Revisiting the Plant CLE Gene Family with a New Method for Predicting and Clustering Short Amino Acid Sequences

Zhe Zhang
Huazhong Agriculture University College of Horticulture and Forestry Sciences https://orcid.org/0000-0001-5265-6069

Lei Liu
Huazhong Agriculture University College of Horticulture and Forestry Sciences

Melis Kucukoglu
University of Helsinki

Dongdong Tian
Huazhong Agriculture University College of Horticulture and Forestry Sciences

Robert M. Larkin
Huazhong Agriculture University College of Horticulture and Forestry Sciences

Xueping Shi
Huazhong Agriculture University College of Horticulture and Forestry Sciences

Bo Zheng (bo.zheng@mail.hzau.edu.cn)
https://orcid.org/0000-0002-5337-5477

Research article

Keywords: Peptide hormone, CLE, Machine learning, Euclidean distance, Gene prediction, Gene clustering, Evolution

Posted Date: January 22nd, 2020

DOI: https://doi.org/10.21203/rs.2.21530/v1

Version of Record: A version of this preprint was published on October 12th, 2020. See the published version at https://doi.org/10.1186/s12864-020-07114-8.

Abstract

Background: The CLV3 / ESR-RELATED (CLE) gene family encodes small secreted peptides (SSPs) and plays vital roles in plant growth and development through cell-to-cell communication. The prediction and classification of CLE genes is challenging because of their low sequence similarity.

Results: We developed a machine learning-aided method for predicting CLE genes by using a CLE motif-specific residual score matrix and a novel clustering method based on the Euclidean distance of the 12 amino acid residues from CLE motifs in a site-weight dependent manner. In total, 2156 CLE candidates—including 627 novel candidates—were predicted from 69 plant species. The results from our CLE motif-based clustering are consistent with previous reports using the entire pre-propeptide. Characterization of CLE candidates provided systematic statistics on protein lengths, signal peptides, relative motif positions, amino acid compositions of different parts of the CLE precursor proteins, and decisive factors of CLE prediction. The approach taken here provides information on the evolution of the CLE gene family and provides evidence that the CLE and IDA/IDL genes share a common ancestor.

Conclusions: Our new approach is applicable to SSPs or other proteins with short conserved domains and hence, provides a useful tool for gene prediction, classification and evolutionary analysis.

Background

Small secreted peptides (SSPs) play vital roles in cell-to-cell communication during plant growth and development [1–4]. The most well understood plant SSPs are encoded by the CLAVATA3 (CLV3)/EMBRYO SURROUNDING REGION (ESR)-RELATED (CLE) gene family [5, 6]. CLE peptides have been widely identified in bryophytes, pteridophytes, gymnosperms and angiosperms [7]. A typical CLE protein contains an N-terminal signal peptide, a non-conserved variable region in the middle, a C-terminal conserved motif (CLE motif) and in some instances, a short C-terminal tail downstream of the CLE motif. CLE motifs are usually composed of 12 to 13 amino acid residues, independent of the flanking sequences. Artificially synthetic peptides representing the CLE motif domains can mimic the over-expression transgenic phenotypes [8–10]. The conserved CLE domains contain hydroxyproline and arabinosylated hydroxyproline residues [11–13]. Interestingly, the influence of these post-translational modifications varies in different species. For instance, post-translational modifications are critical for the activity of the CLV3 peptide in tomato but not in Arabidopsis [14, 15]. Typically, the mature forms of CLE peptides are recognized by plasma membrane-localized leucine-rich repeat receptor-like kinases (LRR-RLK) or receptor-like proteins (LRR-RLP) [16–18]. The extracellular domains of LRR-RLK/RLPs bind cognate CLE peptides as ligands and then transduce the extracellular signals by activating the intracellular domain of the LRR-RLKs or plasma membrane-associated kinases. Various methods have been used to investigate the interactions between CLE peptides and LRR-RLK/RLPs, such as genetic and physiological approaches, direct physical interaction and phosphor-proteomics. However, only a small number of possible ligand-receptor pairs have been identified [19, 20].

CLE peptides regulate the growth and development of various tissues in Arabidopsis, such as the apical and vascular meristems. Well-known CLE peptides include CLV3, CLE40 and TRACHEARY ELEMENT DIFFERENTIATION INHIBITORY FACTOR (TDIF), which regulate the CLAVATA1 (CLV1)–WUSCHEL (WUS) signaling pathway in the shoot apical meristem (SAM) [21], the ARABIDOPSIS CRINKLY4 (ACR4)–WUSCHEL-related Homeobox5 (WOX5) in the root apical meristem (RAM) [22], and the TDIF RECEPTOR/PHLOEM INTERCALATED WITH XYLEM (TDR/PXY)–WOX4 in the vascular cambial meristem (CAM) [10, 17, 23], respectively. In addition, CLE6 promotes gibberellin (GA)-mediated shoot growth [24]. The CLE8–WOX8 signaling pathway regulates endosperm and embryo development [25]. CLE9 and CLE10 regulate the formation of protoxylem by binding BARELY ANY MERISTEM (BAM) [26, 27]. CLE9/CLE10–HAESA-LIKE1 (HSL1)-SOMATIC EMBRYOGENESIS RECEPTOR KINASEs (SERKs) regulate stomatal lineage cell division [27]. CLE20 inhibits root growth and lateral root growth by inhibiting RAM and CAM activity, respectively [28, 29]. CLE25–BAM induces stomatal closure and promotes drought tolerance by controlling abscisic acid accumulation. Additionally, CLE25 promotes phloem initiation by regulating a CLE-RESISTANT RECEPTOR KINASE (CLERK)–CLV2 receptor complex [30, 31]. CLE26 regulates root architecture and protophloem formation [32–34]. CLE45–BAM3 suppresses protophloem differentiation and RAM growth. In contrast, CLE45–STERILITY-REGULATING KINASE MEMBER1/2 (SKM1/SKM2) maintains pollen performance, leading to successful seed production in plants grown in elevated temperatures [35, 36].

Each amino acid of the CLE motif plays different roles [37, 38]. For example, clv3 mutants and plants expressing a CLV3 motif with the Gly residue at the 6th position substituted with Leu, Ile, Val, Phe, Tyr, or Pro are phenotypically similar [38]. Similarly, structural and functional analyses of the TDIF–TDR/PXY (ligand–receptor pair) demonstrate that each amino acid residue of the TDIF motif is important. Indeed, amino acid substitutions at the 1st, 3rd, 4th, 6th, 8th, 9th and 12th positions of the TDIF motif result in reduced or complete loss of function [39]. Although amino acid substitutions at the 2nd, 5th, 10th and 11th sites of the TDIF motif have very little impact on its activity in terms of inhibition of TE differentiation, these sites are important for specifically binding the TDR/PXY receptor [40, 41].

Because of the short coding sequences and the generally low sequence similarity of CLE proteins, the identification and classification of CLE genes has always been a challenge, even in the model plant A. thaliana. Originally, using TBLASTN, 39 typical CLE polypeptides were identified—24 of them were from A. thaliana [42]. Subsequently, CLE40, CLE41 and a nematode CLE (HgCLE) gene were identified using the same approach [43]. The latter one emphasized the possibility of ligand mimicry. CLE41 was the first of a novel class of CLE gene, the TDIF and TDIF-like genes [39, 44]. A total of 32 CLE genes have been identified in A. thaliana defining 26 unique CLE peptides [45]. The Arabidopsis CLE genes have been used to identify CLE homologues in many other plant species, such as Oryza sativa [46], Lotus japonicas [47], Selaginella moellendorfii [48], Medicago truncatula [49, 50], Picea abies [51], Solanum lycopersicum [52], Glycine max [53], Raphanus sativus [54], and Populus trichocarpa [55]. Goad et al. [7] predicted CLE polypeptides from 57 plant species. The classification of CLE gene families has been based on their functions or sequence similarities. Based on the effects of CLE peptides on plant growth, the 22 Arabidopsis CLE polypeptides were classified into two groups [56]. According to their physiological

functions, four classes of CLE peptides were proposed [57]. On the other hand, 13 categories of CLE motifs were generated by clustering of the conserved sequences with the CLANS software [7, 58].

The objective of this study was to develop a novel approach for efficiently and accurately predicting and classifying CLE proteins. The general substitution matrix was replaced with a modified amino acid substitution matrix that is based on the weight of each position of the CLE motif. Machine learning (ML) was used to improve the accuracy of CLE gene predictions. This study helps to define the characteristics of different groups of CLE genes and therefore, to explore the origin and evolution of the CLE gene family.

Results

Developing a new residual score matrix for CLE motifs

To predict CLE genes in plants, we developed a new residual score matrix for CLE motifs by integrating the amino acid substitution matrix, amino acid usage frequency matrix and site weights of the CLE motif.

The amino acid composition of the CLE motif was analyzed in 69 species at different levels that included total proteins, small proteins (≤ 200 residues in length) and CLE precursor proteins (Additional file 1: Figure S1). The amino acid composition of small proteins was similar to total proteins. However, a higher frequency of particular residues was observed in CLE precursors (e.g., proline (P) and histidine (H)) (Additional file 1: Figure S1). The amino acid composition in different regions of CLE proteins was also analyzed (Additional file 1: Figure S1). The frequency of P and H in CLE motifs were both more than fourfold higher than in total proteins, which provides evidence that they are functionally important amino acids for peptide processing or peptide–receptor recognition (Fig. 1A). In contrast, some residues were very scarce in CLE motifs, such as the three aromatic amino acids (phenylalanine (F), tyrosine (Y) and tryptophan (W)) and the two sulfur-containing amino acids (cysteine (C) and methionine (M)) (Fig. 1A; Additional file 1: Figure S1). This strong bias in amino acid composition encouraged us to try and build a CLE-specific score matrix.

We tested three commonly used substitution matrices—BLOSUM62, BLOSUM80 and PAM250—and found that the performance of each was similar (Additional file 2: Figure S2). Nevertheless, the BLOSUM80 matrix provided a slightly better resolution of the motif scores and therefore, was used to develop the new score matrix. For the amino acid usage frequency matrix, 1628 reported CLE genes from 57 species [7] were chosen as references (Additional file 12: Table S1). The amino acid usage frequency at each site of the CLE motif was calculated as a percentage (Additional file 3: Figure S3) and was represented as a Weblogo sequence (Fig. 1D). The conservativities of each site were largely different, as previously reported [38]. Sites with higher conservativity were considered to hold higher weight in our new score matrix. Some sites contained two dominant residues, such as the 8th site (50.03% N and 47.06% D) and the 12th site (66.49% N and 31.38% H). Based on the modified method for evaluating site weight, the weight of the 12th site was set at 1.00, and the 1st, 6th -9th, and 11th sites had weights no less than 0.70 (Fig. 1B).

In the new CLE score matrix, each residue of a candidate CLE motif made a contribution to its total score. Residues at the conserved sites contributed more than those at less conserved sites. Dominant residues contributed more than scarce residues. For example, the proline at the 9th site alone had a score of 6.62, which made the most striking contribut in the score matrix (Fig. 1C). It is worth mentioning that the combination of 12 residues with the highest frequency at each site was “RLVPSGPNPLHN”, found in CLE9/10 in A. thaliana. The combination of residues with the highest score was “RRVPSGPNPLHN”. The total motif score was 38.00.

Machine learning aided the prediction of CLE genes in plants

In addition to the CLE motif score, we also included protein lengths, motif positions and signal peptide scores to predict CLE genes in 69 plant species. Three machine learning algorithms, C4.5, ANN and SVM, were employed to categorize all candidate genes into CLE genes or non-CLE genes using the reported CLE genes in the training data set. Our analysis of the training data set, based on 53 species, yielded 1709 CLE candidate genes, including the 1529 genes that were predicted using the Hidden Markov Model (HMM) (Goad et al 2017) and 180 novel genes. All three machine learning algorithms supported 1475 (96.5%) of the reported CLEs and 106 (58.9%) of the novel CLEs. In total, 94 (5.5%) of the candidate genes were supported by only one algorithm (Fig. 1E). Additionally, machine learning aided in the prediction of CLE genes. Indeed, machine learning identified 447 novel CLE candidates from the 16 species in the testing data set. Therefore, our method identified a total of 2156 CLE candidates in 69 species (Additional file 12: Table S1).

A new CLE classification method based on the Euclidean distance of CLE motifs in a site-weight dependent manner

To group the 2156 CLE motifs, the Euclidean distances (d) between each CLE candidate and the 32 Arabidopsis CLE motifs (AtCLEs) were calculated (Fig. 2 and Fig. 3). Motifs from the top 5% maximum d to AtCLEs were classified into Group “Others”. The rest of the CLE motifs were classified with their closest AtCLE. Consequently, all of the CLE motifs were grouped into six groups, Group1-5 and Others. As a comparison, phylogenetic trees constructed using the A. thaliana CLE motifs (Fig. 2A), CLE proteins without signal peptides (Fig. 2B) and log-normalized rank of all-vs-all BLAST e-values of full-length CLE proteins (Fig. 2C) were constructed using the NJ method, as previously described [7, 59]. The new AtCLE clustering using the HCL method was based on the Euclidean distance between each pair of AtCLE motifs (Fig. 2D). The clustering results were similar to the third phylogenetic tree, except for AtCLE8, AtCLE40 and AtCLE43 (Fig. 2 and Additional file 13: Table S2). In Group3, the AtCLE8 and AtCLE12 motifs, which are “RRVPTGPNPLHH” and “RRVPSGPNPLHH”, respectively, share high sequence similarity. However, clustering of AtCLE40 and AtCLE43 was not consistent among the four methods (Fig. 2D). To determine the reasons for these discrepancies, Weblogos of the appropriate subgroups (Group5A and Group5B) were created. Both subgroups were less conserved relative to other subgroups (Additional file 4: Figure S4 and Additional file 14: Table S3).

A cluster tree of all CLE candidates in 69 species was drawn that includes a heatmap indicating the Euclidean distance between the CLE motifs (Fig. 3). The heatmap demonstrated that the CLE candidates in Group5 and Group “Others” have a higher diversity in residual composition. Based on the cluster tree, the 26 AtCLE subgroups were then combined into 11 subgroups. Weblogos of the final 12 subgroups illustrated the importance of “heavy-weight” sites in the classification of CLE motifs (e.g., the 1st and 8th sites (Additional file 4: Figure S4)). Analysis of tandem CLE genes revealed that Group1 had the highest rate of tandem genes. Besides, candidates from monocots seemed to form clusters with other monocots, and candidates from dicots seemed to form clusters with other dicots. These data indicate a strong specificity among the monocot CLE motifs and the dicot CLE motifs (Fig. 3). Statistical analysis of the different types of CLE motifs showed that monocots and dicots share very few CLE motifs (18 out of the 474 CLE motifs in dicots). Furthermore, there was no common TDIF/TDIF-like motif shared between monocot and dicot species, probably due to the evolution of distinct vascular patterns in monocots and dicots (Additional file 15: Table S4).

Evolution of CLE genes in plants

To understand the evolution of CLE genes in plants, the number of CLE genes in each species was counted (Fig. 4 and Additional file 5: Figure S5). Although three CLE genes had been detected in algae, including one in Dunaliella salina and two in Coccomyxa subellipsoidea, the algal CLE genes were atypical because of their low motif scores, low signal peptide scores and poor motif positions (Additional file 16: Table S5). In contrast to algae, there were nine typical CLE genes in Physcomitrella patens (Fig. 4; Additional file 5: Figure S5 and Additional file 16: Table S5), 11 genes in Sphagnum fallax and eight genes in Marchantia polymorpha (Fig. 4, Additional file 5: Figure S5). These data provide evidence that CLE genes originated in a bryophyte.

The numbers of annotated transcripts in the 62 species of land plants were largely different, ranging from 19287 (M. polymorpha) to 99386 (Triticum aestivum) (Additional file 5: Figure S5). The proportion of CLE genes in different species was not fixed, ranging from 0.015% (Vitis vinifera) to 0.204% (Phaseolus vulgaris). The mean proportion of CLE candidates in dicots was slightly higher than in monocots, which were 0.105% and 0.091%, respectively. Their proportions in the three Bryophytes and the pteridophyte (Selaginella moellendorffii) were 0.027%, 0.041%, 0.041% and 0.036%, respectively, in general lower than in the monocots and dicots.

To further investigate the evolution of CLE genes in different species, the number of CLE genes in each subgroup was counted in each species (Fig. 4). CLE candidates appeared in fewer subgroups in lower plants. For example, the nine CLE candidates in P. patens were all presented in Group3B. Mapoly1011s0001.1 from M. polymorpha was the first candidate identified in Group4. Its motif “HKNPAGPNPIGN” shared high similarity with the CLE motif from Arabidopsis CLE46, a homolog of TDIF. Although none of the Group1 candidates were identified in the bryophytes, two CLE candidates from Group1 were identified in S. moellendorffii.

In addition, to the finding that CLE motifs are most frequently found in monocots and dicots, the number of each motif was counted (Additional file 6: Figure S6 and Additional file 17: Table S6). Our results indicated that the most frequent CLE motif in monocots was “RRVRRGSDPIH”—the same as CLE45 in A. thaliana; and the most frequent CLE motif in dicots was “HEVPSGPNPISN”—the same as CLE41/44 (TDIF) in A. thaliana. Particular CLE motifs have strong bias in monocots and dicots. For example, although the TDIF motif appeared 83 times in dicots, none were found in monocots. In contrast, only a TDIF-like motif “HEVPSGPNPDSN” appeared in monocots (Additional file 17: Table S6).

Statistics analysis of CLE precursor proteins

CLE peptides are derived from nonfunctional precursor proteins by removal of the N-terminal signal peptide from the latter and by enzymatical processing to yield the mature peptide [60]. In order to get a better understanding of CLE protein evolution, various characteristics of different groups were analyzed, including CLE motif score, protein length, relative position of CLE motif, length of the C-terminal tail, signal peptide scores, and correlations among the major variables of the score matrix (Fig. 5, Additional file 3: Figure S3, Additional file 7: Figure S7 and Additional file 8: Figure S8).

Group 3 had the highest median CLE motif score, followed by the rest of the groups in the following order: Group 1, Group 2, Group 4, Group 5 and Group “Others” (Fig. 5A and Additional file 18: Table S7). Although about 90% of the CLE precursor proteins are 50–150 amino acid residues long, most groups had more candidates containing 50 to 100 residues, except for Group 4 (Fig. 5B and Additional file 9: Figure S9). Group 1 to 4, particularly Group 3, had higher values for the relative position of the CLE motifs (i.e., means closer to the C-terminal end). In contrast, the motif positions in Group 5 and “others” were more widely distributed (Fig. 5C). When the number of residues following the CLE motif at the C-terminus was checked, about two thirds of the candidates had a C-terminal tail of 0 to 2 residues. More than 50% of the candidates from both Group 1 and Group 3 did not have a C-terminal tail. The basic amino acids Arginine (R) and lysine (K) dominated at the first amino acid residue position in the short C-terminal tails (1–2 residues), except for the candidates from Group 1 (Fig. 5D and Additional file 7: Figure S7). The presence of a signal peptide in the CLE precursors was predicted online using the SignalP/TargetP server and illustrated with a violin plot. Most genes in Group 1 had high signal peptide scores (Fig. 5E, F). However, about two-thirds of the genes in Grp. 2B and Grp. 5A had SignalP scores lower than the cut-off value (Additional file 19: Table S8). In general, the lengths of the CLE precursor proteins in the bryophytes and S. moellendorffii were slightly longer than the average. Other variables, including the signalP score, motif position and the CLE motif score, were not significantly different between vascular and non-vascular plants (Additional file 8: Figure S8).

To determine how much each variable contributed to each CLE candidate, correlations between the five variables and the decision to define a candidate as a CLE were calculated (Fig. 5G-5I). Motif score and motif position were decisive factors when the length of the CLE proteins was between 50 and 150 residues. Protein length was positively correlated with the decision when the candidates were shorter than 100 residues. However, the correlation was negative for the candidates between 100 and 150 residues in

length (Fig. 5G and 5H). For longer candidates (> 150 residues), the correlation between motif position and the decision was less. For these candidates, the motif score was the only decisive factor (Fig. 5I). It is worth mentioning that the correlation between the signal peptide scores and the decision was less than expected. In addition, we analyzed the gene structures of the CLE candidates in A. thaliana and Zea mays. The results provide evidence that alternative splicing may allow particular CLE genes to concurrently encode proteins with or without the CLE motif (e.g., AT5G59305/CLE46 and GRMZM5G875999) (Additional file 10: Figure S10).

Identification of new types of CLE genes

By applying our new approach, 5% (n = 136) of the CLE candidates that are more distantly related to the Arabidopsis CLEs were clustered into Group “others”. A total of 31 of these candidates were reported previously [7]. Based on the clustering, a novel subgroup of candidates (n = 26) was identified, with an unusual “serine (S)” at the 12th site of the CLE motif (Fig. 6). This subgroup could be further divided into three types, mainly based on the last three residues of their CLE motifs. All members of this subgroup were from monocots and dicots, consistent with their recent evolution.

There were three subsets of CLE candidates containing a motif similar to the Arabidopsis secreted peptide IDA “PIPPSAPSKRHN” [19], that we named the SVPP-type (n = 6), PVPP-type (n = 7) and RIPP-type (n = 8), respectively, based on their first four residues (Fig. 6). Most of the IDA/IDL-like candidates were from the monocots and dicots, except for MA_9094901g0010 and AmTr_v1.0_scaffold00135.62 from M. polymorpha and Amborella trichopoda, respectively. By clustering the IDA/IDL-like candidates together with the Arabidopsis IDA/IDL motifs, using PIP/PIPL motifs [61] as the outgroup, we found that the SVPP- and PVPP-type motifs were grouped with the IDA/IDL family, while the RIPP-type motif was more closely related to the CLV3 motif (Fig. 7A). All of the PVPP-type genes were predicted to encode a potential signaling peptide “PVPPSGPSPCHN” (Fig. 7B).

In addition to the novel CLE candidates from Group “others”, small sets of novel candidates were identified in the major groups. The most common residues at the 1st site of a typical CLE motif are arginine (R) and histidine (H). However, candidates with an initial lysine (K) or tryptophan (W) residue in the CLE motif were identified. These K-type and W-type CLE motifs are the most closely related to CLE16 (Group 3C) and CLE45 (Group 2A), respectively (Additional file 11: Figure S11). The 11 K-type candidates were all from monocots. The 13 W-type candidates were exclusively found in dicots.

Discussion

Small secreted peptides (SSPs) (e.g., CLE peptides) are hard to predict in silico because their conserved motifs are short—usually less than 20 residues in length. The commonly used method for predicting SSPs use BLAST (Basic Local Alignment Search Tool) [62]. However, when using BLAST, some thresholds should be defined, such as the S score, which provides a measure of local similarity for any pair of sequences, and the E-value, which is the probability of finding a segment pair with a score no less than the S score. It is difficult to define an appropriate threshold for E-values when using CLE as query because

it is too short to achieve a high S score and therefore, yields a much greater E-value. When using a CLE precursor protein as query, the signal peptide and the non-conserved variable region will interfere with the BLAST result. Another common method for predicting SSPs uses HMMER [63]—the latest version is HMMER3 [64]. The results from HMMER depend on the training set. Although the public database of small proteins is expanding, it still cannot meet the demand for predicting SSPs.

In this study, we retrieved all of the annotated amino acid sequences for small proteins from 69 plant species. A CLE-specific score matrix was developed because of the hidden information for peptide processing and peptide–receptor interactions in CLE motifs (Fig. 1). Three ML algorithms were applied for predicting CLE genes using multiple variables based on a variety of properties of CLE precursor proteins, in addition to a motif score matrix. A low motif score threshold was set and the union of the ML results was analyzed, in order to keep as many CLE candidates as possible. The “low stringency” strategy allowed us to uncover some candidates that are atypical in that they are less similar to the well-studied AtCLEs. By using our newly developed clustering approach for identifying CLE motifs, we were able to classify the major groups and to identify minor groups of new candidates (Fig. 3, Fig. 4 and Fig. 6). A “high stringency” version of this approach could be developed by simply increasing the threshold of the motif score and changing the ML results from union to intersection.

When a candidate had a low motif score, it probably fell into Group “others” (Fig. 6). Most of the candidates (ca. 78%) in Group “others” have not been previously reported. Several criteria are needed to determine whether a candidate from Group “others” is a CLE, including the number of similar motifs, the number of species containing the candidate, and the number of ML algorithms that support it. Candidate motifs that are identified only in one species are more suspicious than candidate motifs that are identified in more than one species.

Although it is possible to classify CLE genes based on their contributions to particular biological processes, several difficulties impede a comprehensive functional analysis of CLE genes, such as high gene redundancy, specific temporospatial expression patterns and mostly, unknown forms of the mature peptide. Knock-out lines generated using the CRISPR-Cas9 system will shed some light on the biological functions of CLE genes. However, this approach is time consuming. Moreover, transgenic manipulation remains difficult in particular species. In contrast, our new clustering method is more efficient because it considers only the amino acid compositions and site weights of CLE motifs. Functional information embedded in the major residues or the heavyweight sites will be reflected in the score matrix and the one-on-one Euclidean distances between the CLE motifs. The clustering results may in turn be helpful for functional analysis of the CLE genes that are closely grouped.

One of the main purposes of this study was to determine how CLE genes evolved in plants. We were not able to identify any typical CLE genes in the seven species of algae used in this study. The existence of CLE genes in P. patens, S. fallax and M. polymorpha provides evidence that the CLE genes evolved in bryophytes. All of the nine P. patens CLE candidates belong to Group 3B and have a consensus motif sequence of “RXVP(S/T)GPNPLHN”. The motif “RLVPTGPNPLHN” found in P. patens is one of the top10

most frequently used CLE motifs in plants, but it is not common in eudicots. A similar motif “RLVPSGPNPLHN”, found in Arabidopsis CLE9/10 was identified exclusively in eudicots. The CLE9/10 motif is the second most abundant CLE motif in dicots. The involvement of CLE9/10 in the drought response and primary root development in A. thaliana [27, 65] is consistent with the peptide “RLVPTGPNPLHN” helping bryophytes to develop adaptations to survive in more arid environments. Another interesting finding in bryophytes is the evolution of the Group 4 candidate Mapoly1011s0001.1 in M. polymorpha. Its potential motif “HKNPAGPNPIGN” is identical to the Arabidopsis CLE46 motif “HKHPSGPNPTGN” at all of the conserved sites (Fig. 1B and Additional file 5: Figure S5) [40, 41]. CLE46 is highly homologous to CLE41 and CLE44—two TDIF encoding genes in Arabidopsis [39]. However, similar to other liverworts, M. polymorpha has neither vascular tissue nor true roots. Therefore, the presence of a CLE46-like gene in M. polymorpha remains mysterious. Nevertheless, the number of candidate genes in Group 4 rapidly increased in vascular plants, especially genes encoding candidates with the TDIF motif “HEVPSGPNPISN”. The largest number of candidates in dicots contain the TDIF motif (Additional file 17: Table S6).

Besides the CLE gene family, several gene families have been identified that encode SSPs [3, 20]. Among them, the CLEL/GLV/RGF and IDA/IDL motifs share high sequence similarities with the CLE motif [19, 66]. However, our knowledge of the evolutional relationship among these peptide-coding genes remains limited. Based on our less-stringent gene prediction strategy, it is possible to compile a list of atypical CLE genes. We found three types of candidates: true CLE genes, non-peptide-coding genes and novel peptide-coding genes. We identified 21 candidates that belong to three small but conserved groups in Group “others” (Fig. 6). Their potential CLE motifs are highly similar to the IDA/IDAL motifs and thus, appear to represent a transitional type of CLE and IDA/IDL motif. The IDA/IDL genes are involved in floral organ abscission, lateral root emergence and root cap sloughing [67]. Since we have not found any typical IDA/IDL genes in P. patens, S. fallax and M. polymorpha, the transitional CLE/IDA candidate MA_9094901g0010 could be very important for functional and evolutionary studies in the future (Fig. 6 and Fig. 7).

This study was based on a global analysis of the annotated genes from 69 plant species, from single-cell green algae to giant trees. Comparative analysis of CLE gene family sequences from multiple species could increase the reliability of gene prediction and characterization and thus, provide information on how these genes have evolved. There are a few challenges remaining for future work. First, the number of lower plant and lower vascular plant species used in this study was limited. The availability of more genome sequences from bryophyte and pteridophyte species will be useful for understanding the origin and evolution of SSP-encoding genes. Second, the quality of genome annotation varies considerably, mainly due to the complexity of each genome and the quality of genome sequencing, assembly and annotation. Thus, high genome complexity or low genomic sequencing quality will increase the frequency of miss counts of SSP-encoding genes. Furthermore, SSP-encoding genes could not be effectively predicted and/or annotated [68]. It is difficult to distinguish them from non-coding sequences because their coding regions are small. More than this, without a reference gene, there is no effective method to predict an SSP-encoding gene when alternative splicing introduces additional complexity. Regarding

particular types of SSPs that are variable in length, more research is required for determining how to set a gap penalty for SSP prediction. The in silico prediction of SSPs that are present in single-copy or low copy numbers (e.g., Casparian Strip Integrity Factor (CIF) from Arabidopsis) [69, 70] requires a comparative genomics analysis with multiple species. In addition, integration of next-generation sequencing (NGS)-based transcriptomics and mass spectrometry (MS)-based proteomics analyses will provide essential information about SSPs, especially the novel SSPs.

Conclusions

In summary, we developed a novel machine learning-aided method for predicting CLE genes from 69 plant species by using a CLE motif-specific residual score matrix. We found 2156 CLE candidates, including 627 novel CLE candidates. We also developed a novel clustering method based on the Euclidean distance of CLE motifs in a site-weight dependent manner. Our grouping was relatively consistent with the previous reports by Oelkers et al. [58] and Goad et al. [7] Moreover, the advantage of this new clustering method is that it does not require any flanking sequences from the CLE motifs. Characterization of CLE candidates suggested that ca. 90% of the CLE precursor proteins have a protein length of 50 to 150 amino acid residues, about 30% of the CLE candidates may not have a signal peptide targeting them to the secretory pathway, two-thirds of the CLE candidates we identified have a short C-terminal tail (i.e., 0–2 residues) downstream of their CLE motifs, and the CLE motif score was the only decisive factor for identifying candidates longer than 150 residues. These characteristics are important for classifying novel candidates as CLE genes. The approach taken here not only helps us to investigate the evolution of the CLE gene family, but also allows us to discover a potential evolutionary relationship between the CLE and IDA/IDL gene families. The IDA/IDL-like CLE candidates represent a missing link between the two families and provide evidence that the CLE and IDA/IDL genes probably share a common ancestor. Our novel approach for predicting and clustering CLE genes may also be applicable to other SSPs and, therefore may provide a powerful tool for studying the origin and evolution of SSPs.

Methods

Developing a new residual score matrix for CLE motifs in plants

The new residual score matrix for CLE motifs was developed by integrating the amino acid substitution matrix, the amino acid usage frequency matrix of CLE motifs and the site weights of CLE motifs.

To find an optimal amino acid substitution matrix, three commonly used substitution matrices, BLOSUM62, BLOSUM80 and PAM250, were tested using 116 *CLE* candidates from *A. thaliana*, *O. sativa*, *S. moellendorfii* and *P. abies* (Additional file 20: Table S9). The scores of these 116 reported *CLE* genes followed an order from large to small and were fitted to a curve using the Local Polynomial Regression Fitting (LOESS) method. A matrix with the highest sensitivity was chosen to construct the score matrix for the subsequent analyses.

To develop the amino acid usage frequency matrix for CLE motifs, we used the 1628 reported *CLE* genes as references [7]. The percentage of each amino acid residue *S* at each of the 12 sites of the CLE motif was calculated as follows:

$$S_{ij} = \frac{a_{ij}}{n} \times 100\%$$

where S_{ij} represents the percentage of amino acid *i* at site *j*; a_{ij} represents the number of amino acids *i* at site *j* and *n* represents the number of reported *CLE* genes.

The weight of each site (w_j) in the CLE motif was based on the *Bits* value of each site [71]. The modified *Bits* values (*Bits'*) were used to assign a weight to each site of the CLE motif with the following steps: (1) select amino acids with $S_{ij} \geq 25\%$ as the major amino acids for each site, (2) combine S_{ij} values for these amino acids, and (3) calculate the *Bits'* values based on the ratio of the amino acids at each site using the following equation:

$$Bits'(j) = \log_2(m-k+1) - (H_j' + e_m)$$

$$w_j = Bits'(j) / \max(Bits')$$

where *m* represents the types of amino acids (m = 20), *k* represents the number of amino acids with $S_{ij} \geq$ 25% at site *j*, H_j' represents the modified entropy of site *j*, and e_m is the correction number, which was mainly applied when the number of input sequences was less than 20.

A novel residual score matrix *N* was then constructed by integrating the amino acid substitution matrix *M*, the amino acid usage frequency matrix *S* and the site weight w_j:

$$N_{ij} = w_j \times \sum_{k=1}^{n} (S_{jk} \times M_{ik})$$

where M_{ik} represents the substitution score between amino acid *i* and amino acid *k* in the amino acid substitution matrix, S_{jk} represents the frequency of amino acid *k* at site *j* in the amino acid usage frequency matrix. The motif score *v* of each CLE motif was calculated by applying the novel score matrix *N*:

$$v = \sum_{j=1}^{12} N_{ij}$$

where *i* represents an amino acid of the CLE motif at site *j*.

Other variables for the prediction of *CLE* genes

Besides the score of each CLE motif (*v*), other factors were taken into consideration to predict *CLE* genes, including protein length (*L*), signal peptide scores (*SP* for SignalP score/D-value; *TP* for TargetP score/SP-value), and motif position (*P*). Signal peptide scores for each protein sequence was calculated on the SignalP 4.1 Server (http://www.cbs.dtu.dk/services/SignalP/) [72] with a sensitive D-cutoff value (0.34 for SignalP, no TM networks only) and the TargetP 1.1 Server (http://www.cbs.dtu.dk/services/TargetP/) [73] with the plant group. Motif position *P* was calculated as follows:

$$P = \frac{l_s}{L - 11}$$

where *L* represents the length of the corresponding protein and l_s represents the start position of the CLE motif.

Machine learning aided prediction of *CLE* genes in plants

The coding sequences of 68 species were extracted at the whole genome level from Phytozome v12 (https://phytozome.jgi.doe.gov/pz/portal.html) [74]. Coding sequences of *P. abies* were downloaded from PlantGenIE (http://plantgenie.org) [75, 76]. First, we filtered out protein sequences with $L < 30$ and $L > 300$. For the remaining protein sequences ($30 \leq L \leq 300$), we calculated a motif score for any fragment containing 12 amino acid residues. A motif with the highest score was chosen as a potential CLE motif for this protein. The 1529 reported *CLE* genes identified using the HMM algorithms from 53 species in the Phytozome v12.1 database [7] were labeled as *CLE* genes. The number of *CLE* genes (*X*) in a particular species was counted. If $X \leq 10$, 30 candidates with the highest scores were selected. For a species with $X > 10$, $3X$ candidates with the highest scores were selected. All of the *CLE* genes were removed from the list of candidate genes. The remaining genes were defined as non-*CLE* genes. To build the training data set, the *CLE* and non-*CLE* genes were combined.

Three machine learning algorithms, C4.5, Artificial Neural Network (ANN) and Support Vector Machine (SVM), were used to analyze the training dataset using the above-mentioned five variables. All three algorithms were implemented in the R language (R-3.4.0): C4.5, RWeka-0.4-37 package. The Confidence Pruning Factor *C* was set to 0.01 (ANN, nnet-7.3-12 package). The number of neurons in the hidden layer size was set to 20. The maximum number of iterations was set to 1000 (SVM, e1071_1.6-8 package). The default settings were used for the other parameters. Candidate genes from the remaining 16 species were used for the testing data set. Candidate CLEs were supported by at least one of the three classifiers.

Clustering of *CLE* genes in plants

The CLE candidates predicted by machine learning were further clustered using a novel protocol based on the Euclidean distance (*d*). The Euclidean distance between each candidate sequence and each reported Arabidopsis CLE motif was calculated to find its minimum distance (d_{min}). The top 5% of motifs with the maximum d_{min} were categorized into the "others" group. The modified Euclidean distance (*d*) between every two CLE motifs was as follow:

$$d = \sqrt{\sum_{j=1}^{12} d_j^2}$$

$$d_j = \begin{cases} 0 & (a_j = b_j) \\ w_j & (a_j \neq b_j) \end{cases}$$

Where a_j represents the amino acid at site *j* of a candidate CLE motif and b_j represents the amino acid at site *j* of *A. thaliana* CLE motifs. The distance between a_j and b_j was defined as d_j.

For all grouped *CLE* candidate genes, a hierarchical clustering (HCL) method was applied with R (R-3.4.0) to build a clustering tree. Phylogenetic trees of *A. thaliana* CLE motifs, full-length CLE proteins without signal peptides and log-normalized rank of all-vs-all BLAST e-values were constructed using the neighbor-joining (NJ) method with MEGA X [77]. The clustering trees and phylogenetic trees were edited using Evolview (http://www.evolgenius.info/evolview/) [78].

Statistical analysis

To find out the bias in amino acid usage in CLE precursor proteins and CLE motifs, the amino acid composition of all proteins, all small proteins (i.e., proteins with lengths between 50 and 200 amino acid residues) and all CLE candidates were analyzed in 69 plant species. To study the evolution of the *CLE* genes, the numbers of *CLE* candidate genes were counted in each species and in each group. Characterization of CLE precursor proteins was performed by analyzing the distribution of motif scores, protein lengths, motif positions, lengths of C-terminal tails, SignalP and TargetP scores of each CLE candidate by group. Decisive factors in determining a *CLE* candidate gene were uncovered using a correlation analysis between each of the above-mentioned variables and the decision in three ranges of protein lengths, 51-100, 101-150 and > 150 amino acid residues. The clustering trees of CLE candidates in Group "others" and IDA-like candidates were built by applying the HCL method—based on Euclidean distance—to every pair of candidate sequences. The lengths of the C-terminal tails and their corresponding amino acid compositions in each subgroup were evaluated using a heatmap that showed the counts of CLE candidates with different lengths of C-terminal tails and using Weblogos to represent the conserved residues. For *A. thaliana* and *Z. mays*, the gene structures of the *CLE* candidates with

alternative splicing were obtained from the gff3 files of *A. thaliana* and *Zea mays* in Phytozome v12 (https://phytozome.jgi.doe.gov/pz/portal.html).

Weblogo (http://weblogo.threeplusone.com/) [79] was used to create the sequence logo. MEME (http://meme-suite.org/tools/meme) [80] was used to calculate the e-values of each CLE motif. The R package pheatmap-1.0.17, corrplot-0.84 and UpSetR-1.4.0 were applied to create heat maps, correlation maps and upset plots, respectively. Other plots were created using ggplot2-3.2.0. All data were processed using the R language (R-3.4.0 and R-3.6.1).

Abbreviations

CLE
CLV3/ESR-RELATED
SSP
Small Secreted Peptides
TDIF
Tracheary element Differentiation Inhibitory Factor

Declarations

Plant material

No plant materials were used in this study.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

All data generated or analysed during this study are included in this published article and its supplementary information files.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by Fundamental Research Funds for the Central Universities (2662018PY071, 2662015PY130), the National Natural Science Foundation of China (31370673 and 31770639), the National Key Research and Development Program of China (2016YFD0600103), and a Postdoctoral Researcher Grant from Academy of Finland (No. 326036). The funders provide financial support to cover all the running expense of the experiments, student allowance and article-processing charges of this study.

Author Contributions

B.Z. and X.S. designed the research. Z.Z., X.S. and L.L. collected and analyzed the data. Z.Z. drafted the manuscript. All authors discussed the results and made comments on the manuscript.

Acknowledgements

We would like to thank Prof. Lingling Chen and Prof. Manzhu Bao for their constructive suggestions.

References

1. Ryan CA, Pearce G, Scheer J, Moura DS. Polypeptide hormones. Plant Cell. 2002;14 Suppl:S251-S264.
2. Matsubayashi Y, Sakagami Y. Peptide hormones in plants. Annu. Rev. Plant Biol. 2006;57:649-674.
3. Murphy E, Smith S, De Smet I. Small signaling peptides in Arabidopsis development: how cells communicate over a short distance. Plant Cell. 2012;24(8):3198-3217.
4. Matsubayashi Y. Posttranslationally modified small-peptide signals in plants. Annu. Rev. Plant Biol. 2014;65:385-413.
5. Clark SE, Running MP, Meyerowitz EM. CLAVATA1, a regulator of meristem and flower development in Arabidopsis. Development. 1993;119(2):397-418.
6. Fletcher JC, Brand U, Running MP, Simon R, Meyerowitz EM. Signaling of cell fate decisions by CLAVATA3 in Arabidopsis shoot meristems. Science. 1999;283(5409):1911-1914.
7. Goad DM, Zhu C, Kellogg EA. Comprehensive identification and clustering of CLV3/ESR-related (CLE) genes in plants finds groups with potentially shared function. New Phytol. 2017;216(2):605-616.
8. Fiers M, Golemiec E, Xu J, van der Geest L, Heidstra R, Stiekema W et al. The 14-amino acid CLV3, CLE19, and CLE40 peptides trigger consumption of the root meristem in Arabidopsis through a CLAVATA2-dependent pathway. Plant Cell. 2005;17(9):2542-2553.
9. Fiers M, Golemiec E, van der Schors R, van der Geest L, Li KW, Stiekema WJ et al. The CLAVATA3/ESR motif of CLAVATA3 is functionally independent from the nonconserved flanking sequences. Plant Physiol. 2006;141(4):1284-1292.
10. Hirakawa Y, Shinohara H, Kondo Y, Inoue A, Nakanomyo I, Ogawa M et al. Non-cell-autonomous control of vascular stem cell fate by a CLE peptide/receptor system. Proc Natl Acad Sci U S A. 2008; 105(39):15208-15213.

11. Ohyama K, Ogawa M, Matsubayashi Y. Identification of a biologically active, small, secreted peptide in Arabidopsis by in silico gene screening, followed by LC-MS-based structure analysis. Plant J. 2008;55(1):152-160.
12. Ohyama K, Shinohara H, Ogawa-Ohnishi M, Matsubayashi Y. A glycopeptide regulating stem cell fate in Arabidopsis thaliana. Nat. Chem. Biol. 2009;5(8):578-580.
13. Shinohara H, Moriyama Y, Ohyama K, Matsubayashi Y. Biochemical mapping of a ligand-binding domain within Arabidopsis BAM1 reveals diversified ligand recognition mechanisms of plant LRR-RKs. Plant J. 2012;70(5):845-854.
14. Xu C, Liberatore KL, MacAlister CA, Huang Z, Chu YH, Jiang K et al. A cascade of arabinosyltransferases controls shoot meristem size in tomato. Nat. Genet. 2015;47(7):784-792.
15. Kim HJ, Wu CY, Yu HM, Sheen J, Lee H. Dual CLAVATA3 Peptides in Arabidopsis Shoot Stem Cell Signaling. J. Plant Biol. 2017;60(5):506-512.
16. DeYoung BJ, Bickle KL, Schrage KJ, Muskett P, Patel K, Clark SE. The CLAVATA1-related BAM1, BAM2 and BAM3 receptor kinase-like proteins are required for meristem function in Arabidopsis. Plant J. 2006;45(1):1-16.
17. Fisher K, Turner S. PXY, a receptor-like kinase essential for maintaining polarity during plant vascular-tissue development. Curr. Biol. 2007;17(12):1061-1066.
18. Hazak O, Hardtke CS. CLAVATA 1-type receptors in plant development. J. Exp. Bot. 2016;67(16):4827-4833.
19. Butenko MA, Vie AK, Brembu T, Aalen RB, Bones AM. Plant peptides in signalling: looking for new partners. Trends Plant Sci. 2009;14(5):255-263.
20. Olsson V, Joos L, Zhu S, Gevaert K, Butenko MA, De Smet I. Look Closely, the Beautiful May Be Small: Precursor-Derived Peptides in Plants. Annu. Rev. Plant Biol. 2019;70:153-186.
21. Schoof H, Lenhard M, Haecker A, Mayer KF, Jurgens G, Laux T. The stem cell population of Arabidopsis shoot meristems in maintained by a regulatory loop between the CLAVATA and WUSCHEL genes. Cell. 2000;100(6):635-644.
22. Stahl Y, Wink RH, Ingram GC, Simon R. A signaling module controlling the stem cell niche in Arabidopsis root meristems. Curr. Biol. 2009;19(11):909-914.
23. Hirakawa Y, Kondo Y, Fukuda H. Regulation of vascular development by CLE peptide-receptor systems. J. Integr. Plant Biol. 2010;52(1):8-16.
24. Bidadi H, Matsuoka K, Sage-Ono K, Fukushima J, Pitaksaringkarn W, Asahina M et al. CLE6 expression recovers gibberellin deficiency to promote shoot growth in Arabidopsis. Plant J. 2014;78(2):241-252.
25. Fiume E, Fletcher JC. Regulation of Arabidopsis embryo and endosperm development by the polypeptide signaling molecule CLE8. Plant Cell. 2012;24(3):1000-1012.
26. Kondo Y, Hirakawa Y, Kieber JJ, Fukuda H. CLE peptides can negatively regulate protoxylem vessel formation via cytokinin signaling. Plant Cell Physiol. 2011;52(1):37-48.

27. Qian P, Song W, Yokoo T, Minobe A, Wang G, Ishida T et al. The CLE9/10 secretory peptide regulates stomatal and vascular development through distinct receptors. Nat Plants. 2018;4(12):1071-1081.
28. Meng L, Feldman LJ. CLE14/CLE20 peptides may interact with CLAVATA2/CORYNE receptor-like kinases to irreversibly inhibit cell division in the root meristem of Arabidopsis. Planta. 2010;232(5):1061-1074.
29. Zhu Y, Song D, Zhang R, Luo L, Cao S, Huang C et al. A xylem-produced peptide PtrCLE20 inhibits vascular cambium activity in Populus. Plant Biotechnol. J. 2019.
30. Takahashi F, Suzuki T, Osakabe Y, Betsuyaku S, Kondo Y, Dohmae N et al. A small peptide modulates stomatal control via abscisic acid in long-distance signalling. Nature. 2018;556(7700):235-238.
31. Ren SC, Song XF, Chen WQ, Lu R, Lucas WJ, Liu CM. CLE25 peptide regulates phloem initiation in Arabidopsis through a CLERK-CLV2 receptor complex. J. Integr. Plant Biol. 2019.
32. Czyzewicz N, Shi CL, Vu LD, Van De Cotte B, Hodgman C, Butenko MA et al. Modulation of Arabidopsis and monocot root architecture by CLAVATA3/EMBRYO SURROUNDING REGION 26 peptide. J. Exp. Bot. 2015;66(17):5229-5243.
33. Rodriguez-Villalon A, Gujas B, van Wijk R, Munnik T, Hardtke CS. Primary root protophloem differentiation requires balanced phosphatidylinositol-4,5-biphosphate levels and systemically affects root branching. Development. 2015;142(8):1437-1446.
34. Czyzewicz N, De Smet I. The Arabidopsis thaliana CLAVATA3/EMBRYO-SURROUNDING REGION 26 (CLE26) peptide is able to alter root architecture of Solanum lycopersicum and Brassica napus. Plant Signal Behav. 2016;11(1):e1118598.
35. Depuydt S, Rodriguez-Villalon A, Santuari L, Wyser-Rmili C, Ragni L, Hardtke CS. Suppression of Arabidopsis protophloem differentiation and root meristem growth by CLE45 requires the receptor-like kinase BAM3. Proc Natl Acad Sci U S A. 2013;110(17):7074-7079.
36. Endo S, Shinohara H, Matsubayashi Y, Fukuda H. A novel pollen-pistil interaction conferring high-temperature tolerance during reproduction via CLE45 signaling. Curr. Biol. 2013;23(17):1670-1676.
37. Kondo T, Nakamura T, Yokomine K, Sakagami Y. Dual assay for MCLV3 activity reveals structure-activity relationship of CLE peptides. Biochem Biophys Res Commun. 2008;377(1):312-316.
38. Song XF, Guo P, Ren SC, Xu TT, Liu CM. Antagonistic peptide technology for functional dissection of CLV3/ESR genes in Arabidopsis. Plant Physiol. 2013;161(3):1076-1085.
39. Ito Y, Nakanomyo I, Motose H, Iwamoto K, Sawa S, Dohmae N et al. Dodeca-CLE peptides as suppressors of plant stem cell differentiation. Science. 2006;313(5788):842-845.
40. Morita J, Kato K, Nakane T, Kondo Y, Fukuda H, Nishimasu H et al. Crystal structure of the plant receptor-like kinase TDR in complex with the TDIF peptide. Nat. Commun. 2016;7:12383.
41. Zhang H, Lin X, Han Z, Qu LJ, Chai J. Crystal structure of PXY-TDIF complex reveals a conserved recognition mechanism among CLE peptide-receptor pairs. Cell Res. 2016;26(5):543-555.
42. Cock JM, McCormick S. A large family of genes that share homology with CLAVATA3. Plant Physiol. 2001;126(3):939-942.

43. Olsen AN, Skriver K. Ligand mimicry? Plant-parasitic nematode polypeptide with similarity to CLAVATA3. Trends Plant Sci. 2003;8(2):55-57.
44. Kondo T, Sawa S, Kinoshita A, Mizuno S, Kakimoto T, Fukuda H et al. A plant peptide encoded by CLV3 identified by in situ MALDI-TOF MS analysis. Science. 2006;313(5788):845-848.
45. Jun JH, Fiume E, Fletcher JC. The CLE family of plant polypeptide signaling molecules. Cell. Mol. Life Sci. 2008;65(5):743-755.
46. Kinoshita A, Nakamura Y, Sasaki E, Kyozuka J, Fukuda H, Sawa S. Gain-of-function phenotypes of chemically synthetic CLAVATA3/ESR-related (CLE) peptides in Arabidopsis thaliana and Oryza sativa. Plant Cell Physiol. 2007;48(12):1821-1825.
47. Okamoto S, Ohnishi E, Sato S, Takahashi H, Nakazono M, Tabata S et al. Nod factor/nitrate-induced CLE genes that drive HAR1-mediated systemic regulation of nodulation. Plant Cell Physiol. 2009;50(1):67-77.
48. Miwa H, Tamaki T, Fukuda H, Sawa S. Evolution of CLE signaling: origins of the CLV1 and SOL2/CRN receptor diversity. Plant Signal Behav. 2009;4(6):477-481.
49. Mortier V, Den Herder G, Whitford R, Van de Velde W, Rombauts S, D'Haeseleer K et al. CLE peptides control Medicago truncatula nodulation locally and systemically. Plant Physiol. 2010;153(1):222-237.
50. Hastwell AH, de Bang TC, Gresshoff PM, Ferguson BJ. Author Correction: CLE peptide-encoding gene families in Medicago truncatula and Lotus japonicus, compared with those of soybean, common bean and Arabidopsis. Sci Rep. 2017;7(1):15474.
51. Strabala TJ, Phillips L, West M, Stanbra L. Bioinformatic and phylogenetic analysis of the CLAVATA3/EMBRYO-SURROUNDING REGION (CLE) and the CLE-LIKE signal peptide genes in the Pinophyta. BMC Plant Biol. 2014;14:47.
52. Zhang Y, Yang S, Song Y, Wang J. Genome-wide characterization, expression and functional analysis of CLV3/ESR gene family in tomato. BMC Genomics. 2014;15:827.
53. Hastwell AH, Gresshoff PM, Ferguson BJ. Genome-wide annotation and characterization of CLAVATA/ESR (CLE) peptide hormones of soybean (Glycine max) and common bean (Phaseolus vulgaris), and their orthologues of Arabidopsis thaliana. J. Exp. Bot. 2015;66(17):5271-5287.
54. Gancheva MS, Dodueva IE, Lebedeva MA, Tvorogova VE, Tkachenko AA, Lutova LA. Identification, expression, and functional analysis of CLE genes in radish (Raphanus sativus L.) storage root. BMC Plant Biol. 2016;16 Suppl 1:7.
55. Han H, Zhang G, Wu M, Wang G. Identification and characterization of the Populus trichocarpa CLE family. BMC Genomics. 2016;17:174.
56. Whitford R, Fernandez A, De Groodt R, Ortega E, Hilson P. Plant CLE peptides from two distinct functional classes synergistically induce division of vascular cells. Proc Natl Acad Sci U S A. 2008;105(47):18625-18630.
57. Hirakawa Y, Kondo Y, Fukuda H. Establishment and maintenance of vascular cell communities through local signaling. Curr. Opin. Plant Biol. 2011;14(1):17-23.

58. Oelkers K, Goffard N, Weiller GF, Gresshoff PM, Mathesius U, Frickey T. Bioinformatic analysis of the CLE signaling peptide family. BMC Plant Biol. 2008;8:1.
59. Kucukoglu M, Nilsson O. CLE peptide signaling in plants - the power of moving around. Physiol Plant. 2015;155(1):74-87.
60. Tavormina P, De Coninck B, Nikonorova N, De Smet I, Cammue BP. The Plant Peptidome: An Expanding Repertoire of Structural Features and Biological Functions. Plant Cell. 2015;27(8):2095-2118.
61. Vie AK, Najafi J, Liu B, Winge P, Butenko MA, Hornslien KS et al. The IDA/IDA-LIKE and PIP/PIP-LIKE gene families in Arabidopsis: phylogenetic relationship, expression patterns, and transcriptional effect of the PIPL3 peptide. J. Exp. Bot. 2015;66(17):5351-5365.
62. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J. Mol. Biol. 1990;215(3):403-410.
63. Eddy SR. Profile hidden Markov models. Bioinformatics. 1998;14(9):755-763.
64. Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res. 2013;41(12):e121.
65. Zhang L, Shi X, Zhang Y, Wang J, Yang J, Ishida T et al. CLE9 peptide-induced stomatal closure is mediated by abscisic acid, hydrogen peroxide, and nitric oxide in Arabidopsis thaliana. Plant Cell Environ. 2019;42(3):1033-1044.
66. Meng L, Buchanan BB, Feldman LJ, Luan S. CLE-like (CLEL) peptides control the pattern of root growth and lateral root development in Arabidopsis. Proc Natl Acad Sci U S A. 2012;109(5):1760-1765.
67. Shi CL, Alling RM, Hammerstad M, Aalen RB. Control of Organ Abscission and Other Cell Separation Processes by Evolutionary Conserved Peptide Signaling. Plants (Basel). 2019;8(7).
68. Takahashi F, Hanada K, Kondo T, Shinozaki K. Hormone-like peptides and small coding genes in plant stress signaling and development. Curr. Opin. Plant Biol. 2019;51:88-95.
69. Doblas VG, Smakowska-Luzan E, Fujita S, Alassimone J, Barberon M, Madalinski M et al. Root diffusion barrier control by a vasculature-derived peptide binding to the SGN3 receptor. Science. 2017;355(6322):280-284.
70. Nakayama T, Shinohara H, Tanaka M, Baba K, Ogawa-Ohnishi M, Matsubayashi Y. A peptide hormone required for Casparian strip diffusion barrier formation in Arabidopsis roots. Science. 2017;355(6322):284-286.
71. Schneider TD, Stephens RM. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. 1990;18(20):6097-6100.
72. Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat. Methods. 2011;8(10):785-786.
73. Emanuelsson O, Nielsen H, Brunak S, von Heijne G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J. Mol. Biol. 2000;300(4):1005-1016.

74. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J et al. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. 2012;40(Database issue):D1178-D1186.
75. Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin YC, Scofield DG et al. The Norway spruce genome sequence and conifer genome evolution. Nature. 2013;497(7451):579-584.
76. Sundell D, Mannapperuma C, Netotea S, Delhomme N, Lin YC, Sjodin A et al. The Plant Genome Integrative Explorer Resource: PlantGenIE.org. New Phytol. 2015;208(4):1149-1156.
77. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol. Biol. Evol. 2018;35(6):1547-1549.
78. He Z, Zhang H, Gao S, Lercher MJ, Chen WH, Hu S. Evolview v2: an online visualization and management tool for customized and annotated phylogenetic trees. Nucleic Acids Res. 2016;44(W1):W236-W241.
79. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome Res. 2004;14(6):1188-1190.
80. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L et al. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 2009;37(Web Server issue):W202-W208.

Figures

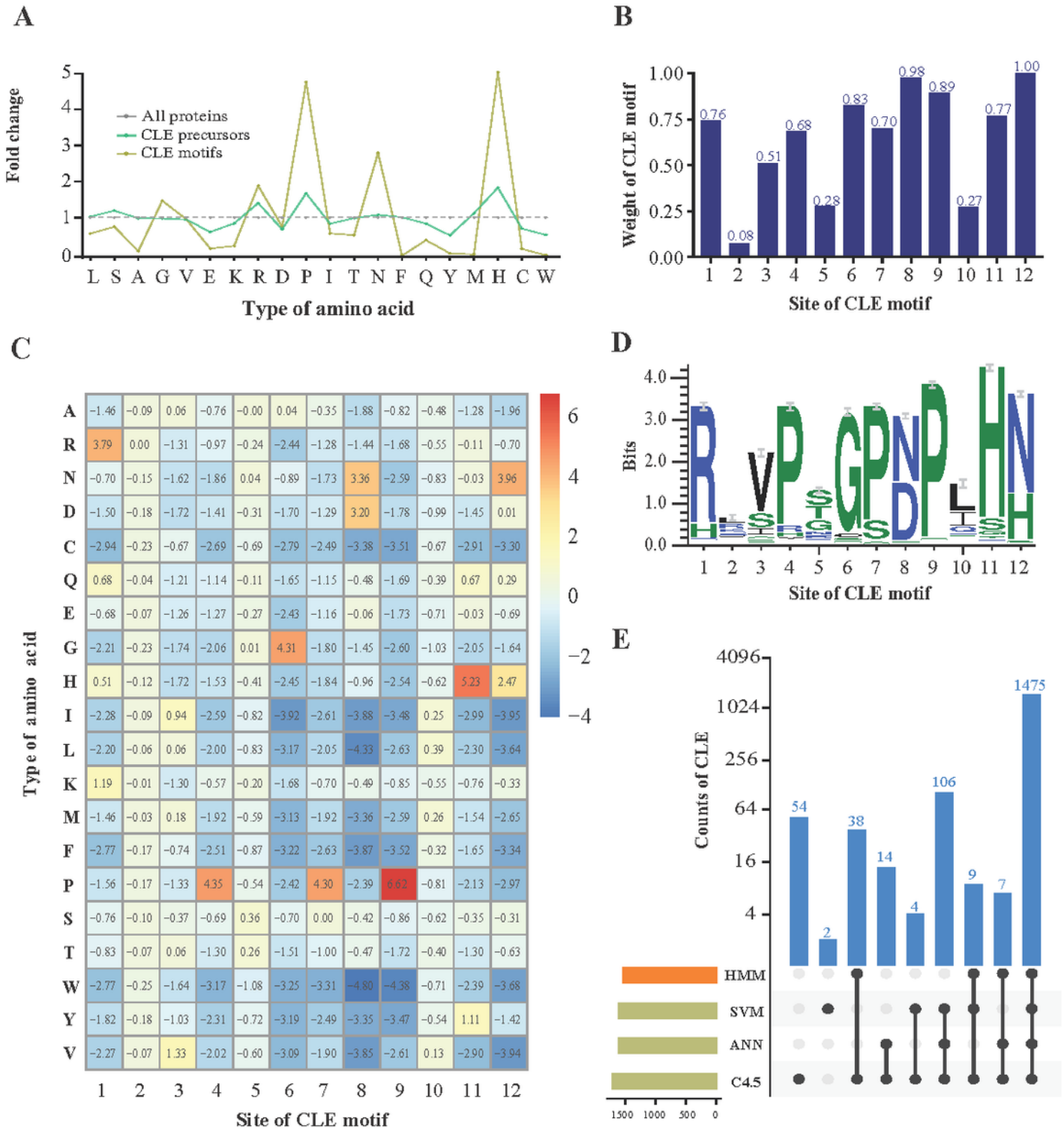

Figure 1

Methods and results for predicting CLE genes. (A) Fold changes of amino acid frequencies of CLE precursors and CLE motifs in 69 species. The amino acid composition of all proteins was used as a control (set to 1.0). The grey, aquamarine and lemon colored lines indicate all proteins, CLE precursors and CLE motifs, respectively. (B) The weight at each site of the CLE motifs. (C) Score matrix of CLE motifs. The amino acids are indicated at the left using the single letter code. The numbers in the grid

represent the score of each amino acid at sites 1 through 12. (D) Weblogo of the 12-residue CLE motif from the 1529 reported CLE genes[7]. (E) UpSet plot for visualizing the intersecting sets of CLE genes predicted by different methods. The number of CLE genes at each intersection was labeled in blue on the top of the appropriate column.

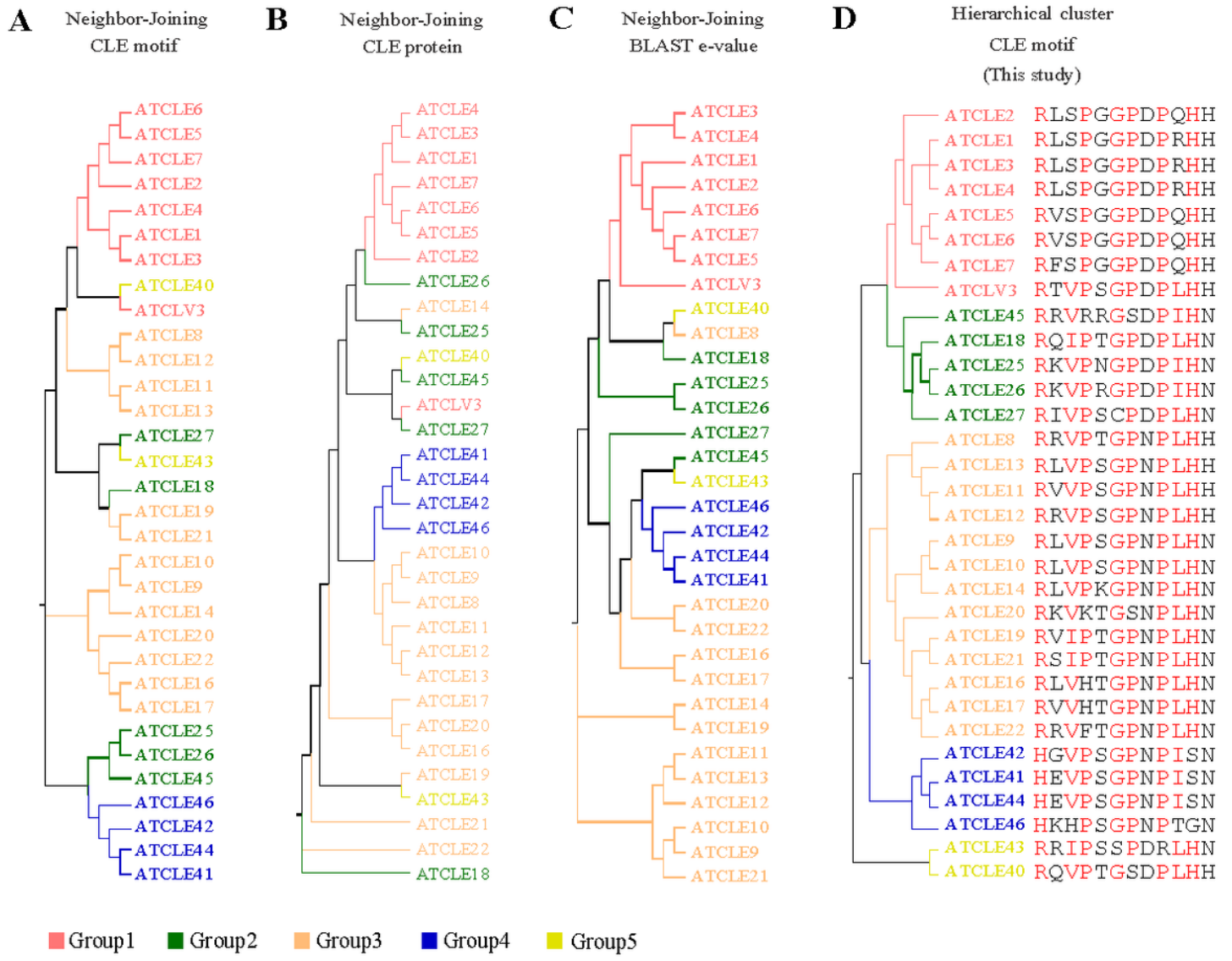

Figure 2

Clustering analysis of Arabidopsis CLE motifs. Phylogenetic tree of AtCLE (A) motifs, (B) full-length proteins without signal peptides and (C) log-normalized rank of all-vs-all BLAST e-values were generated using the NJ method based on the evolutionary distances, which were computed using the Poisson correction method (A, B), and Euclidean distances (C). (D) Clustering of the AtCLE motifs based on the Euclidean distance of each pair of sequences in a site-weight dependent manner. The tree was constructed using the HCL method. The names of the CLE motifs are indicated with different colors.

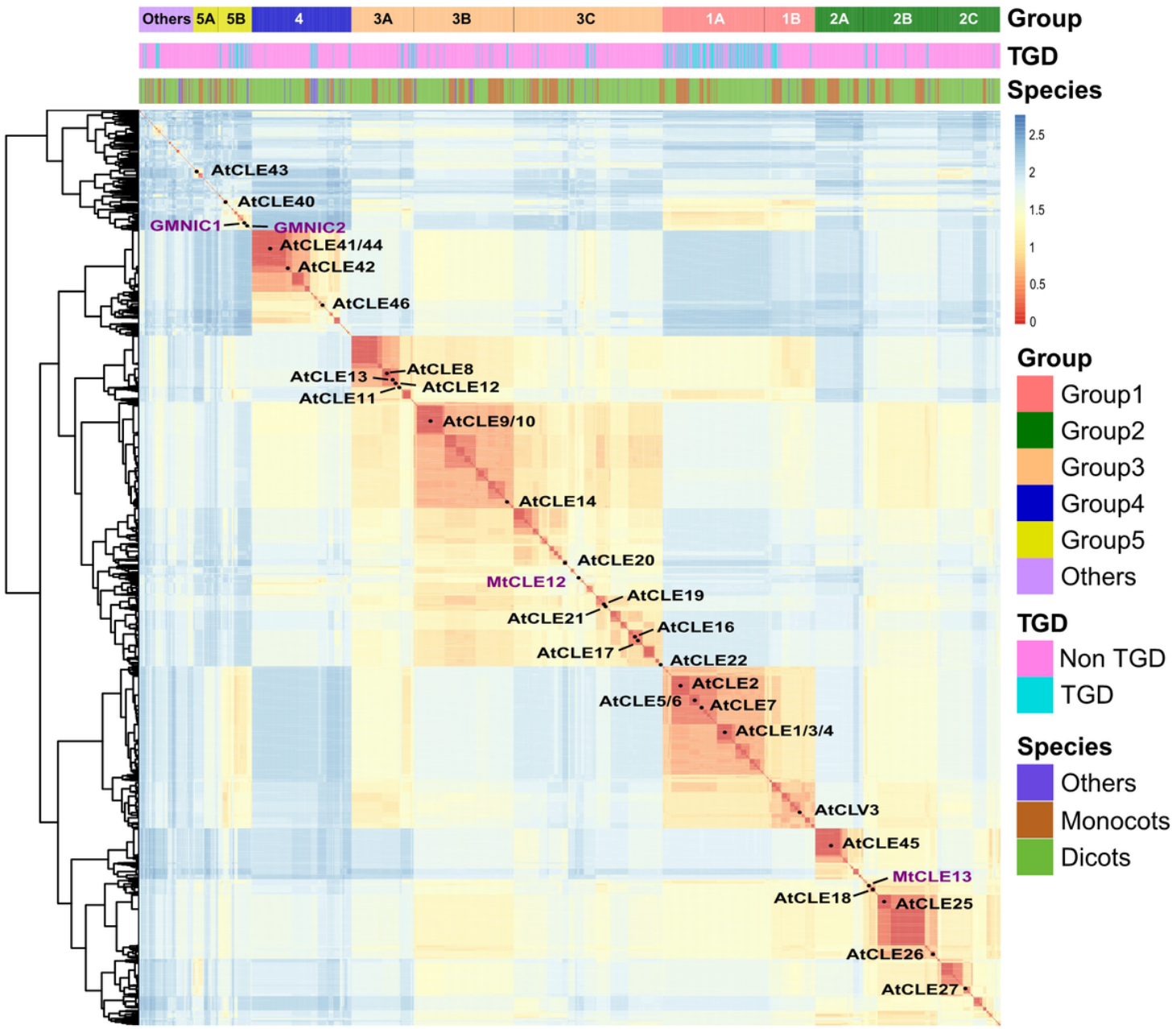

Figure 3

Clustering analysis of CLE motifs in plants. The heat map shows the Euclidean distance of 2156 CLE motifs in 69 plant species. Red represents short distances. Blue represents long distances. A shorter Euclidean distance implies a higher degree of motif similarity. CLE motifs were clustered based on the Euclidean distance of each pair of sequences in a site-weight dependent manner. The clustering tree was generated using the HCL method. The information on the classification of the CLE motifs is shown on the top of the heatmap. All CLE motifs were clustered into six major groups: Group 1-5 and Group “others”. “TGD” and “Non-TGD” indicate whether the motif was from a potential tandem gene duplication (TGD). “Species” indicates that a motif was from a dicot, monocot or other type of plant species.

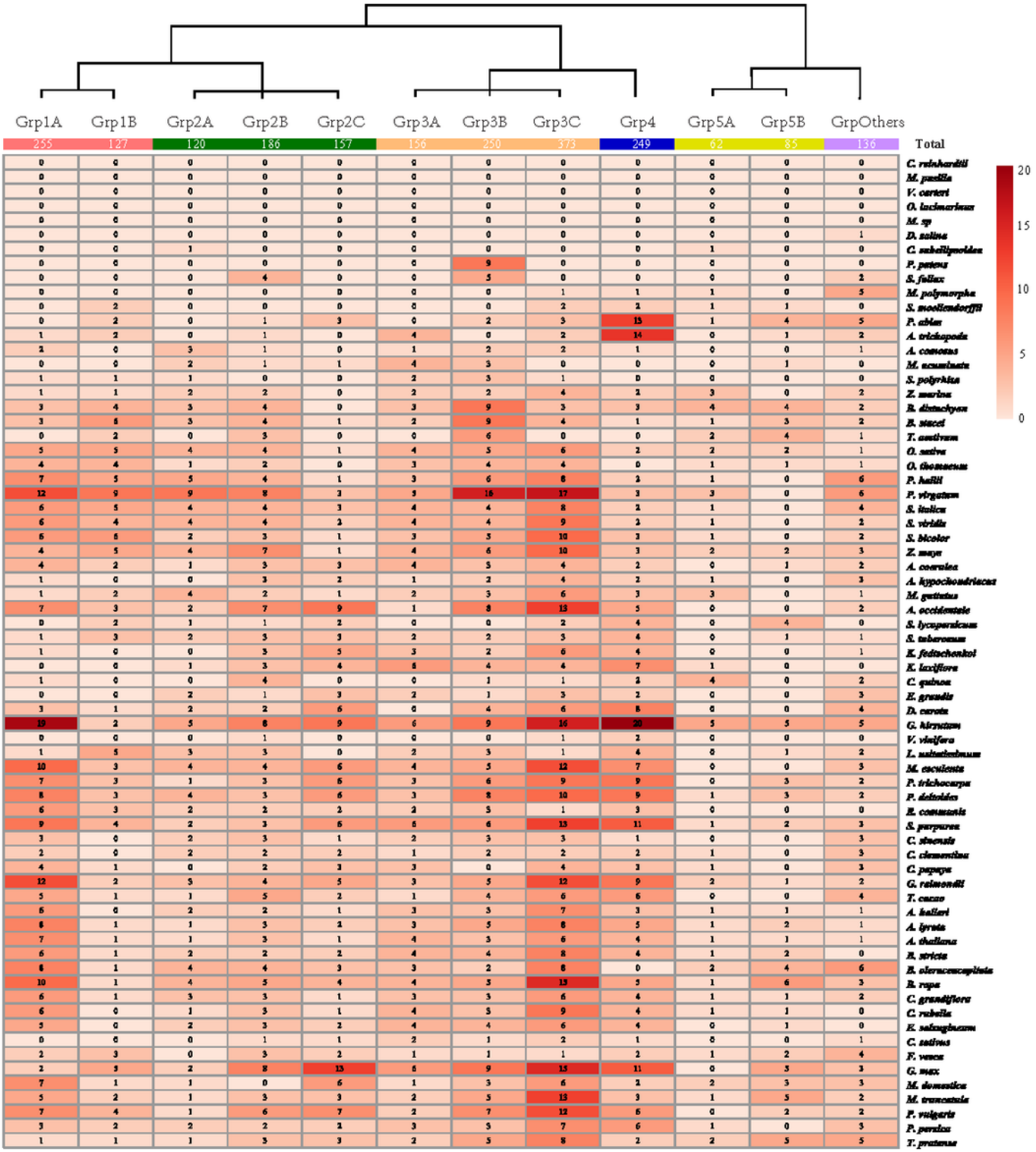

Figure 4

Evolution of CLE genes in plants. The number of CLE candidate genes from each group in each species was counted and indicated in the grid. The 2156 CLE candidates were from 12 groups and 69 species. The abundance of CLE candidates in each group is indicated with different shades of red. A darker shade of red indicates more group members. A lighter shade of red indicates fewer group members. The Latin name of each species is indicated on the right. The group name is indicated at the top of the grid. The

total number of CLE candidates in each subgroup is indicated in the appropriate box. The clustering tree on the top is a simplified version of the tree from Figure 3.

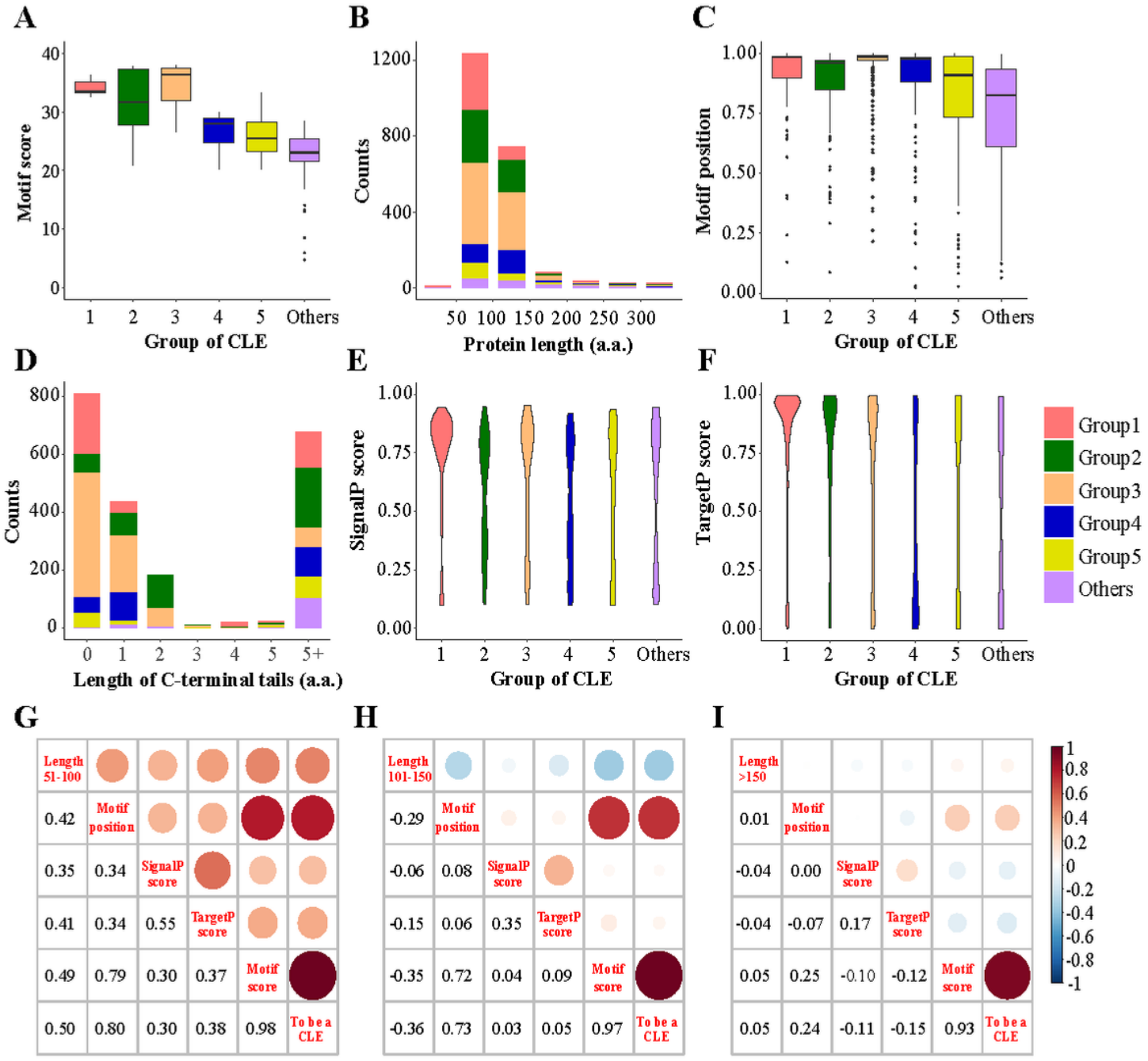

Figure 5

Statistical analysis of the major characteristics of CLE precursors in plants. The major characteristics of 2156 CLE precursor proteins were analyzed, including CLE motif scores (A), protein lengths (B), CLE motif positions (C), lengths of the C-terminal tails (D), and SignalP (E) and TargetP scores (F). Different groups are represented with different colors (A-F). Histogram: the height of the column represents the CLE candidate counts (B, D). The line in the box represents the median value. The upper and lower boundary of the box represents the upper and lower quartile values, respectively. The top and bottom of the line

represents the maximum and minimum value of non-outliers, respectively. The points represent outliers (A, C). The widths of the violins represent the distribution density of the indicated value. The tails of the violins were trimmed to match the range of the data (E, F). (G-I) Correlation between the different characteristics of each CLE candidate in three ranges of protein length: 51-100, 101-150 and >150 amino acid residues, respectively.

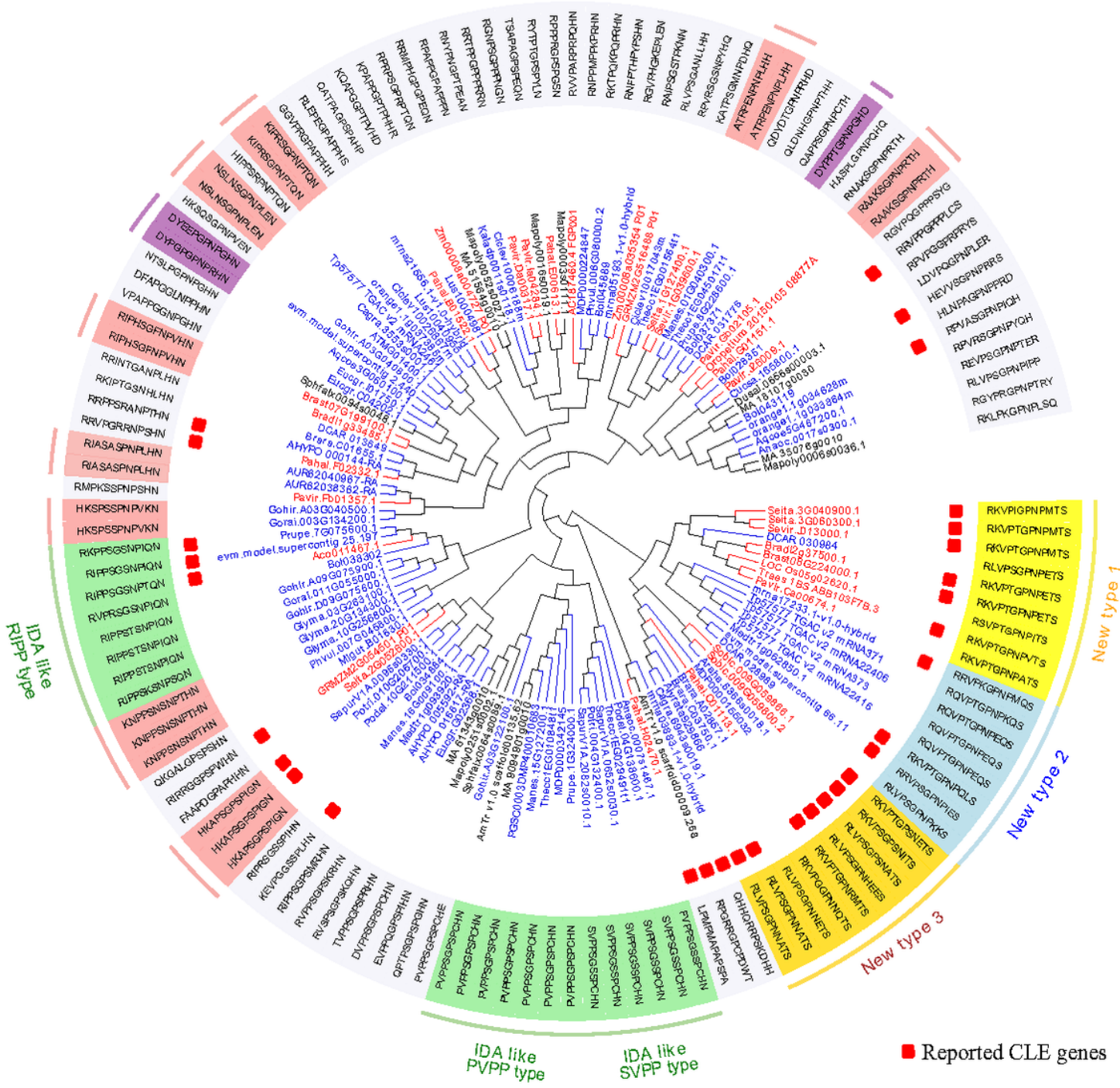

Figure 6

Identification of novel CLE candidates in Group “others”. From the inside to the outside of the ring diagram: clustering tree, gene ID, reporting status, motif sequences, and annotation. The Gene IDs

represented in red, blue and black indicate monocot, dicot and other plant species, respectively. Genes that have been reported are marked with red boxes. Candidate motifs of particular interest are highlighted with different colors. New types1, 2 and 3 are highlighted with yellow, light blue and gold, respectively. IDA-like CLE candidates are highlighted with light green. CLE candidates that appeared more than once in Group “others” are labeled with light red. CLE candidates starting with “DY” are indicated with purple.

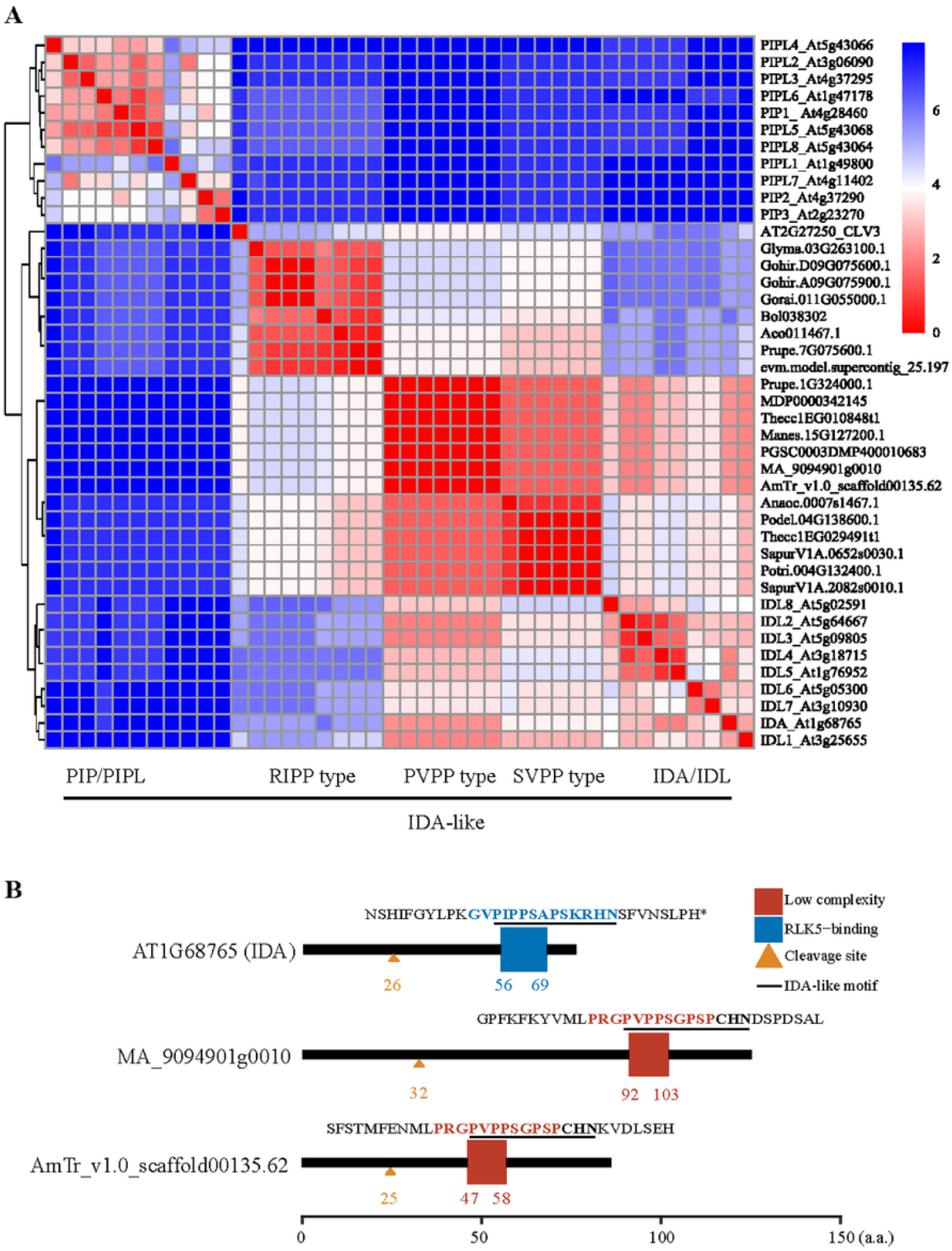

Figure 7

Clustering analysis of IDA-like CLE motifs and Arabidopsis IDA/IDL motifs. (A) Clustering of IDA-like CLE motifs and Arabidopsis IDA/IDL, PIP/PIPL and CLV3 motifs. The heat map indicates the Euclidean distance of each pair of motifs. Red represents short distances. Blue represents long distances. A shorter Euclidean distance implies a higher similarity. (B) Protein domain schematic diagram of Arabidopsis IDA and two “PVPP-type” IDA-like CLE candidates. Protein domains were predicted using SMART. Blue box: RLK5-binding domain; red-brown box: low complexity domain; pale-brown triangle: location of the cleavage site of the signal peptide for the secretory pathway; black underline: IDA or IDA-like motif.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.pdf
- Additionalfile17.xls
- Additionalfile3.pdf
- Additionalfile4.pdf
- Additionalfile14.xls
- Additionalfile13.xls
- Additionalfile18.xls
- Additionalfile8.pdf
- Additionalfile7.pdf
- Additionalfile16.xls
- Additionalfile5.pdf
- Additionalfile12.xls
- Additionalfile15.xls
- Additionalfile11.pdf
- Additionalfile20.xls
- Additionalfile6.pdf
- Additionalfile10.pdf
- Additionalfile9.pdf
- Additionalfile19.xls
- Additionalfile2.pdf