Analysis of Soybean Long Non-Coding RNAs Reveals a Subset of Small Peptide-Coding Transcripts

Xiao Lin, Wengui Lin, Yee-Shan Ku, Fuk-Ling Wong, Man-Wah Li, Hon-Ming Lam, Sai-Ming Ngai, and Ting-Fung Chan

Centre for Soybean Research of the State Key Laboratory of Agrobiotechnology, and School of Life Sciences, The Chinese University of Hong Kong, Shatin NT, Hong Kong

ORCID IDs: 0000-0002-2559-2343 (X.L.); 0000-0003-4859-5683 (M.-W.L.); 0000-0002-6673-8740 (H.-M.L.); 0000-0002-0489-3884 (T.-F.C.).

Long non-coding RNAs (lncRNAs) are defined as non-protein-coding transcripts that are at least 200 nucleotides long. They are known to play pivotal roles in regulating gene expression, especially during stress responses in plants. We used a large collection of in-house transcriptome data from various soybean (Glycine max and Glycine soja) tissues treated under different conditions to perform a comprehensive identification of soybean lncRNAs. We also retrieved publicly available soybean transcriptome data that were of sufficient quality and sequencing depth to enrich our analysis. In total, RNA-sequencing data of 332 samples were used for this analysis. An integrated reference-based, de novo transcript assembly was developed that identified ~69,000 lncRNA gene loci. We showed that lncRNAs are distinct from both protein-coding transcripts and genomic background noise in terms of length, number of exons, transposable element composition, and sequence conservation level across legume species. The tissue-specific and time-specific transcriptional responses of the lncRNA genes under some stress conditions may suggest their biological relevance. The transcription start sites of lncRNA gene loci tend to be close to their nearest protein-coding genes, and they may be transcriptionally related to the protein-coding genes, particularly for antisense and intronic lncRNAs. A previously unreported subset of small peptide-coding transcripts was identified from these lncRNA loci via tandem mass spectrometry, which paved the way for investigating their functional roles. Our results also highlight the present inadequacy of the bioinformatic definition of lncRNA, which excludes those lncRNA gene loci with small open reading frames from being regarded as protein-coding.
protein-coding transcripts and a true set of lncRNAs. For many organisms, such as soybean, a true set of lncRNAs may not be available. As a result, these methods will depend heavily on features learned from protein-coding genes and may thus miss transcripts that encode small peptides. The method PhyloCSF (http://github.com/mlin/PhyloCSF/wiki) represents another type of tool, which is based on a statistical phylogenetic model (Lin et al., 2011). This method is highly dependent on the genome alignment, which is not computed by PhyloCSF itself, and is quite computationally intensive (Lin et al., 2011). In addition, it is only applicable to aligned genomic regions and ignores highly diverse or unaligned ones (Lin et al., 2011). The result generated by PhyloCSF can only suggest whether a region is more likely to be protein-coding or non-coding, and it is insufficient to predict a transcript as an lncRNA only based on this result. To compensate for the inadequacy of computational methods in lncRNA prediction, ribosome profiling and mass spectrometry (MS) data can be used to validate the translation process and peptide products, respectively (Choi et al., 2018).

In view of the aforementioned gap in knowledge on soybean lncRNAs, we took advantage of a large collection of in-house transcriptome data from various soybean tissues treated under various conditions to perform a comprehensive identification of soybean lncRNAs. We also retrieved existing soybean transcriptome data from public sources that are of sufficient quality and sequencing depth to enrich our analysis. A total of 332 RNA-sequencing (RNA-Seq) data samples were used for this analysis. An approach that integrated reference-based and de novo transcriptome assemblies was developed to predict soybean lncRNAs, thus identifying ~69,000 lncRNAs. We demonstrated that lncRNAs differ substantially from protein-coding transcripts in terms of lengths, numbers of exons, transposable element (TE) compositions, and sequence conservation patterns across legume species. Detected lncRNA gene loci are not simply genomic background noise as their transcription start sites (TSSs) and expression levels can be linked to their nearest protein-coding genes. We validated the salt stress responses of some lncRNA genes by reverse transcription quantitative PCR (RT-qPCR). We also showed that a small portion of lncRNAs can, in fact, encode peptides, albeit short ones (with fewer than 100 amino acids), as supported by MS evidence. This reveals their true identity as small peptide genes rather than lncRNAs.

RESULTS
Overview of LncRNA Identification

A total of 332 RNA-Seq data samples were used for lncRNA identification, in a way that integrated both reference-based and de novo methods. Overall, these data samples can be divided into two groups—the cultivated soybean (G. max) and the wild soybean (Glycine soja) datasets (Supplemental Table S1). In the reference-based method, quality- and adapter-trimmed RNA-Seq data from individual samples were mapped to a reference genome of either Williams 82 (Schmutz et al., 2010) or W05 (Fig. 1A; Supplemental Table S2; Xie et al., 2019), followed by transcript assembly (Fig. 1B) and annotation (Fig. 1C). A filtering step was used to detect lncRNAs based on transcript length ($\geq 200$ nt), sequence similarity to soybean protein-coding transcripts (Wucher et al., 2017), and sequence similarity to annotated protein sequences in the nonredundant protein database (Fig. 1D; Pruitt et al., 2007). In total, 46,974 and 35,675 lncRNA gene loci were identified in the G. max and G. soja datasets, encoding 48,752 and 37,599 lncRNA transcripts, respectively (Fig. 1D). The de novo method was predominantly used to cross compare lncRNAs identified in the two lncRNA sets. In this method, quality- and adapter-trimmed RNA-Seq data of individual samples were assembled without a reference genome (Fig. 1E). Transcripts sequenced using PacBio technology (https://www.pacb.com/) served as an independent source of de novo transcripts (Fig. 1F). The combined set of de novo assembled and sequenced transcripts were mapped to the reference genomes of Williams 82 and W05 (Supplemental Table S2) to find their genomic locations (Fig. 1G), which served as an intermediate and allowed us to compare lncRNAs identified among the two lncRNA sets (Fig. 1H). Using such comparisons, ~12,000 lncRNA loci were found to be common in the two lncRNA sets, while the remaining loci were unique to a specific lncRNA set (Fig. 2A).

The entire set of lncRNA gene loci can be categorized into antisense, intergenic, intronic, and overlapping based on their genomic positions relative to the nearest protein-coding genes (Fig. 2B). Antisense lncRNA genes were the most abundant type of the lncRNAs, accounting for ~91% in the G. max (Fig. 2C) and 90% in the G. soja lncRNA sets (Fig. 2D). Similar phenomena have been observed in mammalian species, but to date, the reason is not well understood (Faghihi and Wahlestedt, 2009). The intergenic lncRNA genes were the second most abundant, comprising ~8% of the G. max (Fig. 2C) and 9% of the G. soja lncRNA sets (Fig. 2D). However, intronic and overlapping lncRNA gene loci were rare, accounting for <0.5% of the entire lncRNA sets (Fig. 2, C and D). In addition, PacBio isoform sequencing (Iso-Seq) data could serve as convincing evidence to support the lncRNAs identified by the RNA-Seq data. Overall, 1,495 lncRNA transcripts in the G. max (Fig. 2E) and 1,141 lncRNA transcripts in the G. soja (Fig. 2F) lncRNA set were supported by the Iso-Seq data, with matched strand information and splicing junctions. Among these Iso-Seq–supported lncRNA transcripts, >90% were antisense lncRNAs, whereas the intergenic ones constituted only <9% and the intronic ones <1% (Fig. 2, E and F).
In a recent study by Golicz et al. (2018), >9,000 putative lncRNA loci were identified in various tissues of Williams 82, using only non-strand–specific RNA-Seq data. We compared the lncRNAs detected in the public lncRNA set to those in the lncRNA set produced by Golicz et al. (2018) separately, and found that only ~1,400 loci were shared between our annotation and the one produced by Golicz et al. (2018; Fig. 2G). It is worth noting that the lncRNAs identified in our study are all strand-specific, whereas most of the ones detected by Golicz et al. (2018) are non-strand–specific (Fig. 2G). We have provided annotations for all the lncRNA transcripts identified in this study (Supplemental Table S3) that were absent from the one by Golicz et al. (2018).

### Drastic Differences in Length Detected between LncRNAs and Protein-Coding Transcripts

LncRNAs are conventionally regarded as transcripts longer than 200 nt, without a specified upper limit. LncRNAs (median length: 200–500 nt) are generally much shorter than protein-coding transcripts (median length: >1,500 nt). In addition, the number of protein-coding sequences was much reduced in the set of transcripts measuring >1,000 nt (Supplemental Figs. S1, A–E, and S2, A–E). LncRNAs also tend to possess fewer exons than protein-coding genes, generally only one (Supplemental Figs. S1, F–J, and S2, F–J).

### TE Composition of LncRNAs

LncRNAs are often associated with TEs in animals and plants (Cho, 2018). We explored the TE composition of exons and introns of the two sets of lncRNAs by dividing each exon and intron into 100 bins of equal size and calculating the percentage of nucleotides overlapping with TEs in each bin irrespective of strand. The average TE composition value was taken for all exons or introns in the same lncRNA category for a corresponding bin. In general, the TE composition levels were low for antisense lncRNA exons and introns, as

---

**Figure 1.** Workflow for lncRNA identification. A, Clean reads were mapped to the reference genome using the program HISAT2 (Kim et al., 2015) for each sample. B, Reference-based transcript assembly by the program StringTie (Pertea et al., 2015) was based on the read-mapping result from (A) for each sample. C, An in-house Perl script was used to compare the merged set of transcripts assembled from (B) to the annotations of the reference genome. Only transcripts with expression levels >1 FPKM and reads-per-bp coverage >1 were included for downstream analyses. An assembled transcript from (B) would be categorized as either annotated or novel transcripts of annotated genes, or unannotated transcripts whose gene loci were absent in the reference annotation. The numbers of unannotated genes and transcripts are shown for the two lncRNA sets: G. max and G. soja. D, Unannotated transcripts had to satisfy three criteria to be classified as lncRNAs: (1) ≥200 nucleotides; (2) classified as lncRNAs by FEELnc (Wucher et al., 2017); and (3) with e-values ≥ 0.1 searched against the nonredundant protein database from NCBI using the program BLASTX (Altschul et al., 1997). The numbers of lncRNA genes and transcripts are shown for the two lncRNA sets. E, The same set of clean reads as in (A) were assembled by the program Trinity (Grabherr et al., 2011) using the de novo approach. F, Another set of de novo transcripts generated by PacBio Iso-Seq were retrieved from Xie et al. (2019). G, De novo transcripts from (E) and (F) were mapped to the reference genome with the software GMAP (Wu and Watanabe, 2005). H, Mapping results of de novo transcripts from (G) were used as intermediates to cross compare lncRNAs between the two lncRNA sets, G. max and G. soja.
well as for protein-coding genes, with levels being slightly higher in introns (Supplemental Fig. S3, A and B). The TE composition levels for intergenic, intronic, and overlapping lncRNA exons and introns were higher than those of antisense lncRNAs and protein-coding genes, but still fell below the levels of the genomic background (Supplemental Fig. S3, A and B). It has been suggested that lncRNA gene loci are more than just part of the genomic background and their origin, especially for intergenic lncRNAs. Instead they may be associated with TEs (Johnson and Guigó, 2014). It is worth noting that the TE composition levels for intronic and overlapping lncRNAs, in both their exons and introns, were more variable in both lncRNA sets (Supplemental Fig. S3, A and B), which may be due to the small number of genes in these two lncRNA categories.

**Sequence Conservation of LncRNAs Across Legume Species**

It is well known that many protein-coding transcripts are highly conserved across species, with higher similarities seen among those in more closely related lineages. Therefore, we investigated whether such sequence conservation holds true for soybean lncRNAs. Specifically, we chose five legume species for comparison: *Medicago truncatula*, *Lotus japonicus*, *Phaseolus vulgaris*, *Vigna angularis*, and *Cicer arietinum*. Soybean lncRNAs were aligned to the genomes of these five legume species, and a conservation score was calculated as the product between the percentage of lncRNA sequence coverage and identity. As indicated by the inflated peaks at the conservation score of zero, most soybean lncRNAs could not be aligned well in the
genomes of the five legumes species (Supplemental Figs. S4, A–D, and S5, A–D). Antisense lncRNAs showed a bimodal distribution of the conservation score, which suggests that some of them can be recovered in these legume species (Supplemental Figs. S4A and S5A), which is likely due to the presence of protein-coding genes on the opposite strands. As a control, protein-coding transcripts also exhibited a bimodal distribution of the conservation score, but no extremely inflated peak was observed at the score of zero, which suggests a higher level of sequence conservation (Supplemental Figs. S4E and S5E). The sequences of both lncRNAs and protein-coding transcripts are generally more conserved among soybean, *P. vulgaris*, and *V. angularis* because they belong to the Phaseoid clade, whereas *M. truncatula*, *L. japonicus*, and *C. arietinum* belong to the more distant Hologalegina clade (Lee et al., 2017; Supplemental Figs. S4 and S5).

### Inference of LncRNA TSSs

Because protein-coding transcripts are well known for the presence of promoter regions in their TSSs, we investigated the lncRNA TSSs by considering the TSSs of their nearest protein-coding genes. We first defined the orientations of the lncRNAs relative to their nearest protein-coding genes. The 5' end of each lncRNA gene locus and nearest protein-coding genes were considered as the TSS at the gene locus level for simplification. In this analysis, we considered the nearest protein-coding genes both on the same strand as the lncRNA gene loci and on the opposite strand. Head-to-tail (a) and tail-to-head (b) orientations were used to describe lncRNA protein-coding gene pairs on the same strand in the genome, and head-to-head (c) and tail-to-tail (d) refer to such pairs on different strands (Fig. 3A). The lncRNA protein-coding gene pairs are transcribed in the same direction for head-to-tail (a) and tail-to-head (b) configurations, and they are transcribed in opposite directions for head-to-head (c) and tail-to-tail (d) ones.

We then calculated the distance between the TSSs of each lncRNA protein-coding gene pair for both the *G. max* and *G. soja* lncRNA sets (Supplemental Figs. S6 and S7). The distance value is a positive number if the protein-coding gene TSS is downstream of the lncRNA gene locus TSS, regardless of the four orientations (a) to (d) described in Figure 3A. According to the distribution of these values for antisense (Supplemental Figs. S6B and S7B), intergenic (Supplemental Figs. S6C and S7C), intronic (Supplemental Figs. S6D and S7D), and overlapping (Supplemental Figs. S6E and S7E) lncRNA gene loci, it was observed that the tail-to-tail (d) configuration was enriched for antisense lncRNA gene loci (Supplemental Figs. S6B and S7B), whereas head-to-tail (a) was dominant for intronic ones (Supplemental Figs. S6D and S7D). However, among any adjacent pair of protein-coding genes, either on the same or different strands, no particular configuration was enriched.
Expression Correlation Patterns between LncRNA and Protein-Coding Gene Loci

Because we found that there is enrichment of specific orientation configurations of the TSSs of some lncRNAs, such as those of the antisense and intronic lncRNA genes in relation to those of their nearest protein-coding neighbors, we investigated whether any correlation exists between the expression of these lncRNA and protein-coding genes.

It was hypothesized that the expression levels of lncRNAs would be more correlated with their nearest protein-coding genes than with those farther away. We calculated Spearman correlation coefficients between lncRNA genes and each of their 10 nearest protein-coding genes. For antisense and intergenic lncRNA genes, their expression levels were more correlated with the nearer protein-coding genes than with those farther away, regardless of the strand (Supplemental Figs. S8, A and B, and S9, A and B). For intronic and overlapping lncRNA genes, there was no obvious trend, likely due to their small numbers (Supplemental Figs. S8, C and D, and S9, C and D). In contrast, a protein-coding gene and its neighbors did not show any correlation with respect to distance (Supplemental Figs. S8E and S9E). Thus, we focused on the expression correlation patterns between lncRNAs and their nearest protein-coding gene loci, on both strands. The distribution of Spearman correlation coefficients between the antisense lncRNA genes and their nearest protein-coding genes on opposite strands was shifted toward the positive direction of the Spearman correlation coefficient, compared with the random background distribution (Supplemental Figs. S10A and S11A), and so was that between intronic lncRNA genes and their nearest protein-coding genes on the same strand (Supplemental Figs. S10C and S11C). The distribution of expression correlation coefficients between the other types of lncRNA genes and their nearest protein-coding genes on either strand did not profoundly differ from the background distribution (Supplemental Figs. S10, B and D, and S11, B and D), nor did that between any two adjacent protein-coding genes (Supplemental Figs. S10E and S11E).

We hypothesized that high expression correlation levels may imply functional relationships, such as the regulation of expressions, between the lncRNA and protein-coding gene loci. It would be particularly interesting to look into antisense lncRNA genes, which are transcribed in the opposite direction to their nearest protein-coding genes, and intronic lncRNA genes, which are transcribed in the same direction as their nearest protein-coding genes, because their correlation coefficient distributions were shifted to the positive direction of the Spearman correlation coefficient compared to the background (Supplemental Fig. S10, A and C, and S11, A and C). Such lncRNA protein-coding gene pairs were selected and the distances between their TSSs were analyzed, to infer an effective distance between TSSs of a coexpressed lncRNA protein-coding gene pair. The cutoff for coexpression was set at a Spearman correlation coefficient > 0.5. It can be observed that the distance between an antisense lncRNA gene and its coexpressed nearest protein-coding gene, which is transcribed in the opposite direction, peaked at somewhere slightly larger than 1 kb (Fig. 3B). For an intronic lncRNA gene and its coexpressed nearest protein-coding gene, which is transcribed in the same direction, the distance between their TSSs peaked somewhere between 1 kb and 10 kb (Fig. 3C). It can be concluded that an effective distance between TSSs of antisense, or intronic, lncRNA genes and their nearest protein-coding genes is close to 1 kb (Fig. 3, B and C; Supplemental Fig. S12, B and C). It is worth noting that such close proximity does not always imply a coexpression relationship (Fig. 3, B and C; Supplemental Fig. S12, B and C, gray lines). Thus, it is important to consider both the distance between TSSs and expression correlation when inferring the coexpression relationship between antisense, or intronic, lncRNA genes and their nearest protein-coding genes.

To gain functional insight into these coexpressed (Spearman correlation coefficient > 0.5) lncRNA protein-coding gene pairs, the protein-coding genes were retrieved and used for GO analysis. The result showed that most of these protein-coding genes were related to transcriptional regulation (biological process), nucleus localization (cellular component), and nucleotide binding (molecular function; Fig. 3, D and E; Supplemental Fig. S12, D and E), which implies that these lncRNA genes may be involved in the regulation of gene expression networks.

Stress Response of LncRNA Genes in Different Soybean Germplasms and Tissues

Next, we explored whether these lncRNA genes were responsive to environmental stresses in various soybean germplasms and tissues. From four G. max datasets (datasets 5, 6, 7 and 10; Supplemental Table S1) and one G. soja dataset (dataset 16; Supplemental Table S1), we identified a total of 746 and 116 differentially expressed (DE) lncRNA genes, respectively (Fig. 4A; Supplemental Fig. S13A). We compared these DE lncRNA genes among the four G. max datasets as shown in the Venn diagram and observed that these DE genes were largely specific to each dataset and none
**Figure 4.** Differential expression analysis of lncRNA genes. A, The heatmap shows the scaled expression levels of 746 DE lncRNA genes in various stress conditions from four public datasets of *G. max* (Supplemental Table S1). The expression levels of each gene have been normalized to their own maximum level across all samples, which was set to 1. B, The Venn diagram shows the comparison of these genes across the four datasets. C, The pie chart shows the classification of the 764 DE lncRNA genes, including 130 antisense, 615 intergenic, and one intronic lncRNA genes. D, Time-series RNA-Seq expression patterns of lncRNA genes with reference to the Williams 82 genome. Eighty lncRNA genes that were induced either at early time points in root tissues, or at late time points in leaf tissues upon salt treatment are shown.
was common to all the datasets, which was probably due to the different experimental conditions of these datasets (Fig. 4B). Nevertheless, some DE lncRNA genes were found to be shared by some of the datasets. For example, 48 DE lncRNA genes were common to datasets 6 and 10, both of which investigated drought stress (Fig. 4B; Supplemental Table S1). However, the overlap between DE genes identified in the G. max and G. soja datasets was minimal, with only eight identified (Supplemental Fig. S13B). Greater than 80% of these DE genes were intergenic ones, whereas the antisense and intronic lncRNA genes constituted only a small portion (Fig. 4C; Supplemental Fig. S13C).

We also investigated lncRNA transcriptional responses with respect to time by making use of previously published RNA-Seq datasets corresponding to timed salt treatment of two of our in-house soybean germplasms, C08 (Liu et al., 2018) and W05 (Xie et al., 2019). In a previous study from our group, it was shown that sodium ion started to accumulate in root tissues within the first hour of salt treatment, while it was detected in leaf tissues after 24 h of salt treatment (Liu et al., 2018). In accordance with these observations, we mainly focused on lncRNAs that tended to be induced either at early time points (e.g. 1 and 2 h) in root tissues or at later time points (e.g. 24 and 48 h) in leaf tissues (Fig. 4D; Supplemental Fig. S14) upon salt treatment. These time-dependent expression patterns suggest that a subset of lncRNA genes are likely to participate in stress responses, rather than being transcriptional noise. We selected four such lncRNA genes for detailed expression validation using RT-qPCR. Their expression levels were induced at early time points in the root tissues (Supplemental Fig. S15, A–D). The lncRNA gene Gmax_MSTRG.19570 (Supplemental Fig. S15C) was highly induced specifically in W05, whereas the other three were induced in both C08 and W05 (Supplemental Fig. S15, A, B, and D). In addition, two of these lncRNA gene loci can be found in both this study and in that by Golicz et al. (2018; Supplemental Fig. S15, C and D), whereas the other two can only be found in this study (Supplemental Fig. S15, A and B).

**MS Evidence for LncRNAs Encoding Open Reading Frames**

Numerous ORFs can be predicted from lncRNAs even though by definition they are non-coding, because the “official” definition only considers ORFs longer than 100 amino acids. ORFs encoded by lncRNAs are typically shorter than 100 amino acids (Fig. 5A; Supplemental Fig. S16A), compared to protein sequences, most of which are a few hundred amino acids in length (Fig. 5D; Supplemental Fig. S16D). That may have been the rationale for the current lncRNA definition. ORFs encoded by untranslated regions (UTRs) of protein-coding transcripts were also considered and were found to show similar patterns to those encoded by lncRNAs (Fig. 5, B and C; Supplemental Fig. S16, B and C).

Liquid chromatography–tandem mass spectrometry (LC-MS/MS) was used to validate the translational landscape of these soybean lncRNAs. Total protein was extracted from the root tissues of germplasm C08 (G. max) and fractionated according to hydrophobicity. There were six fractions in total, with each fraction eluted in a certain range of acetonitrile (ACN; i.e. 0% to 9%, 9% to 21%, 21% to 30%, 30% to 39%, 39% to 51%, and 51% to 69%). These MS data were matched to our in-house soybean database and a customized Swiss-Prot database containing archaeal, bacterial, and viral protein sequences (Fig. 5E). The MS data that could be matched to sequences in the customized Swiss-Prot database with an expected value <10 were considered microbial contamination and were eliminated from any downstream analyses (Fig. 5E). Most of the remaining MS data matched known protein sequences, whereas those that matched small ORFs encoded by lncRNAs and UTRs constituted only a small portion (Fig. 5F). Overall, we found 3,212 unique small peptides supported by at least one mass spectrum, encoded by 3,609 lncRNA genes, from both the G. max and G. soja lncRNA sets (Fig. 5E). Cross comparison of these MS-supported small peptides identified 44 unique small peptides encoded by 52 lncRNA genes, with results supported by at least two spectra from different protein fractions (Fig. 5, E and G). A search for small peptides supported by at least two spectra from the same protein fraction identified 115 unique small peptides encoded by 135 lncRNA genes (Fig. 5, E and H). In total, 153 unique small peptides, out of the 3,212 small peptides, encoded by 179 lncRNA genes were identified, with at least two that were MS-supported (Fig. 5E). A search of the amino acid sequences translated from these 153 small peptides against the non-redundant protein database suggests that they do not match the existing sequences well and are potentially novel peptides.

To predict the biological functions of these novel small peptides, coexpression analysis was performed to check whether the lncRNA genes encoding these small peptides were coexpressed with protein-coding genes enriched in certain biological processes. Spearman correlation coefficients between the expression levels of these 179 small peptide-coding lncRNA genes and all protein-coding genes were calculated. The cutoff of coexpression was set at a Spearman correlation coefficient > 0.8. All of these coexpressed protein-coding genes were used for GO analysis. As a result, protein-coding genes of four biological processes, including generation of precursor metabolites and energy (Supplemental Fig. S17A), photosynthesis, light reaction (Supplemental Fig. S17B), ATP synthesis coupled electron transport (Supplemental Fig. S17C), and regulation of defense response (Supplemental Fig. S17D), were found to be particularly enriched, suggesting these small peptides are potentially involved in important biological processes.
DISCUSSION

In this study, we used a large collection of strand-specific RNA-Seq data from multiple tissue types, germplasms, and experimental conditions to identify soybean lncRNAs. By comparing lncRNAs predicted by the G. max and G. soja datasets, ~12,000 loci were found to be common to both sets, while the remaining loci were specific to each set (Fig. 2A). This is likely due to the differences in germplasms, tissue types, and conditions included in the G. max and G. soja datasets (Supplemental Table S1). In contrast, this also suggests the importance of using a large collection of datasets to make comprehensive predictions of lncRNAs. Compared to a recent study on soybean lncRNAs (Golicz et al., 2018), which used a non-strand–specific RNA-Seq dataset, here we can identify not only intergenic lncRNAs, but also antisense and intronic ones. In addition, strand-specific RNA-Seq data can resolve the strand information of all lncRNAs identified, whereas non-strand–specific data will leave most lncRNAs identified without any strand information (Fig. 2E).

Figure 5. Prediction of non-coding RNAs encoding ORFs and analysis of MS data. A to D, Length distributions of ORFs encoded by (A) lncRNAs, (B) 5' UTR, (C) 3' UTR, and (D) protein coding genes of Williams 82 are shown. Numbers of ORFs and median lengths are indicated. The x axis represents the ORF length in number of amino acid residues. The y axis represents the numbers of ORFs. E, The workflow of MS data analysis. F, Identities of MS data from protein samples extracted from different ACN fractions. The x axis indicates the ACN concentrations of different protein fractions. The y axis shows the number of spectra within the different categories. G, HPLC gradient for fractionation with the insert showing the numbers of MS data supporting the presence of small ORFs from different fractions (using the same color scheme as the fractionation gradient). H, Numbers of small ORFs supported by at least two spectra in the same protein fractions. The x axis indicates the ACN concentration of different protein fractions. The y axis shows the number of small ORFs.
By considering the diversity of soybean germplasms, the datasets used and the strand-specific properties, and by applying a set of stringent lncRNA prioritization criteria, we believe that the lncRNAs identified in this study are more complete and more accurate than those from previous studies and can thus serve as a better reference for both cultivated and wild soybean lncRNA catalogs.

Soybean lncRNAs are distinct from protein-coding transcripts in terms of length profiles (Supplemental Figs. S1, A–E, and S2, A–E), exon numbers (Supplemental Figs. S1, F–J, and S2, F–J), TE composition (Supplemental Fig. S3), and sequence conservation levels across legume species (Supplemental Figs. S4 and S5), and are also distinct from genomic background noise in terms of TE composition (Supplemental Fig. S3). Particularly, we noticed that antisense lncRNAs were more conserved than any other types of lncRNAs (Supplemental Figs. S4A and S5A). This observation is as expected because parts of the antisense lncRNA sequences overlap parts of the protein-coding transcripts on the opposite strand. As a result, antisense lncRNAs tend to be more conserved owing to the conserved protein-coding genes on the opposite strand (Supplemental Figs. S4E and S5E). Reciprocal search was performed to examine whether regions in the legume genomes, which were aligned with soybean lncRNAs, could also be mapped back to the corresponding loci in the soybean genomes. Among all those soybean lncRNAs aligned to the legume genomes, ~40% of the transcripts were considered to be reciprocal hits (Supplemental Table S4), suggesting the alignments of these transcripts to the legume genomes were more accurate than the rest of the transcripts. On the other hand, it can be interpreted that the sequence alignment of lncRNAs from one species with those from another species may not always be optimal, which is very likely due to their poor sequence conservation.

Soybean lncRNA gene loci are located closer to their nearest protein-coding genes (Supplemental Figs. S6 and S7), and exhibit higher expression correlation levels to their nearest protein-coding neighbors (Supplemental Figs. S8 and S9), than to protein-coding genes farther away, especially for antisense (Supplemental Figs. S10A and S11A) and intronic (Supplemental Figs. S10C and S11C) lncRNA genes. We observed that such expression correlation was more profound in the G. max datasets (Supplemental Fig. S10, A and C) than in the G. soja ones (Supplemental Fig. S11, A and C), whereas expression levels of neighboring protein-coding genes were more correlated in the G. soja datasets (Supplemental Fig. S11E) than in the G. max ones (Supplemental Fig. S10E). We suspected that it was due to the smaller number of samples in the G. soja datasets than in the G. max ones, which may have affected the robustness of Spearman correlation coefficient calculations in the G. soja datasets. With the addition of more samples in the G. soja datasets, we expect that the expression correlation patterns will be more similar to those seen in the G. max data. We hypothesize that most lncRNA gene loci are transcribed via their own promoters, being independent of their nearest protein-coding genes, because their TSSs tend to be far from those of their nearest protein-coding neighbors. For antisense lncRNA gene loci in tail-to-tail configurations (Supplemental Fig. S6A) with their nearest protein-coding genes, although these pairs are supposedly transcribed in opposite directions to each other, due to the close proximity of their TSSs, it is likely that the pairs share bidirectional promoters (Trinklein et al., 2004). Transcription of intronic lncRNA gene loci in head-to-tail configurations may also be associated with the transcription of the nearest protein-coding genes, given both their transcriptions take place in the same direction, though they may use different promoter elements (Heo and Sung, 2011). We showed that the close proximity between antisense lncRNA gene loci and their nearest protein-coding loci transcribed in the opposite direction, as well as between intronic lncRNA gene loci and their nearest protein-coding loci transcribed in the same direction, were necessary, but not sufficient, for their coexpression (Fig. 3, B and C). The effective distance between the TSSs of lncRNA and protein-coding loci for coexpression was slightly beyond 1 kb (Fig. 3, B and C).

Some lncRNAs have shown tissue-specific expression patterns. For example, in dataset 10 (Fig. 4A) and the time-series salt stress datasets, including datasets 1 and 4 (Fig. 4D; Supplemental Fig. S14), distinct tissue-specific patterns in root, stem, and leaf tissues can be observed. Time-specific expression patterns of lncRNA genes upon salt stress were also observed, showing that some were induced at early time points (Fig. 4D; Supplemental Figs. S14 and S15), while others were induced at late time points (Fig. 4D; Supplemental Fig. S14). By comparing the DE lncRNA genes of different germplasms under different stress conditions, we found that most of these lncRNA genes were unique to certain germplasms under certain stress conditions (Fig. 4B; Supplemental Fig. S13B). Such comparisons revealed 48 commonly induced lncRNA genes in Williams 82 upon drought stress (Fig. 4B, datasets 6 and 10). The expression patterns of lncRNA genes may also be specific to germplasms as we validated that Gmax_MSTRG.19570 was only induced in W05 upon salt stress (Supplemental Fig. S15C). Yet, we are limited by the current datasets to draw the conclusion that some lncRNA genes were specific to germplasms and stress conditions. Nonetheless, by summarizing all these expression patterns, it can still be suggested that lncRNA gene transcriptional responses are likely biologically relevant. In addition, we have validated by RT-qPCR not only the lncRNA genes identified in this study (Supplemental Fig. S15, A and B), but also those reported in Golicz et al. (2018; Supplemental Fig. S15, C and D), confirming the robustness of the lncRNA identification in this study.

Among the hundreds of lncRNA genes differentially expressed upon stress conditions, we found that >80% of them were intergenic and the remaining portion were antisense and intronic lncRNA genes (Fig. 5C; Supplemental Fig. S13C). We checked whether the nearest, or overlapping, protein-coding genes of these
DE lncRNA genes were also differentially expressed. It turns out that the expression correlation between the lncRNA and nearest protein-coding genes were generally weak under the stress conditions tested, for both antisense (Supplemental Fig. S18A) and intergenic ones (Supplemental Fig. S18B). Thus, it is not surprising that these protein-coding genes were not consistent with their nearest, or overlapping, lncRNA genes in terms of differential expression patterns, for both antisense (Supplemental Fig. S18C) and intergenic ones (Supplemental Fig. S18D). Hence it is difficult to infer the functions of these stress-responsive lncRNA genes.

The biological functions of many lncRNA genes remain mostly unknown. By analyzing the TSSs of lncRNA genes and their nearest protein-coding genes (Supplemental Figs. S6 and S7), as well as their coexpression patterns (Supplemental Figs. S10 and S11), we hypothesize that lncRNAs may play a role in transcriptional regulation (Fig. 3, D and E; Supplemental Fig. S12, D and E).

LncRNAs have been arbitrarily defined as transcripts longer than 200 nt that also lack an ORF longer than 100 amino acids. Some lncRNA prediction tools, which calculate the coding potential of a transcript, take the ORF length into consideration. For example, both CONC (Liu et al., 2006) and CPAT (Wang et al., 2013) regard ORF length as a feature in their machine-learning model. However, it is notable that a large portion of lncRNAs are between 200 and 300 nt in length (Supplemental Fig. S1, A–C; Supplemental Fig. S2, A–C), which is consistent with the scarcity of any lncRNAs having an ORF longer than 100 amino acids. However, the subset of lncRNAs with short (<100 amino acids) ORFs identified in this study reveal the inadequacy of the current definition of lncRNA. We have postulated that a portion of lncRNAs actually encode small peptides, because we have observed numerous ORFs, especially those <100 amino acids in length, in our lncRNAs (Fig. 5A; Supplemental Fig. S16A). Furthermore, our MS data revealed a previously unreported set of 179 lncRNAs that encoded 153 small peptides, which were each supported with confidence by at least two spectra (Fig. 5, E, G, and H). We have also tried to eliminate potential contamination from archaeal, bacterial, and viral proteins (Fig. 5E) to reduce false discovery. This potentially novel set of small ORF-encoding lncRNA genes may enable us to further investigate the roles of lncRNAs and will also likely lead to a new definition of lncRNAs. Our pioneering study of soybean lncRNAs has shown that an integrated framework that combines transcriptomics and proteomics can redefine what is conventionally thought of as lncRNAs to be small peptide-encoding genes. With this framework, ribosome profiling (Ingolia et al., 2009) can contribute additional evidence to support translational events surrounding these lncRNAs and further confirm their potential to encode small peptides.

CONCLUSION

Using high-quality soybean transcriptomic data from both publicly available sources and our in-house generated datasets, coupled with an in-house generated pipeline as well as comparisons with the soybean reference genomes Williams 82 and W05, we were able to identify a comprehensive set of lncRNAs in soybean, and demonstrate that they are distinct from the genomic background and protein-coding transcripts, in terms of length, number of exons, TE composition, and sequence conservation level. Moreover, TSS analyses among lncRNAs and their nearest protein-coding genes, and the higher-than-background expression correlations between certain groups of lncRNA protein-coding gene pairs under various experimental conditions, suggest that some lncRNAs may play an active role in transcriptional regulation. They may even be part of the response mechanisms toward salt stress, based on their time-specific, tissue-specific, and germplasm-specific expression patterns, extracted from the published databases and validated by RT-qPCR. Furthermore, the current non-coding part of the definition of lncRNA genes may have to be revisited because it does not take into consideration the possibility of these genes possessing an ORF that is shorter than 100 amino acids. In this study, we have uncovered a subset of soybean lncRNAs that encode small peptides, by extracting and fractionating the total protein from the root tissues of C08 using LC-MS/MS and being able to match the resulting MS data to the predicted small peptides based on those lncRNA sequences with small ORFs. Greater than 100 previously unreported small peptides were identified using this approach. Some of the lncRNA genes encoding them may be related to metabolism and defense response, as revealed by the protein-coding genes with high levels of expression correlation. This finding will also enable us to further investigate the functional roles of these lncRNAs.

MATERIALS AND METHODS

Data Collection and Preprocessing

The datasets used in this study are summarized in Supplemental Table S1. The data were divided into two sets, Glycine max and Glycine soja (Supplemental Table S1). Datasets 1 and 4 were not included for lncRNA identification, and thus are denoted as “NA” (Supplemental Table S1). All datasets are publicly available (Supplemental Table S1). Adapter and quality trimming was performed on the raw reads using the tool Trimomatic 0.36 (Bolger et al., 2014).

Reference Genomes and Annotations

Reference genomes and gene annotations were publicly available for Williams 82 (Wm82a.v1; G. max; Schmutz et al., 2010) and W05 (G. soja; Xie et al., 2019; Supplemental Table S2). We used two soybean reference genomes here because our group has previously reported the substantial genomic differences between G. max and G. soja (Lam et al., 2010; Xie et al., 2019). The softwares RNAmmer v1.2 (Lagesen et al., 2007), tRNAscan-SE v2.0 (Schattner et al., 2005), and Infernal v1.1.2 (Nawrocki and Eddy, 2013) were used for predicting ribosomal RNAs, transfer RNAs, and non-coding RNAs with the program Rfam v13.0 (Kalvari et al., 2018), respectively, in the reference genomes of Williams
Reference-Based Transcriptome Assembly with Strand-Specific RNA-Seq Data

Clean reads for each sample were first mapped to a reference genome with the software HISAT2 v2.1.0 (Kim et al., 2015; Fig. 1A). While reads of G. max samples were mapped to the Williams 82 genome, those of G. soja samples were all mapped to the W05 genome (Supplemental Table S2). Reads alignment results were transferred to the program StringTie v1.3.4d (Pertea et al., 2015) for transcript assembly, without using any gene annotations (Fig. 1B). After the transcriptome assembly, transcripts were merged within the set to which they belong (Supplemental Table S1), using the “merge” function in StringTie, where only transcripts with expression levels >1 fragment per kilobase of transcript per million mapped reads (FPKMs) and reads-per-kbp coverage > 1 were included (Fig. 1B).

de Novo Transcriptome Assembly

The software Trinity v2.6.6.6 (Grabherr et al., 2011) was used for de novo transcriptome assembly (Fig. 1E). de novo transcriptome assembly was performed separately for each sample. All Trinity-assembled transcripts, as well as PacBio Iso-Seq data (Xie et al., 2019), were aligned to the reference genomes of Williams 82 and W05 using GMAP v2018-03-25 (Wu and Watanabe, 2005; Fig. 1, F and G).

Identification of LncRNAs

There were two assemblies generated in total using strand-specific RNA-Seq data. These assemblies were compared with transcripts from gene annotations consisting of protein-coding gene (Supplemental Table S2) and small non-coding RNA annotations from their respective soybean reference genome, Williams 82 or W05, using the software BEDTools v2.27.1 (Quinlan and Hall, 2010) and an in-house Perl script, with the following criteria (Fig. 1C): If an assembled single-exon transcript overlapped a single-exon annotated transcript, or the assembled multiexon transcript shared the exact set of exon junctions with the associated strand information, it would be classified as an annotated transcript; if an assembled transcript overlapped an annotated gene but did not share the exact set of exon–exon junctions with any of the annotated transcripts of the gene with matching strand information, it would be classified as a novel transcript of the corresponding gene; if an assembled transcript cannot be assigned to any single gene, it would be classified as an unannotated transcript; any assembled transcript matching parts of transcripts of more than one annotated genes was considered a suspected transcript and would not be included in the downstream analyses; and if a gene contained both unannotated and suspected transcripts, it would not be included in the downstream analyses.

Three criteria were used to further narrow-down the genes potentially encoding lncRNAs (Fig. 1D): (1) If a gene contained transcripts shorter than 200 nt in length, it would be excluded from downstream analyses. (2) FEELnc (Wucher et al., 2017) was used to evaluate the coding potential of all unannotated transcripts. The parameters were set as “-m ‘shuffle’,” and the shuffled protein-coding transcripts were used as a non-coding training set. If a gene contained transcripts classified as mRNA, it would not be included in the downstream analyses. (3) Sequences of all unannotated transcripts were also searched against the nonredundant protein database from the National Center for Biotechnology Information (NCBI; https://www.ncbi.nlm.nih.gov/) using the software BLASTX v2.6.0+ (Altschul et al., 1997). Sequences with hits having e-values < 0.1 were excluded in the downstream analyses.

Each Trinity assembly, as well as Iso-Seq data, were used to cross compare each unannotated transcript using an in-house Perl script to identify their presence in the two lncRNA sets (G. max and G. soja; Fig. 1H). An lncRNA was considered as a match to a Trinity-assembled transcript or Iso-Seq data only if they were matched in all exon-exon junction sites and were on a matching strand. The method for identifying lncRNAs described above was developed based on a previously published protocol (Lin et al., 2019).

Gene Expression Level Analysis for lncRNAs

Gene expression level analysis was performed with the “ cuffdiff” script from the program Cufflinks v2.2.1 (Trapnell et al., 2012, 2013) for datasets 5, 6, 7, 10 (Song et al., 2016), and 16 (Waldeck et al., 2017; Supplemental Table S1). The entire set of annotations, including all protein-coding and non-coding genes, was used for the calculation. Biological replicates were used to calculate the average expression level of each gene under each experimental condition. All expression levels were calculated as FPKMs. The criteria for determining whether there was differential expression were: fold change > 2 and q value < 0.05. The Spearman correlation coefficient was calculated between lncRNA and protein-coding genes, or two protein-coding genes, to represent pairwise gene expression correlation separately for G. max and G. soja data. Only samples with biological replicates were included in the expression correlation analysis. A random control for the Spearman correlation distribution for lncRNA genes was generated by substituting the nearest protein-coding gene with a random protein-coding gene in the same chromosome. For protein-coding genes, 100,000 random pairs in the same chromosome were generated. These random sampling steps were repeated 100 times.

GO Analysis

Protein-coding genes from neighboring lncRNA protein-coding gene pairs with a Spearman correlation of gene expression levels >0.5 were used for GO analysis performed with the GO Term Enrichment Tool from SoyBase.
all protein-coding and non-coding genes, was used for the calculation. Biological replicas were used for the differential expression analysis of pairwise samples. The criteria for differential expression was $|\log2 FC| > 1$ and $q value < 0.05$. All expression levels were calculated as FPKMs.

### RT-qPCR Validation of IncRNA Expression Patterns

Soybean seeds of germplasms C08 and W05 were germinated in vermiculite moistened with tap water in a greenhouse. The water content of the vermiculite was ~70%. Seven-d-old seedlings were transferred to a hydroponic system with half-strength Hoagland’s solution (Supplemental Tables S7 and S8; Hoagland and Snyder, 1933). When the primary leaves were fully open, the salt treatment (0.9% [w/v] NaCl in half-strength Hoagland’s solution) was applied. The leaf and root tissues were collected at 0, 1, 2, 4, 24, and 48 h after the start of treatment for RNA extraction with TRIzol Reagent (Invitrogen), following the manufacturer’s protocol. There were three biological replicates for each treatment and time-point.

A PrimeScript One Step RT-PCR Kit (Takara) was used for RT-qPCR as per the product manual. Briefly, each 20 μL reaction contained 10 μL of 2× One Step SYBR RT-PCR buffer 4, 0.4 μL of PrimeScript One Step Enzyme Mix 2, 0.4 μL of each primer, and 20 ng of total RNA. Three biological replicates were analyzed for each treatment and time-point. RT-qPCR was performed using a CFX96 Touch Real-Time PCR Detection System (Bio-Rad). The thermal cycler program is shown in Supplemental Table S9. Elf1β was used as the reference gene, as reported by Yim et al. (2015). Primer sequences are listed in Supplemental Table S10.

### Analysis of Non-Coding Regions Encoding ORFs

The RNA sequences in the two IncRNA sets, derived from the 5′ and 3′ UTRs of putative protein-coding genes from the reference genomes of both Williams 82 and W05, were translated into peptide sequences in three reading frames on the positive strand using an in-house Perl script. Each ORF must end with a stop codon but did not have to start with the AUG start codon. ORFs with lengths ≤10 amino acid residues were discarded. The entire set of translated peptides, together with all available soybean protein sequences from the reference genomes of both Williams 82 and W05, and signaling peptides (Choo et al., 2005) of both G. max and G. soja, served as an in-house soybean amino acid sequence database for MS data analyses.

### Analysis of MS Data

Soybean seeds of germplasm C08 were germinated in vermiculite moistened with tap water in the greenhouse. The water content of vermiculite was ~70%. Seven-d-old germinated seedlings were transferred to a hydroponic system with half-strength Hoagland’s solution (Supplemental Tables S7 and S8). Root tissues were collected when the third trifoliate fully opened.

The protein extraction method was modified from Chen et al. (2014). Briefly, 40 g of root tissues was ground into a fine powder in liquid nitrogen. One-hundred microliters of 0.1% (v/v) trifluoroacetic acid (TFA) was then added and blended for another 2 min to produce a crude root extract, which was then centrifuged twice at 10,000g for 20 min at 4°C. After having been sequentially passed through filter paper and a MF-Millipore Membrane Filterer (0.45-μm pore size; Millipore), the supernatant was then loaded onto a Sep-Pak C18 column (Waters), which had been equilibrated with 30 mL of 0.1% (v/v) TFA. Sixty microliters of 0.1% (v/v) TFA was used to wash the column after loading. Then 100 mL of 0.1% (v/v) TFA/60% (v/v) methanol was applied to elute the proteins/peptides. The eluate was vacuum-dried to remove the methanol content, followed by lyophilization. The dried sample was dissolved in 150 μL of 2% (w/v) SDS solution and alkylated as described by a protocol from the University of Washington Proteomics Resource (2011) as follows: The sample was heated at 55°C for 30 min with DTT at a final concentration of 5 mM, followed by alkylation with a final concentration of 14 mM of iodoacetamide in the dark for 30 min. Alkylation was quenched by adding extra DTT to make up to a final concentration of 10 mM and incubated for another 15 min. SDS was removed by a detergent-removal column (Thermo Fisher Scientific). The protein solution was fractionated by HPLC (Agilent Technologies) into six groups based on differential ACN concentration ranges: 0% to 9%, 9% to 21%, 21% to 30%, 30% to 39%, 39% to 51%, and 51% to 69%. Each fraction was lyophilized again and dissolved in 0.05% (v/v) formic acid for downstream LC-MS/MS
Coexpression Network Analysis of Small Peptide-Coding lncRNA Genes

The Spearman correlation coefficient was calculated between the small peptide-coding lncRNA genes and protein-coding genes from the G. max and G. soja datasets separately, to determine coexpression. The lncRNA protein-coding gene pairs with a Spearman correlation of their gene expression levels >0.8 were used for GO analysis with the GO Term Enrichment Tool from SoyBase (Grant et al., 2010). The gene identifications from the G. soja annotations were converted to those from the G. max annotations before performing GO analysis. Briefly, all protein sequences of G. max and G. soja were searched against each other using the program BLASTp. If a pair of G. max and G. soja protein-coding genes got the most number of hits with each other, a corresponding relationship would be assigned to the gene pair. The coexpression networks were visualized with the program Cytoscape 3.7.2 (Shannon et al., 2003).

Accession Numbers

The MS proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository (Perez-Riverol et al., 2019; https://www.ebi.ac.uk/pride/archive/) with the dataset identifier PXD014553.

Statistical Analyses

Unless specified elsewhere, all plots were generated using different functions in the software R (v3.5.1; https://www.r-project.org/), including “pie” for pie charts, “hist” for histograms, and “barplot” or “ggplot2” for bar plots. The density plots were generated using the “density” and “plot” functions. For generating the heatmap, the expression levels of each gene were first normalized to their own maximum level across all samples, which was set to 1. Hierarchical clustering was performed only for the genes using the “hclust” function in the software R, with the distance calculated by 1 − Spearman coefficient. The heatmap was visualized by “heatmap.2” in the “ggplot2” package.

Supplemental Data

The following supplemental materials are available.

Supplemental Figure S1. Statistics on the lengths and exon numbers for the G. max lncRNA set.

Supplemental Figure S2. Statistics on the lengths and exon numbers for the G. soja lncRNA set.

Supplemental Figure S3. Analysis of TE compositions.

Supplemental Figure S4. Sequence conservation levels across five legume species of the G. max lncRNA set.

Supplemental Figure S5. Sequence conservation levels across five legume species of the G. soja lncRNA set.

Supplemental Figure S6. Analysis of lncRNA gene locus TSSs in the G. max lncRNA set.

Supplemental Figure S7. Analysis of lncRNA gene locus TSSs in the G. soja lncRNA set.

Supplemental Figure S8. Expression correlations between lncRNA genes and their ten nearest protein-coding genes in the G. max lncRNA set.

Supplemental Figure S9. Expression correlations between lncRNA genes and their ten nearest protein-coding genes in the G. soja lncRNA set.

Supplemental Figure S10. Expression correlations between lncRNA genes and their nearest protein-coding genes for the G. max lncRNA set.

Supplemental Figure S11. Expression correlations between lncRNA genes and their nearest protein-coding genes for the G. soja lncRNA set.

Supplemental Figure S12. Analysis of lncRNA gene loci TSSs in the G. soja lncRNA set.

Supplemental Figure S13. Differential expression analysis of lncRNA genes.

Supplemental Figure S14. Time-series RNA-Seq expression patterns of lncRNA genes with reference to the W05 genome.

Supplemental Figure S15. Validations of the lncRNA expression patterns.

Supplemental Figure S16. Prediction of non-coding RNAs with ORFs.

Supplemental Figure S17. Coexpression networks of small peptide-coding lncRNA genes and protein coding genes.

Supplemental Figure S18. Coexpression analysis of DE lncRNA genes in response to stress conditions.

Supplemental Table S1. RNA-Seq datasets used in this study.

Supplemental Table S2. The soybean reference genomes and protein-coding gene annotations used in this study.

Supplemental Table S3. Genome annotations of the G. max and G. soja lncRNA sets.

Supplemental Table S4. The percentages of lncRNAs aligned to the genomes of five legume species with reciprocal hits for the G. max and G. soja lncRNA sets.

Supplemental Table S5. Proportions of random DNA sequences in different length ranges.

Supplemental Table S6. The reference genomes of the non-soybean legume species used in this study.

Supplemental Table S7. The recipe of half-strength Hoagland’s solution.

Supplemental Table S8. The recipe of the micronutrient stock solution (1,000×) as mentioned in Supplemental Table S6.

Supplemental Table S9. The thermal cycler program for RT-qPCR.

Supplemental Table S10. Primer sets used in this study.

Supplemental Table S11. Comet parameters for MS data searching.

ACKNOWLEDGMENTS

We thank Dr. Henry T. Nguyen and Dr. Babu Valliyodan for their kind assistance in accessing their published RNA-Seq dataset. Jee Yan Chu copyright-ed the article. We also thank the two reviewers for their constructive comments in improving our work.

Received November 14, 2019; accepted December 13, 2019; published December 27, 2019.

LITERATURE CITED

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402

Bao W, Kojima KK, Kohany O (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. Mol DNA 46: 11

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114–2120

Chen Y-L, Lee C-Y, Cheng K-T, Chang W-H, Huang R-N, Nam HG, Chen Y-R (2014) Quantitative peptidomics study reveals that a wound-induced...
peptide from PT-1 regulates immune signaling in tomato. Plant Cell 26: 4135–4148

Choi J (2018) Transposon-derived non-coding RNAs and their function in plants. Front Plant Sci 9: 600

Choi S-W, Kim H-W, Nam J-W (2018) The small peptide world in long noncoding RNAs. Brief Bioinform 20: 1853–1864

Choo KH, Tan TW, Ranganathan S (2005) SPBp—a signal peptide database. BMC Bioinformatics 6: 249

Eng JK, Jahan TA, Hoopmann MR (2013) Comet: An open-source MS/MS sequence database search tool. Proteomics 13: 22–24

Faghihi MA, Wahledstedt C (2009) The long intergenic noncoding RNA (LincRNA) landscape of the soybean genome. Plant Physiol 176: 2133–2147

Goodstein DM, Shu S, Nelson RT, Cannon SB, Shoemaker RC (2009) CPC: A genomic knowledge base of the genus Vigna based on high-quality, annotated genome sequence of the azuki bean, Vigna angularis (Willd.) Ohwi & Ohashi. Plant Cell Physiol 51: e2

Gollick AA, Singh MB, Bhalla PL (2018) Functional classiﬁcation of long intergenic non-coding RNA (LincRNA) transcripts from RNA-seq data. Methods Mol Biol 2134–2147

Grant D, Nelson CT, Cannon SB, Shoemaker RC (2010) Soybase, the USDA-ARS soybean genetics and genomics database. Nucleic Acids Res 38: D843–D846

Heo JB, Sung S (2011) Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA. Science 331: 76–79

Hoagland DR, Snyder WC (1933) Nutrition of strawberry plant under controlled conditions. Proc Am Soc Hortic Sci 30: 288–294

Hu L, Xu Z, Hu B, Lu Z (2017) COME: A robust coding potential calculation tool for lncRNA identiﬁcation and characterization based on multiple features. Nucleic Acids Res 45: e2

Ingolia NT, Ghemmaghami S, Newman JRS, Weissman JS (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome proﬁling. Science 323: 218–222

Johnson R, Guigó R (2014) The RDIL hypothesis: Transposable elements as functional domains of long noncoding RNAs. RNA 20: 959–976

Kalvari I, Argasinska J, Quinones-Olvera N, Nawrocki EP, Rivas E, Eddy SR (2018) The long intergenic non-coding RNA (LincRNA) landscape of the soybean genome. Plant Physiol 176: 2133–2147

Liu J, Gough J, Rost B (2006) Distinguishing protein-coding from non-coding RNAs through support vector machines. PLoS Genet 2: e29

Marchese FP, Raimondi I, Huarte M (2017) The multidimensional mechanisms of long noncoding RNA function. Genome Biol 18: 206

Nawrocki EP, Eddy SR (2013) Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 29: 3933–3935

Negri TDC, Alves WAL, Bugatti PH, Saito PTM, Domingues DS, Paschoal AR (2018) Pattern recognition analysis on long noncoding RNAs: A tool for prediction in plants. Brief Bioinform 20: 682–689

Pedrioli PGA (2010) Trans-proteomic pipeline: A pipeline for proteomic analysis. Methods Mol Biol 604: 213–238

Perez-Riverol Y, Csordas A, Bai J, Bernal-Llanares M, Hewapathirana S, Kundu DJ, Inuganti A, Griss J, Mayer G, Eisenacher M et al (2019) The PRIDE database and related tools and resources in 2019: Improving support for quantification data. Nucleic Acids Res 47(D1): D442–D450

Pernia A, Pernia GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL (2015) StringTie enables improved reconstruction of a transcriptome based on RNA-seq reads. Nat Biotechnol 33: 290–295

Pruitt KD, Tatusova T, Maglott DR (2009) NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts, and proteins. Nucleic Acids Res 37: D61–D65

Quinlan AR, Hall IM (2010) BEDTools: A ﬂexible suite of utilities for comparing genomic features. Bioinformatics 26: 841–842

Sakai H, Naito K, Takahashi Y, Sato T, Yamamoto T, Muto I, Itoh T, Tomooka N (2016) The Vigna Genome Server, ’VgSOS’: A genomic knowledge base of the genus Vigna based on high-quality, annotated genome sequence of the azuki bean, Vigna angularis (Willd.) Ohwi & Ohashi. Plant Cell Physiol 51: e2

Sato S, Nakamura Y, Kaneko T, Asamizu E, Kato T, Nakao M, Sasamoto S, Watanabe A, Ono A, Kawashima K et al (2008) Genome structure and disease of the legume, Lotus japonicus. DNA Res 15: 227–239

Schattner P, Brooks AN, Lowe TM (2005) The TRNAScan-SE, snocan and sncps web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res 33: W686–W692

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Tholen JJ, Cheng J et al (2010) Genome sequence of the paleopolyploid soybean. Nature 463: 178–183

Schmutz J, McClean PE, Mamidi S, Wu GA, Cannon SB, Grimwood J, Jenkins J, Shu S, Song Q, Chavarro C et al (2014) A reference genome for common bean and genome-wide analysis of dual domestications. Nat Genet 46: 707–713

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Res 13: 2498–2504

Singh U, Khemka N, Rajkumar MS, Garg R, Jain M (2017) PlncPloR for prediction of long non-coding RNAs (lincRNAs) in plants and its application for discovery of abiotic stress-responsive lincRNAs in rice and chickpea. Nucleic Acids Res 45: e183

Song L, Prince S, Valliyodan B, Joshi T, Maldonado dos Santos JV, Wang J, Lin L, Wan J, Wang Y, Xu D, Nguyen HT (2016) Genome-wide transcriptome analysis of soybean primary root under varying water-deﬁcit conditions. BMC Genomics 17: 57

Sun K, Chen X, Jiang P, Song X, Wang H, Sun H (2013a) SeeRNA: Identiﬁcation of long intergenic non-coding RNA transcripts from transcriptome sequencing data. BMC Genomics 14(Suppl 2): S7

Sun L, Luo H, Bu D, Zhao G, Yu K, Zhang C, Liu Y, Chen R, Zhao Y (2013b) Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. Nucleic Acids Res 41: e166

Tang H, Krishnakumar V, Bidwell S, Rosen B, Chan A, Zhou S, Gentzibelt L, Childs KL, Yandell M, Gundlach H, et al (2014) An improved genome release (version Mt4.0) for the model legume Medicago truncatula. BMC Genomics 15: 312

Trappnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JI, Pachter L (2013) Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol 31: 46–53

Trappnell C, Pachter L, Salzberg SL (2009) TopHat: Discovering splice junctions with RNA-Seq. Bioinformatics 25: 1105–1111

Liu A, Xiao Z, Li M-W, Wong F-L, Yang W-S, Ku Y-S, Wang Q, Wang X, Xie M, Yim AK-Y et al (2018) Transcriptomic reprogramming in soybean seedlings under salt stress. Plant Cell Environ 42: 98–114

Plant Physiol. Vol. 182, 2020 1373
Lin et al.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7: 562–578

Trinklein ND, Aldred SF, Hartman SJ, Schroeder DI, Otillar RP, Myers RM (2004) An abundance of bidirectional promoters in the human genome. Genome Res 14: 62–66

UniProt Consortium (2019) UniProt: A worldwide hub of protein knowledge. Nucleic Acids Res 47(D1): D506–D515

University of Washington Proteomics Resource (2011) Protein reduction, alkylation, digestion. UWPR. https://proteomicsresource.washington.edu/docs/protocols03/UWPR_Protocols_Protein_Digestion_Protocols.pdf (April 10, 2011)

Varshney RK, Song C, Saxena RK, Azam S, Yu S, Sharpe AG, Cannon S, Baek J, Rosen BD, Tar’an B, et al (2013) Draft genome sequence of chickpea (Cicer arietinum) provides a resource for trait improvement. Nat Biotechnol 31: 240–246

Vieira LM, Grativil C, Thiebaut F, Carvalho TG, Haroim PR, Hemerly A, Lifschitz S, Ferreira PCG, Walter MEMT (2017) PlantRNA_Sniffer: An SVM-based workflow to predict long intergenic non-coding RNAs in plants. Noncoding RNA 3: 11

Waldeck N, Burkey K, Carter T, Dickey D, Song Q, Taliercio E (2017) RNA-seq study reveals genetic responses of diverse wild soybean accessions to increased ozone levels. BMC Genomics 18: 498

Wang L, Park HJ, Dasari S, Wang S, Kocher J-P, Li W (2013) CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucleic Acids Res 41: e74

Wu TD, Watanabe CK (2005) GMAP: A genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21: 1859–1875

Wucher V, Leguai F, Hédan B, Rizk G, Lagoutte L, Leeb T, Jagannathan V, Cadieu E, David A, Lohi H, et al (2017) FEELnc: A tool for long non-coding RNA annotation and its application to the dog transcriptome. Nucleic Acids Res 45: e57

Xie M, Chung CY-L, Li M-W, Wong F-L, Wang X, Liu A, Wang Z, Leung AK-Y, Wong T-H, Tong S-W, et al (2019) A reference-grade wild soybean genome. Nat Commun 10: 1216

Yang W-C, Katinakis P, Hendriks P, Smolders A, de Vries F, Spee J, van Kammen A, Bisseling T, Franssen H (1993) Characterization of GmE-NOD40, a gene showing novel patterns of cell-specific expression during soybean nodule development. Plant J 3: 573–585

Yim AK-Y, Wong JW-H, Ku Y-S, Qin H, Chan T-F, Lam H-M (2015) Using RNA-seq data to evaluate reference genes suitable for gene expression studies in soybean. PLoS One 10: e0136343