Genes to Cells

Mass spectrum sequential subtraction speeds up searching large peptide MS/MS spectra datasets against large nucleotide databases for proteogenomics

Mohamed Helmy[1,2], Naoyuki Sugiyama[1], Masaru Tomita[1] and Yasushi Ishihama[1,3]*

[1]*Institute for Advanced Biosciences, Keio University, Tsuruoka, Yamagata 997-0017, Japan*
[2]*Systems Biology Program, Graduate School of Media and Governance, Keio University, Fujisawa, Kanagawa 252-0882, Japan*
[3]*Graduate School of Pharmaceutical Sciences, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan*

We have developed a novel bioinformatics method called mass spectrum sequential subtraction (MSSS) to search large peptide spectra datasets produced by liquid chromatography/mass spectrometry (LC-MS/MS) against protein and large-sized nucleotide sequence databases. The main principle in MSSS is to search the peptide spectra set against the protein database, followed by removal of the spectra corresponding to the identified peptides to create a smaller set of the remaining peptide spectra for searching against the nucleotide sequences database. Therefore, we reduce the number of spectra to be searched to limit the peptide search space. Comparing MSSS and conventional search approach using a dataset of 27 LC-MS/MS runs of rice culture cells indicated that MSSS reduced the search queries to 50% and the search time to 75% on average. In addition, MSSS had no effect on the identification false-positive rate (FPR) or the novel peptide sequences identification ability. We used MSSS to analyze another dataset of 34 LC-MS/MS runs, resulting in identifying additional 74 novel peptides. Proteogenomic analysis with these additional peptides yielded 47 new genomic features in 24 rice genes plus 24 intergenic peptides. These results show that the utility of MSSS in searching large databases with large MS/MS datasets for proteogenomics.

Introduction

Recent advances in genome sequencing have resulted in an enormous expansion of sequenced genomes. The genome online database (http://genomesonline.org/) currently exceeds 12,200 completed and ongoing genome-sequencing projects (January 2012). However, the genome sequence alone is not sufficient to elucidate biological functions and, therefore, the primary task in connection with any newly sequenced genome is to annotate the genomic sequence in order to attach biological meaning to it (Ansong *et al.* 2008; Wright *et al.* 2009). The genome annotation has two levels: (i) the structural level, identifying the gene structure; and (ii) the functional level, identifying the biological or biochemical function of the gene product (Koonin & Galperin 2003; Merrihew *et al.* 2008).

To perform genome annotation, genome-sequencing projects usually relay on transcriptional evidence, such as expressed sequence tags (EST) and a variety of *de novo* tools for gene finding and protein prediction (Castellana *et al.* 2008; Stanke *et al.* 2008; Wright *et al.* 2009). Although cDNA and EST can provide evidence for expression of a predicted gene, they still rely on the untranslated mRNA. Thus, they cannot confirm expression at the protein level (Wright *et al.* 2009). Consequently, although the *de novo* tools for gene finding and protein prediction vary in their algorithms and accuracy, they predict large numbers of gene models that suffer from errors in reading frames and exon definition and are sometimes highly redundant (Allen *et al.* 2004; Nielsen & Krogh 2005; Guigo *et al.* 2006; Ansong *et al.* 2008; Coghlan *et al.* 2008). These limitations indicated the

Communicated by: Keiichi I. Nakayama
**Correspondence:* yishiham@pharm.kyoto-u.ac.jp

DOI: 10.1111/j.1365-2443.2012.01615.x

need of another source of information to help correct and confirm the predicted gene models.

Shotgun proteomics approaches using liquid chromatography-tandem mass spectrometry (LC-MS/MS) directly measure the protease-digested peptides derived from the expressed proteins and therefore allows confirmation/correction of the expressed coding regions (Koonin & Galperin 2003; Ansong *et al.* 2008). Typically, the MS/MS spectra are searched against a reference protein database to identify peptides and proteins using algorithms such as Mascot (Perkins *et al.* 1999), SEQUEST (Eng *et al.* 1994), X!Tandem (Craig & Beavis 2004), PepSplice (Roos *et al.* 2007) and pFind (Fu *et al.* 2004; Li *et al.* 2005; Wang *et al.* 2007) which identify the peptides and their parent proteins. This limits the identification to the sequences available in the used database. Thus, the incompleteness of the protein databases causes many high-quality spectra to remain unidentified because of the absence of the corresponding amino acid sequence and limits the identification to known and predicted proteins (Baerenfaller *et al.* 2008; Mo *et al.* 2008; Power *et al.* 2009; Bitton *et al.* 2010).

Searching the MS/MS spectra against the genomic database is a well-known approach that permits wider identification of peptide sequences, as the database will include almost all possible sequences generated by the organism (Yates *et al.* 1995; Choudhary *et al.* 2001a,b; Ansong *et al.* 2008). Since the drafting of the human and *Arabidopsis* genomes, this approach has been widely used in finding novel genomic features using MS/MS data (Choudhary *et al.* 2001a,b; Kuster *et al.* 2001; Baerenfaller *et al.* 2008). In early work, the number of MS/MS spectra in the sample used for searching against the whole genome database was limited. For instance, the whole draft of the human genome (3.3 Gbp) was searched upon a test sample containing a total of 169 MS/MS spectra from 22 proteins (Choudhary *et al.* 2001a,b), which made identification possible even with limited computational resources. However, searching large-scale MS/MS datasets against the genome database remains a major challenge because of the enormous size of the MS/MS data and the databases and as a result of the linear relationship between search time and database size (Edwards 2007).

For instance, the human proteome database is approximately 25 Mbp, whereas the six-frame translation of the genome database is approximately 6 Gbp (Hixson *et al.* 2006). The six-frame translated genome database of rice is over 25 times the size of the rice protein database (Fig. S1 in Supporting Information). Furthermore, in the different updates of the genomic databases, the database size remains almost the same, for example, the first draft of the human genome database and the current version (HG19) are 3.3 and 3.12 Gbp, respectively. Thus, the principal factor that affects the search time and the required computing resources will be the number of MS/MS spectra to be identified. With large numbers of MS/MS spectra and large numbers of database searches, the time required to perform the search can be very long. For example, 260 days of CPU time was required to run 4261 X!Tandem searches against the *Shewanella* genome, even though a PRISM computing cluster with 32 processing nodes was used (Kiebel *et al.* 2006; Turse *et al.* 2010). Thus, several methods have been developed to facilitate this kind of search (Helmy *et al.* 2012).

Sevinsky *et al.* (2008) made it possible to search the whole human genome using common desktop computers although the GENQUEST method, which uses isoelectric focusing of peptides and accurate peptide mass to reduce the search space. Bitton *et al.* (2010) developed an integrated method to search the whole human six-frame translated genome database by splitting the database per chromosome, creating 23 target and decoy databases, then eliminating nonmatching peptides; this significantly reduced the search space.

Edwards successfully reduced the human EST database by 35-fold by means of a sophisticated database compression strategy requiring the EST to be mapped to the vicinity of a known gene, and the peptide to be contained in a 30-amino acid open reading frame (ORF), in which the peptide sequence is confirmed by at least two ESTs, followed by elimination of peptide sequence repetition. This makes search against the human ESTs database possible with affordable computational resources (Edwards 2007). The exon-graph method, proposed by Tanner and colleagues, was successfully used to identify novel genomic features in human and *Arabidopsis* (Tanner *et al.* 2007; Castellana *et al.* 2008). Mo *et al.* (2008) and Power *et al.* (2009) presented methods to identify peptides that overlap the exon-exon junction or exon-skipping events in human, respectively, by creating databases containing only the features that they are targeting. Fermin *et al.* (2006) identified 282 significant ORFs in human by creating an ORFs library for all possible reading frames of the human genome after splitting the genome per chromosome.

Furthermore, to speed up tandem mass spectra identification, some search engines, such as SEQUEST,

pFind and Crux (Eng *et al.* 1994; Wang *et al.* 2007; Park *et al.* 2008) use peptide indexing (Li *et al.* 2010). To create a peptide index, the proteins are digested *in silico* and information such as mass, length, and position are calculated for each peptide. The list is filtered later to remove redundancy and the final list is saved to disk, from which the engine can load it and match it with the spectra (Helmy *et al.* 2012). However, there are several drawbacks, such as the time required to construct the index and the need to reconstruct the whole index if there is any change in the searching parameters. Furthermore, for nonspecific digestion, the time required for constructing the index can be increased by 100-fold (Li *et al.* 2010).

Although these efforts have made it possible to search a large-sized database using MS/MS spectra, all of the procedures are dependent on either preprocessing of the database before searching or searching only a selected portion of the database that contains certain features. The preprocessing, such as splitting the database per chromosome, increases the overall processing time and effort, whereas searching against only certain features, such as ESTs or exons, causes loss of a significant part of the genomic information included in the omitted portion. Therefore, a new bioinformatics method that allows searching large datasets of MS/MS spectra of peptides against large databases, with reasonable computational cost and search time, is required.

In this work, we propose a novel bioinformatics approach, MSSS, that facilitates the identification of novel peptide sequences from large-scale MS/MS peptide spectra datasets and protein sequence and genomic databases. In MSSS, we search the MS/MS spectra against a reference database, for example, a database containing putative protein sequences, then remove the spectra corresponding to all identified peptides, creating a new file that contains only the unidentified spectra. Then, we use the new file to search a nucleotide sequence database to identify novel peptides, which cannot be derived from any annotated protein. The spectra subtraction approach is well-known approach that is optionally used by several database searching tools to exclude the low-quality spectra or as step in complex process aims to reduce the number of the unassigned spectra (Kapp & Schütz 2007; Ning *et al.* 2010). Furthermore, Mascot allows successive rounds of searching the unassigned spectra against different databases with different search parameters (Perkins *et al.* 1999). However, we adopted this approach here to increase the novel peptides identification for applications in proteogenomics and independently from the used search engine or bioinformatics pipeline. Because of this approach, we can reduce the number of spectra to be compared with the database instead of reducing the database size and consequently reduce the number of the queries to be carried out by the search engine, thereby reducing the search time and computational demand.

Results and discussion

For evaluating the MSSS approach shown in Fig. 1A, the published data from shotgun MS-based proteome analyses of rice cultured cells were used, as a typical testing dataset (Helmy *et al.* 2011). The dataset consisting of 152,908 MS/MS spectra were searched against the rice protein, cDNA, transcript and genome databases of Michigan State University (MSU), formerly known as the database of The Institute of Genome Research (TIGR) (Ouyang *et al.* 2007) using Mascot 2.3 (Perkins *et al.* 1999). The four databases and the searching order were carefully selected to provide a novel outcome from each database. The protein database was used as the reference database, because it contains all annotated proteins and peptide sequences. Although the content of the cDNA database corresponds to the protein database content, it allows searching different frame translations of the nucleotide sequence (frame translation is implicitly carried out by Mascot). The transcript database includes the introns, so we can identify intronic and exon-intron spanning peptides. Finally, the genome database includes the intergenic regions. Thus, each of the four databases offers the possibility of identifying unique features.

Four points should be taken into consideration during the assessment of any new method that aims to speed up the peptide sequence identification process (i) improvement of identification time, as the main purpose of the method; (ii) the sequence identification capability should be similar to that of the current methods; (iii) the accuracy should not be impaired; and (iv) the method should be flexible and easy to integrate into current data analysis workflows (Li *et al.* 2010).

To assess the performance of MSSS, its performance was compared with that of the normal peptide identification approach in three respects: 1) the search time required for the identification, 2) the peptide identification capability, and 3) the false-positive rate (FPR) of the identification. The fourth feature, flexibility, is already present because MSSS is an intermediate step that can be easily integrated into any

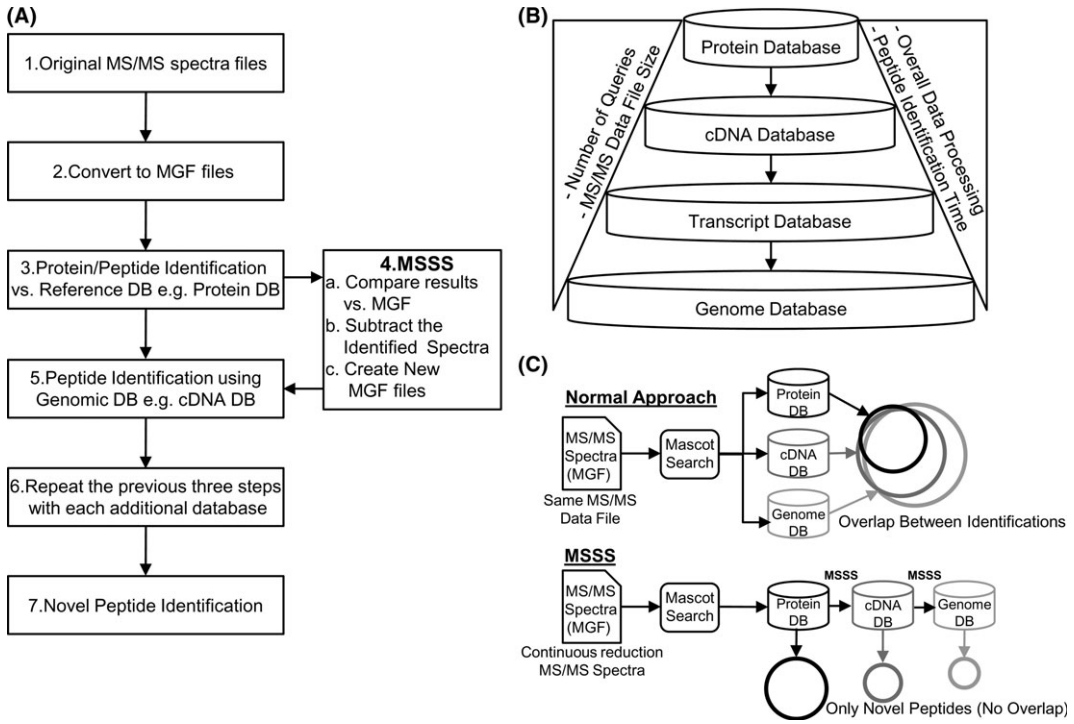

Figure 1 MSSS method. (A) Flowchart of the data analysis in the MSSS method. (B) Advantages of applying MSSS in the peptide identification process (increased peptide search space, decreased search time and decreased overall data processing requirement). (C) Comparison between MSSS and the normal approach. In this figure, we use three databases for demonstration, whereas in the actual work, we used four databases in the same sequence as shown in B.

workflow that supports the Mascot Generic Format (MGF) file format. Figure 1C illustrates the two approaches. We compared the two approaches with four different peptide acceptance criteria based on the identity score of the peptides (Table 1) to find the most suitable condition for MSSS.

Table 1 Peptide acceptance criteria used in the mass spectrum sequential subtraction evaluation

Criteria	Identification confidence (%)	*P* value	Description
C1	95	$P \leq 0.05$	Mascot score confidence $\geq$ 95%
C2	99	$P \leq 0.01$	Mascot score confidence $\geq$ 99%
C3	99.9	$P \leq 0.001$	Mascot score confidence $\geq$ 99.9%
C4	99.99	$P \leq 0.0001$	Mascot score confidence $\geq$ 99.99%

Because the main goal of this method is to provide a new strategy that makes peptide identification from large-scale MS/MS datasets with large-sized databases affordable (i.e., with reasonable computational demands and in a shorter time), it is indispensable to show that the influence of the MSSS approach on the size of MS/MS data files and the number of MS/MS spectra to be identified, because these are the two key factors that determine the time and computational resources required for searching a database.

In contrast to the normal approach, in MSSS, the file size is reduced after each search because of subtraction of the identified spectra. This is reflected in the number of MS/MS spectra per file and, therefore, the number of search queries to be carried out by the search engine. In MSSS, the total file size was reduced by 50% on average because of the sequential subtraction of the identified MS/MS spectra after each database search (Fig. 2A). The reduction in size was proportional to the number of MS/MS spectra remaining in the files, which was reduced by 45% on average (Fig. 2B). This means that the total number of search queries to be carried out by Mascot was

reduced by 45%, resulting in a decrease in search time and required computational resources by 25% on average (Fig. 2C).

The second comparison between MSSS and the normal approach is the peptide identification capability for (i) total nonredundant peptides and (ii) novel nonredundant peptides. We defined a nonredundant peptide as a unique combination of sequence plus modifications. Therefore, total peptides are all peptides identified from the four databases and novel peptides are peptides that cannot be derived from any annotated protein (peptides identified from the cDNA, transcript and genome databases). Both total and novel nonredundant peptides identified through MSSS were the same as those identified through the normal approach with various peptide acceptance criteria (Fig. 3A–B, Fig. S2 in Supporting Information).

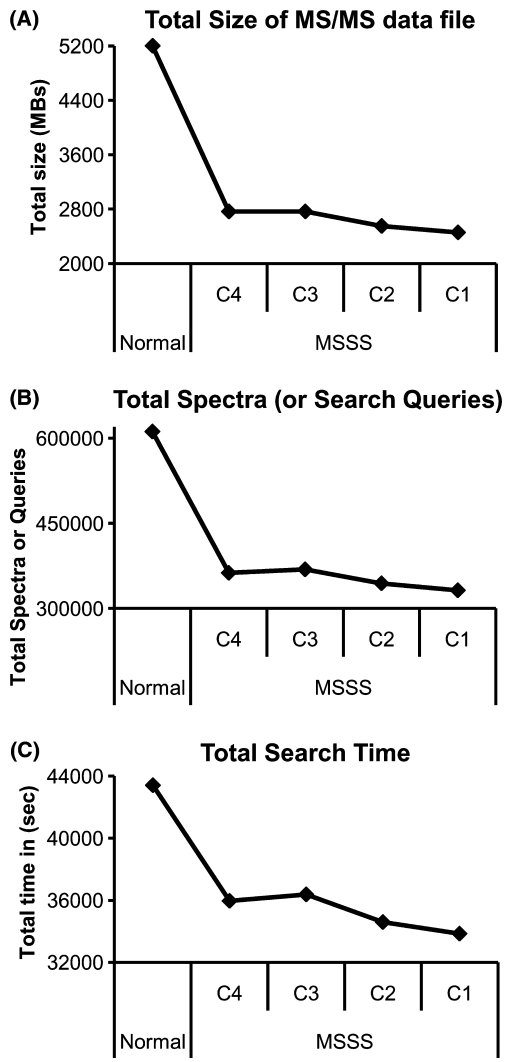

Figure 2 MSSS facilitates the identification of peptides from a large-scale MS/MS peptide spectra dataset and large-sized databases. (A) MS/MS data file size, (B) the number of MS/MS spectra (or the number of search queries carried out by Mascot) and (C) the search time required for peptide identification was reduced significantly after applying MSSS.

Furthermore, we compared the sources of novel peptide identifications at each acceptance level in MSSS and the normal approach to evaluate the contribution of each database. In all cases, the contribution of each database to the novel peptides was similar in both methods (Fig. S3 in Supporting Information). Furthermore, we compared the overlap between the peptides identified in both approaches and we found the same matches in all cases. These results show that the peptide identification capabilities of MSSS and the normal approach are comparable, regardless of the selected peptide acceptance criteria.

Finally, we compared MSSS and the normal approach in terms of the FPR of peptide identification (see Experimental procedures). Because a decoy database is equal in size to the target database, it was practically not possible to append a decoy version to the genome database, as a result of its large size (Fig. S1 in Supporting Information). Therefore, FPR was calculated for the protein, cDNA and transcript databases only (see Experimental procedures). For the protein database and cDNA database identifications, the FPR was the same in MSSS and in the normal approach with all four peptide acceptance criteria (Fig. 3C). In the case of the transcript database identification, the FPR was slightly increased in MSSS. However, this increase in FPR was negligible with the third and fourth criteria (Fig. 3C), which were the two criteria with acceptable FPR [FPR (%) $\leq$ 1%.] Thus, MSSS has a slight effect on the FPR at lower peptide score confidence, but has a negligible effect on the FPR at higher peptide score confidence.

The assessment of MSSS performance thus indicates that MSSS is comparable with the normal identification approach in terms of FPR and peptide identification. However, MSSS offers advantages in terms of reducing the search time, saving computational resources and facilitating peptide identification from large-scale MS/MS datasets and large-sized databases.

Next, we applied MSSS to another dataset of phosphopeptide-enriched rice samples (with total of 185,126 spectra) (Nakagami *et al.* 2010). The dataset of phosphopeptide-enriched samples was shown to extend the peptides coverage in proteogenomic application (Castellana *et al.* 2008). These MS/MS spectra

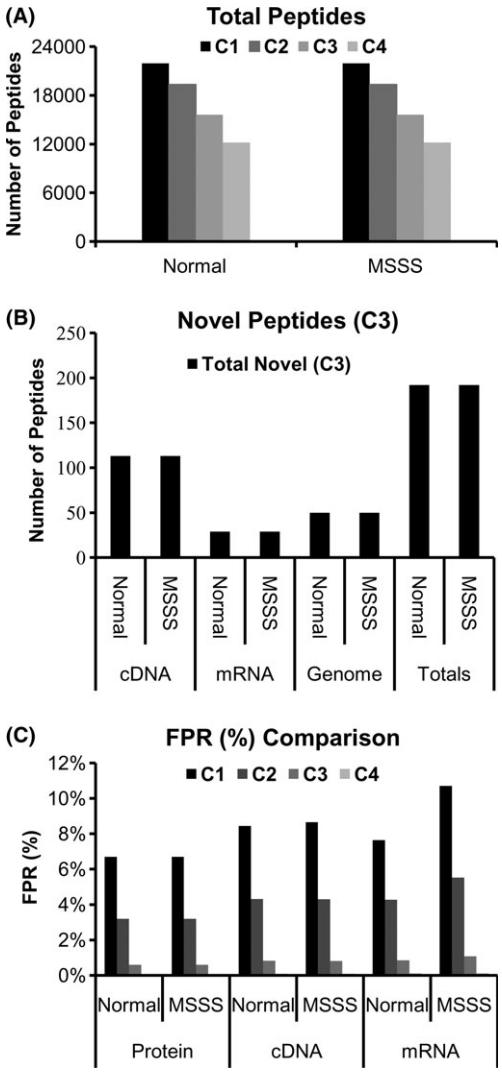

Figure 3 Assessment of MSSS performance. (A) Total nonredundant peptides identification in MSSS and the normal approach. (B) Novel nonredundant peptide sequences identification in MSSS versus the normal approach in C3 (99.9% peptide acceptance criteria). (C) FPR (%) in MSSS versus the normal approach. C1–C4 represent the four different peptide acceptance criteria (Table 1).

from 34 rice phosphoproteomic samples were searched against the same set of databases used in the earlier section (protein, cDNA, transcript and genome databases of MSU v6.1) using the MSSS approach. The MSSS search resulted in 5175, 237, 27, and 31 nonredundant peptides from the protein, cDNA, transcript and genome databases, respectively, when we used the third criterion of Table 1 for peptide identification. Note that the FPR(%) for identified nonredundant phosphopeptides was <0.1% (−0.08%) as was the case of the first test dataset, because we used more stringent filer with the phosphoproteomic samples (see Experimental procedures). The identified peptides were compared with all peptides of the rice proteogenomics database (OryzaPG-DB) (Helmy *et al.* 2011) to exclude the peptides that already exist in the database. The comparison resulted in 3095, 48, 6, and 26 nonredundant peptides identified from the protein, cDNA, transcript and genome databases, respectively, and not existing in OryzaPG-DB.

The 80 novel peptides identified from the cDNA, transcript and genome databases are a useful source of new genomic information, which can be used to refine the genome annotation by applying proteogenomic approaches (Ansong *et al.* 2008). Because we used very strict peptide acceptance criteria (99.9%), all peptides passed the statistical quality filters such as score, *e.value* and delta score. Furthermore, To confirm the identified spectra, the spectral quality was manually verified in terms of peak annotation and identified *b* and *y* ions. As a result, the total of 74 novel peptides (42, 6, and 26 peptides from the cDNA, transcript and genome databases, respectively) was accepted. The final dataset was processed using the following steps to find new genomic features that would help in the genome annotation refinement (Fig. 4A).

The peptides identified from the cDNA and transcript databases are from 39 genes (Table S1 in Supporting Information). Thus, we aligned each peptide to the corresponding unspliced genomic mRNA (transcript). Peptides identified from the genome database were mapped to the genome directly using the proteogenomic mapping tool (Sanders *et al.* 2011) and the mapping coordinates (start and end) were compared with the gene coordinates to map the peptides identified from the genome database to known genes. The mapping resulted in 24 peptides mapped to intergenic regions; the remaining peptides were mapped to known genes (Table S1 in Supporting Information). Peptides mapped to intergenic regions can potentially point to new unannotated genes or coding regions (Fermin *et al.* 2006; Kim *et al.* 2009).

Peptides mapped to known genes can be either from known coding regions, such as exons, or from novel regions, such as introns or untranslated regions (UTR). Peptides mapped to known coding regions are confirmatory peptides, which can be used to validate the current annotation, whereas peptides mapped to novel regions can be used to improve the current annotation by adding novel gene features, novel alternative splicing isoforms or new genes (Castellana & Bafna 2010). To assess the novelty of each peptide,

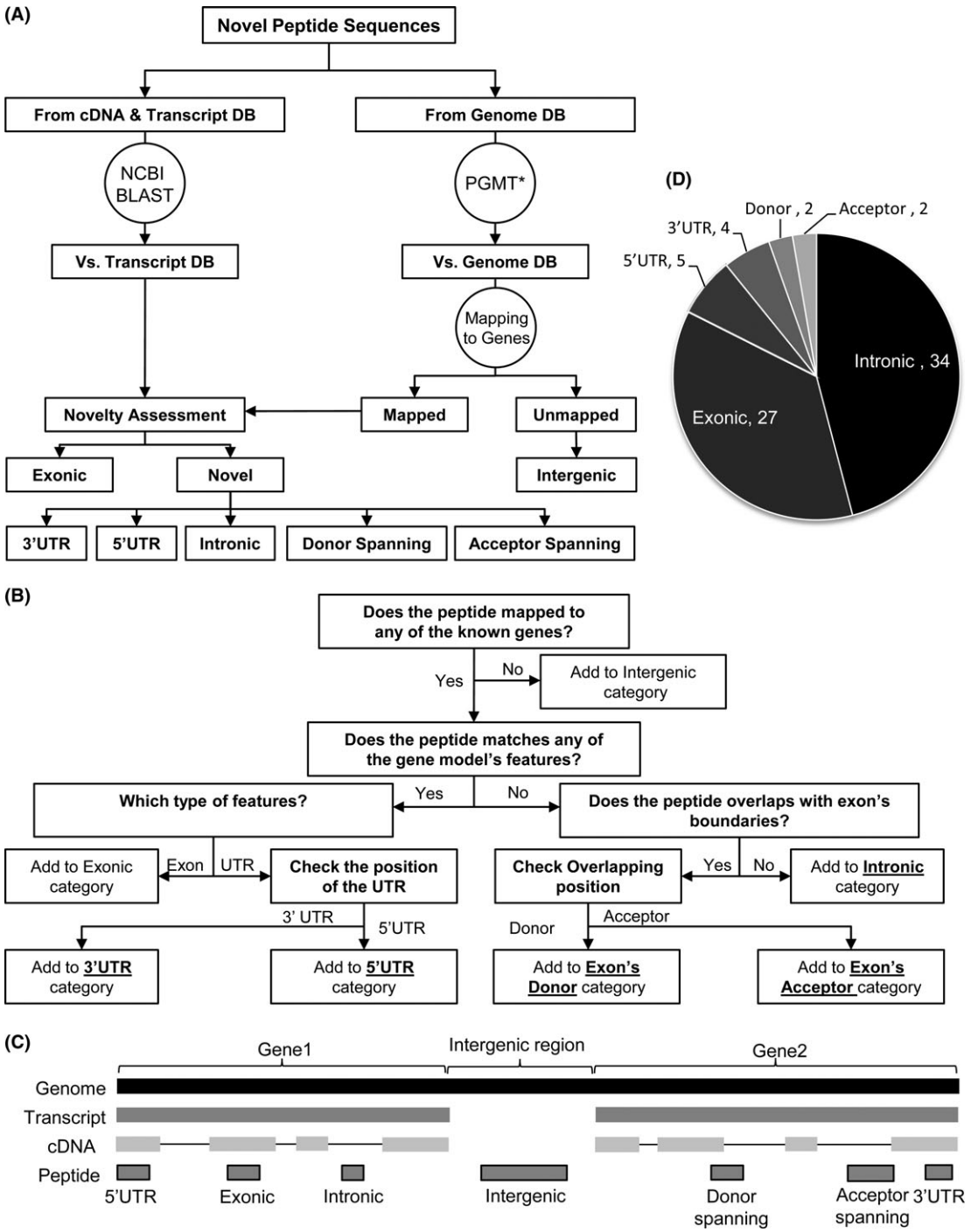

Figure 4 Proteogenomic analysis carried out using the novel peptides identified. (A) Flowchart of the proteogenomic analysis. (B) The updated novelty assessment algorithm used in this study. (C) Schematic illustration of peptide novelty categories. (D) Novel peptides per category. (*) PGMT stands for the proteogenomic mapping tool (Sanders *et al.* 2011).

we used an updated version of our novelty assessment algorithm previously implemented in the Proteo Genomics Features Evaluator (PGFeval) software tool (Helmy *et al.* 2011). The novelty categories of the newer version include intronic, acceptor spanning, donor spanning, exonic and 3′UTR, 5′UTR

(Fig. 4B). Figure 4C illustrates each of the novelty categories.

The proteogenomic analysis showed 47 novel genomic features in 24 genes 22 of them not existing in OryzaPG-DB (Fig. S4 in Supporting Information). The majority of the novel features were intronic peptides (34) and UTR peptides (nine) (Fig. 4D, Table S1 in Supporting Information). Figure 5 shows an example of an intronic peptide with its MS/MS spectra. Table S2 (Supporting Information) shows the comparison between the output of this study and the currently available in OryzaPG-DB.

These results show the utility of MSSS as a novel method to maximize the utility of the MS/MS spectra in proteogenomic studies. For instance, in the above-mentioned *Arabidopsis* study(Baerenfaller *et al.* 2008), 1354 MS/MS runs were carried out and identified using 261 novel peptides, whereas in our MSSS study, we have 34 nano LC-MS/MS runs (−0.025% of the *Arabidopsis* study MS/MS runs) and identified 74 novel peptides (−0.3% of the *Arabidopsis* study novel peptides), although we have to consider that different MS instruments used in the two studies (ion trap in the previous study and ion trap-orbitrap in our study).

In addition to its utility as shown previously, MSSS can be used to find mutated or abnormal peptides related to diseases that cause somatic mutations, such as cancer (Helmy *et al.* 2010). For example, the cancer proteome can be compared against the 'normal' protein database, then the genome database using MSSS. Next, the remaining spectra can be compared against a cancer-driven database, for example, the cancer transcriptome database. In this case, the identified peptides will be related to the disease condition, for example, mutations caused by the disease and, therefore, should be useful to find new biomarkers or new drug targets.

Conclusion

We have developed MSSS as a new bioinformatics method to facilitate the identification of peptides from large-scale MS/MS datasets and large-sized databases and showed that MSSS is useful in maximizing the utility of high-throughput mass spectrometry-based shotgun proteomics and phosphoproteomics data in proteogenomics. MSSS decreased the required search time and computational demands without affecting the accuracy or the peptide identification capability, comparing with the normal approach. Furthermore, it makes searching the whole genome database possible without extra preprocessing, reduction of the database features or splitting the database into smaller databases.

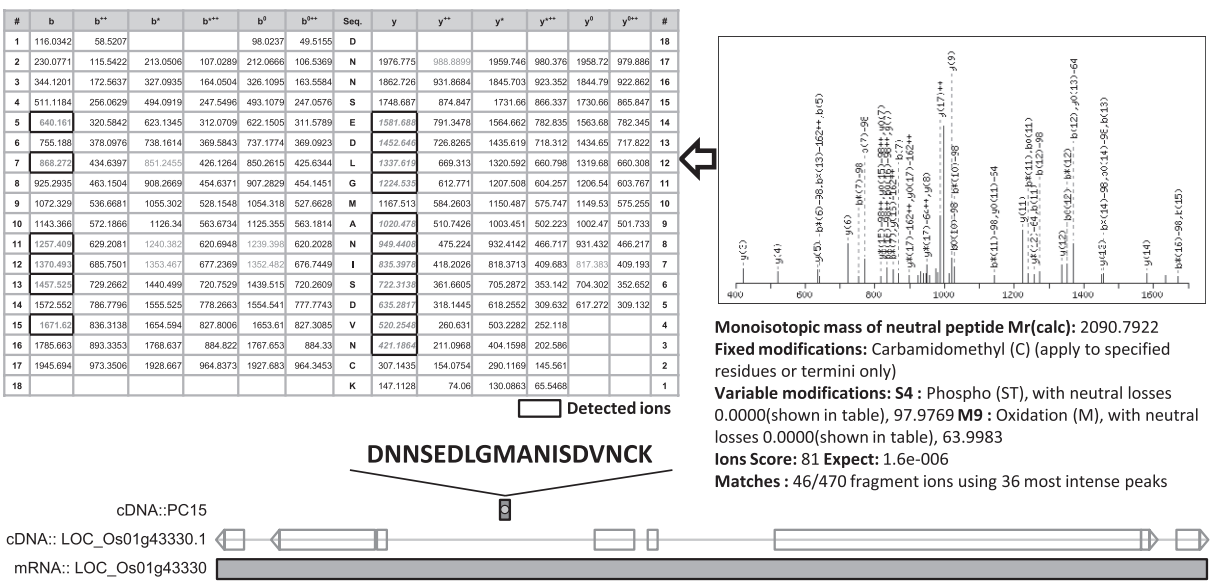

Figure 5 Example of a novel peptide with MS/MS spectra. The peptide was identified from the cDNA database and mapped using PGFeval (Helmy *et al.* 2011) to an intronic region on its genomic mRNA. The peptide's amino acid sequence and spectra are shown in the upper right panel, whereas the matched ions are shown in the upper left panel.

Experimental procedures

Datasets for method development and application

A published dataset from 27 LC-MS/MS analyses of rice cultured cells (Helmy *et al.* 2011) was used for the method development, whereas another published dataset from 34 LC-MS/MS analyses of phosphopeptide-enriched rice tryptic peptides (Nakagami *et al.* 2010) was used for the proteogenomic application of the method.

Database search

Peptides and proteins were identified by Mascot v2.3 (Matrix Science, London, UK) (Perkins *et al.* 1999) against the MSU rice protein, cDNA, transcript (unspliced genomic mRNA) and genome databases (IRGSP 2005; Ouyang *et al.* 2007). Mascot identification parameters were as follows: carbamidomethyl (C) as a fixed modification and acetyl (protein N-term), Gln->pyro-Glu (N-term Q), Glu->pyro-Glu (N-term E), and oxidation (M) as partial modifications for the method development dataset. Phosphorylation (S, T, and Y) as additional partial modifications was used for the application dataset. The product ion mass tolerance was 0.80 Da, whereas the precursor ion mass tolerance was 3 ppm and strict trypsin specificity was used, allowing for two missed cleavages only. In all Mascot searches, peptides were rejected if the Mascot score was below the 95%, 99%, 99.9%, or 99.99% confidence limit based on the identity score of each peptide (Table 1) (see Results and discussion). To increase the identification accuracy and peptide specificity, we accepted peptides with at least seven amino acids (Choudhary *et al.* 2001a,b) and rejected the peptide if the delta score between the first and second hits was <10. For the phosphopeptides identification, we require at least three successive *y*- or *b*-ions with a further two or more *y*-, *b*-, and/or precursor-origin neutral loss ions to be observed. In cases where different identification results were obtained from two databases for the same spectrum, that is, one from the protein database and the second from the cDNA, transcript or genome database, we selected the hit with higher significance (smaller *e.value*).

Mass spectrum sequential subtraction

After obtaining the MS/MS spectra by means of the experimental procedures described previously, we converted the raw data files to MGF (Fig. 1A step 2). Next, we carried out Mascot search against the protein database (reference database) (Fig. 1A step 3), then we compared the identification results with the original MGF files, as each identified peptide corresponds to certain MS/MS spectra. To automate this step, we created an in-house web-based tool written in PHP that performs the comparison, subtracts the identified MS/MS spectra and creates new MGF files containing only the unidentified MS/MS spectra (a basic version of the program, written in perl, is also available upon request). The subtraction significantly reduces the file size and the number of MS/MS spectra in the file (Fig. 1A steps 4a–c). The new MGF files can be searched against another database such as a cDNA, transcript or genome database (Fig. 1A step 5). For each database, we repeat the steps of identification, comparison, MS/MS spectra subtraction and new MGF file construction (Fig. 1A step 6). We end up with novel peptide sequences that do not exist in any of the annotated proteins, identified from searching of multiple genomic databases, with affordable computational demands, reduced search time and reduced overall data processing requirement (Fig. 1B).

MSSS evaluation scheme

To evaluate the performance of MSSS, we used the normal peptide identification approach as a control (shown in Fig. 1C, top). In the normal approach, the whole MS/MS peptide spectra dataset was searched separately and, respectively, against the protein, cDNA, transcript and genome databases to obtain the peptide sequences, using Mascot 2.3 (Fig. 1C, top), then the results of the different searches were combined in an accumulative way (only novel nonredundant peptides from each genomic database are added to the final list). In MSSS, Mascot search was carried out against the same four databases, but after each search the identified MS/MS spectra were subtracted and new MGF files were created, then used to search the next database. The searching order in MSSS was protein database, cDNA database, transcript database then genome database (Fig. 1C, bottom).

Calculating the FPR

To calculate the FPR, each of the protein, cDNA and transcript databases was appended with a decoy database because the use of concatenated target-decoy databases is preferable to separated database searches (Elias & Gygi 2007). For each database search result, we calculated the false-positives (FP) and the true-positives (TP) of the nonredundant set of identified peptides. Next, to calculate the FPR of the protein database search result, we used its own FP and TP ($FPR_{protein} = FP_{protein}/(FP_{protein} + TP_{protein})$) (Elias & Gygi 2007). For the cDNA and mRNA databases search, we calculated an accumulative FPR. The FPR of the cDNA database search result was calculated from $FPR_{cDNA} = FP_{protein+cDNA}/(FP_{protein+cDNA} + TP_{protein+cDNA})$, whereas the FPR of the transcript database search result was calculated from $FPR_{transcript} = FP_{protein+cDNA+transcript}/(FP_{protein+cDNA+transcript} + TP_{protein+cDNA+transcript})$. Therefore, we calculated unbiased FPR for both MSSS and the normal approach, avoiding the effect of any anomalous FPR value.

Bioinformatics analysis

The evaluation of a peptide's novelty and visualization of the genomic features were carried out using 'ProteoGenomic Fea-

tures Evaluator' (PGFeval) (see results) (Helmy *et al.* 2011). Sequence alignment was carried out using a local version of NCBI BLAST and BLS2SEQ Windows version with the default parameters (Altschul *et al.* 1990; Tatusova & Madden 1999) and perl script. Mapping the peptides identified from the genome database to the genome was carried out using the proteogenomic mapping tool (Sanders *et al.* 2011).

Acknowledgements

We would like to thank members of the proteomics group in our institute for their contributions. This study is funded by Yamagata Prefecture and Tsuruoka City grants to Keio University, as well as by the Egyptian Ministry of Higher Education, the Egyptian Bureau of Culture, Science and Education – Tokyo and JSPS Grants-in-Aid for Scientific Research (No. 236172) to M. H., Science and Technology Incubation Program in Advanced Regions from Japan Science and Technology Agency and JSPS Grants-in-Aid for Scientific Research (No. 21310129) to Y.I.

References

Allen, J.E., Pertea, M. & Salzberg, S.L. (2004) Computational gene prediction using multiple sources of evidence. *Genome Res.* **14**, 142–148.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.

Ansong, C., Purvine, S.O., Adkins, J.N., Lipton, M.S. & Smith, R.D. (2008) Proteogenomics: needs and roles to be filled by proteomics in genome annotation. *Brief. Funct. Genomic. Proteomic.* **7**, 50–62.

Baerenfaller, K., Grossmann, J., Grobei, M.A., Hull, R., Hirsch-Hoffmann, M., Yalovsky, S., Zimmermann, P., Grossniklaus, U., Gruissem, W. & Baginsky, S. (2008) Genome-scale proteomics reveals *Arabidopsis thaliana* gene models and proteome dynamics. *Science* **320**, 938–941.

Bitton, D.A., Smith, D.L., Connolly, Y., Scutt, P.J. & Miller, C.J. (2010) An integrated mass-spectrometry pipeline identifies novel protein coding-regions in the human genome. *PLoS One* **5**, e8949.

Castellana, N. & Bafna, V. (2010) Proteogenomics to discover the full coding content of genomes: a computational perspective. *J. Proteomics* **73**, 2124–2135.

Castellana, N.E., Payne, S.H., Shen, Z., Stanke, M., Bafna, V. & Briggs, S.P. (2008) Discovery and revision of Arabidopsis genes by proteogenomics. *Proc. Natl Acad. Sci. USA* **105**, 21034–21038.

Choudhary, J.S., Blackstock, W.P., Creasy, D.M. & Cottrell, J.S. (2001a) Interrogating the human genome using uninterpreted mass spectrometry data. *Proteomics* **1**, 651–667.

Choudhary, J.S., Blackstock, W.P., Creasy, D.M. & Cottrell, J.S. (2001b) Matching peptide mass spectra to EST and genomic DNA databases. *Trends Biotechnol.* **19**, S17–S22.

Coghlan, A., Fiedler, T.J., McKay, S.J., Flicek, P., Harris, T. W., Blasiar, D. & Stein, L.D. (2008) nGASP–the nematode genome annotation assessment project. *BMC Bioinformatics* **9**, 549.

Craig, R. & Beavis, R.C. (2004) TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **20**, 1466–1467.

Edwards, N.J. (2007) Novel peptide identification from tandem mass spectra using ESTs and sequence database compression. *Mol. Syst. Biol.* **3**, 102.

Elias, J.E. & Gygi, S.P. (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **20**, 7–214.

Eng, J.K., McCormack, A.L. & Yates, J.R. (1994) An approach to correlate MS/MS data to amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**, 976–989.

Fermin, D., Allen, B.B., Blackwell, T.W., Menon, R., Adamski, M., Xu, Y., Ulintz, P., Omenn, G.S. & States, D.J. (2006) Novel gene and gene model detection using a whole genome open reading frame analysis in proteomics. *Genome Biol.* **7**, R35.

Fu, Y., Yang, Q., Sun, R., Li, D., Zeng, R., Ling, C.X. & Gao, W. (2004) Exploiting the kernel trick to correlate fragment ions for peptide identification via tandem mass spectrometry. *Bioinformatics* **20**, 1948–1954.

Guigo, R., Flicek, P., Abril, J.F., *et al.* (2006) EGASP: the human ENCODE Genome Annotation Assessment Project. *Genome Biol.* **7**(Suppl. 1), S21–S31.

Helmy, M., Sugiyama, N., Tomita, M. & Ishihama, Y. (2010) Onco-proteogenomics: a novel approach to identify cancer-specific mutations combining proteomics and transcriptome deep sequencing. *Genome Biol.* **11**(Suppl. 1), P17.

Helmy, M., Tomita, M. & Ishihama, Y. (2011) OryzaPG-DB: Rice Proteome Database based on shotgun proteogenomics. *BMC Plant Biol.* **11**, 63.

Helmy, M., Tomita, M. & Ishihama, Y. (2012) Peptide identification by searching large-scale tandem mass spectra against large databases: bioinformatics methods in proteogenomics. *Genes Genom. Genomics* **6**, 76–85.

Hixson, K.K., Adkins, J.N., Baker, S.E., Moore, R.J., Chromy, B.A., Smith, R.D., McCutchen-Maloney, S.L. & Lipton, M. S. (2006) Biomarker candidate identification in Yersinia pestis using organism-wide semiquantitative proteomics. *J. Proteome Res.* **5**, 3008–3017.

IRGSP (2005) The map-based sequence of the rice genome. *Nature* **436**, 793–800.

Kapp, E. & Schütz, F. (2007) Overview of tandem mass spectrometry (MS/MS) database search algorithms. *Current Protocols in Protein Science.* **49**, 25.2.1–25.2.19.

Kiebel, G.R., Auberry, K.J., Jaitly, N., Clark, D.A., Monroe, M.E., Peterson, E.S., Tolic, N., Anderson, G.A. & Smith, R.D. (2006) PRISM: a data management system for high-throughput proteomics. *Proteomics* **6**, 1783–1790.

Kim, W., Silby, M.W., Purvine, S.O., Nicoll, J.S., Hixson, K. K., Monroe, M., Nicora, C.D., Lipton, M.S. & Levy, S.B.

(2009) Proteomic detection of non-annotated protein-coding genes in Pseudomonas fluorescens Pf0-1. *PLoS One* **4**, e8455.

Koonin, E. & Galperin, M. (2003) *Sequence-Evolution-Function: Computational Approaches in Comparative Genomics*. Norwell, MA: Kluwer Academic Publishers.

Kuster, B., Mortensen, P., Andersen, J.S. & Mann, M. (2001) Mass spectrometry allows direct identification of proteins in large genomes. *Proteomics* **1**, 641–650.

Li, D., Fu, Y., Sun, R., Ling, C.X., Wei, Y., Zhou, H., Zeng, R., Yang, Q., He, S. & Gao, W. (2005) pFind: a novel database-searching software system for automated peptide and protein identification via tandem mass spectrometry. *Bioinformatics* **21**, 3049–3050.

Li, Y., Chi, H., Wang, L.H., Wang, H.P., Fu, Y., Yuan, Z. F., Li, S.J., Liu, Y.S., Sun, R.X., Zeng, R. & He, S.M. (2010) Speeding up tandem mass spectrometry based database searching by peptide and spectrum indexing. *Rapid Commun. Mass Spectrom.* **24**, 807–814.

Merrihew, G.E., Davis, C., Ewing, B., Williams, G., Kall, L., Frewen, B.E., Noble, W.S., Green, P., Thomas, J.H. & MacCoss, M.J. (2008) Use of shotgun proteomics for the identification, confirmation, and correction of C. elegans gene annotations. *Genome Res.* **18**, 1660–1669.

Mo, F., Hong, X., Gao, F., Du, L., Wang, J., Omenn, G.S. & Lin, B. (2008) A compatible exon-exon junction database for the identification of exon skipping events using tandem mass spectrum data. *BMC Bioinformatics* **9**, 537.

Nakagami, H., Sugiyama, N., Mochida, K., Daudi, A., Yoshida, Y., Toyoda, T., Tomita, M., Ishihama, Y. & Shirasu, K. (2010) Large-scale comparative phosphoproteomics identifies conserved phosphorylation sites in plants. *Plant Physiol.* **153**, 1161–1174.

Nielsen, P. & Krogh, A. (2005) Large-scale prokaryotic gene prediction and comparison to genome annotation. *Bioinformatics* **21**, 4322–4329.

Ning, K., Fermin, D. & Nesvizhskii, A.I. (2010) Computational analysis of unassigned high-quality MS/MS spectra in proteomic data sets. *Proteomics* **10**, 2712–2718.

Ouyang, S., Zhu, W., Hamilton, J., Lin, H., Campbell, M., Childs, K., Thibaud-Nissen, F., Malek, R.L., Lee, Y., Zheng, L., Orvis, J., Haas, B., Wortman, J. & Buell, C.R. (2007) The TIGR rice genome annotation resource: improvements and new features. *Nucleic Acids Res.* **35**, D883–D887.

Park, C.Y., Klammer, A.A., Kall, L., MacCoss, M.J. & Noble, W.S. (2008) Rapid and accurate peptide identification from tandem mass spectra. *J. Proteome Res.* **7**, 3022–3027.

Perkins, D.N., Pappin, D.J., Creasy, D.M. & Cottrell, J.S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551–3567.

Power, K.A., McRedmond, J.P., de Stefani, A., Gallagher, W. M. & Gaora, P.O. (2009) High-throughput proteomics detection of novel splice isoforms in human platelets. *PLoS One* **4**, e5001.

Roos, F.F., Jacob, R., Grossmann, J., Fischer, B., Buhmann, J.M., Gruissem, W., Baginsky, S. & Widmayer, P. (2007) PepSplice: cache-efficient search algorithms for comprehensive identification of tandem mass spectra. *Bioinformatics* **23**, 3016–3023.

Sanders, W.S., Wang, N., Bridges, S.M., Malone, B.M., Dandass, Y.S., McCarthy, F.M., Nanduri, B., Lawrence, M. L. & Burgess, S.C. (2011) The proteogenomic mapping tool. *BMC Bioinformatics* **12**, 115.

Sevinsky, J.R., Cargile, B.J., Bunger, M.K., Meng, F., Yates, N.A., Hendrickson, R.C. & Stephenson, J.L. Jr (2008) Whole genome searching with shotgun proteomic data: applications for genome annotation. *J. Proteome Res.* **7**, 80–88.

Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. (2008) Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–644.

Tanner, S., Shen, Z., Ng, J., Florea, L., Guigo, R., Briggs, S. P. & Bafna, V. (2007) Improving gene annotation using peptide mass spectrometry. *Genome Res.* **17**, 231–239.

Tatusova, T.A. & Madden, T.L. (1999) BLAST 2 sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol. Lett.* **174**, 247–250.

Turse, J.E., Marshall, M.J., Fredrickson, J.K., Lipton, M.S. & Callister, S.J. (2010) An empirical strategy for characterizing bacterial proteomes across species in the absence of genomic sequences. *PLoS One* **5**, e13968.

Wang, L.H., Li, D.Q., Fu, Y., Wang, H.P., Zhang, J.F., Yuan, Z.F., Sun, R.X., Zeng, R., He, S.M. & Gao, W. (2007) pFind 2.0: a software package for peptide and protein identification via tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **21**, 2985–2991.

Wright, J.C., Sugden, D., Francis-McIntyre, S., Riba-Garcia, I., Gaskell, S.J., Grigoriev, I.V., Baker, S.E., Beynon, R.J. & Hubbard, S.J. (2009) Exploiting proteomic data for genome annotation and gene model validation in *Aspergillus niger*. *BMC Genomics* **10**, 61.

Yates, J.R. III, Eng, J.K. & McCormack, A.L. (1995) Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. *Anal. Chem.* **67**, 3202–3210.

Received: 19 March 2012
Accepted: 14 April 2012

Supporting Information/Supplementary material

The following Supporting Information can be found in the online version of the article:

Figure S1 Comparison of different rice databases in terms of file size and number of residues.

Figure S2 Identification of novel non-redundant peptide sequences by MSSS and the normal approach (C1, C2 and C4 acceptance criteria).

Figure S3 Comparison of the sources of the novel peptide sequences identified in this study.

Figure S4 Comparison between the output of this study and the OryzaPG-DB content.

Table S1 Novel Peptides identified by MSSS approach

Table S2 The output of this study versus the content of OryzaPG-DB

Additional Supporting Information may be found in the online version of this article.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.