bioRxiv preprint doi: https://doi.org/10.1101/2022.02.08.479627; this version posted February 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Title: Reference nodule transcriptomes for *Melilotus officinalis* and *Medicago sativa* cv. Algonquin

Authors: Rui Huang, Wayne A Snedden, George C diCenzo*

Affiliation: Department of Biology, Queen's University, Kingston, Ontario, Canada

*** Corresponding author:** George C. diCenzo
Queen's University, Department of Biology
Biosciences Complex, Room 2433
116 Barrie Street
Kingston, ON, K7P0S7, Canada
george.dicenzo@queensu.ca
+1 (613) 533-6000 x78529

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.08.479627; this version posted February 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

ABSTRACT

Host/symbiont compatibility is a hallmark of the symbiotic nitrogen-fixing interaction between rhizobia and legumes, mediated in part by plant produced nodule-specific cysteine-rich (NCR) peptides and the bacterial BacA membrane protein that can act as a NCR peptide transporter. In addition, the genetic and metabolic properties supporting symbiotic nitrogen fixation often differ between compatible partners, including those sharing a common partner, highlighting the need for multiple study systems. Here, we report high quality nodule transcriptome assemblies for *Medicago sativa* cv. Algonquin and *Melilotus officinalis*, two legumes able to form compatible symbioses with *Sinorhizobium meliloti*. The compressed *M. sativa* and *M. officinalis* assemblies consisted of 79,978 and 64,593 contigs, respectively, of which 33,341 and 28,278 were assigned putative annotations, respectively. As expected, the two transcriptomes showed broad similarity at a global level. We were particularly interested in the NCR peptide profiles of these plants, as these peptides drive bacterial differentiation during the symbiosis. A total of 412 and 308 NCR peptides were predicted from the *M. sativa* and *M. officinalis* transcriptomes, respectively, with approximately 9% of the transcriptome of both species consisting of *NCR* transcripts. Notably, transcripts encoding highly-cationic NCR peptides (isoelectric point > 9.5), which are known to have antimicrobial properties, were ~2-fold more abundant in *M. sativa* than in *M. officinalis*, and ~27-fold more abundant when considering only NCR peptides in the six-cysteine class. We hypothesize that the difference in abundance of highly-cationic NCR peptides explains our previous observation that some rhizobial *bacA* alleles which can support symbiosis with *M. officinalis* are unable to support symbiosis with *M. sativa*.

Keywords: NCR peptides, symbiotic nitrogen fixation, legumes, rhizobia, transcriptomics

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.08.479627; this version posted February 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

INTRODUCTION

Leguminous plants are able to establish symbiotic relationships with a group of soil bacteria known as rhizobia. During the interaction, the rhizobia are located within a specialized organ known as a nodule where they fix atmospheric nitrogen into ammonia in exchange for reduced carbon from their host. Symbiosis is initiated following an exchange of chemical signals in the rhizosphere between compatible partners (1): legumes secrete flavonoids that attract soil rhizobia and induce expression of rhizobial *nod* genes, leading to rhizobial production of chito-oligosaccharide Nod factors that elicit the nodulation process by legumes. This process involves the curling of root hairs to trap rhizobia, and the formation of infection threads within which rhizobia divide and move toward the root cortical layer (2). Rhizobia released from infection threads are endocytosed into the cytoplasm of nodule cells, where they develop into mature N_2-fixing bacteroids. In some legumes, such as those belonging to the Inverted Repeat Lacking Clade (IRLC), the rhizobia undergo an irreversible host-induced process known as terminal differentiation that is largely driven by a unique class of legume proteins known as nodule-specific cysteine-rich (NCR) peptides (3). Terminal differentiation involves cell enlargement, genome endoreplication, and increased membrane permeability, and is thought to increase the efficiency of N_2-fixation (4–6).

Not all rhizobium/legume pairings are compatible (7, 8). Partner compatibility is determined by numerous factors impacting both early and late stages of the symbiotic interaction (9). The flavonoids secreted by legumes vary, as does the ability of rhizobia to respond to different flavonoids (10–13). Similarly, the Nod factor produced by rhizobia differ and legume hosts respond only to Nod factors with specific structures (14). Moreover, legume infection depends on rhizobia producing particular host-compatible exopolysaccharide molecules (15, 16), and variations in exopolysaccharide structure can impact specificity at the level of plant ecotype and bacterial strain (17). In addition, some rhizobia secrete effector proteins that induce effector-triggered immune responses in a cultivar-specific manner, thereby influencing host range (18–20). Moreover, for IRLC legumes, an effective symbiotic interaction requires compatibility between the host-produced NCR peptides and the rhizobial membrane protein BacA (21–24).

NCR peptides are a large class of legume-specific proteins, with ~ 600 members in *Medicago truncatula* (25). These proteins display little conservation in amino acid composition but possess four or six cysteine residues at conserved positions (26). The length of mature NCR peptides varies from about 20 to 50 amino acids and includes two or three disulfide bridges (27). NCR peptides can be classified as either cationic (isoelectric point [pI] $\geqslant 8$), neutral ($6 \leqslant \text{pI} < 8$), or anionic ($\text{pI} < 6$) (27). Highly cationic peptides ($\text{pI} \geq 9.0$) display antimicrobial activity *in vitro*, likely through disrupting microbial membranes, thereby leading to permeabilization and cell lysis (28, 29). *In planta*, NCR peptides are required for rhizobium terminal differentiation and an effective symbiosis in IRLC legumes (3). Deletion of individual *NCR* genes is sufficient to block N_2-fixation (30, 31); however, mutation of other *NCR* genes can result in N_2-fixation in previously incompatible symbioses (21, 22), demonstrating the role of NCR peptides in partner compatibility.

The ability of rhizobia to establish an effective symbiosis with IRLC legumes requires the membrane protein BacA (32). BacA functions as a peptide transporter (33), and *bacA* deletion

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.08.479627; this version posted February 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

mutants are both unable to import NCR peptides and show increased sensitivity to cationic NCR peptides (34–36). In addition, rhizobium *bacA* mutants are unable to fix nitrogen in symbiosis with IRLC legumes; instead, the rhizobia are quickly killed in a NCR peptide-dependent fashion upon release from the infection threads (32, 34). Intriguingly, BacA appears to be a host-range determinant factor in IRLC legumes. For example, studies have shown that introduction of the *bacA* or *bacA*-like genes of *Mesorhizobium loti* and *Bradyrhizobium* species into a *Sinorhizobium meliloti bacA* mutant is insufficient to allow N_2-fixation during interaction with IRLC legumes of the genus *Medicago* (35, 37). Similarly, we previously demonstrated that replacement of the *S. meliloti bacA* with the *bacA* alleles of *Sinorhizobium fredii* NGR234 or *Rhizobium leguminosarum* bv. *viciae* 3841 does not allow for N_2-fixation during symbiosis with *Medicago sativa* (alfalfa) but does support N_2-fixation on the IRLC legumes *Melilotus alba* (white sweet clover) and *Melilotus officinalis* (yellow sweet clover) ((24) and **Table S1**).

In addition to the above-noted comparison, several symbiotic differences have been observed when *S. meliloti* mutants interact with *Medicago* versus *Melilotus* plants (38–41), suggesting that *Melilotus* plants are a valuable secondary model system to study the symbiotic properties of *S. meliloti*. To further develop *M. officinalis* as a model species for studying symbiosis, here we report a reference nodule transcriptome for *M. officinalis*. We further compare the characteristics and the expression of *NCR* genes between *M. officinalis* and *M. sativa* to investigate whether the ability of certain *bacA* alleles to support symbiosis with *Melilotus* but not *Medicago* plants is correlated with differences in the NCR peptide profile of these genera.

MATERIALS AND METHODS

Plant materials and sample collection

M. sativa cv. Algonquin (alfalfa) and *M. officinalis* (yellow blossom sweet clover) seeds were purchased from Speare Seeds Limited (Harriston, Ontario, Canada). Seeds were surface sterilized with 95% ethanol for five minutes followed by 2.5% hypochlorite for 20 minutes, and then soaked in sterile double-distilled water (ddH_2O) for one hour. The sterilized seeds were plated on 1X water agar plates and incubated at room temperature in the dark for two days. Five germinated seeds were placed in autoclaved Leonard Assemblies consisting of two Magenta Jars with a cotton wick extending from the top jar (containing vermiculiate mixed with silica sand [1:1 *w/w*]) into the bottom jar (containing 250 mL of Jensen's media (42)), and then incubated in a Conviron growth chamber for two nights. Wildtype *S. meliloti* strain Rm2011 was grown overnight at 30˚C in LBmc broth (10 g L^{-1} tryptone, 5 g L^{-1} yeast extract, 5 g L^{-1} NaCl, 2.5 mM $CaCl_2$, and 2.5 mM $MgCl_2$), washed with 0.85% NaCl, and diluted to a density of ~ $1x10^7$ CFU mL^{-1} in sterile ddH_2O. Ten mL of cell suspension was then added to each Leonard Assembly. Plants were grown in a Conviron growth chamber with a day (18 hours, 21˚C, light intensity of 300 µmol m^{-2} s^{-1}) and night (6 hours, 17˚C) cycle. Root nodules were collected four weeks post-inoculation and immediately flash frozen with liquid N_2 and stored at –80˚C until use. All nodules collected from plants grown in the same Leonard Assembly were stored in a single tube and treated as one replicate. The shoots from each pot were dried at 60˚C for two weeks prior to measuring shoot dry weight (**Table S2**).

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.08.479627; this version posted February 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

RNA extraction and sequencing

Total RNA from three replicates of frozen *M. sativa* and *M. officinalis* nodule tissue was extracted using Direct-zol RNA miniprep kits (ZYMO Research) according to the manufacturer's protocol. Total RNA samples were treated with DNase I (New England Biolabs) to degrade any contaminating DNA according to the manufacturer's protocol, and the RNA again purified using Direct-zol RNA miniprep kits. Total RNA samples were run on a MOPS-formaldehyde agarose gel (119 mL MOPS buffer [200 mM MOPS, 80 mM sodium acetate, 10 mM EDTA, pH 7.0, in DEPC-treated ddH_2O], 6 mL formaldehyde, 1.25 g agarose) to check the integrity of the RNA (**Figure S1**), and subsequently verified using an Agilent Bioanalyzer chip.

Library preparation and Illumina sequencing were performed at The Centre for Applied Genomics at The Hospital for Sick Children (Toronto, Ontairo, Canada). Libraries were prepared using the NEB Next® Ultra™ II Directional RNA Library Prep Kit for Illumina®. Libraries were then sequenced using one lane of a high throughput flow cell on an Illumina HiSeq 2500 platform, generating 125 bp paired-end reads.

Transcriptome *de novo* assembly and quality control

The nodule transcriptomes of *M. sativa* and *M. officinalis* were *de novo* assembled following the same procedure. First, reads from the triplicate samples were combined, and then preprocessing of the raw reads was performed to ensure only high-quality data was used for *de novo* transcriptome assembly. Read quality was initially evaluated using FastQC version 0.11.9 (43), following which errors in raw reads were identified and corrected by the k-mer based method of Rcorrector version 1.0.4 (44). The FilterUncorrectablePEfastq.py script (github.com/harvardinformatics/TranscriptomeAssemblyTools/) was used to remove any read pair where at least one read had an unfixable error identified by Rcorrector. Adaptors sequences, short reads (< 25 bp), and low-quality reads (Q score < 20) were removed using Trim Galore version 0.6.6 (bioinformatics.babraham.ac.uk/projects/trim_galore/), which is a wrapper calling cutadapt version 3.2 (45) and FastQC (**Table 1**). The processed reads were further trimmed by Trimmomatic version 0.4.0 (46) included in the Trinity software distribution with the following parameters: *SLIDINGWINDOW:5:20 LEADING:3 TRAILING:3 MINLEN:25*. The quality and presence of adaptors in the preprocessed reads were then examined using FastQC. Following preprocessing, 174,707,055 and 119,333,821 paired end reads (~43.7 and ~29.8 Gb, respectively) remained for *M. sativa* and *M. officinalis*, respectively (**Table S3**).

The preprocessed reads were assembled using Trinity version 2.9.0 without genome guidance (47). Then, the assembled contigs were clustered into gene-level clusters using SuperTranscripts (48). Gene isoforms were identified by Corset version 1.09 with the log likelihood ratio threshold set to very high (49). Based on the Corset clusters, Lace version 1.14.1 was used to merge the gene isoforms into single long supertranscripts meant to provide a gene-like view of the transcriptome (48).

Multiple methods were used to examine the quality of the Trinity and SuperTranscript assemblies. First, the alignment rates of the preprocessed reads to the assemblies were inspected

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.08.479627; this version posted February 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

using STAR version 2.7.8a with the two-pass mode that is more sensitive to alternative splicing (50). Second, assembly statistics such as N50 and number of contigs were calculated using the seqstats software (github.com/clwgg/seqstats). Third, the completeness of the assembies was evaluated using BUSCO version 5.1.2, run separately using the OrthoDB v10 'Fabales' and 'Viridiplantae' reference databases (51). The assemblies were also compared to the *S. meliloti* Rm2011 genome (52) using BLASTn version 2.5.0+ (53), which confirmed the absence of contaminating *S. meliloti* transcripts in the assemblies. Finally, the *M. sativa de novo* assembly was aligned to a publicly-available genome of *M. sativa* cultivar XinJiangDaYe (54) with MUMmer version 4.0+, and 87.3% of transcripts were sucessfully aligned to the genome.

Transcriptome annotation

Coding regions within the supertranscripts were predicted by TransDecoder version 5.5.0 (github.com/TransDecoder/TransDecoder), using the results of BLASTp searches (E-value cutoff of 1e-5) against the Uniport database as ORF retention criteria (2021 January release) (55). The functional annotation of the predicted coding sequences then proceeded via three steps. First, BLAST bidirectional best hits between the *M. truncatula* A17 proteome (assembly release r5.0 1.7) (56) and the longest predicted protein isoform of each contig of our transcriptome assemblies were identified using BLASTp (E-value cutoff of 1e-5, culling limit 1). For all bidirectional best hits, the annotations from *M. truncatula* A17 were transferred to the corresponding contigs of the *M. sativa* or *M. officinalis* transcriptome. Second, all predicted protein isoforms of each contig in each transcriptome assembly were annotated using eggNOG-mapper version 2.1.0 with DIAMOND version 2.0.4 and the Viridiplantae dataset (E-value cutoff of 1e-3) (57, 58). Third, for each conting not annotated by BLAST or eggNOG-mapper, the hmmsearch function of HMMER version 3.3.2 was used to search all predicted protein isoforms against the complete set of hidden Markov models (HMMs) from the Pfam version 34.0 database and separately against the TIGRFAM version 15.0 HMM database (E-value cutoff of 1e-5) (59–61), and results were filtered to remove annotations with a Bit-score < 50. For repetitive annotations from isoforms of a gene, only the consensus annotations were retained. For contigs successfully annotated by more than one of the annotation methods, results from the bidirectional BLAST took priority, followed by the results of eggNOG-mapper, then the Pfam searches, and finally the TIGRFAM searches.

NCR peptide identification

Considering the high degree of sequence diversity of NCR peptide sequences, the functional annotation methods described above were not sufficiently sensitive to discover genes encoding NCR peptides in the assemblies. Therefore, the SPADA version 1.0 pipeline was used to identify NCR peptides (62). SPADA is specialized to predict cysteine-rich peptides in plant genomes and is distributed with a *M. truncatula* prediction model. Cysteine-rich peptides in the *M. sativa* and *M. officinalis* assemblies were predicted using the SPADA pipeline with following software: HMMER version 3.0, Augustus version 2.6, GeneWise version 2.2.0, GeneMark.hmm eukaryotic version 3.54, GlimmerHMM version 3.0.1, and GeneID version 1.1 (63–66). The putative NCR

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.08.479627; this version posted February 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

peptide sequences were filtered to remove those without a signal peptide, and then further filtered based on the E-value (cutoff of 1e-5) and hmm score (cutoff of 50). Filtered sequences were then verified via hmmscan searches against the Pfam database, and they were aligned using Clustal Omega version 1.2.4 (67) to ensure the presence of the signature cysteine motif and N terminal signal peptide that are present in *bona fide* NCR peptides.

NCR peptides classification

To predict the lengths of mature NCR peptides, signalP version 4.1g with the notm network was used to predicted cleavage sites and extract mature NCR peptides (68), and the number of cysteine residues in each motif were counted. The pI values of the NCR peptides were predicted using the pIR R package, and the value for each peptide was calculated based on the mean values from all prediction methods excluding the highest and lowest values (69).

Gene-expression level estimation and differential expression analysis

Gene-expression levels of each *M. sativa* and *M. officinalis* replicate transcriptome were estimated by transcript abundance estimation using salmon version 0.12.0 in mapping-based mode (library type automatic, validate Mapping) (70) and the reference transcriptomes produced as described above. R package deseq2 version 1.32.0 (71) was used to perform differential expression analysis between *M. sativa* and *M. officinalis*, using the raw counts from salmon, the length of each gene in each species as an additional parameter during normalization, and limiting the analysis to one-to-one orthologs identified by OrthoFinder version 2.5.2 (72). OrthoFinder was run with default settings using the total predicted *M. sativa* and *M. officinalis* proteomes including all isoforms, following which orthologs were reduced to one per supertransript.

Gene Ontology term analysis

The Gene Ontology (GO) terms for *M. sativa* and *M. officinalis* were obtained from the *M. truncatula* A17 proteome (assembly release r5.0 1.7) and annotations from eggNOG-mapper. For transcripts annotated with GO terms from both sources, the concensus GO term annotations were retained. Then, the GO terms were reduced based on the Generic GO subset (download 10 August 2021).

Software information

All analyses were performed in an Ubuntu 20.04.2 LTS (Linux 5.8.0-48-generic) operation system or on the Compute Canada Graham cluster. Custom scripts were written in Python version 3.8.5 and bash. R version 3.6.3 was used during data analysis (73).

Data availability

All custom scripts to perform the analyses described in this study are available through GitHub (https://github.com/hyhy8181994/Nodule_transcriptome_script). Raw Illumina data are available through the Short Read Archive (SRR15724671, SRR15724670, SRR15724669, SRR15724668,

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.08.479627; this version posted February 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

SRR15724667, and SRR15724666) hosted by the National Center for Biotechnology Information (NCBI). The assembled transcriptomes are available through the Transcriptome Shotgun Assembly Sequence Database (GJLW00000000 and GJLK00000000) hosted by the NCBI.

RESULTS AND DISCUSSION

Reference nodule transcriptomes for *Melilotus officinalis* and *Medicago sativa* cv. Algonquin

To establish reference nodule transcriptomes of *M. sativa* cv. Algonquin and *M. officinalis* during symbiosis with *S. meliloti* Rm2011, the poly-A enriched RNA from triplicate samples was sequenced using Illumina technology (2x125 bp paired-end reads), generating ~50 Gb (~ 202 million paired-end reads) and ~35 Gb (~139 million paired-end reads) of data for *M. sativa* and *M. officinalis*, respectively (see **Table S3** for sequencing statistics). *De novo* assembly of the *M. sativa* sequencing data resulted in 253,871 contigs, while 192,165 *de novo* assembled contigs were produced for *M. officinalis*. Contigs expected to represent splice variants of a single gene were merged into so-called "supertranscripts" using the SuperTranscripts program, resulting in compressed assemblies of 79,978 and 64,593 contigs for *M. sativa* and *M. officinalis*, respectively (**Table 1**). Transcriptomes were annotated as described in the Materials and Methods, resulting in putative annotations for 33,431 *M. sativa* contigs and 28,278 *M. officinalis* contigs (**Datasets S1 and S2**). Of these, ~ 52% (*M. sativa*) and ~ 58% (*M. officinalis*) are high confidence annotations as they were transferred from the *M. truncatula* whole genome annotation following identification of putative orthologs using a BLAST bidirectional best hit approach (**Table S4**). Considering that previous studies have predicted the presence of ~23,000 long non-coding RNAs (lncRNAs) in *M. truncatula* (74) and ~47,000 lncRNAs in the legume *Pisum sativum* (pea) (75), we hypothesize that the majority of the unannotated *M. sativa* and *M. officinalis* transcripts reflect lncRNAs.

All of the examined assembly summary statistics (mean and median contig length, contig N50) were improved in the compressed assemblies compared to the original *de novo* assemblies, indicating that the compressed assemblies are of higher structural quality (**Table 1**). The *M. sativa* transcriptome summary statistics, such as N50 and and transcript length, are consistent with those reported for other *M. sativa de novo* transcriptome assemblies, although the number of transcripts varies likely due to each study examining different tissues (76, 77). In addition, the assemblies appear to be robust; greater than 90% of the filtered reads used for transcriptome assembly could be mapped to the corresponding assemblies by STAR (**Table 1**). Moreover, > 92% and > 83% of the Viridiplantae and Fabales BUSCO marker genes, respectively, were identified as complete and single-copy in the *M. sativa* and *M. officinalis* compressed assemblies (**Figure 1**). The structural quality (e.g. high average and median contig length and N50) and BUSCO benchmark scores described here are in line with those reported for other plant *de novo* transcriptome assemblies (78–80). Taken together, these results indicate that our *M. sativa* cv. Algonquin and *M. officinalis* reference nodule transcriptomes are reliable and of high quality.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.08.479627; this version posted February 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Table 1. Summary statistics from the *de novo* Trinity and compressed (SuperTranscripts) nodule transcriptome assemblies.

	M. sativa		*M. officinalis*	
	Trinity Assembly	SuperTranscripts	Trinity Assembly	SuperTranscripts
Total number of contigs	253,871	79,978	192,165	64,593
Total number of base pairs (bp)	239,421,306	96,211,060	191,200,637	81,371,534
Average contig length (bp)	943	1203	995	1260
Median contig length (bp)	618	734	646	780
Contig N50 (bp)	1,466	1,912	1,584	1968
Minimum contig length (bp)	176	193	183	198
Maximum contig length (bp)	12,655	29,784	14,584	23,941
Overall alignment rate (%)	98.99	92.14	99.15	93.76

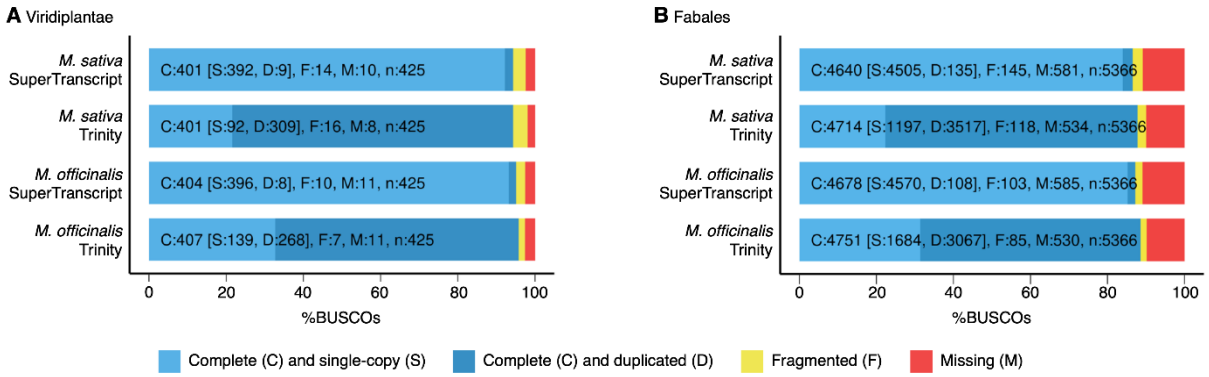

Figure 1. Estimates of nodule transcriptome completeness. Completeness of the *M. sativa* and *M. officinalis* nodule transcriptome assemblies was assessed using BUSCO with the (**A**) Viridiplantae and (**B**) Fabales single-copy marker gene datasets. The fraction of BUSCO genes identified as complete and single-copy (light blue), complete but duplicated (dark blue), fragmented (yellow), and missing (red) is shown.

Comparative transcriptome analysis between *M. sativa* and *M. officinalis*

As an initial examination of the *M. sativa* and *M. officinalis* transcriptomes, the annotated functions of the proteins predicted to be encoded by the supertranscripts were summarized using the Generic GO term subset (**Figure 2, Dataset S3 and S4**). Approximately 18,363 (26.1%) of the *M. sativ*a supertranscripts and 16,674 (28.6%) of the *M. officinalis* supertranscripts were annotated with GO terms. No significant difference in the GO term profiles of the two species was observed, with the five most frequently annotated biological process GO terms being GO:0008150 (biological process), GO:0006950 (response to stress), GO:0006464 (cellular protein modification process), GO:003464 (cellular nitrogen compound metabolic process) and GO:0048856 (anatomical structure development). At this broad scale, the GO term data suggest that there is substantial similarity in the nodule transcriptomes of *M. sativ*a and *M. officinalis*.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.08.479627; this version posted February 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

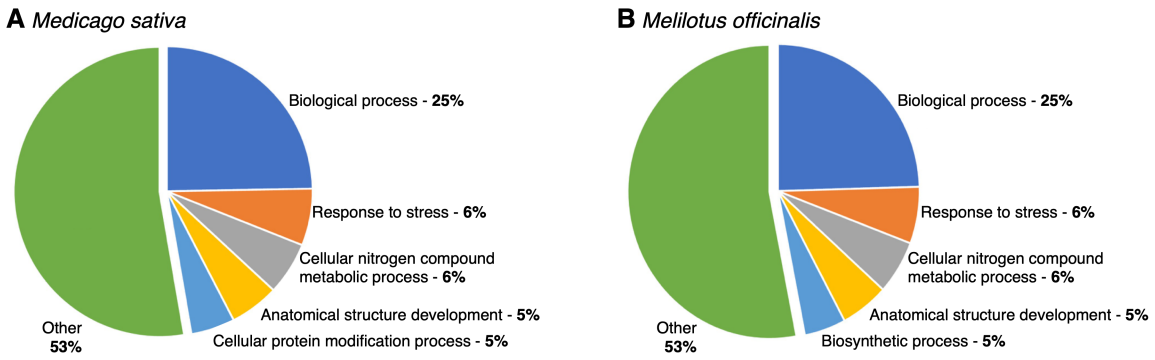

Figure 2. Summary of the Slim GO Biological Processes annotations for the nodule transcriptomes. Transcripts were annotated with Slim GO terms, and the annotations for the biological processes were summarized as pie charts for (**A**) *M. sativa* and (**B**) *M. officinalis*.

We next examined the predicted functions of the proteins encoded by the 50 most abundant transcripts in both species (**Table 2 and 3**). Not surprisingly, these transcripts were enriched in those predicted to encode nodulins and leghaemoglobin-like proteins. Nodulins refer to diverse proteins expressed specifically in nodule tissue, which play various structural or metabolic roles during symbiotic nitrogen fixation. Among the nodulins, are the leghemoglobin proteins that account for up to 40% of the total soluble protein in legume nodules (81). Leghemoglobins play an important role in maintaining the low free-oxygen concentration required to protect the oxygen-sensitive nitrogenase enzyme (82).

To facilitate further comparison of the *M. sativa* and *M. officinalis* transcriptomes, the proteins predicted to be encoded by the supertranscripts of both species were arranged into orthologous groups using OrthoFinder. A total of 20,237 orthologous groups, accounting for 26,304 *M. sativa* and 24,895 *M. officinalis* supertranscripts, were identified. Interestingly, the abundance of the conserved supertranscripts was significantly higher, on average, than that of the species-specific transcripts (p value < 2.2e-16; **Figure 3**). In both plant species, the majority of the most abundant, species-specific annotated transcripts were also nodulins, globin family proteins that are likely species-specific leghaemoglobin isoforms, and some housekeeping genes such as ribonuclease and ribosomal proteins. It is noteworthy that the most abundant *M. sativa*-specific supertranscript is predicted to encode albumin I. Similarly, *M. officinalis* also has a highly-expressed albumin I supertranscript. The albumin I peptide family is known to be highly expressed in legume seeds and play roles in seed protection (83). Expression of albumin I genes has also been observed in *M. truncatula* root nodules, with expression specific to uninfected cells in the nitrogen fixation zone (84). These cells are thought to play essential roles in metabolite transport during symbiosis, and albumin I may have a role in protecting some of the nodule cells from rhizobium infection (84). A phylogenetic analysis of *M. truncaula* nodulins and albumin I peptides indicated that the *M. truncaula* albumin I clustered with a subset of nodulins, reflecting a close evolutionary relationship between these proteins (85).

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.08.479627; this version posted February 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Table 2. The 50 most highly abundant transcripts in the *M. sativa* nodule transcriptome, with the average expression level in transcripts per million (TPM) and the functional annotation.

Gene ID	TPM	Functional Prediction
Cluster-2.43518	30,662	hypothetical protein (hypothetical leghaemoglobin)
Cluster-2.26078	17,769	Putative albumin I
Cluster-2.23447	12,515	hypothetical protein (hypothetical leghaemoglobin)
Cluster-2.23176	8,065	Nodulin-25
Cluster-2.22272	7,992	Putative Late nodulin
Cluster-2.23033	6,778	None
Cluster-2.21873	5,919	hypothetical protein
Cluster-2.22935	5,253	Belongs to the globin family
Cluster-2.33070	5,232	Putative ribonuclease H-like domain-containing protein
Cluster-2.24197	4,245	Component of the replication protein A complex (RPA)
Cluster-2.22983	4,196	Belongs to the globin family
Cluster-2.24546	3,886	Predicted NCR peptide (crp1450_Cluster-2.24546_0M_1)
Cluster-2.19511	3,344	None
Cluster-2.26207	3,296	Extensin-like_protein_repeat
Cluster-2.49512	3,173	Putative Blue (type 1) copper binding protein
Cluster-2.21810	3,050	Predicted NCR peptide (crp1160_Cluster-2.21810_0M_1)
Cluster-2.22936	3,014	Belongs to the globin family
Cluster-2.29430	2,789	Putative Late nodulin
Cluster-2.22881	2,788	Predicted NCR peptide (crp1430_Cluster-2.22881_0M_1)
Cluster-2.22836	2,509	None
Cluster-2.23729	2,271	hypothetical protein
Cluster-2.22458	2,245	Belongs to the globin family
Cluster-2.24829	2,227	hypothetical protein
Cluster-2.25168	2,209	None
Cluster-2.24278	2,093	Nodule-specific_GRP_repeat
Cluster-2.23928	2,038	Late_nodulin_protein
Cluster-2.22042	2,029	None
Cluster-2.23245	2,019	Putative translationally controlled tumor protein
Cluster-2.25794	2,018	Putative protein-synthesizing GTPase
Cluster-2.31376	1,957	Predicted NCR peptide (crp1190_Cluster-2.31376_0M_1)
Cluster-2.21809	1,939	Predicted NCR peptide (crp1160_Cluster-2.21809_0M_1)
Cluster-2.28083	1,854	Predicted NCR peptide (crp1210_Cluster-2.28083_0M_1)
Cluster-2.23524	1,844	Early nodulin-16
Cluster-2.34649	1,839	Putative Late nodulin
Cluster-2.16993	1,808	hypothetical protein
Cluster-2.21813	1,731	None
Cluster-2.22457	1,705	Belongs to the globin family
Cluster-2.28283	1,674	None
Cluster-2.26205	1,621	Predicted NCR peptide (crp1240_Cluster-2.26205_0M_1)
Cluster-2.23310	1,607	asparagine synthetase
Cluster-2.21870	1,596	Nodule-specific_GRP_repeat
Cluster-2.19604	1,583	Predicted NCR peptide (crp1420_Cluster-2.19604_0M_1)
Cluster-2.18485	1,509	hypothetical protein
Cluster-2.26876	1,458	Predicted NCR peptide (crp1410_Cluster-2.26876_0M_1)
Cluster-2.21536	1,441	Ubiquitin_family
Cluster-2.22785	1,408	None
Cluster-2.23374	1,385	Predicted NCR peptide (crp1420_Cluster-2.23374_0M_1)
Cluster-2.28787	1,311	Putative Late nodulin
Cluster-2.22402	1,292	Predicted NCR peptide (crp1520_Cluster-2.22402_0M_1)
Cluster-2.30081	1,289	Late_nodulin_protein

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.08.479627; this version posted February 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Table 3. The 50 most highly abundant transcripts in the *M. officinalis* nodule transcriptome, with the average expression level in transcripts per million (TPM) and the functional annotation.

Gene ID	TPM	Functional Prediction
Cluster-3554.18801	38,953	Belongs to the globin family
Cluster-3554.18778	14,091	Belongs to the globin family
Cluster-3554.16063	10,412	Late_nodulin_protein
Cluster-3554.15387	7,885	Putative Late nodulin
Cluster-3554.16088	6,146	Putative Late nodulin
Cluster-3554.18892	6,104	None
Cluster-3554.15771	5,813	Late_nodulin_protein
Cluster-3554.18802	5,362	Belongs to the globin family
Cluster-3554.18596	5,347	Predicted NCR peptide (crp1430_Cluster-3554.18596_0M_1)
Cluster-3554.15456	3,912	Putative translationally controlled tumor protein
Cluster-3554.18808	3,864	Belongs to the globin family
Cluster-3554.33215	3,302	hypothetical protein
Cluster-3554.18775	3,097	Putative Late nodulin
Cluster-3554.23000	2,990	Two predicted NCR peptide (crp1180_Cluster-3554.23000_0M_1 and crp1180_Cluster-3554.23000_0M_2)
Cluster-3554.36681	2,877	Predicted NCR peptide (crp1500_Cluster-3554.36681_0M_1)
Cluster-3554.21555	2,868	Predicted NCR peptide (crp1430_Cluster-3554.21555_0M_1)
Cluster-3554.16074	2,809	Putative BURP domain-containing protein
Cluster-3554.29297	2,784	hypothetical protein
Cluster-3554.18196	2,723	None
Cluster-3554.18256	2,571	Predicted NCR peptide (crp1440_Cluster-3554.18256_0M_1)
Cluster-3554.27063	2,522	None
Cluster-3554.22577	2,449	Putative Blue (type 1) copper binding protein
Cluster-3554.15361	2,409	eEF1A
Cluster-3554.21311	2,384	Belongs to the globin family
Cluster-3554.18838	2,340	Late_nodulin_protein
Cluster-3554.18700	2,296	Late_nodulin_protein
Cluster-3554.11713	2,261	None
Cluster-3554.18706	2,259	None
Cluster-3554.23445	2,214	None
Cluster-3554.17366	2,141	Predicted NCR peptide (crp1430_Cluster-3554.17366_0M_1)
Cluster-3554.23502	2,073	hypothetical protein
Cluster-3554.25153	2,009	Metallothionein-like protein 2
Cluster-3554.21451	1,953	Belongs to the globin family
Cluster-3554.28600	1,931	hypothetical protein
Cluster-3554.22971	1,849	hypothetical protein
Cluster-3554.18877	1,812	Belongs to the glyceraldehyde-3-phosphate dehydrogenase family
Cluster-3554.13172	1,732	Predicted NCR peptide (crp1440_Cluster-3554.13172_0M_1)
Cluster-3554.18195	1,704	Zinc_knuckle
Cluster-3554.30751	1,653	Putative Late nodulin
Cluster-3554.21774	1,635	Late_nodulin_protein
Cluster-3554.13784	1,629	Predicted NCR peptide (crp1420_Cluster-3554.13784_0M_1)
Cluster-3554.24033	1,611	Prolyl isomerase (PPIase)
Cluster-3554.30294	1,556	metallothionein-like protein
Cluster-3554.24764	1,552	None
Cluster-3554.18368	1,552	Nucleoside diphosphate kinase 1
Cluster-3554.25344	1,537	Metallothionein-like protein 1
Cluster-3554.23646	1,514	None
Cluster-3554.9874	1,512	Belongs to the universal ribosomal protein uL13 family
Cluster-3554.18476	1,503	Late_nodulin_protein
Cluster-3554.31722	1,476	Putative Late nodulin

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.08.479627; this version posted February 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

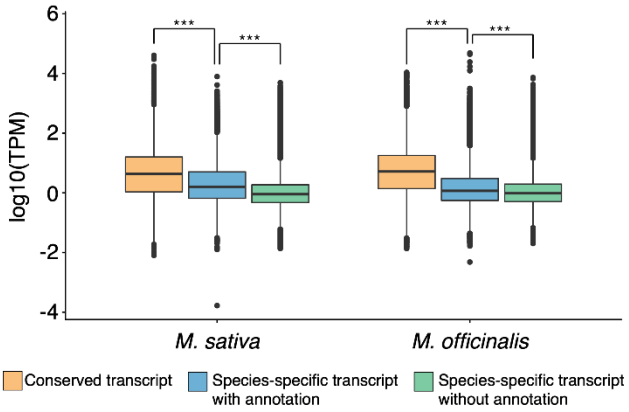

Figure 3. Transcript abundances for conserved and species-specific transcripts. Box plots displaying the distribution of average transcript abundances from triplicate samples, shown separately for genes with orthologs in both *M. sativa* and *M. officinalis* (orange), annotated transcripts found in only *M. sativa* or *M. officinalis* (blue), or transcripts that lack annotations and are found in only *M. sativa* or *M. officinalis* (green). Statistically significant differences between the distributions of a species are indicated with the asterisks (p-value < $1e^{-10}$; pairwise Wilcox tests).

We next compared the abundances of supertranscripts conserved in both *M. sativa* and *M. officinalis*, limiting the analysis to the 15,287 one-to-one orthologs detected by OrthoFinder. Desipite significant variation in the abundance of orthologous transcripts between *M. sativa* and *M. officinalis* – which may reflect limitations of inter-species transcriptome analysis – a clear correlation in the abundance of orthologous transcripts was detected (residual standard error = 0.517; **Figure 4**). Considering the limtations of inter-species differential expression analyses, we restricted our investigation to supertranscripts with absolute $\log_2$ fold changes > 5 and a p-value < 0.05. Using these thresholds, we identified 290 differentially-abundant transcripts, 86 of which were more abundant in *M. sativa*, and 204 of which were more abundant in *M. officinalis*. It should be noted, however, that only 159 of the differentially-abundant transcripts were annotated with the same or similar function in both species, and we focus on these 159 transcripts in the following discussion.

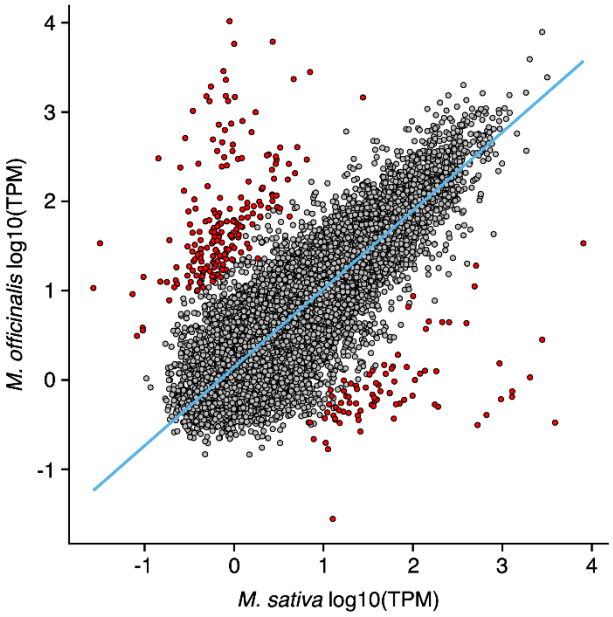

Figure 4. Correlation between transcript abundances of orthologous transcripts in *M. sativa* and *M. officinalis*. Each datapoint represents the transcript abundance of single-copy orthologous transcripts in *M. sativa* and *M. officinalis*. Red datapoints represent transcripts that are differentially abundant between the two species ($|\log_2(\text{fold change})| > 5$, adjusted p-value < 0.01); all other datapoints are in grey. The blue line represents the robust linear regression line, calculated with the rlm function of the MASS package in R.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.08.479627; this version posted February 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Many of the differentially-abundant conserved supertranscripts have annotated functions that suggest the encoded proteins may impact symbiotic nitrogen fixation. These include 21 supertranscripts annotated as encoding nodulins, which include 16 that are more abundant in *M. officinalis* and 5 that are more abundant in *M. sativa*. In addition, 32 supertranscripts encoding proteins predicted to be associated with transcription and translation activity were differentially abundant, with 24 more abundant in *M. sativa* and eight more abundant in *M. officinalis*. We also observed that several supertranscripts encoding proteins predicted to be involved in cell wall synthesis or modification were differentially abundant, with six more highly abundant in *M. sativa* and one more highly abundant in *M. officinalis*. Other differentially-abundant transcripts included those predicted to encode proteins involved in transport (17 transcripts), fatty acid biosynthesis (3 transcripts), flavonoid biosynthesis (3 transcripts), and aromatic compound biosynthesis (1 transcript). Given that this analysis compares two plant species with differing growth rates (**Table S2**), we cannot rule out that some of these transcriptomic differences may also reflect variances in nodule maturity and/or host metabolic activity at the time of harvest.

NCR peptide diversity and expression profile

We previously observed that replacing the *bacA* allele of *S. meliloti* 2011 with the *bacA* alleles of the rhizobia *S. fredii* NGR234 or *R. leguminosarum* bv. viciae 3841 resulted in an inability to fix nitrogen with *M. sativa* while the ability to fix nitrogen with *M. alba* and *M. officinalis* remained ((24) and **Table S1**). We hypothesized that this was due to differences in the NCR peptide profiles of these species (24). To test this hypothesis, supertranscripts encoding NCR peptides were identified in the *M. sativa* and *M. officinalis* transcriptome assemblies using the SPADA pipeline (62). A total of 412 and 308 supertranscripts encoding NCR peptides were identified in the *M. sativa* and *M. officinalis* transcriptomes, respectively, accounting for ~0.5% of all supertranscripts in both assemblies (**Datasets S5 and S6**). The lower count of *NCR* transcripts in *M. officinalis* was offset by a higher median transcript abundance (58.6 transcripts per million [TPM] vs 99.1 TPM; $p < 0.001$; **Figure 5A**), resulting in *NCR* transcripts accounting for roughly 9% of the total nodule transcriptome in both species. In both *M. sativa* and *M. officinalis*, NCR peptides had median lengths of 38 residues, with approximately half of the NCR peptides containing between 30 and 40 residues (**Figure 5B**). Additionally, there was a roughly even number of four and six-cysteine NCR peptides expressed in both plant species, with the four-cysteine class of NCR peptides accounting for 51-55% of the *NCR* transcripts both in terms of number of NCR peptides and expression of *NCR* transcripts as measured by TPM. The NCR peptides from both hosts also showed broadly similar distributions of pI values between approximately 3 to 11, with one peak around a pI of 4 and another around pI 8 (**Figures 5C and 5D**). The pI pattern of the NCR peptides we observed is reminiscent of that reported for other legume species that induce an elongated branched morphology in their microsymbiont, including *M. sativa* and *M. truncatula* (86). Overall, at a global level, the property profiles of NCR peptides for *M. sativa* and *M. officinalis* were very similar, suggesting that the impact of different *bacA* alleles on symbiotic compatibility of *S. meliloti* with *M. sativa* is unlikely a consequence of global differences in the NCR peptide profiles

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.08.479627; this version posted February 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

of these plants and is more likely due to specific NCR peptides. Identifying which NCR peptides functionally correlate with symbiotic compatibility should be the focus of future studies.

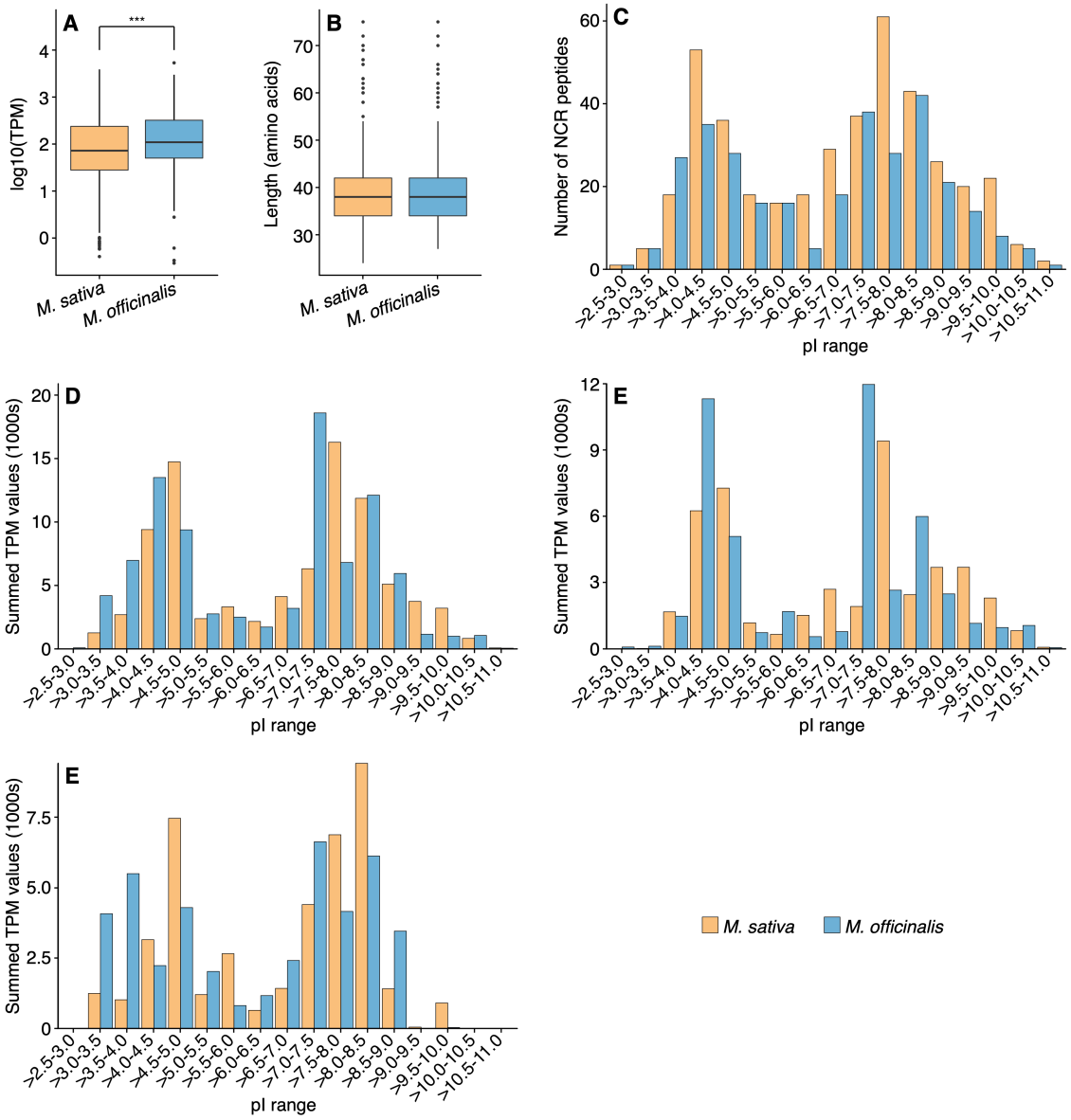

Figure 5. NCR peptide profiles of *Medicago sativa* and *Melilotus officinalis*. NCR peptides were predicted from the *M. sativa* (orange) and *M. officinalis* (blue) transcriptome assemblies, and the properties of the NCR peptides are shown in these graphs. (**A**) Box plots showing the distribution of the abundance (in transcripts per million, TPM) of *NCR* transcripts, based on triplicate samples. The difference in the distributions for the two species was statistically significant (p-value < 0.001; pairwise Wilcox test). (**B**) Box plots showing the distribution of the amino acid lengths of mature NCR peptides. No statistically significant difference in the distributions for the two species was detected. (**C**,**D**) Histograms showing the distributions of the isoelectric points (pI) for the mature NCR peptides. Histograms are based either on the number of NCR peptides with a given pI value (C) or the total abundance of the transcripts encoding NCR peptides with a given pI value (D). **(E,F)** Histograms showing distributions of pI for 4-cysteines (E) and 6-cysteines (F) mature NCR peptides based on total abundance of the transcripts encoding NCR peptides with a given pI value.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.08.479627; this version posted February 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Despite the general similarity in the NCR peptide profiles of *M. sativa* and *M. officinalis*, a key difference emerges when examining the abundance of NCR peptides with extreme pI values; transcripts encoding highly cationic NCR peptides were more abundant in *M. sativa* while transcripts encoding highly anionic NCR peptides were more abundant in *M. officinalis* (**Figure 5D**). Previous work has shown that, in general, only cationic NCR peptides with a pI > 9.0 have antimicrobial activity (87), with anticandidal activity primarily limited to NCR peptides with a pI > 9.5 (88). Here, we observed that transcripts encoding highly cationic NCR peptides (pI > 9.0) were ~2.4-fold more abundant in *M. sativa* than *M. officinalis* (**Figure 5D**). Similarly, transcripts encoding NCR peptides with pI values > 9.5 were ~1.9-fold more abundant in *M. sativa* than *M. officinalis*. Notably, previous work indicated that 4.0% of *M. truncatula NCR* transcripts encode NCR peptides with pI values > 9.5, compared to only 1.8% in the *R. leguminosarum* bv. *viciae* symbiont *P. sativum* (86); this compares to 4.7% and 2.3% for *M. sativa* and *M. officinalis*, respectively (**Figure 5C**). Strikingly, when subdividing the NCR peptides with pI values > 9.5 into those with four or six cysteine residues, we observed that those with six-cysteines were ~27-fold more abundant in *M. sativa* than *M. officinalis* (**Figure 4E and 4F**). Considering these results, we hypothesize that the ability of the *R. leguminosarum bacA* allele to support symbiosis with *M. officinalis* and *P. sativum*, but not *M. sativa*, is a consequence of the elevated abundance of highly cationic (pI > 9.5) NCR peptides in *Medicago* nodules, with six-cysteine NCR peptides possibly being of particular significance. It may be that the BacA proteins of *S. fredii* and *R. leguminosarum* are less capable of transporting these NCR peptides, and consequently, strains with these BacA proteins may be more sensitive to the antimicrobial activities of these cationic NCR peptides.

CONCLUSION

We report high quality nodule transcriptome assemblies for *M. sativa* cv. Algonquin and *M. officinalis* that we expect will serve as valuable resources for the legume research commumity. In particular, we expect that the availability of a nodule transcriptome for *M. officinalis* will help establish this plant as a secondary model system for studies of the symbiotic properties of *S. meliloti*.

We were particularly interested in using these transcriptomes to compare the properties of the NCR peptides encoded by both species. Despite predicting 33% more NCR peptides in *M. sativa* than *M. officinalis*, *NCR* transcripts accounted for roughly 9% of the transcriptome (based on TPM values) in both species. In general, the characteristics of the NCR peptides of *M. sativa* and *M. officinalis* were highly similar. However, transcripts encoding cationic NCR peptides with a pI > 9.5 were ~2-fold more abundant in *M. sativa* than in *M. officinalis*, and 27-fold more abundant when considering only six-cysteine NCR peptides. These results are consistent with previous observations that transcripts encoding cationic NCR peptides with a pI > 9.5 account for ~2-fold more *NCR* transcripts in *M. truncatula* compared to *P. sativum*. Cationic, but not neutral or anionic, NCR peptides display antimicrobial activity through disrupting the integrity of microbial membranes (89). It has been hypothesized that BacA provides protection against these

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.08.479627; this version posted February 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

NCR peptides by importing them into the cytoplasm and thus away from the membrane (90, 91). Considering that the BacA proteins of *S. fredii* and *R. leguminosarum* share less than 60% amino acid identity with the BacA protein of *S. meliloti*, it is reasonable to speculate that they have different substrate specificity and may be less capable of transporting cationic NCR peptides (24). If true, this could explain why the *bacA* alleles of *S. fredii* and *R. leguminosarum* can support symbiotic nitrogen fixation with *M. officinalis* but not *M. sativa*; the increased production of cationic NCR peptides in *M. sativa*, coupled with lower rates of import into the *S. meliloti* cytoplasm, could result in an accumulation of these peptides in the periplasm, resulting in a loss of viability and lack of nitrogen fixation (24). In future work, it will be interesting to test whether *S. meliloti* strains with different *bacA* alleles display differing sensitivities to these highly cationic NCR peptides, or differences in their abilities to transport these peptides.

ACKNOWLEDGEMENTS

We thank Karen Ho and Neda Moradin from The Centre for Applied Genomics (Toronto, Canada) for helpful advice in planning the RNA-seq library preparation strategy. This research was enabled, in part, through computational resources provided by Compute Ontario (computeontario.ca) and Compute Canada (computecanada.ca). Funding for this research was provided by the Natural Sciences and Engineering Research Council of Canada (NSERC) through Discovery Grants to WAS and GCD.

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no conflict of interest.

REFERENCES

1. Oldroyd GED. 2013. Speak, friend, and enter: signaling systems that promote beneficial symbiotic associations in plants. Nat Rev Microbiol 11:252–263.

2. Gage DJ. 2004. Infection and invasion of roots by symbiotic, nitrogen-fixing rhizobia during nodulation of temperate legumes. Microbiol Mol Biol Rev 68:280–300.

3. Van de Velde W, Zehirov G, Szatmari A, Debreczeny M, Ishihara H, Kevei Z, Farkas A, Mikulass K, Nagy A, Tiricz H. 2010. Plant peptides govern terminal differentiation of bacteria in symbiosis. Science (80-) 327:1122–1126.

4. Mergaert P, Uchiumi T, Alunni B, Evanno G, Cheron A, Catrice O, Mausset A-E, Barloy-Hubler F, Galibert F, Kondorosi A. 2006. Eukaryotic control on bacterial cell cycle and differentiation in the *Rhizobium*–legume symbiosis. Proc Natl Acad Sci U S A 103:5230–5235.

5. Lamouche F, Bonadé-Bottino N, Mergaert P, Alunni B. 2019. Symbiotic efficiency of

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.08.479627; this version posted February 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

spherical and elongated bacteroids in the *Aeschynomene-Bradyrhizobium* symbiosis. Front Plant Sci 10:377.

6. Haag AF, Mergaert P. December 2019, 13th. Chapter 9.2.2, Terminal bacteroid differentiation in the *Medicago–Rhizobium* interaction–a tug of war between plant and bacteria. F. de Bruijn (Ed.), Model Legume *Medicago truncatula*. John Wiley & Sons, Inc., New York, NY https://doi.org/10.1002/9781119409144.ch75

7. Wilson JK. 1939. Leguminous plants and their associated organisms. Mem Cornell univ Exp Sta 221.

8. Pueppke SG, Broughton WJ. 1999. *Rhizobium* sp. strain NGR234 and *R. fredii* USDA257 share exceptionally broad, nested host ranges. Mol Plant-Microbe Interact 12:293–318.

9. Walker L, Lagunas B, Gifford ML. 2020. Determinants of Host Range Specificity in Legume-Rhizobia Symbiosis. Front Microbiol 11:85749.

10. Maxwell CA, Hartwig UA, Joseph CM, Phillips DA. 1989. A chalcone and two related flavonoids released from alfalfa roots induce *nod* genes of *Rhizobium meliloti*. Plant Physiol 91:842–847.

11. Recourt K, Schripsema J, Kijne JW, van Brussel AAN, Lugtenberg BJJ. 1991. Inoculation of *Vicia sativa* subsp. *nigra* roots with *Rhizobium leguminosarum* biovar *viciae* results in release of *nod* gene activating flavanones and chalcones. Plant Mol Biol 16:841–852.

12. Kosslak RM, Bookland R, Barkei J, Paaren HE, Appelbaum ER. 1987. Induction of *Bradyrhizobium japonicum* common *nod* genes by isoflavones isolated from Glycine max. Proc Natl Acad Sci U S A 84:7428–7432.

13. Pueppke SG, Bolanos-Vásquez MC, Werner D, Bec-Ferté M-P, Promé J-C, Krishnan HB. 1998. Release of flavonoids by the soybean cultivars McCall and Peking and their perception as signals by the nitrogen-fixing symbiont *Sinorhizobium fredii*. Plant Physiol 117:599–606.

14. D'Haeze W, Holsters M. 2002. Nod factor structures, responses, and perception during initiation of nodule development. Glycobiology 12:79R-105R.

15. Finan TM, Hirsch AM, Leigh JA, Johansen E, Kuldau GA, Deegan S, Walker GC, Signer ER. 1985. Symbiotic mutants of *Rhizobium meliloti* that uncouple plant from bacterial differentiation. Cell 40:869–877.

16. Leigh JA, Signer ER, Walker GC. 1985. Exopolysaccharide-deficient mutants of *Rhizobium meliloti* that form ineffective nodules. Proc Natl Acad Sci U S A 82:6231 LP – 6235.

17. Simsek S, Ojanen-Reuhs T, Stephens SB, Reuhs BL. 2007. Strain-ecotype specificity in *Sinorhizobium meliloti-Medicago truncatula* symbiosis is correlated to succinoglycan oligosaccharide structure. J Bacteriol 189:7733–7740.

18. Yang S, Tang F, Gao M, Krishnan HB, Zhu H. 2010. *R* gene-controlled host specificity in

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.08.479627; this version posted February 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

the legume–rhizobia symbiosis. Proc Natl Acad Sci U S A107:18735–18740.

19. Tsukui T, Eda S, Kaneko T, Sato S, Okazaki S, Kakizaki-Chiba K, Itakura M, Mitsui H, Yamashita A, Terasawa K. 2013. The type III secretion system of *Bradyrhizobium japonicum* USDA122 mediates symbiotic incompatibility with *Rj2* soybean plants. Appl Environ Microbiol 79:1048–1051.

20. Tsurumaru H, Hashimoto S, Okizaki K, Kanesaki Y, Yoshikawa H, Yamakawa T. 2015. A putative type III secretion system effector encoded by the *MA20_12780* gene in *Bradyrhizobium japonicum* Is-34 causes incompatibility with Rj_4 genotype soybeans. Appl Environ Microbiol 81:5812–5819.

21. Yang S, Wang Q, Fedorova E, Liu J, Qin Q, Zheng Q, Price PA, Pan H, Wang D, Griffitts JS. 2017. Microsymbiont discrimination mediated by a host-secreted peptide in *Medicago truncatula*. Proc Natl Acad Sci U S A 114:6848–6853.

22. Wang Q, Yang S, Liu J, Terecskei K, Ábrahám E, Gombár A, Domonkos Á, Szűcs A, Körmöczi P, Wang T. 2017. Host-secreted antimicrobial peptide enforces symbiotic selectivity in *Medicago truncatula*. Proc Natl Acad Sci U S A 114:6854–6859.

23. Wang Q, Liu J, Li H, Yang S, Körmöczi P, Kereszt A, Zhu H. 2018. Nodule-specific cysteine-rich peptides negatively regulate nitrogen-fixing symbiosis in a strain-specific manner in *Medicago truncatula*. Mol Plant-Microbe Interact 31:240–248.

24. diCenzo GC, Zamani M, Ludwig HN, Finan TM. 2017. Heterologous complementation reveals a specialized activity for BacA in the *Medicago–Sinorhizobium meliloti* symbiosis. Mol Plant-Microbe Interact 30:312–324.

25. Young ND, Debellé F, Oldroyd GED, Geurts R, Cannon SB, Udvardi MK, Benedito VA, Mayer KFX, Gouzy J, Schoof H. 2011. The *Medicago* genome provides insight into the evolution of rhizobial symbioses. Nature 480:520–524.

26. Mergaert P, Nikovics K, Kelemen Z, Maunoury N, Vaubert D, Kondorosi A, Kondorosi E. 2003. A novel family in *Medicago truncatula* consisting of more than 300 nodule-specific genes coding for small, secreted polypeptides with conserved cysteine motifs. Plant Physiol 132:161–173.

27. Maróti G, Downie JA, Kondorosi É. 2015. Plant cysteine-rich peptides that inhibit pathogen growth and control rhizobial differentiation in legume nodules. Curr Opin Plant Biol 26:57–63.

28. Maróti G, Kereszt A, Kondorosi E, Mergaert P. 2011. Natural roles of antimicrobial peptides in microbes, plants and animals. Res Microbiol 162:363–374.

29. Tiricz H, Szűcs A, Farkas A, Pap B, Lima RM, Maróti G, Kondorosi É, Kereszt A. 2013. Antimicrobial nodule-specific cysteine-rich peptides induce membrane depolarization-associated changes in the transcriptome of *Sinorhizobium meliloti*. Appl Environ Microbiol

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.08.479627; this version posted February 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

79:6737–6746.

30. Horváth B, Domonkos Á, Kereszt A, Szűcs A, Ábrahám E, Ayaydin F, Bóka K, Chen Y, Chen R, Murray JD. 2015. Loss of the nodule-specific cysteine rich peptide, NCR169, abolishes symbiotic nitrogen fixation in the *Medicago truncatula dnf7* mutant. Proc Natl Acad Sci U S A 112:15232–15237.

31. Kim M, Chen Y, Xi J, Waters C, Chen R, Wang D. 2015. An antimicrobial peptide essential for bacterial survival in the nitrogen-fixing symbiosis. Proc Natl Acad Sci U S A 112:15238–15243.

32. Glazebrook J, Ichige A, Walker GC. 1993. A *Rhizobium meliloti* homolog of the *Escherichia coli* peptide-antibiotic transport protein SbmA is essential for bacteroid development. Genes Dev 7:1485–1497.

33. Marlow VL, Haag AF, Kobayashi H, Fletcher V, Scocchi M, Walker GC, Ferguson GP. 2009. Essential role for the BacA protein in the uptake of a truncated eukaryotic peptide in *Sinorhizobium meliloti*. J Bacteriol 191:1519–1527.

34. Haag AF, Baloban M, Sani M, Kerscher B, Pierre O, Farkas A, Longhi R, Boncompagni E, Hérouart D, Dall'Angelo S. 2011. Protection of *Sinorhizobium* against host cysteine-rich antimicrobial peptides is critical for symbiosis. PLoS Biol 9:e1001169.

35. Guefrachi I, Pierre O, Timchenko T, Alunni B, Barriere Q, Czernic P, Villaécija-Aguilar J-A, Verly C, Bourge M, Fardoux J. 2015. *Bradyrhizobium* BclA is a peptide transporter required for bacterial differentiation in symbiosis with *Aeschynomene* legumes. Mol Plant-Microbe Interact 28:1155–1166.

36. Barrière Q, Guefrachi I, Gully D, Lamouche F, Pierre O, Fardoux J, Chaintreuil C, Alunni B, Timchenko T, Giraud E. 2017. Integrated roles of BclA and DD-carboxypeptidase 1 in *Bradyrhizobium* differentiation within NCR-producing and NCR-lacking root nodules. Sci Rep 7:1–13.

37. Maruya J, Saeki K. 2010. The *bacA* gene homolog, mlr7400, in *Mesorhizobium loti* MAFF303099 is dispensable for symbiosis with *Lotus japonicus* but partially capable of supporting the symbiotic function of *bacA* in *Sinorhizobium meliloti*. Plant Cell Physiol 51:1443–1452.

38. Honma MA, Ausubel FM. 1987. *Rhizobium meliloti* has three functional copies of the *nodD* symbiotic regulatory gene. Proc Natl Acad Sci U S A 84:8558–8562.

39. Zamani M, Cowie A, Finan TM. 2015. Proline auxotrophy in *Sinorhizobium meliloti* results in a plant-specific symbiotic phenotype. Microbiology 161:2341–2351.

40. Zamani M, diCenzo GC, Milunovic B, Finan TM. 2017. A putative 3-hydroxyisobutyryl-CoA hydrolase is required for efficient symbiotic nitrogen fixation in *Sinorhizobium meliloti* and *Sinorhizobium fredii* NGR234. Environ Microbiol 19:218–236.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.08.479627; this version posted February 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

41. Geddes BA, Kearsley JVS, Huang J, Zamani M, Muhammed Z, Sather L, Panchal AK, Finan TM. 2021. Minimal gene set from *Sinorhizobium* (*Ensifer*) *meliloti* pSymA required for efficient symbiosis with *Medicago*. Proc Natl Acad Sci U S A 118: e2018015118.

42. Jensen HL. 1942. Nitrogen fixation in leguminous plants. I. General characters of root-nodule bacteria isolated from species of *Medicago* and *Trifolium* in Australia, p. 98–108. *In* Proc. Linn. Soc. NSW.

43. Andrews S, Krueger F, Seconds-Pichon A, Biggins F, Wingett S. 2015. "FastQC. A quality control tool for high throughput sequence data. Babraham Bioinformatics", Babraham Institute.

44. Song L, Florea L. 2015. Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads. Gigascience 4:s13742-015.

45. Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J 17:10–12.

46. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120.

47. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29:644.

48. Davidson NM, Hawkins ADK, Oshlack A. 2017. SuperTranscripts: a data driven reference for analysis and visualisation of transcriptomes. Genome Biol 18:1–10.

49. Davidson NM, Oshlack A. 2014. Corset: enabling differential gene expression analysis for de novo assembled transcriptomes. Genome Biol 15:1–14.

50. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29:15–21.

51. Seppey M, Manni M, Zdobnov EM. 2019. BUSCO: assessing genome assembly and annotation completeness, p. 227–245. *In* Gene prediction. Springer.

52. Sallet E, Roux B, Sauviac L, Jardinaud M-F, Carrère S, Faraut T, de Carvalho-Niebel F, Gouzy J, Gamas P, Capela D, Bruand C, Schiex T. 2013. Next-Generation Annotation of Prokaryotic Genomes with EuGene-P: Application to *Sinorhizobium meliloti* 2011. DNA Res 20:339–354.

53. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:1–9.

54. Chen H, Zeng Y, Yang Y, Huang L, Tang B, Zhang H, Hao F, Liu W, Li Y, Liu Y, Zhang X, Zhang R, Zhang Y, Li Y, Wang K, He H, Wang Z, Fan G, Yang H, Bao A, Shang Z, Chen J, Wang W, Qiu Q. 2020. Allele-aware chromosome-level genome assembly and efficient transgene-free genome editing for the autotetraploid cultivated alfalfa. Nat

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.08.479627; this version posted February 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Commun 11:2494.

55. 2021. UniProt: The universal protein knowledgebase in 2021. Nucleic Acids Res 49:D480–D489.

56. Pecrix Y, Staton SE, Sallet E, Lelandais-Brière C, Moreau S, Carrère S, Blein T, Jardinaud M-F, Latrasse D, Zouine M, Zahm M, Kreplak J, Mayjonade B, Satgé C, Perez M, Cauet S, Marande W, Chantry-Darmon C, Lopez-Roques C, Bouchez O, Bérard A, Debellé F, Muños S, Bendahmane A, Bergès H, Niebel A, Buitink J, Frugier F, Benhamed M, Crespi M, Gouzy J, Gamas P. 2018. Whole-genome landscape of Medicago truncatula symbiotic genes. Nat Plants 4:1017–1025.

57. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, Mende DR, Letunic I, Rattei T, Jensen LJ, Von Mering C, Bork P. 2019. EggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res 47:D309-D314.

58. Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. Nat Methods 12:59–60.

59. Eddy SR. 2011. Accelerated profile HMM searches. PLoS Comput Biol 7:e1002195.

60. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ. 2021. Pfam: The protein families database in 2021. Nucleic Acids Res 49:D412–D419.

61. Haft DH, Selengut JD, Richter RA, Harkins D, Basu MK, Beck E. 2012. TIGRFAMs and genome properties in 2013. Nucleic Acids Res 41:D387–D395.

62. Zhou P, Silverstein KAT, Gao L, Walton JD, Nallu S, Guhlin J, Young ND. 2013. Detecting small plant peptides using SPADA (small peptide alignment discovery application). BMC Bioinform 14:1–16.

63. Stanke M, Schöffmann O, Morgenstern B, Waack S. 2006. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinform 7:1–11.

64. Birney E, Clamp M, Durbin R. 2004. GeneWise and genomewise. Genome Res 14:988–995.

65. Lukashin A V, Borodovsky M. 1998. GeneMark. hmm: new solutions for gene finding. Nucleic Acids Res 26:1107–1115.

66. Blanco E, Parra G, Guigó R. 2007. Using geneid to identify genes. Curr Protoc Bioinform 18:3–4.

67. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7:539.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.08.479627; this version posted February 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

68. Petersen TN, Brunak S, Von Heijne G, Nielsen H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods 8:785–786.

69. Audain E, Ramos Y, Hermjakob H, Flower DR, Perez-Riverol Y. 2016. Accurate estimation of isoelectric point of protein and peptide based on amino acid sequences. Bioinformatics 32:821–827.

70. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. 2017. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods 14:417–419.

71. Love M, Anders S, Huber W. 2014. Differential analysis of count data–the DESeq2 package. Genome Biol 15:10–1186.

72. Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol 20:1–14.

73. Computing Rf. 2013. R: A language and environment for statistical computing. Vienna R Core Team.

74. Wang T-Z, Liu M, Zhao M-G, Chen R, Zhang W-H. 2015. Identification and characterization of long non-coding RNAs involved in osmotic and salt stress in *Medicago truncatula* using genome-wide high-throughput sequencing. BMC Plant Biol 15:131.

75. Kerr SC, Gaiti F, Beveridge CA, Tanurdzic M. 2017. *De novo* transcriptome assembly reveals high transcriptional complexity in *Pisum sativum* axillary buds and shows rapid changes in expression of diurnally regulated genes. BMC Genomics 18:221.

76. Zhang S, Shi Y, Cheng N, Du H, Fan W, Wang C. 2015. *De novo* characterization of fall dormant and nondormant alfalfa (*Medicago sativa* L.) leaf transcriptome and identification of candidate genes related to fall dormancy. PLoS One 10:e0122170.

77. Arshad M, Gruber MY, Hannoufa A. 2018. Transcriptome analysis of microRNA156 overexpression alfalfa roots under drought stress. Sci Rep 8:1–13.

78. Al-Qurainy F, Alshameri A, Gaafar A-R, Khan S, Nadeem M, Alameri AA, Tarroum M, Ashraf M. 2019. Comprehensive stress-based *de novo* transcriptome assembly and annotation of guar (*Cyamopsis tetragonoloba* (L.) Taub.): an important industrial and forage crop. Int J Genomics 2019.

79. Weisberg AJ, Kim G, Westwood JH, Jelesko JG. 2017. Sequencing and *de novo* assembly of the *Toxicodendron radicans* (poison ivy) transcriptome. Genes (Basel) 8:317.

80. Malovichko Y V, Shtark OY, Vasileva EN, Nizhnikov AA, Antonets KS. 2020. Transcriptomic insights into mechanisms of early seed maturation in the garden pea (*Pisum sativum* L.). Cells 9:779.

81. Nash DT, Schulman HM. 1976. Leghemoglobins and nitrogenase activity during soybean root nodule development. Can J Bot 54:2790–2797.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.08.479627; this version posted February 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

82. Ott T, van Dongen JT, Gu C, Krusell L, Desbrosses G, Vigeolas H, Bock V, Czechowski T, Geigenberger P, Udvardi MK. 2005. Symbiotic leghemoglobins are crucial for nitrogen fixation in legume root nodules but not for general plant growth and development. Curr Biol 15:531–535.

83. Rahioui I, Eyraud V, Karaki L, Sasse F, Carre-Pierrat M, Qin A, Zheng MH, Toepfer S, Sivignon C, Royer C. 2014. Host range of the potential biopesticide Pea Albumin 1b (PA1b) is limited to insects. Toxicon 89:67–76.

84. Limpens E, Moling S, Hooiveld G, Pereira PA, Bisseling T, Becker JD, Küster H. 2013. Cell-and tissue-specific transcriptome analyses of *Medicago truncatula* root nodules. PLoS One 8:e64377.

85. Karaki L, Da Silva P, Rizk F, Chouabe C, Chantret N, Eyraud V, Gressent F, Sivignon C, Rahioui I, Kahn D, Brochier-Armanet C, Rahbé Y, Royer C. 2016. Genome-wide analysis identifies gain and loss/change of function within the small multigenic insecticidal Albumin 1 family of *Medicago truncatula*. BMC Plant Biol 16:63.

86. Montiel J, Downie JA, Farkas A, Bihari P, Herczeg R, Bálint B, Mergaert P, Kereszt A, Kondorosi É. 2017. Morphotype of bacteroids in different legumes correlates with the number and type of symbiotic NCR peptides. Proc Natl Acad Sci U S A 114:5041–5046.

87. Lima RM, Kylarová S, Mergaert P, Kondorosi É. 2020. Unexplored arsenals of legume peptides with potential for their applications in medicine and agriculture. Front Microbiol 11:1307.

88. Ördögh L, Vörös A, Nagy I, Kondorosi É, Kereszt A. 2014. Symbiotic Plant Peptides Eliminate *Candida albicans* Both *In Vitro* and in an Epithelial Infection Model and Inhibit the Proliferation of Immortalized Human Cells. Biomed Res Int 2014:320796.

89. Mikuláss KR, Nagy K, Bogos B, Szegletes Z, Kovács E, Farkas A, Váró G, Kondorosi É, Kereszt A. 2016. Antimicrobial nodule-specific cysteine-rich peptides disturb the integrity of bacterial outer and inner membranes and cause loss of membrane potential. Ann Clin Microbiol Antimicrob 15:43.

90. Quentin N, Quentin B, Nicolas B, Sara D, Dmitrii T, Mickaël B, Romain LB, Claire B, Marie L, Sándor J, Atilla K, Eva K, G. BE, Tatiana T, Benoît A, Peter M, Joerg G. 2022. *Sinorhizobium meliloti* Functions Required for Resistance to Antimicrobial NCR Peptides and Bacteroid Differentiation. mBio 12:e00895-21.

91. F. AMF, Mohammed S, Jon P, L. BK, S. GJ, C. WG, M. AF, Peter M, Philip P. 2022. Genome-Wide Sensitivity Analysis of the Microsymbiont *Sinorhizobium meliloti* to Symbiotically Important, Defensin-Like Host Peptides. mBio 8:e01060-17.

769 **Table 1.** Summary statistics from the *de novo* Trinity and compressed (SuperTranscripts) nodule transcriptome assemblies.

	M. sativa		*M. officinalis*	
	Trinity Assembly	SuperTranscripts	Trinity Assembly	SuperTranscripts
Total number of contigs	253,871	79,978	192,165	64,593
Total number of base pairs (bp)	239,421,306	96,211,060	191,200,637	81,371,534
Average contig length (bp)	943	1203	995	1260
Median contig length (bp)	618	734	646	780
Contig N50 (bp)	1,466	1,912	1,584	1968
Minimum contig length (bp)	176	193	183	198
Maximum contig length (bp)	12,655	29,784	14,584	23,941
Overall alignment rate (%)	98.99	92.14	99.15	93.76

770
771
772

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.08.479627; this version posted February 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

773 **Table 2.** The 50 most highly abundant transcripts in the *M. sativa* nodule transcriptome, with the
774 average expression level in transcripts per million (TPM) and the functional annotation.

Gene ID	TPM	Functional Prediction
Cluster-2.43518	30,662	hypothetical protein (hypothetical leghaemoglobin)
Cluster-2.26078	17,769	Putative albumin I
Cluster-2.23447	12,515	hypothetical protein (hypothetical leghaemoglobin)
Cluster-2.23176	8,065	Nodulin-25
Cluster-2.22272	7,992	Putative Late nodulin
Cluster-2.23033	6,778	None
Cluster-2.21873	5,919	hypothetical protein
Cluster-2.22935	5,253	Belongs to the globin family
Cluster-2.33070	5,232	Putative ribonuclease H-like domain-containing protein
Cluster-2.24197	4,245	Component of the replication protein A complex (RPA)
Cluster-2.22983	4,196	Belongs to the globin family
Cluster-2.24546	3,886	Predicted NCR peptide (crp1450_Cluster-2.24546_0M_1)
Cluster-2.19511	3,344	None
Cluster-2.26207	3,296	Extensin-like_protein_repeat
Cluster-2.49512	3,173	Putative Blue (type 1) copper binding protein
Cluster-2.21810	3,050	Predicted NCR peptide (crp1160_Cluster-2.21810_0M_1)
Cluster-2.22936	3,014	Belongs to the globin family
Cluster-2.29430	2,789	Putative Late nodulin
Cluster-2.22881	2,788	Predicted NCR peptide (crp1430_Cluster-2.22881_0M_1)
Cluster-2.22836	2,509	None
Cluster-2.23729	2,271	hypothetical protein
Cluster-2.22458	2,245	Belongs to the globin family
Cluster-2.24829	2,227	hypothetical protein
Cluster-2.25168	2,209	None
Cluster-2.24278	2,093	Nodule-specific_GRP_repeat
Cluster-2.23928	2,038	Late_nodulin_protein
Cluster-2.22042	2,029	None
Cluster-2.23245	2,019	Putative translationally controlled tumor protein
Cluster-2.25794	2,018	Putative protein-synthesizing GTPase
Cluster-2.31376	1,957	Predicted NCR peptide (crp1190_Cluster-2.31376_0M_1)
Cluster-2.21809	1,939	Predicted NCR peptide (crp1160_Cluster-2.21809_0M_1)
Cluster-2.28083	1,854	Predicted NCR peptide (crp1210_Cluster-2.28083_0M_1)
Cluster-2.23524	1,844	Early nodulin-16
Cluster-2.34649	1,839	Putative Late nodulin
Cluster-2.16993	1,808	hypothetical protein
Cluster-2.21813	1,731	None
Cluster-2.22457	1,705	Belongs to the globin family
Cluster-2.28283	1,674	None
Cluster-2.26205	1,621	Predicted NCR peptide (crp1240_Cluster-2.26205_0M_1)
Cluster-2.23310	1,607	asparagine synthetase
Cluster-2.21870	1,596	Nodule-specific_GRP_repeat
Cluster-2.19604	1,583	Predicted NCR peptide (crp1420_Cluster-2.19604_0M_1)
Cluster-2.18485	1,509	hypothetical protein
Cluster-2.26876	1,458	Predicted NCR peptide (crp1410_Cluster-2.26876_0M_1)
Cluster-2.21536	1,441	Ubiquitin_family
Cluster-2.22785	1,408	None
Cluster-2.23374	1,385	Predicted NCR peptide (crp1420_Cluster-2.23374_0M_1)
Cluster-2.28787	1,311	Putative Late nodulin
Cluster-2.22402	1,292	Predicted NCR peptide (crp1520_Cluster-2.22402_0M_1)
Cluster-2.30081	1,289	Late_nodulin_protein

775

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.08.479627; this version posted February 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Table 3. The 50 most highly abundant transcripts in the *M. officinalis* nodule transcriptome, with the average expression level in transcripts per million (TPM) and the functional annotation.

Gene ID	TPM	Functional Prediction
Cluster-3554.18801	38,953	Belongs to the globin family
Cluster-3554.18778	14,091	Belongs to the globin family
Cluster-3554.16063	10,412	Late_nodulin_protein
Cluster-3554.15387	7,885	Putative Late nodulin
Cluster-3554.16088	6,146	Putative Late nodulin
Cluster-3554.18892	6,104	None
Cluster-3554.15771	5,813	Late_nodulin_protein
Cluster-3554.18802	5,362	Belongs to the globin family
Cluster-3554.18596	5,347	Predicted NCR peptide (crp1430_Cluster-3554.18596_0M_1)
Cluster-3554.15456	3,912	Putative translationally controlled tumor protein
Cluster-3554.18808	3,864	Belongs to the globin family
Cluster-3554.33215	3,302	hypothetical protein
Cluster-3554.18775	3,097	Putative Late nodulin
Cluster-3554.23000	2,990	Two predicted NCR peptide (crp1180_Cluster-3554.23000_0M_1 and crp1180_Cluster-3554.23000_0M_2)
Cluster-3554.36681	2,877	Predicted NCR peptide (crp1500_Cluster-3554.36681_0M_1)
Cluster-3554.21555	2,868	Predicted NCR peptide (crp1430_Cluster-3554.21555_0M_1)
Cluster-3554.16074	2,809	Putative BURP domain-containing protein
Cluster-3554.29297	2,784	hypothetical protein
Cluster-3554.18196	2,723	None
Cluster-3554.18256	2,571	Predicted NCR peptide (crp1440_Cluster-3554.18256_0M_1)
Cluster-3554.27063	2,522	None
Cluster-3554.22577	2,449	Putative Blue (type 1) copper binding protein
Cluster-3554.15361	2,409	eEF1A
Cluster-3554.21311	2,384	Belongs to the globin family
Cluster-3554.18838	2,340	Late_nodulin_protein
Cluster-3554.18700	2,296	Late_nodulin_protein
Cluster-3554.11713	2,261	None
Cluster-3554.18706	2,259	None
Cluster-3554.23445	2,214	None
Cluster-3554.17366	2,141	Predicted NCR peptide (crp1430_Cluster-3554.17366_0M_1)
Cluster-3554.23502	2,073	hypothetical protein
Cluster-3554.25153	2,009	Metallothionein-like protein 2
Cluster-3554.21451	1,953	Belongs to the globin family
Cluster-3554.28600	1,931	hypothetical protein
Cluster-3554.22971	1,849	hypothetical protein
Cluster-3554.18877	1,812	Belongs to the glyceraldehyde-3-phosphate dehydrogenase family
Cluster-3554.13172	1,732	Predicted NCR peptide (crp1440_Cluster-3554.13172_0M_1)
Cluster-3554.18195	1,704	Zinc_knuckle
Cluster-3554.30751	1,653	Putative Late nodulin
Cluster-3554.21774	1,635	Late_nodulin_protein
Cluster-3554.13784	1,629	Predicted NCR peptide (crp1420_Cluster-3554.13784_0M_1)
Cluster-3554.24033	1,611	Prolyl isomerase (PPIase)
Cluster-3554.30294	1,556	metallothionein-like protein
Cluster-3554.24764	1,552	None
Cluster-3554.18368	1,552	Nucleoside diphosphate kinase 1
Cluster-3554.25344	1,537	Metallothionein-like protein 1
Cluster-3554.23646	1,514	None
Cluster-3554.9874	1,512	Belongs to the universal ribosomal protein uL13 family
Cluster-3554.18476	1,503	Late_nodulin_protein
Cluster-3554.31722	1,476	Putative Late nodulin

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.08.479627; this version posted February 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

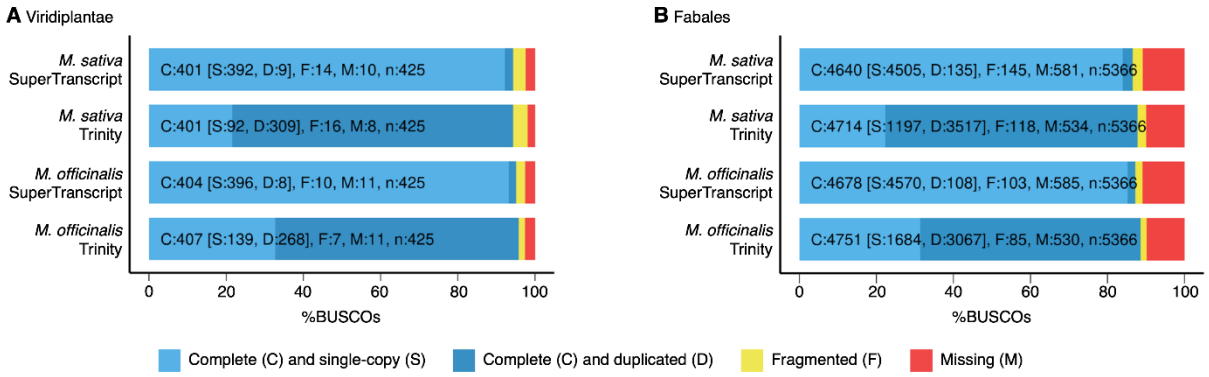

Figure 1. Estimates of nodule transcriptome completeness. Completeness of the *M. sativa* and *M. officinalis* nodule transcriptome assemblies was assessed using BUSCO with the (**A**) Viridiplantae and (**B**) Fabales single-copy marker gene datasets. The fraction of BUSCO genes identified as complete and single-copy (light blue), complete but duplicated (dark blue), fragmented (yellow), and missing (red) is shown.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.08.479627; this version posted February 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

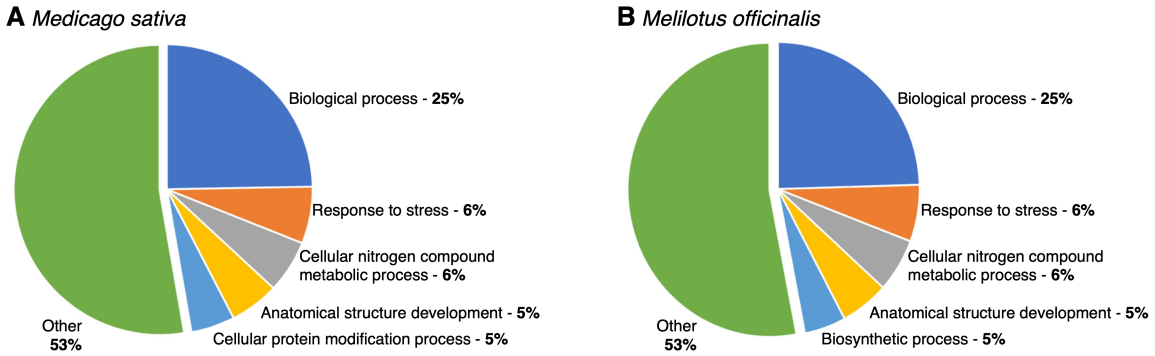

Figure 2. Summary of the Slim GO Biological Processes annotations for the nodule transcriptomes. Transcripts were annotated with Slim GO terms, and the annotations for the biological processes were summarized as pie charts for (**A**) *M. sativa* and (**B**) *M. officinalis*.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.08.479627; this version posted February 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

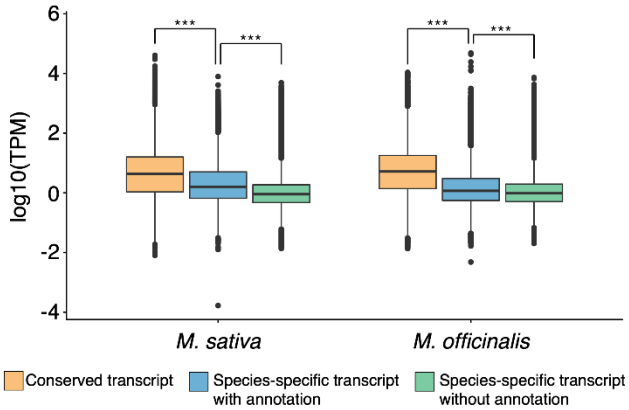

Figure 3. Transcript abundances for conserved and species-specific transcripts. Box plots displaying the distribution of average transcript abundances from triplicate samples, shown separately for genes with orthologs in both *M. sativa* and *M. officinalis* (orange), annotated transcripts found in only *M. sativa* or *M. officinalis* (blue), or transcripts that lack annotations and are found in only *M. sativa* or *M. officinalis* (green). Statistically significant differences between the distributions of a species are indicated with the asterisks (p-value < $1e^{-10}$; pairwise Wilcox tests).

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.08.479627; this version posted February 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

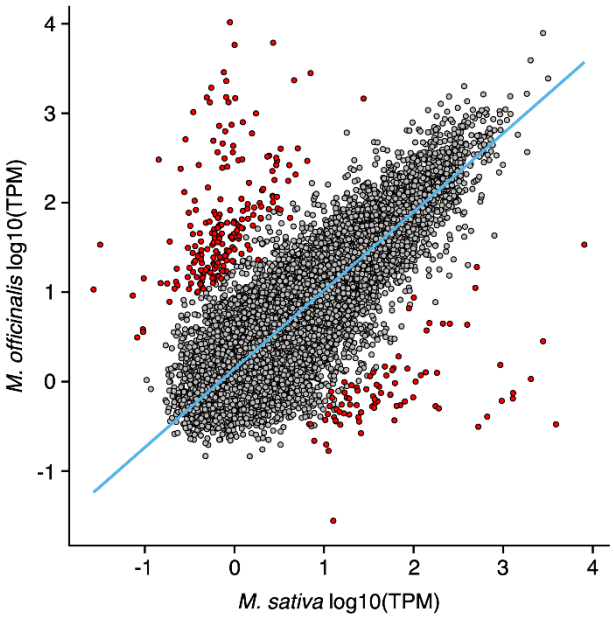

Figure 4. Correlation between transcript abundances of orthologous transcripts in *M. sativa* and *M. officinalis*. Each datapoint represents the transcript abundance of single-copy orthologous transcripts in *M. sativa* and *M. officinalis*. Red datapoints represent transcripts that are differentially abundant between the two species ($|\log_2(\text{fold change})| > 5$, adjusted p-value < 0.01); all other datapoints are in grey. The blue line represents the robust linear regression line, calculated with the rlm function of the MASS package in R.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.08.479627; this version posted February 10, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

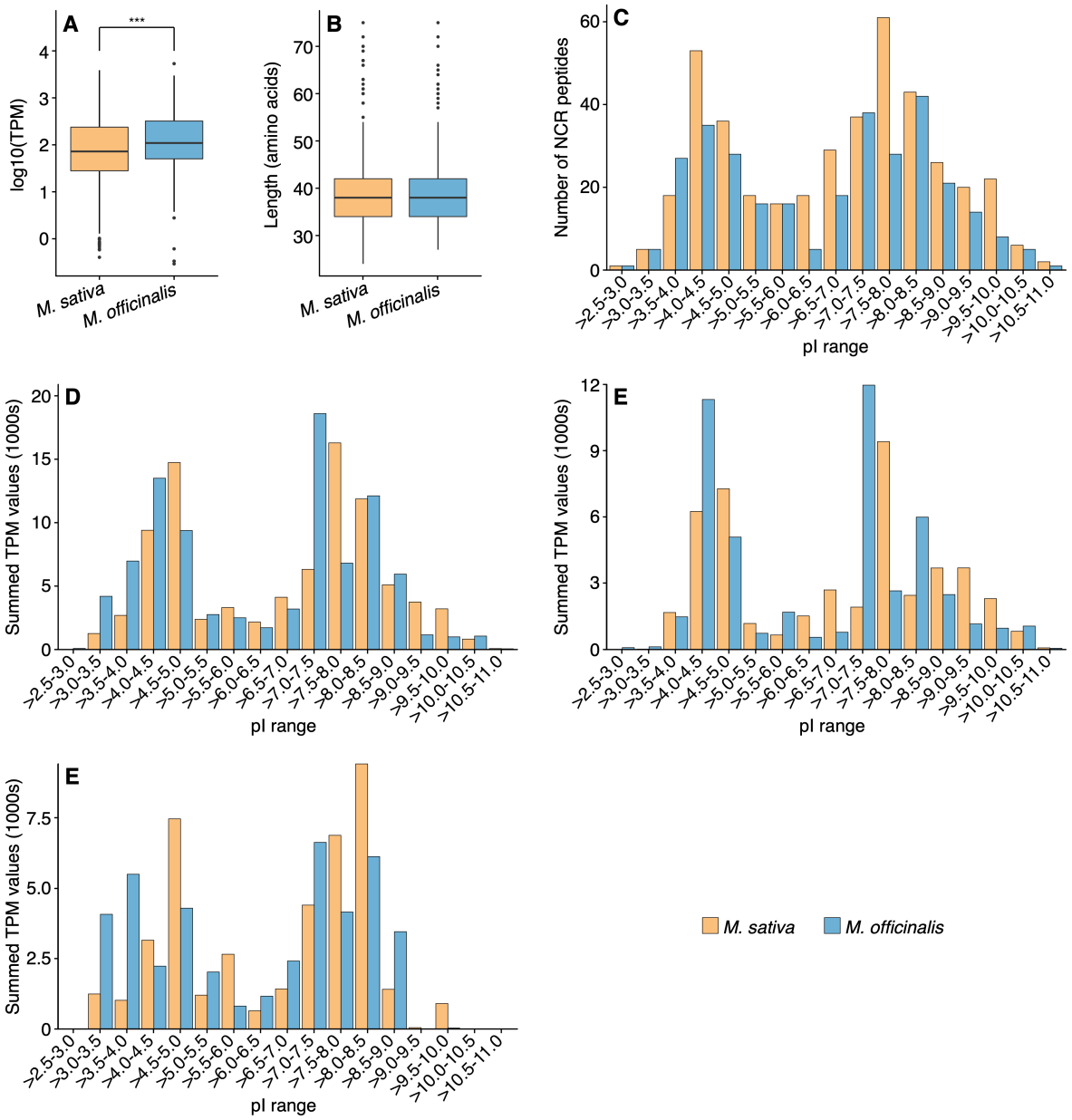

Figure 5. NCR peptide profiles of *Medicago sativa* and *Melilotus officinalis*. NCR peptides were predicted from the *M. sativa* (orange) and *M. officinalis* (blue) transcriptome assemblies, and the properties of the NCR peptides are shown in these graphs. (**A**) Box plots showing the distribution of the abundance (in transcripts per million, TPM) of *NCR* transcripts, based on triplicate samples. The difference in the distributions for the two species was statistically significant (p-value < 0.001; pairwise Wilcox test). (**B**) Box plots showing the distribution of the amino acid lengths of mature NCR peptides. No statistically significant difference in the distributions for the two species was detected. (**C**,**D**) Histograms showing the distributions of the isoelectric points (pI) for the mature NCR peptides. Histograms are based either on the number of NCR peptides with a given pI value (C) or the total abundance of the transcripts encoding NCR peptides with a given pI value (D). **(E,F)** Histograms showing distributions of pI for 4-cysteines (E) and 6-cysteines (F) mature NCR peptides based on total abundance of the transcripts encoding NCR peptides with a given pI value.