CellPress
Partner Journal

Molecular Plant
Resource Article

Large-Scale Discovery of Non-conventional Peptides in Maize and *Arabidopsis* through an Integrated Peptidogenomic Pipeline

Shunxi Wang[1,4], Lei Tian[1,4], Haijun Liu[2,4], Xiang Li[2], Jinghua Zhang[1], Xueyan Chen[1], Xingmeng Jia[1], Xu Zheng[1], Shubiao Wu[3], Yanhui Chen[1], Jianbing Yan[2,*] and Liuji Wu[1,*]

[1]National Key Laboratory of Wheat and Maize Crop Science, Collaborative Innovation Center of Henan Grain Crops, College of Agronomy, Henan Agricultural University, Zhengzhou 450002, China

[2]National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan 430070, China

[3]School of Environmental and Rural Science, University of New England, Armidale, NSW 2351, Australia

[4]These authors contributed equally to this work.

***Correspondence: Jianbing Yan (yjianbing@mail.hzau.edu.cn), Liuji Wu (wlj200120@163.com)**

https://doi.org/10.1016/j.molp.2020.05.012

ABSTRACT

Non-conventional peptides (NCPs), which include small open reading frame-encoded peptides, play critical roles in fundamental biological processes. In this study, we developed an integrated peptidogenomic pipeline using high-throughput mass spectra to probe a customized six-frame translation database and applied it to large-scale identification of NCPs in plants.A total of 1993 and 1860 NCPs were unambiguously identified in maize and *Arabidopsis*, respectively. These NCPs showed distinct characteristics compared with conventional peptides and were derived from introns, 3′ UTRs, 5′ UTRs, junctions, and intergenic regions. Furthermore, our results showed that translation events in unannotated transcripts occur more broadly than previously thought. In addition, we found that dozens of maize NCPs are enriched within regions associated with phenotypic variations and domestication selection, indicating that they potentially are involved in genetic regulation of complex traits and domestication in maize. Taken together, our study developed an integrated peptidogenomic pipeline for large-scale identification of NCPs in plants, which would facilitate global characterization of NCPs from other plants. The identification of large-scale NCPs in both monocot (maize) and dicot (*Arabidopsis*) plants indicates that a large portion of plant genome can be translated into biologically functional molecules, which has important implications for functional genomic studies.

Keywords: non-conventional peptides, small open reading frames, peptidogenomics, mass spectrometry, six-frame translation, plants

Wang S., Tian L., Liu H., Li X., Zhang J., Chen X., Jia X., Zheng X., Wu S., Chen Y., Yan J., and Wu L. (2020). Large-Scale Discovery of Non-conventional Peptides in Maize and *Arabidopsis* through an Integrated Peptidogenomic Pipeline. Mol. Plant. **13**, 1078–1093.

INTRODUCTION

Peptides, typically composed of 2–100 amino acid residues, represent small biological molecules with important roles in biology (Tavormina et al., 2015). Small signaling peptides or peptide hormones, which are a class of short peptides ranging from 5 to 75 amino acids in length, also play critical roles in various biological processes. For example, the discovery and application of the peptide hormone insulin was one of the greatest achievements of the 20th century (Banting and Best, 2007). Studies over the past few decades have mainly focused on conventional peptides (CPs) derived from annotated coding sequences (CDSs) or conventional open reading frames (ORFs). Recently, a novel class of peptides, now defined as non-conventional peptides (NCPs) in this study, has caught significant attention as functionally important endogenous peptides in various organisms (Ma et al., 2014; Couso and Patraquim, 2017; Plaza et al., 2017; Jackson et al., 2018; Chen et al., 2020a). These NCPs are derived from previously unannotated CDSs, such as intergenic regions, untranslated regions (UTRs), introns, and various types of junctions, as well as different reading frames from annotated CDSs.

Published by the Molecular Plant Shanghai Editorial Office in association with Cell Press, an imprint of Elsevier Inc., on behalf of CSPB and IPPE, CAS.

A primary report of an NCP was published more than two decades ago, where a 10-amino-acid peptide was identified to be translated from *ENOD40*, a gene previously annotated as untranslated (van de Sande et al., 1996). Thereafter, the ENOD40 was further proved to play a key role in regulating the response to auxin in flowering plants (Rohrig et al., 2002). In animals and humans, NCPs are known to play important roles in a diverse range of cellular processes, such as calcium transport (Magny et al., 2013), embryogenesis (Kondo et al., 2010), muscle performance (Nelson et al., 2016; Matsumoto et al., 2017), translation control (Hinnebusch et al., 2016; Couso and Patraquim, 2017; Plaza et al., 2017), immune response (Laumont et al., 2016), and stress resistance (Khitun et al., 2019). Functional NCPs, such as POLARIS (Casson et al., 2002), ROTUNDIFOLIA4 (Narita et al., 2004), KOD (Blanvillain et al., 2011), OSIP108 (De Coninck et al., 2013), miPEP165a (Lauressergues et al., 2015), PSEP1, PSEP3, PSEP18, and PSEP25 (Fesenko et al., 2019), CDC26 (Lorenzo-Orts et al., 2019), and vvi-miPEP171d1 (Chen et al., 2020b), have been reported in plants. These studies have demonstrated that NCPs play essential roles in plant development, environmental responses, and translational control. However, due to the limitations of genomic annotation and peptidomic technology, a plethora of NCPs are usually dismissed from further analysis or annotation in plants (Andrews and Rothnagel, 2014; Yin et al., 2019).

The increasing importance of NCPs has led to emerging strategies for their discovery. The advent of next-generation sequencing and developments in bioinformatics has boosted NCP research at a genome-wide scale. Computational approaches based on sequence similarity have been developed to identify potential translational small ORFs (sORFs) in non-coding sequences (Hurst, 2002; Kastenmayer et al., 2006; Hanada et al., 2007; Makarewich and Olson, 2017). However, conservation and homology analysis of sORFs is difficult due to their short sequences and low conservation scores. Another strategy is to perform ribosome profiling by sequencing ribosome-protected fragments, which enables mapping of a genome-wide set of transcripts that are being translated (Ingolia et al., 2009, 2011; Ingolia, 2016; Shiber et al., 2018). In recent years, ribosome profiling has been widely used to confirm the translation of non-annotated ORFs in various species (Ruiz-Orera et al., 2014; Wu et al., 2019; Kurihara et al., 2020). While ribosome profiling itself is an experimental approach, the evaluation of the coding potential of an identified region of interest is in fact mostly computational (Makarewich and Olson, 2017). Existing ribosome profiling techniques have undergone significant modifications and enhancements, which have improved reliably in protein-coding transcript identification (Hsu et al., 2016; Bazin et al., 2017). A different strategy, a mass spectrometry (MS)-based method, is able to detect peptides that are translated from an sORF and can thereby directly validate the protein-coding potential of the transcript (Castellana et al., 2008; Makarewich and Olson, 2017). Recently, a new method referred to as peptidogenomics, which integrates peptidomics (based on high-throughput tandem MS) and genomics, has emerged as a promising strategy for deep analysis of the endogenous NCPs (Kersten et al., 2011; Harvey et al., 2015). Peptidogenomics is an efficient strategy that has been successfully used in micro-organisms and humans (Liu et al., 2011; Slavoff et al., 2013; Mohimani and Pevzner, 2016; Mohimani et al., 2018). However, owing to experimental and computational issues, such as challenges associated with endogenous peptide enrichment, non-specific protease digestion, and lack of complete peptide reference databases, the identification of NCPs using peptidogenomics in plants is still challenging.

Here, we developed an integrated peptidogenomic pipeline for large-scale identification of NCPs in monocot and dicot plants. High-throughput mass spectra of endogenous peptides were used to probe the Ensembl protein database and a customized peptidogenomic database derived from the six-frame translation of genomic sequences. Our results revealed that NCPs could be derived from not only coding sequences but also allegedly non-coding sequences, which show a distribution pattern distinct from that of CPs. The results obtained through large-scale identification of endogenous NCPs in plants thus provide valuable information toward understanding biological functions of these hidden molecules.

RESULTS

An Integrated Peptidogenomic Pipeline for NCP Identification in Plants

Direct detection of NCPs is the most definitive evidence of their existence. To facilitate plant NCP discovery, we developed and applied an integrated peptidogenomic pipeline for large-scale identification of plant NCPs (Figure 1A). For sample preparation, an acid extraction buffer consisting of 1% trifluoroacetic acid (TFA) was utilized based on a previous study (Chen et al., 2014). In addition, heat stabilization in a water bath combined with plant protease inhibitors was applied to diminish non-specific protease digestion. Trichloroacetic acid (TCA)–acetone precipitation was also applied to establish an optimized sample preparation protocol. Plant endogenous peptides were then enriched from larger protein fragments by centrifugation through 10-kDa cutoff filters before being analyzed by liquid chromatography–tandem MS (LC–MS/MS).

To capture the endogenous peptides globally present in maize, we used the Mascot search engine to match the resulting mass spectrum dataset against the Ensembl protein database and a customized peptidogenomic database. The customized peptidogenomic database was constructed using the six-frame translation of maize genomic sequences (Figure 1B). As a result, we obtained a ~5.2-gigabase (Gb) customized peptidogenomic database (containing ~136 million sequences). To avoid an inflated search space for the spectral sequences, we stored the information collected for every peptide (including the encoding schemes and genomic locus) in an index file with the peptide's data. This reduced significantly the digital memory required to store our sequence data. In addition, based on the locus-tracking approach, we used an automated process to map the peptide spectra to genomic loci, which enabled more effective use of the pipeline for large-scale discovery of NCPs.

Large-Scale Identification of CPs and NCPs in Maize

All the reliably identified peptides from the Ensembl protein and customized peptidogenomic databases were combined and

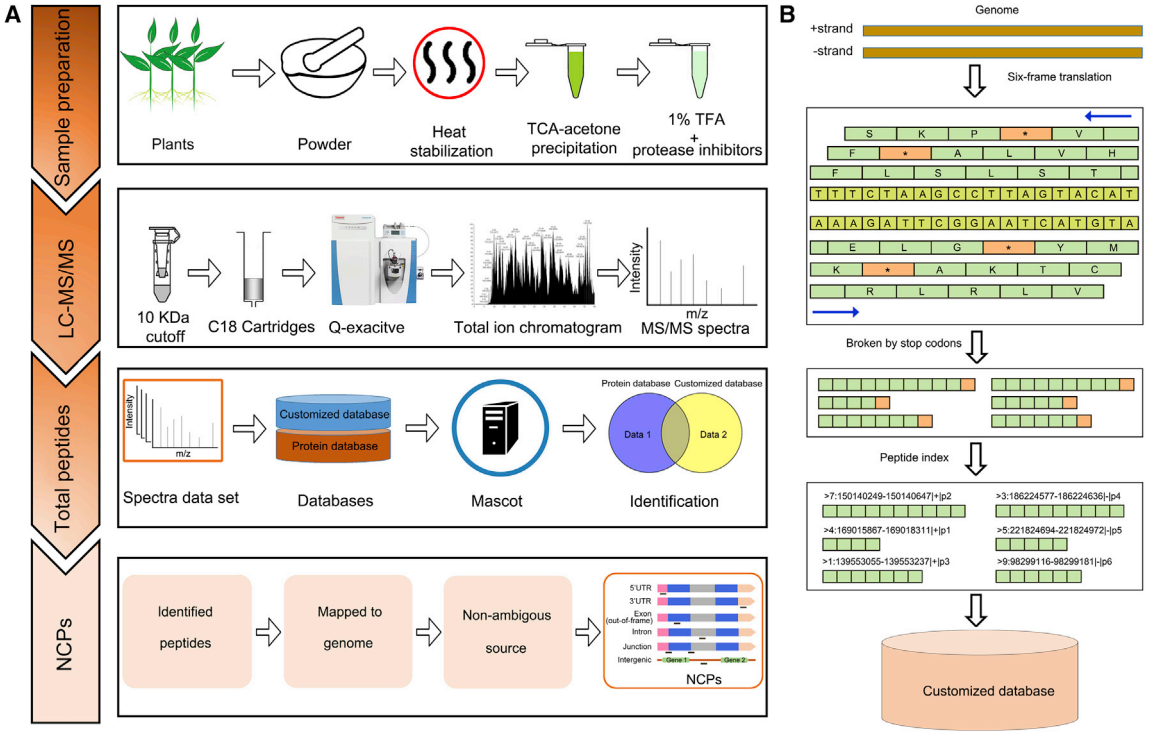

Figure 1. Peptidogenomic Workflow for Plant NCP Identification.

(A) The peptidogenomic workflow for plant NCP identification. Endogenous peptides from plant leaves were extracted using an optimized protocol. Heat stabilization by incubating samples in a water bath at 95°C combined with use of an acid extraction buffer containing 1% TFA and plant protease inhibitors was applied to minimize peptide degradation during peptide extraction. Endogenous plant peptides were enriched from larger protein fragments by centrifugation through 10-kDa cutoff filters. The peptides were analyzed on a high-resolution and high-accuracy mass spectrometer. MS/MS spectra data were searched against a customized peptidogenomic database and the Ensembl protein database using the Mascot search engine. The resulting peptides were used to filter out the CPs and thus obtain the NCPs.

(B) Customized peptidogenomic database construction. The complete maize genomic sequence was downloaded from Ensembl Plants in FASTA format and translated into six-frame format using the EMBOSS:6.6.0 package. The translation of the genomic DNA started from the first, second, and third nucleotides on each strand of each chromosome and ended when a stop codon was encountered. Triplets were translated according to the standard genetic code to assign a one-letter symbol for each amino acid and an asterisk for a stop codon. A peptide index file containing genomic coordinates and orientations (e.g., >7:150140249-150140647|+|p2) was assigned to each peptide sequence.

used to identify both CPs and NCPs. In total, 748 and 3932 non-redundant peptides were identified based on the Ensembl protein database and customized peptidogenomic database, respectively (Figure 2A; Supplemental Tables 1 and 2). Of these, 3315 peptides were specifically identified in the customized peptidogenomic database (Figure 2A). By mapping these peptides to genome loci and applying a series filtering steps (see Methods), a total of 2837 endogenous peptides were then unambiguously assigned to a single genomic locus; 1993 (70.3%) NCPs (Supplemental Table 3) and 844 (29.7%) CPs (Figure 2B and Supplemental Table 4) were identified. The median length of CPs was 16 amino acids while that of NCPs was 12 amino acids, and the difference in length was significant (Figure 2C). Approximately 90% of the CPs and NCPs were less than 23 amino acids and 16 amino acids, respectively (Supplemental Figure 1). Furthermore, the average molecular weight of NCPs was 1325.22 Da, with 99.25% (1978) of peptides having a molecular weight less than 2500 Da. By contrast, the average molecular weight of CPs was 1742.16 Da, with 91.94% (776) of peptides having a molecular weight less than 2500 Da (Figure 2D and 2E). These results indicated that NCPs constitute a significant portion of the plant peptidome and show different characteristics compared with CPs.

CP and NCP Distribution Patterns

Both CPs and NCPs were found to be unevenly distributed on the chromosomes of maize (Figure 3A). For CPs, most peptides were distributed near the telomeres, whereas NCPs were homogeneously located between the centromeres and telomeres of each maize chromosome (Figure 3B). Furthermore, a total of 138 hot regions (defined by 6-Mb windows; see Methods) were discovered (Figure 3A). A total of 58 CP hotspot regions containing 446 (52.84%) peptides were observed, whereas 81 NCP hotspot regions containing 545 (27.35%) peptides were found (Figure 3A). Among these hotspot regions, one located in chromosome 5 was common to both CPs and NCPs. Additionally, the number of NCPs in each chromosome was positively correlated with the chromosomal length (r = 0.07; p = 0.0099), but no correlation between the number of CPs and chromosomal length was detected (Figure 3C).

The interval between two adjacent peptides could be used to accurately define peptide coverage over the genome. We found that 74.88% (632) of CPs were less than 500 kb apart, whereas only 39.74% (792) of NCPs were within 500 kb of each other

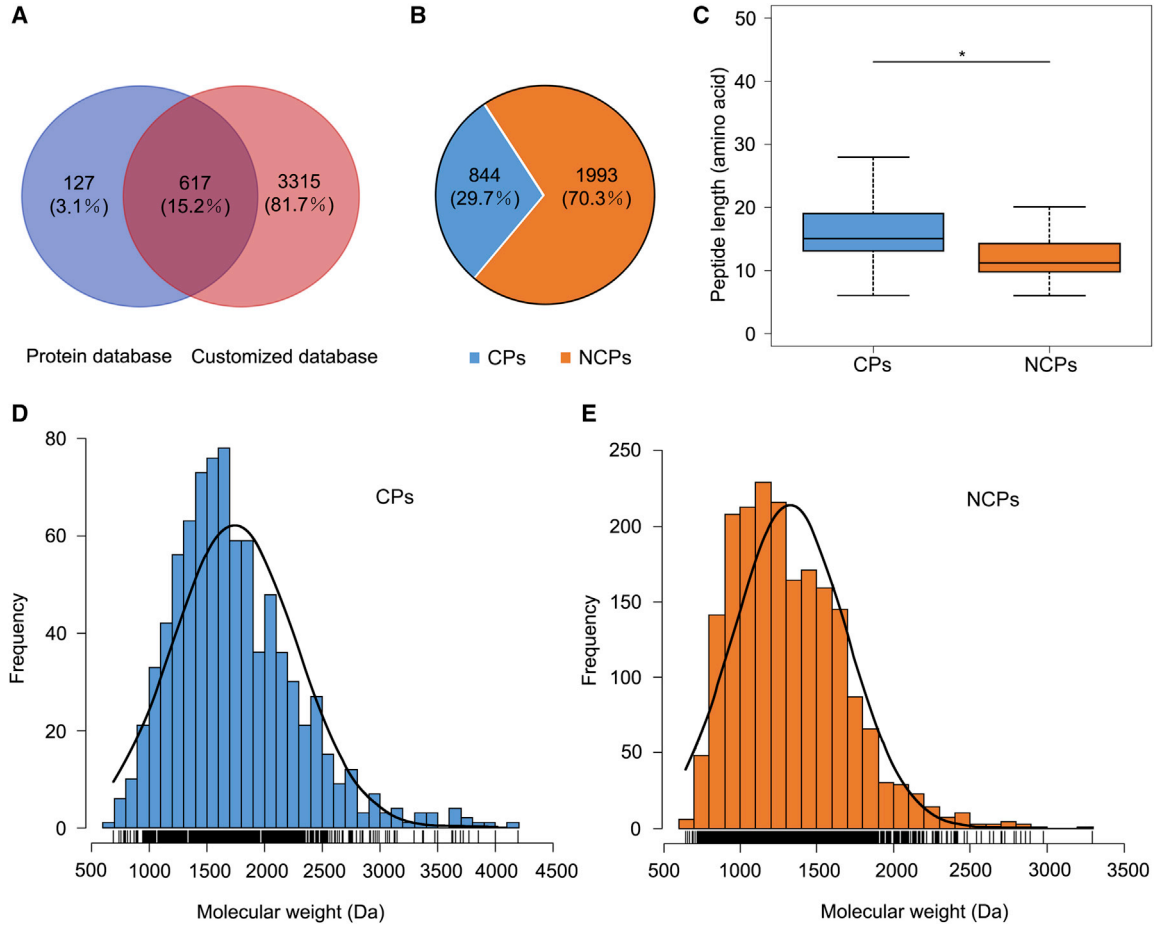

Figure 2. Overview of NCPs Identified in Maize by the Developed Peptidogenomic Pipeline.

(A) Venn diagram showing the number of peptides identified by searches against the Ensembl protein and customized peptidogenomic databases. The areas shown in the diagram are not proportional to the number of peptides in each group.

(B) Number of CPs and NCPs identified through peptidogenomic analysis.

(C) Length of CPs and NCPs. Boxes represent the second and third quartiles and whiskers represent 1.5× interquartile range (IQR). Fisher's exact test was used for hypothesis testing $^{*}p < 0.05$.

(D) Molecular weight distribution of CPs ($n = 844$). The rug plot above the x axis represents the distribution of observations.

(E) Molecular weight distribution of NCPs ($n = 1993$). The rug plot above the x axis represents the distribution of observations.

(Figure 3D). We then compared the locations of these peptides with gene models, whereby 798 (94.55%) CPs were found to be located in regions less than 2 kb from canonical translation start site (TSS); in contrast, this value was 336 (16.86%) for NCPs (Figure 3E). These results reveal the widespread existence of translated NCPs in the genome and the distinct distribution patterns of CPs and NCPs.

To gain further insights into the mechanisms responsible for the generation of CPs and NCPs, we analyzed the nucleotide sequences of CP and NCP source transcripts to predict their TSSs. We observed a preponderance of non-AUG TSSs in both CPs and NCPs (Supplemental Tables 3 and 4). Although it was long thought that eukaryotic translation almost always initiates at the AUG start codon, our results reveal that non-AUG start codons are used at an astonishing frequency. This finding is consistent with the finding of previous peptidomics studies that more than 90% of endogenous peptides start with a non-AUG codon (Chen et al., 2014; Secher et al., 2016; Corbiere et al., 2018). This result is also consistent with those of ribosome profiling and MS studies, which demonstrated that most ORFs contain non-AUG start sites (Ingolia et al., 2011; Slavoff et al., 2013; Na et al., 2018).

NCPs Derived from Coding and Non-coding Sequences

By analyzing their origins, 952 (47.77%) NCPs were assigned to the reverse strand in maize (Figure 4A). Next, by analyzing the location of the NCPs within their respective gene sources, 1708 (85.70%) NCPs were found to be derived from intergenic regions, 139 (6.97%) from introns, 89 (4.47%) from out-of-frame exons, 25 (1.25%) from 3′ UTRs, 18 (0.90%) from 5′ UTRs and 14 (0.70%) from junctions (5′ UTR–exon or intron–exon) (Figure 4B). These results highlight the translation evidence for these allegedly non-coding sequences.

Length analysis showed that the average lengths of NCPs derived from intergenic regions and out-of-frame exons were longer than those derived from junctions (Figure 4C). The NCPs derived from

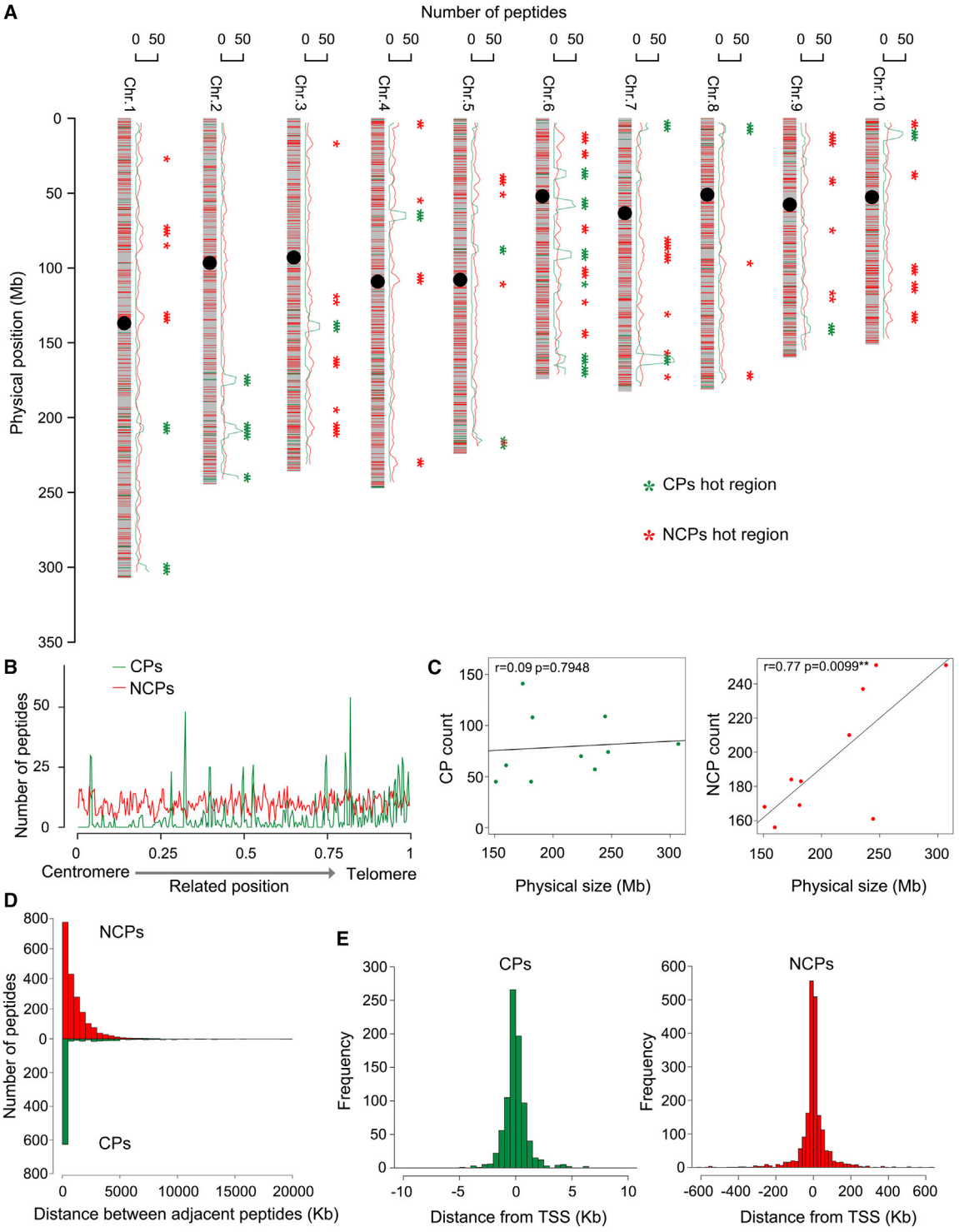

Figure 3. CP and NCP Distribution in Maize.

(A) Genome-wide distribution of CPs (green) and NCPs (red). For each chromosome, the peptide distribution pattern includes three columns. Left: CPs (green) and NCPs (red) mapped onto chromosomes. Black circles are the centromeres. Middle: CP (green) and NCP (red) distribution patterns determined using a window size of 6-Mb and 3-Mb steps based on B73 reference genome. Right: distribution of CP (green) and NCP (red) hotspot regions. Hotspot regions were defined as more than 10 peptides in a window size of 6 Mb.

(B) Normalized distribution of CPs and NCPs shown along the chromosomal arms. The *x* axis represents the normalized length of each arm with the centromere set to "0" and the telomere to "1". The *y* axis reports the number of CPs (green) and NCPs (red).

(C) Correlations between CP or NCP counts and chromosomal length (Pearson correlation: CPs, $r = 0.09$, $p = 0.7948$; NCPs, $r = 0.77$, $^{**}p = 0.0099$).

(D) Histogram of the distances between two of adjacent CPs or NCPs.

(E) Histograms showing the distance from each CP or NCP to the closest TSS.

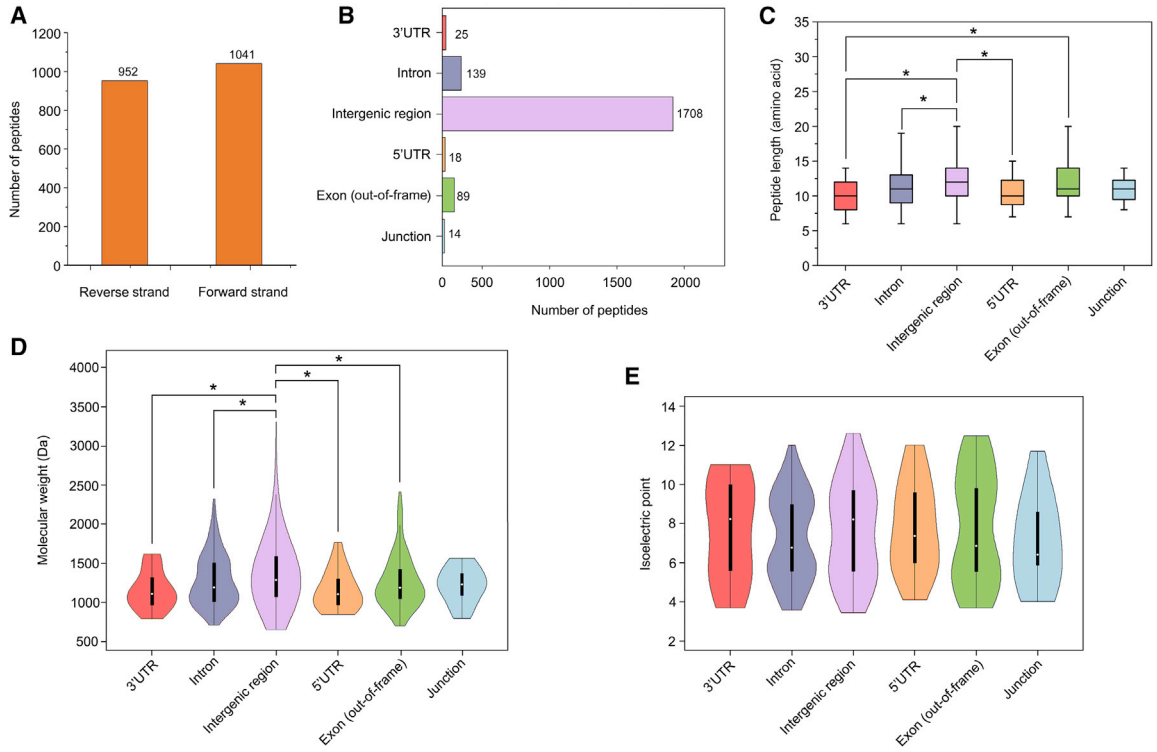

Figure 4. Characteristics of Maize NCPs.

(A) Number of NCPs derived from forward and reverse strands.

(B) Number of NCPs derived from different gene elements.

(C–E) **(C)** Length of NCPs derived from different gene elements. Boxes represent the second and third quartiles and whiskers represent 1.5× IQR. Fisher's exact test was used for hypothesis testing, $*p < 0.05$. Violin plots combine box plot and kernel density trace to describe the distribution patterns of molecular weight **(D)** and isoelectric point **(E)**. Tomato: NCPs derived from 3′ UTRs ($n = 25$); beige: NCPs derived from introns ($n = 139$); lilac: NCPs derived from intergenic regions ($n = 1708$); yellow: NCPs derived from 5′ UTRs ($n = 18$); green: NCPs derived from out-of-frame exons ($n = 139$); light blue: NCPs derived from junctions ($n = 14$). The black bars and thin lines within the violin plots represent the IQRs and the entire data ranges, respectively. White dots in the center indicate the average values. The width of the violin plot represents the density of the distribution. Fisher's exact test was used for hypothesis testing, $*p < 0.05$.

3′ UTRs and 5′ UTRs were the two shortest (Figure 4C). Molecular weight distribution analysis showed that more than 70% (1407) of NCPs were less than 1500 Da. The average molecular weight of NCPs derived from intergenic regions was higher than that of NCPs derived from introns, out-of-frame exons, 5′ UTRs, and 3′ UTRs (Figure 4D and Supplemental Figure 2A). There was no significant difference among the average isoelectric point (p*I*) values of NCPs derived from 3′ UTRs, introns, intergenic regions, 5′ UTRs, out-of-frame exons, and junctions (Figure 4E and Supplemental Figure 2B). Taken together, these results indicate that the identified NCPs have a wide range of physicochemical properties and that NCPs derived from different gene elements show different characteristics.

Verification and Validation of NCPs

To verify these identified NCPs, we assigned these peptides to their respective genomic locus. For example, NCP RMDAHALR was derived from the 5′ UTR of the gene *Zm00001d029555* (Figure 5A), and NCP ILTVNLKP was derived from the 3′ UTR of the gene *Zm00001d050172* (Figure 5B). In addition to NCPs derived from UTRs, we also found a large number of NCPs derived from intergenic regions and introns. For example, NCP QISVELPGVV was derived from the intergenic region between genes *Zm00001d024336* and *Zm00001d024337* (Figure 5C). NCP EGTPKAVGHRQ was derived from the intron of the gene *Zm00001d008363* (Figure 5D). Next, 115 NCPs were synthesized experimentally. MS analysis was performed under the same conditions as were used for peptidogenomic analysis in this study. As shown in Figure 5A–5D, the spectra of synthetic peptides RMDAHALR, ILTVNLKP, QISVELPGVV, and EGTPKAVGHRQ agreed with the spectral data generated from the peptidogenomic analysis. Verification of the other 111 NCPs is shown in Supplemental Data 1.

In addition, we performed transcriptomic analyses using published RNA-sequencing (RNA-seq) data from maize. These RNA-seq data include circular RNAs, long non-coding RNAs (lncRNAs), mRNAs, and small RNAs. Most NCPs (1806, 90.62%) identified in the current study were supported by evidence from these published databases (Supplemental Table 3). Among these NCPs, 1652 were derived from lncRNA and 859 from circular RNA (Supplemental Table 3). The results indicated that these identified NCPs were likely produced from allegedly non-coding sequences.

Lastly, to validate the identified NCPs with independent methods, the available ribosome profiling datasets of maize

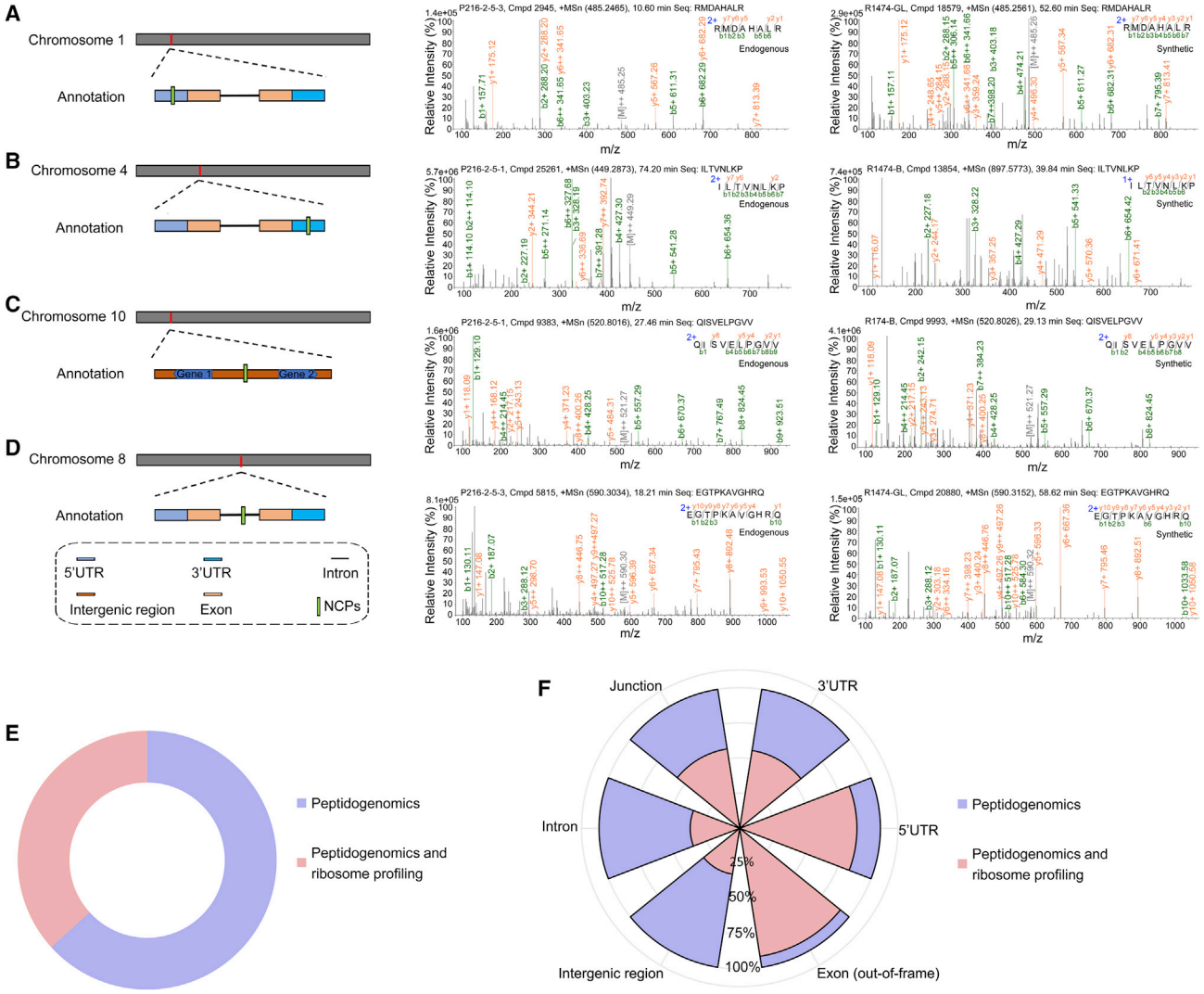

Figure 5. Verification and Validation of Maize NCPs.

(A) NCP RMDAHALR mapped to the 5′ UTR of a gene in chromosome 1 (left). Verification of this NCP by comparing the spectra of the peptide identified by the integrative peptidogenomic pipeline (middle) with those of the synthetic peptide (right).

(B) NCP ILTVNLKP mapped to the 3′ UTR of a gene in chromosome 4 (left). Verification of this NCP by comparing the spectra of the peptide identified by the integrative peptidogenomic pipeline (middle) with those of the synthetic peptide (right).

(C) NCP QISVELPGVV mapped to the intergenic region between two genes in chromosome 10 (left). Verification of this NCP by comparing the spectra of the peptide identified by the integrative peptidogenomic pipeline (middle) with those of the synthetic peptide (right).

(D) NCP EGTPKAVGHRQ mapped to the intron of a gene in chromosome 8 (left). Verification of this NCP by comparing the spectra of the peptide identified by the integrative peptidogenomic pipeline (middle) with those of the synthetic peptide (right).

(E) Percentages of NCPs detected by peptidogenomics and ribosome profiling.

(F) Percentages of NCPs derived from different gene elements detected by peptidogenomics and ribosome profiling.

were analyzed. Ribosome profiling, also known as Ribo-seq (ribosome sequencing), is a method based on deep sequencing of ribosome-protected fragments. In agreement with translation being the intermediate step between transcription and the proteome, ribosome profiling data are more highly predictive of final protein expression than mRNA-seq data (van Heesch et al., 2019). The ribosome profiling analysis showed that 732 (36.73%) NCPs detected by peptidogenomics were also uncovered by ribosome profiling (Figure 5E and Supplemental Table 5). This overlap in validation rate, 36.73%, between these two methods is consistent with previous reports (Samandi et al., 2017; van Heesch et al., 2019; Chen et al., 2020a). Among these NCPs, 564 were derived from intergenic regions, 82 from out-of-frame exons, 49 from introns, 15 from 5′ UTRs, 14 from 3′ UTRs, and eight from junctions. The proportions of the NCPs detected by both methods to those detected by peptidogenomic analysis were 33.02% for NCPs from intergenic regions, 92.13% from out-of-frame exons, 35.25% from introns, 83.33% from 5′ UTRs, 56.00% from 3′ UTRs, and 57.14% from junctions (Figure 5F). These NCPs, which were detected by two different methods, provide a high-confidence collection of NCPs for further studies. We speculate that those NCPs that were detected only by

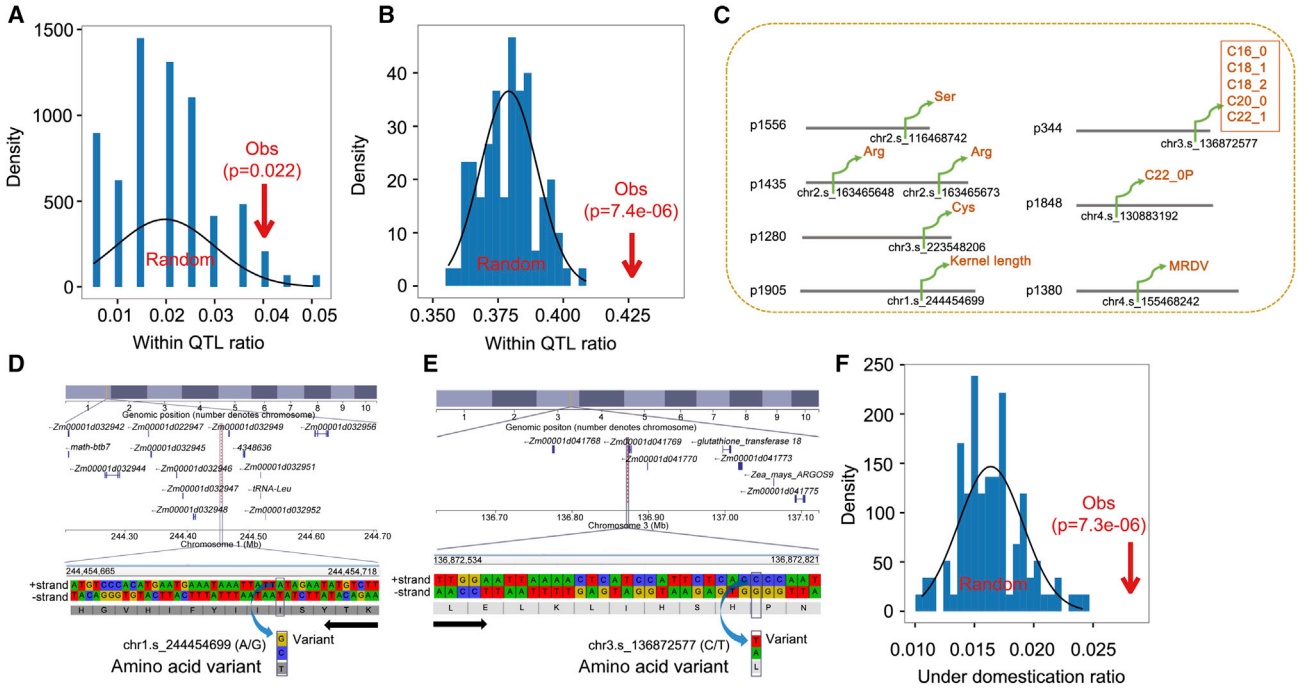

Figure 6. Quantitative Trait Loci Significantly Associated with Phenotypic Traits Linked to Maize NCPs.

(A) Enrichment of NCPs within QTLs.

(B) Enrichment of NCPs located within 20-kb flanking regions of significant SNPs.

(C) Diagram showing the distribution of significant SNPs associated with plant traits within NCPs: one SNP associated with kernel length, one with disease (maize rough dwarf virus, MRDV), two with oil content, and four with amino acid content.

(D) An isoleucine–threonine transition caused by a SNP (chr1.s_244454699, A>G; $p < 9.42e{-}5$) associated with kernel length. Significant SNPs are indicated by red dotted lines. The black arrow indicates the NCP derived from the reverse strand.

(E) Oil content associated with a significant SNP (chr3.s_136872577, C>T; $p < 2.09e{-}07$) that leads to a proline-to-leucine substitution in the NCP. The black arrow shows that the NCP was derived from the forward strand.

(F) Enrichment of NCPs within regions under positive selection during maize domestication. The *x* axis shows the ratio of overlap between the associated SNPs and the NCPs (Obs), and that between the associated SNPs and randomly generated regions (Random). *p* values for upper-tail test were calculated using the "pnorm" function implemented in R (lower.tail = FALSE).

peptidogenomics were either erroneous calls or stable peptides from unstable RNAs.

NCPs Are Enriched in Regions Associated with Phenotypic Variations and Domestication Selection

In maize, coding regions only constitute a small fraction of the whole genome, and the vast majority of the genome has been considered non-coding. Genome-wide association studies and quantitative trait locus (QTL) analysis have identified many functional elements in the non-coding regions in maize (Liu et al., 2017). The fact that 1993 (70.3%) NCPs were derived from non-coding sequences prompts us to believe that they are of significant functional relevance. Therefore, we examined the enrichment of these NCPs within identified QTLs underlying various traits, and within those regions presumed to be under domestication selection.

Compared with randomly selected genomic sequences with same distance distribution and number (see Methods), single-nucleotide polymorphisms (SNPs) significantly associated with plant traits appeared to be significantly enriched within the regions of NCPs ($p < 0.02$, upper-tail test; Figure 6A and Supplemental Table 6). Considering the presence of genetic linkage in association mapping, we further extended the positions of associated SNPs to the flanking 20-kb regions. Statistical analysis showed that these NCPs were more significantly enriched at the QTL regions compared with random regions ($p < 7.4e{-}06$; Figure 6B and Supplemental Table 7). Among the significantly enriched SNPs, several were found to be exactly located within NCPs, and these SNPs showed associations with various phenotypes including kernel length, disease (maize rough dwarf virus [MRDV]), and oil and amino acid contents (Figure 6C and Supplemental Table 6). For instance, an isoleucine–threonine transition at one significant SNP (chr1.s_244454699, A>G; $p < 9.42e{-}5$) associated with kernel length was located within the NCP KTYSIIIYFIHVGH, which was mapped to the non-coding region 13 kb upstream of the gene *Zm00001d032949* (Uncharacterized) (Figure 6D). Another significant SNP (chr3.s_136872577, C > T; $p < 2.09e{-}07$) related to oil content, resulting in a transition from proline to leucine, was associated with the NCP LELKLIHSHPN, which was mapped to the non-coding region 5 kb upstream of the gene *Zm00001d041769* (Figure 6E). These results reveal the potential functions of these NCPs in the regulation of plant phenotypes.

The relationship between domestication and NCPs was also investigated. Compared with randomly selected genomic

sequences with the same distance distribution and number, NCPs were enriched within candidate regions that are associated with domestication selection ($p < 7.3e{-}6$, Upper-tail test; Figure 6F). A total of 55 NCPs were identified within domestication candidate regions (Supplemental Table 8). While further validation is highly needed to explore exactly which domesticated traits are affected and what indeed is the mechanism, this result, for the first time as far as we know, unveils the likely involvement of NCPs during domestication, providing another hidden layer of functional importance in NCPs.

The Applicability of the Peptidogenomics Pipeline to Identify NCPs in *Arabidopsis*

To extend this pipeline to other plants, we used the dicot model plant *Arabidopsis* to test the wider applicability of the peptidogenomic method. As a result, 2353 and 3871 non-redundant peptides were identified using the Ensembl protein database and customized peptidogenomic database (Supplemental Tables 9 and 10), respectively. Of these, 2270 peptides were specifically identified using the customized peptidogenomic database (Figure 7A). In total, 1860 (44.04%) NCPs (Supplemental Table 11) and 2363 (55.96%) CPs were obtained in *Arabidopsis* (Supplemental Table 12). The median length of NCPs was 11 amino acids, which was shorter than that of CPs (13 amino acids) (Figure 7B). Furthermore, the average molecular weight of NCPs (1208.34 Da) was lower than that of CPs (1420.89 Da) (Supplemental Figure 3). In addition, we found that the NCPs identified in *Arabidopsis* were shorter and had a lower molecular weight than those in maize (Supplemental Table 13).

By analyzing the origins of NCPs, we found that 943 (50.70%) NCPs were from the reverse strand (Figure 7C). By analyzing the locations of the NCPs within their respective gene sources, we found that 666 (35.81%) NCPs were derived from intergenic regions, 239 (12.85%) from introns, 651 (35.00%) from out-of-frame exons, 91 (4.89%) from 3′ UTRs, 63 (3.39%) from 5′ UTRs, and 150 (8.06%) from junctions (Figure 7D). The number of NCPs derived from intergenic regions in *Arabidopsis* was lower than that in maize, whereas the number of NCPs from other gene elements in *Arabidopsis* was higher than that in maize (Supplemental Table 13). Length analysis showed that the average length of NCPs derived from 3′ UTRs was the longest and that of NCPs from introns the shortest (Figure 7E). The average molecular weight of NCPs derived from out-of-frame exons was higher than that of NCPs from 5′ UTRs and intergenic regions (Figure 7F and Supplemental Figure 4A). The average p*I* value of NCPs derived from out-of-frame exons and junctions was higher than that of NCPs from introns (Figure 7G and Supplemental Figure 4B). These results together indicate that the developed peptidogenomic pipeline can also be used in dicot plants such as *Arabidopsis*, and that the translation of unannotated transcripts is widespread in both monocot and dicot plants, although they may have different translation patterns.

DISCUSSION

Endogenous peptides are mainly generated by protein degradation, gene encoding, and gene-independent enzymatic formation *in vivo* (Peng et al., 2020). The emergence of peptidomics makes large-scale identification of endogenous peptides extracted from tissues possible (Slavoff et al., 2013; Secher et al., 2016). However, peptidomics studies can be particularly challenging due to non-specific protease digestion during sample preparation (Farrokhi et al., 2008; Secher et al., 2016). Despite the wide use of protease inhibitors in plant peptide extraction, studies in animals and humans have demonstrated that protease inhibitors are not sufficiently effective in preventing peptide degradation (Svensson et al., 2003; Parkin et al., 2005). Recently, heat stabilization, such as focused microwave radiation, integrated with protease inhibitors has been successfully used in animals to minimize proteolytic activity prior to peptide isolation (Secher et al., 2016). However, similar attempts have not been made in plants so far.

Plant cells are more complex than animal cells due to the presence of additional components such as a cell wall, large vacuoles, and chloroplasts, making the isolation of complete endogenous peptides in plants more challenging. In this study, in addition to the combination of heat stabilization using a water bath and plant protease inhibitors to minimize non-specific protease digestion during peptide extraction, TCA–acetone precipitation was also included in the extraction protocol. TCA–acetone precipitation is very useful for removing interfering compounds, such as polysaccharides, polyphenols, pigments, and lipids, in plants (Mechin et al., 2007). Therefore, this step can help limit the interference of non-protein or non-peptide compounds during endogenous peptide extraction. We speculate that the protease-associated non-specific degradation during peptide extraction will be a long-lasting issue as there is no effective extraction protocol to completely prevent this from occurring. Therefore, more efforts should be made to develop a more effective peptide extraction protocol that can retain endogenous peptides in the same states as they are *in vivo* for peptidomics study. In addition, it should be noted that the peptides derived from protein degradation within the cell are also another type of endogenous peptide in addition to those encoded by genes(Peng et al., 2020). Protein degradation ubiquitously occurs in living organisms, and the level of enzymatic degradation of proteins is closely related to precursor protein status and enzyme activity in living organisms (Rubinsztein, 2006). Therefore, peptidomic data are also a good resource for the assessment of the potential protease/peptidase activity involved in the hydrolysis process, although this topic is beyond the scope of this study.

Standard peptidomics approaches identify peptides by matching experimentally observed spectra to databases of predicted spectra based on annotated genes. However, such an approach would not identify NCPs. The most effective strategy to do so is to integrate peptidomics with the six-frame translation of a genome, which is referred to as peptidogenomics (Kersten et al., 2011; Slavoff et al., 2013). A database derived from the six-frame translation of the entire genome can be used to identify peptides encoded in any genomic region (Castellana et al., 2014; Nesvizhskii, 2014; Yang et al., 2018). Peptidogenomics has already proved its value in identifying peptides at the genome scale in micro-organisms and humans (Kersten et al., 2011; Liu et al., 2011; Nguyen et al., 2013; Slavoff et al., 2013; Mohimani and Pevzner,

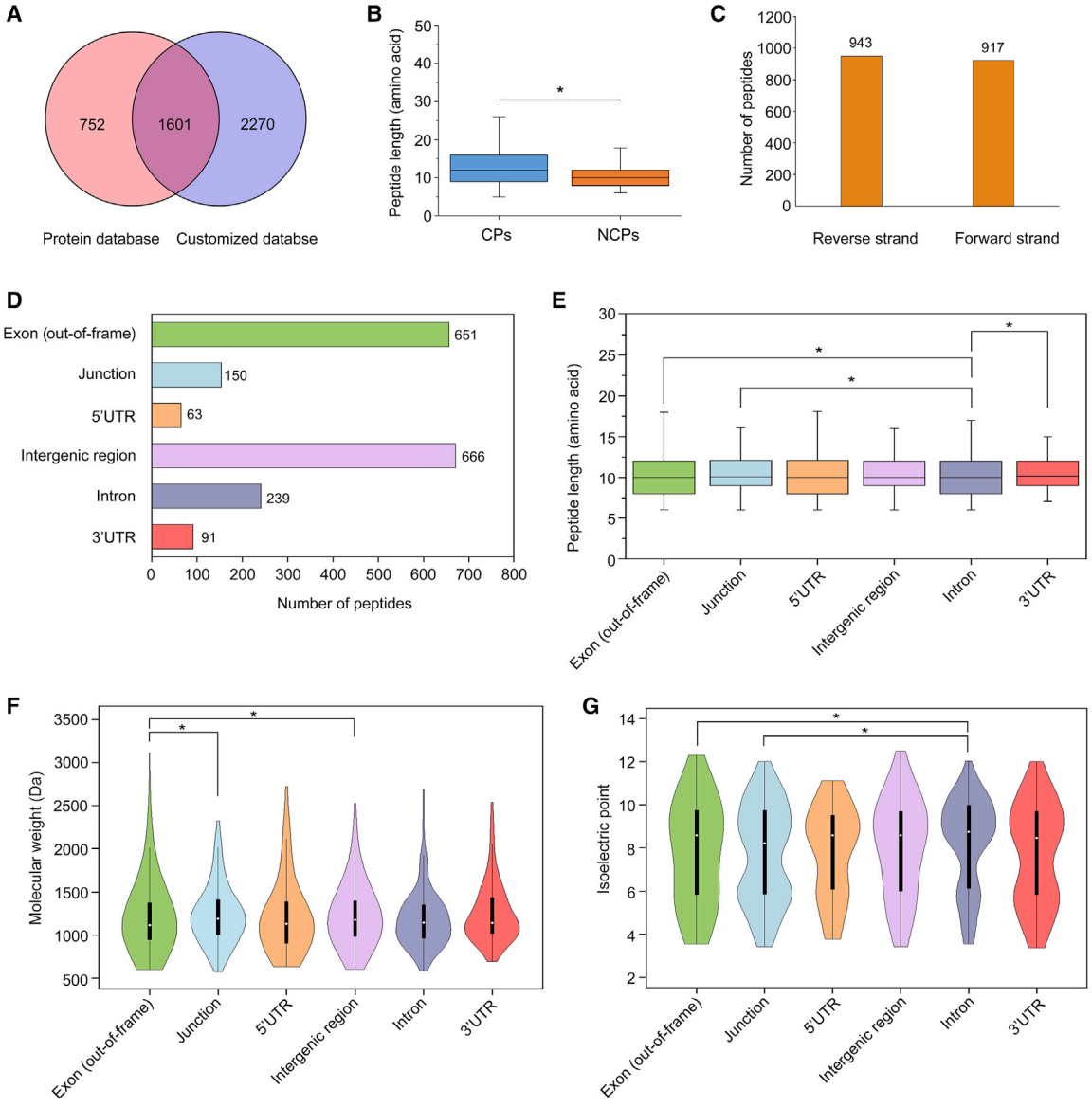

Figure 7. Identification of NCPs in *Arabidopsis*.

(A) Venn diagram showing the number of peptides identified using the Ensembl protein and customized peptidogenomic databases.

(B) Length of CPs and NCPs in *Arabidopsis*. Boxes represent the second and third quartiles, whiskers represent 1.5× IQR. Fisher's exact test was used for hypothesis testing, $*p < 0.05$.

(C) Number of NCPs derived from the forward and reverse strands.

(D) Number of NCPs derived from different gene elements.

(E–G) (E) Length of NCPs derived from different gene elements. Boxes represent the second and third quartiles, whiskers represent 1.5× IQR. Fisher's exact test was used for hypothesis testing, $*p < 0.05$. Violin plot combines box plot and kernel density trace to describe the distribution patterns of molecular weight **(F)** and isoelectric point **(G)**. Tomato: NCPs derived from 3′ UTRs ($n = 91$); beige: NCPs derived from introns ($n = 239$); lilac: NCPs derived from intergenic regions ($n = 666$); yellow: NCPs derived from 5′ UTRs ($n = 63$); green: NCPs derived from out-of-frame exons ($n = 651$); light blue: NCPs derived from junctions ($n = 150$). The black bars and thin lines within the violin plots represent the IQRs and the entire data ranges, respectively. White dots in the center indicate the average values. The width of the violin plot represents the density of the distribution. Fisher's exact test was used for hypothesis testing, $*p < 0.05$.

2016; Mohimani et al., 2018). In this study, we combined peptidomics with a customized peptidogenomic database derived from six-frame translation and the Ensembl protein databases to generate a peptidogenomic pipeline for both maize and *Arabidopsis*. To the best of our knowledge, this is the first report of a peptidogenomic pipeline to analyze NCPs in plants. With this strategy, 1993 and 1860 NCPs were identified in maize and *Arabidopsis*, respectively. The present study demonstrates that integrative peptidogenomic strategies can provide a more holistic overview of the peptidome to identify not only CPs but also NCPs. The results showed that a sizable proportion of peptides are NCPs, indicating that many previously alleged non-coding sequences, including 5′ UTRs, 3′ UTRs, intergenic regions, and introns are actually translatable.

Recently, the translation of lncRNAs has gained increasing attention (Kim et al., 2014; Saghatelian and Couso, 2015; Ransohoff et al., 2018). For example, a peptide encoded by a lncRNA was identified as myoregulin, which acts as an important regulator of calcium uptake in skeletal muscle (Anderson et al., 2015). A peptide encoded by a lncRNA epithelial cell program regulator (*EPR*) controls epithelial proliferation (Rossi et al., 2019). In addition, by overexpression and mutation analysis, peptides encoded by lncRNAs were shown to be involved in the regulation of growth and differentiation in moss (Fesenko et al., 2019). In the present study, 1652 NCPs derived from lncRNAs were identified in maize, and future characterization of these NCPs will be an important milestone in understanding the function of plant lncRNAs.

Upstream ORFs (uORFs) and their encoded peptides have been intensively investigated due to their potential to regulate the translation of downstream main ORFs (Hellens et al., 2016; Hsu and Benfey, 2018). The translation of these uORFs can also be regulated in response to developmental or environmental cues (Starck et al., 2016; Yin et al., 2019). In this study, we identified 18 and 63 NCPs derived from 5′ UTRs in maize and *Arabidopsis*, respectively. Among the 18 maize NCPs, 15 were also uncovered by previous ribosome profiling studies (Lei et al., 2015; Chotewutmontri and Barkan, 2016, 2018; Zoschke et al., 2017; Jiang et al., 2019), further supporting the results of the peptidogenomic analysis in the present study. In contrast to NCPs derived from the 5′ UTRs of genes, NCPs from 3′ UTRs have attracted little attention because they have been considered for a long time to be non-coding (Ingolia et al., 2011). Until only recently, peptides assigned to 3′ UTRs was identified in moss (Fesenko et al., 2019). In our study, we identified 25 and 91 NCPs derived from 3′ UTRs in maize and *Arabidopsis*, respectively, which further suggests that 3′ UTR-encoded peptides deserve much more attention, as these peptides may have vital biological roles.

Many maize QTLs have been found to be highly associated with non-coding regions (Clark et al., 2006; Silvio et al., 2007; Studer et al., 2011; Castelletti et al., 2014; Huang et al., 2018). Recently, we also examined several cases of intergenic QTLs in which the causative locus is a chromatin loop that regulates traits (Li et al., 2019; Peng et al., 2019). Apparently, it is important to study the regulatory elements in the non-coding sequences for a better understanding of the biological mechanisms underlying phenotypic traits. In this study, we found that NCPs were significantly enriched within QTL regions. For example, NCPs were enriched within regions associated with disease resistance, kernel length, and amino acid and oil contents, indicating the important functionality of NCPs in regulating these traits. Domestication is a tractable system for investigating evolutionary changes. Identification of genes involved in domestication will help us to understand the process of domestication and to accelerate the process of domesticating new crops (Wang et al., 2018). Several recent studies have used morphological, genetic, genomic, and archaeological techniques to determine the progressive fixation of different domestication genes in maize (da Fonseca et al., 2015; Liu et al., 2015; Vallebueno-Estrada et al., 2016). However, to date the molecular genetic architecture of maize domestication remains unclear. In this study, statistical analysis was applied for the enrichment anlysis of NCPs in domestication selection regions, suggesting the underlying functional sites for the evolution of maize.

In summary, in contrast to previous attempts using computational approaches or a ribosome profiling strategy to discover unannotated plant coding sequences, we directly and successfully identified plant NCPs on a large scale using an integrated peptidogenomic pipeline. The identification of NCPs reveals that many 5′ UTRs, 3′ UTRs, intergenic regions, introns, and junctions are translated and that some likely express functional peptides. These findings also provide insights into the discovery of novel functional genes or proteins through the characterization of NCPs in a wider array of plants.

METHODS

Sample Preparation

The maize inbred line B73 was grown in a greenhouse under a 15-h light (28°C)/9-h (25°C) dark photoperiod to the three-leaf stage. *Arabidopsis thaliana* (Columbia-0) was grown in a greenhouse under a 16-h light (22°C)/8-h (21°C) dark photoperiod to the four-leaf stage. Three replicates were performed for each species. The collected leaves were quickly frozen in liquid nitrogen and stored at −80°C until analysis.

Peptide Extraction

Maize and *Arabidopsis* leaves (2 g) collected as described above were quickly ground separately in liquid nitrogen. The powder was first heated in water at 95°C for 5 min. The samples were then precipitated in 10% (w/v) trichloroacetic acid/acetone solution at −20°C for 1 h, and the precipitate was washed with cold acetone until the supernatant was colorless. The supernatant was discarded, and the vacuum-dried precipitate was resuspended with a 1% TFA solution containing plant protease inhibitor cocktail (Sigma, USA) and incubated for 1 h at 4°C. It should be noted that TFA cannot be added before heat stabilization because it is a strongly irritating liquid which decomposes and emits toxic fluoride gas when heated. The fractions were ultrasonicated on ice (40 W, 6 s ultrasonic at a time, every 8 s, repeated five times) and then centrifuged at 10 000 *g* for 20 min at 4°C. The supernatants were filtered through a 10-kDa molecular-weight-cutoff centrifuge filter (Millipore, MA, USA) according to the manufacturer's instructions. Peptide mixtures were desalted using C18 cartridges (Empore, SPE Cartridges C18, 7 mm inner diameter, 3 ml volume; Sigma). The peptide fractions were vacuum-evaporated using a vacuum centrifugation concentrator and reconstituted in 40 μl of 0.1% TFA solution for LC–MS/MS analysis.

LC–MS/MS Analysis

For endogenous peptide profiling, MS experiments were performed on a Q Exactive mass spectrometer as described previously (Wang et al., 2019). Five micrograms of peptide mixture was loaded onto a C18 reversed-phase column (Thermo Scientific Easy Column, 10 cm length, 75 μm inner diameter, 3 μm resin) in buffer A (2% acetonitrile and 0.1% formic acid) and separated with a linear gradient of buffer B (80% acetonitrile and 0.1% formic acid) at a flow rate of 250 nl/min controlled by IntelliFlow technology over 120 min. MS data were acquired using a data-dependent top-10 method by dynamically choosing the most abundant precursor ions from the survey scan (300–1800 *m/z*) for higher-energy collisional dissociation (HCD) fragmentation. The determination of the target value was based on predictive automatic gain control. The dynamic exclusion duration was 25 s. Survey scans were acquired at a resolution of 70 000 at *m/z* 200, and the resolution for HCD spectra was set to 17 500 at *m/z* 200. The normalized collision energy was 30 eV and the underfill ratio, which specified the minimum percentage of the target value likely to be reached at maximum fill time, was defined as 0.1%. The instrument was run with peptide recognition mode enabled.

Peptide Database Construction

The complete genomes of maize and *Arabidopsis* were downloaded from Ensembl Plants (ftp://ftp.ensemblgenomes.org/pub/plants/release-41/fasta/zea_mays/dna/ and ftp://ftp.ensemblgenomes.org/pub/plants/release-45/fasta/arabidopsis_thaliana/dna/) in FASTA format. The putative peptide database was derived from the six-frame translation of genomic sequences using EMBOSS:6.6.0. Peptides were terminated whenever a stop codon was encountered. The next peptide was then started at the next nucleotide following the previous stop codon. Instances of ambiguous nucleotides (represented by "N" in the genome sequence) were replaced with random nucleotides; other ambiguous characters were also replaced with random nucleotides depending on their symbol. The genomic coordinates and orientation were recorded for each peptide. Resulting amino acid sequences for each chromosome were recorded in a FASTA-formatted sequence file.

Peptide Identification by Mascot

The Mascot search engine (Matrix Science) was used to search against the Ensembl protein databases for maize (ftp://ftp.ensemblgenomes.org/pub/plants/release-41/fasta/zea_mays) and *Arabidopsis* (ftp://ftp.ensemblgenomes.org/pub/plants/release-45/fasta/arabidopsis_*thaliana*/pep/), and the customized peptidogenomic databases to identify peptides. Mass tolerances on precursor and fragment ions were set to 5 ppm and 0.02 Da, respectively. The Mascot score ($\geq$25) and false discovery rate (FDR; <0.05) were applied to achieve final peptides. The same Mascot score was then applied to the list of peptides identified in the customized peptidogenomic database as described previously (Laumont et al., 2016). Raw data files were converted to peptide maps comprising *m/z* values, charge states, retention time, and intensity for all detected ions above a threshold of 8000 counts.

To obtain quantitative information for the peptides, we analyzed the MS data using MaxQuant software (version 1.3.0.5). The MS data were searched against the identified peptide sequences. An initial search was set at a precursor mass window of 6 ppm, followed by an enzymatic cleavage rule of none and a mass tolerance of 20 ppm for fragment ions. The cutoff of global FDR for peptide identification was set to 0.01. Peptide were quantified by intensities.

Identification of CPs and NCPs

Peptides identified from the Ensembl protein and customized peptidogenomic databases were combined and filtered with a stringent FDR cutoff (score $\geq$25; FDR <0.05). The resulting peptides were assigned to their respective source genes, and their MS/MS spectra manually verified. We then mapped the subset of peptide-encoding regions to discard peptides coming from multiple locations in the genome (1207 peptides for maize and 410 peptides for *Arabidopsis*). To determine the type of sequence (within the source gene) generating each peptide, we used the intersect function of the BEDTools suite to the bed file of the candidates as well as the Ensembl gff file. Peptides derived from previously annotated CDSs or conventional ORFs were classified as CPs. Peptides derived from intergenic regions, UTRs, reading frames different from those of annotated CDSs, introns, and various types of junctions (UTR–exon or exon–intron) were classified as NCPs.

Peptide Distribution at the Genome Level

Peptide density was calculated using a sliding window of 6 Mb with 3-Mb steps. Hotspot regions were defined as 6-Mb regions with a peptide count of more than 10. We downloaded the annotated maize genome from https://plants.ensembl.org/index.html and extracted the physical coordinates of TSSs. We searched for the closest TSS to each peptide to draw a frequency plot of distance between each peptide and its TSS. To accurately estimate the peptide number at the chromosome level, we divided the position of both CPs and NCPs by chromosome arm length.

Verification of NCPs Using Synthetic Peptides

The peptide sequences were chosen from different categories of NCPs identified by the peptidogenomic analysis and synthesized by GL Biochem (Shanghai). Dried peptides were diluted with 0.1% formic acid (Yang et al., 2018), and each synthetic peptide was separately subjected MS analysis with a Q Exactive mass spectrometer with the same parameters as those used for the peptidogenomic analysis.

RNA-Seq and Ribosome Profiling Analysis

RNA-seq datasets were retrieved from the NCBI Short Read Archive database (https://www.ncbi.nlm.nih.gov/sra). These datasets include circular RNAs (Jeck et al., 2013), lncRNAs (Lv et al., 2016; Zhu et al., 2017), mRNAs (Lei et al., 2015; Han et al., 2019), and small RNAs (He et al., 2019). In addition, the publicly available ribosome profiling datasets of maize (Lei et al., 2015; Chotewutmontri and Barkan, 2016, 2018; Zoschke et al., 2017; Jiang et al., 2019) were analyzed. The maize genome sequences and annotation files were obtained from the Ensembl Plants website (https://plants.ensembl.org/Zea_mays/Info/Index). After filtering out the low-quality reads, the remaining reads were mapped to the maize genome. The read count was then calculated for each NCP.

Association Analysis of NCPs with SNP/Regions Associated with a Collection of Traits and the Regions under Domestication Selection

A genome-wide association study was performed using a global germplasm collection of 527 elite maize inbred lines (Li et al., 2013) and a mixed-linear model based on previously reported traits, namely kernel-related yield traits (Liu et al., 2017), diseases (Chen et al., 2015), kernel oil (Li et al., 2013), and amino acid contents (Deng et al., 2017). SNPs called from the whole-genome shotgun (~20× for each line) sequences generated by a recent study (Yang et al., 2019) were used in association analysis. We generated 100 random genomic sets as background, each assigned with the same features as NCPs, including the total number, the number along different chromosomes, and the peptide length distribution (Supplemental Figure 5). The 100 random sets were used to estimate the mean and SD of the normal distribution for background overlapping ratios. The *p* values of enrichment of the observed ratio compared with the normal background distribution were calculated using the "pnorm" function (with lower.tail = FALSE) of R; the *p* values represent the upper-tail *p* value of the test statistic and indicate the probability of an observed value exceeding the expected distribution. Candidate regions associated with domestication were identified by comparing the 527 maize inbred lines with 183 teosinte samples, and the test of enrichment was performed using the same aforementioned test as used for QTL analysis (Supplemental Figure 6).

Data Analysis and Visualization

Unless stated otherwise, analysis and visualization were performed using R. All code is available on request to the corresponding author.

ACCESSION NUMBERS

All raw MS data from this study have been deposited in the ProteomeXchange Consortium via the PRIDE partner repository with dataset identifiers PXD017080 and PXD017081.

SUPPLEMENTAL INFORMATION

Supplemental Information is available at *Molecular Plant Online*.

FUNDING

This work is supported by the National Natural Science Foundation of China (nos. 31872872 and U1804113), National Key Research and Development Program of China (no. 2016YFD0101003), and Henan Association for Science and Technology.

AUTHOR CONTRIBUTIONS

L.W. and J.Y. designed the project. S. Wang, J.Z., L.T., X.C., and X.J. conducted experiments. S. Wang, J.Z., H.L., X.L., X.Z., Y.C., L.T., and S. Wu analyzed the data. S. Wang, L.T., H.L., S. Wu, J.Y., and L.W. wrote the manuscript. L.W. supervised the project. All authors read and approved the manuscript.

ACKNOWLEDGMENTS

We thank Dr. Steven P. Briggs for helpful discussions. We thank Dr. Anguo Sun and Dr. Yanwen Xiang for technical assistance. No conflict of interest declared.

Received: January 19, 2020
Revised: May 4, 2020
Accepted: May 18, 2020
Published: May 20, 2020

REFERENCES

Anderson, D.M., Anderson, K.M., Chang, C.L., Makarewich, C.A., Nelson, B.R., McAnally, J.R., Kasaragod, P., Shelton, J.M., Liou, J., Bassel-Duby, R., et al. (2015). A micropeptide encoded by a putative long noncoding RNA regulates muscle performance. Cell **160**:595–606.

Andrews, S.J., and Rothnagel, J.A. (2014). Emerging evidence for functional peptides encoded by short open reading frames. Nat. Rev. Genet. **15**:193–204.

Banting, F.G., and Best, C.H. (2007). The internal secretion of the pancreas (Reprinted from the Journal of Laboratory and Clinical Medicine, vol 7, pg 251-266, 1922). Indian J. Med. Res. **125**:A251–A266.

Bazin, J., Baerenfaller, K., Gosai, S.J., Gregory, B.D., Crespi, M., and Bailey-Serres, J. (2017). Global analysis of ribosome-associated noncoding RNAs unveils new modes of translational regulation. Proc. Natl. Acad. Sci. U S A **114**:E10018–E10027.

Blanvillain, R., Young, B., Cai, Y.M., Hecht, V., Varoquaux, F., Delorme, V., Lancelin, J.M., Delseny, M., and Gallois, P. (2011). The *Arabidopsis* peptide kiss of death is an inducer of programmed cell death. EMBO J. **30**:1173–1183.

Casson, S.A., Chilley, P.M., Topping, J.F., Evans, I.M., Souter, M.A., and Lindsey, K. (2002). The POLARIS gene of *Arabidopsis* encodes a predicted peptide required for correct root growth and leaf vascular patterning. Plant Cell **14**:1705–1721.

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

Castellana, N.E., Shen, Z., He, Y., Walley, J.W., Cassidy, C.J., Briggs, S.P., and Bafna, V. (2014). An automated proteogenomic method uses mass spectrometry to reveal novel genes in *Zea mays*. Mol. Cell. Proteomics **13**:157–167.

Castelletti, S., Tuberosa, R., Pindo, M., and Salvi, S. (2014). A MITE Transposon insertion is associated with differential methylation at the maize flowering time QTL *Vgt1*. G3 (Bethesda) **4**:805–812.

Chen, G., Wang, X., Hao, J., Yan, J., and Ding, J. (2015). Genome-wide association implicates candidate genes conferring resistance to maize rough dwarf disease in maize. PLoS One **10**:e0142001.

Chen, J., Brunner, A.D., Cogan, J.Z., Nunez, J.K., Fields, A.P., Adamson, B., Itzhak, D.N., Li, J.Y., Mann, M., Leonetti, M.D., et al. (2020a). Pervasive functional translation of noncanonical human open reading frames. Science **367**:1140–1146.

Chen, Q.J., Deng, B.H., Gao, J., Zhao, Z.Y., Chen, Z.L., Song, S.R., Wang, L., Zhao, L.P., Xu, W.P., Zhang, C.X., et al. (2020b). An miRNA-encoded small peptide, vvi-miPEP171d1, regulates adventitious root formation. Plant Physiol.. 10.1104/pp.20.00197

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

Chotewutmontri, P., and Barkan, A. (2016). Dynamics of chloroplast translation during chloroplast differentiation in maize. PLoS Genet. **12**:e1006106.

Chotewutmontri, P., and Barkan, A. (2018). Multilevel effects of light on ribosome dynamics in chloroplasts program genome-wide and psbA-specific changes in translation. PLoS Genet. **14**:e1007555.

Clark, R.M., Tina Nussbaum, W., Pablo, Q., and John, D. (2006). A distant upstream enhancer at the maize domestication gene tb1 has pleiotropic effects on plant and inflorescent architecture. Nat. Genet. **38**:594–597.

Corbiere, A., Walet-Balieu, M.L., Chan, P., Basille-Dugay, M., Hardouin, J., and Vaudry, D. (2018). A peptidomic approach to characterize peptides involved in cerebellar cortex development leads to the identification of the neurotrophic effects of nociceptin. Mol. Cell. Proteomics **17**:1737–1749.

Couso, J.P., and Patraquim, P. (2017). Classification and function of small open reading frames. Nat. Rev. Mol. Cell Biol. **18**:575–589.

da Fonseca, R.R., Smith, B.D., Wales, N., Cappellini, E., Skoglund, P., Fumagalli, M., Samaniego, J.A., Caroe, C., Avila-Arcos, M.C., Hufnagel, D.E., et al. (2015). The origin and evolution of maize in the Southwestern United States. Nat. Plants **1**:14003.

De Coninck, B., Carron, D., Tavormina, P., Willem, L., Craik, D.J., Vos, C., Thevissen, K., Mathys, J., and Cammue, B.P.A. (2013). Mining the genome of *Arabidopsis thaliana* as a basis for the identification of novel bioactive peptides involved in oxidative stress tolerance. J. Exp. Bot. **64**:5297–5307.

Deng, M., Li, D., Luo, J., Xiao, Y., Liu, H., Pan, Q., Zhang, X., Jin, M., Zhao, M., and Yan, J. (2017). The genetic architecture of amino acids dissection by association and linkage analysis in maize. Plant Biotechnol. J. **15**:1250–1263.

Farrokhi, N., Whitelegge, J.P., and Brusslan, J.A. (2008). Plant peptides and peptidomics. Plant Biotechnol. J. **6**:105–134.

Fesenko, I., Kirov, I., Kniazev, A., Khazigaleeva, R., Lazarev, V., Kharlampieva, D., Grafskaia, E., Zgoda, V., Butenko, I., Arapidi, G., et al. (2019). Distinct types of short open reading frames are translated in plant cells. Genome Res. **29**:1464–1477.

Han, L., Mu, Z., Luo, Z., Pan, Q., and Li, L. (2019). New lncRNA annotation reveals extensive functional divergence of the transcriptome in maize. J. Integr. Plant Biol. **61**:394–405.

Hanada, K., Zhang, X., Borevitz, J.O., Li, W.H., and Shiu, S.H. (2007). A large number of novel coding small open reading frames in the intergenic regions of the *Arabidopsis thaliana* genome are transcribed and/or under purifying selection. Genome Res. **17**:632–640.

Harvey, A.L., Edrada-Ebel, R., and Quinn, R.J. (2015). The re-emergence of natural products for drug discovery in the genomics era. Nat. Rev. Drug Discov. **14**:111–129.

He, J., Jiang, Z., Gao, L., You, C., Ma, X., Wang, X., Xu, X., Mo, B., Chen, X., and Liu, L. (2019). Genome-wide transcript and small RNA profiling reveals transcriptomic responses to heat stress. Plant Physiol. **181**:609–629.

Hellens, R.P., Brown, C.M., Chisnall, M.A.W., Waterhouse, P.M., and Macknight, R.C. (2016). The emerging world of small ORFs. Trends Plant Sci. **21**:317–328.

Hinnebusch, A.G., Ivanov, I.P., and Sonenberg, N. (2016). Translational control by 5'-untranslated regions of eukaryotic mRNAs. Science **352**:1413–1416.

Hsu, P.Y., and Benfey, P.N. (2018). Small but mighty: functional peptides encoded by small ORFs in plants. Proteomics **18**:e1700038.

Hsu, P.Y., Calviello, L., Wu, H.L., Li, F.W., Rothfels, C.J., Ohler, U., and Benfey, P.N. (2016). Super-resolution ribosome profiling reveals unannotated translation events in *Arabidopsis*. Proc. Natl. Acad. Sci. U S A **113**:E7126–E7135.

Huang, C., Sun, H.Y., Xu, D.Y., Chen, Q.Y., Liang, Y.M., Wang, X.F., Xu, G.H., Tian, J.G., Wang, C.L., Li, D., et al. (2018). *ZmCCT9* enhances maize adaptation to higher latitudes. Proc. Natl. Acad. Sci. U S A **115**:E334–E341.

Hurst, L.D. (2002). The *Ka/Ks* ratio: diagnosing the form of sequence evolution. Trends Genet. **18**:486.

Ingolia, N.T. (2016). Ribosome footprint profiling of translation throughout the genome. Cell **165**:22–33.

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R.S., and Weissman, J.S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science **324**:218–223.

Ingolia, N.T., Lareau, L.F., and Weissman, J.S. (2011). Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell **147**:789–802.

Jackson, R., Kroehling, L., Khitun, A., Bailis, W., Jarret, A., York, A.G., Khan, O.M., Brewer, J.R., Skadow, M.H., Duizer, C., et al. (2018). The translation of non-canonical open reading frames controls mucosal immunity. Nature **564**:434–438.

Jeck, W.R., Sorrentino, J.A., Wang, K., Slevin, M.K., Burd, C.E., Liu, J., Marzluff, W.F., and Sharpless, N.E. (2013). Circular RNAs are abundant, conserved, and associated with ALU repeats. RNA **19**:141–157.

Jiang, J., Chai, X., Manavski, N., Williams-Carrier, R., He, B., Brachmann, A., Ji, D., Ouyang, M., Liu, Y., Barkan, A., et al. (2019). An RNA chaperone-like protein plays critical roles in chloroplast mRNA stability and translation in *Arabidopsis* and maize. Plant Cell **31**:1308–1327.

Kastenmayer, J.P., Ni, L., Chu, A., Kitchen, L.E., Au, W.C., Yang, H., Carter, C.D., Wheeler, D., Davis, R.W., Boeke, J.D., et al. (2006). Functional genomics of genes with small open reading frames (sORFs) in *S. cerevisiae*. Genome Res. **16**:365–373.

Kersten, R.D., Yang, Y.L., Xu, Y., Cimermancic, P., Nam, S.J., Fenical, W., Fischbach, M.A., Moore, B.S., and Dorrestein, P.C. (2011). A mass spectrometry-guided genome mining approach for natural product peptidogenomics. Nat. Chem. Biol. **7**:794–802.

Khitun, A., Ness, T.J., and Slavoff, S.A. (2019). Small open reading frames and cellular stress responses. Mol. Omics **15**:108–116.

Kim, M.S., Pinto, S.M., Getnet, D., Nirujogi, R.S., Manda, S.S., Chaerkady, R., Madugundu, A.K., Kelkar, D.S., Isserlin, R., Jain, S., et al. (2014). A draft map of the human proteome. Nature **509**:575–581.

Kondo, T., Plaza, S., Zanet, J., Benrabah, E., Valenti, P., Hashimoto, Y., Kobayashi, S., Payre, F., and Kageyama, Y. (2010). Small peptides switch the transcriptional activity of shavenbaby during *Drosophila* embryogenesis. Science **329**:336–339.

Kurihara, Y., Makita, Y., Shimohira, H., Fujita, T., Iwasaki, S., and Matsui, M. (2020). Translational landscape of protein-coding and non-protein-coding RNAs upon light exposure in *Arabidopsis*. Plant Cell Physiol. **61**:536–545.

Laumont, C.M., Daouda, T., Laverdure, J.P., Bonneil, E., Caron-Lizotte, O., Hardy, M.P., Granados, D.P., Durette, C., Lemieux, S., Thibault, P., et al. (2016). Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames. Nat. Commun. **7**:10238.

Lauressergues, D., Couzigou, J.M., Clemente, H.S., Martinez, Y., Dunand, C., Becard, G., and Combier, J.P. (2015). Primary transcripts of microRNAs encode regulatory peptides. Nature **520**:90–93.

Lei, L., Shi, J., Chen, J., Zhang, M., Sun, S., Xie, S., Li, X., Zeng, B., Peng, L., Hauck, A., et al. (2015). Ribosome profiling reveals dynamic translational landscape in maize seedlings under drought stress. Plant J. **84**:1206–1218.

Li, H., Peng, Z.Y., Yang, X.H., Wang, W.D., Fu, J.J., Wang, J.H., Han, Y.J., Chai, Y.C., Guo, T.T., Yang, N., et al. (2013). Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. Nat. Genet. **45**:43–50.

Li, K., Wen, W.W., Alseekh, S., Yang, X.H., Guo, H., Li, W.Q., Wan, L.X., Pan, Q.C., Zhan, W., Liu, J., et al. (2019). Large-scale metabolite quantitative trait locus analysis provides new insights for high-quality maize improvement. Plant J. **99**:216–230.

Liu, H.J., Luo, X., Niu, L.Y., Xiao, Y.J., Chen, L., Liu, J., Wang, X.Q., Jin, M.L., Li, W.Q., Zhang, Q.H., et al. (2017). Distant eQTLs and non-coding sequences play critical roles in regulating gene expression and quantitative trait variation in maize. Mol. Plant **10**:414–426.

Liu, L., Du, Y.F., Shen, X.M., Li, M.F., Sun, W., Huang, J., Liu, Z.J., Tao, Y.S., Zheng, Y.L., Yan, J.B., et al. (2015). *KRN4* controls quantitative variation in maize kernel row number. Plos Genet. **11**:e1005670.

Liu, W.T., Kersten, R.D., Yang, Y.L., Moore, B.S., and Dorrestein, P.C. (2011). Imaging mass spectrometry and genome mining via short sequence tagging identified the anti-infective agent arylomycin in *Streptomyces roseosporus*. J. Am. Chem. Soc. **133**:18010–18013.

Lorenzo-Orts, L., Witthoeft, J., Deforges, J., Martinez, J., Loubery, S., Placzek, A., Poirier, Y., Hothorn, L.A., Jaillais, Y., and Hothorn, M. (2019). Concerted expression of a cell cycle regulator and a metabolic enzyme from a bicistronic transcript in plants. Nat. Plants **5**:184–193.

Lv, Y., Liang, Z., Ge, M., Qi, W., Zhang, T., Lin, F., Peng, Z., and Zhao, H. (2016). Genome-wide identification and functional prediction of nitrogen-responsive intergenic and intronic long non-coding RNAs in maize (*Zea mays* L.). BMC Genomics **17**:350.

Ma, J., Ward, C.C., Jungreis, I., Slavoff, S.A., Schwaid, A.G., Neveu, J., Budnik, B.A., Kellis, M., and Saghatelian, A. (2014). Discovery of human sORF-encoded polypeptides (SEPs) in cell lines and tissue. J. Proteome Res. **13**:1757–1765.

Magny, E.G., Pueyo, J.I., Pearl, F.M.G., Cespedes, M.A., Niven, J.E., Bishop, S.A., and Couso, J.P. (2013). Conserved regulation of cardiac calcium uptake by peptides encoded in small open reading frames. Science **341**:1116–1120.

Makarewich, C.A., and Olson, E.N. (2017). Mining for micropeptides. Trends Cell Biol. **27**:685–696.

Matsumoto, A., Pasut, A., Matsumoto, M., Yamashita, R., Fung, J., Monteleone, E., Saghatelian, A., Nakayama, K.I., Clohessy, J.G., and Pandolfi, P.P. (2017). mTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide. Nature **541**:228–232.

Mechin, V., Damerval, C., and Zivy, M. (2007). Total protein extraction with TCA-acetone. Methods Mol. Biol. **355**:1–8.

Mohimani, H., Gurevich, A., Shlemov, A., Mikheenko, A., Korobeynikov, A., Cao, L., Shcherbin, E., Nothias, L.F., Dorrestein, P.C., and Pevzner, P.A. (2018). Dereplication of microbial metabolites through database search of mass spectra. Nat. Commun. **9**:4035.

Mohimani, H., and Pevzner, P.A. (2016). Dereplication, sequencing and identification of peptidic natural products: from genome mining to peptidogenomics to spectral networks. Nat. Prod. Rep. **33**:73–86.

Na, C.H., Barbhuiya, M.A., Kim, M.S., Verbruggen, S., Eacker, S.M., Pletnikova, O., Troncoso, J.C., Halushka, M.K., Menschaert, G., Overall, C.M., et al. (2018). Discovery of noncanonical translation initiation sites through mass spectrometric analysis of protein N termini. Genome Res. **28**:25–36.

Narita, N.N., Moore, S., Horiguchi, G., Kubo, M., Demura, T., Fukuda, H., Goodrich, J., and Tsukaya, H. (2004). Overexpression of a novel small peptide ROTUNDIFOLIA4 decreases cell proliferation and alters leaf shape in *Arabidopsis thaliana*. Plant J. **38**:699–713.

Nelson, B.R., Makarewich, C.A., Anderson, D.M., Winders, B.R., Troupes, C.D., Wu, F.F., Reese, A.L., McAnally, J.R., Chen, X.W., Kavalali, E.T., et al. (2016). A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle. Science **351**:271–275.

Nesvizhskii, A.I. (2014). Proteogenomics: concepts, applications and computational strategies. Nat. Methods **11**:1114–1125.

Nguyen, D.D., Wu, C.H., Moree, W.J., Lamsa, A., Medema, M.H., Zhao, X.L., Gavilan, R.G., Aparicio, M., Atencio, L., Jackson, C., et al. (2013). MS/MS networking guided analysis of molecule and gene cluster families. Proc. Natl. Acad. Sci. U S A **110**:E2611–E2620.

Parkin, M.C., Wei, H., O'Callaghan, J.P., and Kennedy, R.T. (2005). Sample-dependent effects on the neuropeptidome detected in rat brain tissue preparations by capillary liquid chromatography with tandem mass spectrometry. Anal. Chem. **77**:6331–6338.

Peng, J., Zhang, H., Niu, H., and Wu, R.a. (2020). Peptidomic analyses: the progress in enrichment and identification of endogenous peptides. Trends Analyt Chem. **125**:115835.

Peng, Y., Xiong, D., Zhao, L., Ouyang, W., Wang, S., Sun, J., Zhang, Q., Guan, P., Xie, L., Li, W., et al. (2019). Chromatin interaction maps reveal genetic regulation for quantitative traits in maize. Nat. Commun. **10**:2632.

Plaza, S., Menschaert, G., and Payre, F. (2017). In search of lost small peptides. Annu. Rev. Cell Dev. Biol. **33**:391–416.

Ransohoff, J.D., Wei, Y.N., and Khavari, P.A. (2018). The functions and unique features of long intergenic non-coding RNA. Nat. Rev. Mol. Cell Biol. **19**:143–157.

Rohrig, H., Schmidt, J., Miklashevichs, E., Schell, J., and John, M. (2002). Soybean ENOD40 encodes two peptides that bind to sucrose synthase. Proc. Natl. Acad. Sci. U S A **99**:1915–1920.

Rossi, M., Bucci, G., Rizzotto, D., Bordo, D., Marzi, M.J., Puppo, M., Flinois, A., Spadaro, D., Citi, S., Emionite, L., et al. (2019). LncRNA EPR controls epithelial proliferation by coordinating *Cdkn1a* transcription and mRNA decay response to TGF-beta. Nat. Commun. **10**:1969.

Rubinsztein, D.C. (2006). The roles of intracellular protein-degradation pathways in neurodegeneration. Nature **443**:780–786.

Ruiz-Orera, J., Messeguer, X., Subirana, J.A., and Alba, M.M. (2014). Long non-coding RNAs as a source of new peptides. Elife **3**:e03523.

Saghatelian, A., and Couso, J.P. (2015). Discovery and characterization of smORF-encoded bioactive polypeptides. Nat. Chem. Biol. **11**:909–916.

Samandi, S., Roy, A.V., Delcourt, V., Lucier, J.F., Gagnon, J., Beaudoin, M.C., Vanderperre, B., Breton, M.A., Motard, J., Jacques, J.F., et al. (2017). Deep transcriptome annotation enables the discovery and functional characterization of cryptic small proteins. eLife **6**:e27860.

Secher, A., Kelstrup, C.D., Conde-Frieboes, K.W., Pyke, C., Raun, K., Wulff, B.S., and Olsen, J.V. (2016). Analytic framework for peptidomics applied to large-scale neuropeptide identification. Nat. Commun. **7**:11436.

Shiber, A., Doring, K., Friedrich, U., Klann, K., Merker, D., Zedan, M., Tippmann, F., Kramer, G., and Bukau, B. (2018). Cotranslational assembly of protein complexes in eukaryotes revealed by ribosome profiling. Nature **561**:268–272.

Silvio, S., Giorgio, S., Michele, M., Dwight, T., Xiaomu, N., Fengler, K.A., Robert, M., Ananiev, E.V., Sergei, S., and Edward, B. (2007). Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. Proc. Natl. Acad. Sci. U S A **104**:11376–11381.

Slavoff, S.A., Mitchell, A.J., Schwaid, A.G., Cabili, M.N., Ma, J., Levin, J.Z., Karger, A.D., Budnik, B.A., Rinn, J.L., and Saghatelian, A. (2013). Peptidomic discovery of short open reading frame-encoded peptides in human cells. Nat. Chem. Biol. **9**:59–64.

Starck, S.R., Tsai, J.C., Chen, K.L., Shodiya, M., Wang, L., Yahiro, K., Martins-Green, M., Shastri, N., and Walter, P. (2016). Translation from the 5' untranslated region shapes the integrated stress response. Science **351**:3867.

Studer, A., Zhao, Q., Ross-Ibarra, J., and Doebley, J. (2011). Identification of a functional transposon insertion in the maize domestication gene *tb1*. Nat. Genet. **43**:1160–1164.

Svensson, M., Skold, K., Svenningsson, P., and Andren, P.E. (2003). Peptidomics-based discovery of novel neuropeptides. J. Proteome Res. **2**:213–219.

Tavormina, P., De Coninck, B., Nikonorova, N., De Smet, I., and Cammue, B.P. (2015). The plant peptidome: an expanding repertoire of structural features and biological functions. Plant Cell **27**:2095–2118.

Vallebueno-Estrada, M., Rodriguez-Arevalo, I., Rougon-Cardoso, A., Gonzalez, J.M., Cook, A.G., Montiel, R., and Vielle-Calzada, J.P. (2016). The earliest maize from San Marcos Tehuacan is a partial domesticate with genomic evidence of inbreeding. Proc. Natl. Acad. Sci. U S A **113**:14151–14156.

van de Sande, K., Pawlowski, K., Czaja, I., Wieneke, U., Schell, J., Schmidt, J., Walden, R., Matvienko, M., Wellink, J., van Kammen, A., et al. (1996). Modification of phytohormone response by a peptide encoded by ENOD40 of legumes and a nonlegume. Science **273**:370–373.

van Heesch, S., Witte, F., Schneider-Lunitz, V., Schulz, J.F., Adami, E., Faber, A.B., Kirchner, M., Maatz, H., Blachut, S., Sandmann, C.L., et al. (2019). The translational landscape of the human heart. Cell **178**:242–260.

Wang, M., Li, W., Fang, C., Xu, F., Liu, Y., Wang, Z., Yang, R., Zhang, M., Liu, S., Lu, S., et al. (2018). Parallel selection on a dormancy gene during domestication of crops from multiple families. Nat. Genet. **50**:1435–1441.

Wang, S., Chen, Z., Tian, L., Ding, Y., Zhang, J., Zhou, J., Liu, P., Chen, Y., and Wu, L. (2019). Comparative proteomics combined with analyses of transgenic plants reveal ZmREM1.3 mediates maize resistance to southern corn rust. Plant Biotechnol. J. **17**:2153–2168.

Wu, H.L., Song, G., Walley, J.W., and Hsu, P.Y. (2019). The tomato translational landscape revealed by transcriptome assembly and ribosome profiling. Plant Physiol. **181**:367–380.

Yang, M.K., Lin, X.H., Liu, X., Zhang, J., and Ge, F. (2018). Genome annotation of a model diatom *Phaeodactylum tricornutum* using an integrated proteogenomic pipeline. Mol. Plant **11**:1292–1307.

Yang, N., Liu, J., Gao, Q., Gui, S.T., Chen, L., Yang, L.F., Huang, J., Deng, T.Q., Luo, J.Y., He, L.J., et al. (2019). Genome assembly of a tropical maize inbred line provides insights into structural variation and crop improvement. Nat. Genet. **51**:1052–1059.

Yin, X., Jing, Y., and Xu, H. (2019). Mining for missed sORF-encoded peptides. Expert Rev. Proteomics **16**:257–266.

Zhu, M., Zhang, M., Xing, L., Li, W., Jiang, H., Wang, L., and Xu, M. (2017). Transcriptomic analysis of long non-coding RNAs and coding genes uncovers a complex regulatory network that is involved in maize seed development. Genes (Basel) **8**:274.

Zoschke, R., Chotewutmontri, P., and Barkan, A. (2017). Translation and co-translational membrane engagement of plastid-encoded chlorophyll-binding proteins are not influenced by chlorophyll availability in maize. Front. Plant Sci. **8**:385.