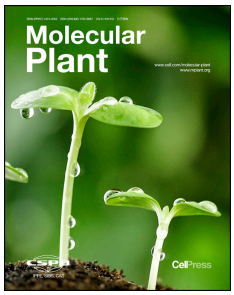
Journal Pre-proof




# Large-scale Discovery of Non-conventional Peptides in Maize and Arabidopsis Through an Integrated Peptidogenomic Pipeline

Shunxi Wang, Lei Tian, Haijun Liu, Xiang Li, Jinghua Zhang, Xueyan Chen, Xingmeng Jia, Xu Zheng, Shubiao Wu, Yanhui Chen, Jianbing Yan, Liuji Wu

PII: S1674-2052(20)30147-7
DOI: https://doi.org/10.1016/j.molp.2020.05.012
Reference: MOLP 939

To appear in: *MOLECULAR PLANT*
Accepted Date: 18 May 2020

Please cite this article as: **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. doi: https://doi.org/10.1016/j.molp.2020.05.012.

# Large-scale Discovery of Non-conventional Peptides in Maize and Arabidopsis Through an Integrated Peptidogenomic Pipeline

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

† These authors contributed equally to this work.

***Corresponding authors**

Liuji Wu (wlj200120@163.com) and Jianbing Yan (yjianbing@mail.hzau.edu.cn)

**Running title:** Large-scale discovery of non-conventional peptides

**SHORT SUMMARY**

This study developed an integrated peptidogenomic pipeline and firstly applied it for large-scale identification of non-conventional peptides (NCPs) in plant. The identified NCPs, which were derived from introns, 3'UTRs, 5'UTRs, junctions and intergenic regions, showed distinct characteristics compared to conventional peptides (CPs). Functional analysis unveiled potential function of NCPs in plant genetic regulation of complex traits and evolution.

## ABSTRACT

Non-conventional peptides (NCPs), which include small open reading frame-encoded peptides, play critical roles in fundamental biological processes. Here 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. Altogether, 1,993 and 1,860 NCPs were unambiguously identified in maize and Arabidopsis, respectively. The NCPs showed distinct characteristics compared to conventional peptides (CPs) and were derived from introns, 3'UTRs, 5'UTRs, junctions and intergenic regions. These results revealed that translation events in unannotated transcripts occurred more broadly than previously thought. In addition, maize NCPs were found to be enriched within regions associated with phenotypic variations and domestication selection, indicating their potential function in plant genetic regulations of complex traits and evolution. Summarily, this study provides an unbiased and global view of plant NCPs. The identification of large-scale NCPs in both monocot and dicot plants reveals that a much larger portion of the plant genome can be translated to biologically functional molecules, which has important implications in functional genomic studies. The present study also provides a useful resource for the characterization of more hidden NCPs in other plants.


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

## INTRODUCTION

Peptides, typically composed of 2 to 100 amino acid residues, represent the small biological molecules with important roles in biology (Tavormina et al., 2015). Small signaling peptides (SSPs) or peptide hormones, which are a class of short peptides ranging from 5 to 75 amino acid 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 in the 20$^{th}$ 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. Recently, a novel class of peptides, now defined as non-conventional peptides (NCPs) in this study, has caught significant attentions 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.

A primary report of the 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 the 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, 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 the research of NCPs at a genome-wide scale. Computational approaches based on sequence similarities have been developed to identify potential translational small open reading frames (sORFs) in noncoding 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 the short sequence and low conservation score. Another strategy is to use ribosome profiling by sequencing ribosome-protected fragments that enables mapping of a genome-wide set of transcripts that are being translated (Ingolia et al., 2009; Ingolia et al., 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). As a different strategy from ribosome profiling, mass spectrometry (MS)-based method is able to detect peptides that are translated from a 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 as peptidogenomics, which integrates peptidomics (based on high throughput MS/MS) and genomics, has emerged as a promising strategy for deep analysis of the endogenous NCPs (Kersten

et al., 2011; Harvey et al., 2015). As an efficient strategy, peptidogenomics has been successfully used in microorganisms 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 endogenous peptide enrichment, nonspecific protease digestion and lack of complete peptide reference databases, the identification of NCPs using peptidogenomics in plant 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 Ensembl protein database and the 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 noncoding sequences. NCPs showed a distinct distribution pattern from that of CPs. In addition, we found that the NCPs were enriched within the genomic regions associated with phenotypic variations and domestication, indicating their potential functions in regulating phenotypes and shaping the evolution of the plants. These results represent a large-scale identification of endogenous NCPs in plants through the integrated peptidogenomic pipeline and thus provide valuable information towards the understanding of the biological function of these hidden molecules.

# RESULTS

## An Integrated Peptidogenomic Pipeline for NCPs Identification in Plants

Directly detecting NCPs is the most definitive evidence of their existence. To facilitate plant NCPs 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 by water bath combined with plant protease inhibitors was applied to diminish nonspecific protease digestion. Trichloroacetic acid (TCA)-acetone

precipitation was also applied to establish an optimized sample preparation protocol. Then, plant endogenous peptides were enriched from larger protein fragments by centrifugation through 10 kDa cutoff filters before they were analyzed with liquid chromatography tandem mass spectrometry (LC-MS/MS).

To capture the endogenous peptides globally present in maize, the Mascot search engine was used to match the resulting mass spectrum data set against Ensembl protein database and customized peptidogenomic database, respectively. 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 the digital memory required to store our sequence data significantly. In addition, based on the locus-tracking approach, we used an automated process to map the peptide spectrum to their genomic loci, which enabled the pipeline for large-scale discovery of NCPs more effectively.

## Large-scale Identification of CPs and NCPs in Maize

All the reliably identified peptides from Ensembl protein and customized peptidogenomic databases were combined and used to identify both CPs and NCPs. In total, 748 and 3,932 non-redundant peptides were identified based on Ensembl protein database and customized peptidogenomic database, respectively (Figure 2A; Tables S1 and S2). Of these, 3,315 peptides were specifically identified by the customized peptidogenomic database (Figure 2A). Then, by mapping these peptides to genome loci and applying series filtering steps (see Methods), a total of 2,837 endogenous peptides were unambiguously assigned to a single genomic locus for each of the peptides. Among them, 1,993 (70.3%) NCPs (Table S3) and 844 (29.7%) CPs (Figure 2B; Table S4) were identified. The median length of CPs was 16 amino acids, while that of NCPs was 12 amino acids, with significant difference (Figure 2C), and approximately 90% of the peptides were less than 23 amino acids for CPs and 16

amino acids for NCPs (Figure S1). Furthermore, the average molecular weight of NCPs was 1325.22 Da, with 99.25% (1,978) 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 constituted a significant portion of plant peptidome, and showed different characteristics compared with CPs.

## Distribution Patterns of CPs and NCPs

Both CPs and NCPs were found unevenly distributed on the chromosomes of maize (Figure 3A). For CPs, most peptides were distributed near the telomeres, whereas NCPs were homogeneously located between 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 CPs hot regions containing 446 (52.84%) peptides were observed, whereas 81 NCPs hot regions containing 545 (27.35%) peptides were present (Figure 3A). Among these hot regions, one hot region located in chromosome 5 was common for 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 peptides 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 3D). We then compared the locations of these peptides to gene models, 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 NCPs translation along 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 CPs and NCPs source transcripts to predict their translation start sites. We observed a preponderance of

non-AUG translation start sites in both CPs and NCPs (Tables S3 and S4). 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 results of previous peptidomics studies that more than 90% endogenous peptides started with non-AUG codon (Chen et al., 2014; Secher et al., 2016; Corbiere et al., 2018). This result also support those of ribosome profiling and mass spectrometric studies, which demonstrate that most ORFs contain non-AUG start sites (Ingolia et al., 2011; Slavoff et al., 2013; Na et al., 2018).

## NCPs Derived from both Coding and Noncoding 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, 1,708 (85.70%) NCPs were 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 of these allegedly noncoding sequences.

Length analysis showed that the average lengths of NCPs derived from intergenic regions and out-of-frame exons were longer than that derived from junctions (Figure 4C). The average lengths of NCPs derived from 3'UTRs and 5'UTRs were the two shortest (Figure 4C). Molecular weight distribution analysis showed that more than 70% (1,407) of NCPs were less than 1500 Da. The average molecular weight of NCPs derived from intergenic regions was higher than that derived from introns, out-of-frame exons, 5'UTRs and 3'UTRs (Figure 4D and Figure S2A). There was no significant difference among the average isoelectric points (PI) values of NCPs derived from 3'UTRs, introns, intergenic regions, 5'UTRs, out-of-frame exons and junctions (Figure 4E and Figure S2B). Taken together, these results indicated that the identified NCPs represented a wide range of physicochemical properties and NCPs derived from different gene elements showed

different characteristics.

## Verification and Validation of NCPs

To verify these identified NCPs, we assigned these peptides to their respective source genomic locus. For example, NCP RMDAHALR was derived from the 5'UTR of gene *Zm00001d029555* (Figure 5A), and NCP ILTVNLKP was derived from the 3'UTR of gene *Zm00001d050172* (Figure 5B). Besides NCPs derived from UTRs, we also found a large number of NCPs 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 gene *Zm00001d008363* (Figure 5D). Next, 115 NCPs were synthesized experimentally. The mass spectrometer analysis was performed under the same conditions as were used for peptidogenomic analysis in this study. As shown in Figure 5A-D, 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 was shown in Supplemental Dataset 1.

In addition, we performed transcriptomic analyses using published RNA-seq data from maize. These RNA-seq data include circular RNAs, lncRNAs, mRNAs and small RNAs. Most NCPs (1,806, 90.62%) identified in the current study received support from these published databases (Table S3). Among these NCPs, 1,652 were from lncRNA and 859 from circular RNA (Table S3). The results indicated that these identified NCPs were likely produced from allegedly noncoding sequences.

Lastly, to validate the identified NCPs with independent methods, the available ribosome profiling datasets of maize 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 has a higher predictive value of final protein than mRNA-seq (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; Table S5). This validation rate of 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 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 out of numbers detected by peptidogenomic analysis were: 33.02% 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 the 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, which were detected only by 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 comprise a small fraction of the whole genome, and the vast majority of the genome has been considered noncoding regions. Genome-wide association study and quantitative trait locus (QTLs) analysis have identified a lot of functional elements in the noncoding regions in maize (Liu et al., 2017). The fact that 1,993 (70.3%) NCPs were derived from noncoding sequences prompts us to believe that they are of significant functional relevance. Therefore, we examined the enrichment of these NCPs with identified QTLs underlying various traits, and with those regions presumed under domestication selection.

Compared to randomly selected genomic sequence with same distance distribution and number (see Methods), it was revealed that significant single nucleotide polymorphisms (SNPs) associated with plant traits appeared to be significantly enriched within the regions of NCPs ($P < 0.02$, Upper-tail test; Figure 6A; Table S6). 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 to the random regions (*P < 7.4e-06*; Figure 6B; Table S7). Among the significant enriched SNPs, several were found exactly located within NCPs, which showed associations to various phenotypes including kernel length, disease (maize rough dwarf virus, MRDV), oil and amino acid contents (Figure 6C; Table S6). 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 13 kb upstream noncoding regions of 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 5 kb upstream noncoding regions of 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 to randomly selected genomic sequences with the same distance distribution and number, it was found that the NCPs were enriched within the 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 the domestication candidate regions (Table S8). While further validations are highly needed to explore which domesticated traits are exactly affected and what's the indeed mechanism, this result, for the first time as far as we know, unveils the likely inclusion of NCPs during domestication, providing another hidden layer of functional importance of NCPs.

## The Applicability of the Peptidogenomics Pipeline to Arabidopsis

To extend this pipeline to other plants, the dicot model plant Arabidopsis was used to test the wider applicability of peptidogenomic method. As a result, 2,353 and 3,871 non-redundant peptides were identified by the Ensembl protein database and customized peptidogenomic database (Tables S9 and S10), respectively. Of these, 2,270 peptides were specifically identified by the customized peptidogenomic

database (Figure 7A). In total, 1,860 (44.04%) NCPs (Table S11) and 2,363 (55.96%) CPs were obtained in Arabidopsis (Table S12). 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) (Figure S3). In addition, we found that the NCPs identified in Arabidopsis have shorter peptide length and lower molecular weight than that in maize (Table 13).

By analyzing the origins of NCPs, 943 (50.70%) NCPs were from the reverse strand (Figure 7C). By analyzing the locations of the NCPs within their respective gene sources, 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 were higher than that in maize (Table S13). Length analysis showed that the average length of NCPs derived from 3'UTRs was the longest and that from introns the shortest (Figure 7E). The average molecular weight of NCPs derived from out-of-frame exons was higher than that from 5'UTRs and intergenic regions (Figure 7F and Figure S4A). The average PI value of NCPs derived from out-of-frame exons and junctions were higher than that from introns (Figure 7G and Figure S4B).

Taken together, these results show that the developed peptidogenomic pipeline can also be used in dicot plants such as Arabidopsis. The translation of unannotated transcripts is widespread in both monocot and dicot plants, though they may have different translation patterns.

## DISCUSSION

Endogenous peptides are formed mainly by protein degradation, gene-encoding and gene-independent enzymatic formation *in vivo* (Peng et al., 2020). The emergence of peptidomics makes it possible for large-scale identification of endogenous peptides extracted from tissues (Slavoff et al., 2013; Secher et al., 2016). However, the study of

the peptidomics can be particularly challenging due to nonspecific 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 effective enough 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 attempt has not been experimented in plants so far.

Plant cells are more complex than animal cells due to the presence of additional components such as cell wall, large vacuoles and chloroplast, making the isolation of complete endogenous peptides in plants more challenging. In this study, in addition to the combination of heat stabilization by water bath and plant protease inhibitors to minimize nonspecific protease digestion in the 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 peptides extraction. We speculate that the protease associated nonspecific 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 states as they were *in vivo* for peptidomics study. In addition, it should be noted that the peptides from protein degradation within the cell is also another type of endogenous peptides in addition to those produced from gene-encoding (Peng et al., 2020). Protein degradation ubiquitously occurs in living organisms and the enzymatic degradation behavior of proteins is closely related to precursor protein status and enzyme activity in living organisms (Rubinsztein, 2006). Therefore, peptidomic data is also a good resource for the assessment of the potential protease/peptidase activity involving in the hydrolysis process, though 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 approach would not identify NCPs. The most effective strategy to do so is to integrate peptidomics with the six-frame translation of genome, which is referred as peptidgenomics (Kersten et al., 2011; Slavoff et al., 2013). 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 proven its value in identifying peptides at the genome-scale in microorganisms and humans (Kersten et al., 2011; Liu et al., 2011; Nguyen et al., 2013; Slavoff et al., 2013; Mohimani and Pevzner, 2016; Mohimani et al., 2018). In this study, we combined the peptidomics with a customized peptidogenomic database derived from six-frame translation and Ensembl protein databases to generate a peptidogenomic pipeline for both maize and Arabidopsis. To the best of our knowledge, this is the first report on a peptidogenomic pipeline to analyze NCPs in plants. With this strategy, 1,993 and 1,860 NCPs have been identified in maize and Arabidopsis, respectively. The present study demonstrates that integrative peptidogenomic strategies can provide a more holistic overview of the peptidome to not only identify CPs but also NCPs. The results showed that a sizeable proportion of peptides was found to be NCPs, indicating that many previously alleged noncoding 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 from 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, 1,652 NCPs

derived from lncRNA have been 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 (mORFs) (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 NCPs were also uncovered by previous ribosome profiling studies (Lei et al., 2015; Chotewutmontri and Barkan, 2016; Zoschke et al., 2017; Chotewutmontri and Barkan, 2018; Jiang et al., 2019), which further supports 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 to be noncoding for a long time (Ingolia et al., 2011). Until only recently, the presence of peptides assigned to 3'UTRs was identified, for example, in moss (Fesenko et al., 2019). In our study, we identified 25 and 91 NCPs that derived from 3'UTRs in maize and Arabidopsis, which further suggests that 3′UTRs encoded peptides deserve much more attention as these peptides may have vital biological roles in organisms.

Many maize QTLs have been found to be highly associated with noncoding 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 that regulate traits by chromatin loops (Li et al., 2019; Peng et al., 2019). Apparently, it is important to study the regulatory elements in the noncoding sequences for a better understanding of the biological mechanisms underlying phenotypic traits. In this study, we found that NCPs were significantly enriched within QTLs regions. For example, NCPs were enriched within regions associated with disease resistance, kernel length, amino acid and oil contents, indicating the important functionality of NCPs in regulating these traits. Domestication is a tractable system

for subsequent 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. The result of statistical analysis in this study showed significant enrichment of NCPs in the domestication selection regions, which may uncover the underlying functional sites for the evolution of the maize due to selection.

Taken together, in contrast to previous attempts of using computational approaches or ribosome profiling strategy to discover unannotated plant coding sequences, we directly and successfully identified large-scale plant NCPs based on the integrated peptidogenomic pipeline. The identification of NCPs reveals that many 5'UTRs, 3'UTRs, intergenic regions, introns, and junctions are translated and 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.

## MATERIALS AND 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 3 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 4 leaf stage. Three replicates were applied for each species. The collected leaves were quickly frozen in liquid nitrogen and stored at -80 °C until analyzed.

### Peptide Extraction

Maize and Arabidopsis leaves (2 g) as described above were quickly grounded in liquid nitrogen, respectively. The powder was firstly heated in the 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, the vacuum-dried precipitate transferred to 1% TFA solution containing plant protease inhibitor cocktail (Sigma, America), and incubated for an hour at 4 °C. It should be noted that TFA can't be added before heat stabilization, because TFA 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, and 5 times) and then centrifuged at 10,000 ×g for 20 min at 4 °C. The supernatants were filtered through 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 μg 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 top10 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 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. Then the next peptide was 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 upon 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 both the Ensembl protein 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 for the Ensembl protein database. The same Mascot score was then applied to the peptide list identified with 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 8,000 counts.

In order to obtain quantitative information for the peptides, the MS data were analyzed 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 intensities were used to indicate quantitative information of peptide.

**Identification of CPs and NCPs**

Peptides identified from Ensembl protein and customized peptidogenomic databases were combined and filtered with the stringent FDR cutoff (score ≥ 25; FDR < 0.05). The resulting peptides were assigned to their respective source genes and their MS/MS spectra were manually verified. Then, we mapped the subset of peptide-encoding regions to discard peptides coming from multiple locations in the genome (1,207 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 annotated CDSs or conventional open reading frames were classified as CPs. Peptides derived from intergenic regions, UTRs, different reading frames from 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. Hot regions were defined as the 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 for 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, position of both CPs and NCPs was divided 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) Ltd. Dried peptides were diluted with 0.1% formic acid (Yang et al., 2018), and each synthetic peptide was separately subjected to Q Exactive mass spectrometer for MS analysis

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 including circular RNAs (Jeck et al., 2013), lncRNAs (Lv et al., 2016; Zhu et al., 2017), mRNAs (Lei et al., 2015; Han et al., 2019), small RNAs (He et al., 2019). In addition, the publicly available ribosome profiling datasets of maize (Lei et al., 2015; Chotewutmontri and Barkan, 2016; Zoschke et al., 2017; Chotewutmontri and Barkan, 2018; Jiang et al., 2019) were also analyzed. The maize genome sequences and annotation files were obtained from the Ensembl Plants (https://plants.ensembl.org/Zea_mays/Info/Index). After filtering out the low-quality reads, the remaining reads were mapped to the maize genome. Then, the read count was 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) using the mixed-linear-model based on previously reported traits, including kernel-related yield traits (Liu et al., 2017), diseases (Chen et al., 2015), as well as 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 (Figure S5). The 100 random sets were used to estimate the mean and standard deviation of the normal distribution for background overlapping ratios. The *p*-values of enrichment of the observed ratio compared to the normal background distribution were calculated using the “pnorm” function (with lower.tail = FALSE) of R, representing the upper tail p-value of the test statistic and indicating the probability of observed value exceeding the expected distribution. Candidate regions associated with domestication were identified by comparing the 527 maize inbred lines to 183 teosinte samples, and the

test of enrichment was estimated using the aforementioned test as QTL analysis (Figure S6).

**Data Analysis and Visualization**

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

## ACCESSION NUMBERS

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

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

## COMPETING INTERESTS

The authors declare no competing interests.

601 ## REFERENCES

602 Anderson, D.M., Anderson, K.M., Chang, C.L., Makarewich, C.A., Nelson, B.R.,
603 McAnally, J.R., Kasaragod, P., Shelton, J.M., Liou, J., Bassel-Duby, R., et al.
604 (2015). A micropeptide encoded by a putative long noncoding RNA regulates
605 muscle performance. Cell 160:595-606.

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

608 Banting, F.G., and Best, C.H. (2007). The internal secretion of the pancreas
609 (Reprinted from the Journal of Laboratory and Clinical Medicine, vol 7, pg
610 251-266, 1922). Indian J. Med. Res. 125:A251-A266.

611 Bazin, J., Baerenfaller, K., Gosai, S.J., Gregory, B.D., Crespi, M., and Bailey-Serres, J.
612 (2017). Global analysis of ribosome-associated noncoding RNAs unveils new
613 modes of translational regulation. Proc. Natl. Acad. Sci. USA
614 114:E10018-E10027.

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

618 Casson, S.A., Chilley, P.M., Topping, J.F., Evans, I.M., Souter, M.A., and Lindsey, K.
619 (2002). The POLARIS gene of Arabidopsis encodes a predicted peptide
620 required for correct root growth and leaf vascular patterning. Plant Cell
621 14:1705-1721.

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

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

628 Castelletti, S., Tuberosa, R., Pindo, M., and Salvi, S. (2014). A MITE Transposon
629 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. https://doi: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.

660 da Fonseca, R.R., Smith, B.D., Wales, N., Cappellini, E., Skoglund, P., Fumagalli, M.,
661 Samaniego, J.A., Caroe, C., Avila-Arcos, M.C., Hufnagel, D.E., et al. (2015).
662 The origin and evolution of maize in the Southwestern United States. Nat.
663 Plants 1:14003.
664 De Coninck, B., Carron, D., Tavormina, P., Willem, L., Craik, D.J., Vos, C., Thevissen,
665 K., Mathys, J., and Cammue, B.P.A. (2013). Mining the genome of
666 *Arabidopsis thaliana* as a basis for the identification of novel bioactive
667 peptides involved in oxidative stress tolerance. J. Exp. Bot. 64:5297-5307.
668 Deng, M., Li, D., Luo, J., Xiao, Y., Liu, H., Pan, Q., Zhang, X., Jin, M., Zhao, M., and
669 Yan, J. (2017). The genetic architecture of amino acids dissection by
670 association and linkage analysis in maize. Plant Biotechnol. J. 15:1250-1263.
671 Farrokhi, N., Whitelegge, J.P., and Brusslan, J.A. (2008). Plant peptides and
672 peptidomics. Plant Biotechnol. J. 6:105-134.
673 Fesenko, I., Kirov, I., Kniazev, A., Khazigaleeva, R., Lazarev, V., Kharlampieva, D.,
674 Grafskaia, E., Zgoda, V., Butenko, I., Arapidi, G., et al. (2019). Distinct types
675 of short open reading frames are translated in plant cells. Genome Res.
676 29:1464-1477.
677 Han, L., Mu, Z., Luo, Z., Pan, Q., and Li, L. (2019). New lncRNA annotation reveals
678 extensive functional divergence of the transcriptome in maize. J. Integr. Plant
679 Biol. 61:394-405.
680 Hanada, K., Zhang, X., Borevitz, J.O., Li, W.H., and Shiu, S.H. (2007). A large
681 number of novel coding small open reading frames in the intergenic regions of
682 the *Arabidopsis thaliana* genome are transcribed and/or under purifying
683 selection. Genome Res. 17:632-640.
684 Harvey, A.L., Edrada-Ebel, R., and Quinn, R.J. (2015). The re-emergence of natural
685 products for drug discovery in the genomics era. Nat. Rev. Drug Discov.
686 14:111-129.
687 He, J., Jiang, Z., Gao, L., You, C., Ma, X., Wang, X., Xu, X., Mo, B., Chen, X., and
688 Liu, L. (2019). Genome-wide transcript and small RNA profiling reveals
689 transcriptomic responses to heat stress. Plant Physiol. 181:609-629.

690 Hellens, R.P., Brown, C.M., Chisnall, M.A.W., Waterhouse, P.M., and Macknight,
691 R.C. (2016). The emerging world of small ORFs. Trends Plant Sci.
692 21:317-328.
693 Hinnebusch, A.G., Ivanov, I.P., and Sonenberg, N. (2016). Translational control by
694 5'-untranslated regions of eukaryotic mRNAs. Science 352:1413-1416.
695 Hsu, P.Y., and Benfey, P.N. (2018). Small but mighty: Functional peptides encoded by
696 small ORFs in plants. Proteomics 18:e1700038.
697 Hsu, P.Y., Calviello, L., Wu, H.L., Li, F.W., Rothfels, C.J., Ohler, U., and Benfey, P.N.
698 (2016). Super-resolution ribosome profiling reveals unannotated translation
699 events in Arabidopsis. Proc. Natl. Acad. Sci. USA 113:E7126-E7135.
700 Huang, C., Sun, H.Y., Xu, D.Y., Chen, Q.Y., Liang, Y.M., Wang, X.F., Xu, G.H., Tian,
701 J.G., Wang, C.L., Li, D., et al. (2018). *ZmCCT9* enhances maize adaptation to
702 higher latitudes. Proc. Natl. Acad. Sci. USA 115:E334-E341.
703 Hurst, L.D. (2002). The *Ka/Ks* ratio: diagnosing the form of sequence evolution.
704 Trends Genet. 18:486.
705 Ingolia, N.T. (2016). Ribosome footprint profiling of translation throughout the
706 genome. Cell 165:22-33.
707 Ingolia, N.T., Ghaemmaghami, S., Newman, J.R.S., and Weissman, J.S. (2009).
708 Genome-wide analysis in vivo of translation with nucleotide resolution using
709 ribosome profiling. Science 324:218-223.
710 Ingolia, N.T., Lareau, L.F., and Weissman, J.S. (2011). Ribosome profiling of mouse
711 embryonic stem cells reveals the complexity and dynamics of mammalian
712 proteomes. Cell 147:789-802.
713 Jackson, R., Kroehling, L., Khitun, A., Bailis, W., Jarret, A., York, A.G., Khan, O.M.,
714 Brewer, J.R., Skadow, M.H., Duizer, C., et al. (2018). The translation of
715 non-canonical open reading frames controls mucosal immunity. Nature
716 564:434-438.
717 Jeck, W.R., Sorrentino, J.A., Wang, K., Slevin, M.K., Burd, C.E., Liu, J., Marzluff,
718 W.F., and Sharpless, N.E. (2013). Circular RNAs are abundant, conserved, and
719 associated with ALU repeats. RNA 19:141-157.

720 Jiang, J., Chai, X., Manavski, N., Williams-Carrier, R., He, B., Brachmann, A., Ji, D.,
721 Ouyang, M., Liu, Y., Barkan, A., et al. (2019). An RNA chaperone-like protein
722 plays critical roles in chloroplast mRNA stability and translation in
723 Arabidopsis and maize. Plant Cell 31:1308-1327.
724 Kastenmayer, J.P., Ni, L., Chu, A., Kitchen, L.E., Au, W.C., Yang, H., Carter, C.D.,
725 Wheeler, D., Davis, R.W., Boeke, J.D., et al. (2006). Functional genomics of
726 genes with small open reading frames (sORFs) in *S. cerevisiae*. Genome Res.
727 16:365-373.
728 Kersten, R.D., Yang, Y.L., Xu, Y., Cimermancic, P., Nam, S.J., Fenical, W., Fischbach,
729 M.A., Moore, B.S., and Dorrestein, P.C. (2011). A mass spectrometry-guided
730 genome mining approach for natural product peptidogenomics. Nat. Chem.
731 Biol. 7:794-802.
732 Khitun, A., Ness, T.J., and Slavoff, S.A. (2019). Small open reading frames and
733 cellular stress responses. Mol. Omics 15:108-116.
734 Kim, M.S., Pinto, S.M., Getnet, D., Nirujogi, R.S., Manda, S.S., Chaerkady, R.,
735 Madugundu, A.K., Kelkar, D.S., Isserlin, R., Jain, S., et al. (2014). A draft map
736 of the human proteome. Nature 509:575-581.
737 Kondo, T., Plaza, S., Zanet, J., Benrabah, E., Valenti, P., Hashimoto, Y., Kobayashi, S.,
738 Payre, F., and Kageyama, Y. (2010). Small peptides switch the transcriptional
739 activity of shavenbaby during drosophila embryogenesis. Science
740 329:336-339.
741 Kurihara, Y., Makita, Y., Shimohira, H., Fujita, T., Iwasaki, S., and Matsui, M. (2020).
742 Translational landscape of protein-coding and non-protein-coding RNAs upon
743 light exposure in Arabidopsis. Plant Cell Physiol. 61:536-545.
744 Laumont, C.M., Daouda, T., Laverdure, J.P., Bonneil, E., Caron-Lizotte, O., Hardy,
745 M.P., Granados, D.P., Durette, C., Lemieux, S., Thibault, P., et al. (2016).
746 Global proteogenomic analysis of human MHC class I-associated peptides
747 derived from non-canonical reading frames. Nat. Commun. 7:10238.
748 Lauressergues, D., Couzigou, J.M., Clemente, H.S., Martinez, Y., Dunand, C., Becard,
749 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.

780 Ma, J., Ward, C.C., Jungreis, I., Slavoff, S.A., Schwaid, A.G., Neveu, J., Budnik, B.A.,
781 Kellis, M., and Saghatelian, A. (2014). Discovery of human sORF-encoded
782 polypeptides (SEPs) in cell lines and tissue. J. Proteome Res. 13:1757-1765.
783 Magny, E.G., Pueyo, J.I., Pearl, F.M.G., Cespedes, M.A., Niven, J.E., Bishop, S.A.,
784 and Couso, J.P. (2013). Conserved regulation of cardiac calcium uptake by
785 peptides encoded in small open reading frames. Science 341:1116-1120.
786 Makarewich, C.A., and Olson, E.N. (2017). Mining for micropeptides. Trends Cell
787 Biol. 27:685-696.
788 Matsumoto, A., Pasut, A., Matsumoto, M., Yamashita, R., Fung, J., Monteleone, E.,
789 Saghatelian, A., Nakayama, K.I., Clohessy, J.G., and Pandolfi, P.P. (2017).
790 mTORC1 and muscle regeneration are regulated by the LINC00961-encoded
791 SPAR polypeptide. Nature 541:228-232.
792 Mechin, V., Damerval, C., and Zivy, M. (2007). Total protein extraction with
793 TCA-acetone. Methods Mol. Biol. 355:1-8.
794 Mohimani, H., Gurevich, A., Shlemov, A., Mikheenko, A., Korobeynikov, A., Cao, L.,
795 Shcherbin, E., Nothias, L.F., Dorrestein, P.C., and Pevzner, P.A. (2018).
796 Dereplication of microbial metabolites through database search of mass
797 spectra. Nat. Commun. 9 :4035.
798 Mohimani, H., and Pevzner, P.A. (2016). Dereplication, sequencing and identification
799 of peptidic natural products: from genome mining to peptidogenomics to
800 spectral networks. Nat. Prod. Rep. 33:73-86.
801 Na, C.H., Barbhuiya, M.A., Kim, M.S., Verbruggen, S., Eacker, S.M., Pletnikova, O.,
802 Troncoso, J.C., Halushka, M.K., Menschaert, G., Overall, C.M., et al. (2018).
803 Discovery of noncanonical translation initiation sites through mass
804 spectrometric analysis of protein N termini. Genome Res. 28:25-36.
805 Narita, N.N., Moore, S., Horiguchi, G., Kubo, M., Demura, T., Fukuda, H., Goodrich,
806 J., and Tsukaya, H. (2004). Overexpression of a novel small peptide
807 ROTUNDIFOLIA4 decreases cell proliferation and alters leaf shape in
808 *Arabidopsis thaliana*. Plant J. 38:699-713.
809 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. USA 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 in 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 Bio. 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. USA 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. USA 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. USA 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.

# FIGURE CAPTIONS

**Figure 1. Peptidogenomic Workflow for Plant NCPs Identification.**

(A) Peptidogenomic workflow for plant NCPs identification. Endogenous peptides from plant leaves were extracted using an optimized protocol. Heat stabilization by water bath at 95 °C combined with an acid extraction buffer containing 1% TFA and plant protease inhibitors was applied to minimize the peptide degradation during peptide extraction. Plant endogenous 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 the customized peptidogenomic database and Ensembl protein database using Mascot searching 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 then translated into six-frame using 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 a '*' symbol 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.

**Figure 2. Overview of the Peptidogenomic Results.**

(A) Venn diagram showing the number of peptides identified by Ensembl protein and customized peptidogenomic databases. The areas shown in the diagram are not proportional to the number of peptides in each group. (B) The number of CPs and NCPs identified through peptidogenomic analysis. (C) Length of CPs and NCPs. Boxes represent the second and third quartiles, whiskers represent 1.5 ×interquartile range. Fisher's exact test was used for hypothesis testing, * $p < 0.05$. (D) The molecular weight distribution of CPs (n=844). (E) Molecular weight distribution of NCPs (n=1,993). The rug plot above the x-axis represents the frequency at each

exposure level.

**Figure 3. CPs and NCPs Distribution in Maize.**

(A) The 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: CPs (green) and NCPs (red) distribution patterns by using a window size of 6 Mb and 3 Mb steps based on B73 reference genome. Right: hot region distribution of CPs (green) and NCPs (red). Hot regions were defined as more than 10 peptides in a window size of 6 Mb. (B) The normalized distribution of CPs and NCPs was 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 both 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) The histogram of the distances between two of adjacent CPs or NCPs. (E) The histograms showing the distance from each CP or NCP to the closest TSS.

**Figure 4. Characteristics of NCPs.**

(A) Number of NCPs derived from both forward and reverse strands. (B) Number of NCPs derived from different gene elements. (C) Length of NCPs derived from different gene elements. Boxes represent the second and third quartiles, whiskers represent 1.5 × the interquartile ranges. Fisher's exact test was used for hypothesis testing, * $p < 0.05$. Violin plots that 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=1,708); 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 interquartile ranges 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

975 hypothesis testing, * $p < 0.05$.

976 **Figure 5. Verification and Validation of NCPs.**

977 (A) NCP RMDAHALR mapped to the 5'UTR of a gene in chromosome 1 (left).
978 Verification of this NCP by comparing the spectra of the peptide identified by the
979 integrative peptidogenomic pipeline (middle) to that of synthetic peptide (right). (B)
980 NCP ILTVNLKP mapped to the 3'UTR of a gene in chromosome 4 (left). Verification
981 of this NCP by comparing the spectra of the peptide identified by the integrative
982 peptidogenomic pipeline (middle) to that of synthetic peptide (right). (C) NCP
983 QISVELPGVV mapped to the intergenic region between two genes in chromosome
984 10 (left). Verification of this NCP by comparing the spectra of the peptide identified
985 by the integrative peptidogenomic pipeline (middle) to that of synthetic peptide (right).
986 (D) NCP EGTPKAVGHRQ mapped to the intron of a gene in chromosome 8 (Left).
987 Verification of this NCP by comparing the spectra of the peptide identified by the
988 integrative peptidogenomic pipeline (middle) to that of synthetic peptide (right). (E)
989 Percentages of NCPs detected by peptidogenomics and ribosome profiling. (F)
990 Percentages of NCPs derived from different gene elements detected by
991 peptidogenomics and ribosome profiling.

992 **Figure 6. Quantitative Trait Loci (QTLs) Associated Significantly with**
993 **Phenotypic Traits Linked to NCPs.**

994 (A) The enrichment of NCPs within QTLs. (B) The enrichment of NCPs located
995 within 20 kb flanking regions of significant SNPs. (C) Diagram showing the
996 distribution of significant SNPs associated with plant traits within NCPs, one SNP
997 associated with kernel length, one with disease (maize rough dwarf virus, MRDV),
998 two with oil content, and four with amino acid content. (D) An isoleucine–threonine
999 transition caused by a SNP (chr1.s_244454699, A > G; $P < 9.42e\text{-}5$) associated with
1000 kernel length. Significant SNPs are indicated by red dotted lines. The black arrow
1001 indicates the NCP derived from the reverse strand. (E) The oil content associated
1002 significant SNP (chr3.s_136872577, C > T; $P < 2.09e\text{-}07$) that leads to a proline to
1003 leucine substitution in the NCP. The black arrow shows that the NCP was derived
1004 from the forward strand. (F) The enrichment of NCPs within regions under positive

selection during maize domestication. The x-axis shows the ratio of overlapping 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).

**Figure 7. Identification of NCPs in Arabidopsis.**

(A) Venn diagram showing the number of peptides identified by 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 × the interquartile ranges. 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) Length of NCPs derived from different gene elements. Boxes represent the second and third quartiles, whiskers represent 1.5 × the interquartile ranges. 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 interquartile ranges 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, * $p < 0.05$.

## SUPPLEMENTAL INFORMATION

**Figure S1. Length Distribution of CPs (A) and NCPs (B) in Maize.**

**Figure S2. Molecular Weight and Isoelectric Point Distribution of NCPs in Maize.**

(A) Molecular weight of NCPs derived from different gene elements in maize. (B) Isoelectric point distribution of NCPs derived from different gene elements in maize.

Tomato: NCPs derived from 3'UTR (n=25); beige: NCPs derived from introns (n=139); lilac: NCPs derived from intergenic regions (n=1,708); 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 rug plot above the x-axis represents the frequency at each exposure level.

**Figure S3. Molecular Weight Distribution of CPs and NCPs in Arabidopsis.**

(A) Molecular weight distribution of CPs (n=2,363). (B) Molecular weight distribution of NCPs (n=1,860). The rug plot above the x-axis represents the frequency at each exposure level.

**Figure S4. Molecular Weight and Isoelectric Point Distribution of NCPs in Arabidopsis.**

(A) Molecular weight of NCPs derived from different gene elements in Arabidopsis. (B) Isoelectric point distribution of NCPs derived from different gene elements in Arabidopsis. Tomato: NCPs derived from 3'UTR (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 rug plot above the x-axis represents the frequency at each exposure level.

**Figure S5. Enrichment Analysis of SNPs within NCPs in Maize.**

Analysis of the ratios of the numbers of NCPs containing SNPs associated with plant traits. As a control, we also collected similar ratios in the "random regions" by randomly shifting the genomic sequence 100 times in the same chromosome with the same number and same distance distribution. Statistics analysis was conducted for the within ratio between NCPs and random sequences.

**Figure S6. Enrichment Analysis of NCPs within the Domestication Selection Regions in Maize.**

To further explore the relationship between NCPs and domestication, we selected the NCPs with at least 1 bp overlapped with the domestication candidate region. As the control, we randomly shifted the genomic sequence 100 times in the same chromosome to generate random sequences with the same number and same distance

distribution.

**Table S1. Non-redundant Peptides Identified by Protein Database in Maize.**

**Table S2. Non-redundant Peptides Identified by the Customized Peptidogenomic Database in Maize.**

**Table S3. NCPs Identified in Maize.**

**Table S4. CPs Identified in Maize.**

**Table S5. NCPs Detected by Both Peptidogenomics and Ribosome Profiling in Maize.**

**Table S6. SNPs Significantly Associated with NCPs in Maize.**

**Table S7. SNPs Located within the 20 kb Flanking Regions of NCPs in Maize.**

**Table S8. Colocalization of Domestication and NCPs Regions in Maize.**

**Table S9. Non-redundant Peptides Identified by Protein Database in Arabidopsis.**

**Table S10. Non-redundant Peptides Identified by the Customized Peptidogenomic Database in Arabidopsis.**

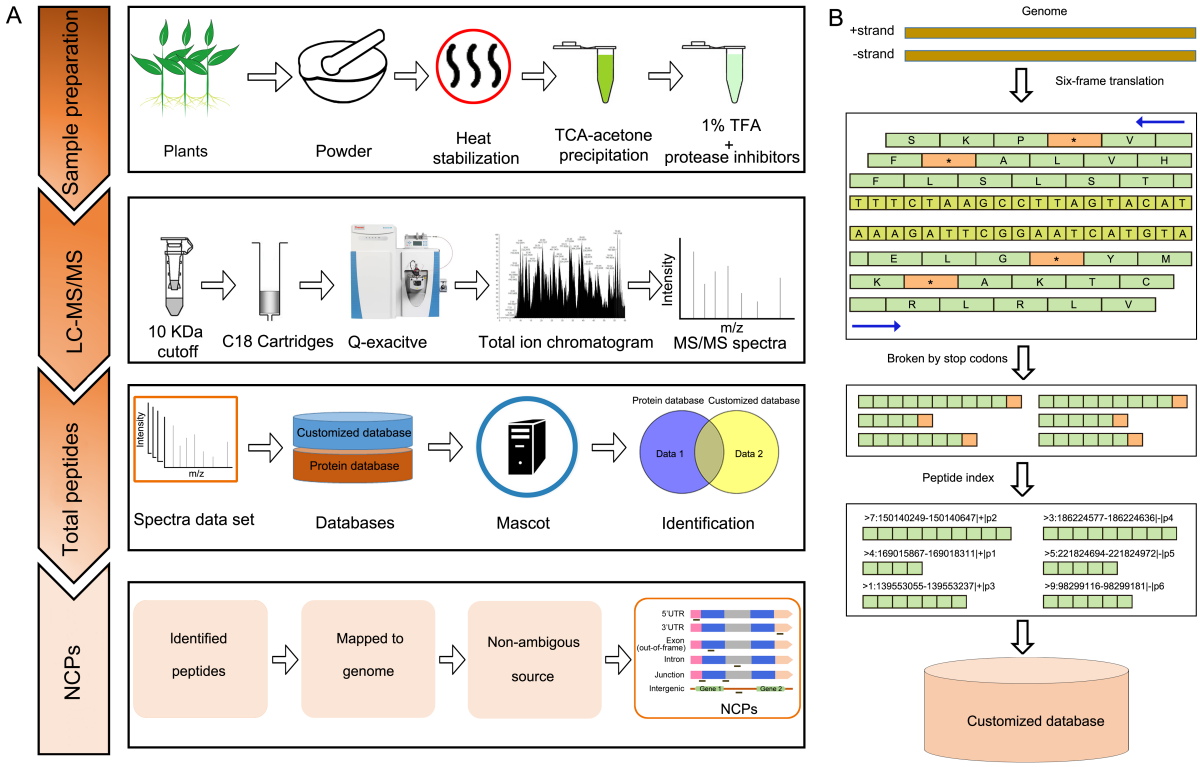
**Table S11. NCPs Identified in Arabidopsis.**

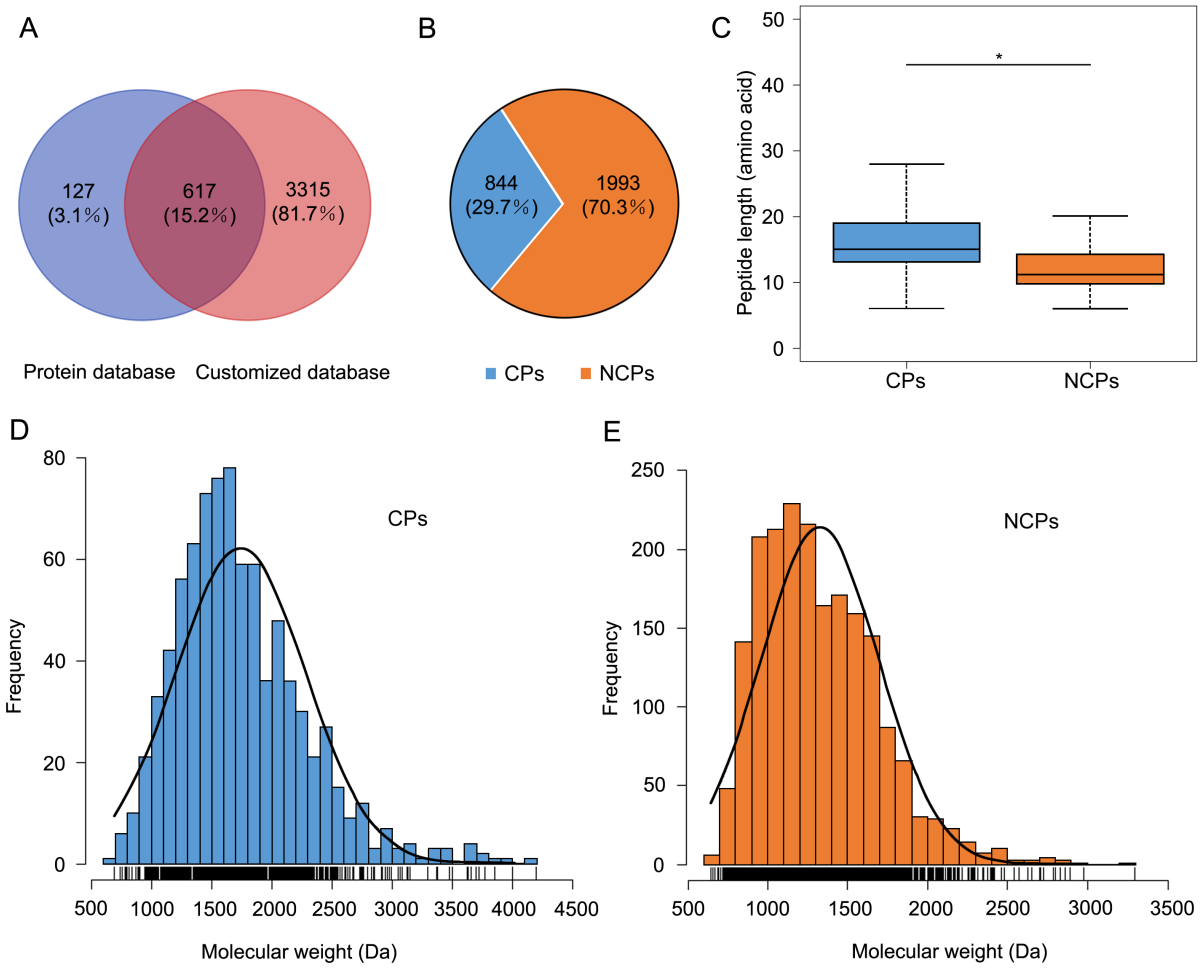
**Table S12. CPs Identified in Arabidopsis.**

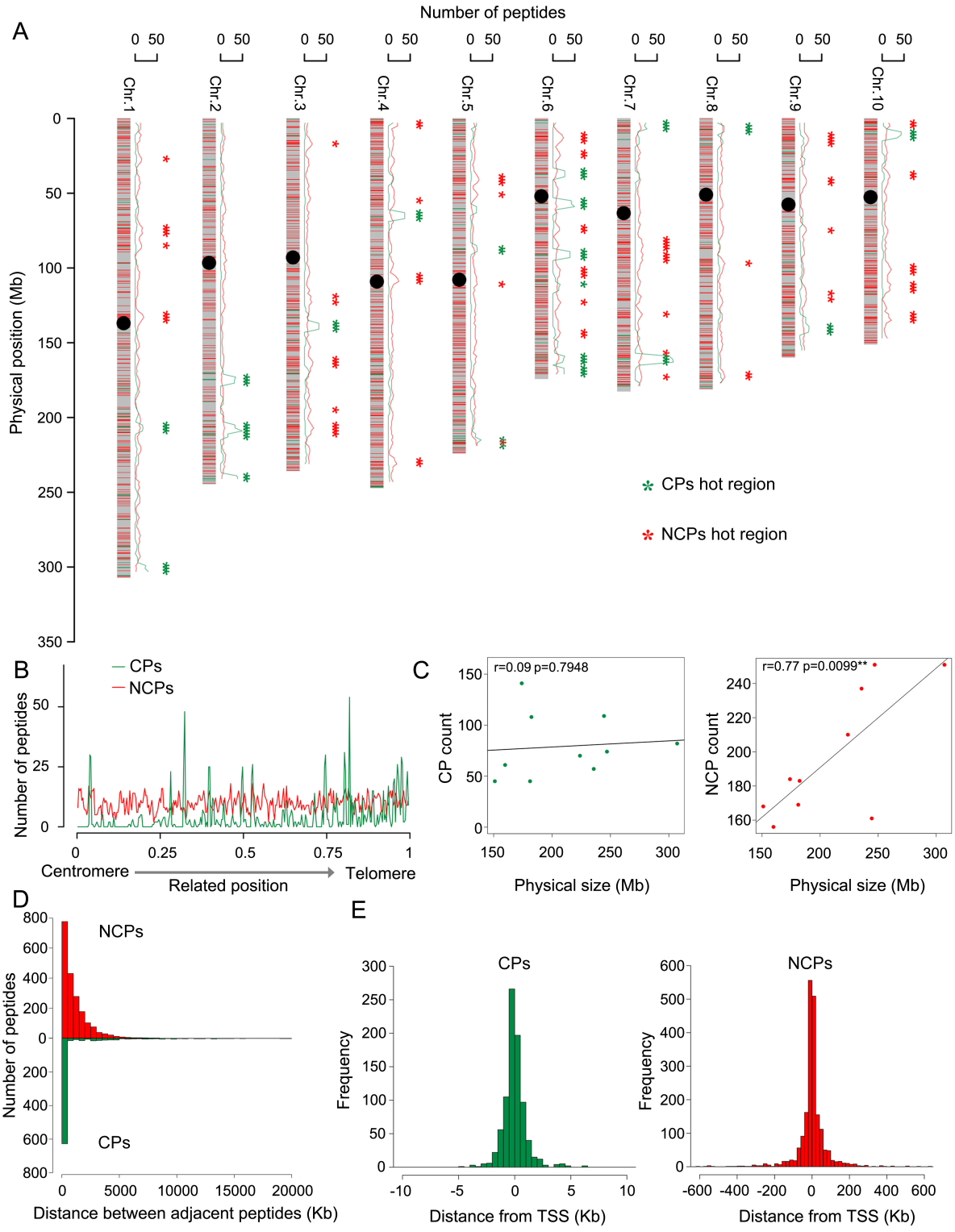
**Table S13. Comparisons between the NCPs Identified in Maize and Arabidopsis.**

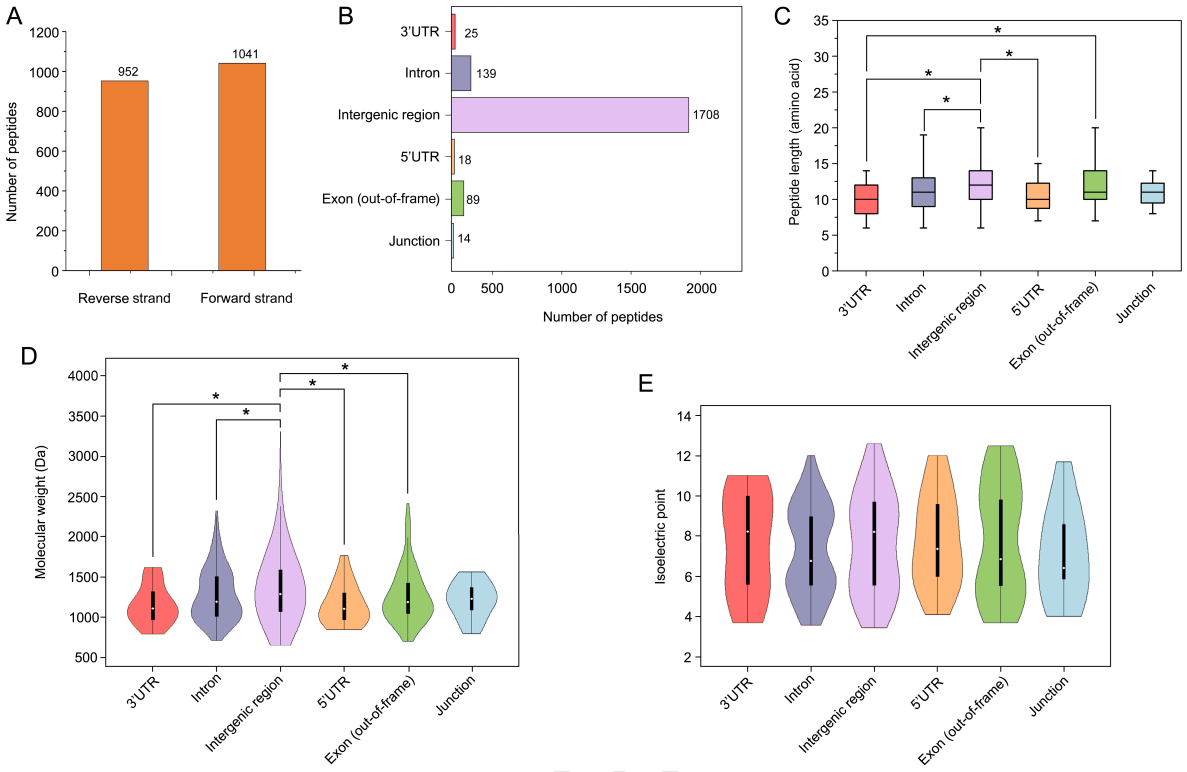
**Supplemental Dataset 1. Verification of the other 111 NCPs.**

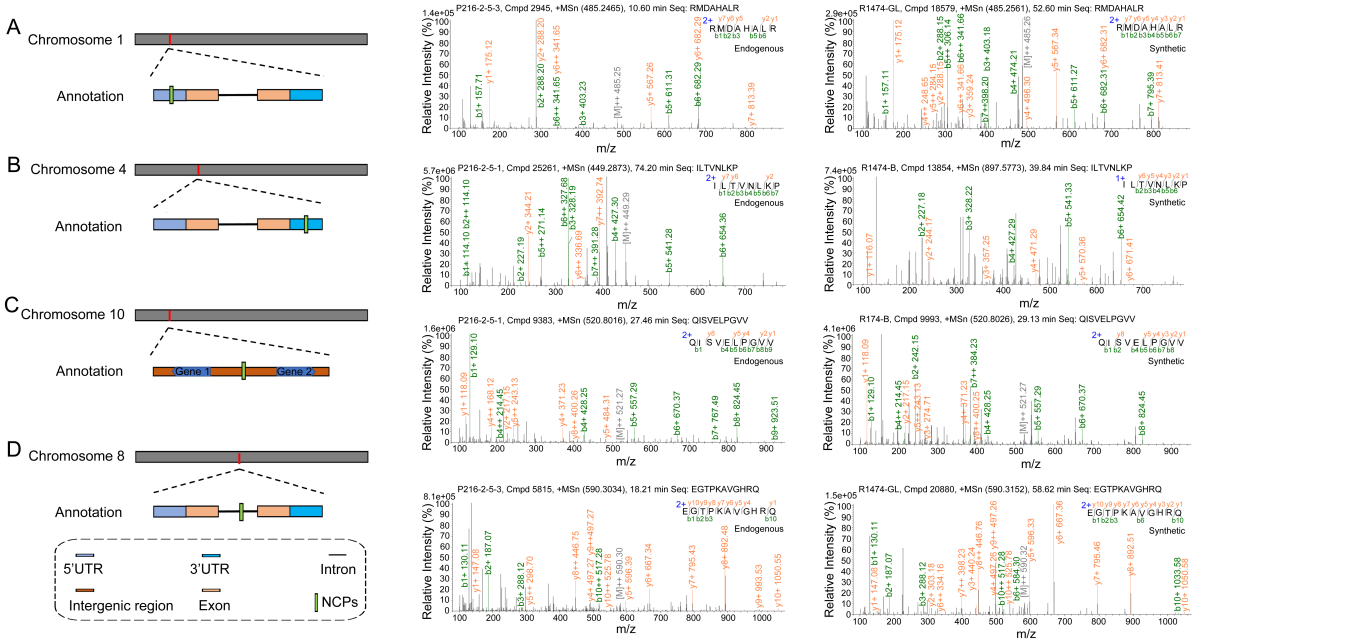
The other 111 NCPs were verified by comparing the spectra of the endogenous NCPs identified by the integrative peptidogenomic pipeline to that of synthetic peptides.

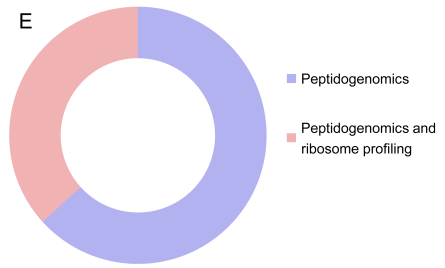

A
Sample preparation
LC-MS/MS
Total peptides
NCPs
Plants
Powder
Heat stabilization
TCA-acetone precipitation
1% TFA + protease inhibitors
10 KDa cutoff
C18 Cartridges
Q-exacitve
Total ion chromatogram
Intensity
m/z
MS/MS spectra
Intensity
m/z
Spectra data set
Customized database
Protein database
Databases
Mascot
Protein database
Customized database
Data 1
Data 2
Identification
Identified peptides
Mapped to genome
Non-ambigous source
5'UTR
3'UTR
Exon (out-of-frame)
Intron
Junction
Intergenic
Gene 1
Gene 2
NCPs
B
Genome
+strand
-strand
Six-frame translation
Broken by stop codons
Peptide index
>7:150140249-150140647|+|p2
>3:186224577-186224636|-|p4
>4:169015867-169018311|+|p1
>5:221824694-221824972|-|p5
>1:139553055-139553237|+|p3
>9:98299116-98299181|-|p6
Customized database

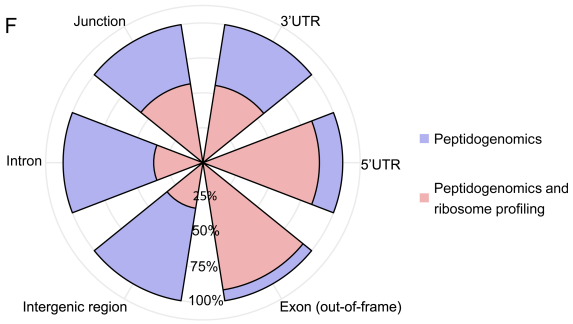

A
127
(3.1%)
617
(15.2%)
3315
(81.7%)
Protein database
Customized database
B
844
(29.7%)
1993
(70.3%)
CPs
NCPs
C
Peptide length (amino acid)
*
CPs
NCPs
D
CPs
Frequency
Molecular weight (Da)
E
NCPs
Frequency
Molecular weight (Da)

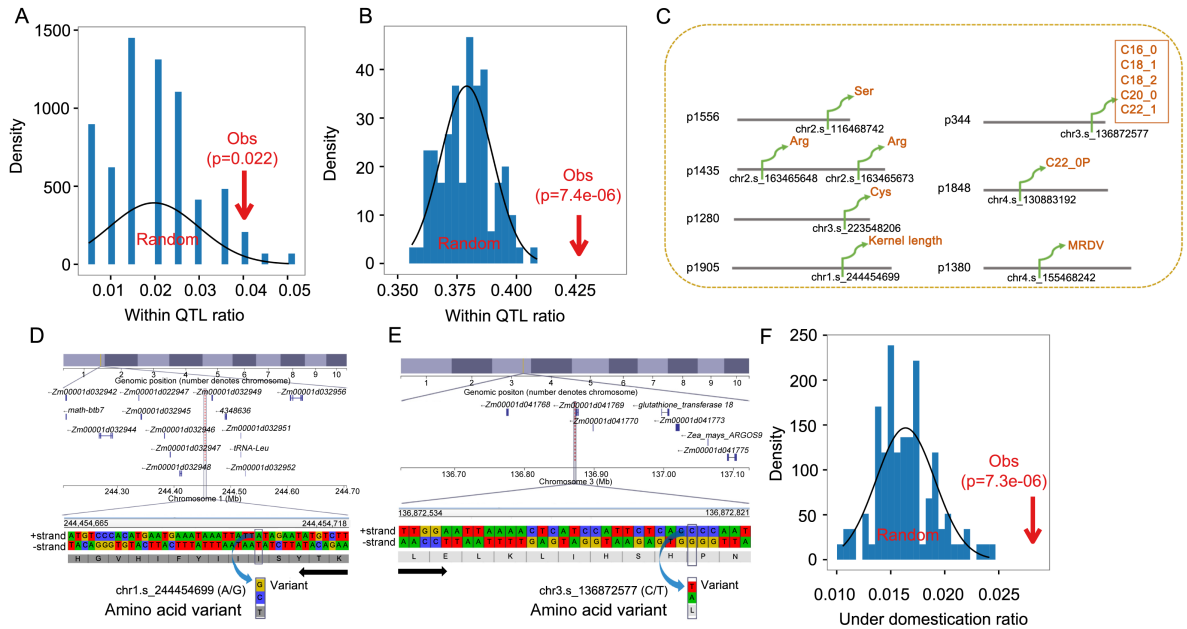

A
Number of peptides
Chr.1
Chr.2
Chr.3
Chr.4
Chr.5
Chr.6
Chr.7
Chr.8
Chr.9
Chr.10
Physical position (Mb)
CPs hot region
NCPs hot region
B
CPs
NCPs
Number of peptides
Centromere
Related position
Telomere
C
r=0.09 p=0.7948
CP count
Physical size (Mb)
r=0.77 p=0.0099**
NCP count
Physical size (Mb)
D
NCPs
CPs
Number of peptides
Distance between adjacent peptides (Kb)
E
CPs
Frequency
Distance from TSS (Kb)
NCPs
Frequency
Distance from TSS (Kb)

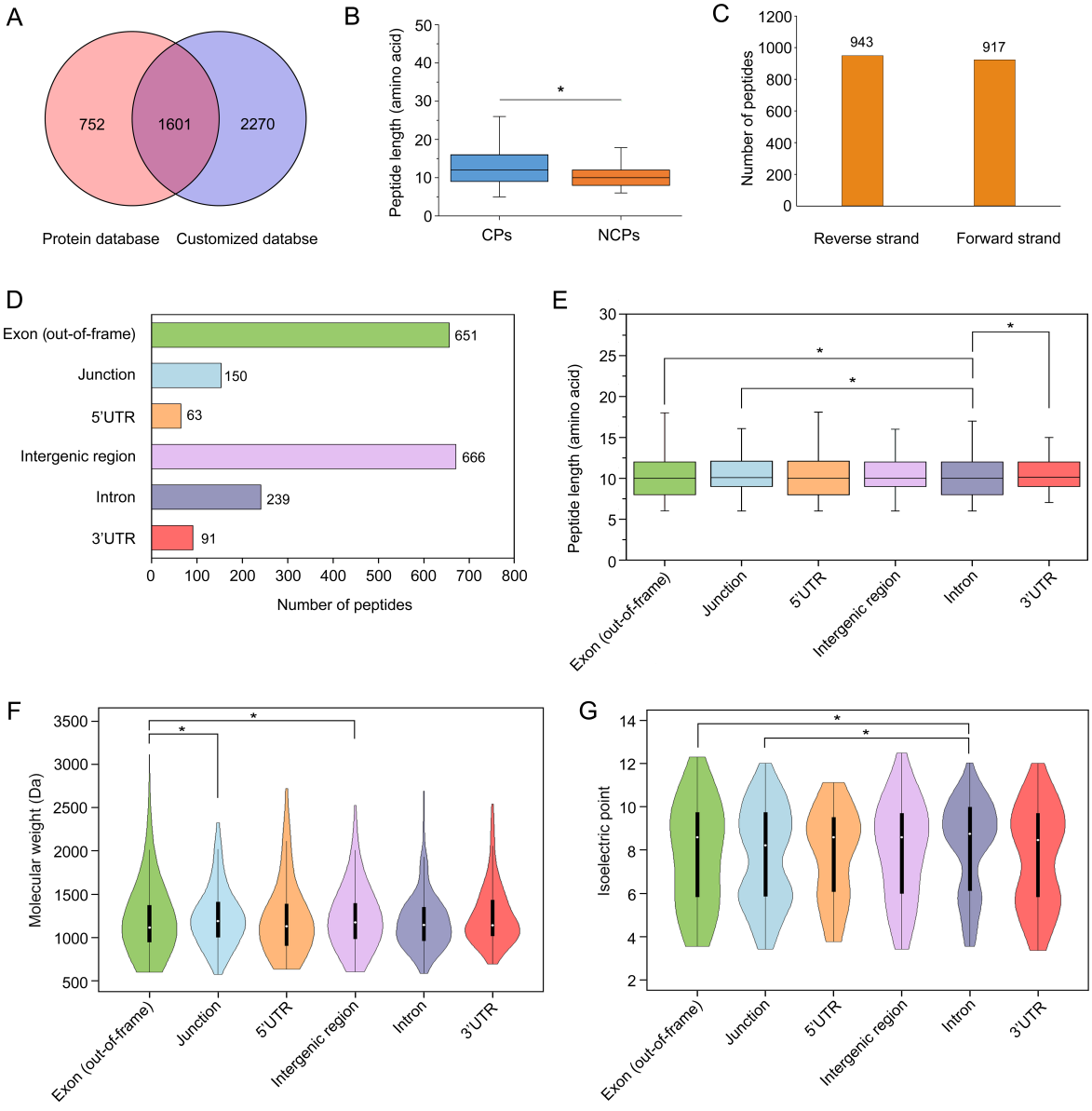



A
Chromosome 1
Annotation
B
Chromosome 4
Annotation
C
Chromosome 10
Annotation
Gene 1
Gene 2
D
Chromosome 8
Annotation
5'UTR
3'UTR
Intron
Intergenic region
Exon
NCPs
P216-2-5-3, Cmpd 2945, +MSn (485.2465), 10.60 min Seq: RMDAHALR
R1474-GL, Cmpd 18579, +MSn (485.2561), 52.60 min Seq: RMDAHALR
P216-2-5-1, Cmpd 25261, +MSn (449.2873), 74.20 min Seq: ILTVNLKP
R174-B, Cmpd 13854, +MSn (897.5773), 39.84 min Seq: ILTVNLKP
P216-2-5-1, Cmpd 9383, +MSn (520.8016), 27.46 min Seq: QISVELPGVV
R174-B, Cmpd 9993, +MSn (520.8026), 29.13 min Seq: QISVELPGVV
P216-2-5-3, Cmpd 5815, +MSn (590.3034), 18.21 min Seq: EGTPKAVGHRQ
R1474-GL, Cmpd 20880, +MSn (590.3152), 58.62 min Seq: EGTPKAVGHRQ
Endogenous
Synthetic
Relative Intensity (%)
m/z


E
Peptidogenomics
Peptidogenomics and ribosome profiling


F
Junction
3'UTR
Intron
5'UTR
Intergenic region
Exon (out-of-frame)
25%
50%
75%
100%
Peptidogenomics
Peptidogenomics and ribosome profiling




A
752
1601
2270
Protein database
Customized database
B
Peptide length (amino acid)
50
40
30
20
10
0
*
CPs
NCPs
C
Number of peptides
1200
1000
800
600
400
200
0
943
917
Reverse strand
Forward strand
D
Exon (out-of-frame)
651
Junction
150
5'UTR
63
Intergenic region
666
Intron
239
3'UTR
91
0
100
200
300
400
500
600
700
800
Number of peptides
E
Peptide length (amino acid)
30
25
20
15
10
5
0
*
*
*
Exon (out-of-frame)
Junction
5'UTR
Intergenic region
Intron
3'UTR
F
Molecular weight (Da)
3500
3000
2500
2000
1500
1000
500
*
*
Exon (out-of-frame)
Junction
5'UTR
Intergenic region
Intron
3'UTR
G
Isoelectric point
14
12
10
8
6
4
2
*
*
Exon (out-of-frame)
Junction
5'UTR
Intergenic region
Intron
3'UTR