Genome-wide analysis of flanking sequences reveals that Tnt1 insertion is positively correlated with gene methylation in Medicago truncatula

Article *in* The Plant Journal · February 2019

DOI: 10.1111/tpj.14291

CITATIONS
0

READS
134

9 authors, including:

Liang Sun
Noble Research Institute
46 PUBLICATIONS **233** CITATIONS
SEE PROFILE

Upinder Gill
University of Florida, Gulf Coast Education and Research Center
36 PUBLICATIONS **239** CITATIONS
SEE PROFILE

Raj Nandety
Noble Research Institute
20 PUBLICATIONS **516** CITATIONS
SEE PROFILE

Perdeep Mehta
Noble Research Institute
53 PUBLICATIONS **2,728** CITATIONS
SEE PROFILE

DR LIANG SUN (Orcid ID : 0000-0002-0321-9222)

DR RAJA SEKHAR SEKHAR NANDETY (Orcid ID : 0000-0002-1129-0790)

DR KIRANKUMAR S MYSORE (Orcid ID : 0000-0002-9805-5741)

Article type : Resource

Genome-wide analysis of flanking sequences reveals that *Tnt1* insertion is positively correlated with gene methylation in *Medicago truncatula*

Liang Sun[1], Upinder S. Gill[1,2], Raja Sekhar Nandety[1], Soonil Kwon[1], Perdeep Mehta[1], Rebecca Dickstein[3], Michael K. Udvardi[1], Kirankumar S. Mysore[1*], Jiangqi Wen[1*]

[1]Noble Research Institute, 2510 Sam Noble Parkway, Ardmore, OK, 73401, USA
[2]Current address: Gulf Coast Research and Education Center, University of Florida, Wimauma, FL, 33598, USA

[3]Department of Biological Sciences, University of North Texas, 1155 Union Circle #305220, Denton, TX, 76203, USA

[*]Correspondence should be addressed to Kirakumar S. Mysore (ksmysore@noble.org) or to Jiangqi Wen (jwen@noble.org)

Running title: *Tnt1* insertion is correlated with gene methylation

Keywords: *Tnt1*, FST, insertional mutagenesis, methylation, random distribution, *Medicago truncatula*

Abstract

From a single transgenic line harboring five *Tnt1* transposon insertions, we generated a near-saturated insertion population in *Medicago truncatula*. Using thermal asymmetric interlaced (TAIL) PCR followed by sequencing, we recovered 388,888 flanking sequence tags (FSTs) from 21,741 insertion lines in this population. FST recovery from 14 *Tnt1* lines using the whole genome sequencing and/or *Tnt1*-capture sequencing approaches suggests an average of 80 insertions per line, which is more than previous estimation of 25 insertions. Analysis of the distribution pattern and preference of *Tnt1* insertions showed that *Tnt1* is overall randomly distributed throughout the *M. truncatula* genome. At the chromosomal level, *Tnt1* insertions occurred on both arms of all chromosomes, with insertion frequency negatively correlated with the GC content. Based on 174,546 filtered FSTs that show exact insertion locations in the *M. truncatula* genome version 4.0 (Mt4.0), 0.44 *Tnt1* insertions occurred per kb and 19,583 genes contained *Tnt1* with an average of 3.43 insertions per gene. Pathway and gene ontology analyses revealed that *Tnt1*-inserted genes are significantly enriched in processes associated with "stress", "transport", "signaling" and "stimulus response". Surprisingly, gene groups with higher methylation frequency were more frequently targeted for insertion. Analysis of 19,583 *Tnt1*-inserted genes revealed that 59% (1,265) of 2,144 transcription factors, 63% (765) of 1,216 receptor kinases and 56% (343) of 616 NBS-LRR genes harbored at least one *Tnt1* insertion, compared to the overall 38% of *Tnt1*-inserted genes out of 50,894 annotated genes in the genome.

Introduction

With more than 18,000 species, legumes are second only to grasses in agricultural importance as food and feed (Graham and Vance, 2003). Legumes are rich sources of dietary proteins, oils, fibres, and/or other beneficial nutrients for humans and livestock (Wang *et al.*, 2003). Legumes play a critical role in sustainable agriculture because they establish symbiosis with rhizobia that fix atmospheric nitrogen and with arbuscular mycorrhizal fungi that provide plant root cells with minerals and water. Globally-cultivated legumes fix 40-60 million tonnes of nitrogen each year (Smil, 1999). Understanding the biology and genetics of symbiotic nitrogen fixation in legumes may enable us to transfer this process to non-legumes, therefore, reduce the massive use of industrially-produced fertilizers (Ane *et al.*, 2008).

Medicago truncatula is a diploid model legume closely related to the tetraploid forage crop, alfalfa (*Medicago sativa*). *M. truncatula* has a relatively small genome compared to other cultivated legumes and a relatively short life cycle, making it a good model and reference species for legume research (Doyle, 1990, Handberg and Stougaard, 1992). Over the years, efforts from the legume community have led to sequencing of the *M. truncatula* genome (Young *et al.*, 2011) and successful development

of protocols for *M. truncatula* regeneration and transformation (Somers *et al.*, 2003), paving the way for large-scale insertional mutagenesis.

Induced mutagenesis has been used widely in plants for applied and fundamental research. Although T-DNA has been used successfully as a biological mutagen in Arabidopsis and rice (Alonso *et al.*, 2003, Sha *et al.*, 2004), it is not feasible in *M. truncatula* due to the lack of a high throughput transformation system and the fact that T-DNA insertion has low copy number in plant genomes. To reach near-saturation mutagenesis, a large number of transformants are required for T-DNA insertion mutagenesis. However, the class I transposable elements (retrotransposons) have gained much attention as an alternative insertional mutagen. Transposition of retrotransposons is activated by various stresses including tissue culture, which has enabled generation of large scale insertion mutant populations in model plants, such as *Transposon of Oryza sativa 17* (*Tos17)* in rice (Yamazaki *et al.*, 2001), *Lotus retrotransposon 1 (LORE1)* in *Lotus japonicus* (Malolepszy *et al.*, 2016) and *Transposable element of tobacco (Nicotiana tabacum) cell type 1 (Tnt1)* or *Tobacco retrotransposon 1* (*Tto1*) in Arabidopsis (Courtial *et al.*, 2000, Okamoto and Hirochika, 2000).

Tnt1 is one of the few well-characterized long terminal repeat retrotransposons in plants (Grandbastien *et al.*, 1989). *Tnt1* is very active in *M. truncatula* during tissue culture but is stable during seed-to-seed propagation (d'Erfurth *et al.*, 2003), therefore, it was chosen to initiate near-saturation mutagenesis in *M. truncatula* (Tadege *et al.*, 2005, Tadege *et al.*, 2008). The high efficiency of *Tnt1* transposition during tissue culture results in multiple insertions of *Tnt1* in regenerated insertion mutant lines (d'Erfurth *et al.*, 2003). The most significant advantage of insertional mutagenesis (such as T-DNA and *Tnt1*) is the ease of lesion identification. Since the sequence of *Tnt1* is known and the *M. truncatula* reference genome is sequenced (Young *et al.*, 2011), the flanking sequences adjacent to *Tnt1* insertion sites can be easily determined in the genome. Using thermal asymmetric interlaced (TAIL)-PCR (Liu *et al.*, 1995, Liu *et al.*, 2005) followed by either Sanger sequencing or Illumina's Miseq next generation sequencing, we sequenced 388,888 *Tnt1* flanking sequence tags (FSTs) from the *M. truncatula Tnt1* insertion population (Cheng *et al.*, 2017b). We previously claimed that *Tnt1* inserts into the *Medicago* genome in a random fashion based on an analysis of 2,461 FSTs (Tadege *et al.*, 2008). The significantly increased number of FSTs prompted us to re-visit our previous conclusions. A comprehensive bioinformatics analysis of *Tnt1* insertion characteristics was performed in this study using currently available FSTs. Specific attention was paid to the coverage of *Tnt1* insertions in different gene groups and, surprisingly, we found that *Tnt1* insertion is positively correlated with methylation frequency in gene groups.

Results

1. Generation of *Tnt1* insertion population, identification and utilization of mutants

After selecting the transgenic line *Tnk88-7-7* generated by d'Erfurth et al. (d'Erfurth *et al.*, 2003) as the starting material, we regenerated a large *Tnt1* insertion population in *M. truncatula* ecotype R108 in a 10-year effort at Noble Research Institute to achieve a near saturation mutagenesis in *M. truncatula*. The population comprises 21,741 independent *Tnt1* insertion lines that are available for functional genomics studies in *M. truncatula*. With a previously estimated average of 25 *Tnt1* insertions per line (Tadege *et al.*, 2005), the *Tnt1* insertion population encompasses more than 500,000 insertions in the *M. truncatula* genome. To promote the use of *Tnt1* mutants, the first research community forward/phenotypic screening workshop was organized in 2005 (Yarce *et al.*, 2013) and continued annually for 13 years. The phenotypic description of lines that display aberrant phenotypes and photographs of all screened lines are available to the public through a web-accessible database (https://medicago-mutant.noble.org/).

From the 21,741 *Tnt1* insertion lines, 388,888 FSTs were recovered by using TAIL-PCR with a *Tnt1*-specific primer and an arbitrary degenerate (AD) primer. The PCR products representing each individual *Tnt1* line were either directly cloned into the pGEM T-easy vector and sequenced by Sanger sequencing or indexed, arranged in two-dimensional pools and subsequently sequenced by paired-end sequencing using the Illumina Miseq platform (Cheng *et al.*, 2017b). All FSTs were compared with the *M. truncatula* A17 reference genome Mt4.0 (pseudomolecules of the R108 genome are not yet available) and mapped to the reference genome. The coordinates of the mapped FSTs on chromosomes are viewable using GBrowse (https://gb.noble.org/cgi-bin/gb2/gbrowse/medicago_4/). Users can check if there are *Tnt1* insertions in their genes of interest by BLAST searching the *Tnt1* FST database (https://medicago-mutant.noble.org/mutant/blast/blast.php) using the R108 genomic sequences of genes of interest as queries. Once a true match between the query sequence and a specific FST (at least 98% sequence identity) is found, the seeds of the *Tnt1* line can be requested through an online ordering system. Although nearly 400,000 FSTs have been sequenced from all the *Tnt1* insertion lines using TAIL-PCR-coupled FST sequencing, there are more insertions in the *Tnt1* population that have not yet been recovered by TAIL-PCR. To make full use of the mutant population, a PCR-based reverse screening approach was developed to identify *Tnt1* insertions for genes of interest in pooled genomic DNA samples (Cheng *et al.*, 2014). Overall, a greater than 85% success rate has been achieved from the reverse screening of 1,305 genes (Cheng *et al.*, 2014). This is an underestimation of the current success rate since many screens were carried out several years ago when the population was smaller than the current population.

From 2006 to January 2019, more than 12,000 mutant lines (from both database searches and reverse screening) have been distributed to 203 research groups in 24 countries. By January 2019, more than 135 peer-reviewed papers have been published using *Tnt1* mutants by these and our own research groups.

2. Distribution of *Tnt1* in the genome and chromosome level

2.1 Tnt1 insertion distribution

Of the 388,888 FSTs, 33,000 were recovered by Sanger sequencing, and the rest were from Illumina sequencing. The average sequence length of these FSTs is 363 bp. Since *Tnt1* insertion lines are in the *M. truncatula* R108 background, all recovered FSTs were first compared with the *M. truncatula* R108 v1.0 genome (Moll *et al.*, 2017) with an e-value ≤ 1.00E-5 and at least 90% similarity with the R108 reference genome sequence. 384,087 (98.8%) FST sequences were successfully mapped to the R108 reference genome.

During the assembly of Illumina Miseq reads, we intentionally kept 28 bp of the *Tnt1* terminal sequence as the signature in the resultant FSTs for ease of identifying exact *Tnt1* insertion sites. To identify the signature sequence in FSTs, we processed all FSTs (see Methods) and obtained 245,381 FST sequences with the signature at either 5' or 3' end. The remaining 143,507 FSTs likely resulted from AD primer end sequencing in the Illumina sequencing. If a TAIL-PCR product is longer than 500 bp, there will be no overlap sequence between the two 250-bp paired-end reads. Since the *Tnt1* signature only exists in reads that were sequenced from the *Tnt1*-specific primer end, the FSTs assembled from reads that were sequenced from the AD primer end will not have the *Tnt1* signature. Out of the 245,381 FSTs, 244,907 (99.8%) were successfully mapped to the *M. truncatula* R108 v1.0 reference genome. The unmapped 474 FSTs presumably correspond to un-sequenced regions of the *M. truncatula* R108 genome.

Due to the unavailability of pseudomolecules, genome annotation, information of pathways, gene ontology and specific gene databases in the R108 genome, we decided to BLAST search the 245,381 FSTs that have *Tnt1* signature sequences (see Methods) against the latest *M. truncatula* A17 version 5 (Mt5.0) reference genome (Pecrix et al., 2018). Among them, 235,056 FST sequences (95.79%) were successfully aligned to the Mt5.0 reference genome with an e-value ≤ 1.00E-5. Although Mt5.0 is a more complete version of the *M. truncatula* reference genome, the gene IDs in Mt5.0 are different from the previous version (Mt4.0). Existing databases such as pathways and gene ontology are all based on the gene IDs in Mt4.0. Therefore, we BLAST search the 245,381 FSTs against the *M. truncatula* A17 version 4 (Mt4.0) reference genome. From the BLAST results, we only chose the FST sequences that have at least 90% sequence identity with Mt4.0 reference sequences and the start position of BLAST hits must be at the first 28 bp of FST sequence to ensure the insertion sites are

accurate and reliable. Ultimately, 174,546 (72.7% of aligned FSTs and 44.9% of total FSTs) filtered FST sequences that meet the above two criteria were used for the following downstream analysis.

From the BLAST results, the exact *Tnt1* insertion site for each of the 174,546 FST sequences was extracted. Based on the insertion positions on chromosomes, the *Tnt1* insertion frequencies in the whole genome were determined by using a window size of 500 kb in Circos software (Krzywinski *et al.*, 2009). We found that *Tnt1* inserts in both arms of each chromosome more frequently than in the centromere regions (Fig. 1). Overall, the *M. truncatula* genome is AT-rich (Young *et al.*, 2011), however, the GC content is relatively high near the centromere regions. To further explore if there is a link between *Tnt1* insertion frequency and GC content at the genome level, the same window size of 500 kb was used to calculate the GC content along the Mt4.0 genome and visualized in Circos (Fig. 1A). Interestingly, the GC content was significantly negatively correlated with *Tnt1* insertion frequency by Pearson correlation analysis (r = -0.56, p-value < 2.2e-16) (Fig. 1B).

To examine how *Tnt1* insertions are distributed on each chromosome, the *Tnt1* insertion numbers for each chromosome were determined. Out of the 174,546 filtered FSTs, 170,899 FSTs were mapped to the eight chromosomes of *M. truncatula*. The remaining 3,647 FSTs were mapped to unanchored scaffolds. Results showed that *Tnt1* insertion frequencies are significantly and positively correlated with the chromosome length (R^2 = 0.87, p = 0.0007) (Fig. 2A). This observation suggests that *Tnt1* is equally and randomly distributed at the genome level but is unequally distributed at the chromosome level (prefers both arms of chromosomes). Based on the mapped 170,899 FSTs, we found that the density of *Tnt1* insertion in the *M. truncatula* genome is 0.44 *Tnt1* insertions per kilo-base pair (Table 1). The actual density of *Tnt1* insertions will be significantly higher than 0.44 because we only used a fraction (170,899 out of 388,888 FSTs) of total *Tnt1* FSTs to calculate the chromosome insertion frequencies and we only recovered 34% of *Tnt1* insertions from the *Tnt1* insertion population by TAIL-PCR (see results in Section 3).

2.2 Tnt1 insertion has no strong target site specificity in the *M. truncatula* genome

From the 174,546 filtered FSTs with exact insertion site, 40 bp genomic sequences upstream and downstream of the insertion sites were extracted. The average occurrence frequencies of A, T, G and C in the 81 bp regions of 174,546 FSTs was calculated. Subsequently, the occurrence frequencies of A, T, G and C at each position from -5 to +5 (with the insertion site designated as position 0) were compared to the average occurrence frequencies of each base pairs in the 81 bp regions. Results showed that although the occurrence frequencies of some bases at specific position are statistically significantly greater (e.g. A at position 0 and G at position -3) or less (e.g. A at position -1 and T at position -3) than the average occurrence frequencies, the target site specificity of *Tnt1* insertion in *M.*

truncatula genome is not strong (Fig. S1). This conclusion is consistent with our previous results from an analysis of much lower number of FSTs (Tadege *et al.*, 2008).

2.3 Tnt1 prefers to insert in exons and gene-rich regions

With greatly increased numbers of *Tnt1* insertion lines and FST sequences, we re-evaluated *Tnt1* insertion preference in the *M. truncatula* genome. The insertion positions of 174,546 FSTs that has at least 90% identity to the *M. truncatula* A17 genome were annotated using ANNOVAR (Wang *et al.*, 2010, Yang and Wang, 2015). It was found that 33.2%, 22.6%, 20.1% and 23.7% FSTs insert in exon, intron + UTR, upstream/downstream regions of a gene, and intergenic regions, respectively (Table 2). Overall 56.1% and 43.9% of *Tnt1* insertions were found in genic and non-genic regions, where non-genic region includes intergenic, upstream (1-kb region upstream of transcription start site) and downstream regions (1-kb region downstream of transcription end site). The percentages of *Tnt1* insertions in exon (33.2%), genic (56.1%) and non-genic (43.9%) regions are very close to those of our previous analysis based on 964 FSTs (34.1%, 57.3% and 42.6%, respectively) (Tadege *et al.*, 2008), further confirming our previous conclusion that *Tnt1* prefers inserting in exons and gene-rich regions.

3. Distribution of *Tnt1* insertions in individual mutant line

Though the overall distribution of *Tnt1* insertions in the *M. truncatula* genome is random (Fig. 1A), this does not necessarily mean that the distribution of *Tnt1* insertions is random in individual lines. To examine the distribution pattern of *Tnt1* insertions in individual lines, the whole genome sequencing (WGS) and *Tnt1*-capture sequencing (see Methods) approaches were used to maximize the identification of *Tnt1* insertion sites in a few selected *Tnt1* insertion lines. WGS and *Tnt1*-capture were also important to determine the average percentage of FSTs recovered by TAIL-PCR. We sequenced four lines by WGS and 10 lines by *Tnt1*-capture. Two of these lines were sequenced by both WGS and *Tnt1*-capture to compare the efficacy of the two approaches. Results indicated that all the 164 insertions identified by WGS were also recovered by *Tnt1*-capture in NF1962. In NF11044, there is only one insertion that was identified by WGS but not recovered by *Tnt1*-capture and vice-a-versa, and the remaining 16 insertions were recovered by both WGS and *Tnt1*-capture (Table 3), suggesting that both WGS and *Tnt1*-capture recover almost all insertions. We also used WGS data from three additional *Tnt1* lines that have already been published elsewhere (see Table 3 for sources) for our analyses.

It is noteworthy that the average FST number recovered by TAIL-PCR in the 15 listed lines is 27, which is very close to our previously estimated 25 *Tnt1* insertions per line. However, the average of 27 FSTs recovered by TAIL-PCR is only 34% of actual FSTs present in the 15 sequenced lines (Table 3). Therefore, our new estimation of average *Tnt1* insertions per line is 80, which is substantially higher than the number previously estimated (Tadege *et al.*, 2008). In summary, TAIL-PCR from either the forward and/or reverse direction recovers roughly one third of total flanking sequences in a given line. Although we are cautious to give an actual average number of insertions per line in the *Tnt1* population based on the limited number of lines examined by WGS and *Tnt1*-capture sequencing, it is safe to conclude that the number of insertions is much greater than previously estimated 25. We now estimate that the 21,741 *Tnt1* insertion lines contain ~1.7 million insertions.

In order to examine the distribution pattern of *Tnt1* insertions in individual lines, eleven individual *Tnt1* insertion lines from Table 3 with low, intermediate and high insertion numbers were selected and insertions in each line were mapped to their locations on chromosomes. As shown in Fig. S2 to Fig. S12, regardless of the insertion numbers, in most cases, *Tnt1* inserted randomly in each individual line, though occasionally clusters of insertions were observed in specific chromosomes. The random distribution pattern was especially apparent in those lines with high *Tnt1* insertion numbers (Fig. S6-S12). Therefore, we conclude that *Tnt1* insertions are randomly distributed at both the population and individual line levels.

4. Genes containing *Tnt1*

In the *M. truncatula* Mt4.0 reference genome, 50,894 genes were predicted and annotated. From 174,546 FSTs with exact *Tnt1* insertion sites, 19,583 (38% of total genes) genes were found to contain *Tnt1* in at least one insertion line and 6,105 genes with at least four insertions. An average of 3.43 insertions per gene was observed, with the most frequently hit gene (Medtr5g099060) containing 60 *Tnt1* insertions in the population as a whole (Fig. 3A and Supplementary file1). In fact, the filtered 174,546 FSTs used for this analysis only account for 44.9% of the total FSTs that were recovered by TAIL-PCR, and we were only able to recover about 34% of all *Tnt1* insertions in the population by TAIL-PCR (Table 3). Therefore, it is reasonable to assume that most genes in the *M. truncatula* genome are disrupted by *Tnt1* in the insertion population.

4.1 Systematic analysis of inserted genes

MapMan (Thimm *et al.*, 2004) was used to enrich the 6,105 genes that are inserted more than the average insertion number (i.e. 3.43 insertions per gene). Results showed that these frequently inserted genes fall into the following five out of the 35 MapMan pathways: “stress”, “signaling”, “secondary metabolism”, “transport”, and “nucleotide metabolism” (Fig. 3B and Supplementary file 2).

Gene ontology analysis is widely used to enrich genes to biological processes in transcriptomic studies. We also employed AgriGO v2.0 (Tian *et al.*, 2017) to enrich the frequently inserted genes in gene ontology categories. The 6,105 frequently inserted genes were assigned to 59 GO terms (Supplementary file 3). To better visualize the significantly affected GO terms, ReviGO (Supek *et al.*, 2011) was used to remove redundant GO terms and the remaining terms was visualized in semamtic similarity-based scatterplots. The significant GO terms under the biological processes are "response to stress", "response to stimulus", "defense response", "protein phosphorylation", and "transmembrane transport". The significant GO terms under molecular functions are "ATP binding", "active transmembrane transporter activity", "protein tyrosine kinase activity", and "transporter activity". The results of GO enrichment analysis were very similar to the pathway analysis results.

From pathway and GO analyses, we found that some gene groups were significantly enriched with *Tnt1* insertions. We were interested in deciphering the underlying mechanism of this enrichment. DNA methylation has been reported to play a vital role in the inactivation of transposons (Saze *et al.*, 2012). Therefore, we first explored the relationship between the *Tnt1* insertion and methylation frequency in these gene groups. All gene groups were either downloaded from public databases or extracted from the *M. truncatula* 4.0 genome annotation file (see Methods for details). The DNA methylation data of *M. truncatula* wild-type R108 (Niederhuth *et al.*, 2016) were used to calculate methylation frequency, where genes with "CG", "CHG" or "CHH" methylation were considered as methylated genes. A statistical analysis of the occurrence frequency of methylated genes and the occurrence frequency of genes with more than the average *Tnt1* insertion number (> = 4) in 44 gene function groups was summarized in Figure 4A and Supplementary file 4. Gene groups "transporter", "receptor kinase", "secondary metabolism" and "cytochrome_P450" that show high methylation frequencies were also more frequently inserted by *Tnt1*, while nodule-related and vesicle-related genes with low methylation frequencies were less frequently inserted (Figure 4A). The *Tnt1* insertion frequency was positively correlated with the methylation frequency in gene groups ($R^2 = 0.57$) (Figure 4B).

To further unveil the mechanism of *Tnt1* insertion preference, correlations between *Tnt1* insertion frequency and gene expression levels or GC contents in the aforementioned gene groups were investigated. Five microarray data sets were downloaded from *M. truncatula* gene expression atlas (see Methods). We first examined the relationship between the *Tnt1* insertion frequency and the expression level of individual genes. Though no clear correlation was observed in the analysis (Figure 4C), we did observe an intriguing phenomenon that the most frequently inserted genes are always lowly expressed, whereas highly expressed genes are always less frequently inserted by *Tnt1* (Figure 4C). Next, we calculated the average expression level of each group of genes that have four or more *Tnt1* insertions. Correlation analysis indicated that there was no statistically significant correlation between the *Tnt1* insertion frequency and the average gene expression (Figure 5A). Furthermore, we

calculated the average GC content in each gene group that was used for methylation and gene expression analysis and analyzed its relationship with the *Tnt1* insertion frequency. Results showed that no significant correlation was observed (Figure 5B).

From these systematic analyses, we found that *Tnt1* insertion has a significant correlation with gene methylation but no correlation with gene expression levels or GC contents.

4.2 Important gene groups inserted by *Tnt1*

It was noted from the aforementioned analyses on pathways and gene ontology that several groups of genes that play important roles in stress responses and signaling are significantly affected by *Tnt1* insertions. To get more insights, a detailed analysis was performed on three large gene groups that are involved in "stress" and "signaling": transcription factors, protein kinases and disease resistant genes.

Transcription factors (TF)

Transcription factors are master regulators that switch gene expression on or off, or at least modulate expression, by binding to promoter regions of genes. Manipulation of TF expression (e.g. knockout or over-expression) often drastically affects plant phenotypes (Riano-Pachon *et al.*, 2007). Through forward genetics screenings, several *M. truncatula* developmental defects involving nodules, roots, leaves and flowers have been identified and in many cases, the causative *Tnt1* is inserted in a gene encoding a transcription factor (Wang *et al.*, 2008, Chen *et al.*, 2010, Tadege *et al.*, 2011, Zhou *et al.*, 2011, Cheng *et al.*, 2012, Couzigou *et al.*, 2012, Pislariu *et al.*, 2012, Uppalapati *et al.*, 2012, Verdier *et al.*, 2012, Zhou *et al.*, 2014, Niu *et al.*, 2015, Cerri *et al.*, 2016, Duan *et al.*, 2017, Gou *et al.*, 2018)[34-47]. To evaluate how many transcription factors (TFs) are disrupted by *Tnt1* insertions in *M. truncatula Tnt1* insertion population, we downloaded *M. truncatula* TFs from the Plant Transcription Factor database (PlantTFDB). There are 57 annotated TF families and 2,144 putative TFs in the current database; 53 out of the 57 families (93%) and 1,265 out of the 2,144 genes (59%) contain *Tnt1* insertions in the mutant population, based on 174,546 filtered FSTs (Fig. 6A, C). The percentage (59%) of disrupted TFs is far greater than that of all genes (38%). The gene (Medtr5g099060) with most *Tnt1* insertions (60 insertions) is a plant-specific TF containing the RWP-RK motif. The detailed list of TF genes inserted by *Tnt1* is in Supplementary file 5 & 6, which includes the list of TF families, their gene members, insertion frequencies in our *Tnt1* database and top 100 most frequently inserted TF genes. This list will assist researchers to identify TFs of interest for their reverse genetics analyses. Five TF families "RAV", "HRT-like", "LFY", "SAP","DBB" have *Tnt1* insertions in all family members. TF families with >= 80% members that are inserted with *Tnt1* are listed in Supplementary Table S1. Although the frequencies in the three largest TF families ERF,

MYB and bHLH are not very high, many of the important TFs with known function in these three families are inserted by *Tnt1* with different insertion numbers and are highlighted in Supplementary file 6.

Receptor kinases and NBS-LRR genes

Receptor kinases are key players involved in important growth and development processes, responses to environmental cues, and hormone perception and pathogen recognition (Song *et al.*, 1995, Li and Chory, 1997, Gomez-Gomez and Boller, 2000, Limpens *et al.*, 2003, Haruta *et al.*, 2014). Like TFs, many characterized *Tnt1* insertion mutants have insertions in protein kinases (Berrabah *et al.*, 2014, Huault *et al.*, 2014, Boivin *et al.*, 2016, Cheng *et al.*, 2017a) . Most of the disease resistance genes (*R* genes) in plants cloned to date encode nucleotide-binding site leucine-rich repeat (NBS-LRR) proteins, which are thought to monitor the status of plant proteins that are targeted by pathogen effectors, thus are also called adaptable guards (McHale *et al.*, 2006). NBS-LRR genes are specifically enriched in the *M. truncatula* genome, more than other sequenced plant genomes like Arabidopsis, rice, and poplar (Young *et al.*, 2011). Therefore, we examined the frequencies of *Tnt1* insertions in receptor kinases and NBS-LRR genes in the *Tnt1* insertion population. In total, 1,216 and 616 receptor kinases and NBS-LRR genes, respectively, were found from the Mt4.0 reference genome annotation. Based on the 174,546 filtered FSTs in our *Tnt1* insertion population, 765 (62.9%) receptor kinases and 343 (55.6%) NBS-LRR genes are inserted by *Tnt1* (Figure 6B). The coverages of *Tnt1* insertions in these two gene groups are also markedly higher than the overall coverage (38%). In the top 15 most inserted genes, five are receptor kinases with at least 30 *Tnt1* insertions per gene. The detailed inserted receptor kinase and NBS-LRR genes are listed in Supplementary files 7 & 8.

Discussion

Insertional mutagenesis is a powerful approach for mutant generation and subsequent mutation identification. Large-scale mutant populations have been generated successfully in different model plant species using different insertional mutagens, such as Arabidopsis by T-DNA, rice by T-DNA and *Tos17*, and *Lotus japonicus* by *LORE1* (Alonso *et al.*, 2003, Sha *et al.*, 2004, Malolepszy *et al.*, 2016). The most significant advantage of insertional mutagenesis is the convenience for reverse genetics because the known sequences of insertional mutagens facilitate the ease of identifying mutated or disrupted genes. From a single *M. truncatula* R108 starter line harboring five *Tnt1* insertions (Tadege *et al.*, 2008), a near-saturated insertion population with 21,741 independent lines was generated through tissue culture and regeneration. Most (98.8%) of the recovered 388,888 FSTs can be mapped onto the R108 genome v1.0. The small portion (1.2%) of unmapped FSTs probably represents insertions recovered from the un-sequenced regions in the genome. From the 388,888

FSTs, 174,546 filtered FSTs that have the *Tnt1* end sequence signature were used for *Tnt1* insertion feature analysis. Our results indicate that *Tnt1* insertions are randomly distributed at both genome and chromosome levels in the entire population and in individual lines. The random distribution feature of *Tnt1* in *M. truncatula* is consistent with our previous analysis from only 1,935 FSTs (Tadege *et al.*, 2008). This feature of *Tnt1* insertion in *M. truncatula* is different from another retrotransposon *LORE1*, which not only avoids intergenic regions but also shows a non-random distribution pattern in the gene space of the *L. japonicus* genome (Malolepszy *et al.*, 2016).

Unlike T-DNA insertions, which occur more randomly in the Arabidopsis genome (Alonso et al., 2003), *Tnt1* prefers the AT rich region at the genome level (Fig 1A). The *M. truncatula* genome is overall AT rich (Young *et al.*, 2011), but near the centromere of each chromosome, the GC content is apparently higher than that in both arms (Fig. 1A). The *Tnt1* insertion frequency is lower in centromere region than in both arms of each chromosome. Similarly, the rice genome is also AT rich, the insertion of *Tos17* also avoids the GC-rich heterochromatin regions. Similar to *Tnt1*, insertion events of *Tos17* are more frequent in genic regions than in intergenic regions. The frequency of *Tos17* insertions in genic regions of the rice genome is three times of that in intergenic regions (Miyao *et al.*, 2003). Although both *Tnt1* and *LORE1* are retrotransposons and *M. truncatula* and *L. japonicus* are both legumes, the insertion characteristics of *Tnt1* and *LORE1* is apparently different. While *Tnt1* randomly inserts in the *M. truncatula* genome, 74% of *LORE1* insertions non-randomly target genic regions of the *L. japonicus* genome (Malolepszy *et al.*, 2016). *LORE1* is an endogenous retrotransposon of *L. japonicus* and the transposition of *LORE1* is achieved through germ-line activation (the activation takes place in generative tissues (Fukai *et al.*, 2010)). *Tnt1*, however, was isolated from tobacco, thus a heterologous transposable element in the *M. truncatula* genome. *Tnt1* is activated during the early callus induction stage of tissue culture (Tadege *et al.*, 2008) and cannot be activated during germ-line propagation. Though the exact mechanism of the difference in insertion characteristics of *Tnt1* and *LORE1* is not fully understood, the difference in the origination and the activation feature may contribute to some extent.

Tnt1 insertion has neither genome-wide hot spots nor target site specificity in the *M. truncatula* genome (Tadege *et al.*, 2008; Fig. S1). Target site specificity is a typical feature of many LTR retrotransposons (Miyao *et al.*, 2003, Naito *et al.*, 2006). For example, short palindromic consensus sequences flanking the insertion sites are observed in LTR retrotransposons *Tos17* and *LORE1* and in many retroviruses that carry LTRs (Miyao *et al.*, 2003, Naito *et al.*, 2006). However, *Tnt1* heterologous insertions in *M. truncatula* has a different scenario. Although we observed statistically significant base preference at several positions up- or down-stream of the insertion sites from small- (Tadege *et al.*, 2008) and large-scale FST analyses (Fig. S1), the base preference was not strong enough to establish any unique consensus motifs or signatures for *Tnt1* insertion.

We previously estimated that the average insertion number per line in the *Tnt1* mutant population is 25 based on Southern blot analysis and recovered FST numbers using TAIL-PCR (Tadege *et al.*, 2005, Tadege *et al.*, 2008). This number is significantly higher than that of *LORE1* (average 4.76 insertions) and *Tos17* (average 3.37 insertions in rice cultivar Nipponbare) (Piffanelli *et al.*, 2007, Malolepszy *et al.*, 2016). Using WGS and *Tnt1*-capture approaches, we identified a markedly increased average number of *Tnt1* insertions per line, suggesting that the previously estimated average number of 25 insertions per line was an underestimation. This underestimation is mostly because of the low sensitivity of Southern blot analysis as well as the limitation of TAIL-PCR. Whether a TAIL-PCR is successful or not largely depends on whether there is a suitable binding site for AD primers near the insertion location. If there are only a few insertions in a genome, the competition and/or interference between primers is limited, therefore, most insertions can be amplified. However, if there are tens or a few hundred insertions in one line, the competition or interference between specific primers and/or AD primers is inevitable. From the results of 15 *Tnt1* lines sequenced by WGS and/or *Tnt1*-capture, the average recovery rate by TAIL-PCR was determined to be about 34%. It is now estimated that the *M. truncatula Tnt1* insertion lines have an average of ~80 insertions per line. Therefore, 21,741 independent regenerated *Tnt1* insertion lines encompass ~1.7 million insertions. Based on the equation, $P = 1 - (1-[x/y])^{R_{tr}.R_{gs}.N}$ (Tadege *et al.*, 2008, Krzywinski *et al.*, 2009), where P is the probability of finding one *Tnt1* insert within a given gene, x is the average length of a gene transcript (1.7 kb), y is the total length of transcribed region in the gene space (190 Mb from Mt4.0), Rtr is the probability of insertion into a transcription region within the gene space (0.561), Rgs is the probability of insertion into gene space in the whole genome (0.988), and N is the total number of *Tnt1* inserts required, we found that the probability of finding a *Tnt1* insert in any given gene of *M. tuncatula* in our population is ~ 99%. However, this estimate is based on the assumption that the length of a gene is 1.7 kb. For genes that are less than 1 kb, the probability of finding a *Tnt1* insert will slightly reduce. Nevertheless, using *Tnt1* as an insertional mutagen, a near-saturation mutagenesis has been achieved in *M. truncatula*.

In addition to confirming the conclusions we made from a small-scale FST analysis (Tadege *et al.*, 2008, Krzywinski *et al.*, 2009), we further examined *Tnt1* insertions in some specific gene groups from pathway and gene ontology enrichment analyses. Several pathways, such as “stress”, “signaling” and “nucleotide metabolism”, which are related to stress or stimulus responses, stand out from the analyses. We speculate that this could be ascribed to the procedure we used to generate the *Tnt1* insertion lines. Unlike other characterized retrotransposons such as *LORE1* and *Tos17*, *Tnt1* is activated during the early callus induction stage from leaf explants in *M. truncatula* (Tadege *et al.*, 2008, Krzywinski *et al.*, 2009). Compared to the homeostatic hormonal state in plants, leaf explants on the callus induction medium are exposed to a dramatic shocking environment, especially the hormone changes. In this environment, leaf explant cells experience various stresses and stimuli.

From a specialized and ordered organ (leaf) to pluripotent cell mass (callus), callus formation represents an important cell reprogramming process (Xu *et al.*, 2018). In Arabidopsis, Fan et al. (2012) (Fan *et al.*, 2012) reported that auxin-induced LATERAL ORGAN BOUNDARIES DOMAIN (LBD) transcription factors act as the master regulators to trigger auxin-induced callus formation. During the reprogramming process, genes that are involved in stress responses, phosphorylation, and transcription regulation might be actively modified by methylation. When *Tnt1* is activated during the reprogramming process, these genes may be more frequently targeted. We did observe a strong positive correlation between the *Tnt1* insertion frequency and the methylation frequency in these gene groups (Fig. 4B) as well as significantly increased frequencies of *Tnt1* insertions in transcription factors, receptor kinases and NBS-LRR genes (Fig. 6). In contrary to these results, *LORE1* of *L. japonicus* prefers to insert in CHG-hypomethylated genes (Malolepszy *et al.*, 2016). This is consistent with the idea that chromatin accessibility for retrotransposons is correlated with the methylation status of DNA (Jin *et al.*, 2017). It is intriguing why *Tnt1* inserts in a hyper-methylated region whose chromatin would be less accessible compared to other regions of the chromosome. Possible explanation for this would be that the methylation status of DNA might vary for different tissues and conditions in which the plant is growing. Since *Tnt1* transposition occurs during callus formation, the methylation status of genes during this condition could be entirely different from a soil grown plant from which the methylation data were generated. It warrants further investigation to determine why *Tnt1* prefers to insert in highly methylated genes. Nevertheless, this study using comprehensive analyses of FSTs showed preference for *Tnt1* insertion and this will be highly useful for the applicability of *Tnt1*-mediated insertional mutagenesis not only in *M. truncatula* but also in other plant species.

Methods

FST isolation and sequencing

Genomic DNA was isolated as described (Doyle, 1990). FSTs were recovered by TAIL-PCR (Liu *et al.*, 1995, Liu *et al.*, 2005) using *Tnt1*-specific primers in combination with five arbitrary degenerate primers (AD1-3, AD5-6, Cheng et al., 2017). Two rounds of TAIL-PCR were used to amplify specific amplicons. For the first about 2,650 *Tnt1* lines, TAIL-PCR was carried out using genomic DNA from individual lines as the templates. For each line, the secondary PCR products amplified from *Tnt1*-specific primer in combination with one of five AD primers were pooled, purified and cloned into T-easy vector (Promega, Madison, WI) and the cloned plasmids were sequenced using the traditional Sanger sequencing method and the sequences of each line were manually analyzed (Cheng *et al.*, 2014). For the remaining 19,050 lines, we used a 2-dimensional pooling strategy to pool genomic DNA samples from 400 lines into 20 X-pools and 20 Y-pools. Each pool was used as a TAIL-PCR template. *Tnt1*-specific primers in combination with one of the five AD primers were used for TAIL-

PCR. The *Tnt1*-specific secondary primers were indexed for each X and Y pools. Secondary TAIL-PCR products from 40 pooled templates were pooled, purified and sequenced using the Illumina Miseq platform. The sequencing reads were de-multiplexed, index-sorted, trimmed, and assembled using Trinity software as described (Cheng *et al.*, 2017b).

Tnt1-capture sequencing

For *Tnt1*-capture sequencing, a total of four 5'-biotinylated xGen Lockdown Probes (Integrated DNA Technologies, Inc., Skokie, IL) of 120 bp were synthesized using the end sequence of *Tnt1*-LTR region. Bar-coded Illumina libraries were prepared individually for 10 *M. truncatula Tnt1* insertion lines. Illumina libraries were pooled before hybridization-based capture of *Tnt1* specific fragments. For hybridization capture, protocol instructions and reagents provided by IDT (Integrated DNA Technologies, Inc., Skokie, IL) were used with some modifications. We used 10 µl of SeqCap EZ Developer Reagent (Roche Sequencing, Pleasanton, CA) instead of 5 µg of human Cot-1 DNA as a blocking reagent. After hybridization capture, *Tnt1* fragment-enriched-libraries were evaluated for enrichment efficiency by quantitative PCR prior to sequencing. Captured DNA fragments were sequenced using Illumina NextSeq. All Illumina sequences were trimmed by Trimmomatic (Bolger et al., 2014; LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:30). *Tnt1* insertions were identified by ITIS (Jiang et al., 2015; total supporting reads at detected insertion t = 10).

FST sequence preprocessing and annotating

To accurately identify *Tnt1* insertion sites in the *M. truncatula* genome, all FST sequences with length shorter than 50bp or without the *Tnt1* signature sequence ("CCCAACA", "CATCATCA" or "TGATGATGTCC") or the *Tnt1* signature sequence not within 28bp from the beginning or at the end of FST sequences were discarded. The preprocessed reliable FST sequences were aligned to the *M. truncatula* A17 version4 (Mt4.0) or version5 (Mt5.0) reference genome using BLASTN with an e-value threshold ≤ 1.00E-5. The FST sequences with best hit from BLAST analysis were further processed to filter incorrect alignments if the similarity score is less than 90%. The final FST sequences with the Mt4.0 genome alignment information were annotated by using ANNOVAR (Wang *et al.*, 2010, Yang and Wang, 2015).

Visualization and Statistics of *Tnt1* insertion Frequency

GC content was calculated as following:

$$\text{GC content} = \frac{G+\text{C}}{\text{A+T+G+C}} \times 100\%$$

Circos was used to visualize the GC content and *Tnt1* insertion frequencies.

Pearson correlation (cor.test) was used to calculate correlation values between the GC content and the *Tnt1* insertion frequency in Rstudo (Version 0.99.489). Linear regression model was used to test the relationship between the chromosome length and *Tnt1* insertion frequencies.

Functional gene group analysis

All *M. truncatula* transcription factors (http://planttfdb.cbi.pku.edu.cn/index.php?sp=Mtr) were downloaded from PlantTFDB database (Jin *et al.*, 2017)[6]. We compare the *Tnt1* inserted genes in our databases with all *M. truncatula* transcription factors in PlantTFDB database. All receptor kinases and NBS-LRR related genes were extracted from *M. truncatula* A17 reference genome annotation file (https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Mtruncatula). If gene annotation contains key words "receptor" and "kinase", or "NBS-LRR", we will consider this gene as a receptor kinase or NBS-LRR gene, respectively. All other gene groups are chosen based on their significance and their role in plant biology and also by their broad conservation across genomes. Several groups of genes in secondary metabolites were chosen as they are represented in majority or unique to *M. truncatula* compared to other reference genomes such as Arabidopsis or rice. All genes in each group were manually curated based on the key words in the Mt4.0 gene annotation file. The resulting gene numbers in each group were also cross checked with *M. truncatula* Genome Database web pages with same verbal cues and was found to yield same number of sequences as obtained through the stand alone search.

Gene expression analysis

All the gene expression datasets were downloaded from *M. truncatula* gene expression atlas (https://mtgea.noble.org/v3/). The following five different gene expression datasets were downloaded using the differential gene expression analysis datasets with search criteria of greater of 0.01 fold change in all of the target experiments; a) Root Macrophomina infected 48 hpi (Ref Expt. #128) with 0 hpi control (Ref Expt. #125) b) root hairs 3 dpi Sm1021 (Ref Expt. #259) with 1 dpi Sm1021 control (Ref Expt. #258) c) Root 2d Sdl culture 180mM NaCl 48h (Ref Expt. #82) with Root 2d Sdl culture 180mM NaCl 0h (Ref Expt. #79) d) Root 2wk Sdl hydroponic 200mM NaCl 24h (Ref Expt. #87) with Root 2wk Sdl hydroponic 200mM NaCl 0h (Ref Expt. #84) e) cell suspension methyl jasmonate treated 24h post treatment (Ref Expt. #5) with control treatment (Ref Expt. #6). Basal expression values from each of those datasets were used as expression basal values before fold change (FC) calculations for the purpose of *Tnt1* insertion analysis. The probesetID in the experiments were compared with the *M. truncatula* gene ID and the comparative gene ID was extracted along with the basal expression values and used for the *Tnt1* gene insertion analysis.

Acknowledgements

Development of the *M. truncatula Tnt1* mutant population and FST recovery was funded by the National Science Foundation, USA (DBI-0703285, IOS-1127155, and IOS-1733470) and by Noble Research Institute. We thank Dr. Million Tadege, Dr. Hee-Kyung Lee, Miki Hartwell, Janie Gallaway, Trina Cottrell and Haidi Tu for their excellent technical assistance in tissue culture and plant maintenance in the greenhouse. We also thank Dr. Xiaofei Cheng and Shulan Zhang for their assistance in FST recovery.

Author contributions

L.S., K.S.M and J.W. conceived and designed the study; L.S., J.W., R.D., U.S.G., R.N., and M.P. performed data analysis; L.S. and S.K. conducted the statistical analyses; L.S., M.K.U., J.W. and K.S.M. wrote the manuscript.

Competing interests: The authors declare no competing interests.

Supporting legends

Supplementary Figure S1. *Tnt1* insertion has no strong preference on target sites in the genome.

Supplementary Figure S2. Distribution of 16 *Tnt1* insertions in line NF11044.

Supplementary Figure S3. Distribution of 17 *Tnt1* insertions in line NF13918.

Supplementary Figure S4. Distribution of 30 *Tnt1* insertions in line NF8208.

Supplementary Figure S5. Distribution of 41 *Tnt1* insertions in line NF2085.

Supplementary Figure S6. Distribution of 91 *Tnt1* insertions in line NF10739.

Supplementary Figure S7. Distribution of 93 *Tnt1* insertions in line NF2258.

Supplementary Figure S8. Distribution of 116 *Tnt1* insertions in line NF11579.

Supplementary Figure S9. Distribution of 136 *Tnt1* insertions in line NF20764.

Supplementary Figure S10. Distribution of 144 *Tnt1* insertions in line NF0707.

Supplementary Figure S11. Distribution of 160 *Tnt1* insertions in line NF1692.

Supplementary Figure S12. Distribution of 246 *Tnt1* insertions in line NF0365.

Supplementary Table S1. The most frequently targeted TF families by *Tnt1* insertion.

Supplementary File 1. Genes inserted in the *Medicago truncatula Tnt1* FST database.

Supplementary File 2. Pathway analysis of *Tnt1*-inserted genes with at least four insertions.

Supplementary File 3. GO analysis of *Tnt1*-inserted genes with at least four insertions.

Supplementary File 4. *Tnt1* insertion and methylation frequencies in different groups of genes.

Supplementary File 5. Transcription factor families inserted by *Tnt1*.

Supplementary File 6. Transcription factor genes inserted by *Tnt1*.

Supplementary File 7. NBS-LRR genes inserted by *Tnt1*.

Supplementary File 8. Receptor kinase genes inserted by *Tnt1*.

Reference

Alonso, J.M., Stepanova, A.N., Leisse, T.J., Kim, C.J., Chen, H., Shinn, P., Stevenson, D.K., Zimmerman, J., Barajas, P., Cheuk, R., Gadrinab, C., Heller, C., Jeske, A., Koesema, E., Meyers, C.C., Parker, H., Prednis, L., Ansari, Y., Choy, N., Deen, H., Geralt, M., Hazari, N., Hom, E., Karnes, M., Mulholland, C., Ndubaku, R., Schmidt, I., Guzman, P., Aguilar-Henonin, L., Schmid, M., Weigel, D., Carter, D.E., Marchand, T., Risseeuw, E., Brogden, D., Zeko, A., Crosby, W.L., Berry, C.C. and Ecker, J.R. (2003) Genome-wide insertional mutagenesis of Arabidopsis thaliana. *Science*, **301**, 653-657.

Ane, J.M., Zhu, H. and Frugoli, J. (2008) Recent Advances in Medicago truncatula Genomics. *Int J Plant Genomics*, **2008**, 256597.

Berrabah, F., Bourcy, M., Eschstruth, A., Cayrel, A., Guefrachi, I., Mergaert, P., Wen, J.Q., Jean, V., Mysore, K.S., Gourion, B. and Ratet, P. (2014) A nonRD receptor-like kinase prevents nodule early senescence and defense-like reactions during symbiosis. *New Phytol*, **203**, 1305-1314.

Boivin, S., Kazmierczak, T., Brault, M., Wen, J., Gamas, P., Mysore, K.S. and Frugier, F. (2016) Different cytokinin histidine kinase receptors regulate nodule initiation as well as later nodule developmental stages in Medicago truncatula. *Plant Cell Environ*, **39**, 2198-2209.

Cerri, M.R., Frances, L., Kelner, A., Fournier, J., Middleton, P.H., Auriac, M.C., Mysore, K.S., Wen, J., Erard, M., Barker, D.G., Oldroyd, G.E. and de Carvalho-Niebel, F. (2016) The Symbiosis-Related ERN Transcription Factors Act in Concert to Coordinate Rhizobial Host Root Infection. *Plant Physiol*, **171**, 1037-1054.

Chen, J., Yu, J., Ge, L., Wang, H., Berbel, A., Liu, Y., Chen, Y., Li, G., Tadege, M., Wen, J., Cosson, V., Mysore, K.S., Ratet, P., Madueno, F., Bai, G. and Chen, R. (2010) Control of dissected leaf morphology by a Cys(2)His(2) zinc finger transcription factor in the model legume Medicago truncatula. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 10754-10759.

Cheng, X., Gou, X., Yin, H., Mysore, K.S., Li, J. and Wen, J. (2017a) Functional characterisation of brassinosteroid receptor MtBRI1 in Medicago truncatula. *Sci Rep*, **7**, 9327.

Cheng, X., Krom, N., Zhang, S., Mysore, K.S., Udvardi, M. and Wen, J. (2017b) Enabling Reverse Genetics in Medicago truncatula Using High-Throughput Sequencing for Tnt1 Flanking Sequence Recovery. *Methods Mol Biol*, **1610**, 25-37.

Cheng, X., Peng, J., Ma, J., Tang, Y., Chen, R., Mysore, K.S. and Wen, J. (2012) NO APICAL MERISTEM (MtNAM) regulates floral organ identity and lateral organ separation in Medicago truncatula. *New Phytol*, **195**, 71-84.

Cheng, X.F., Wang, M.Y., Lee, H.K., Tadege, M., Ratet, P., Udvardi, M., Mysore, K.S. and Wen, J.Q. (2014) An efficient reverse genetics platform in the model legume Medicago truncatula. *New Phytol*, **201**, 1065-1076.

Courtial, B., Feuerbach, F., Eberhard, S., Rohmer, L., Chiapello, H., Camilleri, C. and Lucas, H. (2000) Tnt1 transposition events are induced by in vitro transformation of Arabidopsis thaliana, and transposed copies integrate into genes. *Molecular Genetics and Genomics*, **265**, 32-42.

Couzigou, J.M., Zhukov, V., Mondy, S., Abu el Heba, G., Cosson, V., Ellis, T.H., Ambrose, M., Wen, J., Tadege, M., Tikhonovich, I., Mysore, K.S., Putterill, J., Hofer, J., Borisov, A.Y. and Ratet, P. (2012) NODULE ROOT and COCHLEATA maintain nodule development and are legume orthologs of Arabidopsis BLADE-ON-PETIOLE genes. *Plant Cell*, **24**, 4498-4510.

d'Erfurth, I., Cosson, V., Eschstruth, A., Lucas, H., Kondorosi, A. and Ratet, P. (2003) Efficient transposition of the Tnt1 tobacco retrotransposon in the model legume Medicago truncatula. *Plant J*, **34**, 95-106.

Doyle, J.J.a.D., J.L. (1990) Isolation of plant DNA from fresh tissue. *Focus*, **12**, 13-15.

Duan, M., Zhang, R., Zhu, F., Zhang, Z., Gou, L., Wen, J., Dong, J. and Wang, T. (2017) A Lipid-Anchored NAC Transcription Factor Is Translocated into the Nucleus and Activates Glyoxalase I Expression during Drought Stress. *Plant Cell*, **29**, 1748-1772.

Fan, M., Xu, C., Xu, K. and Hu, Y. (2012) LATERAL ORGAN BOUNDARIES DOMAIN transcription factors direct callus formation in Arabidopsis regeneration. *Cell Res*, **22**, 1169-1180.

Fukai, E., Umehara, Y., Sato, S., Endo, M., Kouchi, H., Hayashi, M., Stougaard, J. and Hirochika, H. (2010) Derepression of the plant Chromovirus LORE1 induces germline transposition in regenerated plants. *PLoS Genet*, **6**, e1000868.

Gomez-Gomez, L. and Boller, T. (2000) FLS2: an LRR receptor-like kinase involved in the perception of the bacterial elicitor flagellin in Arabidopsis. *Mol Cell*, **5**, 1003-1011.

Gou, J., Debnath, S., Sun, L., Flanagan, A., Tang, Y., Jiang, Q., Wen, J. and Wang, Z.Y. (2018) From model to crop: functional characterization of SPL8 in M. truncatula led to genetic improvement of biomass yield and abiotic stress tolerance in alfalfa. *Plant Biotechnol J*, **16**, 951-962.

Graham, P.H. and Vance, C.P. (2003) Legumes: Importance and constraints to greater use. *Plant Physiol*, **131**, 872-877.

Grandbastien, M.A., Spielmann, A. and Caboche, M. (1989) Tnt1, a mobile retroviral-like transposable element of tobacco isolated by plant cell genetics. *Nature*, **337**, 376-380.

Handberg, K. and Stougaard, J. (1992) Lotus japonicus, an autogamous, diploid legume species for classical and molecular-genetics. *Plant J*, **2**, 487-496.

Haruta, M., Sabat, G., Stecker, K., Minkoff, B.B. and Sussman, M.R. (2014) A peptide hormone and its receptor protein kinase regulate plant cell expansion. *Science*, **343**, 408-411.

Huault, E., Laffont, C., Wen, J., Mysore, K.S., Ratet, P., Duc, G. and Frugier, F. (2014) Local and systemic regulation of plant root system architecture and symbiotic nodulation by a receptor-like kinase. *PLoS Genet*, **10**, e1004891.

Jin, J., Tian, F., Yang, D.C., Meng, Y.Q., Kong, L., Luo, J. and Gao, G. (2017) PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Res*, **45**, D1040-D1045.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J. and Marra, M.A. (2009) Circos: an information aesthetic for comparative genomics. *Genome Res*, **19**, 1639-1645.

Li, J. and Chory, J. (1997) A putative leucine-rich repeat receptor kinase involved in brassinosteroid signal transduction. *Cell*, **90**, 929-938.

Limpens, E., Franken, C., Smit, P., Willemse, J., Bisseling, T. and Geurts, R. (2003) LysM domain receptor kinases regulating rhizobial Nod factor-induced infection. *Science*, **302**, 630-633.

Liu, Y.G., Chen, Y. and Zhang, Q. (2005) Amplification of genomic sequences flanking T-DNA insertions by thermal asymmetric interlaced polymerase chain reaction. *Methods Mol Biol*, **286**, 341-348.

Liu, Y.G., Mitsukawa, N., Oosumi, T. and Whittier, R.F. (1995) Efficient isolation and mapping of Arabidopsis thaliana T-DNA insert junctions by thermal asymmetric interlaced PCR. *Plant J*, **8**, 457-463.

Malolepszy, A., Mun, T., Sandal, N., Gupta, V., Dubin, M., Urbanski, D., Shah, N., Bachmann, A., Fukai, E., Hirakawa, H., Tabata, S., Nadzieja, M., Markmann, K., Su, J., Umehara, Y., Soyano, T., Miyahara, A., Sato, S., Hayashi, M., Stougaard, J. and Andersen, S.U. (2016) The LORE1 insertion mutant resource. *Plant J*, **88**, 306-317.

McHale L, Tan X, Koehl P, Michelmore RW. (2006) Plant NBS-LRR proteins: adaptable guards. *Genome Biol* **7**, 212.

Miyao, A., Tanaka, K., Murata, K., Sawaki, H., Takeda, S., Abe, K., Shinozuka, Y., Onosato, K. and Hirochika, H. (2003) Target site specificity of the Tos17 retrotransposon shows a preference for insertion within genes and against insertion in retrotransposon-rich regions of the genome. *Plant Cell*, **15**, 1771-1780.

Moll, K.M., Zhou, P., Ramaraj, T., Fajardo, D., Devitt, N.P., Sadowsky, M.J., Stupar, R.M., Tiffin, P., Miller, J.R., Young, N.D., Silverstein, K.A.T. and Mudge, J. (2017) Strategies for optimizing BioNano and Dovetail explored through a second reference quality assembly for the legume model, Medicago truncatula. *BMC genomics*, **18**.

Naito, K., Cho, E., Yang, G., Campbell, M.A., Yano, K., Okumoto, Y., Tanisaka, T. and Wessler, S.R. (2006) Dramatic amplification of a rice transposable element during recent domestication. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 17620-17625.

Niederhuth, C.E., Bewick, A.J., Ji, L., Alabady, M.S., Kim, K.D., Li, Q., Rohr, N.A., Rambani, A., Burke, J.M., Udall, J.A., Egesi, C., Schmutz, J., Grimwood, J., Jackson, S.A., Springer, N.M. and Schmitz, R.J. (2016) Widespread natural variation of DNA methylation within angiosperms. *Genome biology*, **17**, 194.

Niu, L., Lin, H., Zhang, F., Watira, T.W., Li, G., Tang, Y., Wen, J., Ratet, P., Mysore, K.S. and Tadege, M. (2015) LOOSE FLOWER, a WUSCHEL-like Homeobox gene, is required for lateral fusion of floral organs in Medicago truncatula. *Plant J*, **81**, 480-492.

Okamoto, H. and Hirochika, H. (2000) Efficient insertion mutagenesis of Arabidopsis by tissue culture-induced activation of the tobacco retrotransposon Tto1. *Plant J*, **23**, 291-304.

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

Piffanelli, P., Droc, G., Mieulet, D., Lanau, N., Bes, M., Bourgeois, E., Rouviere, C., Gavory, F., Cruaud, C., Ghesquiere, A. and Guiderdoni, E. (2007) Large-scale characterization of Tos17 insertion sites in a rice T-DNA mutant library. *Plant Mol Biol*, **65**, 587-601.

Pislariu, C.I., Murray, J.D., Wen, J., Cosson, V., Muni, R.R., Wang, M., Benedito, V.A., Andriankaja, A., Cheng, X., Jerez, I.T., Mondy, S., Zhang, S., Taylor, M.E., Tadege, M., Ratet, P., Mysore, K.S., Chen, R. and Udvardi, M.K. (2012) A Medicago truncatula tobacco retrotransposon insertion mutant collection with defects in nodule development and symbiotic nitrogen fixation. *Plant Physiol*, **159**, 1686-1699.

Riano-Pachon, D.M., Ruzicic, S., Dreyer, I. and Mueller-Roeber, B. (2007) PlnTFDB: an integrative plant transcription factor database. *BMC bioinformatics*, **8**, 42.

Saze, H., Tsugane, K., Kanno, T. and Nishimura, T. (2012) DNA methylation in plants: relationship to small RNAs and histone modifications, and functions in transposon inactivation. *Plant Cell Physiol*, **53**, 766-784.

Sha, Y., Li, S., Pei, Z., Luo, L., Tian, Y. and He, C. (2004) Generation and flanking sequence analysis of a rice T-DNA tagged population. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik*, **108**, 306-314.

Smil, V. (1999) Nitrogen in crop production: An account of global flows. *Global Biogeochemical Cycles*, **13**, 647-662.

Somers, D.A., Samac, D.A. and Olhoft, P.M. (2003) Recent advances in legume transformation. *Plant Physiol*, **131**, 892-899.

Song, W.Y., Wang, G.L., Chen, L.L., Kim, H.S., Pi, L.Y., Holsten, T., Gardner, J., Wang, B., Zhai, W.X., Zhu, L.H., Fauquet, C. and Ronald, P. (1995) A receptor kinase-like protein encoded by the rice disease resistance gene, Xa21. *Science*, **270**, 1804-1806.

Supek, F., Bosnjak, M., Skunca, N. and Smuc, T. (2011) REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One*, **6**, e21800.

Tadege, M., Lin, H., Bedair, M., Berbel, A., Wen, J., Rojas, C.M., Niu, L., Tang, Y., Sumner, L., Ratet, P., McHale, N.A., Madueno, F. and Mysore, K.S. (2011) STENOFOLIA regulates blade outgrowth and leaf vascular patterning in Medicago truncatula and Nicotiana sylvestris. *Plant Cell*, **23**, 2125-2142.

Tadege, M., Ratet, P. and Mysore, K.S. (2005) Insertional mutagenesis: a Swiss Army knife for functional genomics of Medicago truncatula. *Trends Plant Sci*, **10**, 229-235.

Tadege, M., Wen, J., He, J., Tu, H., Kwak, Y., Eschstruth, A., Cayrel, A., Endre, G., Zhao, P.X., Chabaud, M., Ratet, P. and Mysore, K.S. (2008) Large-scale insertional mutagenesis using the Tnt1 retrotransposon in the model legume Medicago truncatula. *Plant J*, **54**, 335-347.

Thimm, O., Blasing, O., Gibon, Y., Nagel, A., Meyer, S., Kruger, P., Selbig, J., Muller, L.A., Rhee, S.Y. and Stitt, M. (2004) MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J*, **37**, 914-939.

Tian, T., Liu, Y., Yan, H., You, Q., Yi, X., Du, Z., Xu, W. and Su, Z. (2017) agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. *Nucleic Acids Res*, **45**, W122-W129.

Uppalapati, S.R., Ishiga, Y., Doraiswamy, V., Bedair, M., Mittal, S., Chen, J., Nakashima, J., Tang, Y., Tadege, M., Ratet, P., Chen, R., Schultheiss, H. and Mysore, K.S. (2012) Loss of abaxial leaf epicuticular wax in Medicago truncatula irg1/palm1 mutants results in reduced spore differentiation of anthracnose and nonhost rust pathogens. *Plant Cell*, **24**, 353-370.

Verdier, J., Zhao, J., Torres-Jerez, I., Ge, S., Liu, C., He, X., Mysore, K.S., Dixon, R.A. and Udvardi, M.K. (2012) MtPAR MYB transcription factor acts as an on switch for proanthocyanidin biosynthesis in Medicago truncatula. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 1766-1771.

Wang, H., Chen, J., Wen, J., Tadege, M., Li, G., Liu, Y., Mysore, K.S., Ratet, P. and Chen, R. (2008) Control of compound leaf development by FLORICAULA/LEAFY ortholog SINGLE LEAFLET1 in Medicago truncatula. *Plant Physiol*, **146**, 1759-1772.

Wang, K., Li, M. and Hakonarson, H. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*, **38**, e164.

Wang, T.L., Domoney, C., Hedley, C.L., Casey, R. and Grusak, M.A. (2003) Can we improve the nutritional quality of legume seeds? *Plant Physiol*, **131**, 886-891.

Xu, C., Cao, H., Xu, E., Zhang, S. and Hu, Y. (2018) Genome-Wide Identification of Arabidopsis LBD29 Target Genes Reveals the Molecular Events behind Auxin-Induced Cell Reprogramming during Callus Formation. *Plant Cell Physiol*, **59**, 744-755.

Yamazaki, M., Tsugawa, H., Miyao, A., Yano, M., Wu, J., Yamamoto, S., Matsumoto, T., Sasaki, T. and Hirochika, H. (2001) The rice retrotransposon Tos17 prefers low-copy-number sequences as integration targets. *Mol Genet Genomics*, **265**, 336-344.

Yang, H. and Wang, K. (2015) Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nature protocols*, **10**, 1556-1566.

Yarce, J.C., Lee, H.K., Tadege, M., Ratet, P. and Mysore, K.S. (2013) Forward genetics screening of Medicago truncatula Tnt1 insertion lines. *Methods Mol Biol*, **1069**, 93-100.

Young, N.D., Debelle, F., Oldroyd, G.E.D., Geurts, R., Cannon, S.B., Udvardi, M.K., Benedito, V.A., Mayer, K.F.X., Gouzy, J., Schoof, H., Van de Peer, Y., Proost, S., Cook, D.R., Meyers, B.C., Spannagl, M., Cheung, F., De Mita, S., Krishnakumar, V., Gundlach, H., Zhou, S.G., Mudge, J., Bharti, A.K., Murray, J.D., Naoumkina, M.A., Rosen, B., Silverstein, K.A.T., Tang, H.B., Rombauts, S., Zhao, P.X., Zhou, P., Barbe, V., Bardou, P., Bechner, M., Bellec, A., Berger, A., Berges, H., Bidwell, S., Bisseling, T., Choisne, N., Couloux, A., Denny, R., Deshpande, S., Dai, X.B., Doyle, J.J., Dudez, A.M., Farmer, A.D., Fouteau, S., Franken, C., Gibelin, C., Gish, J., Goldstein, S., Gonzalez, A.J., Green, P.J., Hallab, A., Hartog, M., Hua, A., Humphray, S.J., Jeong, D.H., Jing, Y., Jocker, A., Kenton, S.M., Kim, D.J., Klee, K., Lai, H.S., Lang, C.T., Lin, S.P., Macmil, S.L., Magdelenat, G., Matthews, L., McCorrison, J., Monaghan, E.L., Mun, J.H., Najar, F.Z., Nicholson, C., Noirot, C., O'Bleness, M., Paule, C.R., Poulain, J., Prion, F., Qin, B.F., Qu, C.M., Retzel, E.F., Riddle, C., Sallet, E., Samain, S., Samson, N., Sanders, I., Saurat, O., Scarpelli, C., Schiex, T., Segurens, B., Severin, A.J., Sherrier, D.J., Shi, R.H., Sims, S., Singer, S.R., Sinharoy, S., Sterck, L., Viollet, A., Wang, B.B., Wang, K.Q., Wang, M.Y., Wang, X.H., Warfsmann, J., Weissenbach, J., White, D.D., White, J.D., Wiley, G.B., Wincker, P., Xing, Y.B., Yang, L.M., Yao, Z.Y., Ying, F., Zhai, J.X., Zhou, L.P., Zuber, A., Denarie, J., Dixon, R.A., May, G.D., Schwartz, D.C., Rogers, J., Quetier, F., Town, C.D. and Roe, B.A. (2011) The Medicago genome provides insight into the evolution of rhizobial symbioses. *Nature*, **480**, 520-524.

Zhou, C., Han, L., Hou, C., Metelli, A., Qi, L., Tadege, M., Mysore, K.S. and Wang, Z.Y. (2011) Developmental analysis of a Medicago truncatula smooth leaf margin1 mutant reveals context-dependent effects on compound leaf development. *Plant Cell*, **23**, 2106-2124.

Zhou, C., Han, L., Li, G., Chai, M., Fu, C., Cheng, X., Wen, J., Tang, Y. and Wang, Z.Y. (2014) STM/BP-Like KNOXI Is Uncoupled from ARP in the Regulation of Compound Leaf Development in Medicago truncatula. *Plant Cell*, **26**, 1464-1479.

Legends

Figure 1. A. Distribution of *Tnt1* insertion and GC content. The outer layer represents the *Tnt1* insertion, and the inner layer represents GC content distribution. ( ) Indicates centromere location on each chromosome. B. Correlation of *Tnt1* insertion frequency and GC content in the *M. truncatula* genome. *Tnt1* insertion frequency and GC content were calculated in 500 kb window size.

Figure 2. *Tnt1* insertion frequency is positively correlated with chromosome length in *M. truncatula*. Total *Tnt1* insertion number in each chromosome was plotted against the length (kb) of chromosomes.

Figure 3. Systematic analysis of inserted genes (A). *Tnt1* insertions in the *M. truncatula* genome. The blue bar in panel A shows the average insertion per gene is 3.43. (B) The significantly affected pathways by *Tnt1* insertions. (C). Biological processes that are more frequently inserted. (D). Molecular functions that are more frequently inserted.

Figure 4. *Tnt1* insertion and methylation frequency and gene expression in *M. truncatula*. A. Line plot of *Tnt1* insertion and methylation frequencies in different groups of genes with at least four insertions.

B. *Tnt1* insertion frequency is positively correlated with methylation frequency. C. Correlation between *Tnt1* insertion frequency and gene expression level of all inserted genes.

Figure 5. *Tnt1* insertion and expression levels and GC contents of genes with at least four insertions in *M. truncatula*. A. Correlation of *Tnt1* insertion and average gene expression level in different gene groups. B. Correlation of *Tnt1* insertion and average GC content in different gene groups.

Figure 6. Important gene groups inserted by *Tnt1*. The number (A) and frequency (B) of *Tnt1* insertions in transcription factors, receptor kinases and NBS-LRR genes. (C) The distribution of different families of transcription factors in *M. truncatula* and *Tnt1* insertion numbers in each family.

Table 1. *Tnt1* insertion distribution on the *M. truncatula* genome

Table 2. *Tnt1* insertion distribution in *M. truncatula* genome

Table 3. Comparison of FST recovery by different approaches.

Table 1. *Tnt1* insertion distribution on the *M. truncatula* genome

	Chr1	**Chr2**	**Chr3**	**Chr4**	**Chr5**	**Chr6**	**Chr7**	**Chr8**	**Total**
***Tnt1* insertions**	24661	21100	24186	26540	21834	11929	21010	19639	170899
Chromosome length (kb)	52991	45730	55515	56582	43631	35276	49172	45570	384467
Frequency (*Tnt1*/kb)	0.465381	0.461404	0.435666	0.469054	0.500424	0.338162	0.427276	0.430963	0.444509

Table 2. *Tnt1* insertion distribution in different regions of the *M. truncatula* genome

Insertion Site	Number of FSTs	Percentage
exon	58078	33.2
splicing[a]	512	0.3
UTR5[b]	3786	2.2
UTR3[c]	4829	2.7
intron	30890	17.7
upstream	17138	9.8
downstream	17991	10.3
intergenic	41322	23.7

[a] Insertion is within 2bp of a splicing junction. This only refers to the 2bp in the intron that is close to an exon.
[b] 5' untranslated region
[c] 3' untranslated region

Table 3. Comparison of FST recovery by different approaches.

Line ID	# of FSTs by TAIL-PCR[a]	# of FSTs by *Tnt1*-capture	# of FSTs by WGS	% recovered by TAIL	Source
NF0054	29 (25 S)		65	44.6	Jiang et al. 2015
NF0365	16 (16 S)	255		6.7	This work
NF0707	53 (16 S)	145		36.5	This work
NF1962	16	164	164	9.8	This work
NF2085	20 (20 S)	44		45.5	This work
NF2258	29	105		27.6	This work
NF2801	5		51	9.8	This work
NF8208	13	32		40.6	This work
NF10547	10		65	15.4	Veerappan et al. 2016
NF10739	41	96		42.7	This work
NF11044	7	17	17	41.2	This work
NF11217	37		97	38.1	Veerappan et al. 2016
NF11579	30	119		25.2	This work
NF13918	10	19		52.6	This work
NF20764	94		136	69.1	This work
Average	**27.4**	**99.6**	**85**	**33.7**	

[a] The numbers in parenthesis indicate the numbers of FSTs recovered by Sanger (S) sequencing