Short Title: The *Medicago truncatula* Small Secreted Peptide Database

**Corresponding author:** Patrick Xuechun Zhao ([pzhao@noble.org](mailto:pzhao@noble.org))

MtSSPdb: the *Medicago truncatula* Small Secreted Peptide Database

Clarissa Boschiero ([cboschiero@noble.org](mailto:cboschiero@noble.org))\(^1\),†, Xinbin Dai ([xdai@noble.org](mailto:xdai@noble.org))\(^1\),†, Peter Knut Lundquist ([pklundq@msu.edu](mailto:pklundq@msu.edu))\(^1\), Sonali Roy ([sroy@noble.org](mailto:sroy@noble.org))\(^1\), Thomas Christian de Bang ([tdb@plen.ku.dk](mailto:tdb@plen.ku.dk))\(^2\), Shulan Zhang ([shzhang@noble.org](mailto:shzhang@noble.org))\(^1\), Zhaohong Zhuang ([zzhuang@noble.org](mailto:zzhuang@noble.org))\(^1\), Ivone Torres-Jerez ([itjerez@noble.org](mailto:itjerez@noble.org))\(^1\), Michael K. Udvardi ([mudvardi@noble.org](mailto:mudvardi@noble.org))\(^1\), Wolf-Rüdiger Scheible ([wrscheible@noble.org](mailto:wrscheible@noble.org))\(^1\), and Patrick Xuechun Zhao ([pzhao@noble.org](mailto:pzhao@noble.org))\(^1\)

†Contributed equally

\(^1\)Noble Research Institute, 2510 Sam Noble Parkway, Ardmore, OK 73401, USA
\(^2\)Department of Biochemistry & Molecular Biology, Plant Resilience Institute, Michigan State University, East Lansing, MI 48824, USA
\(^3\)Department of Plant and Environmental Sciences and Copenhagen Plant Science Center, Faculty of Science, University of Copenhagen, DK-1871 Frederiksberg C, Denmark

One-sentence summary: MtSSPdb hosts a compendium of small secreted peptide sequences with annotations and an RNA-seq-based gene expression atlas for *Medicago truncatula*, a plant small secreted peptide prediction tool, and phenotyping data from synthetic peptide screens *in planta*.

Footnotes:

\(^1\)This work was funded by the National Science Foundation (NSF IOS #1444549), the Oklahoma Center for the Advancement of Science and Technology (OCAST PS18-012), and the Noble Research Institute. Financial support to T.C.d.B. was provided by Novo Nordisk Fonden (NNF17OC0024884).

\(^2\)Corresponding author email: [pzhao@noble.org](mailto:pzhao@noble.org)

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantphysiol.org) is: Patrick Xuechun Zhao ([pzhao@noble.org](mailto:pzhao@noble.org)).
Author contributions: C.B., X.D., and P.X.Z designed, developed, and implemented the database. C.B. constructed the database website, performed bioinformatics analyses, and wrote the manuscript. P.K.L., S.R., and T.C.d.B. generated case studies and peptide library data, provided usage feedback, and revised the manuscript. S.Z. performed the peptide screening for root-related phenotypes. Z.Z. performed data analyses. I.T.J. performed RT-qPCR validations and hormone treated RNA-seq experiments. P.X.Z., W.R.S., and M.K.U. conceived and coordinated the project, revised the manuscript, and provided overall supervision. All authors read and approved the final manuscript.

ABSTRACT

A growing number of small secreted peptides (SSPs) in plants are recognized as important regulatory molecules with roles in processes such as growth, development, reproduction, stress tolerance, and pathogen defense. Recent discoveries further implicate SSPs in regulating root nodule development, which are of particular significance for legumes. SSP-coding genes are frequently overlooked due to genome annotation pipelines generally ignoring small open reading frames (sORFs), which are those most likely to encode SSPs. Also, SSP-coding sORFs are often expressed at low levels or only under specific conditions, thus are under-represented in non-tissue-targeted or non-condition-optimized RNA-seq projects. We previously identified 4,439 SSP-encoding genes in the model legume *Medicago truncatula*. To support systematic characterization and annotation of these putative SSP-encoding genes, we developed the *Medicago truncatula* Small Secreted Peptide Database or MtSSPdb (https://mtsspdb.noble.org/). MtSSPdb currently hosts 1) a compendium of *Medicago truncatula* SSP candidates with putative function and family annotations; 2) a large-scale *Medicago truncatula* RNA-seq-based gene expression atlas integrated with various analytical tools, including differential expression, co-expression, and pathway enrichment analyses; 3) an online plant SSP prediction tool capable of analyzing protein sequences at the genome-scale using the same protocol for the identification of SSP genes; and 4) information about a library of synthetic peptides and root and nodule phenotyping data from synthetic peptide screens *in planta*. These datasets and analytical tools make MtSSPdb a unique and valuable resource for the plant research community. MtSSPdb also has the potential to become the most complete database of SSPs in plants.
INTRODUCTION

Plant small secreted peptides (SSPs) are crucial intercellular messenger molecules that regulate a multitude of processes (Matsubayashi, 2014). SSPs are typically encoded within preproteins of 100-250 amino acids, that are subsequently processed into shorter bioactive peptides of ca. 5-50 residues (Breiden and Simon, 2016, de Bang et al., 2017, Lease and Walker, 2006), which act at very low, often nanomolar physiological concentrations (Murphy et al., 2012).

SSPs have emerged as an important class of regulatory molecules involved in plant growth, development, plant-microbe interactions, and stress tolerance (Czyzewicz et al., 2013, Nakaminami et al., 2018, Takahashi et al., 2018). This is of particular significance for legumes, since recent discoveries show that SSPs regulate symbiotic root nodulation (Djordjevic et al., 2015, Nishida et al., 2018, Kereszt et al., 2018), and root development (Araya et al., 2016, Patel et al., 2018). SSPs are also involved in reproductive development, embryogenesis, and pathogen interaction, among many other plant processes (Breiden and Simon, 2016, Matsubayashi, 2014). Due to their various effects in plants, SSPs are of interest as potential tools to improve plant performance, including as supplements to improve fertilizer-use efficiency for instance.

Legumes are key components of sustainable agricultural systems since they form symbioses with soil bacteria that fix atmospheric nitrogen, reducing dependency on synthetic nitrogen fertilizers, with clear benefits to agricultural producers and the environment (Graham and Vance, 2003, Valentine et al., 2017). *Medicago truncatula* has been chosen as a premier model legume because it is closely related to economically important forage species such as alfalfa (Young and Udvardi, 2009), and it is invaluable for cross-legume genomic comparison studies (Tang et al., 2014). The *Medicago truncatula* sequencing project began in 2003; its bacterial artificial chromosome (BAC) based genomic assembly was released in 2011 (Young et al., 2011) (Mt3.5), and an optical map-based assembly using Illumina and 454 sequences was released in 2014 (Tang et al., 2014) (Mt4.0). The Mt4.0 assembly has 50,894 genes (31,661 with high confidence and 19,233 with low confidence), with an ~82% overlap with the previous genome annotation (Mt3.5), but there are still gaps, unanchored scaffolds, and 13,367 genes annotated as encoding hypothetical proteins (Tang et al., 2014). More recently, the *Medicago truncatula* genome assembly (MtrunA17r5.0-ANR) based on high-depth PacBio sequencing, in which a total of 51,316 gene models that also included a significant number of non-coding genes, was published (Pecrix et al., 2018). Yet, this assembly was not focused on SSP discovery, and many SSP-coding genes are not included.
It is important to note that short open reading frames (ORFs) were largely overlooked and omitted in both Mt3.5 and Mt4.0 annotations. Newly generated RNA-seq data can provide expression evidence for such omitted genes. With the intent to mine for previously omitted genes, we re-annotated the *Medicago truncatula* genome (both Mt3.5v5 and Mt4.0v1 genome assemblies) using 64 RNA-seq libraries (de Bang et al., 2017). In addition, Hidden Markov models (HMMs) of known SSP families were used to scan both genome assemblies for SSP genes. Relying on the improved procedure, 4,439 SSP-coding genes were identified in *Medicago truncatula*, including 2,455 novel SSPs which were previously not reported in the literature (de Bang et al., 2017). The multi-step analytical procedure employed in this study (Boschiero et al., 2019) was particularly successful in the prediction of new small ORFs that were overlooked in standard genome annotation (Zhou et al., 2013).

To host the re-annotated genes, SSP families, and related knowledge (de Bang et al., 2017), we developed a comprehensive database named MtSSPdb - the *Medicago truncatula* Small Secreted Peptide Database. The main highlights and features of MtSSPdb are: a) a compendium of 48 known SSP gene families and more than 200 putative SSP families, which were curated from 4,439 potential or confirmed SSP-coding genes from the above-mentioned re-annotation procedure; b) an online prediction tool that is able to predict SSPs for user-submitted large-scale protein sequences using a protocol similar to that described in (de Bang et al., 2017); c) a comprehensive transcriptome database for SSP genes with analytical tools; and d) a catalog of trait information for a collection of SSPs tested on roots and nodulated plants. MtSSPdb also hosts all novel gene models in *Medicago truncatula* that were identified by the re-annotation procedure (de Bang et al., 2017). MtSSPdb is an important resource for the plant scientific community and has potential to become the most complete database of small secreted peptides in plants.

**DATABASE CONTENT**

**Medicago truncatula** SSP genes and families

The *Medicago truncatula* genome (both Mt3.5v5 and Mt4.0v1 genome assemblies) was recently re-annotated, and in total, 70,094 non-redundant genes were predicted including 7,771 newly annotated gene loci (de Bang et al., 2017). The re-annotation corrected many previously predicted gene models, and helped to identify additional genes (de Bang et al., 2017). From the non-redundant
genes, 4,425 gene loci were flagged as candidate SSP-coding genes based on different criteria, including protein length, signal peptide prediction, and homology with previously known SSPs or HMMs identified in previously known SSP families. A total of 1,970 of these SSP-coding genes were homologs of 46 previously established SSP gene families, while an additional 2,455 candidate SSP-coding genes were named the “Focal List,” containing potential novel SSPs. Importantly, from the potentially novel SSPs, 56% were found to have a putative ortholog in at least one of 16 plant species, including many which appear to be legume specific (de Bang et al., 2017).

Among this Focal List, a new gene family was identified called Peptide Suppressing Nodulation (PSN) encompassing four members (de Bang et al., 2017). Moreover, 14 new SSPs in Medicago truncatula were added from the recently described IRON MAN (IMA) family (Grillet et al., 2018), making a grand total of 4,439 putative SSP-coding genes (1,988 SSP homologs and 2,451 putative novel SSPs) and 48 SSP gene families (Figure 1A).

MtSSPdb provides HMMs for most SSP families. All 4,439 SSP-coding genes belong to 262 SSP gene families (48 known and 214 putative SSP families) which are well-described in the MtSSPdb, including family HMM models, profile logos for visualization, and their gene members. The 48 known gene families include 1,988 genes, and the 214 putative SSP families include 2,451 potentially novel SSPs that were previously unreported in the literature. The smallest SSP families are Plant Elicitor Peptides (PEPs), Casparian Strip Integrity Factor (CIF), and Subtilisin-embedded Plant Elicitor Peptide (SUBPEP) with only one Medicago truncatula gene; and the largest families are Nodule-specific Cysteine Rich Group B (NCR-B) and Group A (NCR-A) with 428 and 361 genes, respectively, totaling 789 genes in MtSSPdb.

It is worth mentioning that the 24 known SSP gene families that were searched for (de Bang et al., 2017) but not identified in Medicago truncatula, were nonetheless included in MtSSPdb. These families were discovered in several other plant species, including Arabidopsis (Arabidopsis thaliana), maize (Zea mays), or tobacco (Nicotiana tabacum), and can be useful in the study of other species.

**RNA-seq gene expression data**

The SSP Gene Expression Atlas (SSP-GEA) currently hosts 16 RNA-seq experiments (ten publicly available datasets and six datasets from in-house experiments) comprising 681 RNA-seq samples.
from 192 treatments or plant organs (Figure 1B) covering drought, plant hormone treatments, macronutrient deficiencies, nodule/root development, symbiotic interactions, salt stress, and various plant organs for all genes, not only for SSP-coding genes. SSP-GEA will be updated twice a year depending on newly available data and user suggestions. Currently, no other gene atlas curating published RNA-seq data for Medicago truncatula is available. It is important to mention that the previous Medicago truncatula Gene Expression Atlas (MtGEA) includes only microarray data (739 arrays from 274 experiments) and the last update was in 2015 (Benedito et al., 2008).

Synthetic peptide library data

Another section in MtSSPdb is the peptide library that currently lists 155 synthetic peptides derived from 104 Mt genes, including 95 SSP-coding genes from 20 known SSP families, six putative SSPs from five putative SSP families, and three non-SSPs. In this section, users can find the peptides grouped by gene family for ease of use. For each peptide, detailed information about chemical composition [e.g., molecular weight, pI (isoelectric point), and GRAVY (grand average of hydropathicity)], and phenotype description (Supplemental Figure S1) is provided. The synthetic peptides were tested on three different species (Medicago truncatula, Arabidopsis, and Panicum virgatum) for 24 root- and nodule-related phenotypes based on five categories: i) descriptive traits, ii) primary root traits, iii) lateral root traits, iv) total root traits, and v) nodule traits (Table 1). Images are available showing root phenotypes from Medicago truncatula plants treated with 91 different synthetic peptides compared to untreated control roots. Metadata information for each peptide is available for download, and users can request aliquots of available synthetic peptides for their research via the provided contact form. Detailed information about the synthetic peptides, SSP families, and their annotations are shown in Supplemental Table S1.

USER INTERFACE AND UTILITY

MtSSPdb structure is categorized into three main sections: resources, Gene Expression Atlas, and tools (Figure 2). MtSSPdb provides three analytical tools - a search function for genes, representative transcripts, functional annotations, or SSP gene families; a BLAST search function; and an online plant SSP prediction tool that identifies genes that potentially encode SSPs in the
precursor proteins. These sections are described below in more detail to show their utility and usage.

**Search tool and gene/family card information**

The search tool is a primary function to search *Medicago truncatula* genes by gene or transcript IDs, keywords of annotation, or SSP gene family name (Figure 3). There is an option to preselect only SSP-coding genes, which narrows down search results to SSP-coding genes. For each gene, there is a Gene Card page available with detailed information, including genomic coordinates, SignalP D-score, protein length, SSP type, annotation, and sequences. Additionally, gene expression examples are shown for different experiments (Figure 3).

MtSSPdb contains SSP gene families divided into two groups, i.e. “known” and “putative” SSPs. There are currently 48 known SSP gene families containing 1,988 SSP-coding genes identified in *Medicago truncatula* with gene family card information for each of these (Figure 4A). An additional 24 SSP gene families are included although they were not yet identified in *Medicago truncatula*. HMM profiles are available for 37 known families and for 59 putative families with at least five members. HMM profiles can be downloaded or visualized by sequence logos (Figure 4B). Users can download FASTA or alignment files and an HMM profile file for each family, and visualize all gene family members (Figure 4C). Figure 4 presents an example of a gene family card for Clavata/Embryo Surrounding Region (CLE) with 52 gene members.

MtrunA17r5.0-ANR gene IDs (Pecrix et al., 2018) are also available on the Gene Card page. To facilitate searching SSPs in the MtrunA17r5.0-ANR genome assembly, we conducted gene mapping. The total of 44,623 MtrunA17r5.0-ANR transcripts were queried against the 70,094 Noble genome re-annotation transcripts (de Bang et al., 2017). We successfully mapped 43,018 (96.4%) of the MtrunA17r5.0-ANR transcripts to the Noble *Medicago truncatula* genome re-annotation, including ~39% of small genes (< 200 amino acids) and ~77% of the SSP-coding genes (Supplemental Table S2). We further conducted SSP gene prediction on the remaining 1,605 MtrunA17r5.0-ANR genes that were not mapped to the Noble genome re-annotation using our integrated online plant SSP prediction tool. Among the 1,605 MtrunA17r5.0-ANR genes, only five belong to known SSP families, 11 likely belong to known SSP families, and 183 were identified as
putative SSPs. These results show that the Noble *Medicago truncatula* genome re-annotation was optimized for identification of small genes (de Bang et al., 2017).

**BLAST**

A BLAST tool was implemented with different search options (BLASTN, BLASTP, BLASTX, TBLASTN, or TBLASTX) (Figure 3). We developed a web interface for NCBI BLAST, which enables users to search their sequences against hosted sequences (Camacho et al., 2009). Users can select two different target libraries (all *Medicago truncatula* genes or only SSP-coding genes) and different output formats (Figure 3). The BLAST tool allows users to upload up to 500 MB data per sequence search. In the output, a link to the respective gene card page is provided for each gene.

**SSP Prediction Tool**

Due to a lack of SSP prediction tools in the public domain, we developed and implemented such a tool as part of MtSSPdb. The tool predicts if a given protein sequence is likely to encompass an SSP based on several criteria (de Bang et al., 2017), including (1) protein length (≤ 200, 230, or 250) (Breiden and Simon, 2016, de Bang et al., 2017, Lease and Walker, 2006); (2) presence of a signal peptide cleavage site(Petersen et al., 2011); (3) presence of a sequence pattern characteristic of HMMs of known SSP gene families; (4) homology with known SSP-coding genes previously identified; and (5) absence of transmembrane (TM) helices. The prediction pipeline is suited to the analysis of protein sequences from multiple plant species, since its reference sequences and HMMs are built based on sequences from 35 diverse plant species such as Arabidopsis, *Medicago truncatula*, soybean (*Glycine max*), maize, rice (*Oryza sativa*), poplar (*Populus trichocarpa*), tobacco, *Lotus japonicus*, grapevine (*Vitis vinifera*), Amborella trichopoda, chickpea (*Cicer arietinum*), common bean (*Phaseolus vulgaris*), tomato (*Solanum lycopersicum*), clementine (*Citrus clementina*), and others (Ghorbani et al., 2015, Grillet et al., 2018).

Preproteins comprising SSPs contain an N-terminal signal peptide that directs the preprotein to the endoplasmic reticulum for cleavage, maturation, and sorting. SignalP 4.1 has been shown to be an effective predictor of N-terminal signal peptides of proteins from a wide array of species, including prokaryotes and eukaryotes (Petersen et al., 2011). It relies on a *D*-score, which is a combined value from signal peptide and cleavage site prediction networks, and is used to discriminate signal
peptides from non-signal peptides. The default $D$-score cut-off for signal peptide is 0.45 as applied in SignalP 4.1 server (Petersen et al., 2011).

After the sequence’s homology analysis, prediction of TM helices is then performed with putative SSP-coding genes passing the above criteria. Any gene predicted to harbor at least one TM helix is considered not to be an SSP, since membrane-anchoring is not characteristic of SSPs that act in cell-cell signaling, but note that TM predictions can vary depending on the tool used (Ganapathiraju et al., 2008, Tsirigos et al., 2015).

The final output table (Figure 3) presents the calculated values of each of the five individual features plus a cumulative prediction that places the protein within one of three types of SSPs – “known”, “likely known”, or “putative” SSP. A known SSP has a protein length of ≤ 200 amino acids, a SignalP $D$-score of > 0.25, and homology with previous SSPs, while a putative SSP has a protein length of ≤ 230 amino acids, SignalP $D$-score of > 0.45, no TM domains, and no significant homologies with known SSPs or hits with only one type of homology. We included an additional SSP type defined as “likely known SSPs” with significant homologies to known SSPs and a small protein length (≤ 250 amino acids). In this category, there are for example several CLE peptides, including CLE2, 8, 19, 27, 34, 36, 41, and 48. The details of criteria were described in our previously published paper (de Bang et al., 2017). All details about input, output, and criteria used are provided on the Help page of MtSSPdb.

In the output result page, users can filter the results by adjusting various cutoff thresholds ("protein length", "SignalP $D$-score", "HMM homology e-value", and "Smith-Waterman homology e-value") using our filter function. In addition, users can filter the results by the SSP classification; for example, users can choose to display only known, likely known, or putative SSPs.

When we analyzed all Medicago truncatula re-annotated genes (70,094) with our SSP prediction tool, our prediction generated a 98.6% matching classification with those previously produced by our group (de Bang et al., 2017). The differences are primarily due to manual curation. We recommend that SSPs predicted by our tool be further confirmed by expression evidence and subsequent experimental validation.

SSP Gene Expression Atlas (SSP-GEA)
The SSP Gene Expression Atlas (SSP-GEA) is a major component of MtSSPdb, and provides several tools to analyze and display (1) an expression profile with gene search by keyword, ID, and expression pattern over conditions, (2) differential gene expression, (3) gene co-expression, (4) Gene Ontology (GO) term, and (5) Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment for a selected group of genes (Figure 5). SSP-GEA allows users to choose the RNA-seq dataset of interest, and all the above analyses are performed “on-the-fly”. SSP-GEA allows users to query expression levels of any Medicago truncatula gene or transcript of interest. Queries can also be done using SSP gene family name or annotation keywords. Users begin by selecting experiment(s) of interest (Figure 1B), and the experimental conditions. After submission, the raw read counts of selected samples are extracted from the database and normalized for further analysis, and then users can search genes by keyword or expression patterns. A bar chart or line chart is generated on the web page. Users have the option to download a “.CSV” table with all the expression values of the queried genes from expression profile, differential expression (DE), or co-expression analyses.

Figure 5 shows the different outputs obtained from the analytical tools. In SSP-GEA, the DE analysis tool was designed for differential expression analysis of RNA-seq data from different experiments. Users can select the desired experiment, the two conditions to be compared (numerator and denominator), and the \( p \)-value cut-off. Users also have the option to filter out any gene with a low number of normalized mean counts or filter the results based on log2-fold-changes or adjusted \( p \)-values. For co-expression analysis, users select the relevant experiments and conditions. Once the analysis is complete, users can specify their genes of interest to extract a list of co-expressed genes. GO enrichment and KEGG pathway analyses are either available separately or are integrated with expression profile, DE, and co-expression analyses where the output results from these upstream analyses can be directly imported into the GO/KEGG analysis module to further identify enriched GO terms or KEGG pathways in the list of genes. Additional filters are available prior to the GO and KEGG analyses, such as \( p \) or \( q \) value cut-off and number of top terms or pathways.

CASE STUDIES

The following case studies demonstrate the usefulness of MtSSPdb in identifying novel candidate peptides and their biological functions.
Bioactivity of CEP9 requires proline hydroxylation

The C-terminally Encoded Peptide 1 (CEP1) peptide of *Medicago truncatula* was previously shown to inhibit lateral root formation when applied exogenously to germinated seedlings (Imin et al., 2013). The *Medicago truncatula* genome encodes 17 CEP genes, 15 of which are represented in the Synthetic Peptide Library of MtSSPdb (Figure 6A). Like the CEP1 gene (*MT35v5_contig_59554_1*), CEP9 (*MT35v5_AC233112_1015*) harbors two highly similar CEP domains. Peptides representing both domains are present in the peptide library, and these peptides were synthesized with or without hydroxyproline (Hyp) residues on residues 4 and 11 (Figure 6B), which have been shown to be important for bioactivity in CEP1. The screening data in the peptide library clearly show that both CEP9 peptides reduce lateral root density, like CEP1, and that this activity depends on Hyp at residues 4 and 11 (Figure 6C). These results indicate the CEP9 peptide may share overlapping function with CEP1 in *Medicago truncatula* roots.

Identification of a novel SSP transcript with support from MtSSPdb

Visualization of RNA-seq data revealed that *Medtr6g027320* was incorrectly annotated in Mt v4.0 (Figure 7A). Two transcripts can be seen with widely differing expression patterns. Both corrected transcripts within this locus have ORFs with ~60% of identity, strongly predicted signal peptides, and Plant Peptide Containing Sulfated Tyrosine (PSY) domains near the C-terminus. Accordingly, these transcripts were renamed as *PSY7* and *PSY8*. Extracting tissue expression plots from the SSP Gene Expression Atlas section of MtSSPdb, we found that *PSY7* is the dominantly expressed transcript and is primarily expressed in aerial organs such as leaf and petiole (Figure 7B). In contrast, *PSY8* is expressed at low levels throughout much of the plant, including root, nodule, and stem. DE analysis, using the SSP-GEA, revealed that *PSY7* transcript levels in leaves exceed those in roots by over 1,000-fold, while *PSY8* transcript levels were about three-fold lower in shoots than in roots. To test for activity of *PSY7*, the predicted 19-residue peptide was synthesized, including the expected sulfate group at Tyr2 and Hyp at Pro13 and Pro16 (Amano et al., 2007), and employed in our root phenotyping screen. Compared to the mock-treated control, the *PSY7* peptide enhanced primary root growth (Figure 7C), consistent with the characterized role of AtPSY in cell expansion (Amano et al., 2007). Additionally, a slight suppression of lateral root density is also observed.
The bioactivity found for this synthetic peptide provides additional support for the identification of this novel transcript.

**SSPs regulated during mycorrhizal and rhizobial symbioses**

The SSP-GEA contains expression data from symbiotic interactions with nitrogen-fixing bacteria known as rhizobia and arbuscular mycorrhizal fungi (Luginbuehl et al., 2017). In legumes, these two symbionts are known to share common signaling components during development of nodules (root organs hosting the rhizobia) and mycorrhizal colonization. To investigate if any SSP gene transcripts were commonly regulated during both nodulation and mycorrhizal colonization, DE analyses were performed in SSP-GEA based on nodules at 4, 10, 14, and 28 days post rhizobia inoculation compared to controls (de Bang et al., 2017), and mycorrhizal roots at 8, 14, and 27 days post inoculation compared to non-mycorrhizal control roots (Luginbuehl et al., 2017). In total, 341 to 1,173 differentially expressed SSPs were identified in nodules, compared to 29 to 108 in mycorrhizal roots (adjusted p-value < 0.1) (Figure 8A). Collectively, 1,292 individual SSPs were identified to be differentially expressed during nodule development, whereas only 144 individual SSPs were found to be differentially expressed in mycorrhizal roots. Despite this large discrepancy in the number of differentially expressed SSPs between rhizobial and mycorrhizal roots, 107 SSPs were found to be shared, of which many showed a similar response to nodulation and mycorrhization, while others responded in opposite directions (Figure 8B). Hierarchical clustering based on transcriptional changes of the 107 commonly regulated SSPs grouped the SSPs into four different clusters (Figure 8C). SSPs in Cluster I were highly induced in nodules and during later stages of mycorrhizal symbiosis, and included 10 NCR peptides, five leginsulins, five Nodule-specific Glycine Rich Peptides (NodGRPs), and three plant defensins (Supplemental Table S3).

The high expression of supposedly nodule-specific NCRs and NodGRPs in mycorrhizal roots could possibly indicate that the analyzed roots were also nodulated. However, the well-studied nodulation-marker SSP transcripts CLE12 and CLE13 were not induced, leading to the conclusion that the expression of this subset of NCRs and NodGRP is not nodule-specific. Cluster II contained SSPs generally upregulated in both mycorrhizal roots and nodules. A group of these were strongly upregulated in mycorrhizal roots compared to nodulation, which included five plantacyanins (PCY). Cluster III constituted 13 SSP transcripts with reduced abundance in nodules, but with moderately induced expression in mycorrhizal roots at 8 dpi. Six of these belonged to the Root Cap (RC)
family. SSPs moderately upregulated by mycorrhiza and downregulated during nodulation grouped into Cluster IV-a that contained four SSPs from the Pro-rich Protein Group 669 (PRP669), and five from the subtilisin inhibitor (SubIn) families, respectively. Cluster IV-b represented 30 SSPs with significant downregulation (adjusted \( p\)-value < 0.1) during nodulation, but an inconsistent response to mycorrhizal colonization. Nine of these were PCYs and four were non-specific lipid transfer proteins (LTPs).

**CAPE16 is implicated in rhizobial persistence within nodules**

SSPs in the CAP-derived Peptide (CAPE) family are derived from functional precursor proteins involved in the pathogen defense pathway in leaves (Chen et al., 2014). The CAPE peptides are embedded at the C-terminus of larger Pathogenesis-related protein 1 proteins and are cleaved into an 11-residue peptide prior to secretion. Several CAPE peptides in Arabidopsis are induced by salt treatment (Chien et al., 2015), but no functional studies have been carried out. Analysis with SSP-GEA in MtSSPdb revealed that several CAPE gene transcripts are abundant in nodule tissue (Figure 9A). Investigation of the co-expression patterns of selected CAPE gene transcripts revealed that CAPE16 (Medtr5g018770), in particular, was strongly enriched for co-expression with other SSP-coding genes. Filtering co-expressing genes with > 0.8 Pearson correlation coefficient showed that 23% of CAPE16 co-expressed gene transcripts were SSP-coding genes, compared to 5-9% of four other CAPE gene transcripts (Figure 9B). The co-expressing SSPs were predominantly NCR, leginsulin, and plant defensin genes (Figure 9C). NCR genes in particular encode nodule-specific SSPs with roles in development and maintenance of rhizobia within symbiotic nodules in *Medicago truncatula* (Kereszt et al., 2018). Thus, the expression patterns discerned from MtSSPdb may indicate a role in nodulation for CAPE16, but further experiments should be conducted to validate these findings.

**DISCUSSION**

There are only three published databases dedicated to plant small peptides (Table 2). PlantSSPdb hosts a collection of small secretory peptides from 32 plant species, including 820 *Medicago truncatula* SSP-coding genes (Ghorbani et al., 2015). Besides PlantSSPdb, there is SPdb, a signal peptide database containing signal sequences of archaea, prokaryotes and eukaryotes, but it includes
only 17 SSPs described in *Medicago truncatula* (Choo et al., 2005). Also, there is a database of small secreted peptides predicted in Arabidopsis with very limited and outdated information (Lease and Walker, 2006).

In PlantSSPdb, it is possible to browse SSP genes and download family HMMs and protein sequences for five pillar species (Arabidopsis, rice, poplar, grapevine, and maize). PlantSSPdb uses similar criteria to ours to identify SSPs within their reference pillar species. However, in additional species such as *Medicago truncatula*, the SSP identification relies solely on automated, unsupervised searches on each family’s HMM built from the five pillar species. Because SSPs rapidly evolve and can be species- or genera-specific, HMM-based searches that rely on evidence from only the five pillar species are of limited value for alternative species, in particular *Medicago truncatula* and other legumes which are known to have a number of legume-specific SSPs.

In contrast to PlantSSP, MtSSPdb relies on iterative searches and extensive manual curation throughout the *Medicago truncatula* genome (de Bang et al., 2017). These steps have resulted in high-quality SSP gene models for *Medicago truncatula*. The improved analytical procedures, described in (de Bang et al., 2017), identified over 4,000 predicted SSPs, including almost 2,000 members of known SSP families, and a novel legume-specific SSP family named PSN (de Bang et al., 2017). Most of the 820 *Medicago truncatula* SSP genes found in the PlantSSP database (Ghorbani et al., 2015) are included in MtSSPdb, but only ~30% of the SSP-coding genes could be associated with the HMMs from the PlantSSP database (de Bang et al., 2017). This is likely a reflection of the extensive manual curation underlying MtSSPdb, and highlights the greatly improved identification of putative *Medicago truncatula* SSPs.

Furthermore, no gene expression or annotation data is available for the SSP genes in PlantSSPdb (Ghorbani et al., 2015). The gene expression data and related analytical functions, such as profiling, differential expression, co-expression, and pathway enrichment analysis, are helpful tools for exclusion of false positive predictions and the identification of biological functions for the SSP genes.

Online analysis tools are a convenient way to search SSP genes from user-submitted sequences. PlantSSPdb provides a web-based BLAST search tool against SSP reference sequences with an 8 MB limit for data upload. NCBI BLAST is a heuristic search algorithm which compromises on sensitivity to sequences with lower similarity to obtain faster search performance (Camacho et al., 2009). This feature may cause the loss of SSP candidates, since SSPs rapidly evolve and...
conservation can be limited to short sequences. To address this issue, MtSSPdb integrates a comprehensive SSP prediction tool which utilizes HMM and Smith-Waterman searches. In addition, our database also provides the information from SignalP analysis and protein length. The SSP prediction tool was able to identify more accurate results of SSPs compared to BLAST from PlantSSPdb (Ghorbani et al., 2015). For example, using as input *Medicago truncatula* protein sequences from Mt3.4v4 (*n* = 64,152), we predicted 1,218 known SSPs with the SSP tool, but using BLAST/PlantSSPdb with a stringent *e*-value (< 1e-07), we obtained more than 7,000 best hits, and most of them with low identity.

The MtSSPdb prediction tool accepts user-submitted protein sequences up to 500 MB, which is enough for most genome-wide analyses. This prediction tool enables users to easily utilize the knowledge of known SSP families to identify new SSP proteins with high confidence.

In addition, MtSSPdb has gene expression information for each SSP gene, including expression profiles from common biological conditions, including various plant tissues and treatments, such as hormone and plant macro-nutrition treatments, which provide insights for the functional characterization of these SSP genes.

It is worth mentioning that MtSSPdb focuses on *Medicago truncatula*, however, it will be important to expand this database to other relevant model plant species and legumes such as alfalfa and soybean.

**CONCLUSIONS**

MtSSPdb is the first plant SSP database that integrates gene expression, an SSP prediction online tool, and synthetic peptide information. MtSSPdb hosts large-scale genomics and transcriptomics data in the model legume, *Medicago truncatula*, and provides multiple functions to search, retrieve, analyze, and visualize different datasets. It also hosts, under the synthetic peptide library, phenotyping data from synthetic peptide screens *in planta*. Compared to the previously published database (Ghorbani et al., 2015), MtSSPdb contains more comprehensive and up-to-date data for *Medicago truncatula*, resulting in a valuable resource for the plant research community. In addition, the integrated SSP prediction tool is the first web-based tool for the identification of plant SSPs from users’ submissions using multiple SSP characteristics. This tool also allows users to submit protein sequences on a genome scale for data analysis. To the best of our knowledge, no such
comprehensive resource focusing on small peptide-coding genes, which are numerous and often still
unannotated, exists for any plant species. It is worth mentioning that the database contains all
known SSP family information, including 24 families which have not been identified in *Medicago
truncatula*. Thus, this database can be expanded to other relevant model plant species, e.g.
Arabidopsis and *Brachypodium distachyon*, and legume species such as alfalfa and soybean.
MtSSPdb has potential to become the most comprehensive database of small secreted peptides in
plants. The MtSSPdb is available at https://mtsspdb.noble.org/.

MATERIALS AND METHODS

*Medicago truncatula* SSP genes and families

The *Medicago truncatula* genome (both Mt3.5v5 and Mt4.0v1 genome assemblies) was re-
annotated using the generic genome annotation tool MAKER pipeline (Cantarel et al., 2008). Gene
model expression evidences include 64 RNA-seq libraries that were mainly sequenced after the
release of MTv4.0 and protein/EST sequences that are publicly available in legumes (de Bang et al.,
2017). The SPADA pipeline (Zhou et al., 2013) and sORF Finder (Hanada et al., 2010) were used
to identify short genes. The former was optimized by including HMM SSP models from
PlantSSPdb (Ghorbani et al., 2015). The gene models were annotated using plant UniProt
(https://www.uniprot.org/) as reference database with BLASTP for GO and KEGG (*e*-values < 1e-
05). HMMs were established in two steps. Firstly, we generated a multi-alignment file for each
family using representative member sequences in *Medicago truncatula*. Secondly, the multi-
alignment files were converted into HMM model files using HMMER software (Finn et al., 2011).
The interactive gene family logos for HMM profiles were built using the Skylign tool (Wheeler et
al., 2014). BLAT (Kent, 2002) was used to map MtrunA17r5.0-ANR genes (P eerix et al., 2018).

Development of MtSSPdb web portal

The web portal was developed using the Python Flask framework and MySQL.

RNA-seq data analysis

RNA-seq data produced in-house were generated using Illumina technology representing different
organs (leaves, shoot, petioles, buds, flowers, pods, roots, and nodules). More information about the
experimental methods is available (de Bang et al., 2017). RNA-seq datasets were mapped against
the representative transcripts of *Medicago truncatula* genes to estimate raw counts and effective
transcript lengths using the Sailfish/Salmon tool (Patro et al., 2017). These results were uploaded into the database.

Normalization and differential expression analysis of raw counts are performed using DESeq2 (Love et al., 2014). The co-expression module was developed based on WGCNA (Langfelder and Horvath, 2008), which generates a co-expression matrix for the entire genome using biweight midcorrelation approach (also called bicor). The matrix can be processed to generate co-expressed functional gene modules (Langfelder and Horvath, 2008). The enriched GO terms or KEGG pathways in differentially expressed genes or co-expressed gene modules are detected by using \( p \)-values from hypergeometric distribution and Benjamini-Hochberg adjustment.

**SSP Prediction Tool**

The presence of a signal peptide cleavage site is predicted by SignalP 4.1 (Petersen et al., 2011). A sequence’s homology to previously identified SSPs is determined in two ways. First, by searching user-submitted protein sequences against a collection of 37 curated HMMs of known SSP families (de Bang et al., 2017) and 4,780 HMMs from PlantSSP (Ghorbani et al., 2015) using HMMER (Finn et al., 2011), and second, by searching user-submitted sequences against 3,402 known plant SSP protein sequences with SSearch (Ropelewski et al., 2004), a fast implementation of the Smith-Waterman search algorithm. \( E \)-values of \( \leq 0.01 \) were used for significant homologies. Prediction of TM helices is performed with TMHMM Server v.2.0 (Krogh et al., 2001), and excludes predicted N-terminal signal peptides.

**ACCESSION NUMBERS**


**SUPPLEMENTAL MATERIAL**

**Supplemental Figure S1.** MtSSPdb screenshot showing an overview of the Peptide Library section.

**Supplemental Table S1.** Gene information of synthetic peptides tested in three species.
Supplemental Table S2. Mapping results of MtrunA17r5.0-ANR transcripts against the Noble Medicago truncatula genome re-annotation.

Supplemental Table S3. SSPs commonly regulated between nodulation and mycorrhiza.

DATA AVAILABILITY

The authors declare that all data supporting the findings of this study are available within the article and Supplemental Material online or are available upon request from the corresponding author.

ACKNOWLEDGMENTS

We thank our collaborators who provided us valuable feedbacks during the development of MtSSPdb. The work was funded by the National Science Foundation (NSF IOS #1444549), the Oklahoma Center for the Advancement of Science and Technology (OCAST PS18-012), and the Noble Research Institute. Financial support to T.C.d.B. was provided by Novo Nordisk Fonden (NNF17OC0024884).

COMPETING INTERESTS

The authors declare that they have no competing interests.
Table 1. Root and nodule-related traits evaluated in *Medicago truncatula*, *Arabidopsis thaliana*, and *Panicum virgatum* for 155 synthetic peptides.

<table>
<thead>
<tr>
<th>Trait</th>
<th>Species tested</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Descriptive traits</strong></td>
<td></td>
</tr>
<tr>
<td>Root phenotype</td>
<td><em>Mt</em></td>
</tr>
<tr>
<td>Nodule phenotype</td>
<td><em>Mt</em></td>
</tr>
<tr>
<td>Ca-Spike Assay</td>
<td><em>Mt</em></td>
</tr>
<tr>
<td><strong>Primary Root Traits</strong></td>
<td></td>
</tr>
<tr>
<td>Primary root length (cm)</td>
<td><em>Mt, At, Pv</em></td>
</tr>
<tr>
<td>Lateral root density (n/cm)</td>
<td><em>Mt, At, Pv</em></td>
</tr>
<tr>
<td>Primary root mean diameter (cm)</td>
<td><em>Mt, At, Pv</em></td>
</tr>
<tr>
<td>Primary root surface area (cm²)</td>
<td><em>Mt, At, Pv</em></td>
</tr>
<tr>
<td>Primary root volume (cm³)</td>
<td><em>Mt, At, Pv</em></td>
</tr>
<tr>
<td>Primary root straightness (vector length/total length)</td>
<td><em>Mt, At, Pv</em></td>
</tr>
<tr>
<td><strong>Lateral Root Traits</strong></td>
<td></td>
</tr>
<tr>
<td>Total number of lateral roots</td>
<td><em>Mt</em></td>
</tr>
<tr>
<td>Total lateral root length (cm)</td>
<td><em>Mt, At</em></td>
</tr>
<tr>
<td>Number of secondary lateral roots</td>
<td><em>Mt, At, Pv</em></td>
</tr>
<tr>
<td>Total length of secondary lateral roots</td>
<td><em>Mt, At, Pv</em></td>
</tr>
<tr>
<td>Mean length of secondary lateral roots</td>
<td><em>Mt, At, Pv</em></td>
</tr>
<tr>
<td>Mean diameter of secondary lateral roots (cm)</td>
<td><em>Mt, At, Pv</em></td>
</tr>
<tr>
<td>Mean diameter of all lateral roots (cm)</td>
<td><em>Mt</em></td>
</tr>
<tr>
<td>Secondary lateral roots surface area (cm²)</td>
<td><em>Mt, At, Pv</em></td>
</tr>
<tr>
<td>Secondary lateral roots volume (cm³)</td>
<td><em>Mt, At, Pv</em></td>
</tr>
<tr>
<td>Secondary lateral roots insertion angle (°)</td>
<td><em>Mt, At, Pv</em></td>
</tr>
<tr>
<td><strong>Total Root Traits</strong></td>
<td></td>
</tr>
<tr>
<td>Total root length (cm)</td>
<td><em>Mt, At, Pv</em></td>
</tr>
<tr>
<td>Total root surface area system (cm²)</td>
<td><em>Mt, At, Pv</em></td>
</tr>
<tr>
<td>Total root volume system (cm³)</td>
<td><em>Mt, At, Pv</em></td>
</tr>
<tr>
<td><strong>Nodule Traits</strong></td>
<td></td>
</tr>
<tr>
<td>Nodule number</td>
<td><em>Mt</em></td>
</tr>
<tr>
<td>Nodule density</td>
<td><em>Mt</em></td>
</tr>
</tbody>
</table>

*Medicago truncatula* (*Mt*), *Arabidopsis thaliana* (*At*), and *Panicum virgatum* (*Pv*).
Table 2. A list of the available small secreted and signaling peptides databases in plants.

<table>
<thead>
<tr>
<th>SSP databases</th>
<th>MtSSPdb</th>
<th>PlantSSPdb</th>
<th>SPdb</th>
<th>Arabidopsis Unannotated Secreted Peptide db</th>
</tr>
</thead>
<tbody>
<tr>
<td>Species</td>
<td><em>M. truncatula</em> (to be expanded soon to <em>A. thaliana</em> and <em>B. distachyon</em>)</td>
<td>32 species, including <em>M. truncatula</em></td>
<td>different species including plants and <em>M. truncatula</em></td>
<td><em>A. thaliana</em></td>
</tr>
<tr>
<td><em>M. truncatula</em> SSPs (genome assembly)</td>
<td>4,439 (Mt3.5v5 and Mt4.0v1)</td>
<td>820 (Mt3.5v4)</td>
<td>17</td>
<td>not available</td>
</tr>
<tr>
<td>Annotation</td>
<td>available</td>
<td>not available</td>
<td>not available</td>
<td>not available</td>
</tr>
<tr>
<td>Gene family</td>
<td>262 available with detailed information (function, reference, HMM profile logos, genes, etc.)</td>
<td>334 available for <em>M. truncatula</em> with limited information</td>
<td>not available</td>
<td>not available</td>
</tr>
<tr>
<td>BLAST (input sequence size)</td>
<td>available (500 Mb)</td>
<td>available (8 Mb)</td>
<td>not available</td>
<td>not available</td>
</tr>
<tr>
<td>Expression data</td>
<td>available for 16 experiments and 192 conditions</td>
<td>not available</td>
<td>not available</td>
<td>not available</td>
</tr>
<tr>
<td>Gene Expression Atlas</td>
<td>available with multiple analyses</td>
<td>not available</td>
<td>not available</td>
<td>not available</td>
</tr>
<tr>
<td>SSP Prediction tool</td>
<td>available across multiple plant species and genome scale sequences</td>
<td>not available</td>
<td>not available</td>
<td>not available</td>
</tr>
<tr>
<td>Synthetic Peptide Library</td>
<td>available with 155 peptides tested on 3 species for root and nodule-related traits, and SSP order option available</td>
<td>not available</td>
<td>not available</td>
<td>not available</td>
</tr>
</tbody>
</table>
FIGURE LEGENDS

Figure 1. *Medicago truncatula* Small Secreted Peptide Database (MtSSPdb) content. **a** Proportion of *Medicago truncatula* re-annotated genes identified as known or putative small secreted peptides (SSPs) in the MtSSPdb. **b** List of RNA-seq datasets available at Medicago SSP Gene Expression Atlas (SSP-GEA).

Figure 2. Framework of the MtSSPdb showing the structure, available resources, and tools.

Figure 3. Workflow of the tools available on the MtSSPdb. Top: Three different tools - Search, SSP Prediction, and BLAST - are available to search genes or analyze nucleotide/protein sequences. Bottom: The resulting gene search is displayed as a list, and gene cards are available for each gene. BLAST output is provided. In addition, a table is displayed as a result of the SSP Prediction tool with SSP classification from protein sequences.

Figure 4. MtSSPdb screenshot showing an overview of a gene family card for the Clavata/Embryo Surrounding Region (CLE) family. **a** Gene family summary information. **b** An Hidden Markov model (HMM) profile logo showing the C-terminal region of the protein. **c** Five of the gene members and their associated information.

Figure 5. Available tools in the SSP Gene Expression Atlas showing different outputs. Top: a bar chart is displayed as a result of the expression profile analysis of *Medtr4g117040.1* transcript for different nitrogen levels in shoots and roots; a bar chart is displayed as a result of the differential expression analysis for nitrogen levels in roots (low nitrogen levels x full nutrition control; adjusted \( p\)-value < 0.1); and a line chart is displayed as a result of the co-expression analysis of the gene *Medtr6g452990* for nitrogen levels in shoots and roots (low nitrogen levels, full nutrition control, and nitrogen re-supply). Bottom: a directed acyclic graph (DAG) graph is displayed as a result of the Gene Ontology (GO) enrichment analysis from five top GO terms from genes differentially expressed for different nitrogen levels in roots; and a table is displayed with five top significantly
(q-value < 0.05) enriched Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways from
genes differentially expressed for nitrogen levels in roots.

**Figure 6.** Surveying the peptide library data reveals the importance of proline hydroxylation of C-
terminally Encoded Peptide (CEP) peptides. 

- **a** Hybrid screenshot of the peptide library page of the MtSSPdb showing the four CEP9 peptides investigated here. 
- **b** CEP9 encodes a 151-amino acid polypeptide encoding a signal peptide (hashed lines) and containing two CEP domains, designated CEP9_1 (red) and CEP9_2 (blue). The synthetic peptides used for screening are presented below, assuming the presence of hydroxyproline (Hyp). Peptides lacking Hyp held standard Pro residues at both positions. 
- **c** Synthetic peptides from both the upstream and downstream domains strongly inhibit lateral root formation, but only when proline residues 4 and 11 are hydroxylated. Percentage values indicate the average number of secondary lateral roots in the presence of the indicated peptide, relative to the mock-treated control. Plots show the median value (center lines), box limits indicate the 25th and 75th percentiles, and the bars indicate the maximum and minimum values among the data points collected (Students t-test). \( n = 10 \) for peptide-treated plants and \( n = 20 \) for mock-treated plants.

**Figure 7.** Re-annotation of a Plant Peptide Containing Sulfated Tyrosine (PSY) reveals a novel bioactive PSY SSP-coding gene. 

- **a** RNA-seq evidence indicates that Medtr6g027320.1 is annotated incorrectly in Mt v4.0. RNA-seq reads mapped to each locus are illustrated as a Sashimi plot representing the number of mapped reads at each nucleotide. Two distinct transcripts emerge when RNA-seq mapped reads are separated by expression in shoots or roots. Reads from one representative biological replicate are presented above. The new transcripts uncover an additional upstream PSY SSP-coding gene, and have been renamed PSY7 and PSY8 accordingly. 
- **b** A plot of tissue expression levels of the newly uncovered PSY7 and PSY8 genes. PSY7 is the dominantly expressed transcript and is mostly expressed in leaves and shoots. In contrast, PSY8 demonstrates low-level expression throughout roots, nodules, and stems/petioles. Inset, expression changes from shoot vs. root highlight the differential expression of the two PSY transcripts. 
- **c** Synthetic PSY7 peptide shows bioactivity when tested in the root screening platform, leading to enhanced primary root length (consistent with its characterized role in cell division/expansion), while also suppressing lateral root density. Plots show the median value (center lines), box limits indicate the 25th and 75th
percentiles, and the bars indicate the maximum and minimum values among the data points collected (Students t-test). $n = 10$ for peptide-treated plants and $n = 20$ for mock-treated plants.

Figure 8. SSPs regulated by mycorrhiza colonization and nodulation. a Number of differentially expressed SSPs during nodule development at 4, 10, 14, and 28 days post inoculation with rhizobia, and 8, 14, and 27 days post inoculation with mycorrhizal fungi (adjusted $p$-value < 0.1). b Venn diagram illustrating the number of commonly regulated SSPs between mycorrhiza and nodulation. SSPs were included if they were differentially expressed in just one of the treatments. c Hierarchical clustering of the 107 commonly regulated SSPs from log2 fold-changes were calculated based the respective control samples, i.e., non-mycorrhizal or non-rhizobial.

Figure 9. CAP-derived Peptide 16 (CAPE16) expression predominates in symbiotic nodules and co-expresses with many SSPs. a RNA-seq data from the MtSSPdb. Five CAPE peptides, which have been shown previously to be induced by macronutrient deficiencies, were interrogated for expression patterns across 12 different organs. b CAPE16 shows a strong enrichment in co-expressing SSP-coding genes. Co-expressing SSPs were defined as those with a Pearson correlation coefficient above 0.8 across 122 different experimental conditions. Thirty-five out of a total 149 co-expressors of CAPE16 were SSP genes, amounting to 23%, versus 5-9% of the other four CAPE genes. c Pie chart showing the distribution of the CAPE16 co-expressing SSP-coding genes, most of which are Nodule Cys-rich (NCR) peptides, Leginsulins, or Plant Defensins (PDF). ENOD40, Early Nodulin 40; LP, LEED..PEED; nsLTP, non-specific Lipid Transfer Protein; PCY, Plantacyanin; nodGRP, nodule-specific Gly-rich protein; CLE, CLAVATA/Embryo Surrounding Region.
LITERATURE CITED


KENT, W. J. 2002. BLAT--the BLAST-like alignment tool. Genome Res, 12, 656-64.


**Figure 1.** *Medicago truncatula* Small Secreted Peptide Database (MtSSPdb) content. a Proportion of *Medicago truncatula* re-annotated genes identified as known or putative small secreted peptides (SSPs) in the MtSSPdb. b List of RNA-seq datasets available at Medicago SSP Gene Expression Atlas (SSP-GEA).
Figure 2. Framework of the MtSSPdb showing the structure, available resources, and tools.
Figure 3. Workflow of the tools available on the MtSSPdb. Top: Three different tools - Search, plant SSP prediction, and BLAST - are available to search genes or analyze nucleotide/protein sequences. Bottom: The resulting gene search is displayed as a list, and gene cards are available for each gene. BLAST output is provided. In addition, a table is displayed as a result of the plant SSP prediction tool with SSP classification from protein sequences.
Figure 4. MtSSPdb screenshot showing an overview of a gene family card for the Clavata/Embryo Surrounding Region (CLE) family. a Gene family summary information. b An Hidden Markov model (HMM) profile logo showing the C-terminal region of the protein. c Five of the gene members and their associated information.
Figure 5. Available tools in the SSP Gene Expression Atlas showing different outputs. Top: a bar chart is displayed as a result of the expression profile analysis of Medtr4g117040.1 transcript for different nitrogen levels in shoots and roots; a bar chart is displayed as a result of the differential expression analysis for nitrogen levels in roots (low nitrogen levels x full nutrition control; adjusted p-value < 0.1); and a line chart is displayed as a result of the co-expression analysis of the gene Medtr6g452990 for nitrogen levels in shoots and roots (low nitrogen levels, full nutrition control, and nitrogen re-supply). Bottom: a directed acyclic graph (DAG) graph is displayed as a result of the Gene Ontology (GO) enrichment analysis from five top GO terms from genes differentially expressed for different nitrogen levels in roots; and a table is displayed with five top significantly (q-value < 0.05) enriched Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways from genes differentially expressed for nitrogen levels in roots.
**A**

**PEPTIDE LIBRARY**

Here you can find a list of 155 small synthetic peptides grouped by SSP gene family. For each synthetic peptide you can find additional information about their main description, and 24 root and nodule-related phenotypes observed in 3 different species (*M. truncatula, A. thaliana* and Switchgrass).

Please contact us if you would like to order any of these synthetic peptides.

<table>
<thead>
<tr>
<th>PEPTIDE NAME</th>
<th>GENE ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>CEP9_1</td>
<td>MT35v5_AC233112_1015</td>
</tr>
<tr>
<td>CEP9_1</td>
<td>MT35v5_AC233112_1015</td>
</tr>
<tr>
<td>CEP9_2</td>
<td>MT35v5_AC233112_1015</td>
</tr>
<tr>
<td>CEP9_2</td>
<td>MT35v5_AC233112_1015</td>
</tr>
</tbody>
</table>

**B**

**CEP9 (MT35_v5_AC233112_1015.1)**

- **CEP9_1**: AFR(Hyp)TTPGSS(Hyp)GVGH
- **CEP9_2**: AFRPTTPGSSPGVGH
- **CEP9_2**: AFRKTPYPNHSPGVGH
- **CEP9_2**: AFRKTPYPNHSPGVGH

**C**

Figure 6. Surveys the peptide library data reveals the importance of proline hydroxylatation of C-terminally Encoded Peptide (CEP) peptides. 

- **a** Hybrid screenshot of the peptide library page of the MtSSPdb showing the four CEP9 peptides investigated here.
- **b** CEP9 encodes a 151-amino acid polypeptide encoding a signal peptide (hashed lines) and containing two CEP domains, designated CEP9_1 (red) and CEP9_2 (blue). The synthetic peptides used for screening are presented below, assuming the presence of hydroxyproline (Hyp). Peptides lacking Hyp held standard Pro residues at both positions.
- **c** Synthetic peptides from both the upstream and downstream domains strongly inhibit lateral root formation, but only when proline residues 4 and 11 are hydroxylated. Percentage values indicate the average number of secondary lateral roots in the presence of the indicated peptide, relative to the mock-treated control. Plots show the median value (center lines), box limits indicate the 25th and 75th percentiles, and the bars indicate the maximum and minimum values among the data points collected (Students t-test). $n = 10$ for peptide-treated plants and $n = 20$ for mock-treated plants.
Figure 7. Re-annotation of a Plant Peptide Containing Sulfated Tyrosine (PSY) reveals a novel bioactive PSY SSP-coding gene. a RNA-seq evidence indicates that Medtr6g027320.1 is annotated incorrectly in Mt v4.0. RNA-seq reads mapped to each locus are illustrated as a Sashimi plot representing the number of mapped reads at each nucleotide. Two distinct transcripts emerge when RNA-seq mapped reads are separated by expression in shoots or roots. Reads from one representative biological replicate are presented above. The new transcripts uncover an additional upstream PSY SSP-coding gene, and have been renamed PSY7 and PSY8 accordingly. b A plot of tissue expression levels of the newly uncovered PSY7 and PSY8 genes. PSY7 is the dominantly expressed transcript and is mostly expressed in leaves and shoots. In contrast, PSY8 demonstrates low-level expression throughout roots, nodules, and stems/petioles. Inset, expression changes from shoot vs. root highlight the differential expression of the two PSY transcripts. c Synthetic PSY7 peptide shows bioactivity when tested in the root screening platform, leading to enhanced primary root length (consistent with its characterized role in cell division/expansion), while also suppressing lateral root density. Plots show the median value (center lines), box limits indicate the 25th and 75th percentiles, and the bars indicate the maximum and minimum values among the data points collected (Students’ t-test, $\eta = 10$ for peptide-treated plants and $n = 20$ for mock-treated plants).
Figure 8. SSPs regulated by mycorrhiza colonization and nodulation. a Number of differentially expressed SSPs during nodule development at 4, 10, 14, and 28 days post inoculation with rhizobia, and 8, 14, and 27 days post inoculation with mycorrhizal fungi (adjusted *p*-value < 0.1). b Venn diagram illustrating the number of commonly regulated SSPs between mycorrhiza and nodulation. SSPs were included if they were differentially expressed in just one of the treatments. c Hierarchical clustering of the 107 commonly regulated SSPs from log2 fold-changes were calculated based the respective control samples, i.e., non-mycorrhizal or non-rhizobial.
Figure 9. CAP-derived Peptide 16 (CAPE16) expression predominates in symbiotic nodules and co-expresses with many SSPs. a RNA-seq data from the MtSSPdb. Five CAPE peptides, which have been shown previously to be induced by macronutrient deficiencies, were interrogated for expression patterns across 12 different organs. b CAPE16 shows a strong enrichment in co-expressing SSP-coding genes. Co-expressing SSPs were defined as those with a Pearson correlation coefficient above 0.8 across 122 different experimental conditions. Thirty-five out of a total 149 co-expressors of CAPE16 were SSP genes, amounting to 23%, versus 5-9% of the other four CAPE genes. c Pie chart showing the distribution of the CAPE16 co-expressing SSP-coding genes, most of which are Nodule Cys-rich (NCR) peptides, Leginsulins, or Plant Defensins (PDF). ENOD40, Early Nodulin 40; LP, LEED..PEED; nsLTP, non-specific Lipid Transfer Protein; PCY, Plantacyanin; nodGRP, nodule-specific Gly-rich protein; CLE, CLAVATA/Embryo Surrounding Region.


Ab separates the following text:


Ab separates the following text:


Ab separates the following text:


Ab separates the following text:


Ab separates the following text:


Ab separates the following text:


Ab separates the following text:


Ab separates the following text:


Ab separates the following text:


Ab separates the following text:
