Improvements to the Rice Genome Annotation Through Large-Scale Analysis of RNA-Seq and Proteomics Data Sets

Authors

Zhe Ren, Da Qi, Nina Pugh, Kai Li, Bo Wen, Ruo Zhou, Shaohang Xu, Siqi Liu, and Andrew R. Jones

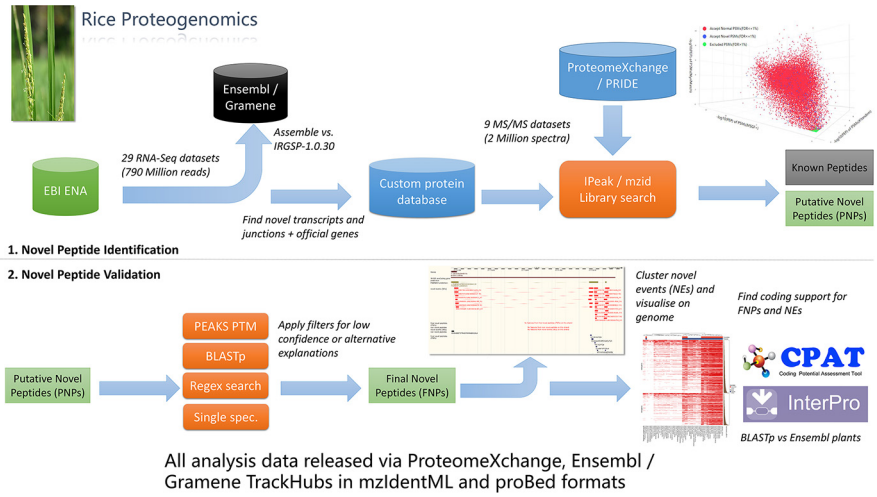
Correspondence

siqiliu@genomics.cn; andrew.jones@liverpool.ac.uk

In Brief

The genome of rice has been sequenced in the past, serving as a starting point to search for useful genetic traits. However, the process of annotating all the genes is a challenging and ongoing process. We have re-analyzed large amounts of publicly available data about rice proteins, to correct errors in gene sequences and find new genes. Our new results are presented in a simple format to allow users of databases to see the correspondence between genes and the proteins.

Graphical Abstract



Highlights

- We have mapped public proteomics data against the rice genome/transcriptome.
- We discovered 1584 novel peptides not currently explained by gene models.
- 101 new loci were matched by novel peptides, not currently annotated as genes.
- Data are made persistently available for simple visualization on genome browsers.

Varland et al., 2019, Molecular & Cellular Proteomics *18*, 86–98
January 2019

https://doi.org/10.1074/mcp.RA118.000832

⌘ *Author's Choice*

Improvements to the Rice Genome Annotation Through Large-Scale Analysis of RNA-Seq and Proteomics Data Sets*[S]

Zhe Ren‡§§, Da Qi‡§§, Nina Pugh§, Kai Li‡, Bo Wen‖¶, Ruo Zhou‡, Shaohang Xu‡, Siqi Liu‡**, and Andrew R. Jones§††

Rice (*Oryza sativa*) is one of the most important worldwide crops. The genome has been available for over 10 years and has undergone several rounds of annotation. We created a comprehensive database of transcripts from 29 public RNA sequencing data sets, officially predicted genes from Ensembl plants, and common contaminants in which to search for protein-level evidence. We re-analyzed nine publicly accessible rice proteomics data sets. In total, we identified 420K peptide spectrum matches from 47K peptides and 8,187 protein groups. 4168 peptides were initially classed as putative novel peptides (not matching official genes). Following a strict filtration scheme to rule out other possible explanations, we discovered 1,584 high confidence novel peptides. The novel peptides were clustered into 692 genomic loci where our results suggest annotation improvements. 80% of the novel peptides had an ortholog match in the curated protein sequence set from at least one other plant species. For the peptides clustering in intergenic regions (and thus potentially new genes), 101 loci were identified, for which 43 had a high-confidence hit for a protein domain. Our results can be displayed as tracks on the Ensembl genome or other browsers supporting Track Hubs, to support re-annotation of the rice genome. ***Molecular & Cellular Proteomics 18: 86–98, 2019. DOI: 10.1074/mcp.RA118.000832.***

The development of next-generation and third generation sequencing technologies mean that genome sequences are now being routinely generated for an ever-expanding range of species, strains, breeds, and even individuals within populations. For the genome to be useful for fundamental and applied research requires high-quality annotation. Following genome assembly, annotation involves the discovery of the start codons for all genes, and their exon splicing patterns, which is a highly challenging task. Gene finding in most genome projects is performed via software that makes *ab initio* predictions of coding sequences and, where possible, uses homology to other annotated genomes. Experimental data in the form of large-scale RNA Sequencing (RNA-Seq)[1] is also commonly used to find mRNAs and align reads that cross-intron junctions to infer splicing. Undoubtedly the use of large-scale RNA-Seq data vastly improves genome annotation but nevertheless, all genomes suffer from some proportion of mistaken annotation, such as incorrect translation initiation sites, incorrect splicing or pseudogenes called as protein-coding.

It is now becoming widely recognized that inference of the protein-coding elements of the genome can be greatly improved using large-scale mass spectrometry (MS) data on peptide sequences, in so-called *proteogenomics* approaches (1). In a typical proteogenomics pipeline, MS/MS spectra are searched against a customized protein sequence database, produced from curated gene predictions, as well as incorporating predicted possible sequences from *ab initio* gene finders and/or aligned RNA-Seq derived transcripts. Therefore, proteogenomics not only provides expression-level evidence of protein-coding genes but also has the potential to improve the protein-coding gene sets *i.e.* proteogenomics can provide evidence that novel transcripts or alternative predictions (for known genes) have supporting evidence at the protein sequence level. There have been several proteogenomics studies on plants that have shown the ability to discover novel protein-coding genes and predict or improve splicing annotation. For instance, in 2008, Castellana *et al.* performed a proteogenomics analysis on *Arabidopsis* tissues (2). They successfully identified 778 novel genes and made 695 gene model refinements. Later in 2014, they developed an automatic method of proteogenomics and performed analysis on *Zea mays*, finding 165 novel protein-coding genes and proposing updated models for 741 additional genes (3).

From the ‡BGI-Shenzhen, Shenzhen 518083, China; §Institute of Integrative Biology, University of Liverpool, Liverpool, L69 7ZB, UK; ‖Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas 77030; ¶Lester and Sue Smith Breast Center, Baylor College of Medicine, Houston, Texas 77030

Received May 3, 2018, and in revised form, August 31, 2018
Published, MCP Papers in Press, October 5, 2018, DOI 10.1074/mcp.RA118.000832

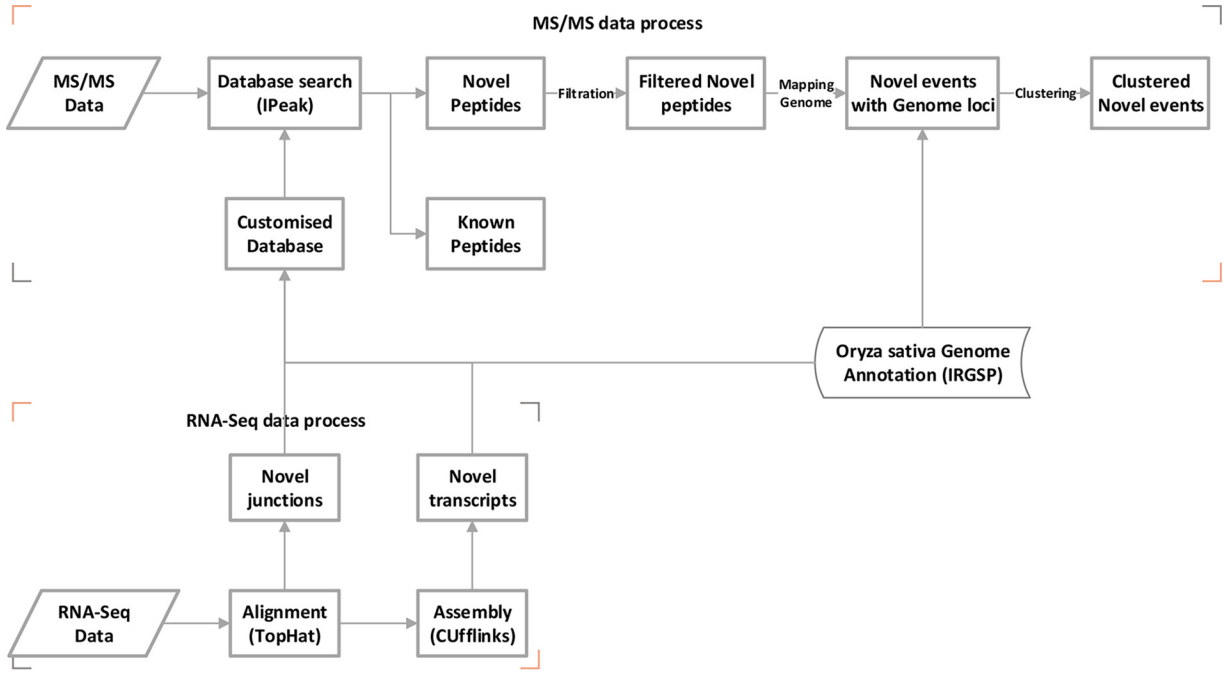


FIG. 1. **An overview of the analysis workflow for proteogenomics in this study.**

Rice (*Oryza sativa*) is the staple food for half the world's population. The completion of the rice genome sequencing, and several rounds of annotation, have provided a base for molecular and genetic studies (4–6). Comprehensive genomic and transcriptomic studies of rice have been conducted worldwide, serving as a base for research aimed at matching the demand of increasing food supplies (7–12). A previous effort at proteogenomics analysis on rice has been performed, and a database produced (although no longer searchable online), in which LC-MS data sets were queried against gene predictions made from the relevant genome build at that time (13). Herein, we have performed a comprehensive proteogenomics analysis on rice through collecting public genomics, transcriptomics and proteomics data, to discover novel protein-coding genes and new splice sites.

With the development of genomics, transcriptomics and proteomics techniques, the ability to detect ever higher proportions of the transcribed genes and evidence for translated proteins has become possible via proteogenomics. In addition, tools and strategies, such as customProDB and SpliceDB (14), have effectively improved the performance of proteogenomics by facilitating improved design of the search database. The construction of "novel event" candidates (*i.e.* new exons or splice junctions) is one of the most important steps in proteogenomics. Some studies aim to be comprehensive, such as using six frame translation of whole genome sequence (15, 16), although these approaches are likely to contain exonic sequences for all possible genes, they suffer from a lack of statistical power (because of overall database size) and splicing information. An alternative is to use *ab initio* gene predictions from gene finding software (17). In this case, the constructed candidates will contain predicted splice events, but rely on the accuracy of gene finding software that is not generally high, meaning that some possible splice sites will be missed. A third alternative is to create a search database from mapping RNA-Seq data onto the genome. The use of RNA-Seq results can balance these aspects, keeping a relative comprehensive search space but without the size expansion of six frame translations.

We designed our proteogenomics pipeline as follows. First, for database construction, we used transcriptomics data aligned onto the genome. For the translation step, we kept only the longest frame to control the overall database size. Second, for database searching we used multiple search engines via our previously published IPeak approach (18). IPeak combines the machine learning approach of Percolator and the FDRScore algorithm for search engine integration, which has been demonstrated to improve sensitivity over using a single search engine (19). IPeak is available as part of the mzid Library and ProteoAnnotator projects (20, 21). Third, we performed extensive filtration to ensure that identified peptides not matching the official annotation (novel peptides) were high confidence and the corresponding spectra could not be explained by other causes. Fourth, to validate and annotate the resulting novel peptides and corresponding novel events, we aligned our novel peptides back onto the genome for visualization against other tracks of evidence.

[1] The abbreviations used are: RNA-Seq, RNA sequencing; MS/MS, tandem Mass Spectrometry; CDS, protein-coding sequences; ORF, open reading frame; PSM, peptide spectrum match; PTM, post-translational modification; FDR, false discovery rate; PEP, posterior error probability; PSI, Proteomics Standards Initiative.

Last, to standardize the presentation of results, we use standard formats from the Proteomics Standards Initiative (PSI) - mzIdentML (22) and proBed (23), which allow for rapid and automated visualization of the results via public genome browsers. Using 29 data sets of RNA-Seq data (789,141,453 reads), and 9 MS/MS data sets (2,051,418 spectra), this study represents one of the most comprehensive proteogenomics efforts undertaken on rice.

EXPERIMENTAL PROCEDURES

An overview of the pipeline used for rice proteogenomics is summarized in Fig. 1. The workflow is mainly divided to two parts for the processing of the RNA-Seq and MS/MS data, as follows.

Data Collection—In this study, raw RNA-Seq data that was generated from the Illumina platform in paired end mode was collected from the European Nucleotide Archive (ENA, http://www.ebi.ac.uk/ena) database. A total of 29 runs (153,907,936,648 bases/ 789,141,453 reads) was contained in the data sets, and the full details are listed in SupplementaryFile1.xlsx (tab: "RNA Data Collection"). Data sets from various sources were merged to provide a comprehensive database of possible transcripts to match against.

To search for peptide evidence, MS/MS data acquired from high-resolution mass spectrometers (LTQ Orbitrap XL, LTQ Orbitrap Velos, TripleTOF 5600 and Q Exactive) was used, regardless of whether the data was generated from profiling or enrichment studies. The raw MS/MS data for this study was collected from the ProteomeXchange (PX, http://www.proteomexchange.org/) database, including a total of nine data sets and 2,051,418 MS/MS spectra. Detailed information about data sets is listed in Table I and further in SupplementaryFile1.xlsx (tab: "MSMS Data Collection"). As shown in SupplementaryFile1.xlsx, in most cases, MS/MS data sets were source from *O. sativa* Japonica, which is considered the reference genome. However, to further increase coverage, several proteomics data sets were sourced from *O. sativa* Indica, and one from *O. sativa* KDML105 (Thai Jasmine) rice. Results are only presented if they map 100% to the Japonica reference, and the results are also presented demonstrating the source data set for each novel peptide found (file SupplementaryFile1.xlsx, tab: "Putative Novel Peptides"), enabling filtration of those observed only in certain strains.

Construction of the Customized Database Based On RNA-Seq Data—The RNA-Seq reads from each run were individually aligned using TopHat (v2.0.12) against the *Oryza sativa* genome (IRGSP-1.0.30). The accepted matches in Bam format and the junctions in BED format were produced by TopHat. The parameters used in TopHat mapping were set as: the alignment sensitivity at "very sensitive," read mismatches at 2, the expected inner distance between mate pairs at 150, library type at fr-unstranded and other parameters at default. All the accepted reads from each run were individually sent for assembly into transcript sequences by Cufflinks (v2.2.1). Afterward, Cuffmerge was employed to combine the transcripts from each run to form longer transcripts in GTF format. The longer transcripts marked with class code " = " are from the transcripts completely matched to the known exons, termed as *known transcripts*, whereas those with other class codes are from the transcripts partially or totally mismatched to the known exons (IRGSP-1.0.30), assigned as *novel transcripts (NTs)*. All the *novel transcripts* in GTF format were taken for construction of the customized database. All the junctions were first de-duplicated and aligned against the "official junction sites" from the *Oryza sativa* annotation (IRGSP-1.0.30) to filter out the *known junctions* by custom scripts, and the remaining junctions were considered as *novel junctions (NJs)* for further construction of the customized database. All the novel junctions and novel transcripts are collectively called *novel events (NEs)*.

TABLE I

The raw MS/MS data collected for this study from ProteomeXchange database. Detailed search parameters used are provided in supplementary File S1

Data set Identifier	Title	Instrument	Publication	Announce Date	Spectra count
PXD000265	Oryza sativa egg, sperm, callus, pollen and seedling proteome	LTQ Orbitrap XL; LTQ Orbitrap XL	Abiko *et al.* 2013 (32)	2013/8/2	253743
PXD000313	Quantitative proteomics study on rice embryo during embryogenesis by using isobaric tags for relative and absolute quantification (iTRAQ)	Q Exactive	Zi *et al.* 2013 (33)	2013/8/6	976822
PXD000923	Rice Pistil LC-MSMS	TripleTOF 5600	Wang *et al.* 2014 (34)	2014/7/31	87502
PXD001030	Proteomic analysis of proteins related to rice grain chalkiness using iTRAQ based on a notched-belly mutant with white-belly	TripleTOF 5600	Lin *et al.* 2014 (35)	2014/6/17	301027
PXD001058	Unravelling the proteomic profile of rice meiocytes during early meiosis	LTQ Orbitrap Velos	Collado-Romero, Alós & Prieto 2014 (36)	2014/8/13	92162
PXD002291	Rice (Oryza sativa) lysine-acetylation LC-MS/MS	Q Exactive	Xiong *et al.* 2016 (37)	2016/2/24	94400
PXD002739	Acetylome analyses in the germinating rice seed	Q Exactive	He *et al.* 2016 (38)	2016/1/26	34344
PXD002740	Succinylome analyses in the germinating rice seed	Q Exactive	He *et al.* 2016 (38)	2016/1/26	25862
PXD003156	Gel-free/label-free proteomic analysis of developing rice grains under heat stress	LTQ Orbitrap	Timabud *et al.* 2015 (39)	2015/12/17	185556

TABLE II
The overall counts of spectra, PSMs, percent of spectra identified, total identified peptides and total novel peptides per dataset

ProteomeXchange ID	Spectra Count	Total Identified Spectra (FDR 1%)	Total Spectra Identification Rate (FDR 1%)	Total Identified Peptides (FDR 1%)	Putative Novel Peptides (FDR 1%)
PXD000265	253743	83083	32.74%	21563	1500
PXD000313	976822	72516	7.42%	17802	1182
PXD000923	87502	16584	18.95%	6344	705
PXD001030	301027	29902	9.93%	7678	466
PXD001058	92162	24634	26.73%	9458	605
PXD002291	94400	16471	17.45%	2654	172
PXD002739	34344	9151	26.65%	2913	173
PXD002740	25862	10683	41.31%	2934	196
PXD003156	185556	158889	85.63%	8213	314
Total	2,051,418	421913	20.57%	47663	4168

All the *NEs* were matched back to their corresponding genome loci. Reads mapping to multiple locations were not filtered at this stage, but peptides mapping to multiple loci were handled explicitly (see below). The matched genomic fragments were translated in six reading frames. An accepted novel translation product from six reading frame translation was judged by two criteria, more than 5 amino acids (15 nucleotides) at least, and only the longest product being taken for a transcript. All the accepted novel translation products were added in to the list of the *Oryza sativa* proteins annotated from IRGSP-1.0.30, as well as contaminants from cRAP (http://www.thegpm.org/crap/), to generate a new protein database for MS/MS searching.

Peptide Search Based On MS/MS Data (IPeak search)—IPeak, a Java-based open source software package, was employed for peptide search, which uses the Percolator to re-score peptide-spectrum matches (PSMs) from MS-GF+ (v9733), MyriMatch (v2.2.8634) and X! Tandem (v2009.10.01.1). IPeak incorporates the FDRScore algorithm to combine the results from different search engines. All the MS/MS data collected from nine data sets were converted into MGF format using ProteoWizard (v3.0.4238) (24), then were searched with IPeak against the customized database, in which the minimum lengths of amino acids in sequences were no less than six. The PSMs with FDRScores less 0.01, corresponding to q-value (global FDR) < 0.01, were initially used to create the list of peptides identified. Most search parameters in the original publications associated with each of the nine data sets were used in the IPeak search. The exact parameters we used for the search are in SupplementaryFile1.xlsx (tab: "Search Parameters"). All the identified peptides through IPeak derived from the known rice proteins were marked as *known peptides*, whereas those not from those proteins were denoted as *putative novel peptides (PNPs*).

Mapping PNPs to Genomic Positions—The *PNPs* were mapped back to the genome to locate their positions on the chromosome by custom scripts to generate the GTF files, which record the genomic positions of the corresponding *NEs* of *PNPs*. The positional information of *PNPs* and *NEs* were imported into the original mzIdentML file using the proteogenomics encoding described in the mzIdentML version 1.2 specifications (25).

Filtration To Remove the PNPs With Potentially Incorrect Assignments Or Low Confidence—The *PNPs* obtained from the IPeak search were further filtered to remove the potential for alternative explanations of the corresponding spectra being more likely. PEAKS PTM (26) can identify many types of post-translational modifications (PTMs) and chemical modifications to peptides. To reduce the possibility of misidentification of novel peptides because of modifications (*e.g.* where a novel peptide has the same mass as a known peptide with a common modification), the MS/MS spectra that were matched to the novel peptides were re-searched with PEAKS PTM against the official annotation. For any novel peptide whose spectrum was confidently identified by PEAKS PTM as the modified or alternative form of a known peptide (q-value < 0.01), was flagged for removal from the novel peptide set.

To further check that PNP sequences could not be explained by other types of biochemical events (missed cleavage, proteolysis or single amino acid substitutions) on known peptides, all the *PNPs* were aligned by BLASTp and custom scripts (using a regular expression) against the known protein sequences (IRGSP-1.0.30). The BLASTp search was conducted in two modes, default and short sequence optimized, with results filtered allowing for a maximum of one mismatch, zero gap, and exact sequence length between query and hit. Any *PNPs* with a confident match by BLASTp (either mode) were excluded from the result set. To further increase the confidence of *PNPs*, the peptides with only a single spectrum support were also removed. After all these filtration steps, the remained *PNPs* were called the *final novel peptides (FNPs*).

FNPs were created initially compared with IRGSP-1.0.30 (*i.e.* absent from), however we also mapped them to the updated IRGSP-1.0.38 and to a different set of gene prediction from MSU (RGAP version 7, http://rice.plantbiology.msu.edu/) to determine if FNPs were present in those annotation sets.

Clustering of FNPs and NEs on the Genomic Landscape—The *FNPs* and their related *NEs* were parsed from mzIdentML into GFF format by custom scripts and were input to BEDTools (v2.25) to cluster the NEs onto the corresponding genomic loci. During clustering, 100bp was set as the maximum distance allowed between the *NEs*.

Homology Analysis of the FNPs With Other Plant Proteins—To provide further evidence supporting the annotation of *novel peptides*, the sequences of FNPs were aligned using BLASTp (short sequence optimized mode) to all the plant proteomes stored in Ensembl plants (release-38) with tolerance of only one residue mismatched or deleted in BLASTp. A more permissive analysis allowing for one gap and two mismatches is displayed in the supplementary results (SupplementaryResults.docx).

All the transcripts corresponding to peptides located within intergenic region, and thus potentially new genes, were first translated to amino acid sequences, then the translation products were searched with InterProScan to identify potential protein domains, using filtration of e-value $< 10^{-5}$.

Visualization of the NEs and FNPs—To visualize the *NEs* and *FNPs*, phpMs (27) was employed to transfer mzIdentML to proBed format, which is compatible with genome browsers supporting BED, such as IGV, UCSC, and Ensembl. To allow downstream comparisons, the two proBed files were combined from nine proteomics data sets: one contains all novel peptides identified and the other contains all

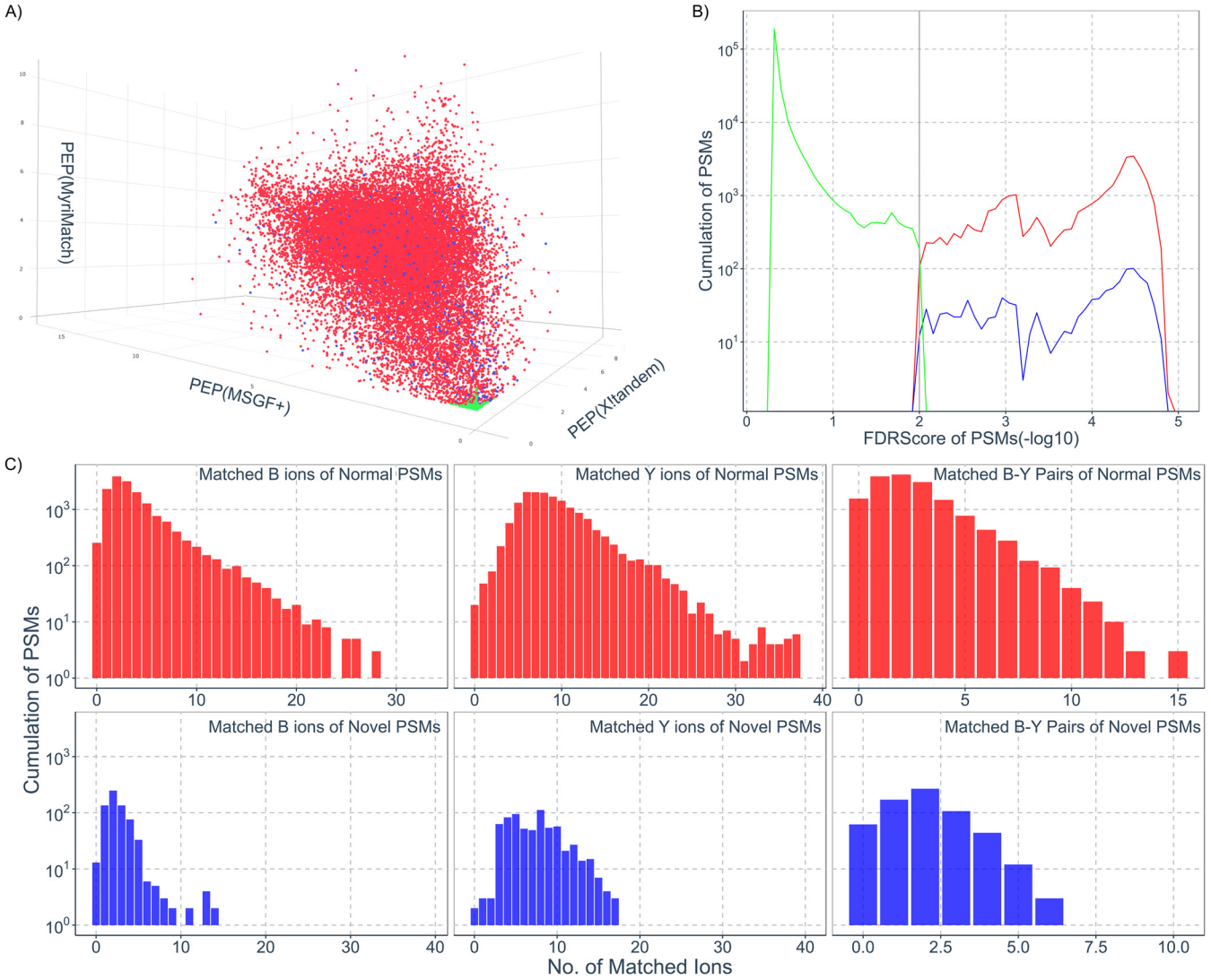


FIG. 2. **Evaluation of the novel peptides derived from MS/MS data.** *A*, The overall scoring integration of the novel peptides identified by three search engines. The axis in the three-dimensional scatter plot stand for -log 10 (Posterior Error Probability) values from X!Tandem, MS-GF+ and MyriMatch, respectively. The color points represent the normal peptides (red), the novel peptides (blue) and the excluded peptides including decoy peptides and peptides with FDRScore over 1% (green), respectively. *B*, The distribution of FDRScores for the peptides with density plots. The color curves have the same meaning as described in *A*. *C*, The distribution of matched ions for b, y and b-y pairs with bar plots. The colors represent the same meaning as described in *A* (ions from excluded peptides are not shown).

peptides (nonnovel) detected in the known *Oryza sativa* protein database. Track hubs are web-accessible directories of omics (genomic and proteomic in our project) data that can be viewed alongside official genome annotation through Genome Browser interfaces. Following the steps in https://genome.ucsc.edu/goldenpath/help/hgTrackHubHelp.html#Setup, one hub (containing two tracks) was generated, by converting the proBed files into BigBed format. The hub is hosted on our server and made publicly accessible via the track hub registry (http://trackhubregistry.org/ for "rice proteogenomic"). Each novel peptide and its cluster has its own link to Gramene/Ensembl browser that can be found in Supplementary File1.xlsx (tab "Final Novel Peptides").

Coding Potential Estimation—The coding potential (CP) of identified novel events clusters was estimated by CPAT (28). Rice known coding transcripts (42,132) and noncoding transcripts (total 53,250) from IRGSP-1.0.30 were used for training following the guide in (http://rna-cpat.sourceforge.net/). The three features, ORF size, Fickett score and Hexamer score, were taken to assess the training result. The training and thresholding for CPAT is exemplified in supplemental Figs. S4 and S5.

ProteomeXchange Submission—Raw files from all nine proteomic data sets and our result mzIdentML v1.2 files have been deposited to ProteomeXchange Consortium via the PRIDE (29) partner repository with the data set identifier PXD008960. Versions of the mzIdentML files containing only PNPs are also present in the PRIDE repository with the suffix "_removeA.mzid" for simple visualization in mzIdentML compatible software.

RESULTS

Defining the Novel Events Based On RNA-Seq Data—In IRGSP-1.0.30, there are 91,080 genes with 97,751 annotated

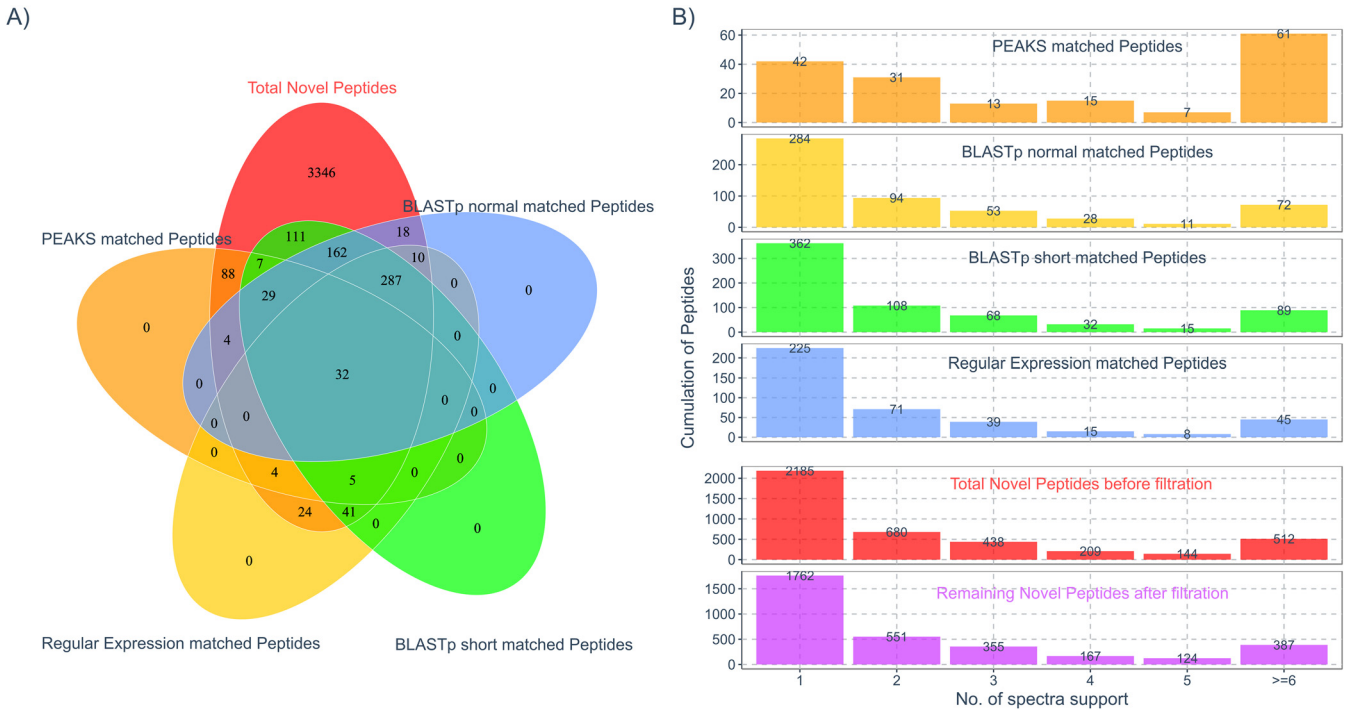


FIG. 3. **Multiple filtration to treat the putative novel peptides.** *A*, Venn diagram indicating the overlap results of multiple filtration schemes for the putative novel peptides. The filtration includes PEAKS PTM (orange), BLASTp short search (yellow), BLASTp normal search (green) and regular expression match (blue). After the filtration, a total of 3,346 of the putative novel peptides (red) remain. *B*, The spectrum/peptide distribution passing and removed by filtration. The upper panel (four blocks): the color bars indicate the distribution of spectra and peptides removed by the different filtration methods as in 3*A*. Lower panel (two blocks): comparison of the distribution of spectra and peptides for the total novel peptides with and without filtration.

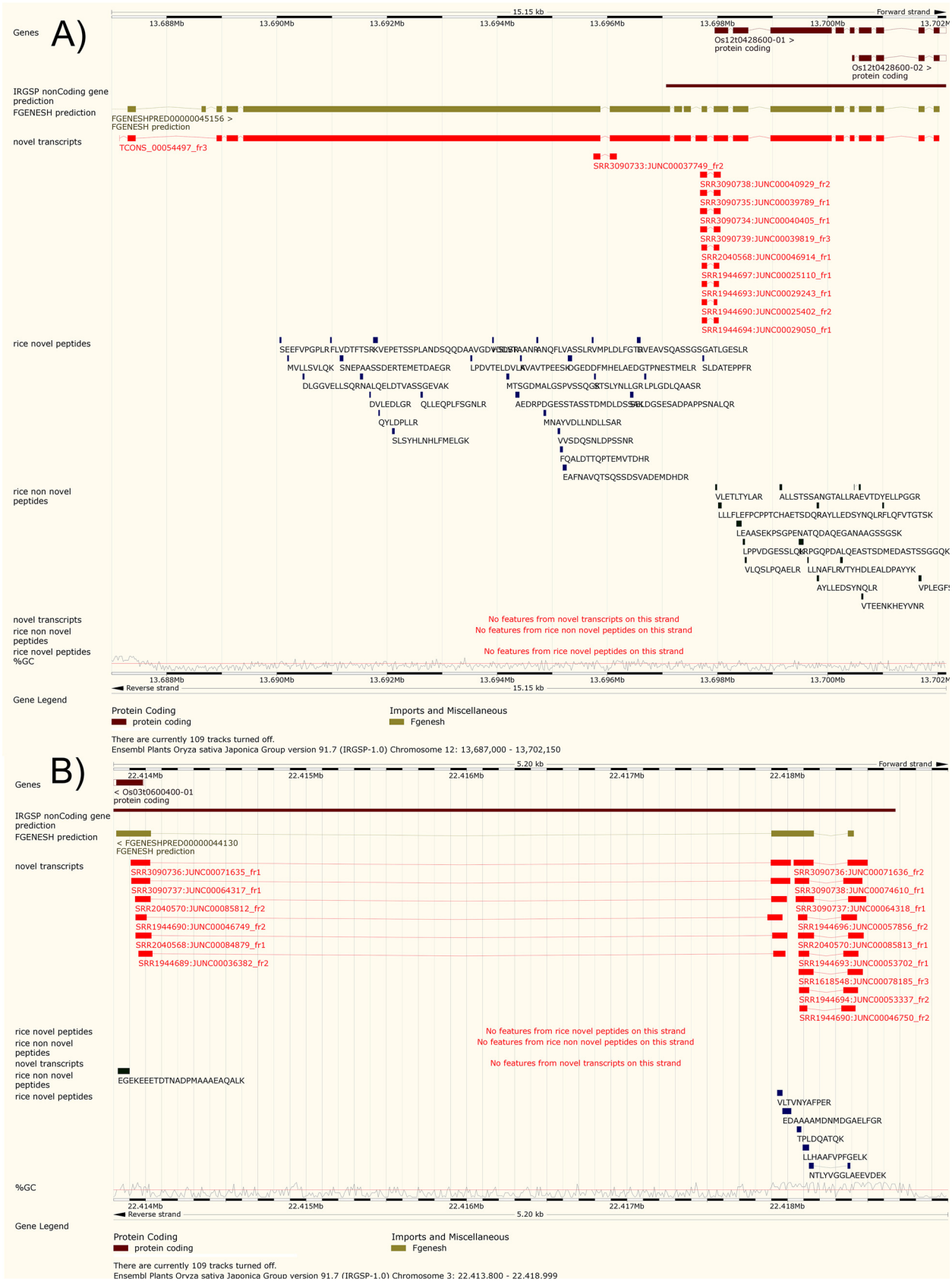
TABLE III
Classification of the novel peptides and novel events based on their type and localization

		Total	Intergenic	Intragenic
Novel peptides	Final Novel Peptides (FNPs)	1548	963	585
	Unique mapping Final Novel Peptides (Unique mapping FNPs)	1514	944	570
Novel Events	Novel transcripts candidates (NT)	530	69	461
	Novel junction candidates (NJ)	115	28	87

transcripts, including 35,679 protein coding genes, with 42,132 potentially protein coding transcripts. Within the annotation there are 150,594 annotated official junction sites of which over 98% of the sites (147,696) are contributed by the coding transcripts. The RNA-Seq data collected for this study from 29 runs exhibited an average mapping rate to the genome of 78.99%, with an average coverage to the annotated transcripts of 86.56% and to the (protein) coding transcripts of 97.48% (supplemental Table S2 and supplemental Fig. S1). High coverage of the RNA-Seq data demonstrated the customize database was a relatively comprehensive collection of coding production of *Oryza sativa*.

The RNA-Seq data was aligned to the genome by TopHat. As a result, a total of 2,940,788 junctions (355,323 junction sites) were identified. By comparing with the official junction sites, 1,047,488 junctions (56,432 nonredundant junction sites) in the RNA-Seq data exactly matched to those annotated junctions, whereas the remaining 1,893,300 junctions (298,891 junction sites) were marked as *NJs*. Details of all the junctions and the corresponding sites are shown in supplemental Table S3. The mapped reads were sent to assembly, leading to 201,360 transcripts constructed by Cuffmerge, in which 103,707 transcripts were exactly matched with annotated transcripts and the other 106,653 transcripts were regarded as *NTs*. Details of the assembled transcripts in each run are shown in supplemental Fig. S2. Thus, the *NE* set consists of 1,997,007 events.

A six frame translation of a genome sequence is a common method of looking for novel coding regions. However, a six-frame translation of the IRGSP-1.0.30 genome generates 25,761,390 small ORF candidates using the length cut-off of 6 or more amino acids, which would be an excessively large


A)
Genes
15.15 kb
Forward strand
13.688Mb
13.690Mb
13.692Mb
13.694Mb
13.696Mb
13.698Mb
13.700Mb
Os12t0428600-01 > protein coding
Os12t0428600-02 > protein coding
IRGSP nonCoding gene prediction
FGENESH prediction
FGENESHPRED00000045156 > FGENESH prediction
novel transcripts
TCONS_00054497_fr3
SRR3090733:JUNC00037749_fr2
SRR3090738:JUNC00040929_fr2
SRR3090735:JUNC00039789_fr1
SRR3090734:JUNC00040405_fr1
SRR3090739:JUNC00039819_fr3
SRR2040568:JUNC00046914_fr1
SRR1944697:JUNC00025110_fr1
SRR1944693:JUNC00029243_fr1
SRR1944690:JUNC00025402_fr2
SRR1944694:JUNC00029050_fr1
rice novel peptides
SEEFVPGPLR
MVLLSVLQK
SNEPAASSDERTEMETDAEGR
DLGGVELLSQRNALQELDTVASSGEVAK
DVLEDLGR
QLLEQPLFSGNLR
QYLDPLLR
SLSYHLNHLFMELGK
LPDVTELDVLK
AEDRPDGESSTASSTDMDLDSSEK
MNAYVDLLNDLLSAR
VVSDQSNLDPSSNR
FQALDTTQPTEMVTDHR
EAFNAVQTSQSSDSVADEMDHDR
rice non novel peptides
VLETLTYLAR
ALLSTSSANGTALLRAEVTDYELLPGGR
LLLFLEFPCPPTCHAETSDQRAYLLEDSYNQLRFLQFVTGTSK
LEAASEKPSGPENATQDAQEGANAAGSSGSK
VLQSLPQAELR
LLNAFLRVTYHDLEALDPAYYK
AYLLEDSYNQLR
VTEENKHEYVNR
novel transcripts
rice non novel peptides
rice novel peptides
No features from novel transcripts on this strand
No features from rice non novel peptides on this strand
No features from rice novel peptides on this strand
%GC
Reverse strand
Gene Legend
Protein Coding
protein coding
Imports and Miscellaneous
Fgenesh
There are currently 109 tracks turned off.
Ensembl Plants Oryza sativa Japonica Group version 91.7 (IRGSP-1.0) Chromosome 12: 13,687,000 - 13,702,150
B)
Genes
5.20 kb
Forward strand
22.414Mb
22.415Mb
22.416Mb
22.417Mb
22.418Mb
< Os03t0600400-01 protein coding
IRGSP nonCoding gene prediction
FGENESH prediction
< FGENESHPRED00000044130 FGENESH prediction
novel transcripts
SRR3090736:JUNC00071635_fr1
SRR3090737:JUNC00064317_fr1
SRR2040570:JUNC00085812_fr2
SRR1944690:JUNC00046749_fr2
SRR2040568:JUNC00084879_fr1
SRR1944689:JUNC00036382_fr2
SRR3090736:JUNC00071636_fr2
SRR3090738:JUNC00074610_fr1
SRR3090737:JUNC00064318_fr1
SRR1944696:JUNC00057856_fr2
SRR2040570:JUNC00085813_fr1
SRR1944693:JUNC00053702_fr1
SRR1618548:JUNC00078185_fr3
SRR1944694:JUNC00053337_fr2
SRR1944690:JUNC00046750_fr2
rice novel peptides
rice non novel peptides
novel transcripts
rice non novel peptides
rice novel peptides
No features from rice novel peptides on this strand
No features from rice non novel peptides on this strand
No features from novel transcripts on this strand
EGEKEEETDTNADPMAAAEAQALK
VLTVNYAFPER
EDAAAAMDNMDGAELFGR
TPLDQATQK
LLHAAFVPFGELK
NTLYVGGLAEEVDEK
%GC
Reverse strand
Gene Legend
Protein Coding
protein coding
Imports and Miscellaneous
Fgenesh
There are currently 109 tracks turned off.
Ensembl Plants Oryza sativa Japonica Group version 91.7 (IRGSP-1.0) Chromosome 3: 22,413,800 - 22,418,999

search space, leading to reduced statistical power. On the bases of the customized database derived from RNA-Seq data described above, only about 1/26 small ORF candidates (970, 204) were generated with the same length cut-off, 869,872 from *novel junctions* and 100,332 from *novel transcripts*. Further comparison of the theoretical digestion with trypsin (without missed cleavage), generates 31,088,508 peptides in a standard six frame translation database compared with 2,045,553 in the customized database, the search space for the latter being only 1/15 of the former. As such, we can conclude that the assembled customized database should allow for a much higher statistical power than a standard six frame translation.

Identification of the Novel Peptides Based On the Collected MS/MS Data Sets—Through IPeak search at 1% PSM FDR against a total 9 ProteomeXchange data sets, 421,913 peptide spectra matches (PSMs) were found, corresponding to 47,663 peptides identified. The detailed statistics for the identified spectra and peptide are shown in Table II. The PSM rates in these data sets were diverse, with generally higher rates in the data sets from enrichment studies, whereas lower rates in the data sets from profiling studies. Of the matched peptides, 43,495 peptides were derived from proteins in the IRGSP-1.0.30 annotation assigned as known peptides, corresponding to 8,187 protein groups, whereas the other 4168 peptides were only mapped to the NEs, then termed as putative novel peptides (PNPs). The details of PNPs are listed in supplementary File S1 (tab "Putative Novel Peptides"). The full results for both novel and nonnovel (matching canonical peptides) are included as supplementary data in proBed format (23), which as a tab-separated file can be simply visualized in spreadsheet software, as well as being loadable on genome browsers.

The PSMs for *PNPs* and known peptides we refer to as nPSMs and kPSMs, respectively. Although we took the global FDR to control the false discovery rate of total identified PSMs, the nPSMs might suffer from a higher false positive rate than the global FDR estimated because of the method of database construction (30). To confirm the confidence of the nPSMs, we used three search engines, X!Tandem, MS-GF+ and MyriMatch, and attempted to address whether the distribution of Posterior Error Probabilities (PEP) for kPSMs and nPSMs was different from that generated from the decoy database. As shown in Fig. 2*A*, the PEPs of kPSMs and nPSMs are evenly distributed along the three axes with the similar patterns and are scattered away from the origin, whereas the distribution of PEPs of decoy PSMs is completely different from that of kPSMs and nPSMs, and is narrowly located around the origin. Using the FDRScore to integrate search engine scores, Fig. 2*B* reveals that the distribution of -log (FDRScore) of the nPSMs is comparable with that of kPSMs, but is significantly different from that of decoy PSMs. The matched ions, including b ions, y ions and b-y ion pairs, of nPSMs are compared with those from kPSMs as presented in Fig. 2*C*, calculated using the PDV software (31). Although the right tail of the distribution of matched ions and ion pairs of kPSMs is somewhat higher than that of kPSMs, the distribution apex and shape are comparable. Overall, the evaluation illustrated in Fig. 2 provides support that the MS/MS quality of the nPSMs are similar to kPSMs.

Filtration to Remove the PNPs With Potentially Incorrect Assignments Or Low Confidence—As mentioned under Experimental Procedures, we filtered PNPs by searching their spectra with PEAKS PTM to search for evidence of known peptides with modifications or amino acid substitutions potentially explaining the same spectra. The 14,164 spectra corresponding to 4168 PNPs were searched with PEAKS PTM, resulting in 740 spectra matching 169 peptides from canonical rice proteins with single amino acid mutation or common modifications at 1% FDR (supplementary File S1, tab "PEAKS PTM Identifications"). To further ensure the PNP sequences could not have been derived from known proteins, we performed several filtration processes, regular expression matching, standard BLASTp and short sequence optimized BLASTp. First, a regular expression matching was conducted using custom scripts. A total of 403 peptides were matched by regular expression matching, including 397 peptides matched with a peptide in the official database generated without tryptic termini. Second, a BLASTp search with at most one mismatch was conducted to remove the peptides with potential single amino acid variations. Using standard BLASTp and short sequence optimized BLASTp, 542 and 674 peptides were matched. Combining all the filtration results, a total 904 distinct peptides were filtered out, which were matched by at least one of the filtration conditions. Fig. 3*A* contains a Venn diagram displaying the detailed information regarding the filtration results. Fig. 3*B* displays the peptides filtered in relation to the number of spectra supporting the identification, showing in most cases that there were few differences across different matched sets. One observation from Fig. 3*B* is that the PEAKS PTM filter remove a relatively large proportion of peptides with >6 spectra support, likely indicating that these were otherwise confidently identified abundant peptides. For Fig. 3*A* and 3*B*, we conclude that the

FIG. 4. **Genomic landscapes of the two typical NEs found in this study.** *A*, Both evidence types, transcriptomics and proteomics, support the encoding regions located upstream of the of the Os12t0428600 gene. A screen shot is obtained from Gramene, showing the genomic information of position 13687144–13702026 on rice chromosome 12. The track names on the left illustrate evidence from different sources, including predicted genes, NEs and novel peptides. On the right, the genomic landscape for the gene Os12t0428600 from a theoretical prediction and official annotation, and the location of the transcripts and peptides identified on the same gene based on the official annotation and the NEs. *B*, A screen shot is obtained from Gramene, showing the genomics information of 22413908–22418499 on rice chromosome 3. The three novel peptides and especially the junction peptide (NTLYVGGLAEEVDEK) suggest a novel encoding region and confirmed its junction site.

different filtration mechanisms are complementary, because no single method was able to identify all possible causes of potential incorrect assignment.

After these filtrations, we found a large portion of the PNPs still supported by only a single spectrum (bottom panel of Fig. 3*B*). To increase the confidence of PNPs, we adopted a more stringent criterion that an identified peptide must be supported by at least two MS/MS spectra. Thus, a total 1762 peptide were further removed, leaving 1584 peptides passing all filters, and thus marked Final Novel Peptides (*FNPs*, see supplementary File S1, "Final Novel Peptides" tab and "FNP Detail" tab).

Clustering of FNPs and NEs On the Genomic Landscape—FNPs and their corresponding NEs were mapped back to genomic coordinates. Compared with IRGSP-1.0.30, 68 FNPs were mapped to multiple genome loci, whereas the other 1514 were mapped to unique genome loci, called unique mapping FNPs. Among the 1514 unique mapping FNPs, 944 novel peptides were in intergenic regions, whereas the other 570 novel peptides are located in intragenic regions *e.g.* within introns, different frames or different splices of existing genes. However, the relatively large number of intergenic peptides does not necessarily indicate entirely new genes being discovered, because the matching transcripts might be located close to existing genes, and thus be additional or alternative exons. This point is further addressed by the clustering process described below. The details of the FNPs mapped to the rice genome are summarized in Table III. In addition, 682 out of 1514 unique mapping FNPs span multiple exons potential junction sites. A further alignment of FNPs to the latest IRGSP and MSU Rice Genome Annotation implied that our analysis for discovery of novel peptides is valuable, 62 FNPs have been independently confirmed as protein-coding by the up-to-date annotation in IRGSP-1.0.38, whereas 286 FNPs have also since been independently mapped to nonputative protein sequences of the MSU Rice Genome Annotation version 7.0 (supplementaryfile1.xlsx tab: "FNP mapping MSU").

For the sake of tracing all the FNPs back to rice genome, we first clustered the NEs based on their genomic loci, because the NJs and NTs generated from different algorithms might have some overlapping sites or regions on the genome. We then localized all the FNPs to these correspondingly clustered NE regions. A total 686 clusters are found on the genome, of which 645 contain at least one unique mapping FNP. The 645 clusters are further divided to two groups according to whether a cluster contains at least one novel transcript or not, the former termed as novel transcript clusters (NT clusters) totalling 530 clusters, and the others denoted as novel junction clusters (NJ clusters) totalling 115 clusters. In the NT clusters, the novel transcripts suggest the putative gene models of the region, and the novel junction are considered as a complementary part of the gene model. If an NT cluster is mapped to an intergenic region, it would be considered as a novel gene candidate, whereas if it is overlapped with an intragenic region, it would be denoted as a complementary to a known gene. Once the 530 NT clusters were mapped onto IRGSP-1.0.30, 69 NT clusters were mapped in the intergenic regions, whereas the other 461 are matched in the intragenic region. The NJ clusters (*i.e.* with only novel junctions) suggest an incomplete gene model, indicated as putative splicing sites of an underlying model. Similarly, they were mapped onto IRGSP-1.0.30 as well, resulting in 28 and 87 NJ clusters localized in the intergenic and intragenic region, respectively. The details of the NT clusters and NJ clusters mapped to rice genome are summarized in Table III, and the two typical examples of NT and NJ cluster are depicted in Fig. 4*A* and 4*B*.

Evaluation Toward the NEs With the Evidence of FNPs—After the FNPs and their corresponding NEs had been clustered, we compared the NEs supported by FNPs with coding and noncoding genes from IRGSP-1.0.30. First, the length statistics of the NT clusters was estimated, *i.e.* selecting the longest transcript from all the same clustered transcripts. As shown in a violin-plot of transcript length in Fig. 5*A*, the length distribution of the coding and noncoding transcripts in IRGSP-1.0.30 is substantially different from each other. The distribution of NEs is also like that of coding transcripts but very distinct from noncoding transcripts. Second, the junction sites (nucleotide pairs on either site of the site junction) mapped from FNPs were compared with known junction sites from IRGSP-1.0.30 for coding and noncoding genes (supplemental Fig. S6). The distributions of junction site distributions for NJs was near identical to junctions in the official annotation, and substantially different to the distribution of splice junction sites for noncoding junctions.

The CPAT pipeline is a well-accepted method for estimating the coding potential of transcripts. We used CPAT to evaluate all the NTs for their coding potential. As shown in Fig. 5*B*, the evaluation suggests that 86.42% of the NTs supported by FNPs (530) have fully confident CPs (scored as 100% by CP), whereas less than 1% of that NTs have CPs below the CP coding threshold (see supplemental Figs. S4 and S5 for training data). The analysis of NTs without evidence from FNPs (97123), shows that only 50% of transcripts are scored as coding with full confidence (CP = 100%). We thus demonstrate a strong enrichment for transcripts supported by FNPs to be coding.

Support for Novel Peptides in Other Annotations—We performed a BLASTp ("BLASTp short") search of novel peptides against proteomes from Ensembl plants to detect if these regions had been predicted as genes in other species or varieties of rice (Supplementary File 1, tab "BLASTp final novel peptides"). Out of 1,584 FNPs, only 321 are not matched to other species following stringent thresholding, with 1,263 having at least one hit to another species, indicating that the majority of FNPs are annotated at least once in another species. A heat map and distribution of hits to the

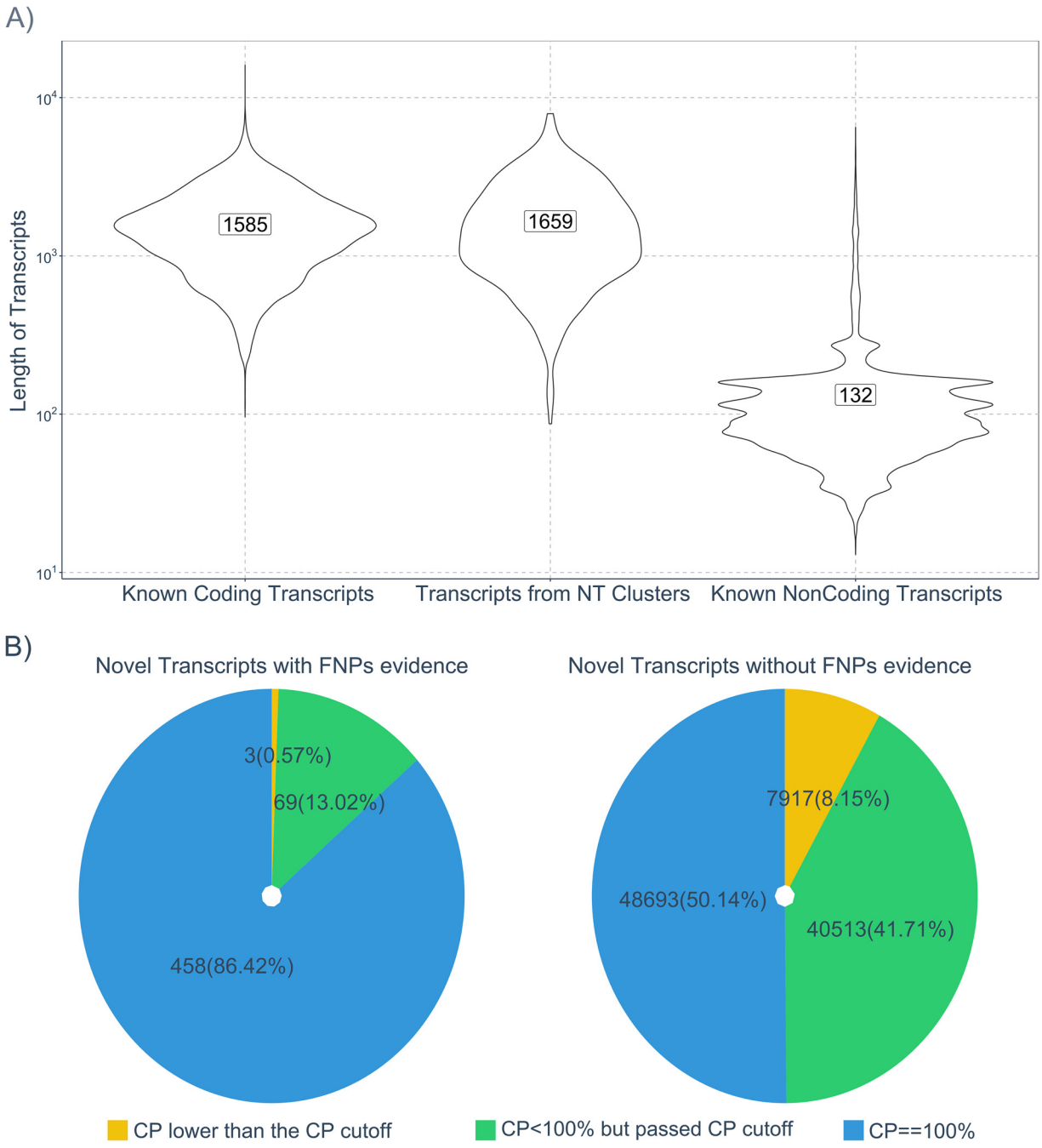


FIG. 5. **Evaluation to the NEs with the evidence of FNPs.** *A*, The length distribution of novel transcripts from NEs with evidence of FNPs (left) compared with coding (middle) and noncoding transcripts (right) from IRGSP-1.0.30. *B*, The coding potential (CP) of novel transcripts from NEs with evidence of FNPs and without evidence of FNPs.

species is displayed in Fig. 6. The novel peptides could mainly be found in *poaceae* plant (grasses), especially varieties of rice *oryza* and related species. As the peptides' length increase, the matches to distant species are seen less often, presumably because of the stringent nature of the thresholding we are applying (only one mismatch or gap allowed). A more permissive analysis is displayed in the supplemental Fig. S7 (allowing two mismatches or one gap) showing a higher proportion of longer peptides mapped to other species.

To further support that NEs supported by FNPs are protein-coding, we performed InterProScan analysis on the translation sequences of 101 intergenic transcripts from 69 NT clusters (full details in supplementary File S1, tab "InterProScan summary". Out of 101 sequences, 43 sequences had a significant (e-value $<$ E-05) match to an existing

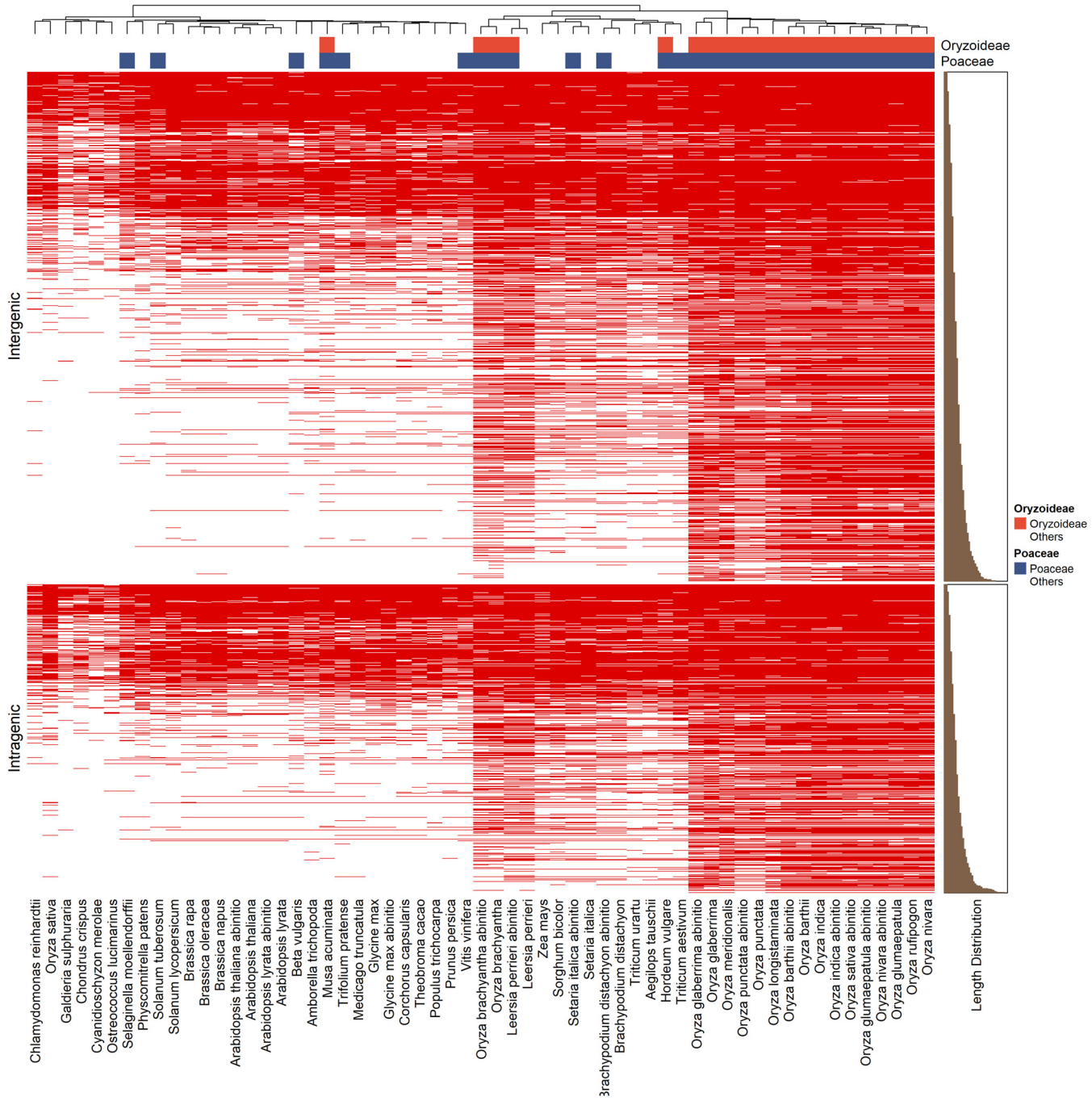


FIG. 6. **Identification of novel peptides in annotations from other plants.** The heatmap represents hierarchical analysis of the final novel peptides mapped against the proteins encoded by the 44 plant genomes from Ensembl (red = positive match, white = no match from BLASTp, allowing no gaps and one mismatch). The novel peptides are divided into two groups, intergenic (upper panel) and intragenic (lower panel), and are ranked by peptide length for hierarchical analysis.

protein domain or family signature, indicating strong evidence that these are indeed missed genes in the annotation. We found evidence for new genes with functions including DNA repair, hydrolase, urease, thiolase-like, ATP synthase and proteinase inhibitor.

Visualization of Results and Public Availability—To support on-going annotation efforts, we have made results available as permanent Track Hubs. In supplementary File S1 (tab "Final Novel Peptides"), there is a link from each cluster supported by one or more novel peptides to a visualization of the corresponding region viewed on the Ensembl genome. There are instructions in the supplemental material (supplemental Fig. S3) for how to configure the tracks to make them visible, which must be followed first.

CONCLUSIONS

We have performed a rigorous proteogenomics analysis on rice, to support future efforts to improve the annotation of the genome. Our results have been openly released in standard formats, which can be easily viewed directly as tracks on genome browsers. As well as providing confirmatory evidence for over 8000 genes as being protein coding, we have strong evidence that the annotations can be improved for >600 loci. A total of 101 loci were identified in intergenic regions with strong peptide support, of which 43 had a positive prediction of a functional domain, indicating new genes that can be added to the rice genome. Because the results are made persistently available as genomic tracks, we anticipate that future curation efforts will use these data sets for determining the protein-coding gene set of rice.

DATA AVAILABILITY

Data are available via ProteomeXchange/PRIDE with identifier PXD008960. The data are also available as permanent TrackHubs, visit http://trackhubregistry.org/ and search for "rice proteogenomic", to display data on the Ensembl plants or Gramene genome browsers.

* This work was supported by Biotechnology and Biological Sciences Research Council (BBSRC) [BB/N013743/1, BB/L024128/1], the Ministry of Science and Technology of the People's Republic of China (2011DFA33220), the International Science & Technology Cooperation Program of China (2014DFB30020) and National Key Basic Research Program of China (2014CBA02002, 2014CBA02005).

[S] This article contains supplemental material.

** To whom correspondence may be addressed. E-mail: siqiliu@genomics.cn.

†† To whom correspondence may be addressed. E-mail: andrew.jones@liverpool.ac.uk.

§§ Joint first author.

The author(s) declare(s) that they have no competing interests.

Author contributions: Z.R., D.Q., B.W., S.L., and A.R.J. designed research; Z.R. and D.Q. performed research; Z.R. and D.Q. contributed new reagents/analytic tools; Z.R., D.Q., P.N., K.L., B.W., R.Z., S.X., S.L., and A.R.J. analyzed data; Z.R., D.Q., S.L., and A.R.J. wrote the paper.

REFERENCES

1. Nesvizhskii, A. I. (2014) Proteogenomics: concepts, applications and computational strategies. *Nat. Meth.* **11,** 1114–1125
2. Castellana, N. E., Payne, S. H., Shen, Z., Stanke, M., Bafna, V., and Briggs, S. P. (2008) Discovery and revision of Arabidopsis genes by proteogenomics. *Proc. Natl. Acad. Sci. U.S.A.* **105,** 21034–21038
3. Castellana, N. E., Shen, Z., He, Y., Walley, J. W., Cassidy, C. J., Briggs, S. P., and Bafna, V. (2014) An automated proteogenomic method uses mass spectrometry to reveal novel genes in Zea mays. *Mol. Cell Proteomics* **13,** 157–167
4. International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. *Nature* **436,** 793–800
5. Goff, S. A., Ricke, D., Lan, T. H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H., Hadley, D., Hutchison, D., Martin, C., Katagiri, F., Lange, B. M., Moughamer, T., Xia, Y., Budworth, P., Zhong, J., Miguel, T., Paszkowski, U., Zhang, S., Colbert, M., Sun, W. L., Chen, L., Cooper, B., Park, S., Wood, T. C., Mao, L., Quail, P., Wing, R., Dean, R., Yu, Y., Zharkikh, A., Shen, R., Sahasrabudhe, S., Thomas, A., Cannings, R., Gutin, A., Pruss, D., Reid, J., Tavtigian, S., Mitchell, J., Eldredge, G., Scholl, T., Miller, R. M., Bhatnagar, S., Adey, N., Rubano, T., Tusneem, N., Robinson, R., Feldhaus, J., Macalma, T., Oliphant, A., and Briggs, S. (2002) A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). *Science* **296,** 92–100
6. Yu, J., Hu, S., Wang, J., Wong, G. K., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X., Cao, M., Liu, J., Sun, J., Tang, J., Chen, Y., Huang, X., Lin, W., Ye, C., Tong, W., Cong, L., Geng, J., Han, Y., Li, L., Li, W., Hu, G., Huang, X., Li, W., Li, J., Liu, Z., Li, L., Liu, J., Qi, Q., Liu, J., Li, L., Li, T., Wang, X., Lu, H., Wu, T., Zhu, M., Ni, P., Han, H., Dong, W., Ren, X., Feng, X., Cui, P., Li, X., Wang, H., Xu, X., Zhai, W., Xu, Z., Zhang, J., He, S., Zhang, J., Xu, J., Zhang, K., Zheng, X., Dong, J., Zeng, W., Tao, L., Ye, J., Tan, J., Ren, X., Chen, X., He, J., Liu, D., Tian, W., Tian, C., Xia, H., Bao, Q., Li, G., Gao, H., Cao, T., Wang, J., Zhao, W., Li, P., Chen, W., Wang, X., Zhang, Y., Hu, J., Wang, J., Liu, S., Yang, J., Zhang, G., Xiong, Y., Li, Z., Mao, L., Zhou, C., Zhu, Z., Chen, R., Hao, B., Zheng, W., Chen, S., Guo, W., Li, G., Liu, S., Tao, M., Wang, J., Zhu, L., Yuan, L., and Yang, H. (2002) A draft sequence of the rice genome (Oryza sativa L. ssp. indica). *Science* **296,** 79–92
7. Li, J. Y., Wang, J., and Zeigler, R. S. (2014) The 3,000 rice genomes project: new opportunities and challenges for future rice research. *Gigascience* **3,** 8
8. 3,000 Rice Genomes Project. (2014) The 3,000 rice genomes project. *Gigascience* **3,** 7
9. Rice Annotation, P., Tanaka, T., Antonio, B. A., Kikuchi, S., Matsumoto, T., Nagamura, Y., Numa, H., Sakai, H., Wu, J., Itoh, T., Sasaki, T., Aono, R., Fujii, Y., Habara, T., Harada, E., Kanno, M., Kawahara, Y., Kawashima, H., Kubooka, H., Matsuya, A., Nakaoka, H., Saichi, N., Sanbonmatsu, R., Sato, Y., Shinso, Y., Suzuki, M., Takeda, J., Tanino, M., Todokoro, F., Yamaguchi, K., Yamamoto, N., Yamasaki, C., Imanishi, T., Okido, T., Tada, M., Ikeo, K., Tateno, Y., Gojobori, T., Lin, Y. C., Wei, F. J., Hsing, Y. I., Zhao, Q., Han, B., Kramer, M. R., McCombie, R. W., Lonsdale, D., O'Donovan, C. C., Whitfield, E. J., Apweiler, R., Koyanagi, K. O., Khurana, J. P., Raghuvanshi, S., Singh, N. K., Tyagi, A. K., Haberer, G., Fujisawa, M., Hosokawa, S., Ito, Y., Ikawa, H., Shibata, M., Yamamoto, M., Bruskiewich, R. M., Hoen, D. R., Bureau, T. E., Namiki, N., Ohyanagi, H., Sakai, Y., Nobushima, S., Sakata, K., Barrero, R. A., Sato, Y., Souvorov, A., Smith-White, B., Tatusova, T., An, S., An, G., S. O. O., Fuks, G., Fuks, G., Messing, J., Christie, K. R., Lieberherr, D., Kim, H., Zuccolo, A., Wing, R. A., Nobuta, K., Green, P. J., Lu, C., Meyers, B. C., Chaparro, C., Piegu, B., Panaud, O., and Echeverria, M. (2008) The Rice Annotation Project Database (RAP-DB): 2008 update. *Nucleic Acids Res.* **36,** D1028–D1033
10. Zhang, G., Guo, G., Hu, X., Zhang, Y., Li, Q., Li, R., Zhuang, R., Lu, Z., He, Z., Fang, X., Chen, L., Tian, W., Tao, Y., Kristiansen, K., Zhang, X., Li, S., Yang, H., Wang, J., and Wang, J. (2010) Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. *Genome Res.* **20,** 646–654
11. Lu, T., Lu, G., Fan, D., Zhu, C., Li, W., Zhao, Q., Feng, Q., Zhao, Y., Guo, Y., Li, W., Huang, X., and Han, B. (2010) Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq. *Genome Res.* **20,** 1238–1249
12. Kawahara, Y., de la Bastide, M., Hamilton, J. P., Kanamori, H., McCombie, W. R., Ouyang, S., Schwartz, D. C., Tanaka, T., Wu, J., Zhou, S., Childs, K. L., Davidson, R. M., Lin, H., Quesada-Ocampo, L., Vaillancourt, B., Sakai, H., Lee, S. S., Kim, J., Numa, H., Itoh, T., Buell, C. R., and Matsumoto, T. (2013) Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. *Rice* **6,** 4
13. Helmy, M., Tomita, M., and Ishihama, Y. (2011) OryzaPG-DB: Rice Proteome Database based on Shotgun Proteogenomics. *BMC Plant Biol.* **11,** 63
14. Burset, M., Seledtsov, I. A., and Solovyev, V. V. (2001) SpliceDB: database of canonical and noncanonical mammalian splice sites. *Nucleic Acids Res.* **29,** 255–259
15. Fermin, D., Allen, B., Blackwell, T., Menon, R., Adamski, M., Xu, Y., Ulintz, P., Omenn, G., and States, D. (2006) Novel gene and gene model detection using a whole genome open reading frame analysis in proteomics. *Genome Biol.* **7,** R35
16. Khatun, J., Yu, Y., Wrobel, J. A., Risk, B. A., Gunawardena, H. P., Secrest, A., Spitzer, W. J., Xie, L., Wang, L., Chen, X., and Giddings, M. C. (2013) Whole human genome proteogenomic mapping for ENCODE cell line data: identifying protein-coding regions. *BMC Genomics* **14,** 141

17. Brosch, M., Saunders, G. I., Frankish, A., Collins, M. O., Yu, L., Wright, J., Verstraten, R., Adams, D. J., Harrow, J., Choudhary, J. S., and Hubbard, T. (2011) Shotgun proteomics aids discovery of novel protein-coding genes, alternative splicing, and "resurrected" pseudogenes in the mouse genome. *Genome Res.* **21,** 756–767
18. Wen, B., Du, C., Li, G., Ghali, F., Jones, A. R., Kall, L., Xu, S., Zhou, R., Ren, Z., Feng, Q., Xu, X., and Wang, J. (2015) IPeak: An open source tool to combine results from multiple MS/MS search engines. *Proteomics* **15,** 2916–2920
19. Jones, A. R., Siepen, J. A., Hubbard, S. J., and Paton, N. W. (2009) Improving sensitivity in proteome studies by analysis of false discovery rates for multiple search engines. *Proteomics* **9,** 1220–1229
20. Ghali, F., Krishna, R., Lukasse, P., Martinez-Bartolome, S., Reisinger, F., Hermjakob, H., Vizcaino, J. A., and Jones, A. R. (2013) A toolkit for the mzIdentML standard: the ProteoIDViewer, the mzidLibrary and the mzidValidator. *Mol. Cell Proteomics*, mcp.O113.029777
21. Ghali, F., Krishna, R., Perkins, S., Collins, A., Xia, D., Wastling, J., and Jones, A. R. (2014) ProteoAnnotator – Open source proteogenomics annotation software supporting PSI standards. *Proteomics* **14,** 2731–2741
22. Jones, A. R., Eisenacher, M., Mayer, G., Kohlbacher, O., Siepen, J., Hubbard, S., Selley, J., Searle, B., Shofstahl, J., Seymour, S., Julian, R., Binz, P.-A., Deutsch, E. W., Hermjakob, H., Reisinger, F., Griss, J., Vizcaino, J. A., Chambers, M., Pizarro, A., and Creasy, D. (2012) The mzIdentML data standard for mass spectrometry-based proteomics results. *Mol. Cell. Proteomics* **11,** M111.014381
23. Menschaert, G., Wang, X., Jones, A. R., Ghali, F., Fenyö, D., Olexiouk, V., Zhang, B., Deutsch, E. W., Ternent, T., and Vizcaíno, J. A. (2018) The proBAM and proBed standard formats: enabling a seamless integration of genomics and proteomics data. *Genome Biol.* **19,** 12
24. Chambers, M. C., Maclean, B., Burke, R., Amodei, D., Ruderman, D. L., Neumann, S., Gatto, L., Fischer, B., Pratt, B., Egertson, J., Hoff, K., Kessner, D., Tasman, N., Shulman, N., Frewen, B., Baker, T. A., Brusniak, M.-Y., Paulse, C., Creasy, D., Flashner, L., Kani, K., Moulding, C., Seymour, S. L., Nuwaysir, L. M., Lefebvre, B., Kuhlmann, F., Roark, J., Rainer, P., Detlev, S., Hemenway, T., Huhmer, A., Langridge, J., Connolly, B., Chadick, T., Holly, K., Eckels, J., Deutsch, E. W., Moritz, R. L., Katz, J. E., Agus, D. B., MacCoss, M., Tabb, D. L., and Mallick, P. (2012) A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotech.* **30,** 918–920
25. Vizcaino, J. A., Mayer, G., Perkins, S. R., Barsnes, H., Vaudel, M., Perez-Riverol, Y., Ternent, T., Uszkoreit, J., Eisenacher, M., Fischer, L., Rappsilber, J., Netz, E., Walzer, M., Kohlbacher, O., Leitner, A., Chalkley, R. J., Ghali, F., Martinez-Bartolome, S., Deutsch, E. W., and Jones, A. R. (2017) The mzIdentML data standard version 1.2, supporting advances in proteome informatics. *Mol. Cell Proteomics* **16,** 1275–1285
26. Han, X., He, L., Xin, L., Shan, B., and Ma, B. (2011) PeaksPTM: Mass Spectrometry-Based Identification of Peptides with Unspecified Modifications. *J. Proteome Res.* **10,** 2930–2936
27. Collins, A., and Jones, A. R. (2018) phpMs: A PHP-Based Mass Spectrometry Utilities Library. *J. Proteome Res.* **17,** 1309–1313
28. Wang, L., Park, H. J., Dasari, S., Wang, S., Kocher, J. P., and Li, W. (2013) CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res.* **41,** e74
29. Vizcaino, J. A., Csordas, A., Del-Toro, N., Dianes, J. A., Griss, J., Lavidas, I., Mayer, G., Perez-Riverol, Y., Reisinger, F., Ternent, T., Xu, Q. W., Wang, R., and Hermjakob, H. (2016) 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* **44,** 11033
30. Krug, K., Carpy, A., Behrends, G., Matic, K., Soares, N. C., and Macek, B. (2013) Deep coverage of the Escherichia coli proteome enables the assessment of false discovery rates in simple proteogenomic experiments. *Mol. Cell Proteomics* **12,** 3420–3430
31. Li, K., Vaudel, M., Zhang, B., Ren, Y., and Wen, B. (2018) PDV: an integrative proteomics data viewer. *Bioinformatics* submitted, doi: 10.1093/bioinformatics/bty770
32. Abiko, M., Furuta, K., Yamauchi, Y., Fujita, C., Taoka, M., Isobe, T., and Okamoto, T. (2013) Identification of proteins enriched in rice egg or sperm cells by single-cell proteomics. *PloS One* **8,** e69578
33. Zi, J., Zhang, J., Wang, Q., Zhou, B., Zhong, J., Zhang, C., Qiu, X., Wen, B., Zhang, S., Fu, X., Lin, L., and Liu, S. (2013) Stress responsive proteins are actively regulated during rice (Oryza sativa) embryogenesis as indicated by quantitative proteomics analysis. *PLOS ONE* **8,** e74229
34. Wang, K., Zhao, Y., Li, M., Gao, F., Yang, M. K., Wang, X., Li, S., and Yang, P. (2014) Analysis of phosphoproteome in rice pistil. *Proteomics* **14,** 2319–2334
35. Lin, Z., Zhang, X., Yang, X., Li, G., Tang, S., Wang, S., Ding, Y., and Liu, Z. (2014) Proteomic analysis of proteins related to rice grain chalkiness using iTRAQ and a novel comparison system based on a notched-belly mutant with white-belly. *BMC Plant Biol.* **14,** 163–163
36. Collado-Romero, M., Alós, E., and Prieto, P. (2014) Unravelling the proteomic profile of rice meiocytes during early meiosis. *Frontiers Plant Sci.* **5,** 356
37. Xiong, Y., Peng, X., Cheng, Z., Liu, W., and Wang, G. L. (2016) A comprehensive catalog of the lysine-acetylation targets in rice (Oryza sativa) based on proteomic analyses. *J. Proteomics* **138,** 20–29
38. He, D., Wang, Q., Li, M., Damaris, R. N., Yi, X., Cheng, Z., and Yang, P. (2016) Global Proteome Analyses of Lysine Acetylation and Succinylation Reveal the Widespread Involvement of both Modification in Metabolism in the Embryo of Germinating Rice Seed. *J. Proteome Res.* **15,** 879–890
39. Timabud, T., Yin, X., Pongdontri, P., and Komatsu, S. (2016) Gel-free/label-free proteomic analysis of developing rice grains under heat stress. *J. Proteomics* **133,** 1–19