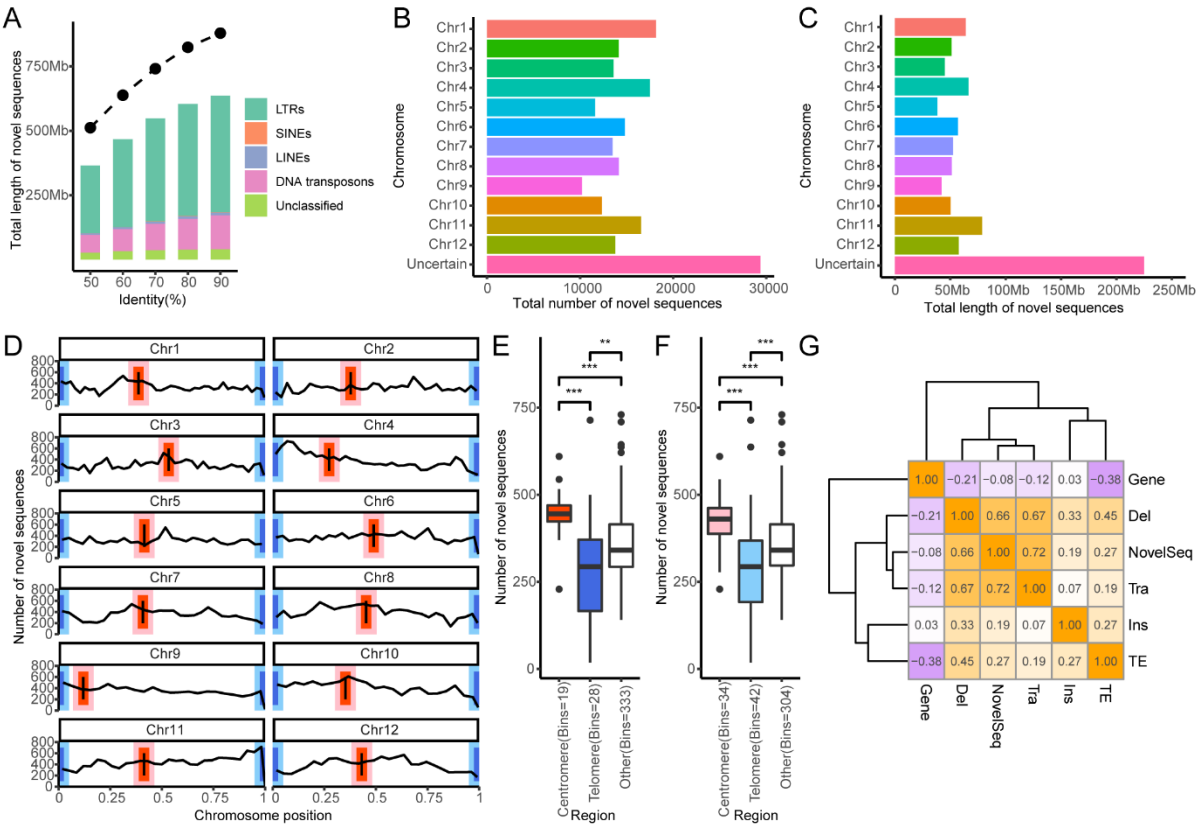

Figure S4. The non-redundant novel sequences of 111 rice accessions. (A) The total sizes and their TE components of novel suquences constructed with different parameter settings, especially the global identity percentages set in the redundancy removing step (B) The numbers of novel sequences and (C) The total lengths of novel sequences in different chromosomes. (D) The distribution of novel sequences in different chromosomes. (E-F) The number of novel sequences near centromere/telomere/other regions. Red/pink: upstream and downstream 2.5% / 5% regions near the centromeres. Blue / lightblue: upstream or downstream 2.5% / 5% regions near telomeres. White: Other regions. (G) The spearman correlation coefficients between any two distributions of genes / deletions / novel sequence / translocation breakpoints / insertions / TEs.

Supplemental Figure S5

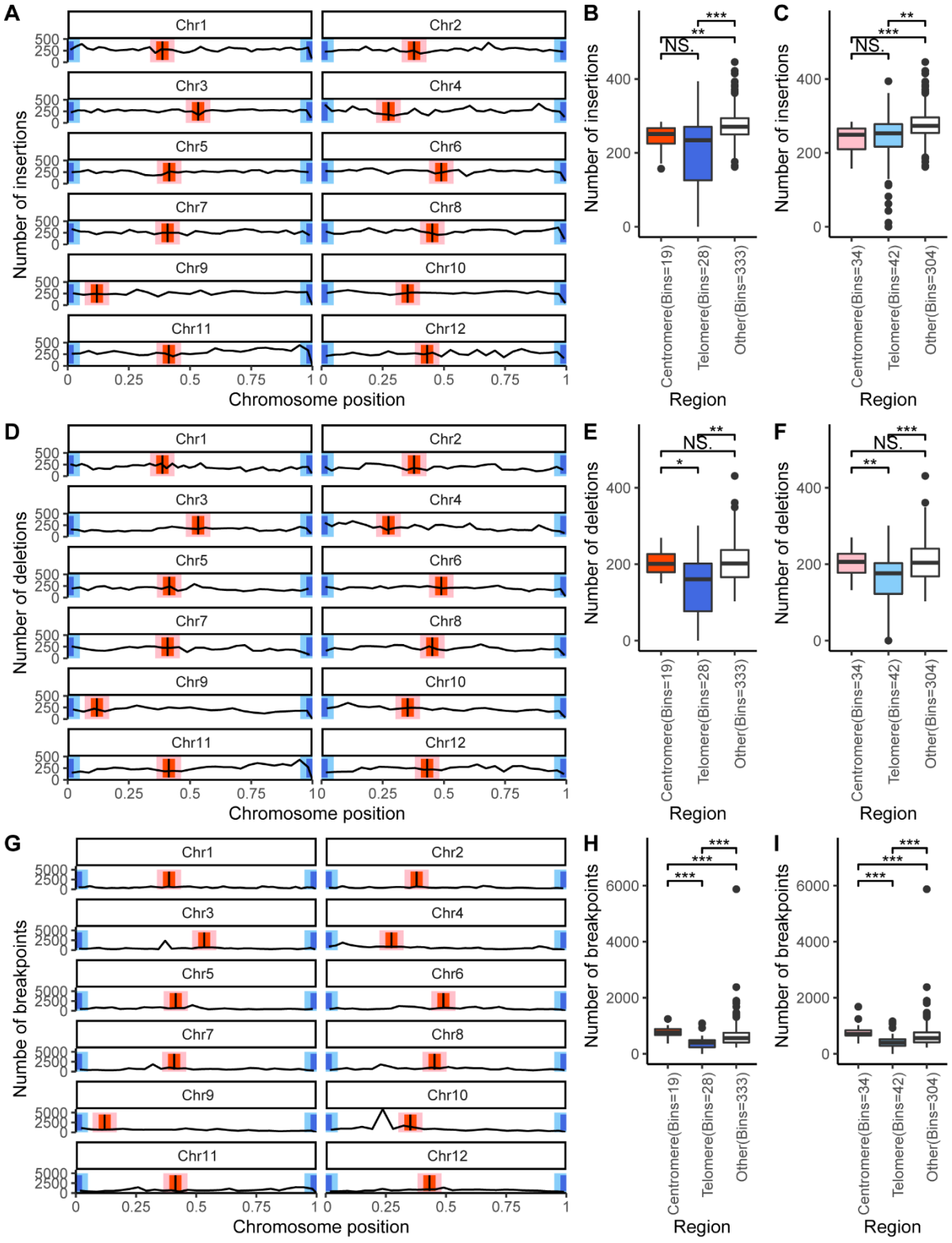

Figure S5. The non-redundant SVs of 111 rice accessions. (A-C) insertions. (D-F) deletions. (G-I) translocation breakpoints. The color scheme is the same as in Supplemental Figure 4D-4F.

Supplemental Figure S6

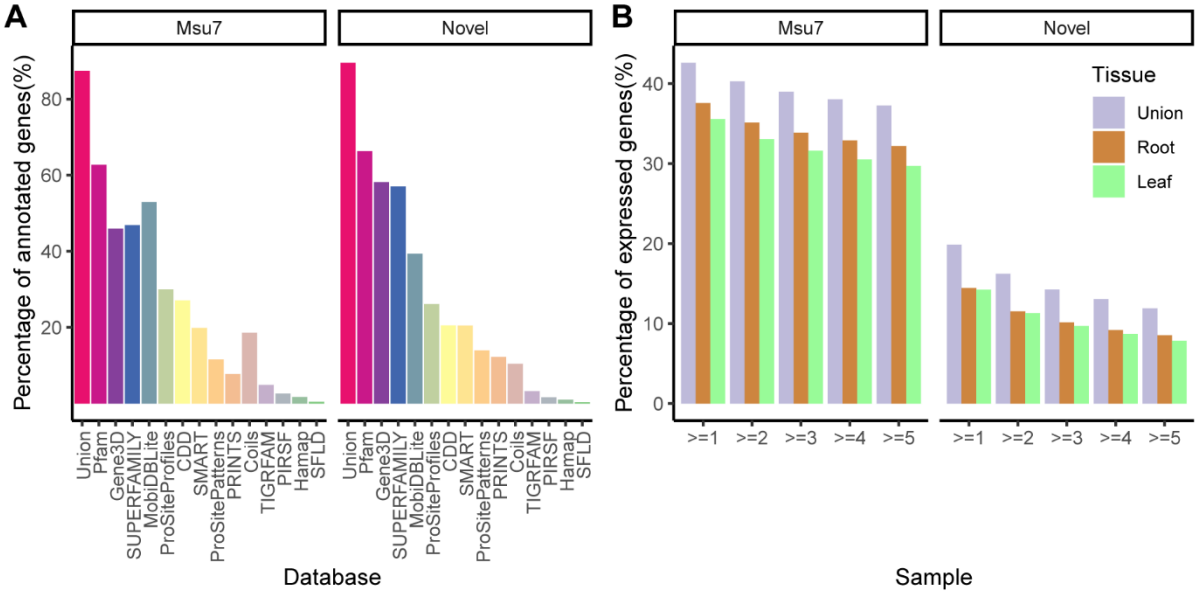

Figure S6. The percentage of MSU7 genes (n=55,986) and novel genes (n=19,319) with domain annotation and RNA-seq evidence. (A) MSU7 genes and novel genes with domain annotation. ‘Union’ means the union of database Pfam, Gene3D, SUPERFAMILY, MobiDBLite, ProSiteProfiles, CDD, SMART, ProSitePatterns, PRINTS, Coils, TIGRFAM, PIRSF, Hamap and SFLD. (B) MSU7 genes and novel genes with RNA-seq evidence. ‘Union’ means the union of tissue root and leaf.

Supplemental Figure S7

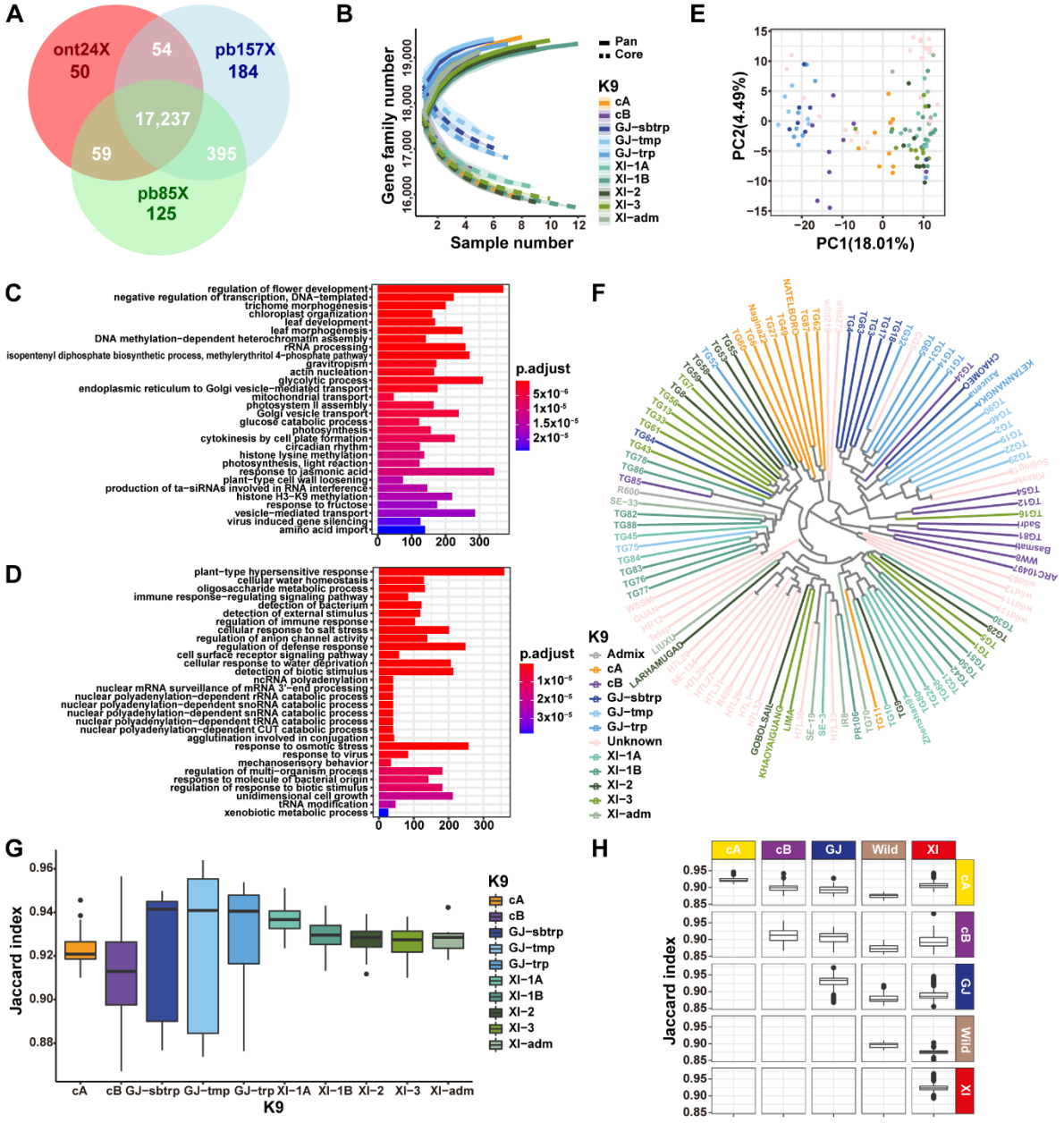

Figure S7. Gene family PAVs of 111 rice accessions. (A) Gene family level overlapping of rice accession IR64 sequenced with Nanopore 24× (ont24X), PacBio 157× (pb157X) and PacBio 85× (pb85X). (B) The pan-genome size estimation using111 rice accessions for K9 subpopulations (Admix and GJ-adm were ignored due to their small sample sizes (one sample and no sample respectively). (C) The biological process GO enrichment terms of core and softcore genes. (D) The biological process GO enrichment terms of distributed genes. (E) The PCA analysis of 111 rice accessions of K9 subpopulations using gene family PAVs. (F) The clustering of 111 rice accessions using gene family PAVs. (G) The similarity of gene family PAVs in K9 subpopulations. (H) The similarity of gene family PAVs in and between K5 and wild subpopulations.

Supplemental Figure S8

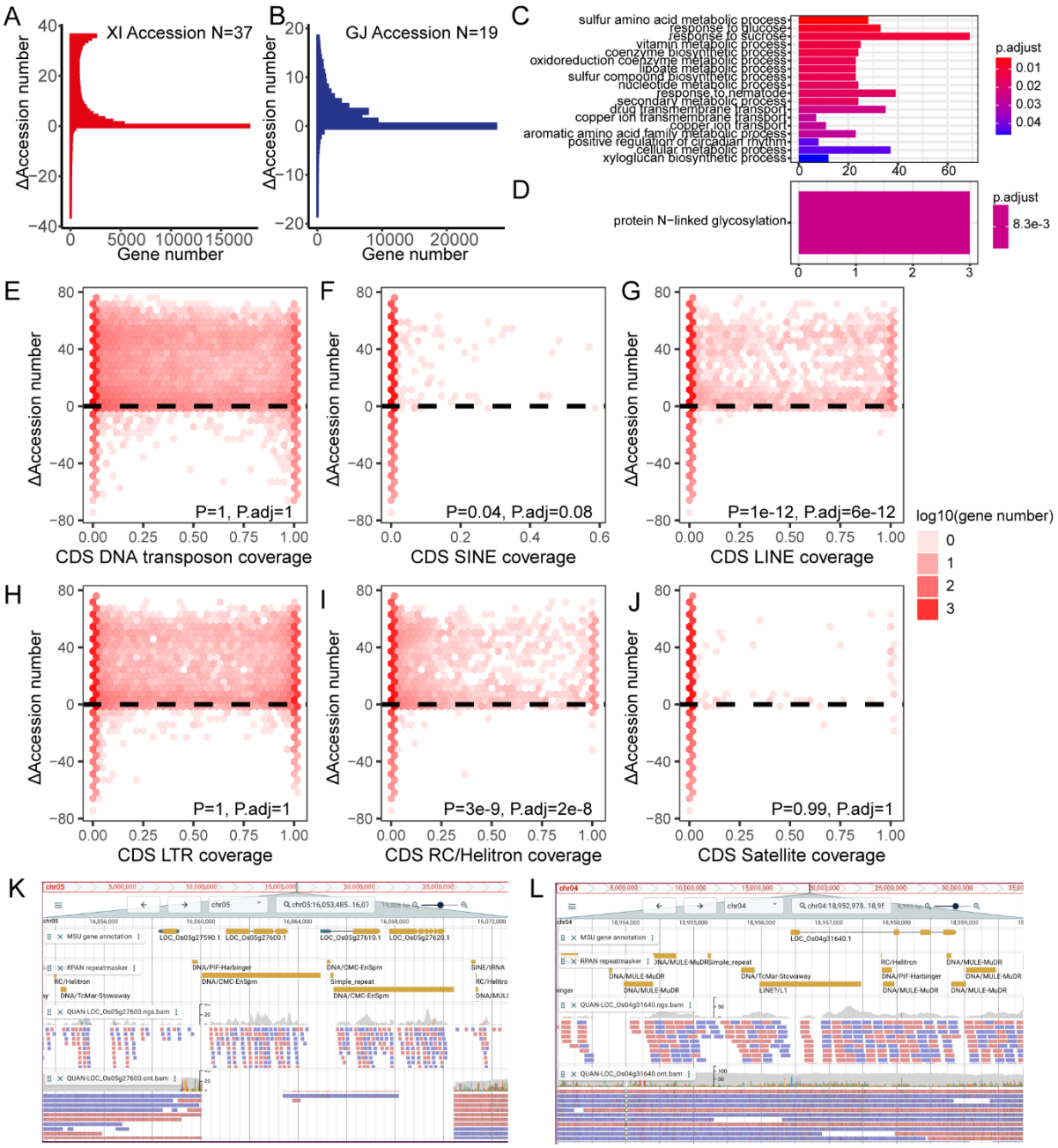

Figure S8. The differences of gene PAVs from NGS and TGS data. (A-B) The relationship between the numbers of accessions with different number of gene PAVs detected (for each gene, △Accession number = TGS detected accession number - NGS detected accession number) and gene numbers in (A) XI population (Accession N=37) and (B) GJ population (Accession N=19)**.** (C-D) The biological process GO enrichment terms of (C) TGS-preferred genes and (D) NGS-perferred genes. (E-J) The relationship between different number of detected accessions and (E) CDS DNA transposons coverage, (F) CDS SINEs coverage,

(G) CDS LINEs coverage, (H) CDS LTR coverage, (I) CDS RC/Helitron coverage, (J) CDS satellite coverage. Some NGS-preffered genes have CDS regions fully overlapped with DNA transposons or LTRs. One side Wilcoxon Rank Sum test (alternative hypothesis: TGS-preferred genes have higher repeat element coverage in the CDS regions. Adjust method = "Holm") was applied to exam which group (TGS preferred genes (△Accession number > 0) or NGS preferred genes (△Accession number < 0)) have a higher percentage of CDS regions covered with repeat elements. (K) The read alignment of a NGS-preferred gene LOC_Os05g27600 (△Accession number = -66) in rice accession QUAN from NGS and TGS. (L) The read alignment of a TGS-preferred gene LOC_Os04g31640 (△Accession number = 75) in rice accession QUAN from NGS and TGS.

Supplemental Figure S9

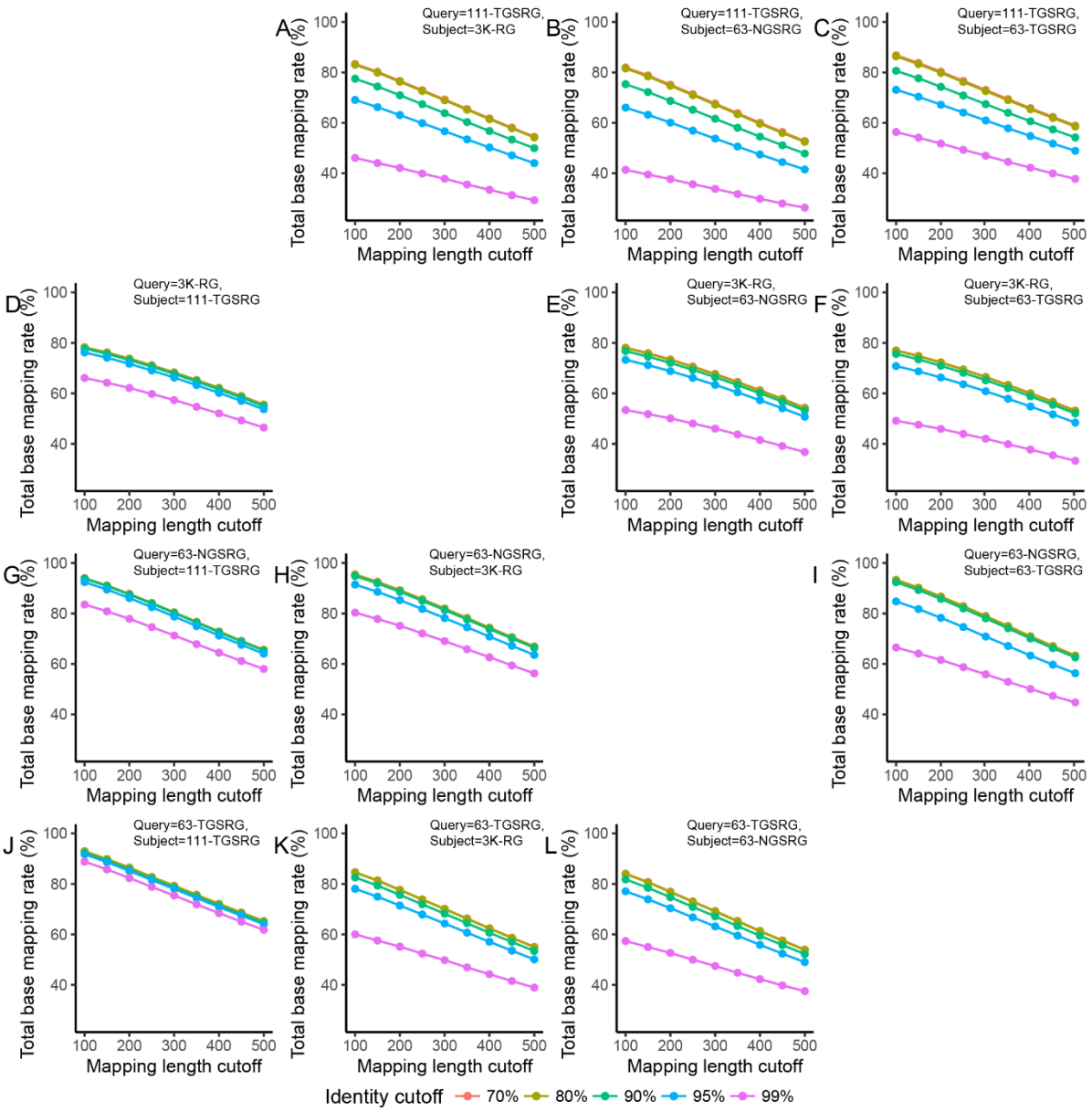

Figure S9. Overlapping of sequences in pan-genomes derived from 111-TGSRG, 3K-RG, 63-NGSRG, 63-TGSRG. (A-C) 111-TGSRG novel sequences (non-redundant, non-repeat) mapped to the pan-genome derived from (A) 3K-RG, (B) 63-NGSRG and (C) 63-TGSRG. (D-F) 3K-RG novel sequences mapped to the pan-genome derived from (D) 111-TGSRG, (E) 63-NGSRG and (F) 63-TGSRG. (G-I) 63-NGSRG novel sequences mapped to the pan-genome derived from (G) 111-TGSRG, (H) 3K-RG and (I) 63-TGSRG. (J-L) 63-TGSRG novel sequences mapped to the pan-genome derived from (J) 111-TGSRG, (K) 3K-RG and (L) 63-NGSRG.

Supplemental Figure S10

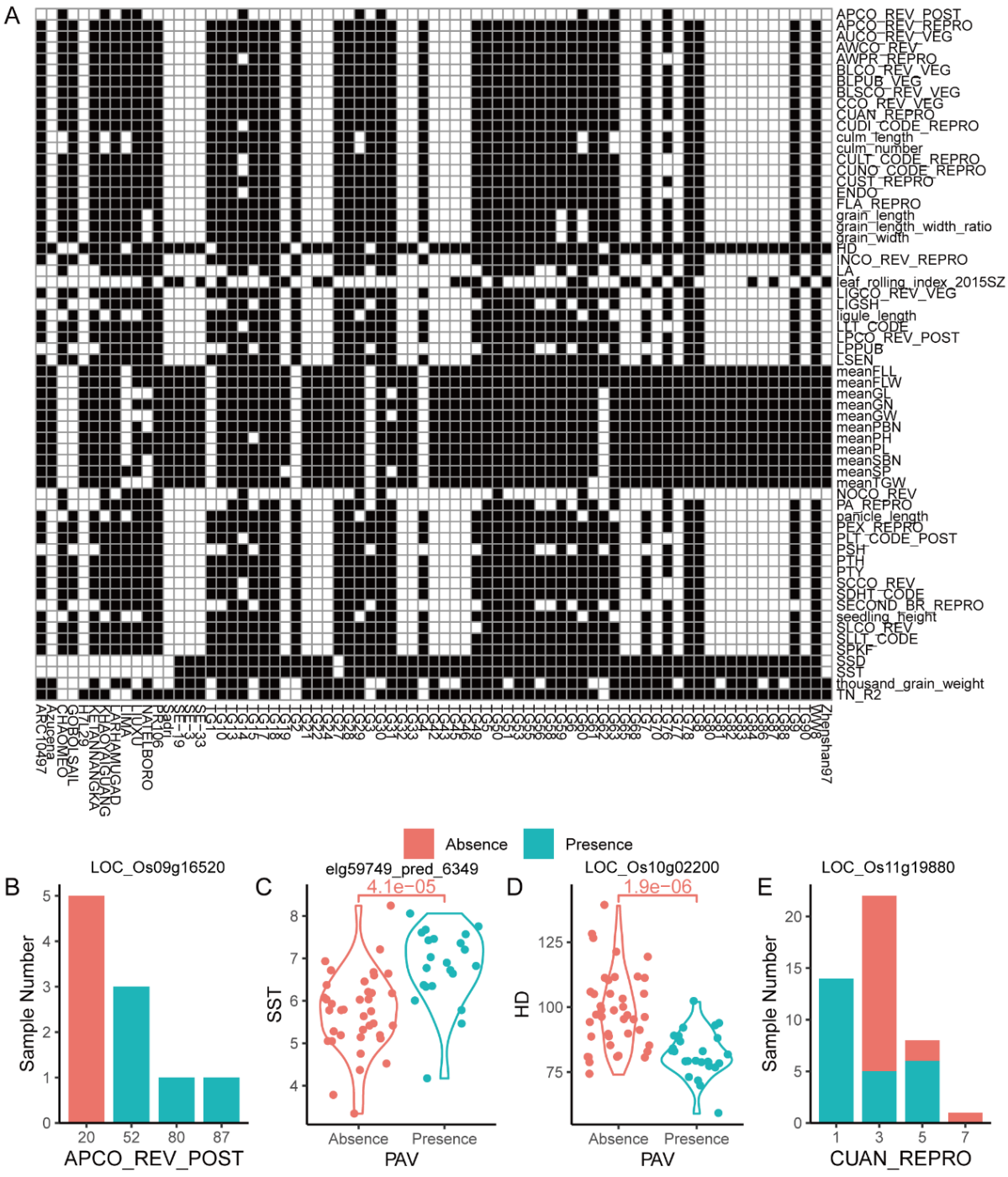

Figure S10. Associations between phenotypes and gene PAVs. (A) The available values (black) and the missing values (white) in phenotypes of rice accessions. (B-E) Some examples with low p-values but high FDRs. (B) APCO_REV_REPRO: apiculus color at reproductive, 20-straw; 52-brown (tawny); 80-purple; 87-purple apex. LOC_Os09g16520 ($p=8.0\times10^{-3}$, FDR=0.29). (C) SST: Score of salt toxicity of leaves. elg59749_pred_6349 ($p=4.1\times10^{-5}$, FDR=0.77). (D) HD: heading date. LOC_Os10g02200 ($p=2.0\times10^{-5}$, FDR=0.07). (E) CUAN_REPRO: culm angle at reproductive, 1-erect (<15°); 3-semi erect (intermediated

~20°); 5-open (~40°); 7-spreading (>60°-80°, culms not resting on the grounds).

LOC_Os11g19880 ($p=2.0\times10^{-6}$, FDR=0.07).

Supplemental Tables

Supplemental Table S1

Table S1. Summary of sample information from 113 samples of 111 rice accessions. In file Supplemental_Table_S1.xlsx.

Supplemental Table S2

Table S2. Summary of raw nanopore sequencing data from 69 cultivated and 6 wild rice accessions. In file Supplemental_Table_S2.xlsx.

Supplemental Table S3

Table S3. Summary of raw illumina short reads for 69 cultivated and 6 wild rice accessions. In file Supplemental_Table_S3.xlsx.

Supplemental Table S4

Table S4. Summary of assembly metrics from 69+13 cultivated and 6 wild rice accessions. In file Supplemental_Table_S4.xlsx.

Supplemental Table S5

Table S5. Summary of sequencing data of 25 rice accessions from public databases. In file Supplemental_Table_S5.xlsx.

Supplemental Table S6

Table S6. Summary of gap-filled high-quality reference genomes of rice accessions. In file Supplemental_Table_S6.xlsx.

Supplemental Table S7

Table S7. Mapping rates of novel genes to the reference genome and three different pan-genomes (identity >= 95%, transcript coverage >= 95%, all high-scoring segment pairs' lengths are more than 28bps). In file Supplemental_Table_S7.xlsx.

Supplemental Table S8

Table S8. Mapping rates of novel gene transcripts to transcripts from different genomes or pan-genomes (global identity >= 95%). In file Supplemental_Table_S8.xlsx.

Supplemental Table S9

Table S9. Mapping rate of novel gene proteins to different proteins of genomes or pan-genomes (global identity >= 95%). In file Supplemental_Table_S9.xlsx.

Supplemental Table S10

Table S10. Mapping rate of novel gene proteins to different proteins of genomes or pan-genomes (global identity >= 50%). In file Supplemental_Table_S10.xlsx.

Supplemental Table S11

Table S11. Assocation of phenotypes and genes (only results with P<5e-2 and FDR<5e-2 are listed). In file Supplemental_Table_S11.xlsx.

References

Alonge M, Soyk S, Ramakrishnan S, Wang X, Goodwin S, Sedlazeck FJ, Lippman ZB, Schatz MC. 2019. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome Biol* **20**: 224.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114-2120.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10**: 421.

Chikhi R, Medvedev P. 2014. Informed and automated k-mer size selection for genome assembly. *Bioinformatics* **30**: 31-37.

Chin CS, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, Dunn C, O'Malley R, Figueroa-Balderas R, Morales-Cruz A et al. 2016. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods* **13**: 1050-1054.

De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. 2018. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* **34**: 2666-2669.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. *Genome Biol* **5**: R12.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357-359.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094-3100.

Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. 2018. Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics* **34**: i142-i150.

Seppey M, Manni M, Zdobnov EM. 2019. BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods Mol Biol* **1962**: 227-245.

Vaser R, Sovic I, Nagarajan N, Sikic M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res* **27**: 737-746.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**: e112963.

Wang W, Mauleon R, Hu Z, Chebotarov D, Tai S, Wu Z, Li M, Zheng T, Fuentes RR, Zhang F et al. 2018. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature* **557**: 43-49.