

Special issue: Microproteins

Review

Exploring the world of small proteins in plant biology and bioengineering

Louise Petri^{1,3}, Anne Van Humbeeck^{2,3}, Huanying Niu^{2,3}, Casper Ter Waarbeek^{2,3}, Ashleigh Edwards^{1,3}, Maurizio Junior Chiurazzi^{1,3}, Ylenia Vittozzi^{1,3}, and Stephan Wenkel^{1,3} ^{2,*}

Small proteins are ubiquitous in all kingdoms of life. MicroProteins, initially characterized as small proteins with protein interaction domains that enable them to interact with larger multidomain proteins, frequently modulate the function of these proteins. The study of these small proteins has contributed to a greater comprehension of protein regulation. In addition to sequence homology, sequence-divergent small proteins have the potential to function as microProtein mimics, binding to structurally related proteins. Moreover, a multitude of other small proteins encoded by short open reading frames (sORFs) and peptides, derived from diverse sources such as long noncoding RNAs (lncRNAs) and miRNAs, contribute to a variety of biological processes. The potential of small proteins is evident, offering promising avenues for bioengineering that could revolutionize crop performance and reduce reliance on agrochemicals in future agriculture.

MicroProteins, key regulators of plant and animal development

It is widely recognized that a greater number of ORFs are translated beyond those annotated in genome databases [1]. Historically underestimated due to their small size and the associated detection challenges, proteins encoded by **sORFs** (see [Glossary](#)) have recently garnered increasing attention for their roles as regulators of plant growth and development [2] as well as human health and disease [3]. **Microproteins** are encoded by sORFs and represent a diverse group of proteins, unified by their relatively small size (typically comprising approximately 100 amino acids or fewer). **MicroProteins**, one subclass of microproteins, are defined as small single-domain proteins that, similarly to miRNAs, exert a dominant-negative effect on their targets through protein–protein interactions ([Figure 1](#)). These microProteins function as post-translational regulators capable of modulating biological processes dependent on the formation of protein complexes, which may be homodimeric, heterodimeric, or multimeric [4–6]. The inhibitor of DNA binding (Id) protein was the first small protein to be identified that fit the microProtein definition. ID is a 16 kDa protein, which contains a helix–loop–helix (HLH) domain that enables it to interact with the basic (b)HLH transcriptional regulator MyoD and thereby control muscle differentiation [7].

Uncovering the complexity of sORFs as independent transcription units

The first plant microProteins to be identified were the LITTLE ZIPPER proteins, which comprise ZPR1, ZPR2, ZPR3, and ZPR4 in *Arabidopsis thaliana*. All four *ZPR* genes are encoded as individual transcription units within the *Arabidopsis* genome and thus have been classified as **trans-microProteins**. ZPR microProteins possess a leucine zipper domain and engage in homotypic interactions, regulating class III homeodomain-leucine zipper (HD-ZIPIII) transcription factors. They play a significant role in maintaining stem cells in the shoot apical

Highlights

Short open reading frames (sORFs) encoding microproteins, microProteins, and other small proteins are ubiquitous across all life forms. Gaining insight into their functions and bioengineering potential represents a major challenge.

A defining feature of microProteins is the presence of a conserved protein–protein interaction domain. This enables them to interact with a diverse range of proteins.

MicroProteins may be encoded by individual genes (*trans*-microproteins) or produced via alternative transcription (*cis*-microproteins).

Small proteins produced from long non-coding RNAs or miRNAs, including sORF-encoded peptides and miRNA-encoded peptides, have been identified as potent developmental regulators.

Advances in protein prediction have the potential to enhance the identification and application of these molecules in synthetic biology and biotechnology.

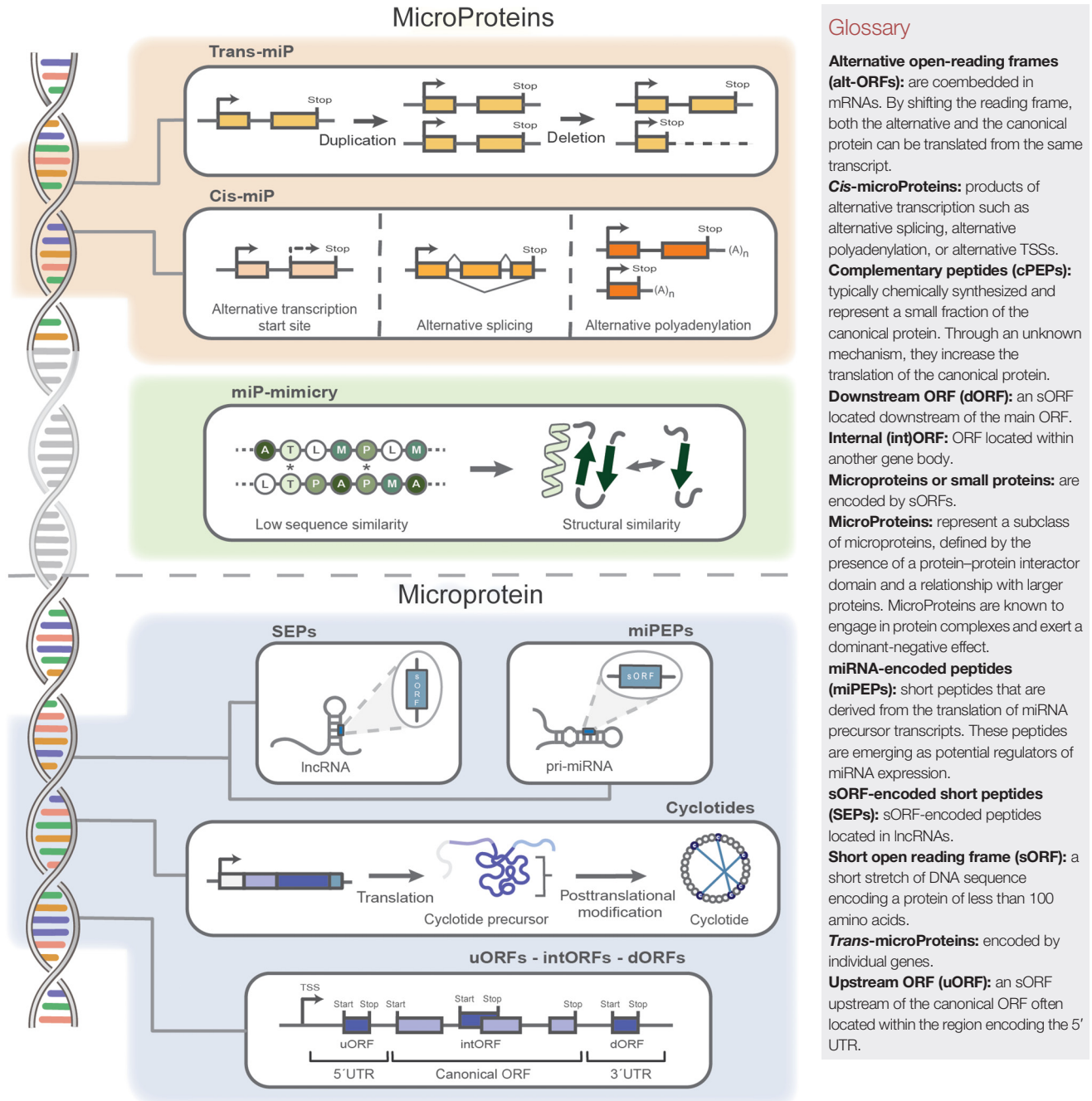
¹Department of Plant and Environmental Sciences, University of Copenhagen, Thorvaldsensvej 40, 1871 Frederiksberg, Denmark

²Umeå Plant Science Centre, Department of Plant Physiology, Umeå University, Linnaeus väg 6, 90736 Umeå, Sweden

³These authors contributed equally

*Correspondence: stephan.wenkel@umu.se (S. Wenkel).





Glossary

Alternative open-reading frames (alt-ORFs): are coembedded in mRNAs. By shifting the reading frame, both the alternative and the canonical protein can be translated from the same transcript.

Cis-microProteins: products of alternative transcription such as alternative splicing, alternative polyadenylation, or alternative TSSs.

Complementary peptides (cPEPs): typically chemically synthesized and represent a small fraction of the canonical protein. Through an unknown mechanism, they increase the translation of the canonical protein.

Downstream ORF (dORF): an sORF located downstream of the main ORF.

Internal (int)ORF: ORF located within another gene body.

Microproteins or small proteins: are encoded by sORFs.

MicroProteins: represent a subclass of microproteins, defined by the presence of a protein-protein interactor domain and a relationship with larger proteins. MicroProteins are known to engage in protein complexes and exert a dominant-negative effect.

miRNA-encoded peptides

(miPEPs): short peptides that are derived from the translation of miRNA precursor transcripts. These peptides are emerging as potential regulators of miRNA expression.

sORF-encoded short peptides

(SEPs): sORF-encoded peptides located in lncRNAs.

Short open reading frame (sORF): a short stretch of DNA sequence encoding a protein of less than 100 amino acids.

Trans-microProteins: encoded by individual genes.

Upstream ORF (uORF): an sORF upstream of the canonical ORF often located within the region encoding the 5' UTR.

Figure 1. Overview presenting the different types of small proteins. MicroProteins (miPs) regulate larger proteins through shared protein-protein interaction domains. In the case of sequence-dependent miPs, some exist as individual genes (*trans*-miPs), whereas others are generated by alternative transcription, splicing, or polyadenylation (*cis*-miPs). Furthermore, the interaction between a miP and its target can also be solely dependent on structural similarities (miP mimicry). Other types of microproteins originate from noncoding RNAs containing short open reading frames (sORFs). The specific functions of peptides resulting from long noncoding RNAs (lncRNAs) remain to be elucidated. miRNA-encoded peptides (miPEPs) are derived from primary miRNAs (pri-miRNAs) and are postulated to function by increasing transcription of their own pri-miRNA transcript, thereby establishing a positive feedback mechanism. Cyclotides are

(Figure legend continued at the bottom of the next page.)

meristem and influencing leaf development through interaction with the HD-ZIP IIIs. This interaction establishes a negative feedback loop [8,9].

Id-like microProteins that contain solely an HLH domain and interact with bHLH transcription factors have also been identified in plants, although to their animal counterparts are of independent evolutionary origin. Plant Id-like microProteins have been shown to regulate plant responses to environmental stimuli, particularly light. As sessile organisms, plants depend on light for energy production. However, they frequently encounter obstacles in accessing optimal light conditions due to neighboring vegetation [10]. Transcription factors such as PHYTOCHROME INTERACTING FACTORS (PIFs) that are bHLHs facilitate shade-avoidance responses, whereas LONG HYPOCOTYL IN FAR-RED 1 (HFR1), an atypical bHLH, suppresses excessive shade responses by sequestering PIFs [11]. In addition, HFR1 interacts with the microProtein KIDARI (KDR), an Id-like HLH protein, which modulates HFR1 activity [12]. Thus, PIFs, HFR1, and KDR form a regulatory triangle [13]. Other Id-like proteins have been identified that interact with bHLH transcription factors to regulate developmental programs and responses to hormones such as brassinosteroids (BRs) [14,15]. In arabidopsis, the PACLOBUTRAZOL-RESISTANT (PRE) family of microProteins has been shown to suppress the dwarf phenotype associated with BR insensitivity when overexpressed. Similarly, in rice, overexpression of INCREASED LEAF ANGLE INCLINATION1 (ILI1), a PRE1 homolog, results in phenotypes that resemble those observed in plants exposed to high levels of BR. ILI1 interacts with ILI1 BINDING BHLH 1 (IBH1), a bHLH factor suppressed by BR, indicating the existence of a regulatory module involving HLH/bHLH dimers in hormone signaling across plant species [14,15]. This module is analogous to the Id/MyoD regulatory module in animals [7].

MicroProteins, BBX31 and BBX30, also known as miP1a and miP1b, are microProteins containing a B-Box zinc finger domain to interact with their targets. The CONSTANS (CO) transcription factor, a critical regulator of flowering time, is among the miP1a/b targets [5]. CO binds to the promoter of *FLOWERING LOCUS T (FT)*, also known as florigen, activating its expression [16] and thereby inducing flowering in response to photoperiod. Additionally, miP1a and miP1b have been found to interact with the TOPLESS/TOPLESS-RELATED (TPL/TPR) corepressor proteins through their carboxy-terminal PFVFL motif, forming a trimeric complex with TPL and CO that acts to delay flowering [5,17].

Microproteins can also act by altering the subcellular localization pattern of their targets. For example, when transiently co-expressed with ZINC-FINGER HOMEODOMAIN 5 (ZHD5), MINI FINGER 1 (MIF1), a small zinc finger *trans*-microProtein, prevents nuclear import of ZHD5 to retain it in the cytoplasm [18]. This example demonstrates that *trans*-microProteins are involved in a multitude of biological processes. Not only do they regulate transcription factors by engaging them in biologically inactive protein complexes or protein complexes with altered biological function, but they can also sequester transcription factors in the cytoplasm and potentially other subcellular compartments.

sORFs hiding within other genes

Recent evidence suggests that the genome is pervasively translated [19,20]. **Upstream ORFs (uORFs)** in the 5' untranslated region (UTR) of mRNAs are found in 20–30% of eukaryotic

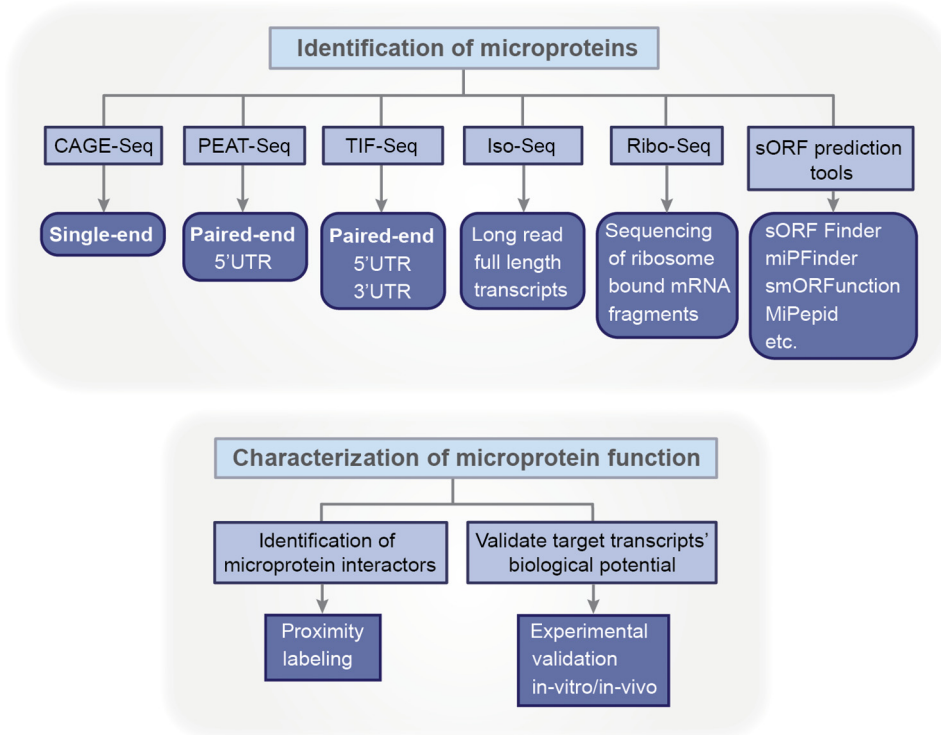
generated by cleavage of protein precursors, and it is likely that the future will reveal the vast diversity of microproteins. Furthermore, sORFs have been identified in the upstream and downstream gene untranslated regions (UTRs; uORFs and dORFs, respectively) as well as within the canonical coding sequence [internal (intORFs)].

transcripts, including plants [21]. Although the mechanism has not been fully elucidated, uORFs are believed to influence the translation of the downstream main ORF (mORF), potentially by inducing ribosome stalling [22]. In addition, a recent ribosome profiling study identified 209 ORFs in the 3' UTR of annotated genes [23]. These **downstream ORFs (dORFs)** were enriched in genes that are highly expressed, although it remains to be seen whether they directly influence the translation of their upstream mORF [23]. Furthermore, the study by Wu *et al.* identified a considerable number of sORFs and uORFs, several of which are involved in processes such as translational repression and splicing. By comparing ribosome profiling data with public CAGE sequencing data, which analyzes transcription start sites (TSSs), the authors identified a subset of dORFs and **internal ORFs (intORFs)** that are likely the result of independent transcription events (Figure 1) [23]. In tomato (*Solanum lycopersicum*), ribosome profiling using lactimidomycin and puromycin to stall ribosomes at translation initiation sites (TISs) and depletes elongation revealed that 10–15% of TISs occur within annotated coding sequences [24]. Moreover, at least 20% of identified TISs were located at near-cognate start codons, such as CUG, GUG, or UUG, which are often overlooked by gene prediction models [24]. Demonstrating the stability of these noncanonical translation products will be crucial to discern those that function at the level of translation, through regulating expression of the mORF, for example, from those encoding function post-translationally, like microProteins. Wu *et al.* (2022) employed mass spectrometry-based proteomics to validate translation from some of the noncanonical ORFs they discovered [23]. However, current proteomic methods tend to favor larger and more abundant proteins; many small proteins lack tryptic peptides and thus remain undetected, and even when alternative digestion methods are employed, the low abundance of small proteins typically hinders their detection [25]. Other RNA-sequencing approaches can provide a starting point for the investigation of microProteins and other small proteins. In a recent publication, 5' paired end analysis of transcription start site sequencing (5'PEAT-Seq) was employed to identify novel **cis-microProtein** candidates in arabidopsis. This approach revealed 377 *cis*-microProtein candidate transcripts, hypothesized to contain a single protein–protein interaction domain upon translation [26]. Further characterization of one such microProtein candidate, ATHB2miP, showed it comprises a single leucine zipper domain and mimics the phenotype of an *athb2* knockout when overexpressed. Gene expression analysis in plants either lacking ATHB2 or overexpressing *ATHB2miP* revealed common misregulation patterns, supporting that ATHB2miP functions like a canonical microProtein [6,26].

Experimentally differentiating their function from the larger protein isoform with which they share a coding sequence poses a significant challenge when studying *cis*-microProteins. Mutations in the *cis*-microProtein coding sequence will inevitably impact the larger protein isoform. Fortunately, valuable insights can be gained from the field of RNA biology. For instance, the translation potential of sORFs encoded by the primary miRNA (pri-miRNA), pri-miR171b, was validated *in planta* by fusion of the start codon and upstream promoter sequence to the β -glucuronidase reporter gene [27]. In wheat, a study that identified the lncRNA VAS used mutations affecting the start codons of potential ORFs in the noncoding transcript as a method to disprove their significance and confirm the function of the noncoding RNA component over the ORF [28]. This same approach could be applied to validate the coding potential of *cis*-microProtein candidates, with minimal impact to the larger isoform, to explore their effects on plant physiology (Figure 2).

MicroProtein mimicry: different and yet the same

Protein–protein interaction studies have traditionally relied on sequence similarity to identify potential interacting partners. This approach has also been common in the microProtein field, where most microProteins have been discovered this way [29–32]. However, the focus on sequence similarity may overlook proteins that are not sequence conserved, potentially creating a blind spot in our understanding of protein interactions (Figure 1).



Trends in Genetics

Figure 2. Methods that can be employed to investigate microproteins. A plethora of RNA-sequencing technologies are available for the purpose of identifying novel microprotein transcripts. Once a candidate microprotein has been identified, its existence on the protein level must be validated. This can be achieved by combining short open reading frame (sORF) prediction and ribosome profiling (Ribo-Seq). Once a validated candidate has been identified, its function can be explored by proteomic approaches and confirmed experimentally. This may entail investigating the interaction partners and the molecular function of the candidate. Further details on each technology are provided in [Table 1](#) in the main text. Abbreviations: CAGE-Seq, cap analysis gene expression sequencing; Iso-Seq, isoform sequencing; PEAT-Seq, paired end analysis of transcription start site sequencing; TIF-Seq, transcript isoform sequencing; UTR, untranslated region.

Advancements in protein structure prediction over the past decade, driven by developments in sequence technology and bioinformatics (see [Table 1](#)), have demonstrated that sequence similarity is not essential for protein–protein interactions or functional similarities. The discovery of heterotypic microProtein interactions is an example of this shift in understanding. Research into structural evolution suggests that structural cores evolve significantly slower than sequences, typically three to ten times slower [33]. It has been shown that non-homologous and even unrelated proteins can share conserved 3D structures, such as DNA-binding domains, including the histone fold, the helix–turn–helix motif, and the zinc finger [34].

The assumption that protein–protein interaction domains are more sequence conserved than other regions of the protein has been debated. Although multiple sequence alignments are commonly used to indicate the conservation of amino acid residues at protein interaction interfaces, this phenomenon is not observed when novel surface patch definition methods are employed [35]. It seems therefore plausible that some small proteins may not show sequence conservation to larger proteins but might contain a structural fold that facilitates protein–protein interactions. One intriguing example from humans is the pTINCR microprotein [36], which is upregulated during differentiation and cellular stress in epithelial tissues. Despite not sharing high sequence similarity, pTINCR is structurally similar to ubiquitin and can interact with

Table 1. Overview of different state-of-the-art technologies available for identification of noncanonical gene products, including advantages and disadvantages of different technologies, estimated price, and examples of their use^a

Goal	Name	Technology	Advantage	Disadvantages	Price	Refs
Discovery of short transcripts	CAGE-Seq	Illumina single-end sequencing of capped, reverse-transcribed, and restriction site-labeled, purified, digested mRNA	Precise mapping of TSSs compared with older technologies	Not appropriate for short transcripts due to single-end sequencing Bias toward long-lived transcripts Needs experimental validation to discriminate between coding and noncoding transcripts	\$	[69]
	PEAT-Seq	Illumina paired-end sequencing of capped, restriction-digested, reverse-transcribed mRNA	Higher mapping accuracy of short and overlapping transcripts	Short reads. Assembly is required, and it can lead to loss of small transcripts Only 5' UTR is sequenced Needs experimental validation to discriminate between coding and noncoding transcripts	\$\$	[26,70,71]
	TIF-Seq	Illumina paired-end sequencing of capped and uncapped, reverse-transcribed, barcoded (at both 5' and 3' ends) mRNA	Sequencing of both the 5' UTR and 3' UTR regions Bias toward short transcripts	Short reads. Assembly is required, and it can lead to loss of small transcripts Needs experimental validation to discriminate between coding and noncoding transcripts	\$\$\$	[72,73]
	Iso-Seq	PacBio HiFi long-reads	Full-length short and long transcripts No assembly; no information is lost	High sequence error rate. It is usually coupled with Illumina reads/RNA-Seq for sequence confirmation Needs experimental validation to discriminate between coding and noncoding transcripts	\$\$\$\$	[74,75]
Discovery of short coding transcripts	Ribo-Seq	Deep sequencing of RNase-protected, ribosome-bound mRNA fragments	Identification of actively translated transcripts	Coupling with RNA-Seq from the same sample is required to filter out the UTRs and correctly map the ORFs It provides a snapshot of the translation machinery at a specific point in time	\$\$\$\$\$	[76–78]
Discovery of sORFs	sORFs prediction tools	Bioinformatics and machine learning (ML) tools for the identification of sORFs	Usually publicly available and free	To be coupled with sequencing data. Tools must be chosen and adapted to the sequencing data available Tools may require appropriate expertise		[29,31,32]
Discovery of microproteins and peptides	Proximity labeling and proteomics	Engineered peroxidases and ligases in the proteomics experimental pipeline	Proximity labeling allows the direct identification of small proteins and peptides and their specific interactors	In order to identify small proteins, all proteomic experimental advancements still rely on the curation of both the sORF annotation data and the spectrum data generated in the proteomics pipeline	\$\$\$\$\$	[59,79,80]

^aAbbreviations: CAGE-Seq, cap analysis gene expression sequencing; Iso-Seq, isoform sequencing; PEAT-Seq, paired end analysis of transcription start site sequencing; TIF-Seq, transcript isoform sequencing.

SUMO through its SUMO-interacting motif. It influences CDC42 SUMOylation and plays a role in epithelial cancer formation and tumor suppression [36].

Given these challenges, structure-based approaches to identifying microProteins are needed, necessitating accurate structure prediction methods such as AlphaFold [37], which can reveal structural similarities absent in sequence data. Despite tools such as BLAST and HHBLITS failing

to identify sequence similarities between certain proteins, structure prediction has shown significant structural similarities between them. Effectively implementing a structure-based search requires advanced tools such as FoldSeek, which offers improved sensitivity and speed compared with older tools such as Dali and TM-align [38]. This enables faster, more efficient searches for potential microProteins (Figure 2). Yet, challenges remain, especially with intrinsically disordered small proteins.

Bringing order to chaos: microproteins and other small proteins

In addition to microProteins, which interact with their targets through protein–protein interaction domains by sequence similarity or structural mimicry, a diverse array of other functional small proteins, here referred to as ‘microproteins,’ also exist. These microproteins originate from sORFs and do not necessarily function by interacting with larger proteins (Figure 1). sORF-encoded microproteins contain typically fewer than 100 amino acids, although much smaller proteins or even peptides are frequently observed. They have been found profusely, but not exclusively, in noncoding RNA transcripts including lncRNAs and pri-miRNAs. These transcripts include **sORF-encoded short peptides (SEPs)** and **miRNA-encoded peptides (miPEPs)** [27,39].

Until the early 2000s, lncRNAs were defined as RNA transcripts exceeding 200 nucleotides that lacked protein-coding capabilities. They typically functioned through mechanisms such as antisense interference or scaffolding [40]. However, the discovery of the lncRNA *EARLY NODULIN 40 (ENOD40)*, which harbors two sORFs, challenged this definition within the plant kingdom. Initially classified as an mRNA, *ENOD40* is now recognized as a lncRNA with protein-coding regions that give rise to peptides, called SEPs. *ENOD40* primarily facilitates cell division in cortical cells during nodule formation in legumes [41]. It was shown that the sORFs in *ENOD40* are essential for fully inducing nodule formation in *Medicago truncatula* [42]. Further studies have shown that the RNA structures predicted between the sORFs are also crucial for their functionality, suggesting that *ENOD40* works on both RNA and protein levels [42,43]. Despite these insights, the broader potential of noncoding RNAs containing sORFs was not fully appreciated until recently, when the use of ribosome profiling demonstrated that numerous lncRNAs were associated with ribosomes in a manner similar to protein-coding RNAs [44–48].

A considerable number of bioinformatic endeavors have been undertaken with the objective of predicting and confirming the presence of sORFs within lncRNAs in a variety of plants. For instance, the translation of over 100 sORFs from lncRNAs was demonstrated by comparing mass spectrometry data on total protein from soybean with the predicted sequences of sORFs embedded within lncRNA transcripts [49]. New methods are being developed continuously, including elements of machine learning [50,51]. These studies, although providing insight into the prevalence of sORFs, rarely provide concrete examples of their biological activity, leaving the specific functions and, consequently, potential applications of sORFs largely unexplored.

miPEPs are sORF-encoded peptides hidden within pri-miRNA transcripts (Figure 1). These transcripts are known to produce 21–23-nucleotide-long miRNAs that regulate the level of their target mRNA through various mechanisms, including directed cleavage by AGO1, leading to mRNA degradation [52]. In *M. truncatula*, pri-miR171b encodes both *miR171b* and miPEP171b [27]. Exogenously administered miPEP171b was shown to enhance the levels of *pri-miR171b*, implying a positive feedback mechanism whereby the miPEP influences the transcript level of its own pri-miRNA. On the basis of experiments in arabidopsis, the increase in transcript level of pri-miRNAs is suggested to be a result of altered transcription in contrast to transcript stability [27]. The precise mechanism by which this process occurs is still not fully understood. However, key features of the interaction between miPEPs and pri-miRNA transcripts have been elucidated.

These include the interaction between the two moieties, which may occur in a sequence-dependent manner, as the sequence of the pri-miRNA has been shown to be essential for miPEP activity [53]. Nevertheless, while miPEPs are identified in a variety of species, the degree of sequence conservation between species is relatively low, with a high degree of specificity to the species of origin. This suggests that miPEPs may have a universal yet specialized function across different organisms [27,53,54].

It should be noted that not all microproteins originate from sORFs. Cyclotides (Figure 1), for instance, are a notable exception, deriving from larger precursor proteins from which they are enzymatically cleaved to produce a distinct structure of approximately 30 amino acids [55]. This structure is characterized by a cyclized backbone and three disulfide bonds formed by six cysteines, collectively known as a cyclic cysteine knot (CKK). This unique structural configuration grants cyclotides remarkable thermal and enzymatic stability [56]. Beyond their natural biological functions, cyclotides exhibit substantial potential as pharmaceutical agents [57].

Finally, evidence is beginning to emerge that microproteins can also be produced by means of being coembedded within the same gene but translated from the corresponding mRNA in a distinct reading frame. Such **alternative open reading frames (alt-ORFs)** encode proteins that are markedly different in sequence and length from their canonical counterparts. The alteration of the translation frame of mRNAs has the potential to yield numerous microproteins that are not yet annotated in protein databases. In the animal field, alternative proteins have been identified by bio-orthogonal noncanonical amino acid tagging [58] and affinity labeling followed by mass spectrometry [59]. The latter approach may be employed to enrich specific subcellular compartments for the purpose of identifying alt-proteins. Similarly, affinity purification followed by mass spectrometry may also be employed to identify alt-proteins that are part of specific protein complexes. The identification and characterization of alt-proteins is still in its infancy; however, some may act as important regulators of cellular processes.

Molecular LEGO: using small proteins in bioengineering approaches

As previously stated, ZPR microProteins, which are distinguished by a single leucine zipper (ZIP) domain, have been demonstrated to interact with HD-ZIPIII transcription factors via the ZIP domain [8,9]. The identification of the ZPR proteins and their ability to interfere with protein function at the post-translational level has led to the development of synthetic microProtein techniques, which aim to target specific protein domains through the synthesis of microProteins.

The ectopic expression of individual protein–protein interaction domains of multidomain proteins can impact plant development. Transgenic plants overproducing respective synthetic microProteins exhibited phenotypes similar to those observed in loss-of-function mutants. In a study, three well-studied proteins, each playing a crucial role in a distinct pathway, were targeted: DICER-LIKE1 (DCL1), BRASSINOSTEROID INSENSITIVE 1 (BRI1), and CRYPTOCHROME1 (CRY1) [60]. Overexpression of a synthetic microProtein composed solely of the PAZ domain of DCL1 in arabidopsis induced multiple developmental abnormalities, including smaller rosettes and abnormal cotyledons, similar to those observed in *dcl1* mutants. A synthetic microProtein derived from the transmembrane-juxtamembrane domain of BRI1 produced a severe dwarf phenotype in arabidopsis comparable with that observed in *bri1* mutant plants. Targeting the PH subdomain of CRY1 with a synthetic microProtein resulted in elongated hypocotyls when grown under blue light conditions, mirroring the phenotype observed in *cry1* mutants [60]. This approach has also proved successful in rice with Hd1miP, which comprises the B-BOX domain of HEADING DATE 1 (Hd1) and significantly alters flowering time by interacting with Hd1 [61].

A strategy was proposed with the objective of enhancing the use of the synthetic microProtein approach for the purpose of controlling potential multidomain proteins. This strategy entailed the integration of computational and genome-engineering approaches. First, the strategy involved the screening of microProteins that are present in monocotyledonous plants but absent in dicotyledonous plants. Second, it was investigated whether the dicotyledonous model plant (here, arabidopsis) possessed the ancestral, larger proteins that are potentially subject to microProtein control in the monocotyledonous system [62,63]. Finally, the use of synthetic microProteins in the dicot model system enabled the screening for biological activity, and the genome engineering of *de novo* microProteins allowed specific biological processes to be addressed [62].

Another potential application involves spraying of synthetic peptides on plants to influence agronomical traits of interest such as fruit ripening, root growth, and disease resistance. A peptide corresponding to the nuclear localization signal motif of ETHYLENE INSENSITIVE 2 was synthesized and named NUCLEAR LOCALIZATION SIGNAL OCTAPEPTIDE 1 (NOP-1). NOP-1 competes with EIN2 for binding to the ethylene receptor protein ETHYLENE RESPONSE1 (ETR1) and also reduces ethylene-induced growth responses in dark-grown arabidopsis seedlings [64]. Moreover, surface treatment of developing green tomato fruits with NOP-1 significantly delayed fruit ripening [65].

By analogy to miPEPs, it was recently shown that the exogenous application of **complementary peptides (cPEPs)** can activate the translation of target proteins [66]. Here, the cPEP is a peptide that represents a fraction of the protein that it regulates and through an unknown mechanism boosts the translation. It is imaginable that combined treatment with different cPEPs can affect multiple traits simultaneously.

The few examples presented here demonstrate the significant potential of microProteins and other small proteins to modify traits of interest through microprotein engineering. The diversity of small proteins allows the development of tailored approaches. The concurrent application of orthogonal strategies in conjunction with genome-engineering techniques is likely to represent a significant turning point in the years to come.

Concluding remarks

Small proteins are pervasive in biological systems [67]. Many of them play crucial roles in growth and development, as well as in cross-kingdom communication and disease. Their small size enables secretion and long-distance transport; yet, it also presents a challenge in their identification due to technical limitations (see [Outstanding questions](#)). Recent advances in deep sequencing and our ability to identify mRNA isoforms and translation events have led to the identification of thousands of novel small proteins. The potential for generating microproteins through proteolytic processes has only recently started to be explored [68]. The study concludes that at least 1% of all proteins produce shorter proteoforms. Nevertheless, this could represent merely the tip of the iceberg, because the present study does not account for the spatial and temporal expression patterns of all proteins. It seems reasonable to assume that there are a greater number of cell type and stage-specific alternative proteoforms that have yet to be identified. Many of these proteoforms and alternative proteins can also have important functions and may be useful in future bioengineering approaches.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT4.0 and DeepL in order to standardize the writing style and to eliminate any instances of grammatical error. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Outstanding questions

Several small proteins display cell type-specific expression patterns. The question thus arises regarding how these patterns are established.

It is unclear whether the majority of small proteins act in a cell-autonomous or non-cell-autonomous manner. What strategies can be employed to prevent the movement of small proteins between cells? Does evidence exist for the active shuttling of small proteins between cells?

What are the mechanisms regulating small proteins at the protein level? Does the small size confer increased stability due to the absence of proteolysis recognition motifs? What other post-translational modifications affect the activity of small proteins?

miPEPs and cPEPs act by recognizing and binding to their cognate coding sequences, thereby enhancing their own expression. What are the mechanisms underlying the interactions of cPEPs/miPEPs with their corresponding coding regions? It is of interest to determine the mechanisms by which miPEPs/cPEPs affect transcription and translation.

Acknowledgments

We acknowledge funding through the Novo Nordisk Foundation (grants 2019OC53580, NNF18OC0034226, and NNF20OC0061440 to S.W.) and the Independent Research Fund Denmark (0136- 00015B and 0135-00014B) to S.W.

Declaration of interests

The authors have no conflicts of interest to declare.

References

- Fesenko, I. *et al.* (2019) Distinct types of short open reading frames are translated in plant cells. *Genome Res.* 29, 1464–1477
- Kushwaha, A.K. *et al.* (2022) Plant microProteins: small but powerful modulators of plant development. *iScience* 25, 105400
- Rathore, A. *et al.* (2018) Small, but mighty? Searching for human microproteins and their potential for understanding health and disease. *Expert Rev. Proteomics* 15, 963–965
- Eguen, T. *et al.* (2015) MicroProteins: small size-big impact. *Trends Plant Sci.* 20, 477–482
- Graeff, M. *et al.* (2016) MicroProtein-mediated recruitment of CONSTANS into a TOPLESS trimeric complex represses flowering in Arabidopsis. *PLoS Genet.* 12, e1005959
- Staudt, A.-C. and Wenkel, S. (2011) Regulation of protein function by microProteins. *EMBO Rep.* 12, 35–42
- Benezra, R. *et al.* (1990) The protein Id: a negative regulator of helix-loop-helix DNA binding proteins. *Cell* 61, 49–59
- Wenkel, S. *et al.* (2007) A feedback regulatory module formed by LITTLE ZIPPER and HD-ZIP III genes. *Plant Cell* 19, 3379–3390
- Kim, Y.-S. *et al.* (2008) HD-ZIP III activity is modulated by competitive inhibitors via a feedback loop in Arabidopsis shoot apical meristem development. *Plant Cell* 20, 920–933
- Martinez-Garcia, J.F. *et al.* (2014) The shade avoidance syndrome in Arabidopsis: the antagonistic role of phytochrome a and B differentiates vegetation proximity and canopy shade. *PLoS One* 9, e109275
- Hornitschek, P. *et al.* (2009) Inhibition of the shade avoidance response by formation of non-DNA binding bHLH heterodimers. *EMBO J.* 28, 3893–3902
- Hyun, Y. and Lee, I. (2006) KIDARI, encoding a non-DNA binding bHLH protein, represses light signal transduction in *Arabidopsis thaliana*. *Plant Mol. Biol.* 61, 283–296
- Galstyan, A. *et al.* (2011) The shade avoidance syndrome in Arabidopsis: a fundamental role for atypical basic helix-loop-helix proteins as transcriptional cofactors. *Plant J.* 66, 258–267
- Wang, H. *et al.* (2009) Regulation of Arabidopsis brassinosteroid signaling by atypical basic helix-loop-helix proteins. *Plant Cell* 21, 3781–3791
- Zhang, L.-Y. *et al.* (2009) Antagonistic HLH/bHLH transcription factors mediate brassinosteroid regulation of cell elongation and plant development in rice and Arabidopsis. *Plant Cell* 21, 3767–3780
- Samach, A. *et al.* (2000) Distinct roles of CONSTANS target genes in reproductive development of Arabidopsis. *Science* 288, 1613–1616
- Rodriguez, V.L. *et al.* (2021) A microProtein repressor complex in the shoot meristem controls the transition to flowering. *Plant Physiol.* 187, 187–202. <https://doi.org/10.1093/plphys/kiab235>
- Hong, S.-Y. *et al.* (2011) Nuclear import and DNA binding of the ZHD5 transcription factor is modulated by a competitive peptide inhibitor in Arabidopsis. *J. Biol. Chem.* 286, 1659–1668
- Chen, J. *et al.* (2020) Pervasive functional translation of non-canonical human open reading frames. *Science* 367, 1140–1146
- Sruthi, K.B. *et al.* (2022) Pervasive translation of small open reading frames in plant long non-coding RNAs. *Front. Plant Sci.* 13, 975938
- Kawaguchi, R. and Bailey-Serres, J. (2005) mRNA sequence features that contribute to translational regulation in Arabidopsis. *Nucleic Acids Res.* 33, 955–965
- Hiragori, Y. *et al.* (2023) Genome-wide identification of Arabidopsis non-AUG-initiated upstream ORFs with evolutionarily conserved regulatory sequences that control protein expression levels. *Plant Mol. Biol.* 111, 37–55
- Wu, H.L. *et al.* (2024) Improved super-resolution ribosome profiling reveals prevalent translation of upstream ORFs and small ORFs in Arabidopsis. *Plant Cell* 36, 510–539
- Li, Y.R. and Liu, M.J. (2020) Prevalence of alternative AUG and non-AUG translation initiators and their regulatory effects across plants. *Genome Res.* 30, 1418–1433
- Wacholder, A. and Carvunis, A.R. (2023) Biological factors and statistical limitations prevent detection of most noncanonical proteins by mass spectrometry. *PLoS Biol.* 21, e3002409
- Edwards, A. *et al.* (2024) A shade-responsive microProtein in the Arabidopsis *ATHB2* gene regulates elongation growth and root development. *bioRxiv*, Published online February 15, 2024. <https://doi.org/10.1101/2024.02.01.578400>
- Laressergues, D. *et al.* (2015) Primary transcripts of microRNAs encode regulatory peptides. *Nature* 520, 90–93
- Xu, S. *et al.* (2021) The vernalization-induced long non-coding RNA VAS functions with the transcription factor TaRF2b to promote TaVRN1 expression for flowering in hexaploid wheat. *Mol. Plant* 14, 1525–1538
- Hanada, K. *et al.* (2010) sORF finder: a program package to identify small open reading frames with high coding potential. *Bioinformatics* 26, 399–400
- Ji, X. *et al.* (2020) smORFunction: a tool for predicting functions of small open reading frames and microproteins. *BMC Bioinforma* 21, 455
- Leong, A.Z. *et al.* (2022) Short open reading frames (sORFs) and microproteins: an update on their identification and validation measures. *J. Biomed. Sci.* 29, 19
- Zhu, M. and Gribskov, M. (2019) MiPepid: MicroPeptide identification tool using machine learning. *BMC Bioinforma.* 20, 559
- Illegård, K. *et al.* (2009) Structure is three to ten times more conserved than sequence – a study of structural response in protein cores. *Proteins* 77, 499–508
- Sousounis, K. *et al.* (2012) Conservation of the three-dimensional structure in non-homologous or unrelated proteins. *Hum. Genomics* 6, 10
- Caffrey, D.R. *et al.* (2004) Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci.* 13, 190–202
- Boix, O. *et al.* (2022) pTINCR microprotein promotes epithelial differentiation and suppresses tumor growth through CDC42 SUMOylation and activation. *Nat. Commun.* 13, 6840
- Jumper, J. *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589
- van Kempen, M. *et al.* (2024) Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* 42, 243–246
- Sharma, A. *et al.* (2020) Primary transcript of miR858 encodes regulatory peptide and controls flavonoid biosynthesis and development in Arabidopsis. *Nat. Plants* 6, 1262–1274
- Ulveling, D. *et al.* (2011) When one is better than two: RNA with dual functions. *Biochimie* 93, 633–644
- Ganguly, P. *et al.* (2021) The natural antisense transcript DONE40 derived from the lncRNA ENOD40 locus interacts with SET domain protein ASHR3 during inception of symbiosis in *Arachis hypogaea*. *Mol. Plant-Microbe Interact.* 34, 1057–1070
- Sousa, C. *et al.* (2001) Translational and structural requirements of the early nodulin gene enod40, a short-open reading frame-containing RNA, for elicitation of a cell-specific growth response in the alfalfa root cortex. *Mol. Cell. Biol.* 21, 354–366
- Gulyaev, A.P. and Roussis, A. (2007) Identification of conserved secondary structures and expansion segments in enod40 RNAs reveals new enod40 homologues in plants. *Nucleic Acids Res.* 35, 3144–3152
- Ruiz-Orera, J. *et al.* (2020) Evolution of new proteins from translated sORFs in long non-coding RNAs. *Exp. Cell Res.* 391, 111940

45. Bazin, J. *et al.* (2017) Global analysis of ribosome-associated noncoding RNAs unveils new modes of translational regulation. *Proc. Natl. Acad. Sci. U. S. A.* 114, E10018–E10027
46. Guo, Y. *et al.* (2023) The translational landscape of bread wheat during grain development. *Plant Cell* 35, 1848–1867
47. Hsu, P.Y. *et al.* (2016) Super-resolution ribosome profiling reveals unannotated translation events in Arabidopsis. *Proc. Natl. Acad. Sci. U. S. A.* 113, E7126–E7135
48. Li, S. *et al.* (2016) Biogenesis of phased siRNAs on membrane-bound polysomes in Arabidopsis. *Elife* 5, e22750
49. Lin, X. *et al.* (2020) Analysis of soybean long non-coding RNAs reveals a subset of small peptide-coding transcripts. *Plant Physiol.* 182, 1359–1374
50. Chen, Z. *et al.* (2023) sORFPred: a method based on comprehensive features and ensemble learning to predict the sORFs in plant lncRNAs. *Interdiscip. Sci.* 15, 189–201
51. Zhao, S. *et al.* (2023) Identification of small open reading frames in plant lncRNA using class-imbalance learning. *Comput. Biol. Med.* 157, 106773
52. Hutvagner, G. and Zamore, P.D. (2002) A microRNA in a multiple-turnover RNAi enzyme complex. *Science* 297, 2056–2060
53. Laressergues, D. *et al.* (2022) Characterization of plant microRNA-encoded peptides (miPEPs) reveals molecular mechanisms from the translation to activity and specificity. *Cell Rep.* 38, 110339
54. Kang, M. *et al.* (2020) Identification of miPEP133 as a novel tumor-suppressor microprotein encoded by miR-34a pri-miRNA. *Mol. Cancer* 19, 143
55. Gillon, A.D. *et al.* (2008) Biosynthesis of circular proteins in plants. *Plant J.* 53, 505–515
56. Craik, D.J. *et al.* (1999) Plant cyclotides: a unique family of cyclic and knotted proteins that defines the cyclic cystine knot structural motif. *J. Mol. Biol.* 294, 1327–1336
57. Gründemann, C. *et al.* (2012) Do plant cyclotides have potential as immunosuppressant peptides? *J. Nat. Prod.* 75, 167–174
58. Cao, X. *et al.* (2022) Nascent alt-protein chemoproteomics reveals a pre-60S assembly checkpoint inhibitor. *Nat. Chem. Biol.* 18, 643–651
59. Na, Z. *et al.* (2022) Mapping subcellular localizations of unannotated microproteins and alternative proteins with MicroID. *Mol. Cell* 82, 2900–2911.e7
60. Dolde, U. *et al.* (2018) Synthetic microProteins: versatile tools for post-translational regulation of target proteins. *Plant Physiol.* 176, 3136–3145
61. Eguen, T. *et al.* (2020) Control of flowering in rice through synthetic microProteins. *J. Integr. Plant Biol.* 62, 730–736
62. Hong, S.Y. *et al.* (2020) Heterologous microProtein expression identifies LITTLE NINJA, a dominant regulator of jasmonic acid signaling. *Proc. Natl. Acad. Sci. U. S. A.* 117, 26197–26205
63. Straub, D. and Wenkel, S. (2017) Cross-species genome-wide identification of evolutionary conserved microProteins. *Genome Biol. Evol.* 9, 777–789
64. Bisson, M.M. and Groth, G. (2015) Targeting Plant ethylene responses by controlling essential protein-protein interactions in the ethylene pathway. *Mol. Plant* 8, 1165–1174
65. Bisson, M.M. *et al.* (2016) Peptides interfering with protein-protein interactions in the ethylene signaling pathway delay tomato fruit ripening. *Sci. Rep.* 6, 30634
66. Ormancey, M. *et al.* (2023) Complementary peptides represent a credible alternative to agrochemicals by activating translation of targeted proteins. *Nat. Commun.* 14, 254
67. Nevers, Y. *et al.* (2023) Protein length distribution is remarkably uniform across the tree of life. *Genome Biol.* 24, 135
68. McWhite, C.D. *et al.* (2024) Alternative proteoforms and proteoform-dependent assemblies in humans and plants. *Mol. Syst. Biol.* 20, 933–951
69. Takahashi, H. *et al.* (2012) 5' end-centered expression profiling using cap-analysis gene expression and next-generation sequencing. *Nat. Protoc.* 7, 542–561
70. Ni, T. *et al.* (2010) A paired-end sequencing strategy to map the complex landscape of transcription initiation. *Nat. Methods* 7, 521–527
71. Morton, T. *et al.* (2014) Paired-end analysis of transcription start sites in Arabidopsis reveals plant-specific promoter signatures. *Plant Cell* 26, 2746–2760
72. Pelechano, V. *et al.* (2014) Genome-wide identification of transcript start and end sites by transcript isoform sequencing. *Nat. Protoc.* 9, 1740–1759
73. Thomas, Q.A. *et al.* (2020) Transcript isoform sequencing reveals widespread promoter-proximal transcriptional termination in Arabidopsis. *Nat. Commun.* 11, 2589
74. Rhoads, A. and Au, K.F. (2015) PacBio sequencing and its applications. *Genomics Proteomics Bioinforma.* 13, 278–289
75. Schaarschmidt, S. *et al.* (2020) Utilizing PacBio iso-seq for novel transcript and gene discovery of abiotic stress responses in *Oryza sativa* L. *Int. J. Mol. Sci.* 21, 8148
76. Gelsinger, D.R. *et al.* (2020) Ribosome profiling in archaea reveals leaderless translation, novel translational initiation sites, and ribosome pausing at single codon resolution. *Nucleic Acids Res.* 48, 5201–5216
77. Ingolia, N.T. *et al.* (2012) The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat. Protoc.* 7, 1534–1550
78. Wu, H.L. *et al.* (2019) The tomato translational landscape revealed by transcriptome assembly and ribosome profiling. *Plant Physiol.* 181, 367–380
79. Chu, Q. *et al.* (2017) Identification of microprotein-protein interactions via APEX tagging. *Biochemistry* 56, 3299–3306
80. Stekhoven, D.J. *et al.* (2014) Proteome-wide identification of predominant subcellular protein localizations in a bacterial model organism. *J. Proteome* 99, 123–137