Loss-of-function phenomics, ncORFs, and ambiguity of mutant phenotypes in *Medicago truncatula*

Umut Çakır[1,2], Noujoud Gabed[3], Selen Kaya[4], Vagner A. Benedito[5], Marie Brunet[6,7], Xavier Roucou[7,8], Igor S. Kryvoruchko[9]

[1]Clinical Neuroscience Research Group, Max Planck Institute for Multidisciplinary Sciences, City Campus, Göttingen 37075, Germany

[2]Georg-August-University, Göttingen 37073, Germany

[3]Cellular and Molecular Biology Department, Oran High School of Biological Sciences (ESSBO), Oran 31000, Algeria

[4]Department of Molecular Biology and Genetics, Çanakkale Onsekiz Mart University, Çanakkale 17020, Turkey

[5]School of Agriculture and Natural Sciences, University of Maryland Eastern Shore, Princess Anne, MD 21853, USA

[6]Department of Pediatrics, Medical Genetics Service, Université de Sherbrooke, Sherbrooke QC J1E4K8, Canada

[7]Centre de Recherche du Centre Hospitalier Universitaire de Sherbrooke (CRCHUS), Sherbrooke QC J1E4K8, Canada

[8]Department of Biochemistry and Functional Genomics, Université de Sherbrooke, Sherbrooke QC J1E4K8, Canada

[9]Department of Biology, United Arab Emirates University, P.O. Box 15551, Al Ain, UAE

Correspondence

I. S. Kryvoruchko, Department of Biology, College of Science, United Arab Emirates University, Building F1, Room 0169, P.O. Box 15551, Al Ain, UAE, Tel: +971 3 7136724
E-mail: igor.kryvoruchko@uaeu.ac.ae

Running title

Loss-of-function phenomics and ncORFs

Abbreviations

ncORF, non-canonical open reading frame; refORF, reference open reading frame; ncProt, non-canonical protein; refProt, reference protein; mRNA, messenger RNA; ncRNA, non-coding RNA; rRNA, ribosomal RNA; tRNA, transfer RNA; CDS, coding sequence; UTR, untranslated region; aa, amino acids; bp, base pairs; nt, nucleotides; MS, mass spectrometry

SUMMARY

Non-canonical open reading frames (ncORFs) are an emerging area of research that is quickly gaining momentum. Many peptides and proteins missed in initial annotation efforts (ncProts) were subsequently shown to be crucial for a wide range of biological processes. The discovery of ncORFs continues to improve the accuracy of loss-of-function studies because they often occupy the same genomic spaces as annotated ORFs. While databases of mutant phenotypes linked to genomic loci are available in a few species, none of these databases integrate the information on ncORFs present in already characterized loci. In this study, we introduce a nearly comprehensive loss-of-function phenomics dataset of *Medicago truncatula* (673 loci

characterized over the past 30 years), which should become an integral part of the genome browser of this organism. We used this dataset to provide a critical analysis of the potential contribution of ncORFs to published phenotypes. We detected mass spectrometry (MS)-validated ncORFs in 10 functionally characterized genes, including major regulators of development and symbiotic relationships. We also found conserved ncORFs in 113 characterized genes, including four genes with highly conserved ncORFs. We show that in some studies, the contribution of ncORFs can be ruled out, while in others it cannot. Using real examples, we systematized ambiguities associated with ncORFs. Furthermore, we highlighted little-known trans effects of insertional mutagenesis on splicing as contributors to that ambiguity. Finally, our meta-analysis of published phenotypes indicates that different protein classes have significantly different (unique) proportions of unconditional, conditional, and neutral phenotypes, potentially reflecting their relative functional importance.

Keywords

non-canonical open reading frame, mutant phenotype, phenomics, functional genomics, mass spectrometry proteomics, *MthGSHSb*, *MtKASII*, *MtPHO2-A*, *MtbZIP60b*, purifying selection

Significance statement

This study is the first to merge a nearly comprehensive inventory of loss-of-function studies in a eukaryotic organism with the information on novel MS-validated and conserved ncORFs.

INTRODUCTION

For a long time, functional genomics and genome annotation efforts have primarily been focused on annotated reference proteins (refProts) translated from canonical (reference) open reading frames (refORFs), often overlooking additional coding potential within the same transcripts (Kozak, 1999). Mass spectrometry (MS) proteomics challenged this paradigm through the discovery of the non-canonical proteome (Vanderperre et al., 2013; Mouilleron et al., 2016; Orr et al., 2020; Cardon et al., 2021; Wright et al., 2022). It was deduced from the abundance of high-quality mass spectra that cannot be assigned to refProts (Ning et al., 2010; Pathan et al., 2017). Such a hidden proteome exists even in the smallest culturable bacterium (*Mycoplasma pneumoniae*) with only 729 ORFs (Lluch-Senar et al., 2016). A considerable part of the non-annotated proteome is thought to originate from translation of non-canonical ORFs (ncORFs) to non-canonical peptides and proteins (ncProts). The recognition of the biological relevance of these ncProts is relatively recent (Mouilleron et al., 2016; Orr et al., 2020; Mudge et al., 2022) and has been accelerating in the past few years (Ardern, 2023; Álvarez-Urdiola and Riechmann, 2025; Çakır et al., 2025b; Cardon et al., 2025; Pavesi, 2025; Chang et al., 2025).

The ultimate goal of functional genomics studies is to establish accurate relationships between phenotypes and genotypes for the benefit of biomedical research, agriculture, and biotechnology. We can refer to the multitude of such phenotype-genotype units as a phenome. Deducing a true phenome from functional data can be very complex due to the polycistronic nature of eukaryotic transcripts (Mouilleron et al., 2016). Specifically, it introduces a new level of complexity to the interpretation of all functional genomics studies. A ncORF with a biological function can be present on the same RNA molecule together with the refORF. For example, it can be located in a sequence annotated as an untranslated region (UTR) based on the refORF, even though this region may contain a functional coding sequence in any reading frame. For such transcripts, loss-of-function approaches based on post-transcriptional gene silencing, such as RNA interference (RNAi) (Sharp & Zamore, 2000; Baulcombe, 2004; Chaudhary et al., 2024), yield phenotypes that cannot be unequivocally attributed to refProts

alone because an RNAi construct simultaneously affects both refProts and ncProts. Insertion-deletion mutagenesis systems such as CRISPR/Cas9 (for Clustered Regularly Interspaced Short Palindromic Repeats [CRISPR] and CRISPR-associated protein 9 [Cas9]; Doudna & Charpentier, 2014; Guo et al., 2023) and retrotransposons (e.g., *Tnt1*, for Transposon of *Nicotiana tabacum*; Tadege et al., 2008; Lee et al., 2018; Kaur et al., 2021) may seem to be more proof against such ambiguity because they can be more selective regarding the target region of a transcript. However, these systems, too, can affect both a refProt and a ncProt, especially if their ORFs overlap and share the same genetic space. In this study, we described a scenario in which insertional mutagenesis affects both ORFs, even if the insertion is located only in a refORF, acting on a ncORF in trans through the effect on splicing. Likewise, gain-of-function studies introduce similar ambiguity because an overexpression of a refORF simultaneously overexpresses an embedded ncORF. These considerations call for an objective re-evaluation of published phenotypes in light of the potential contribution of ncORFs.

Although very comprehensive phenomics databases have been established for mouse (Wang et al., 2015; Baldarelli et al., 2024), *Arabidopsis thaliana* (Lloyd and Meinke, 2012), and other eukaryotes (Courtier-Orgogozo et al., 2020), none of them considered the contribution of ncORFs so far. Our analysis was conducted using *Medicago truncatula* because it stands out as a model organism. In contrast to *Arabidopsis*, it can establish symbiotic nitrogen fixation (SNF) in association with soil bacteria (rhizobia) and arbuscular mycorrhizal (AM) symbiosis with fungi (Barker et al., 1990; Kang et al., 2016; Bandyopadhyay et al., 2019; Nandety et al., 2022). *M. truncatula* has been used as a model for functional genomics since 1995 (Sagan et al., 1995). Despite 30 years in this status, the model has no comprehensive inventory of loss-of-function studies so far. Likewise, a large-scale detection of ncORFs on annotated transcripts had not been conducted in *M. truncatula* until very recently (Çakır et al., 2025b). Even though much attention was paid to the inventory of bioactive peptides in *M. truncatula* (Mergaert et al., 2003; Zhang et al., 2006; Volkening et al., 2012; de Bang et al., 2017; Kereszt et al., 2018; Boschiero et al., 2020), most of them should only conditionally be categorized as ncProts because they originate from transcripts with no annotated long refORFs. Their short ORFs act as refORFs of corresponding transcripts (see supporting data, for example, Data S12). In this study, we filled the knowledge gap about ncORFs, primarily mRNA-ncORFs, in conjunction with published loss-of-function phenotypes attributed to refORFs. We collected information on nearly all such phenotypes linked to specific loci or groups of genes and merged it with data on MS-validated and conserved ncORFs. For the MS validation, we analyzed three large datasets available on ProteomeXchange (Deutsch et al., 2023). They correspond to 16 samples of 10 different organs (PXD002692, Marx et al., 2016; PXD013606, Shin et al., 2021; and PXD022278, Castañeda et al., 2021). To identify conserved ncORFs, we conducted a three-frame global amino acid sequence similarity search of the whole transcriptome of *M. truncatula* using UniProt v. 2020_02 as a subject database (UniProt Consortium, 2019). We deployed a simple, “naïve” approach based on the difference in the number of hits between ncORFs (conservation at the nucleotide level) and corresponding ncProts (conservation at the amino acid level). This way, we detected highly conserved ncORFs under purifying selection in four well-characterized genes, including the master regulator of phosphate homeostasis *MtPHO2-A* (Curtin et al., 2017; Čermák et al., 2017; Huertas et al., 2023). We exemplified the applicability of this “naïve” approach for the large-scale detection of ncORFs that undergo purifying selection. Our analysis permitted discrimination between studies that can rule out the contribution of ncORFs to reported phenotypes and studies in which such ambiguity calls for additional experiments, such as differential mutagenesis of ncORFs and refORFs. By straightforward sequence comparisons, we explained the likely origin of highly conserved ncORFs in five well-characterized genes

138 and showed that strong conservation may not necessarily imply translation of a given ncORF.
139 Instead, some conserved ncORFs are better understood in the context of alternative splicing
140 or gene model structure. Thus, our approach is not only informative for identifying potentially
141 functional ncORFs, but also valuable for refining current gene annotations. Due to space
142 limitations, we placed much of the valuable information about MS-validated and conserved
143 ncORFs into files called Supporting Results and Supporting Discussion. We refer readers to
144 these files for the deep coverage of ncORFs in transcripts characterized by loss-of-function
145 approaches.

146 Our phenomics resource will be useful to the *M. truncatula* community regardless of the
147 ncORF dimension, because it serves the following basic purposes: (1) avoiding the duplication
148 of efforts in studying genes that have already been targeted by other groups; (2) consistency
149 of naming new loci and avoiding redundancy; (3) correction of gene models. It should be noted
150 that despite the very impressive accuracy of the current genome annotation (v. 5.1.9, Pecrix
151 et al., 2018), we found surprisingly large number of cases in which annotation accuracy can
152 be further improved.

153 Lastly, our inferential meta-analysis of loss-of-function phenotypes indicates that diverse
154 protein classes have unique proportions of altered (non-conditional), conditional, and neutral
155 phenotypes, which are significantly different from the overall and specific proportions of
156 functional protein groups in the dataset. These distinct frequencies of phenotypes may reflect
157 the degree of biological importance or functional redundancy of specific groups of proteins.
158 This information can help navigate searches to genes with strong or weak but clear
159 phenotypes by focusing on specific protein classes with high proportions of non-conditional
160 phenotypes. Likewise, it can help avoid misallocation of time and resources in functional
161 genomics or breeding efforts when the likelihood of a given gene to produce a notable
162 phenotype is low. For this reason, our resource is also relevant for applied research and
163 industry, because it can be mined for trait-associated genes, including those involved in
164 symbiosis. This type of information is not available even in the most comprehensive plant
165 phenomics database developed for *A. thaliana* (Lloyd and Meinke, 2012).

166 **RESULTS**

167 **Structure, scope, and general features of the loss-of-function phenomics dataset**

168 Year 2025 marked the 30th anniversary of *M. truncatula* as a model for loss-of-function
169 genomics (Sagan et al., 1995). We conducted a very detailed human curation of 565 loss-of-
170 function studies published between 1995 and 2025. We adopted the style and structure of the
171 first large list of loss-of-function phenotypes of *M. truncatula*, which was compiled by Roy et
172 al. (2020). The list contained 115 loci with SNF phenotypes. We extended the list to 380 loci
173 with SNF phenotypes and 293 loci with non-SNF phenotypes, 673 loci in total (Figure 1).
174 Additional reputable resources contributed to the compilation of our non-SNF list. Especially
175 useful were the review by Kang et al. (2016) and multiple chapters in two books: Cañas and
176 Beltrán (2018) and Sinharoy et al. (2022). Since *M. truncatula* is primarily a model for studies
177 on SNF, we kept SNF-based lists separately and contrasted them with non-SNF lists in some
178 datasets. While importing the information about the 115 loci from Supplemental Dataset 1 of
179 Roy et al. (2020), we rechecked each phenotype using an approach centered on *M. truncatula*,
180 which resulted in some minor differences between our dataset and the dataset of Roy et al.
181 (2020). We added a column with the current locus identifiers (genome annotation v. 5.1.9).
182 Deducing this information was often difficult because not all studies provided an easy means
183 of identifying the targeted genes. We had to conduct careful analysis of primer sequences,
184 RNAi constructs, *Tnt1* flanking sequence tags (FSTs), and other data to find the exact
185 correspondence of studied loci with gene models of the current genome release. Despite these

efforts, it was not possible to ascertain the locus identifiers of seven characterized genes: *MtCIR2*, *MtCRA1*, *MtGLYI*, *MtLAL*, *MtLHCB1.4*, *MtLYK2bis*, and *MtLYK5bis*. Three of them are likely to exist only in the ecotype R108: *MtGLYI*, *MtLYK2bis*, and *MtLYK5bis*. The sequence of *MtCIR2* was kindly provided by the authors (Zhao et al., 2023). However, it has no corresponding gene model in the latest release of the genome annotation. Locus *MtuORF1p* is a special case. It is an upstream ORF (uORF) located in the 5'-UTR of a characterized transcription factor *MtNF-YA1/MtHAP2-1* (Combier et al., 2006, 2008). These eight genetic units, specifically, *MtCIR2*, *MtCRA1*, *MtGLYI*, *MtLAL*, *MtLHCB1.4*, *MtLYK2bis*, *MtLYK5bis*, and the uORF *MtuORF1p* found in the transcript of *MtNF-YA1/MtHAP2-1*, were excluded from the count of unique locus identifiers (673 unique loci listed but 681 cistrons described; uORF MtuORF1p and MtNF-YA1/MtHAP2-1 are two different cistrons of a single locus). Consequently, these eight loci are also excluded from all figures and most datasets. Details of corresponding phenotypes of these and other genes can be found in Data S1-S6. It should be noted that studies describing mutants with no linked genetic loci were beyond the scope of our analysis. Thus, the few mutants mentioned above (for example, mutants of *MtCIR2* and *MtCRA1*) are an exception because their corresponding loci were thought to be known. Data S1 (non-SNF refORFs) and Data S2 (non-SNF ncORFs) contain information on refORFs and ncORFs associated with non-SNF phenotypes, respectively. Data S3 (SNF refORFs) and Data S4 (SNF ncORFs) provide the corresponding information for SNF phenotypes. Data S5 lists genes studied both for SNF and non-SNF phenotypes. Data S6 combines the information on all 681 genes studied by loss-of-function approaches. Data S7 contains references for loss-of-function studies. Data S8 is reserved for references describing gain-of-function phenotypes of ncORFs. Our core dataset consists of six Excel files described in Table 1. Although our study focused on loss-of-function phenotypes, an exception was made for characterized ncORFs. To provide a deeper functional description of ncORFs, we also included gain-of-function phenotypes in Data S2 and S4. However, all other datasets in our study are based exclusively on loss-of-function phenotypes. The complete list of datasets pertinent to this study is shown in Table 2 (Data S7-S19). Data S10 and S11 are of special importance because they can be used as "cheat-sheets" by functional genomics specialists and researchers in the industry. Specifically, the lists can predict the likelihood of a gene giving a scorable phenotype. Data S12 is a master table listing and analyzing characterized loci that have MS-validated and/or conserved ncORFs.

The critical need for a comprehensive loss-of-function phenomics dataset in *M. truncatula*

The progress of functional genomics in *M. truncatula* has been greatly accelerated by the insertional mutagenesis using tobacco retrotransposon *Tnt1* (Tadege et al., 2008; Lee et al., 2018; Kaur et al., 2021), although the toolkit is not limited to this system (Grandbastien et al., 1989; Mazier et al., 2007; Cui et al., 2013; Vives et al., 2016). Over the past five years, on average, nearly 47 new genes have been characterized by loss-of-function approaches each year (Figure 1). With the advent of CRISPR/Cas9 (Doudna & Charpentier, 2014; Guo et al., 2023), this number will hopefully rise further each year. The total number of genes with information on mutant phenotypes reached 673 by 16 June 2025, which is only a fraction of the corresponding number of genes characterized in *Arabidopsis* (Lloyd and Meinke, 2012). Nevertheless, 673 is the number large enough to require systematic recording, curation, and meta-analysis. The lack of a standardized protocol for collecting and displaying information on newly characterized genes compromised the status of *M. truncatula* as a functionally annotated model. When a researcher identifies a gene that looks interesting for the functional analysis, in the first place, a question arises whether the gene has already been targeted by other groups. Unfortunately, this information is not readily available for 60% of genes out of 673 characterized so far (Figure 2a; Data S6). In this context, it is important to emphasize the

central role of the *M. truncatula* genome browser v. 5 developed in 2018 (Pecrix et al., 2018). This indispensable resource is the hub linking many types of data and displaying them in a well-organized, user-friendly format for the maximal benefit of the *Medicago* community. However, integrating old and new functional genomics data into the genome browser is a major technical challenge. This is because the truly comprehensive and error-free process of collecting such data cannot be easily automated. For example, it is very unlikely to be efficiently outsourced to artificial intelligence. The team curating the browser made an enormous effort and accurately referenced loss-of-function studies for 269 genes, which is 40% of the total (Figure 2a; Data S6, "Yes" in Columns V and W). Unfortunately, 173 genes (26%) completely evaded the functional annotation (Figure 2b; Data S6, "No" in Columns V and W), while 218 genes (32%) were provided with references that do not inform the users about the corresponding loss-of-function analyses (Figure 2a, categories "Yes/YN" and "No/YN"; Data S6, "NY" in Column W). Naturally, newer studies tend to be missed at a higher rate compared to older ones. For example, more than half of genes functionally characterized since 2023 are not labeled as such by the end of July 2025 (Data S6 Column W). However, the population of 173 "missed" genes ("No/No" category in Figure 2a) extends in time back to 2001, with 81 genes characterized more than five years ago (Figure 2b; Data S6). This lack of integration into the browser, even years after the original publication, reflects the difficulty of manual curation of such data. We found no bias toward any method, phenotype, or protein class among those "missed" genes. They were overlooked simply due to the lack of a centralized system integrating all functional data into the browser. These facts strongly suggest the need for such a system. Possibly, it could be developed with a Wikipedia-like format, with authors encouraged or obligatorily requested to enter all relevant information about genes at the time they are submitted for publication. One reason for such an open format is the difficulty of allocating this task to a specific researcher/curating team member. Because dozens of new genes are characterized each year, keeping functional data up to date requires significant and continuous effort. For example, for our team, it took nearly three years to collect and systematize information about 673 genes. This problem can be efficiently addressed by agreeing upon a set of rules and implementing these rules for all researchers who publish functional characterization studies on *M. truncatula* genes. Our dataset can be the basis for such a system, which can be further developed in real time as more genes are characterized each year.

Specific examples where a unified loss-of-function database could help avoid redundancy and confusion

An important question is whether this lack of functional annotation had any negative impact on studies in *M. truncatula*. Can we find examples of such impact through the meta-analysis of our dataset? Unfortunately, yes. Column X (Notes) in Data S6 describes a few such instances. They are categorized below.

Multiple names for the same gene: potential duplication of efforts

The first loss-of-function study on locus MtrunA17_Chr5g0422541 (*MtAOC1*) was published in 2005 (Isayenkov et al., 2005). Recently, a different team characterized this gene very completely, too, and named it *MtAOC2* (Yang et al., 2023). Both studies were published in Plant Physiology and covered different aspects. The second team, and the editor of the journal, were unaware of the 18-year-old study on this gene. It is evident that no researcher or editor would ever wish to miss such important information by the time of the intended publication, or ideally much earlier. Had it been available through a centralized resource, it would greatly facilitate planning, experimental design, and manuscript writing for the second team and would have avoided confusion for anyone wishing to study this gene in the future.

The fact that this gene was targeted by two high-quality studies is still not reflected in the genome browser, which exemplifies the difficulty of manual curation of such data. To further complicate the situation, other loci were called *MtAOC2* (MtrunA17_Chr7g0220911) by Isayenkov et al. (2005) and *MtAOC1* (MtrunA17_Chr4g0034431) by Yang et al. (2023), neither of which were targeted by loss-of-function studies so far. Developing a consistent and non-redundant set of gene names is another important and challenging task that can be addressed efficiently by a unified resource.

Another locus with a similar fate is MtrunA17_Chr8g0343011. It was called *MtLOX9* in the first loss-of-function study (Li et al., 2022), *MtLOX24* in the second, unrelated study (Xu et al., 2024), and *MtLOX4* in the genome browser v. 5.1.9. Like in the first example, the authors of the second study were unaware of the previous work on this gene, which is not referenced in the genome browser. It may be argued that the problem is minor because the groups studied different processes. However, under certain circumstances, funding, time, and material resources could be wasted by duplicating earlier efforts. This possibility should serve as a motivation for the research community to establish a system that integrates all loss-of-function data into corresponding genome browsers, not only for *M. truncatula*.

One name for multiple genes: ambiguity

A case of the opposite type can be exemplified by studies that gave the same name to different and sometimes unrelated loci. *MtCHS* was a name given to three different genes: MtrunA17_Chr7g0252331 (Zhang et al., 2009), MtrunA17_Chr7g0220211 (Samac et al., 2011), and MtrunA17_Chr1g0201391 (Cook et al., 2025). None of these loci are named or referenced in the genome browser v. 5.1.9. Other examples are theloci called *MtMYB1* and *MtbZIP60* in original articles. *MtMYB1* is MtrunA17_Chr7g0242521 in Floss et al. (2017) but MtrunA17_Chr3g0080861 in Li et al. (2022). *MtbZIP60* is MtrunA17_Chr1g0171291 in Ribeiro et al. (2022) but MtrunA17_Chr4g0035414 in Xing et al. (2024). Only one of them is correctly referenced in the genome browser v. 5.1.9 (*MtMYB1*, Floss et al., 2017). To avoid confusion, we added lowercase letters to the names of these seven genes, in the chronological order (*MtCHSa*, *MtCHSb*, *MtCHSc*; *MtMYB1a*, *MtMYB1b*; *MtbZIP60a*, *MtbZIP60b*, respectively). *MtSymSCL1* was a name given to locus MtrunA17_Chr1g0183641 by Kim & Nam (2013) and locus MtrunA17_Chr4g0057891 by Rey et al. (2017). In the current dataset, we use alternative names of these genes: *MtSCL2* and *MtRAD1*, respectively, which are correctly named and referenced in the genome browser v. 5.1.9.

Converging acronyms: even more ambiguity

The loci *MtCRA1* and *MtCRA2* fall into a special sub-category. In contrast to the genes mentioned above, these names were given to unrelated loci. A locus with still unknown identity (Data S6) was called *MtCRA1* for COMPACT ROOT ARCHITECTURE 1 (Scholte et al. 2002; Laffont et al., 2010). Locus MtrunA17_Chr3g0140861 was named *MtCRA2* (the same acronym) in Huault et al. (2014) and several subsequent studies. However, two other genes were named *MtCRA1* and *MtCRA2* for CHITINASE-RELATED AGGLUTININ 1 and 2, respectively (Data S6), although they have not been targeted by loss-of-function studies so far (Tian et al., 2013; Zhang et al., 2016a). Thus, their names are not revised and were not included in the current dataset.

Incorrect locus naming in the genome browser v. 5.1.9

We found numerous examples of gene names that conflict with naming from original loss-of-function studies. We summarized these examples in Tables 3 and S1. At least 11 names were incorrectly assigned to other loci in the genome browser v. 5.1.9. These names were already used for genes characterized by loss-of-function approaches. Two names, *MtChOMT1* and

MtPRP1, were misassigned to two genes each. For example, MtrunA17_Chr8g0366751 is the locus ID of a gene called *MtMYC2* by Guo et al. (2024). However, in the genome browser, a different member of this family was named *MtMYC2*: MtrunA17_Chr5g0411341. This gene was called *MtBHLH2* by Deng et al. (2019) (see Data S1, Data S3, and Table S1). These examples further emphasize the need for a unified resource that would implement and regulate gene naming based on original publications or some other systematic principle.

Splitting one characterized gene model into two

Genome annotation is a dynamic process. Naturally, some older gene models characterized some time ago do not correspond to a single locus in the current genome annotation (v. 5.1.9). Such models are now split into two loci each. This is relevant to the annotation of these loci because it introduces functional ambiguity. Deciding which of two resulting loci was affected in an original study is not a trivial task. In some cases, both loci might have been affected. In a unified system proposed here, such a decision may require an independent expert who re-examines the original data. Alternatively, it can be based on the manual input of a team that produced the data. In this article, we simply draw attention to such models, with a few examples.

Functional analysis of v4 locus Medtr7g113890 (*MtGSHSa*) was conducted more than 20 years ago (Frendo et al., 2001, 2005). Now this gene model is split into two loci: MtrunA17_Chr7g0273151 and MtrunA17_Chr7g0273161. Which of these two transcripts was affected by two antisense constructs in the original study? Probably both, even though they share no sequence similarity. Although the article provides no sequence data of their constructs, the mRNA of locus MtrunA17_Chr7g0273151 was likely to be co-affected with a related locus MtrunA17_Chr7g0273131 by the shorter of the two antisense constructs (1.1 kb) because this pair has 76% identity over the length of 859 nt. The mRNA of the second locus, MtrunA17_Chr7g0273161, was probably co-affected with a related locus MtrunA17_Chr7g0273141 by the longer of antisense constructs (1.8 kb) because this pair has 75% identity over the length of 815 nt. Thus, the reported phenotype can be attributed to all four loci unless a dedicated study shows otherwise. Unfortunately, this information is not reflected in the genome browser v. 5.1.9, nor is the fact that the loci were targeted by a loss-of-function approach. In this and a few other cases, we attempted to fill this gap by assigning unique but related names to these loci (*MtGSHSa* and *MthGSHSa*; *MtGSHSb and MthGSHSb*, respectively). Specifically for these homologs, we modified the original names introduced by Frendo et al. (2001, 2005), where “h” stands for “homo”. They are fully explained in Supporting Results. This group of genes presents special interest to our study and will be discussed in detail in the context of ncORFs (Data S12 to S14 and Data S16).

The second v4 locus of this type is Medtr4g088055 (*MtROP7*, Wang et al., 2021), which is now split into MtrunA17_Chr4g0046211 and MtrunA17_Chr4g0046221. These two loci are separated by only one nucleotide (A) in the plus-strand and share no sequence identity. We propose names *MtROP7a* and *MtROP7b*, respectively, for these loci. RNAi constructs used in the original study (Wang et al., 2021) target the first locus (*MtROP7a*) together with many other genes, including *MtROP8* and *MtROP10*. The alignment of these RNAi constructs with the second locus (*MtROP7b*) yields no obvious similarity, except for a 40-nt-long region having 88% identity with the RNAi construct produced by primers Forward (MROP10) 1 and Reverse (MROP10) 1 (Supplemental Table 2 in Wang et al., 2021). Is this very local similarity sufficient to affect *MtROP7b*? Probably not, because much longer sequences (>130 nt) are typically used for efficient RNAi-mediated silencing in plants (Kakiyama et al., 2019). Only the *MtROP7b*-specific qRT-PCR analysis can tell. With so little evidence linking this gene to the conditional phenotype reported by Wang et al. (2021), *MtROP7b* would be removed from our

379 dataset as being unaffected by the loss-of-function experiment. However, the following
380 consideration shows how important it is to look deeper into such examples and to inspect the
381 original qRT-PCR data on a case-by-case basis. Analysis of *MtROP7*-designed qRT-PCR
382 primers used by Wang et al. (2021) clearly shows that the new model is wrong and the old
383 one (v4) is correct (Supplemental Table 2 in Wang et al., 2021). The forward primer
384 (TTATGCTCCAAATGTGCCTA) is located in *MtROP7a* whereas the reverse primer
385 (TGTTTTAGAGCTGCACTCA) resides in *MtROP7b*. They can form an amplicon of 269 bp
386 exclusively from the cDNA of the old model. If the signal reported in Figure S5 of Wang et al.
387 (2021) had come from contamination with genomic DNA, the amplicon would have been 401
388 bp long because of Intron 4. The melting curve analysis would certainly have alerted the
389 researchers about such contamination. Thus, the reported data is very likely to be correct and
390 unambiguous. This is not the only example of a "longer" original model making more sense
391 because of being supported either by transcriptomic or by MS proteomic data (see the
392 example of *MthGSHSb* in Data S16 Slide 7). Nevertheless, it is important to look deeper into
393 the reason for rejecting the old model by the current genome annotation. The 58-bp-long
394 junction between the ORFs of *MtROP7a* and *MtROP7b* is free of stop codons only in frame 3
395 relative to the sequences, which is not the reading frame of the annotated ORFs. Thus, the
396 sequences are separated by in-frame stop codons. Possibly, the junction is a spliced-out
397 intron, which could explain why in-frame stop codons in this region were tolerated by natural
398 selection. It is worth noting that neither of these two loci is currently provided with the correct
399 reference in the genome browser v. 5.1.9 (Data S6).

400 Another example of this sort is v4 locus Medtr4g073400 (*MtSYT1*, Gavrin et al., 2017), which
401 is now split into MtrunA17_Chr4g0037111 and MtrunA17_Chr4g0037121 (*MtSYT1a* and
402 *MtSYT1b*, respectively, in our dataset). As in the case of *MtROP7*, the loci are separated by
403 a single nucleotide (A) in the plus-strand. Likewise, qRT-PCR primers MtSyt1-F
404 AGGAGGCCCGTTGGAATTTT and MtSyt1-R GTGATGCTTCCCCTCAACA (Supplementary
405 Table 1 in Gavrin et al., 2017) span the junction between the two loci, forming an amplicon of
406 625 bp from both cDNA and genomic DNA, according to the new model. The length of 625 bp
407 is clearly suboptimal for qRT-PCR (normal range is between 60 and 150 bp, Udvardi et al.,
408 2008). Evidently, the qRT-PCR signal shown in Supplementary Figure 7 of Gavrin et al. (2017)
409 could come from the cDNA corresponding to the "old" model, if the junction that contains in-
410 frame stop codons was spliced out. We doubt that the qRT-PCR settings used in that high-
411 quality study would permit the generation of a signal exclusively from the contamination with
412 genomic DNA, which would result in this unusually long amplicon (625 bp). For this reason, in
413 our dataset, we consider *MtSYT1a* and *MtSYT1b* as part of the same functional unit, as it was
414 described in the original study. If the new model is correct, it is impossible to find out if both
415 loci were affected by the generic RNAi construct designed for simultaneous silencing of
416 *MtSYT1* and *MtSYT3*, which share no sequence identity. This is because RNAi primers
417 MtSyt1-F TGGATCCATCACAGGCCATG and MtSyt1-R CGCAAACGGATTTGTGTGGT
418 (Supplementary Table 1 in Gavrin et al., 2017) do not perfectly match the cDNA of the split
419 models MtrunA17_Chr4g0037111 and MtrunA17_Chr4g0037121. Thus, functional ambiguity
420 does not permit exclusion of either locus from the reported RNAi phenotype even though only
421 MtrunA17_Chr4g0037111 is currently provided with the correct reference in the genome
422 browser v. 5.1.9 (Data S6).

423 The last example of a split locus is v4 model Medtr1g102190 (*MtLICK1*, Wang et al., 2025). It
424 corresponds to two adjacent loci in the genome assembly v. 5.1.9: MtrunA17_Chr1g0204221
425 and MtrunA17_Chr1g0204231 (*MtLICK1a* and *MtLICK1b*, respectively, in our dataset). They
426 are separated by a nucleotide triplet ACA in the minus-strand. If this newer model is correct,
427 only MtrunA17_Chr1g0204221 was affected by the *Tnt1* insertion in Wang et al., 2025. Based
428 on FST NF9810_low_9 (the *Tnt1* mutant collection, Tadege et al., 2008; Lee et al., 2018; Kaur

et al., 2021), the insertion is located in the 5'-UTR of this gene. However, locus MtrunA17_Chr1g0204221 is a direct continuation of locus MtrunA17_Chr1g0204231. The high similarity of exon-intron structures shown in Extended Data Fig. 2k and i (Wang et al., 2025) suggests that the new model may be incorrect. Thus, the insertion is predicted to be in the last exon of the v4 model. If it were the 5'-UTR, the mutant phenotype would probably not differ from the wild type. The study provided further proof for the validity of the original model by MS proteomics. Specifically, Table 1 in Wang et al. (2025), which illustrates MtLICK1 phosphorylation sites, lists a peptide MSI(pT)VGAAEGLAYLHHEANPHIIHR with the original position of the phosphorylation site at T151. It maps to the 3'-UTR of MtrunA17_Chr1g0204231 in such a way that 306 nt located between MSITVGAAEGLA and YLHHEANPHIIHR in the current model are missing. The same concerns peptides RM(pS)ITVGAAEGLAYLHHEANPHIIHR (S149) and RMSITVGAAEGLA(pY)LHHEANPHIIHR (Y160). This indicates that this part of the junction between MtrunA17_Chr1g0204231 and MtrunA17_Chr1g0204221 is a spliced-out intron, like in the cases of *MtROP7* and possibly *MtSYT1*. In contrast, peptides DIKA(pS)NVLLDTEFQAK (S177) and ASNVLLD(pT)EFQAK (T183) map exactly to the junction site between MtrunA17_Chr1g0204231 and MtrunA17_Chr1g0204221, which means this region is a translated exon. Thus, we consider MtrunA17_Chr1g0204231 and MtrunA17_Chr1g0204221 as a single functional unit in our dataset, as it was described in the original study. Unfortunately, neither of these loci is provided with the correct reference in the genome browser v. 5.1.9 (Data S6).

Duplicated gene models

A different type of challenge in assigning a mutant phenotype to an annotated locus is exemplified by a v4 gene Medtr4g125520. It corresponds to two entries in the genome assembly v. 5.1.9: MtrunA17_Chr4g0070611 and MtrunA17_Chr4g0070641. In contrast to the splitting of an old model into two, as described in the previous section, these two loci are separated by several other genes and repeat elements. Although the gDNA sequences of MtrunA17_Chr4g0070611 and MtrunA17_Chr4g0070641 are identical base-for-base over the shared region, the loci are located in separate genomic positions. The shorter locus (MtrunA17_Chr4g0070641) entirely lacks the region corresponding to the end of Exon 3 of MtrunA17_Chr4g0070611. The gene model presented in Figure 5a of Liu et al. (2014) corresponds to MtrunA17_Chr4g0070611. Thus, only this longer locus is listed in the current dataset. However, if both models are correct, it is impossible to tell which one was affected by the *Tnt1* insertions because the gDNA sequences around the insertion sites are identical in these two genes. We attempted to solve this ambiguity by analyzing FST sequences of *Tnt1* insertions corresponding to lines *myb14-1* and *myb14-2* (Figure 5a of Liu et al., 2014). These are FSTs NF14565_low_42 and NF0084-INSERTION-2, respectively, both in reverse orientation. The first FST is not informative for our purpose because its end reaches only 14 bp upstream of the translational start (ATG) codon of MtrunA17_Chr4g0070611 and MtrunA17_Chr4g0070641. These 14 bp do not match the gDNA of either locus in the Jemalong A17-based genome browser v. 5.1.9. They are likely to present a sequencing error because they also do not match a corresponding model in the ecotype R108 (the background of the *Tnt1* mutant population; Tadege et al., 2008; Lee et al., 2018; Kaur et al., 2021). The second FST, NF0084-INSERTION-2, gives more information because its end reaches 612 bp upstream of the ATG. While corresponding regions in the Jemalong A17 genome are identical between MtrunA17_Chr4g0070611 and MtrunA17_Chr4g0070641, there is a single matching sequence in the R108 genome. It is immediately upstream of a gene nearly identical to MtrunA17_Chr4g0070611 at the gDNA level (only three mismatches over the entire length of MtrunA17_Chr4g0070611). Thus, we conclude that the second (short) copy may exist in the ecotype Jemalong A17 but is absent from the ecotype R108, which suggests a very recent

duplication. This is further supported by the qRT-PCR-mediated test for gene copy number presented in Supplemental Figure S11 of Liu et al. (2014). Primers MtMYB14qPCR1-F AGGCAACATATCCTCAGATGAAG and MtMYB14qPCR1-R TCGTGTTCCAATAATTCTTGATTTCA (Supplemental Table 4, Liu et al., 2014) reside in a region shared between MtrunA17_Chr4g0070611 and MtrunA17_Chr4g0070641. The signal generated by these primers corresponds to a single genomic copy. This is an example leaving a big question about the functional status of both loci in Jemalong A17: which of them is a functional ortholog of the single copy in R108? Even with a targeted mutagenesis tool such as CRISPR/Cas9 it may be challenging to discriminate between these loci because of the high nucleotide identity within, upstream, and downstream of the sequences. Future functional annotation of such loci should probably indicate ambiguities related to the differences between the genomes of ecotypes A17 and R108.

Loss-of-function methodology and biological processes studied in *M. truncatula*

Diverse methods were used for loss-of-function analyses of 673 *M. truncatula* genes (22 approaches, Figure S1). More than half of the studies applied one (or a combination) of just two methods: insertional mutagenesis of tobacco retrotransposon *Tnt1* (359 genes) (Tadege et al., 2008; Lee et al., 2018; Kaur et al., 2021) and post-transcriptional silencing mediated by RNAi (265 genes) (Sharp & Zamore, 2000; Baulcombe, 2004; Chaudhary et al., 2024). As a new method, gene editing with the aid of CRISPR/Cas9 (Doudna & Charpentier, 2014; Guo et al., 2023) was used for only 87 genes so far. The popularity of RNAi as a tool is remarkable given the known limitations and ambiguity associated with this method. Technically, it is a simple and inexpensive approach when the generation of transgenic roots (so-called hairy roots, Nilsson & Olsson, 1997; Limpens et al., 2004), rather than the entire transgenic plants, is sufficient for the analysis. Out of 117 genes first characterized in 2023-2025, 30 were studied by RNAi, which is almost the same as the number of genes studied by CRISPR/Cas9 in the same period (31 genes). *Tnt1* was used for more than half of the 117 genes (67 genes), followed by FNB (11 genes), EMS (four genes), and AS (one gene) in the same period. FNB, EMS, and AS stand for Fast Neutron Bombardment mutagenesis (Li et al., 2001), Ethyl Methanesulfonate mutagenesis (Sega, 1984), and Antisense-mediated gene silencing (Waterhouse et al.,1998), respectively. From this data, it is evident that the *Tnt1* resource plays a pivotal role in the functional genomics of *M. truncatula* (Data S6, Figure S1).

M. truncatula was used for studies on 97 biological processes (Data S9, Figure S2). However, this legume was primarily established as a model for SNF. More than half of 673 genes were tested for SNF (380 genes), followed by root development (166 genes), and AM (141 genes). Among non-symbiotic processes other than root development, *M. truncatula* was frequently tested for leaf development (104 genes), flower development (60 genes), control of body size (48 genes), floral transition (48 genes), and pod development (40 genes). Seed-related studies are also numerous in *M. truncatula* (93 genes): seed development (47 genes), seed composition (27 genes), seed germination (8 genes), seed longevity (5 genes), seed coat function (3 genes), and seed dormancy (3 genes).

More than half of the genes studied in *M. truncatula* are transcription factors and enzymes

Transcription factors (TFs) and enzymes are two protein groups prevailing in the current dataset (224 and 205 genes, respectively). Counts of these and other protein groups are shown in Figure 3, Figure S3, Data S10, and Data S11. Among TFs, the top four most represented families are AP2-EREBP (21 genes), GRAS (19 genes), MYB-HB-like (18 genes), and SBP (17 genes). Among enzymes, the most represented classes are kinases (48 genes), oxidoreductases (45 genes), transferases (43 genes), and hydrolases (39 genes). Given the

527 pivotal role of transport proteins in SNF, AM, and other biological processes, transporters (73
528 genes) and channels (13 genes) are also well-represented in the current dataset. Remarkably,
529 transport substrates were found for many of these proteins (60 transporters and 7 channels),
530 which is usually a challenging task.

531 **Non-conditional loss-of-function phenotypes are found in 62% of genes studied in *M.***
532 ***truncatula***

533 We have categorized phenotypes of 673 genes studied by loss-of-function approaches into
534 three categories. This classification was based exclusively on the genetic conditions.
535 Specifically, it did not consider conditions such as different growth media or temperature
536 regimes. Phenotypes caused by single genes were categorized as non-conditional (A for
537 altered). Phenotypes caused by more than one gene were regarded as conditional (C for
538 conditional). Neutral phenotypes (N for neutral) were those that did not differ from the wild
539 type, regardless of the number of genes affected in each line. The overall proportions of A-,
540 C-, and N-phenotypes were 62%, 27%, and 11%, respectively (Figure S4). However, these
541 proportions differed slightly between SNF phenotypes and non-SNF phenotypes (compare
542 Figures S4a and S4b). The proportion of C-phenotypes was higher among genes tested for
543 SNF (31%), compared to non-SNF phenotypes (26%). At the same time, the proportions of N-
544 phenotypes did not differ much (16% and 15% for SNF and non-SNF, respectively). Only 53%
545 of genes tested for SNF gave non-conditional phenotypes (A-phenotypes), which suggests
546 functional redundancy due to the robust structure of the genetic program controlling SNF. This
547 relatively low "success rate" of loss-of-function studies on SNF reflects the challenge in
548 identifying crucial genes that can be used for the improvement of SNF in legume crops and
549 synthetic biology efforts to make non-legume plants undergo SNF. With some effort, Data S6
550 can be used by the research community as a resource for predicting the "success rate" of
551 planned loss-of-function studies for specific biological processes, as we showed for SNF. For
552 example, all 11 genes tested for cold stress showed an A-phenotype for this biological process
553 (higher sensitivity or lower sensitivity). Hence, other genes possibly involved in cold stress are
554 likely to have scorable A-phenotypes, too. In contrast, out of 60 genes tested for flower
555 development, not more than half exhibited A-phenotypes. Because Data S6 organizes
556 phenotypes at the gene level, additional processing is required to examine them in the context
557 of individual biological processes tested in each study. In the next section, we focus on
558 statistically significant differences in this "success rate" among different protein classes and
559 groups of classes. Extracting this information from Data S6 is simple because phenotypic
560 categories are linked to genes. Moreover, we find this type of data more relevant for academic
561 research and breeding efforts because it can tell which gene groups are more likely to yield a
562 scorable phenotype.

563 In the remaining part of this section, we present data on the agreement between phenotypic
564 categories when the same genes are tested for SNF and non-SNF phenotypes (Figure S5).
565 Among 219 genes tested for SNF and at least one non-SNF phenotype (Data S5), 40% had
566 A-phenotypes for both processes (group AA). For conditional (CC) and neutral (NN)
567 phenotypes, this proportion was 25% and 5%, respectively (Figure S5). Cases of discordance
568 regarding phenotypic categories were not very prominent among genes tested for both types
569 of biological processes. The largest discordance group (10%) was represented by genes that
570 exhibited A-phenotypes when tested for non-SNF processes and N-phenotypes when tested
571 for SNF (group AN in Figure S5). The reciprocal group, NA, accounts for only 5% of 219 genes,
572 which emphasizes that purely SNF-specific genes are not very abundant. These observations
573 suggest the relatively large number of genes with pleiotropic effects, as can be seen also for
574 various non-SNF phenotypes in Data S6. In summary, the chance that a gene selected for
575 analysis will give an A-phenotype for at least one biological process is 62% (Figure S4c). If a

576 gene shows an A-phenotype for SNF, in 40% of cases, it will also have a scorable A-phenotype
577 for at least one non-SNF process (Figure S5b). This may reflect the adoption of non-symbiotic
578 genes for symbiotic processes reported in many studies (Limpens et al., 2003; Soyano et al.,
579 2019; Huisman & Geurts, 2019; Liu & Bisseling, 2020; Hawkins & Oresnik, 2022).

580 **Unique proportions of phenotypic categories in different protein classes may reflect**
581 **their relative functional importance**

582 As shown in Figure 3, the proportions of phenotypic categories are not the same for different
583 protein classes. Likewise, different groups of combined protein classes have different
584 proportions of phenotypic categories (Figure S3). The natural question is whether these
585 differences are statistically significant. Data S10 and S11 present statistical analyses of these
586 differences for protein classes and larger groups, respectively. Sheet "Sorted by count" in Data
587 S10 shows which classes have A:C:N ratios significantly different from the overall ratio of
588 62:27:11 (see also Figure S4c). Four classes differ significantly from that ratio: TFs of the SBP
589 family and the C2C2-Dof family, channels with unknown substrates, and protease inhibitors.
590 TFs of the SBP family have the ratio of 6:94:0, which corresponds to 16 out of 17 tested SBP
591 genes having C-phenotypes. This, however, is a special case because nearly all these SBP
592 genes were affected by miRNA in a single study (Wang et al., 2019). A somewhat similar
593 situation is observed in TFs of the C2C2-Dof family (the ratio of 8:0:92), where 12 out of 13
594 genes were targeted in a single study (Zhang et al., 2019). In contrast to the SBP family, these
595 genes were targeted in separate *Tnt1* lines and exhibited N-phenotypes. Channels with
596 unknown substrates fall into the same category (the ratio of 17:0:83), with four out of seven
597 genes targeted in a single study (Charpentier et al., 2016). Protease inhibitors have the ratio
598 of 20:0:80, with four out of five genes exhibiting N-phenotypes. Much more interesting
599 differences are observed when individual classes are cross-compared with each other rather
600 than with the overall ratio of 62:27:11. Sheet "Example comparisons" in Data S10 highlights
601 highly significant differences between related classes such as kinases and oxidoreductases,
602 transferases and hydrolases, TFs of the AP2-EREBP family and the MYB-HB-like family, etc.
603 For example, kinases have a much higher proportion of A-phenotypes compared to
604 oxidoreductases and a much lower proportion of N-phenotypes, which reflects their crucial
605 roles in various processes. Likewise, TFs of the MYB-HB-like family have a much higher
606 proportion of A-phenotypes compared to other TF groups, such as the AP2-EREBP family and
607 the GRAS family. TFs of the AP2-EREBP family have an unusually high proportion of C-
608 phenotypes (57%) but no N-phenotypes among 21 tested genes. This information can be of
609 use for practical purposes because it provides a quantitative measure of relative functional
610 importance and redundancy.

611 When combined functional groups are considered (Data S11), similar significant differences
612 are evident. Channels stand out as a group with the A:C:N ratio of 62:0:38, which is
613 significantly different from the overall ratio of 62:27:11. They have an abnormally low
614 proportion of C-phenotypes and an abnormally large proportion of N-phenotypes. Likewise,
615 the combined group of protein classes designated as "Other" in Data S11 (see Data S10 for
616 the classes included in this group) has the ratio of 73:12:15, which is significantly different
617 from the overall ratio of 62:27:11. It has an abnormally high proportion of A-phenotypes. Sheet
618 "Example comparisons" in Data S11 exemplifies significant differences between combined
619 groups. For example, the ratio of 71:18:11 for transporters is significantly different from the
620 ratio of 53:34:13 for TFs, which literally translates into a higher chance of observing a scorable
621 A-phenotype when genes under study are transporters.

622 It should be noted that 73% of A-phenotypes (group Other) is the largest proportion only within
623 Dataset S11 (differences between combined groups). Much larger proportions can be found

624 in individual classes with at least 5 genes per class (Data S10). The following classes have
625 proportions of A-phenotypes larger than 73%: Cytochrome P450 (100%, n=6), TF (bZIP
626 family) (100%, n=5); TF (NAM family) (83%, n=6), TF (C2H2 family) (80%, n=10); TF
627 (Homeobox-WOX family) (80%, n=5); and TF (MYB-HB-like family) (78%, n=18). TFs exhibit
628 a very large variation in proportions of A-phenotypes. For example, the following TF families
629 have rather low proportions: TF (SBP family) (6%, n=17); TF (C2C2-Dof family) (8%, n=13);
630 TF (Homeodomain-TALE-KNOX family) (11%, n=9); TF (ARF family) (22%, n=9); TF
631 (Hap3/NF-YB family) (27%, n=11); TF (HD-SAD family) (40%, n=5); and TF (AP2-EREBP
632 family) (43%, n=21).

633 Lastly, it is worth noting protein classes with the largest proportions of N-phenotypes (Data
634 S10): TF (C2C2-Dof family) (92%, n=13); Channel (substrate unknown) (83%, n=6); Protease
635 inhibitor (80%, n=5); Transporter (proton efflux) (60%, n=5); Membrane protein (PLAT domain
636 containing) (40%, n=5); TF (Hap3/NF-YB family) (36%, n=11); Peptide (CEP family) (25%,
637 n=8); Transporter (substrate unknown) (23%, n=13); and Enzyme (oxidoreductase) (22%,
638 n=45). Channels, regardless of the substrate, are also prominent in Dataset 11, with the
639 largest proportion of N-phenotypes among combined groups (38%, n=13).

640 In practical terms, the meta-analysis presented in this section may encourage the research
641 community to focus on classes with the higher "success rate". However, it should not bias
642 researchers and funding agencies against candidate genes with relatively low "success rates"
643 as they can still be highly informative and useful in breeding programs and synthetic biology
644 efforts (e.g. *MtDASH*, *MtDMI1*, *MtHA1*, and many other genes in Data S6).

645 **NcORFs in *M. truncatula* genes targeted by loss-of-function approaches**

646 Due to space limitations, the entire Results section dedicated to ncORFs was placed into
647 Supporting Results, which are available in the online version of this article as a separate file.
648 This section is crucial to the message of this paper, which can be summarized as follows:
649 ncORFs cannot be ignored in functional studies as they introduce ambiguity regarding the
650 interpretation of mutant phenotypes. The section references Data S12-S18 and has its own
651 list of literature sources.

652 **DISCUSSION**

653 **The genome annotation of *M. truncatula* requires better integration of loss-of-function**
654 **data**

655 The latest release of the *M. truncatula* genome (v. 5.1.9) is an indispensable reference for
656 plant biologists, especially for studies that take advantage of the symbiotic properties of this
657 model organism (Pecrix et al., 2018). However, the resource lacks proper integration of
658 information on loss-of-function phenotypes. This concerns not only the detailed description of
659 phenotypes but also relevant references to loss-of-function studies (Figure 2). Consequently,
660 a researcher who considers targeting a locus of interest may be unaware of any previous work
661 on that locus. This limitation has multiple negative effects on the progress of functional
662 genomics in *M. truncatula*, including duplication of efforts, assigning multiple names to the
663 same genes, and being unable to benefit from the evidence generated in previous
664 experiments. It could be argued that obtaining such information is the responsibility of
665 researchers. However, a well-annotated and user-friendly resource should aim to facilitate
666 information retrieval, not merely leave it to the users. The tool should minimize not only the
667 effort but also the ambiguity as to which locus was affected, with little space left for
668 misinterpretation. During our meta-analysis, we have noticed that establishing a correct link
669 between a locus described in a loss-of-function study and the latest identifier of that locus is
670 not always a trivial task. Often, it requires careful sequence analysis of primers, RNAi

segments, and even the conversion of DNA/protein sequences from image to text. This concerns not only old studies conducted before the advent of the genome browser but also new studies. Our dataset can be integrated into the genome annotation of *M. truncatula* for the maximal benefit of the research community.

Keeping the loss-of-function domain of the browser up to date may require a human-curated protocol

Once the data becomes part of the genome browser, updating it in real time, or at least frequently, presents a considerable challenge. We have noticed that many details of the phenotypes and other relevant information summarized in Data S6 may be extremely difficult to extract automatically. A human-curated protocol allowing direct submission may be the only adequate way to record and consolidate error-free data on loss-of-function phenotypes. This, however, is not a task that can be handled in a background mode by a dedicated staff member of the genome browser. Integrating numerous functional genetic studies (Figure 1) can be highly time-consuming, particularly given the steadily increasing number of publications each year. Therefore, we propose that such information be requested directly from the authors of loss-of-function studies. A convention can be established as to the rules and the format of entering the information on newly characterized genes directly into the browser shortly before a publication is transferred to the print stage by a journal. Each submission can have a unique identifier that researchers can cite in their publications. If successful, this concept can become a standard not only for *M. truncatula* but also for other species. Generalizing such a submission system, however, would be the next challenge, this time of an organizational type. It would require coordination between multiple institutions, resources, and publishers in such a way that the journals make it compulsory for authors to register their mutant data. Importantly, this should concern not only phenotypes of genes that were clearly in the focus of a study (for example, their names were mentioned in a title or in an abstract) but also phenotypes of mutants that had already been published in the same or a different context earlier. Typically, such mutants have an auxiliary role in a study. However, they often provide additional details unnoticed or not addressed in the original articles. Many phenotypes in our dataset fall into this category. Data S1-S6 also list genes that were mutagenized only once; however, they did not constitute the core of articles describing their phenotypes. Thus, such genes can be overlooked in the literature because the fact of their mutagenesis was not highlighted anywhere except for a few sentences and/or supplementary figures. Extracting such information and assigning it accurately to specific loci requires a human effort that cannot be easily automated.

The network linking the genome browsers and the journals should be the future of functional genomics. It would eliminate many limitations mentioned in the previous section, thus facilitating the progress of biological research. At the same time, one format described in the next section can serve as a model for other resources (Baldarelli et al., 2024). It was successfully implemented with an option of direct submission for mouse mutants.

Strong points and limitations of existing loss-of-function databases

Historically, the first attempt to create a multi-species genotype/phenotype database was made at the dawn of the genomic era, before the *M. truncatula* genome became well-annotated. This database, called PhenomicDB, included mouse, zebrafish, fruit fly, *Caenorhabditis elegans*, yeast, and *A. thaliana* (Kahraman et al., 2005; Groth et al., 2007). It contained 399,772 phenotypes associated with 77,400 genes. Unfortunately, the URL of PhenomicDB, http://www.phenomicdb.de, has been offline since 2010. It is not available in an alternative form.

Currently, the most advanced loss-of-function resource for animal organisms is Mouse Genome Informatics (MGI) (Baldarelli et al., 2024). It is closely associated with MUTAGENETIX™, a database of N-ethyl-N-nitrosourea (ENU)-induced mutations that cause phenotypes in mice (Wang et al., 2015). MGI is a well-structured, comprehensive database containing information on 20,583 genes with phenotype annotations (as of 25 February 2026). It enables filtering phenotypes by the mutagenesis method (the methods are not limited to the treatment with ENU), allele attribute (conditional, dominant negative, no functional change, etc.), type of mutation (deletion, insertion, duplication, etc.), and even the anatomical system affected by the phenotype. It shows human homologs of affected genes and homologs in other mammalian organisms. Importantly, the genome browser of MGI has a comprehensive portfolio for each locus describing all possible aspects of corresponding mutants, including references to original experimental studies. The resource accepts direct submissions of phenotype data, which makes it ideal for efficient functional analysis. Its format and depth should be the model for browsers of other organisms. The only drawback of the MGI browser is the absence of a track for ncORFs (as of 25 February 2026), which does not permit convenient checking for ncORFs potentially co-affected in mutagenesis studies. The *M. truncatula* has such a track, with information on chimeric and non-chimeric ncORF MS-validated in our study (Çakır et al., 2025b).

The most comprehensive analysis of loss-of-function phenotypes in a plant organism was conducted in *A. thaliana* (Lloyd and Meinke, 2012). There is little to add to the deep coverage of that study, which is much more versatile compared to the basic analysis we present in this manuscript. The current genome browser of *A. thaliana* (Araport11, Cheng et al., 2017) contains a field that summarizes loss-of-function and other important data for each affected locus (Curator_summary). Unfortunately, that field contains no references to original studies that established the link between a locus and its phenotype. Likewise, no details are provided as to the genes potentially contributing to the phenotype. Thus, a user does not learn if the phenotype is neutral, genetically conditional (more than one gene involved), or caused by a single gene. Regarding the references, the *M. truncatula* genome browser is one step ahead of the *A. thaliana* browser as it provides relevant references to at least some genes targeted by loss-of-function approaches (Figure 2). Like the genome browser of *M. truncatula*, the *A. thaliana* browser has a separate track for ncORFs of the following types: uORFs, dORFs, and lncRNA-ORFs, according to the system of Mudge et al. (2022). In contrast to the *M. truncatula* browser, the ncORF track does not include intORFs (as of 25 February 2026), presumably because they are difficult to discriminate from translated refORFs by ribosome profiling (Ribo-Seq). The ncORF data in the *A. thaliana* browser were generated by Wu et al. (2024a).

Gephebase is an excellent resource for evolutionary studies (Courtier-Orgogozo et al., 2020). It focuses on genotype-phenotype relationships in natural variation observed in animals, yeast, and plants. Unfortunately, the database excludes lab mutants. Other databases are dedicated to mutations associated with human diseases or are much smaller in their scope compared with the resources mentioned above.

Several platforms were developed for *in silico* mutant screening in plants. They include, but are not limited to, the database of *M. truncatula* mutants (Tadege et al., 2008; Lee et al., 2018; Kaur et al., 2021), TOMATOMA for tomato (Saito et al., 2011; Shikata et al., 2016), and MiRiQ for rice (Kubo et al., 2024). However, these platforms cannot substitute genome browsers with integrated loss-of-function data, such as MGI and Araport11, because most of the mutants listed in those databases have not been characterized yet.

These considerations position our resource as smaller in scale than the *A. thaliana* database, but still essential for plant biologists, as it represents the most comprehensive loss-of-function

766 dataset available after *A. thaliana*. Importantly, the integration of ncORF information with
767 functional data gives our dataset a unique value among comparable resources. Specifically,
768 among all standard genome browsers, only the browsers of *A. thaliana* and *M. truncatula* have
769 the ncORF tracks (Wu et al., 2024a; Çakır et al., 2025b). In contrast to the *Arabidopsis*
770 browser, our ncORF track in the *Medicago* browser includes MS-validated ncORFs that are
771 embedded in refORFs or overlap with refORFs (ncORF designated as "intORFs", "uoORFs",
772 and "doORFs" in the system of Mudge et al., 2022), along with all other major ncORF types.
773 Lastly, the *Medicago* ncORF track is currently the only one (globally) that includes chimeric
774 ncORFs (ncORFs that are translated to chimeric peptides via programmed ribosomal
775 frameshifting, PRF).

776 **About the content of Supporting Discussion**

777 Due to space limitations, many important considerations about loss-of-function data, as well
778 as about ambiguities associated with ncORFs and *Tnt1*, were placed into a separate file
779 named Supporting Discussion. The file references Figure 4, supporting Figures S6-S9,
780 supporting Tables S2 and S3, Data S19, and has its own list of literature sources. Figure 4
781 introduces seven types of ambiguity associated with translated ncORFs in loss-of-function
782 studies, with real examples.

783 **The take-home message**

784 Functional genomics relies upon comprehensive and accurate annotation of genes available
785 as a genome browser. At the same time, a truly comprehensive genome browser should
786 encompass the information on loss-of-function phenotypes. While the mouse, *A. thaliana*, and
787 *M. truncatula* genomes are among the few to link such information to specific loci, there is
788 currently no standard pipeline for systematic recording, storage, and displaying phenotypic
789 data for other organisms. Here, we propose developing a centralized system that should make
790 it obligatory for researchers to report such information at the publication stage. We also
791 showed the limitations of the most comprehensive genome browsers, which can be
792 summarized as the lack of information on genes co-affected in original studies (multiple
793 knockouts or knockdowns) and missing references to such studies. Besides, the main purpose
794 of this work was to show that two fundamental aspects of functional genome annotation have
795 been largely overlooked in standard genome browsers. These aspects are (1) translated and
796 conserved ncORFs and (2) trans effects of insertional mutagenesis on splicing. Although both
797 phenomena have been well documented, they remain beyond consideration in nearly all
798 functional analyses. We provided examples of how overlooking such information may result
799 in major misinterpretation of loss-of-function phenotypes.

800 Importantly, this ambiguity is not unique to *M. truncatula*. It represents a pervasive, often
801 ignored variable in plant functional genomics. Whether using RNAi, which inevitably targets
802 co-transcribed ncORFs, or gene editing via CRISPR/Cas9, which can disrupt overlapping
803 coding sequences or splicing motifs, the "one gene, one phenotype" assumption is frequently
804 violated. Our data highlight that a significant proportion of "verified" mutants likely requires re-
805 evaluation to rule out ncORF contributions. Thus, the framework presented here serves not
806 merely as a curatorial resource, but as a necessary quality-control step for distinguishing
807 causal refORF mutations from ncORF-mediated artifacts in genetic studies.

808 We also showed how ncORFs can help rectify doubtful gene models. Despite the relatively
809 modest scope of the genome browser of *M. truncatula*, the integration of loss-of-function and
810 ncORF data in one resource can make it even more useful. In addition to summarizing the
811 information on phenotypes of 673 genes in the present manuscript, we mapped 805 non-
812 chimeric and 156 chimeric MS-validated ncORFs to corresponding loci in the genome

813 browser. To the best of our knowledge, after the ncORF study conducted in *Arabidopsis* (Wu
814 et al., 2024a), this is the second study to merge loss-of-function data with ncORFs in any
815 organism. At the same time, it is the first attempt to reconsider published loss-of-function
816 phenotypes in the light of knowledge about translated ncORFs expressed from the same loci.
817 We expect that this resource will serve as the model for browsers of other organisms, which
818 will ultimately improve the quality of functional analyses for the maximal benefit of biomedical
819 research, sustainable agriculture, and biotechnology.

820 Our meta-analysis of loss-of-function studies produced unique data on family-wise proportions
821 of altered, conditional, and neutral phenotypes in different protein classes and families. This
822 type of resolution is so far unprecedented because other studies mostly reported frequencies
823 of essential genes only for heterogeneous classes that combine members of diverse families.
824 Our list of tripartite phenotype proportions can be used by legume biologists and breeders as
825 a cheat sheet to focus on candidate genes that are likely to give a scorable phenotype. We
826 explicitly state that such a resource may have validity only within the legume family or even
827 *M. truncatula* alone. Like any cheat sheet, it should be used constructively: gene families with
828 low proportions of altered phenotypes should not be deprived of due attention.

829 The manuscript also coins several ideas in the field of ncORFomics. To the best of our
830 knowledge, these ideas are novel. First, based on earlier reviews and our new data, we
831 developed, discussed, and exemplified the classification of ambiguities associated with
832 ncORFs in loss-of-function studies. Second, we explained why the analysis of taxonomic
833 ranges of ncProts' conservation should become a standard component of genome
834 annotations. Third, we advocated the potential biological relevance of ncProts whose
835 conservation signatures are limited to the species of origin. Fourth, we showed how certain
836 translated ncORFs can systematically remain undetected due to regions of 100% similarity to
837 refProts in the species of origin. Lastly, we provided examples in which very strong
838 conservation of ncORFs at the amino acid level may not be a reliable marker of their
839 translation. These considerations will help better understand the biological roles of ncORFs,
840 separately from refORFs, as an integral component of the organismal translatome.

841 **MATERIALS AND METHODS**

842 **Meta-analysis of loss-of-function studies**

843 Literature on loss-of-function studies in *M. truncatula* was collected manually using Google
844 Scholar, PubMed, and other search engines. Main texts and supplementary data were
845 examined for information on locus identifiers, short and long forms of locus names,
846 mutagenesis systems, loss-of-function phenotypes, and protein classes. The protein class of
847 each entry was reconfirmed with the *M. truncatula* genome browser (Pecrix et al., 2018) and
848 the UniProt reference protein database (UniProt Consortium, 2019). The correspondence
849 between the latest locus identifiers and identifiers from earlier versions of the genome release
850 was established in various ways, depending on the availability of the primary information in
851 original studies. For some entries, the current locus identifiers were deduced using BLASTN
852 (Altschul et al., 1990; Boratyn et al., 2013) analysis of primers and/or RNAi segments, BLASTP
853 (Altschul et al., 1990; Boratyn et al., 2013) analysis of protein sequences, and sometimes
854 required the conversion of images to text for the extraction of sequence information. All steps
855 were manually curated, free of artificial intelligence.

856 **Detection of ncORFs**

857 Conserved and MS-validated ncORFs were detected using MosaicProt (Çakır et al., 2025c).
858 In short, each annotated transcript was scanned for the presence of stop-free regions longer
859 than 59 nt in all forward reading frames. Such regions were *in silico* translated, supplied with

unique informative identifiers, and saved in FASTA format. RefProts were separated from ncProts by matching each sequence to the currently annotated list of refProts.

For conservation analysis, all ncProts were compared with UniProt v. 2020_02 entries (UniProt Consortium, 2019) using DIAMOND v. 0.9.14 (Buchfink et al., 2021). NcProts with at least one match, including matches to *M. truncatula* sequences only, were categorized as conserved if the e-value of the top match was not more than 0.001. We explicitly declare that our definition of conserved ncProts is conditional in this study. It does not imply multi-species conservation, although some ncProts are conserved across hundreds of species.

For MS-validation of ncProts, they were compared to MS peptides from three publicly deposited proteomic datasets (PXD002692, Marx et al., 2016; PXD013606, Shin et al., 2021; and PXD022278, Castañeda et al., 2021), 16 samples in total, using SearchGUI v. 4.0.41 (Barsnes & Vaudel, 2018) and its partner tool PeptideShaker v. 2.0.33 (Vaudel et al., 2015). Each sample was searched independently using the target-decoy approach with FDR below 1%. MS peptides matching refProts and/or common contaminants were filtered out. The detection of NCR169 serves as a positive control for our procedure. NCR169 is a well-characterized peptide crucial for SNF (Starker et al., 2006; Domonkos et al., 2013; Farkas et al., 2014; Horváth et al., 2015, 2023; Güngör et al., 2023; Li et al., 2025).

MS-validated non-chimeric and chimeric ncProts (805 and 156, respectively) were mapped to the current genome and are available in the genome browser as two separate tracks (Çakır et al., 2025b). For chimeric MS peptides, not only 145 primary sources but also 101 potential alternative sources were mapped. The primary sources are loci in which chimeric MS peptides were identified in our analysis. Alternative sources of the same peptides were found using the ChiMSource software (Çakır et al., 2025a). Which of the loci are the true sources of chimeric MS-validated peptides was discussed in Çakır et al. (2025b). For 805 non-chimeric MS-validated peptides, we found 833 potential sources, which are mapped in the genome browser.

Locating mutations and insertions relative to refORFs and ncORFs of affected genes

To determine whether refORFs and selected ncORFs were affected in the original studies, we analyzed images and sequence data presented by the authors. In some cases, the exact positions of *Tnt1* insertions were deduced from FST sequences downloaded from the *Tnt1* mutant database of *M. truncatula* (Tadege et al., 2008; Lee et al., 2018; Kaur et al., 2021). Exon-intron maps of gene models were prepared by aligning cDNA to genomic DNA in Geneious® (MUSCLE alignment), with the manual correction of alignment artifacts at the splicing sites.

AUTHOR CONTRIBUTIONS

Umut Çakır: Conceptualization, Methodology, Software, Validation, Formal analysis, Resources, Data Curation, Investigation, Writing - Original Draft, Writing - Review & Editing, Visualization, Funding acquisition. **Noujoud Gabed**: Conceptualization, Validation, Formal analysis, Resources, Investigation, Writing - Review & Editing. **Selen Kaya**: Investigation, Data Curation. **Vagner A. Benedito**: Conceptualization, Writing - Review & Editing. **Marie Brunet**: Methodology. **Xavier Roucou**: Methodology, Writing - Review & Editing. **Igor S. Kryvoruchko**: Conceptualization, Methodology, Validation, Formal analysis, Resources, Data Curation, Investigation, Software, Writing - Original Draft, Writing - Review & Editing, Visualization, Supervision, Project administration, Funding acquisition.

ACKNOWLEDGMENTS

This work was supported by the Scientific and Technological Research Council of Turkey (TÜBİTAK) grants and Boğaziçi University standard research grant (BAP-P) to UÇ and IK (TÜBİTAK 1001 120Z514, TÜBİTAK 1002 120Z247, and BAP-P 18841), a Canada Research Chair in functional proteomics and discovery of novel proteins to XR, and a Canadian Institutes of Health Research Project Grant PJT-175322 to XR and MB. MB was supported by a Fonds de Recherche du Québec en Santé (FRQS) Junior 1 award (307936). Computational analysis was conducted using the server of the Turkish National e-Science e-Infrastructure (TRUBA) center. We sincerely thank Dr. Yun Kang for her critical feedback on the manuscript. The completion of this study was made possible by support for IK from the United Arab Emirates University and for UÇ from the IMPRS-Genome Science PhD program.

DATA AVAILABILITY STATEMENT

The annotated models of 123 genes listed in Data S12 are available in Geneious® format in the online version of this article (Data S18). They include cDNA and gDNA sequences, refORFs and ncORFs mapped to cDNA sequences, and alignments of cDNA and gDNA, which outline exon-intron structures of the models. The entire set of files for the conservation evidence and MS-validation of ncProts will be published later in a paper describing all conserved and translated ncProts detected *in M. truncatula*. On 8 January 2025, our data were integrated into the *M. truncatula* genome portal MtrunA17r5.0-ANR as a separate track. In total, 156 MS-validated chimeric peptides were mapped to 246 primary and alternative source loci in the genome browser (non-repeat-element alternative sources only). In addition, 805 MS-validated non-chimeric ncProts were mapped to 833 genomic loci. The complete set of data related to this article is available from the digital repository, Zenodo.

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990) Basic local alignment search tool. Journal of molecular biology, 215, 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.

Álvarez-Urdiola, R., & Riechmann, J. L. (2025) The 'non-conventional' peptidome: A new layer in plant regulatory mechanisms. Plant communications, 101437. Advance online publication. https://doi.org/10.1016/j.xplc.2025.101437.

Ardern, Z. (2023) Alternative Reading Frames are an Underappreciated Source of Protein Sequence Novelty. Journal of molecular evolution, 91, 570–580. https://doi.org/10.1007/s00239-023-10122-3.

Baldarelli, R. M., Smith, C. L., Ringwald, M., Richardson, J. E., Bult, C. J., & Mouse Genome Informatics Group (2024) Mouse Genome Informatics: an integrated knowledgebase system for the laboratory mouse. Genetics, 227, iyae031. https://doi.org/10.1093/genetics/iyae031.

Bandyopadhyay, K., Verdier, J., & Kang, Y. (2019) The model legume Medicago truncatula: past, present, and future. In *Plant biotechnology: progress in genomic era* (Khurana, S., & Gaur, R., eds). Singapore: Springer, pp. 109-130. https://doi.org/10.1007/978-981-13-8499-8_5.

Barker, D. G., Bianchi, S., Blondon, F., Dattée, Y., Duc, G., Essad, S. *et al*. (1990) Medicago truncatula, a model plant for studying the molecular genetics of theRhizobium-legume symbiosis. Plant Molecular Biology Reporter, 8, 40–49. https://doi.org/10.1007/bf02668879.

Barsnes, H., & Vaudel, M. (2018) SearchGUI: A Highly Adaptable Common Interface for Proteomics Search and de Novo Engines. Journal of proteome research, 17, 2552–2555. https://doi.org/10.1021/acs.jproteome.8b00175.

Baulcombe, D. (2004) RNA silencing in plants. Nature, 431, 356–363. https://doi.org/10.1038/nature02874.

Boratyn, G. M., Camacho, C., Cooper, P. S., Coulouris, G., Fong, A., Ma, N. *et al*. (2013) BLAST: a more efficient report with usability improvements. Nucleic acids research, 41, W29–W33. https://doi.org/10.1093/nar/gkt282.

Boschiero, C., Dai, X., Lundquist, P. K., Roy, S., Christian de Bang, T., Zhang, S. *et al*. (2020) MtSSPdb: The Medicago truncatula Small Secreted Peptide Database. Plant physiology, 183, 399–413. https://doi.org/10.1104/pp.19.01088.

Buchfink, B., Reuter, K., & Drost, H. G. (2021) Sensitive protein alignments at tree-of-life scale using DIAMOND. Nature methods, 18, 366–368. https://doi.org/10.1038/s41592-021-01101-x.

Çakır, U., Gabed, N., Yurtseven, A., & Kryvoruchko, I. (2025a) ChiMSource improves the accuracy of studies on novel amino acid sequences by predicting alternative sources of mass spectrometry-derived peptides. Computational and Structural Biotechnology Journal, 27, 3704–3709. https://doi.org/10.1016/j.csbj.2025.08.023.

Çakır, U., Gabed, N., Köroğlu, Y. E., Kaya, S., Sinharoy, S., Benedito, V. A. *et al*. (2025b) Discovery of diverse chimeric peptides in a eukaryotic proteome sets the stage for experimental validation of the mosaic translation hypothesis. Computational and structural biotechnology journal, 27, 4048–4064. https://doi.org/10.1016/j.csbj.2025.09.019.

Çakır, U., Gabed, N., Yurtseven, A., & Kryvoruchko, I. (2025c) A universal pipeline MosaicProt enables large-scale modeling and detection of chimeric protein sequences for studies on programmed ribosomal frameshifting. Computational and Structural Biotechnology Journal, 27, 5105–5115. https://doi.org/10.1016/j.csbj.2025.11.023.

Cañas, L. A. & Beltrán, J. P. (2018) Functional Genomics in Medicago truncatula 1st edn. New York: Humana. https://doi.org/10.1007/978-1-4939-8633-0.

Cardon, T., Fournier, I., & Salzet, M. (2021) Shedding Light on the Ghost Proteome. Trends in biochemical sciences, 46, 239–250. https://doi.org/10.1016/j.tibs.2020.10.003.

Cardon, T., Fournier, I., & Salzet, M. (2025) Chasing the Ghost Proteome in the Dark matter. Molecular & cellular proteomics, 101076. Advance online publication. https://doi.org/10.1016/j.mcpro.2025.101076.

Castañeda, V., González, E. M., & Wienkoop, S. (2021) Phloem Sap Proteins Are Part of a Core Stress Responsive Proteome Involved in Drought Stress Adjustment. Frontiers in plant science, 12, 625224. https://doi.org/10.3389/fpls.2021.625224.

Čermák, T., Curtin, S. J., Gil-Humanes, J., Čegan, R., Kono, T. J. Y., Konečná, E. *et al*. (2017) A Multipurpose Toolkit to Enable Advanced Genome Engineering in Plants. The Plant cell, 29, 1196–1217. https://doi.org/10.1105/tpc.16.00922.

Chang, Y., Lei, T., Zhou, F., Jiang, J., Huang, Y., Zhu, Z. *et al*. (2025) Evolutionary remodeling of non-canonical ORF translation in mammals. eLife, 14:RP109128. https://doi.org/10.7554/elife.109128.1.

Charpentier, M., Sun, J., Vaz Martins, T., Radhakrishnan, G. V., Findlay, K., Soumpourou, E. *et al*. (2016) Nuclear-localized cyclic nucleotide-gated channels mediate symbiotic calcium oscillations. Science, 352, 1102–1105. https://doi.org/10.1126/science.aae0109.

Chaudhary, D., Jeena, A. S., Rohit, Gaur, S., Raj, R., Mishra, S. *et al*. (2024). Advances in RNA Interference for Plant Functional Genomics: Unveiling Traits, Mechanisms, and Future Directions. Applied biochemistry and biotechnology, 196, 5681–5710. https://doi.org/10.1007/s12010-023-04850-x.

Cheng, C. Y., Krishnakumar, V., Chan, A. P., Thibaud-Nissen, F., Schobel, S., & Town, C. D. (2017) Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. The Plant journal, 89, 789–804. https://doi.org/10.1111/tpj.13415.

Combier, J. P., Frugier, F., de Billy, F., Boualem, A., El-Yahyaoui, F., Moreau, S. *et al*. (2006) MtHAP2-1 is a key transcriptional regulator of symbiotic nodule development regulated by microRNA169 in Medicago truncatula. Genes & development, 20, 3084–3088. https://doi.org/10.1101/gad.402806.

Combier, J. P., de Billy, F., Gamas, P., Niebel, A., & Rivas, S. (2008) Trans-regulation of the expression of the transcription factor MtHAP2-1 by a uORF controls root nodule development. Genes & development, 22, 1549–1559. https://doi.org/10.1101/gad.461808.

Cook, N. M., Gobbato, G., Jacott, C. N., Marchal, C., Hsieh, C. Y., Lam, A. H. C. *et al*. (2025) Autoactive CNGC15 enhances root endosymbiosis in legume and wheat. Nature, 638, 752–759. https://doi.org/10.1038/s41586-024-08424-7.

Courtier-Orgogozo, V., Arnoult, L., Prigent, S. R., Wiltgen, S., & Martin, A. (2020) Gephebase, a database of genotype-phenotype relationships for natural and domesticated variation in Eukaryotes. Nucleic acids research, 48, D696–D703. https://doi.org/10.1093/nar/gkz796.

Cui, Y., Barampuram, S., Stacey, M. G., Hancock, C. N., Findley, S., Mathieu, M. *et al*. (2013) Tnt1 retrotransposon mutagenesis: a tool for soybean functional genomics. Plant physiology, 161, 36–47. https://doi.org/10.1104/pp.112.205369.

Curtin, S. J., Tiffin, P., Guhlin, J., Trujillo, D. I., Burghart, L. T., Atkins, P. *et al*. (2017) Validating Genome-Wide Association Candidates Controlling Quantitative Variation in Nodulation. Plant physiology, 173, 921–931. https://doi.org/10.1104/pp.16.01923.

de Bang, T. C., Lundquist, P. K., Dai, X., Boschiero, C., Zhuang, Z., Pant, P. *et al*. (2017) Genome-Wide Identification of Medicago Peptides Involved in Macronutrient Responses and Nodulation. Plant physiology, 175, 1669–1689. https://doi.org/10.1104/pp.17.01096.

Deutsch, E. W., Bandeira, N., Perez-Riverol, Y., Sharma, V., Carver, J. J., Mendoza, L. *et al*. (2023) The ProteomeXchange consortium at 10 years: 2023 update. Nucleic acids research, 51, D1539–D1548. https://doi.org/10.1093/nar/gkac1040.

Domonkos, A., Horvath, B., Marsh, J. F., Halasz, G., Ayaydin, F., Oldroyd, G. E. *et al*. (2013) The identification of novel loci required for appropriate nodule development in Medicago truncatula. BMC plant biology, 13, 157. https://doi.org/10.1186/1471-2229-13-157.

Dong, W., Zhu, Y., Chang, H., Wang, C., Yang, J., Shi, J. *et al*. (2021) An SHR-SCR module specifies legume cortical cell fate to enable nodulation. Nature, 589, 586–590. https://doi.org/10.1038/s41586-020-3016-z.

Doudna, J. A., & Charpentier, E. (2014) Genome editing. The new frontier of genome engineering with CRISPR-Cas9. Science, 346, 1258096. https://doi.org/10.1126/science.1258096.

Erickson, B. J., Staples, N. C., Hess, N., Staples, M. A., Weissert, C., Finkelstein, R. R. *et al*. (2020) PRPs localized to the middle lamellae are required for cortical tissue integrity in Medicago truncatula roots. Plant molecular biology, 102, 571–588. https://doi.org/10.1007/s11103-019-00960-5.

Farkas, A., Maróti, G., Durgő, H., Györgypál, Z., Lima, R. M., Medzihradszky, K. F. *et al*. (2014) Medicago truncatula symbiotic peptide NCR247 contributes to bacteroid differentiation through multiple mechanisms. Proceedings of the National Academy of Sciences of the United States of America, 111, 5183–5188. https://doi.org/10.1073/pnas.1404169111.

Floss, D. S., Gomez, S. K., Park, H. J., MacLean, A. M., Müller, L. M., Bhattarai, K. K. *et al*. (2017) A Transcriptional Program for Arbuscule Degeneration during AM Symbiosis Is Regulated by MYB1. Current biology, 27, 1206–1212. https://doi.org/10.1016/j.cub.2017.03.003.

Frendo, P., Jiménez, M. J., Mathieu, C., Duret, L., Gallesi, D., Van de Sype, G. *et al*. (2001) A Medicago truncatula homoglutathione synthetase is derived from glutathione synthetase by gene duplication. Plant physiology, 126, 1706–1715. https://doi.org/10.1104/pp.126.4.1706.

Frendo, P., Harrison, J., Norman, C., Hernández Jiménez, M. J., Van de Sype, G., Gilabert, A. *et al*. (2005) Glutathione and homoglutathione play a critical role in the nodulation process of Medicago truncatula. Molecular plant-microbe interactions, 18, 254–259. https://doi.org/10.1094/MPMI-18-0254.

Gavrin, A., Kulikova, O., Bisseling, T., & Fedorova, E. E. (2017) Interface Symbiotic Membrane Formation in Root Nodules of Medicago truncatula: the Role of Synaptotagmins MtSyt1, MtSyt2 and MtSyt3. Frontiers in plant science, 8, 201. https://doi.org/10.3389/fpls.2017.00201.

Grandbastien, M. A., Spielmann, A., & Caboche, M. (1989) Tnt1, a mobile retroviral-like transposable element of tobacco isolated by plant cell genetics. Nature, 337, 376–380. https://doi.org/10.1038/337376a0.

Groth, P., Pavlova, N., Kalev, I., Tonov, S., Georgiev, G., Pohlenz, H. D. *et al*. (2007) PhenomicDB: a new cross-species genotype/phenotype resource. Nucleic acids research, 35, D696–D699. https://doi.org/10.1093/nar/gkl662.

Güngör, B., Biró, J. B., Domonkos, Á., Horváth, B., & Kaló, P. (2023) Targeted mutagenesis of Medicago truncatula Nodule-specific Cysteine-Rich (NCR) genes using the Agrobacterium rhizogenes-mediated CRISPR/Cas9 system. Scientific reports, 13, 20676. https://doi.org/10.1038/s41598-023-47608-5.

Guo, C., Ma, X., Gao, F., & Guo, Y. (2023) Off-target effects in CRISPR/Cas9 gene editing. Frontiers in bioengineering and biotechnology, 11, 1143157. https://doi.org/10.3389/fbioe.2023.1143157.

Guo, D., Li, J., Liu, P., Wang, Y., Cao, N., Fang, X. *et al*. (2024) The jasmonate pathway promotes nodule symbiosis and suppresses host plant defense in Medicago truncatula. Molecular plant, 17, 1183–1203. https://doi.org/10.1016/j.molp.2024.06.004.

Hawkins, J. P., & Oresnik, I. J. (2022) The Rhizobium-Legume Symbiosis: Co-opting Successful Stress Management. Frontiers in plant science, 12, 796045. https://doi.org/10.3389/fpls.2021.796045.

Horváth, B., Domonkos, Á., Kereszt, A., Szűcs, A., Ábrahám, E., Ayaydin, F. *et al*. (2015) Loss of the nodule-specific cysteine rich peptide, NCR169, abolishes symbiotic nitrogen fixation in the Medicago truncatula dnf7 mutant. Proceedings of the National Academy of Sciences of the United States of America, 112, 15232–15237. https://doi.org/10.1073/pnas.1500777112.

Horváth, B., Güngör, B., Tóth, M., Domonkos, Á., Ayaydin, F., Saifi, F. *et al*. (2023) The Medicago truncatula nodule-specific cysteine-rich peptides, NCR343 and NCR-new35 are required for the maintenance of rhizobia in nitrogen-fixing nodules. The New phytologist, 239, 1974–1988. https://doi.org/10.1111/nph.19097.

Huault, E., Laffont, C., Wen, J., Mysore, K. S., Ratet, P., Duc, G. *et al*. (2014) Local and systemic regulation of plant root system architecture and symbiotic nodulation by a receptor-like kinase. PLoS genetics, 10, e1004891. https://doi.org/10.1371/journal.pgen.1004891.

Huertas, R., Torres-Jerez, I., Curtin, S. J., Scheible, W., & Udvardi, M. (2023) Medicago truncatula PHO2 genes have distinct roles in phosphorus homeostasis and symbiotic nitrogen fixation. Frontiers in plant science, 14, 1211107. https://doi.org/10.3389/fpls.2023.1211107.

Huisman, R., & Geurts, R. (2019) A Roadmap toward Engineered Nitrogen-Fixing Nodule Symbiosis. Plant communications, 1, 100019. https://doi.org/10.1016/j.xplc.2019.100019.

Isayenkov, S., Mrosk, C., Stenzel, I., Strack, D., & Hause, B. (2005) Suppression of allene oxide cyclase in hairy roots of Medicago truncatula reduces jasmonate levels and the degree of mycorrhization with Glomus intraradices. Plant physiology, 139, 1401–1410. https://doi.org/10.1104/pp.105.069054.

Kahraman, A., Avramov, A., Nashev, L. G., Popov, D., Ternes, R., Pohlenz, H. D. *et al*. (2005) PhenomicDB: a multi-species genotype/phenotype database for comparative phenomics. Bioinformatics, 21, 418–420. https://doi.org/10.1093/bioinformatics/bti010.

Kakiyama, S., Tabara, M., Nishibori, Y., Moriyama, H., & Fukuhara, T. (2019) Long DCL4-substrate dsRNAs efficiently induce RNA interference in plant cells. Scientific reports, 9, 6920. https://doi.org/10.1038/s41598-019-43443-9.

Kang, Y., Li, M., Sinharoy, S., & Verdier, J. (2016) A Snapshot of Functional Genetic Studies in Medicago truncatula. Frontiers in plant science, 7, 1175. https://doi.org/10.3389/fpls.2016.01175.

Kaur, P., Lui, C., Dudchenko, O., Nandety, R. S., Hurgobin, B., Pham, M. *et al*. (2021) Delineating the Tnt1 Insertion Landscape of the Model Legume Medicago truncatula cv. R108 at the Hi-C Resolution Using a Chromosome-Length Genome Assembly. International journal of molecular sciences, 22, 4326. https://doi.org/10.3390/ijms22094326.

Kim, G. B., & Nam, Y. W. (2013) A novel GRAS protein gene MtSymSCL1 plays a role in regulating nodule number in Medicago truncatula. Plant Growth Regulation, 71, 77–92. https://doi.org/10.1007/s10725-013-9814-7.

Kereszt, A., Mergaert, P., Montiel, J., Endre, G., & Kondorosi, É. (2018) Impact of Plant Peptides on Symbiotic Nodule Development and Functioning. Frontiers in plant science, 9, 1026. https://doi.org/10.3389/fpls.2018.01026.

Kiirika, L. M., Bergmann, H. F., Schikowsky, C., Wimmer, D., Korte, J., Schmitz, U. *et al*. (2012) Silencing of the Rac1 GTPase MtROP9 in Medicago truncatula stimulates early mycorrhizal and oomycete root colonizations but negatively affects rhizobial infection. Plant physiology, 159, 501–516. https://doi.org/10.1104/pp.112.193706.

Kozak, M. (1999) Initiation of translation in prokaryotes and eukaryotes. Gene, 234, 187–208. https://doi.org/10.1016/s0378-1119(99)00210-3.

Kubo, T., Yamagata, Y., Matsusaka, H., Toyoda, A., Sato, Y., & Kumamaru, T. (2024) MiRiQ Database: A Platform for In Silico Rice Mutant Screening. Plant & cell physiology, 65, 169–174. https://doi.org/10.1093/pcp/pcad134.

Laffont, C., Blanchet, S., Lapierre, C., Brocard, L., Ratet, P., Crespi, M. *et al*. (2010) The compact root architecture1 gene regulates lignification, flavonoid production, and polar auxin transport in Medicago truncatula. Plant physiology, 153, 1597–1607. https://doi.org/10.1104/pp.110.156620.

Lee, H. K., Mysore, K. S., & Wen, J. (2018) Tnt1 Insertional Mutagenesis in Medicago truncatula. Methods in molecular biology, 1822, 107–114. https://doi.org/10.1007/978-1-4939-8633-0_7.

Li, X., Song, Y., Century, K., Straight, S., Ronald, P., Dong, X. *et al*. (2001) A fast neutron deletion mutagenesis-based reverse genetics system for plants. The Plant journal, 27, 235–242. https://doi.org/10.1046/j.1365-313x.2001.01084.x.

Li, H., Schilderink, S., Cao, Q., Kulikova, O., & Bisseling, T. (2021a) Plant-specific histone deacetylases are essential for early and late stages of Medicago nodule development. Plant physiology, 186, 1591–1605. https://doi.org/10.1093/plphys/kiab140.

Li, W., Ma, Q., Yin, P., Wen, J., Pei, Y., Niu, L. *et al*. (2021b) The GA 20-Oxidase Encoding Gene MSD1 Controls the Main Stem Elongation in Medicago truncatula. Frontiers in plant science, 12, 709625. https://doi.org/10.3389/fpls.2021.709625.

Li, M., Xu, L., Zhang, L., Li, X., Cao, C., Chen, L. *et al*. (2022) Overexpression of Mtr-miR319a Contributes to Leaf Curl and Salt Stress Adaptation in Arabidopsis thaliana and

Medicago truncatula. International journal of molecular sciences, 24, 429. https://doi.org/10.3390/ijms24010429.

Li, F., Kumar, A., Xu, P., & Murray, J. D. (2025) Symbiotic nitrogen fixation suppresses root nitrate uptake in Medicago truncatula under nitrate limitation. The New phytologist, 246, 1905–1911. https://doi.org/10.1111/nph.70115.

Limpens, E., Franken, C., Smit, P., Willemse, J., Bisseling, T., & Geurts, R. (2003) LysM domain receptor kinases regulating rhizobial Nod factor-induced infection. Science, 302, 630–633. https://doi.org/10.1126/science.1090074.

Limpens, E., Ramos, J., Franken, C., Raz, V., Compaan, B., Franssen, H. *et al*. (2004) RNA interference in Agrobacterium rhizogenes-transformed roots of Arabidopsis and Medicago truncatula. Journal of experimental botany, 55, 983–992. https://doi.org/10.1093/jxb/erh122.

Liu, C., Jun, J. H., & Dixon, R. A. (2014) MYB5 and MYB14 Play Pivotal Roles in Seed Coat Polymer Biosynthesis in Medicago truncatula. Plant physiology, 165, 1424–1439. https://doi.org/10.1104/pp.114.241877.

Liu, J., & Bisseling, T. (2020) Evolution of NIN and NIN-like Genes in Relation to Nodule Symbiosis. Genes, 11, 777. https://doi.org/10.3390/genes11070777.

Lloyd, J., & Meinke, D. (2012) A comprehensive dataset of genes with a loss-of-function mutant phenotype in Arabidopsis. Plant physiology, 158, 1115–1129. https://doi.org/10.1104/pp.111.192393.

Lluch-Senar, M., Mancuso, F. M., Climente-González, H., Peña-Paz, M. I., Sabido, E., & Serrano, L. (2016) Rescuing discarded spectra: Full comprehensive analysis of a minimal proteome. Proteomics, 16, 554–563. https://doi.org/10.1002/pmic.201500187.

Mazier, M., Botton, E., Flamain, F., Bouchet, J. P., Courtial, B., Chupeau, M. C. *et al*. (2007) Successful gene tagging in lettuce using the Tnt1 retrotransposon from tobacco. Plant physiology, 144, 18–31. https://doi.org/10.1104/pp.106.090365.

Marx, H., Minogue, C. E., Jayaraman, D., Richards, A. L., Kwiecien, N. W., Siahpirani, A. F. *et al*. (2016) A proteomic atlas of the legume Medicago truncatula and its nitrogen-fixing endosymbiont Sinorhizobium meliloti. Nature biotechnology, 34, 1198–1205. https://doi.org/10.1038/nbt.3681.

Mergaert, P., Nikovics, K., Kelemen, Z., Maunoury, N., Vaubert, D., Kondorosi, A. *et al*. (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 physiology, 132, 161–173. https://doi.org/10.1104/pp.102.018192.

Mouilleron, H., Delcourt, V., & Roucou, X. (2016) Death of a dogma: eukaryotic mRNAs can code for more than one protein. Nucleic acids research, 44, 14–23. https://doi.org/10.1093/nar/gkv1218.

Mudge, J. M., Ruiz-Orera, J., Prensner, J. R., Brunet, M. A., Calvet, F., Jungreis, I. *et al*. (2022) Standardized annotation of translated open reading frames. Nature biotechnology, 40, 994–999. https://doi.org/10.1038/s41587-022-01369-0.

Nandety, R. S., Wen, J., & Mysore, K. S. (2022) Medicago truncatula resources to study legume biology and symbiotic nitrogen fixation. Fundamental research, 3, 219–224. https://doi.org/10.1016/j.fmre.2022.06.018.

Nilsson, O. & Olsson, O. (1997) Getting to the root: The role of the Agrobacterium rhizogenes rol genes in the formation of hairy roots. Physiologia Plantarum, 100, 463-473. https://doi.org/10.1111/j.1399-3054.1997.tb03050.x

Ning, K., Fermin, D., & Nesvizhskii, A. I. (2010) Computational analysis of unassigned high-quality MS/MS spectra in proteomic data sets. Proteomics, 10, 2712–2718. https://doi.org/10.1002/pmic.200900473.

Orr, M. W., Mao, Y., Storz, G., & Qian, S. B. (2020) Alternative ORFs and small ORFs: shedding light on the dark proteome. Nucleic acids research, 48, 1029–1042. https://doi.org/10.1093/nar/gkz734.

Pathan, M., Samuel, M., Keerthikumar, S., & Mathivanan, S. (2017) Unassigned MS/MS Spectra: Who Am I?. Methods in molecular biology, 1549, 67–74. https://doi.org/10.1007/978-1-4939-6740-7_6.

Pavesi, A. (2025) Systematic Detection of Alternative Open Reading Frames (altORFs) in Cancer Driver Genes. Journal of molecular evolution, 10.1007/s00239-025-10284-2. Advance online publication. https://doi.org/10.1007/s00239-025-10284-2.

Pecrix, Y., Staton, S. E., Sallet, E., Lelandais-Brière, C., Moreau, S., Carrère, S. *et al*. (2018) Whole-genome landscape of Medicago truncatula symbiotic genes. Nature plants, 4, 1017–1025. https://doi.org/10.1038/s41477-018-0286-7.

Rey, T., Bonhomme, M., Chatterjee, A., Gavrin, A., Toulotte, J., Yang, W. *et al*. (2017) The Medicago truncatula GRAS protein RAD1 supports arbuscular mycorrhiza symbiosis and Phytophthora palmivora susceptibility. Journal of experimental botany, 68, 5871–5881. https://doi.org/10.1093/jxb/erx398.

Ribeiro, B., Erffelinck, M. L., Lacchini, E., Ceulemans, E., Colinas, M., Williams, C. *et al*. (2022) Interference between ER stress-related bZIP-type and jasmonate-inducible bHLH-type transcription factors in the regulation of triterpene saponin biosynthesis in Medicago truncatula. Frontiers in plant science, 13, 903793. https://doi.org/10.3389/fpls.2022.903793.

Roy, S., Liu, W., Nandety, R. S., Crook, A., Mysore, K. S., Pislariu, C. I. *et al*. (2020) Celebrating 20 Years of Genetic Discoveries in Legume Nodulation and Symbiotic Nitrogen Fixation. The Plant cell, 32, 15–41. https://doi.org/10.1105/tpc.19.00279.

Sagan, M., Morandi, D., Tarenghi, E., & Duc, G. (1995) Selection of nodulation and mycorrhizal mutants in the model plant Medicago truncatula (Gaertn) after g-ray mutagenesis. Plant Science, 111, 63–71. https://doi.org/10.1016/0168-9452(95)04229-N.

Saito, T., Ariizumi, T., Okabe, Y., Asamizu, E., Hiwasa-Tanase, K., Fukuda, N. *et al*. (2011) TOMATOMA: a novel tomato mutant database distributing Micro-Tom mutant collections. Plant & cell physiology, 52, 283–296. https://doi.org/10.1093/pcp/pcr004.

Samac, D. A., Peñuela, S., Schnurr, J. A., Hunt, E. N., Foster-Hartnett, D., Vandenbosch, K. A. *et al*. (2011) Expression of coordinately regulated defence response

genes and analysis of their role in disease resistance in Medicago truncatula. Molecular plant pathology, 12, 786–798. https://doi.org/10.1111/j.1364-3703.2011.00712.x.

Scholte, M., d'Erfurth, I., Rippa, S., Mondy, S., Cosson, V., Durand, P. *et al*. (2002) T-DNA tagging in the model legume Medicago truncatula allows efficient gene discovery. Molecular Breeding, 10, 203–215. https://doi.org/10.1023/A:1020564612093.

Sega, G. A. (1984) A review of the genetic effects of ethyl methanesulfonate. Mutation research, 134, 113–142. https://doi.org/10.1016/0165-1110(84)90007-1.

Sharp, P. A. & Zamore P. D. (2000) RNA Interference. Science, 287, 2431–2433. https://doi.org/10.1126/science.287.5462.2431.

Shikata, M., Hoshikawa, K., Ariizumi, T., Fukuda, N., Yamazaki, Y., & Ezura, H. (2016) TOMATOMA Update: Phenotypic and Metabolite Information in the Micro-Tom Mutant Resource. Plant & cell physiology, 57, e11. https://doi.org/10.1093/pcp/pcv194.

Shin, J., Marx, H., Richards, A., Vaneechoutte, D., Jayaraman, D., Maeda, J. *et al*. (2021) A network-based comparative framework to study conservation and divergence of proteomes in plant phylogenies. Nucleic acids research, 49, e3. https://doi.org/10.1093/nar/gkaa1041.

Sinharoy, S., Kang, Y., & Benedito, V. (2023) The Medicago truncatula Genome 1st edn. Cham: Springer. https://doi.org/10.1007/978-3-030-90757-0.

Soyano, T., Shimoda, Y., Kawaguchi, M., & Hayashi, M. (2019) A shared gene drives lateral root development and root nodule symbiosis pathways in Lotus. Science, 366, 1021–1023. https://doi.org/10.1126/science.aax2153.

Starker, C. G., Parra-Colmenares, A. L., Smith, L., Mitra, R. M., & Long, S. R. (2006) Nitrogen fixation mutants of Medicago truncatula fail to support plant and bacterial symbiotic gene expression. Plant physiology, 140, 671–680. https://doi.org/10.1104/pp.105.072132.

Tadege, M., Wen, J., He, J., Tu, H., Kwak, Y., Eschstruth, A. *et al*. (2008). Large-scale insertional mutagenesis using the *Tnt1* retrotransposon in the model legume *Medicago truncatula*. The Plant journal, 54, 335–347. https://doi.org/10.1111/j.1365-313X.2008.03418.x.

Tian, Y., Liu, W., Cai, J., Zhang, L. Y., Wong, K. B., Feddermann, N. *et al*. (2013) The nodulation factor hydrolase of Medicago truncatula: characterization of an enzyme specifically cleaving rhizobial nodulation signals. Plant physiology, 163, 1179–1190. https://doi.org/10.1104/pp.113.223966.

Udvardi, M. K., Czechowski, T., & Scheible, W. R. (2008) Eleven golden rules of quantitative RT-PCR. The Plant cell, 20, 1736–1737. https://doi.org/10.1105/tpc.108.061143.

UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. Nucleic acids research, 47, D506–D515. https://doi.org/10.1093/nar/gky1049.

Vanderperre, B., Lucier, J. F., Bissonnette, C., Motard, J., Tremblay, G., Vanderperre, S. *et al*. (2013) Direct detection of alternative open reading frames translation products in human significantly expands the proteome. PloS one, 8, e70698. https://doi.org/10.1371/journal.pone.0070698.

Vaudel, M., Burkhart, J. M., Zahedi, R. P., Oveland, E., Berven, F. S., Sickmann, A. *et al*. (2015) PeptideShaker enables reanalysis of MS-derived proteomics data sets. Nature biotechnology, 33, 22–24. https://doi.org/10.1038/nbt.3109.

Vives, C., Charlot, F., Mhiri, C., Contreras, B., Daniel, J., Epert, A. *et al*. (2016) Highly efficient gene tagging in the bryophyte Physcomitrella patens using the tobacco (Nicotiana tabacum) Tnt1 retrotransposon. The New phytologist, 212, 759–769. https://doi.org/10.1111/nph.14152.

Volkening, J. D., Bailey, D. J., Rose, C. M., Grimsrud, P. A., Howes-Podoll, M., Venkateshwaran, M. *et al*. (2012). A proteogenomic survey of the Medicago truncatula genome. Molecular & cellular proteomics, 11, 933–944. https://doi.org/10.1074/mcp.M112.019471.

Wang, T., Zhan, X., Bu, C. H., Lyon, S., Pratt, D., Hildebrand, S. *et al*. (2015) Real-time resolution of point mutations that cause phenovariance in mice. Proceedings of the National Academy of Sciences of the United States of America, 112, E440–E449. https://doi.org/10.1073/pnas.1423216112.

Wang, H., Lu, Z., Xu, Y., Kong, L., Shi, J., Liu, Y. *et al*. (2019) Genome-wide characterization of SPL family in Medicago truncatula reveals the novel roles of miR156/SPL module in spiky pod development. BMC genomics, 20, 552. https://doi.org/10.1186/s12864-019-5937-1.

Wang, M., Feng, H., Xu, P., Xie, Q., Gao, J., Wang, Y. *et al*. (2021) Phosphorylation of MtRopGEF2 by LYK3 mediates MtROP activity to regulate rhizobial infection in Medicago truncatula. Journal of integrative plant biology, 63, 1787–1800. https://doi.org/10.1111/jipb.13148.

Wang, X., Zhang, J., Liu, X., Kong, Y., & Han, L. (2023) The Roles of the PSEUDO-RESPONSE REGULATORs in Circadian Clock and Flowering Time in Medicago truncatula. International journal of molecular sciences, 24, 16834. https://doi.org/10.3390/ijms242316834.

Wang, D., Jin, R., Shi, X., Guo, H., Tan, X., Zhao, A. *et al*. (2025) A kinase mediator of rhizobial symbiosis and immunity in Medicago. Nature, 643, 768–775. https://doi.org/10.1038/s41586-025-09057-0.

Waterhouse, P. M., Graham, M. W., & Wang, M. B. (1998) Virus resistance and gene silencing in plants can be induced by simultaneous expression of sense and antisense RNA. Proceedings of the National Academy of Sciences of the United States of America, 95, 13959–13964. https://doi.org/10.1073/pnas.95.23.13959.

Wright, B. W., Yi, Z., Weissman, J. S., & Chen, J. (2022) The dark proteome: translation from noncanonical open reading frames. Trends in cell biology, 32, 243–258. https://doi.org/10.1016/j.tcb.2021.10.010.

Wu, H. L., Ai, Q., Teixeira, R. T., Nguyen, P. H. T., Song, G., Montes, C. *et al*. (2024a) Improved super-resolution ribosome profiling reveals prevalent translation of upstream ORFs and small ORFs in Arabidopsis. The Plant cell, 36, 510–539. https://doi.org/10.1093/plcell/koad290.

Wu, W., Zhuang, Y., Chen, D., Ruan, Y., Li, F., Jackson, K. *et al*. (2024b) Methylated chalcones are required for rhizobial nod gene induction in the Medicago truncatula rhizosphere. The New phytologist, 242, 2195–2206. https://doi.org/10.1111/nph.19701.

Xing, J., Wang, J., Cao, J., Li, K., Meng, X., Wen, J. *et al*. (2024) Identification of a Novel Gene MtbZIP60 as a Negative Regulator of Leaf Senescence through Transcriptome Analysis in Medicago truncatula. International journal of molecular sciences, 25, 10410. https://doi.org/10.3390/ijms251910410.

Xu, L., Xu, Y., Lv, H., Xu, Y., Wen, J., Li, M. *et al*. (2024). Transcriptomic Analysis Reveals the Mechanism of MtLOX24 in Response to Methyl Jasmonate Stress in Medicago truncatula. Agriculture, 14, 1076. https://doi.org/10.3390/agriculture14071076.

Yang, L., Sun, Q., Geng, B., Shi, J., Zhu, H., Sun, Y. *et al*. (2023) Jasmonate biosynthesis enzyme allene oxide cyclase 2 mediates cold tolerance and pathogen resistance. Plant physiology, 193, 1621–1634. https://doi.org/10.1093/plphys/kiad362.

Yin, P., Ma, Q., Wang, H., Feng, D., Wang, X., Pei, Y. *et al*. (2020) SMALL LEAF AND BUSHY1 controls organ size and lateral branching by modulating the stability of BIG SEEDS1 in Medicago truncatula. The New phytologist, 226, 1399–1412. https://doi.org/10.1111/nph.16449.

Zhang, K., McKinlay, C., Hocart, C. H., & Djordjevic, M. A. (2006) The Medicago truncatula small protein proteome and peptidome. Journal of proteome research, 5, 3355–3367. https://doi.org/10.1021/pr060336t.

Zhang, J., Subramanian, S., Stacey, G., & Yu, O. (2009) Flavones and flavonols play distinct critical roles during nodulation of Medicago truncatula by Sinorhizobium meliloti. The Plant journal, 57, 171–183. https://doi.org/10.1111/j.1365-313X.2008.03676.x.

Zhang, L. Y., Cai, J., Li, R. J., Liu, W., Wagner, C., Wong, K. B. *et al*. (2016a) A single amino acid substitution in a chitinase of the legume Medicago truncatula is sufficient to gain Nod-factor hydrolase activity. Open biology, 6, 160061. https://doi.org/10.1098/rsob.160061.

Zhang, Z., Hu, X., Zhang, Y., Miao, Z., Xie, C., Meng, X. *et al*. (2016b) Opposing Control by Transcription Factors MYB61 and MYB3 Increases Freezing Tolerance by Relieving C-Repeat Binding Factor Suppression. Plant physiology, 172, 1306–1323. https://doi.org/10.1104/pp.16.00051.

Zhang, L., Jiang, A., Thomson, G., Kerr-Phillips, M., Phan, C., Krueger, T. *et al*. (2019) Overexpression of Medicago MtCDFd1_1 Causes Delayed Flowering in Medicago via Repression of MtFTa1 but Not MtCO-Like Genes. Frontiers in plant science, 10, 1148. https://doi.org/10.3389/fpls.2019.01148.

Zhao, M., Tian, R., Sun, X., & Zhang, W. H. (2023) lncRNA MtCIR2 positively regulates plant-freezing tolerance by modulating CBF/DREB1 gene clusters. Plant, cell & environment, 46, 2450–2469. https://doi.org/10.1111/pce.14611.

Zhou, S., Yang, T., Mao, Y., Liu, Y., Guo, S., Wang, R. *et al*. (2021) The F-box protein MIO1/SLB1 regulates organ size and leaf movement in Medicago truncatula. Journal of experimental botany, 72, 2995–3011. https://doi.org/10.1093/jxb/erab033.

Table 1. Description of the core datasets

Dataset	Locus type	Phenotypes	Type of functional characterization	Number of loci†, LOF	Number of loci†, GOF only
Data S1	RefORFs	Non-SNF	LOF	508 (501)	-
Data S2	NcORFs	Non-SNF	LOF, GOF	11 (11)	16 (15)
Data S3	RefORFs	SNF	LOF	354 (351)	-
Data S4	NcORFs	SNF	LOF, GOF	30 (30)	18 (17)
Data S5‡	RefORFs and ncORFs	Non-SNF and SNF	LOF	222 (219)	-
Data S6§	RefORFs and ncORFs	Non-SNF and SNF	LOF	681 (673)	-

†Numbers in brackets correspond to unique loci for which the identifiers in v5.1.9 are known.
‡Data S5 lists only those genes that were tested for both non-SNF and SNF phenotypes.
§Data S6 lists all genes.
Abbreviations:
RefORFs – reference open reading frames
NcORFs – non-canonical open reading frames
LOF – loss-of-function phenotypes
GOF – gain-of-function phenotypes
SNF – symbiotic nitrogen fixation phenotypes

1350

1351 **Table 2**. Description of the supporting datasets

Dataset	Description
Data S7	References for LOF studies only (refORFs and ncORFs)
Data S8	References for GOF studies only (ncORFs only)
Data S9	The complete list of 97 biological processes examined in LOF studies
Data S10	The complete list of protein classes, corresponding phenotype counts, and statistical analysis of proportions
Data S11	The complete list of combined protein groups, corresponding phenotype counts, and statistical analysis of proportions
Data S12	A master table that summarizes various parameters of 148 ncORFs found in 123 genes targeted by functional analyses
Data S13	Visual demonstration of conservation at the amino acid level for eight selected ncORFs
Data S14	Evidence for purifying selection of the eight selected ncORFs
Data S15	24 nucleotide alignments that demonstrate evidence for purifying selection of the eight selected ncORFs
Data S16	Sequence analysis demonstrating the origin of highly conserved ncORFs in five selected genes
Data S17	Nucleotide alignments used in Data S16
Data S18	Gene models and mapped ncORFs of 123 loci targeted by functional analyses in *M. truncatula* (Data S12)
Data S19	NcProts grouped with refProts during the MS analysis

1352

1353 **Table 3**. Loci incorrectly named in the genome browser v. 5.1.9

Name†	Actual locus ID (publication)	Publication
MtCbf4	MtrunA17_Chr1g0203861	Zhang et al., 2016b
MtChOMT1‡	MtrunA17_Chr1g0200491	Wu et al., 2024b
MtGA20ox1	MtrunA17_Chr1g0204181	Li et al., 2021b
MtHDT3	MtrunA17_Chr7g0265761	Li et al., 2021a
MtMYC2	MtrunA17_Chr8g0366751	Guo et al., 2024
MtPRP1‡	MtrunA17_Chr4g0013391	Erickson et al., 2020
MtPRR7	MtrunA17_Chr1g0181811	Wang et al., 2023
MtRac1	MtrunA17_Chr6g0486041	Wang et al., 2021
MtROP9	MtrunA17_Chr5g0405791	Kiirika et al., 2012
MtSCR	MtrunA17_Chr7g0245601	Dong et al., 2021
MtSLB1	MtrunA17_Chr5g0447381	Yin et al., 2020; Zhou et al., 2021

†The table lists correct names and corresponding locus identifiers. These names were incorrectly assigned to other loci in the genome browser v. 5.1.9.
‡Names *MtChOMT1* and *MtPRP1* were misassigned to two genes each in the genome browser v. 5.1.9.
Table S1 provides an extended version of this list.

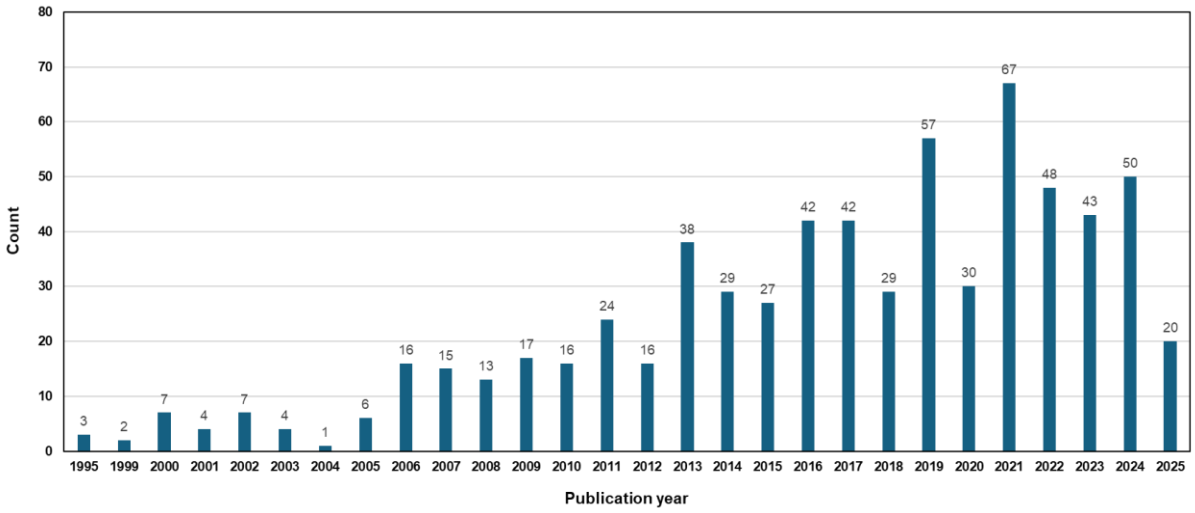

Figure 1. Timeline of loss-of-function studies in Medicago truncatula. The graph is based on the first publication of each genetic locus, 673 loci in total. 219 of these genes were studied in the context of both symbiotic nitrogen fixation (SNF) and other processes. 380 loci were tested for SNF phenotypes and 512 for non-SNF phenotypes. On average, ca. 47 new genes have been characterized each year since 2020 (2025 excluded). The complete dataset consists of 565 published works and bioRxiv preprints as of 16 June 2025.

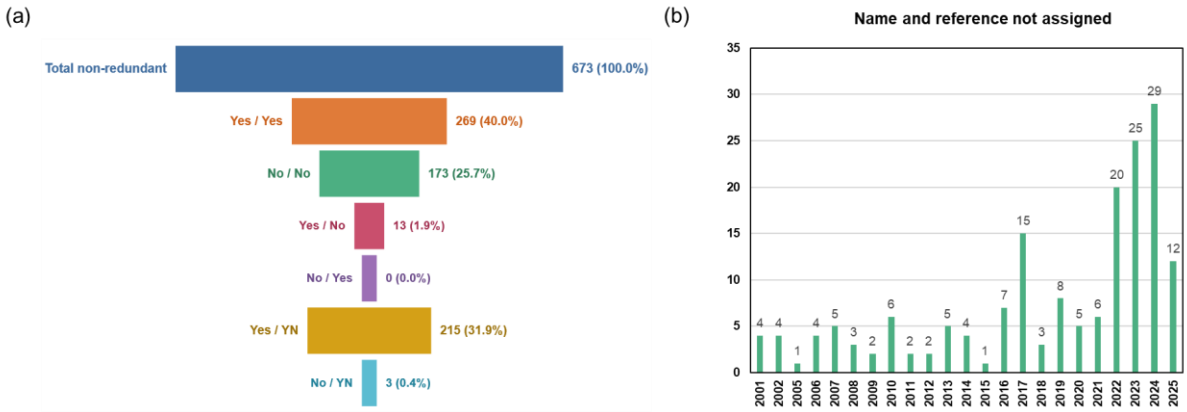

Figure 2. Functional annotation status of 673 genes studied by loss-of-function approaches in M. truncatula, based on the genome browser v. 5.1.9. (a) Six categories of loci based on their annotation status. The first element in each category's code corresponds to a gene name ("Yes" - name assigned, "No" - name not assigned). More than one quarter (26%) of characterized genes are not named (Data S6 Column V). The second element in each category's code corresponds to a reference ("Yes" - reference assigned, "No" - reference not assigned, "YN" for "Yes and No" - an irrelevant reference assigned). More than one quarter (27%) of characterized genes have no reference (Data S6 Column W). Among genes with a reference, 32% have only irrelevant references or references to transcriptomic studies/reviews instead of correct references to loss-of-function studies. (b) The distribution of 173 "missed"

genes over their publication years (category “No/No”, no name and no reference in the genome browser v. 5.1.9). Only the year of the first loss-of-function publication is counted for each gene.

Total count	A, count	C, count	N, count	A, proportion	C, proportion	N, proportion	Protein class
48	34	13	1	71	27	2	Enzyme (kinase)
45	25	10	10	56	22	22	Enzyme (oxidoreductase)
43	29	13	1	67	30	2	Enzyme (transferase)
39	22	16	1	56	41	3	Enzyme (hydrolase)
21	9	12	0	43	57	0	TF (AP2-EREBP family)
19	11	6	2	58	32	11	TF (GRAS family)
18	14	4	0	78	22	0	TF (MYB-HB-like family)
17	1	16	0	6	94	0	TF (SBP family)
13	1	0	12	8	0	92	TF (C2C2-Dof family)
13	5	5	3	38	38	23	Transporter (substrate unknown)
12	8	3	1	67	25	8	TF (MADS-MIKC family)
11	3	4	4	27	36	36	TF (Hap3/NF-YB family)
10	5	3	2	50	30	20	F-box protein
10	8	1	1	80	10	10	TF (C2H2 family)
9	6	2	1	67	22	11	Peptide (cysteine-rich, NCR family)
9	2	7	0	22	78	0	TF (ARF family)
9	6	2	1	67	22	11	TF (bHLH family)
9	1	7	1	11	78	11	TF (Homeodomain-TALE-KNOX family)
8	4	4	0	50	50	0	Enzyme (ligase)
8	1	5	2	13	63	25	Peptide (CEP family)
302	224	46	32	74	15	11	Other
673	419	179	75	62	27	11	All

Figure 3. Different protein classes have different proportions of phenotypes, which may reflect their relative functional importance. Letters A, C, and N stand for altered, conditional, and neutral phenotypes, respectively. The graph shows only the top-20 protein classes. The complete list of protein classes and corresponding statistical analysis are available in Data S10. The summary on phenotypes of combined protein groups is available in Data S11.

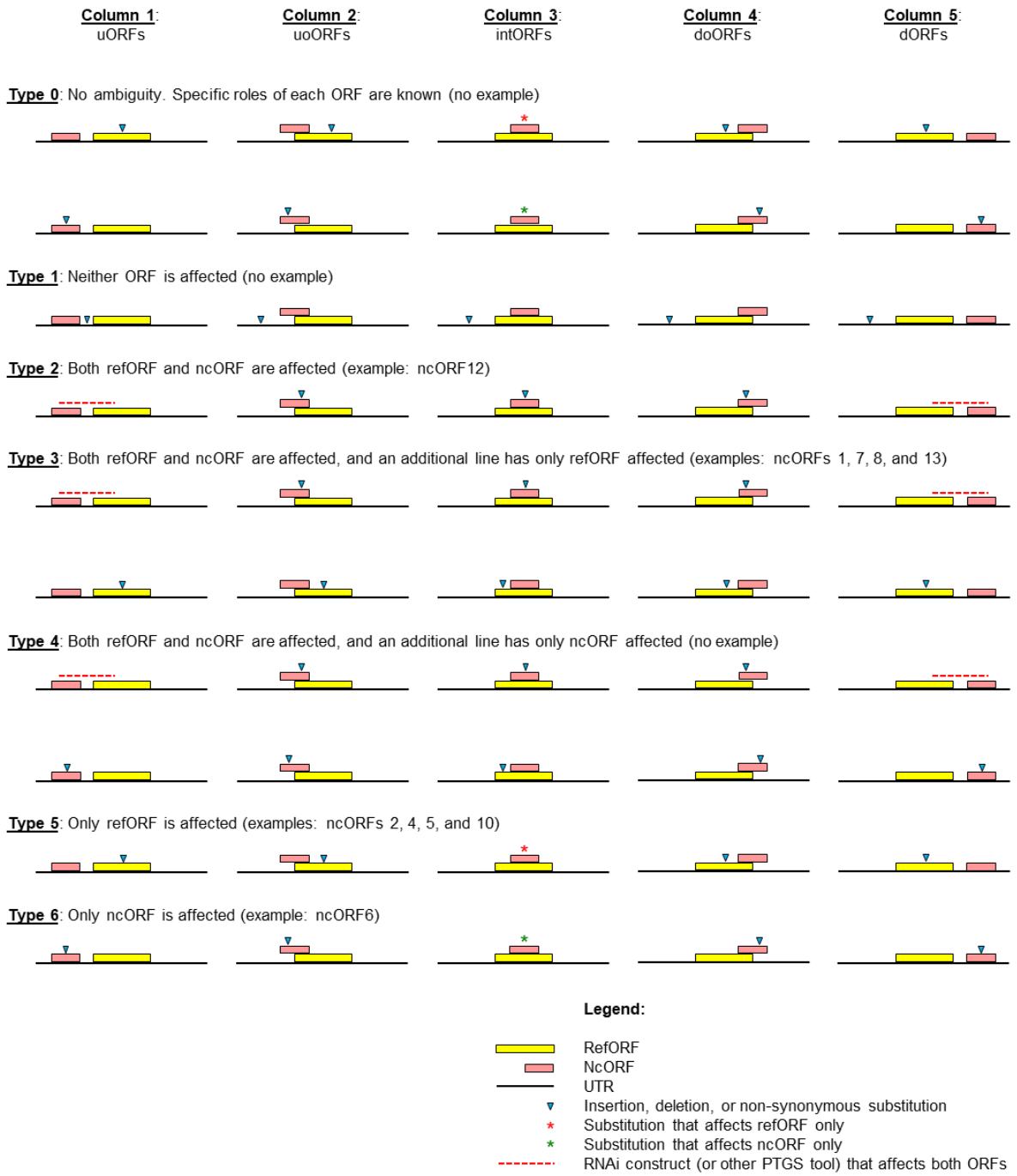

Figure 4. Main types of ambiguity associated with translated ncORFs in loss-of-function studies. Ambiguity types are exemplified based on the assumption that currently annotated gene models are correct. Each column corresponds to a specific location of a ncORF relative to its refORF (uORFs, uoORFs, intORFs, doORFs, and dORFs, respectively, Mudge et al., 2022). Each row corresponds to a specific scenario leading to the inactivation of one or two ORFs. The graph is a simplified illustration of the most common scenarios. It does not apply to the following situations: (1) a mutation outside both ORFs affects their transcription or translation without affecting their amino acid sequences; (2) an insertion at one location interferes with splicing at a different location, thus affecting ORFs in trans. Abbreviations: UTR, untranslated region; RNAi, RNA interference; PTGS, post-transcriptional gene silencing.

SUPPORTING INFORMATION

1394 Additional Supporting Information may be found in the online version of this article.

1395 **Figure S1**. Loss-of-function approaches used for studying 673 genes in *Medicago truncatula*.

1396 **Figure S2**. Biological processes examined in loss-of-function studies on 673 genes in
1397 *Medicago truncatula*.

1398 **Figure S3**. Different groups of protein classes have different proportions of phenotypes, which
1399 may reflect their relative functional importance.

1400 **Figure S4**. Partitioning of phenomics data into altered (A), conditional (C), and neutral (N)
1401 phenotypes.

1402 **Figure S5**. Phenotypes of 219 genes studied in the context of both symbiotic nitrogen fixation
1403 (SNF) and other processes.

1404 **Figure S6**. Aberrant transcript forms recovered from two *Tnt1* mutant lines of *MtSWEET11*
1405 (Kryvoruchko et al., 2016).

1406 **Figure S7**. Frequencies of aberrant transcript forms recovered from two *Tnt1* mutant lines
1407 shown in Figure S6.

1408 **Figure S8**. A ClustalW protein alignment of the mutant version of *MtSWEET11* (Type X) with
1409 the wild-type version from ecotype R108 (the background of the *Tnt1* mutant population).

1410 **Figure S9**. Biological and biosafety implications of induced insertional mutagenesis.

1411 **Table S1**. Loci incorrectly named in the genome browser v. 5.1.9 (an extended version of
1412 Table 3).

1413 **Table S2**. Five v4 gene models that may be correct without splitting.

1414 **Table S3**. MS-validated ncORFs that introduce ambiguity into the interpretation of loss-of-
1415 function studies.

1416 **Data S1**. A master table that summarizes loss-of-function studies on mRNA-refORFs and
1417 ncRNAs tested for phenotypes other than symbiotic nitrogen fixation.

1418 **Data S2**. A master table that summarizes loss-of-function and gain-of-function studies on
1419 ncORFs tested for phenotypes other than symbiotic nitrogen fixation.

1420 **Data S3**. A master table that summarizes loss-of-function studies on mRNA-refORFs and
1421 ncRNAs tested for symbiotic nitrogen fixation phenotypes.

1422 **Data S4**. A master table that summarizes loss-of-function and gain-of-function studies on
1423 ncORFs tested for symbiotic nitrogen fixation phenotypes.

1424 **Data S5**. A master table that summarizes loss-of-function studies on mRNA-refORFs,
1425 ncRNAs, and ncORFs tested for both symbiotic nitrogen fixation (SNF) and non-SNF
1426 phenotypes.

1427 **Data S6**. A master table that summarizes loss-of-function studies on 673 loci in *Medicago*
1428 *truncatula*.

1429 **Data S7**. A list of 565 loss-of-function studies on mRNA-refORFs, ncRNAs, and ncORFs in
1430 *Medicago truncatula*.

1431 **Data S8**. A list of 37 gain-of-function and other non-mutant studies on ncORFs/peptides in
1432 *Medicago truncatula.*

1433 **Data S9**. The complete list of 97 biological processes examined in loss-of-function studies on
1434 673 genes in *Medicago truncatula*.

1435 **Data S10**. The complete list of protein classes, corresponding phenotype counts, and
1436 statistical analysis of proportions.

1437 **Data S11**. The complete list of combined protein groups, corresponding phenotype counts,
1438 and statistical analysis of proportions.

1439 **Data S12**. A master table that summarizes various parameters of 148 ncORFs found in 123
1440 genes targeted by functional analyses in *Medicago truncatula*.

1441 **Data S13**. Visual demonstration of conservation at the amino acid level for eight selected
1442 ncORFs.

1443 **Data S14**. Evidence for purifying selection of eight selected ncORFs.

1444 **Data S15**. 24 nucleotide alignments that demonstrate evidence for purifying selection of eight
1445 selected ncORFs.

1446 **Data S16**. Sequence analysis demonstrating the origin of highly conserved ncORFs in five
1447 selected genes.

1448 **Data S17**. Nucleotide alignments used in Data S16.

1449 **Data S18**. Gene models and mapped ncORFs of 123 loci targeted by functional analyses in
1450 *Medicago truncatula* (Data S12).

1451 **Data S19**. NcProts grouped with refProts during the MS analysis.