10-26-2021 11:30 AM

Developing a toolbox to engineer quantitative trait variation in legume species using CRISPR/Cas technologies.

Petar Miletic, *The University of Western Ontario*

Supervisor: Szczyglowski, Krzysztof, *Agriculture and Agri-Food Canada (AAFC-AAC)*
Co-Supervisor: Grbic, Vojislava, *The University of Western Ontario*
A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Biology

Abstract

The impact of intensive agriculture on the environment is immense. This is especially dire with regard to the natural nitrogen (N) cycle, where the human driven interference, primarily associated with industrial fertilization, has reached unsustainable levels. Unlike cereals, legumes, such as soybean, alfalfa and common bean, have the ability to use atmospheric nitrogen, which limits the need for industrial fertilization. A more wide-spread use of legumes could alleviate some of the negative impacts on the biogeochemical cycle while also providing a useful alternative to meat consumption, an important factor in sustainability. To reach this goal, further improvements of legume crops with regard to their nitrogen economy and yield are essential. In the model legume, *Lotus japonicus*, the HAR1 receptor kinase plays a central role in the plant nitrogen and phosphate nutrition by regulating beneficial symbioses and root system architecture. In this thesis study, I used the *HAR1* locus as a paradigm for the development of a CRISPR/Cas-based toolkit, with the ultimate goal of generating a range of synthetic variation at agriculturally important traits. While genome modifications at the *HAR1* locus are yet to be demonstrated, the toolbox required to perform these experiments was developed. It should facilitate rapid expansion on the repertoire of alleles available for accelerated breeding of new, high yielding legume varieties that are better attuned with the natural environment.

Keywords: Crop nutrition, nitrogen, legumes, CRISPR/Cas, *Lotus japonicus*, rhizobia, Hypernodulation Aberrant Root Formation 1 (HAR1), hairy roots, engineered trait variation

Summary for lay audience

There is an urgent need to improve crop yields while simultaneously reducing the use of nitrogen fertilizers. One way to address this problem is to turn to legumes, which can use atmospheric nitrogen. They achieve this by forming a symbiotic relationship with soil bacteria known as rhizobia. However, legumes are currently not able to produce as much grain as cereals do. Breeders can use natural variation to improve the yields of legumes, but the levels of natural variation that is available is relatively small. Using a modern gene editing technology, called CRISPR, it is possible to expand on the existing variation. This expansion allows crop breeders to identify beneficial variants much quicker than they could using traditional methods. Using CRISPR to expand on the variation in genes related to nitrogen uptake, for example, will help reduce the need for industrial nitrogen inputs.

Improving legumes' yields requires the knowledge of gene interactions which regulate nitrogen use. In this thesis study, gene editing tools were developed in order to expand on the natural variation that exist in the model legume, *Lotus japonicus*. A gene, called HAR1, which regulates the plant nitrogen economy was chosen as a paradigm. The gene editing toolbox, which can be applied to any agriculturally relevant plant function has been developed and testing these new tools is in progress. If successful, the knowledge gained by working in the model plant organism will be transferable to crop plants. It should facilitate accelerated breeding of new, high yielding legume varieties that are better attuned with the natural environment.

Acknowledgments

First off, I would like to thank my supervisor, Dr. Krzysztof Szczyglowski, for his support throughout my entire graduate experience. He provided a lab environment where failures weren't scorned, rather they were seen as opportunities to learn. This attitude towards mentoring has primed me for success in my future endeavours. Dr. Szczyglowski was always available to me and provided me with excellent suggestions and invaluable insights. His expertise and guidance have greatly improved my abilities to perform scientific inquiry and communicate my findings. Dr. Szczyglowski exhibited great patience whenever I struggled, and he used these opportunities to teach me to never give up. Thank you Dr. Szczyglowski.

I would like to extend special gratitude to my co-supervisor, Dr. Vojislava Grbic. Her advice and support were an immense help in the completion of this thesis. I would also like to extend gratitude to my advisory committee members, Dr. Yuhai Cui and Dr. Marc-Andre LaChance. Your mentorship and guidance have helped me discover problems, and solutions, which I did not have the foresight to anticipate.

Next, I would like to thank Dr. Szczyglowski's laboratory technicians, Loretta Ross and Terry Huebert. When I first started in Dr. Szczyglowski's lab, Loretta taught me the standard protocols utilized by our team. If I ever had questions regarding protocols, safety, anything thesis related, or just wanted to chat, Loretta and Terry were always available. They have helped me analyze results, troubleshoot problems, and suggested solutions when I was stuck. They have both been a true pleasure to work with. I am also thankful to Angelo Kaldis, who was always available to train me in the safe use of equipment and machines in the London Research and Development Centre.

Thank you to my lab mates Jasmine, Arina, Mark, and Laurel. Discussing our projects and *Lotus* biology has been extremely fun. I would also like to thank my colleagues Ramtin, Alexandria, and Nishat. The days we spent working and laughing together have been amazing. I want to thank you all for being such great friends!

I want to thank my Parents, Aleksandra and Dusan Miletic, for their never ending love and support my whole life. I would not be where I am today without the two of you. I also want to thank my girlfriend, Camille Thierry, for being there for me through this whole journey. Without your support I would not have been able to complete this work.

Finally, I would like to extend a special thank you to Dr. Duanmu Deqiang for generously providing the binary vectors used in the Dual Promoter system described in this thesis.

Table of Contents

List of Tables

List of Figures

List of Appendices

Chapter 1: Introduction

1. The natural nitrogen cycle

Nitrogen is most abundant in its inert, dinitrogen (N_2) state, which is the main constituent of the Earth's atmosphere, accounting for approximately 78% of its composition (Figure 2). N_2 is converted by both biological nitrogen fixation, a prokaryotic (archaea and bacteria) process and, to a much lesser extent, by physical processes (e.g. atmospheric fixation by lightning) into reactive molecules, ammonia and nitrogen oxides, respectively (Noxon, 1976; Bruijn, 2014). Ammonia is rapidly converted by nitrifying bacteria, to nitrites (NO_2^-) and nitrates (NO_3^-). These bioactive nitrogen-containing molecules can readily be utilized by plants and indirectly used by animals, including humans (Figure 2). Various decomposers, such as soil bacteria and fungi, add to the global nitrogen cycle, by degrading nitrogen-containing organic molecules, such as proteins, amino acids and chlorophyll, hence further contributing nitrogen to the environment in the form of ammonia. In contrast, denitrifying bacteria reduce an excess of nitrites and nitrates to gaseous forms, primarily N_2, which returns to atmosphere, closing the nitrogen cycle (Figure 2). As further discussed below, industrial fertilization has greatly disturbed this natural cycle.

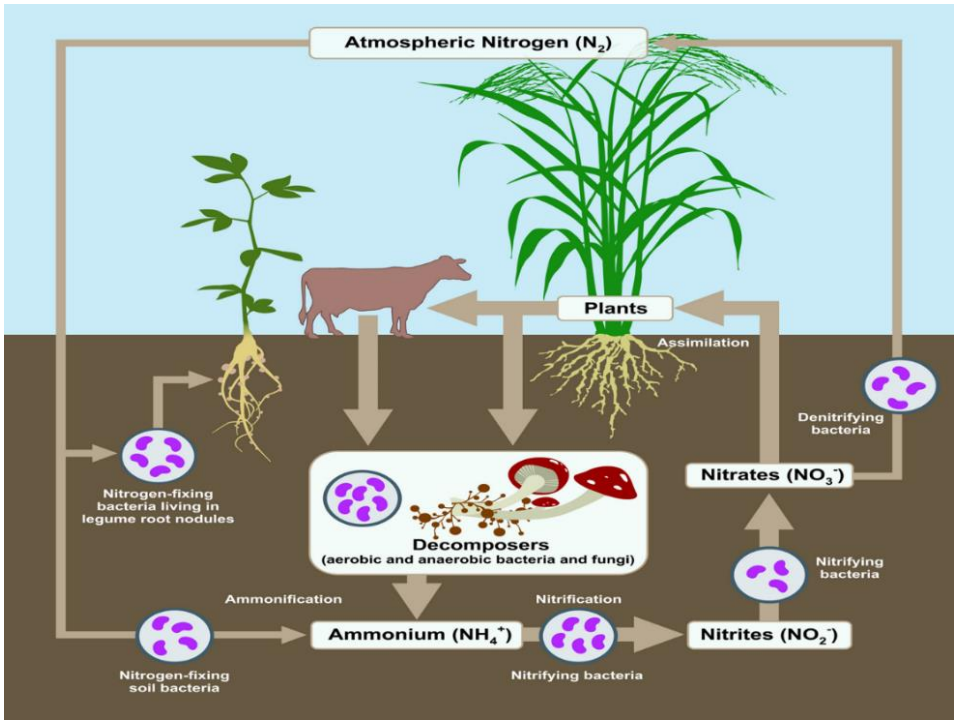


Figure 1. The natural nitrogen cycle is depicted. For further details, see text. The schematic has been modified from https://en.wikipedia.org/wiki/Nitrogen_cycle.

2. The Green Revolution and how it led to the nitrogen problem.

Enhancing crop yields ensured food security for a rapidly growing human population. As technology improved, the techniques to maximize yield became more sophisticated. Perhaps the biggest step forward in the productivity of crops was a result of the Green Revolution, which started in the 1960's (Hedden, 2003). This involved breeding improved, semi-dwarf varieties of rice and wheat, often referred to as Green Revolution Varieties (GRVs), which were more resistant to pests and lodging and were very productive (Peng et. al., 1999). However, industrial fertilization was required to maximize the yield of the GRVs (Egli, 2008). In fact, the use of chemical

fertilizers, along with pesticides, improved the crop yields so drastically that it played a major role in driving the exponential growth of the human population (Figure 1).

The use of nitrogen fertilizer was made possible through the invention of a process that allowed industrial fixation of atmospheric N_2 into ammonia, known as the Haber-Bosch process (Haber, 1920). This process combines N_2 and hydrogen gasses under extremely high heat (about $500^{O}C$) and pressure (about 200 atmospheres) along with an iron catalyst, to produce liquid ammonia (Haber, 1920). The conditions required for the chemical synthesis of ammonia to proceed efficiently can only be achieved through the use of large quantities of energy that are produced by burning fossil fuels, mainly natural gas, which generates copious amounts of greenhouse gasses (Vicente & Dean, 2017). Ammonium derived from the Haber-Bosch process now accounts for 60% of the fertilizers used globally, which has resulted in a massive impact on the nitrogen cycle and several other planetary conditions, including biodiversity and global warming (Wang et. al., 2018).

The natural nitrogen cycle, long self-sustainable and balanced, has been significantly disrupted by anthropogenic (i.e. human-driven) addition of combined nitrogen, primarily associated with agricultural production, causing significant nitrogen enrichment in the environment, the consequences of which are disastrous (Smith & Schindler, 2009; see also, https://e360.yale.edu/features/the-nitrogen-problem-why-global-warming-is-making-it-worse). In fact, as of 2010, agricultural practices accounted for the addition of approximately 207 million metric tons of bioactive nitrogen to the environment, as compared to approximately 111 million metric tons produced by natural processes (Canfield et. al., 2010). Any excess of bioavailable nitrogen runs off into the water systems that surround the application areas. Increased nitrogen concentrations in the water, along with phosphate, allows for some algal species to thrive, resulting

in harmful algal blooms (Sellner et. al., 2003; https://www.theguardian.com/environment/2020/jan/04/lethal-algae-blooms-an-ecosystem-out-of-balance). These blooms consume the available oxygen resulting in the suffocation of the ecosystems, a process known as eutrophication (Smith & Schindler, 2009). As of 2018, every continent, including Antarctica, had coastal zones of eutrophication (Figure 3). Not only are fertilizers directly damaging to water and soil, but their production and use are also a major source of greenhouse gas emissions, including nitrous oxide, an ozone-depleting gas that has a potential to be 300 times more influential on climate change compared to carbon dioxide (Woods et. al., 2010). The destruction of aquatic ecosystems through eutrophication along with the impact of N fertilization on global warming are causing irreparable changes to global ecosystems. Therefore, an immediate action is needed to halt these negative trends (Smith & Schindler, 2009).

Figure 2. Coastal areas marked by red dots represent locations where oxygen concentrations [O_2] have declined to hypoxic conditions (≤63 mmol O_2 per liter). This decline was either exacerbated or caused by anthropogenic nutrient runoffs. Oxygen starved environments suffer from loss of biodiversity. The image was modified from Breitburg, D. et. al. (2018).

3. The nitrogen bottleneck in green revolution varieties.

The use of nitrogen fertilizers was more of a double-edged sword than anyone could have predicted in the 1960's and 1970's, when its application became widespread. The benefits of nitrogenous fertilizer use were clear. Much greater yields could be achieved in the same amount of space if fertilizer was applied as compared to no fertilization (Egli, 2008). There was however a massive oversight with respect to important drawbacks that accompanied the use of fertilizers. The environmental issues that arose due to the excessive use of fertilizer, such as eutrophication of coastal areas (Figure 3), are now being widely discussed (Breitburg et. al., 2018). Moreover, new molecular data showed that the high yielding GRVs of rice and wheat suffered from diminished nitrogen use efficiency (NUEs; Wu et. al., 2020). This mainly resulted from the selection of partial dwarfism that prevented crop logging while enhancing the yield (Hedden, 2003). However, the underlying alleles that were selected for during the breeding process also resulted in increased accumulation of the DELLA protein, which was shown to negatively regulate the plant nitrogen economy (Wu et. al., 2020).

DELLAs are a family of GRAS (GIBBERELIC ACID INSENSITIVE REPRESSOR OF *ga1-3* SCARECROW) transcriptional regulators that are plant specific (Yoshida et. al., 2014). They act as master regulators that counteract growth promoting gibberellins, hence limiting vegetative plant growth (Wu et. al., 2020). The DELLA proteins are normally degraded by the activity of gibberellins (Peng et. al., 1999). GRVs that had increased levels of the DELLA protein were found to have gibberellins that were either present in reduced quantities or were less effective at degrading the mutant DELLA protein compared to their parental lines (Wu et. al., 2020). The resulting phenotypes of the crops were semi-dwarf plants, which had

an immense production capacity but required high levels of nitrogen input to achieve maximum yield (Li et. al., 2018).

In an interesting twist, DELLA proteins were found to be responsible for decreasing NUE of the GRVs by counteracting the function of the GROWTH-REGULATING FACTOR 4 (GRF4), a transcriptional regulator that promotes nitrogen uptake and metabolism (Li et. al. 2018). A progressive selection for varieties that contained greater and greater concentrations of DELLA proteins led to a bottleneck in the crops' abilities to uptake and assimilate nitrogen. This is the main reason why the increased fertilization of GRVs was required to reach top yield. Considering that close to 50% of the applied nitrogen globally is being lost through runoff and volatilization (Vidal et. al., 2020), the invention and global use of highly productive GRVs came at a significant price to the environment.

Given the hugely detrimental effect of reactive nitrogen, minimizing the use of industrial fertilizers is essential. However, stopping the use of fertilizers abruptly would lead to serious food shortages (Crews & Peoples, 2003). Avenues that could close the gap between low nitrogen inputs and high crop yields are being explored. Better understanding and usage of natural mechanisms constitutes one path toward reaching this goal. New genomic resources and novel approaches, such as the rapidly emerging field of synthetic biology, are also expected to have a significant impact in this context.

4. Plants have accumulated various adaptations to deal with nitrogen limitations.

Low levels of bioavailable nitrogen have been common for most of life's history on the planet. In fact, nitrogen and phosphorus availability are the primary limiting factors to plant growth following the availability of water (Agren et. al., 2012). Thus, ancient relatives of modern plants

had to adapt to nutrient scarcity over the course of their evolutionary histories. The earliest plants that emerged on land were most likely bryophytes that did not have any roots and grew by spreading along surfaces (Kenrick & Crane, 1997; Brundrett, 2002; Morris et. al., 2018). Modern bryophytes include mosses, liverworts, and hornworts. According to the fossil record the relationship between modern mosses and vesicular-arbuscular mycorrhizas (VAM) is incredibly similar to that of early plants colonized by VAM hyphae (Brundrett, 2002). This symbiosis is likely to have been instrumental for the terrestrial colonization by ancient plants as it allowed the rootless species to combat desiccation and utilize the phosphate (Helgason & Fitter, 2005), and possibly nitrogen (Buckling & Kafle, 2015) that was trapped in the soil using fungal hyphae (Pirozinski and Malloch, 1975).

Roots evolved later in the history of land plants, following the emergence of plant vasculature (Kenrick & Crane, 1997; Kenrick & Strullu-Derrien, 2014). It is likely that roots originated as a result of subterranean stems growing from early vascular plants (Brundrett, 2002). High levels of root morphological divergence first emerged during an era of rapid plant diversification in the mid-Devonian (Brundrett, 2002). They eventually also gained phenotypic plasticity, which is the ability to modify their growth patterns in response to changing soil conditions, including the availability of nutrients (Hodge, 2004). Morphological variation in response to nitrogen supply is an excellent example of such developmental root plasticity (Forde, 2014).

The plasticity of extant plants' root systems governs their ability to forage for nitrogen and other nutrients, an important trait especially in natural environments (Schneider and Lynch, 2020). Roots will have increased length and overall biomass as a result of low nitrogen availability, however, their growth is stunted if they are severely deprived of nitrogen (Araya et. al., 2014; Kiba

& Krapp, 2016). These and other growth patterns are mediated by local and systemic regulatory mechanisms that integrate and convert external and internal information into cohesive root responses that maximize the capture and utilization of limited nutrients, including nitrogen (Wang et al., 2018).

After root plasticity, the root associated microbial community is the most important adaptation plants have when it comes to scavenging for nutrients. Transfer of nutrients between microorganism and plants is considered an ancient phenomenon that was already in existence before plants invaded the land and also as exemplified by the ancient VAM symbiosis (Taylor et al., 1995; Strullu-Derrien et. al., 2018). However, plants associate with a wide range of microorganisms, including highly diverged fungal and bacterial species, which are collectively known as plant microbiomes (Fitzpatrick et. al., 2018; Trivedi et al., 2020). As with the human gut microbiome, the root microbiomes have been shown to benefit the host with respect to several key functions, such as protection against pathogens (Mousa et al., 2016) and the already mentioned nutrition. An interesting example of the latter has been described for two varieties of rice grown in Asia, namely *Indica* and *Japonica*. Species belonging to the *Indica* variety of rice recruit microbiomes that are more diverse and characterized by enhanced nitrogen assimilation and metabolism functions as compared to those present in the microbiomes of *Japonica* rice species. This results in better NUE and growth of the *Indica* species (Zhang et. al., 2019). Interestingly, a sequence variation at the rice *NRT1.1B* gene, encoding a nitrate transceptor (i.e. nitrate sensor and transporter) was shown to be mainly responsible for the microbiome variation between the two rice varieties, underscoring the importance of the host genome in the recruitment of a beneficial microbial community (Zhang et. al., 2019).

Much like the plants themselves, the root microbiomes are very susceptible to fluctuations in soil nitrogen content. Among many edaphic factors, elevated levels of anthropogenic N and P were shown to influence taxonomic and functional traits of soil microbial communities (Leff et. al., 2015; Pan et al., 2014). Furthermore, a long-term exposure to industrial fertilization has also been indirectly linked to favoring less effective microbial mutualists (Johnson, 1993; for current review see Vandenkoornhuyse et. al. 2015). However, engineered root microbiomes that consist of species which maintain active nitrogen assimilation and metabolism in the presence of exogenous nitrogen inputs may improve plant productivity (Bloch et. al., 2020). Therefore, maximizing benefits as provided by plant microbiomes while limiting anthropogenic inputs will likely constitute an essential avenue in improving NUE and saving the environment.

In the same context, legume plants such as garden pea, soybean, and alfalfa, have the ability to form endo-symbiotic relationships with a subgroup of nitrogen-fixing bacteria commonly known as rhizobia. This intimate relationship is restricted in nature (Griesmann et. al., 2018) and is also unique, as rhizobia reside inside the living root cells (Parniske, 2018), where they facilitate the assimilation of atmospheric nitrogen by the host legume in a process called symbiotic nitrogen fixation (SNF) (Figure 2). To be able to fix atmospheric N_2 rhizobia require a hypoxic environment because the N_2 fixing enzyme, nitrogenase, is extremely sensitive to oxygen (Rees & Howard, 2000). Once engaged in symbiosis, legumes develop specialized root derived organs, called root nodules, and produce an oxygen carrier molecule, leghemoglobin. Root nodules combined with leghemoglobin facilitate the hypoxic environment necessary for rhizobial nitrogen fixation (Ott et. al., 2005). Engaging in this mutualistic relationship reduces or entirely eliminates the need for soil nitrogen (Franche et. al., 2009; Udvardi and Poole, 2013). Hence, the symbiosis and the pertinent

regulatory mechanisms have been of great interest and are considered to be crucial in minimizing the requirement for nitrogen fertilization globally (Bailey-Serres et. al., 2019).

5. Symbiotic nitrogen fixation is likely to play an important role in the development of sustainable agriculture.

Although most plant microbiomes contain bacterial species that can facilitate nitrogen nutrition of their host plants, none can do so as efficiently as the endo-symbiotic rhizobia. The legume-*Rhizobium* relationship, where the bacteria are hosted intracellularly in symbiotic organs, called root nodules, provides the most sophisticated and direct route for nitrogen nutrition of the host that is independent from soil nitrogen. The symbiotic relationship starts with a chemical cross-talk between the partners that initiates the infection process at the root epidermis (Crespi & Galvez, 2000). Concomitantly, a signaling cascade to the subtending root cortex stimulates nodule organogenesis (Soyano et. al., 2013). Nodules host rhizobia intracellularly and provide hypoxic conditions (Ott et. al., 2005), allowing the bacterial nitrogenase enzyme to reduce N_2 to ammonia. The ammonia is secreted to and assimilated by plant cells, which supports plant growth (Figure 4). In return, the host supplies its symbiotic partner with photosynthetic carbon (Vance, 2008). In this relationship, the host plant is in control and will restrict the number of nodules which are formed to maintain homeostasis (Reid et. al., 2015). This so called Autoregulation of Nodulation (AON), represents an important systemic, root-to-shoot-to-root, regulatory feed-back mechanism (Magori et. al., 2009; Ferguson et. al., 2010; Suzaki et. al., 2015). AON serves to preserve the balance between the need for nitrogen and the extent to which supporting the symbiosis through the supplementation of photosynthates is still beneficial to the host (Oka-Kira and Kawaguchi, 2006).

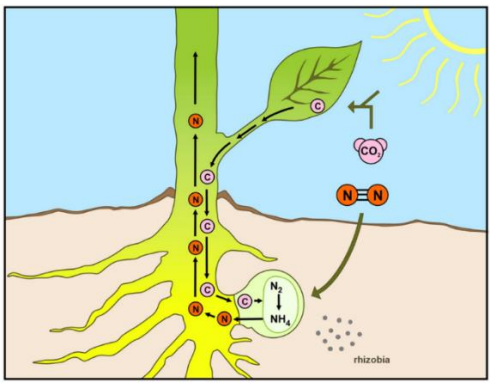


Figure 3. Schematic depicting the symbiotic relationship between legume plants and rhizobia that facilitates biological nitrogen fixation in legume species. Root nodules are formed as a response to rhizobial signaling and provide a hypoxic environment that allows the bacterial enzymatic complex, the nitrogenase, to convert atmospheric nitrogen to ammonium (N_2 to NH_4^+). The rhizobia reside in root nodules, where nitrogen is fixed and supplied to the host legume in exchange for photosynthates.

Different legumes form relationships with different species of rhizobia. The model legume, *Lotus japonicus* (Handberg & Stougaard, 1992; Szczyglowski & Stougaard, 2008), the subject of my thesis work, forms root nodules to accommodate its nitrogen-fixing symbiont, *Mesorhizobium loti* (Kaneko et. al., 2000). The extent of nodule formation in *L. japonicus* is regulated by a central gene in the AON pathway, called *HYPERNODULATION ABBERANT ROOT FORMATION 1* (*HAR1*) (Wopereis et. al., 2001; Krusell et al., 2002; Nishimura et al., 2002). *HAR1* encodes a leucine-rich repeat receptor kinase (LRR-RK) that is likely orthologous to the *A. thaliana CLAVATA1* LRR-RK (Okamoto & Kawaguchi, 2015). The activity of *HAR1* prevents excessive nodule formation by generating a systemic, root-shoot-root feedback signaling mechanism that restricts subsequent nodulation events. The initial nodule formation events trigger expression of the CLV3/ESR-related (CLE) peptides in *L. japonicus* roots. These are translocated to the shoot

and are perceived by HAR1, which results in a shoot-to-root signaling cascade (Okamoto et. al., 2013). This cascade involves the microRNA miR2111 (Tsikou et. al., 2018) and a Kelch repeat-containing F-box protein called TOO MUCH LOVE (TML) (Takahara et. al., 2013), which effectively inhibit the formation of new nodules. In the absence of HAR1, as in the *har1-1* mutant background, the root develops an excessive number of nodules, a phenomenon that is referred to as hypernodulation. This significantly restricts the root growth and also impedes the shoot development (Wopereis et al., 2001), demonstrating that maintenance of the symbiotic homeostasis is vital for the host.

Interestingly, Wopereis et. al. (2001) showed that *HAR1* was not only involved in regulating nodule formation during symbiosis with *M. loti* but was also responsible for controlling the root system architecture, by balancing root elongation and lateral root formation. Other research has shown that *HAR1* and its predicted orthologues in different legume species regulate nitrogen uptake and metabolism (Lagunas et. al., 2019). A phosphate acquiring root-mycorrhiza fungi symbiosis was also shown to be regulated by *NODULE AUTOREGULATORY RECEPROR KINASE* (*NARK*), the soybean orthologue of *HAR1* (Schaarschmidt et al., 2013). The functional attributes of *HAR1* and its orthologues in different legume crop species have made this locus an attractive subject of research that aims to further our basic understanding of plant growth regulation in the context of N and P nutrition. The same locus has also become a potential target for breeding approaches directed toward improving nutrient acquisition in legume crop plants.

6. Employing breeding strategies to improve plant nutrition to save the environment.

Domestication and breeding have significantly reduced the pool of available diversity that remains among elite crops varieties (Smale, 1997; Shi & Lai, 2015). These dwindled pools of

variation are now creating significant challenges in furthering crop improvements through the use of traditional breeding methods. Finding new, useful alleles has become difficult, and for some agronomic traits beneficial alleles may not exist, either because they did not arise naturally or were lost as a result of domestication (Scheben & Edwards, 2018). The genetic variations (i.e. new alleles) that exist in wild relatives and land races can be used for further improvement of elite varieties. However, these alleles need to be introgressed and the elite genetic background must be re-constituted by sequential backcrosses to the original elite genotype. This process is time consuming, and a positive outcome is uncertain (Tanksley &Nelson, 1996). Furthermore, improvements to agronomically important traits must be made more rapidly, given that the exponentially growing human population must be fed without causing further destruction to the environment (Scheben & Edwards, 2018).

A new breeding method, which is based on targeted genome modifications, presents a massive shortcut to the traditional method. Using genome editing, it is possible to make surgical changes or create a continuum of trait variation, hence significantly enriching the breeding process (Rodriguez-Leal et. al., 2017). Expanding on natural variation in regulatory sequences, such as gene promoters underlying important agronomic traits, is promising for rapid crop improvement (Scheben & Edwards, 2018). This expansion of natural variation is effectively generating what can be called synthetic variation. While several methods have been used to induce directed changes to plant genomes, including zinc finger nucleases (ZFNs) (Weinthal et. al., 2010) and transcription activator-like effector nucleases (TALENs) (Chen & Gao, 2013), the recently discovered type II Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)-associated (Cas) enzyme systems (Shan et. al., 2013; Koonin & Makarova, 2019) provide the most direct route for crop improvements.

7. CRISPR and the use of CRISPR-based genome editing to expand on the natural variation that exists in agriculturally important species.

CRISPR/Cas systems arose in bacteria and archaea as an adaptation against phages (Jinek et. al., 2012; Koonin & Makarova, 2019). They have the capability to create double-stranded breaks at very precise locations in genomic DNA. These systems require two components to function, a Cas protein and crisprRNA:tracrRNA heteroduplex. The Cas protein is an enzyme that contains two nuclease domains, RuvC and HNH (Shan et. al., 2013). The RNA heteroduplex consists of short RNA molecules, the crisprRNA, also known as spacer or the guide RNA (gRNA) and the trans-activating crisprRNA (tracrRNA). The gRNA contains sequence that is complementary to target sites in phages or plasmid DNAs, while the tracrRNA acts as a handle that is needed to activate gRNA/Cas protein based DNA cleavage (Doudna & Charpentier, 2014). The gRNA and tracrRNA get processed together in the microbial cells to generate a gRNA:tracrRNA heteroduplex that can associate with the Cas protein (Jinek et. al., 2012). The Cas protein then scans DNA, both native and foreign, for target sites as guided by any given gRNA. Once the gRNA forms Watson-Crick base pairing with the target DNA strand, the Cas enzyme will induce a double stranded break at the site. However, the double stranded cleavage will only occur if the so called protospacer adjacent motif (PAM) site, a short sequence that is recognized by the Cas protein, is present at the 3' end of the target site (Figure 5). These double-stranded breaks neutralize the invading phage or plasmid DNAs (Shan et. al., 2013). The requirement for a PAM sequence at the 3' end of the target sites (i.e. protospacers) prevents the cleavage of the native DNA, thus avoiding damage to the host genome.

CRISPR/Cas systems arose in many different microbial species and some have been adapted for the purpose of genome editing in eukaryotes. A chimeric RNA molecule that contains

both the gRNA and tracrRNA sequences, known as a single guide RNA (sgRNA), was developed (Jinek et. al., 2012). Next, Cas proteins were codon optimized for use in eukaryotes (Shan et. al., 2013). The Cas9 enzyme from *Streptococcus pyogenes* and Cas12a from *Lachnospiraceae bacterium* were shown to be effective for use in eukaryotic cells (Shan et. al., 2013; Ha et. al., 2020). Together these two components allow for the generation of mutations at highly specific sites anywhere in eukaryotic genomes where PAMs are present. These CRIPSR/Cas systems for genome editing are more efficient at generating site specific mutations and are less costly than ZFNs and TALENs (Doudna & Charpentier, 2014).

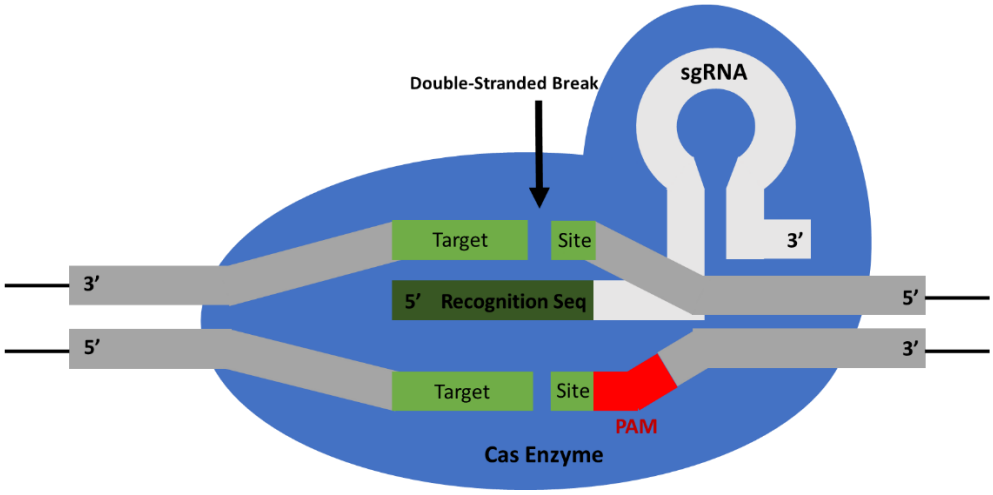


Figure 4. A schematic showing a CRISPR associated (Cas) enzyme and a single guide RNA (sgRNA) associated with each other. A Cas enzyme scans the genomic DNA for sequences complementary to the recognition sequence of the sgRNA. The genomic target site must contain a protospacer adjacent motif (PAM) at its 3'end, and each Cas enzyme has a unique PAM sequence. The single-stranded recognition sequence of the sgRNA base pairs with the recognition site in the genomic DNA, and the two magnesium-dependent nuclease domains present in the Cas enzyme cleave the genomic DNA. The double stranded break is subject to endogenous cellular repair mechanisms which often generate mutations at the target sites.

CRISPR/Cas based systems have been used to conduct different types of genetic studies in model and crop plant species (Wang et. al., 2016; Ji et. al., 2019). There is also evidence to suggest that precise gene knock-in or replacement experiments are possible in plants, but these technologies are still being developed (Lu et. al., 2020). However, one of the most exciting avenues that CRIPSR/Cas systems provide in the study of crop improvement is the possibility of generating a range of alleles that result in quantitative trait variations. Thus, rather than studying gene function alone, this method provides the opportunity to generate synthetic variation that can be used by

breeders when trying to improve agronomic traits. For example, by using a set of sgRNAs targeting the promoter region of the *Solanum lycopersicum CLAVATA3* gene, a range of variants with differing fruit size and seed number was generated (Rodriguez-Leal et. al., 2017). Similar approaches should guide the development of synthetic variation at many agriculturally relevant loci directly in elite crop varieties, hence aiding the necessary, rapid improvement in crop productivity and their environmental performance.

8. Changing agricultural practices and improving crops for sustainable food production.

Different changes can be envisaged to current food production systems in order to reduce their impacts on the environment. Further improvements in the management of food production are going to be crucial moving forward, but they will likely not be sufficient to reduce the environmental impacts caused by agriculture (Gomiero et. al., 2011). Shifting to a more plant based diet globally is predicted to help in tackling several pressing issues, including climate change and biodiversity. This is because of the disproportionately negative impact that animal production has on biodiversity, land use and the environment (Sakadevan & Nguyen, 2017). However, the necessary increase in plant production has its own challenges, primarily associated with excessive use of industrial inputs, fertilizers and pesticides. Work on improving nitrogen use efficiency in cereal crops, such as rice, wheat and maize, offers an important glimpse at possible approaches to achieve greater sustainability (Li et al., 2018; Zang et al., 2019; Geddes et. al., 2018; Wu et al., 2020). Engineering cereals to fix their own nitrogen is another very promising, albeit likely a more distant solution (Mus et. al., 2016; Soumare et. al., 2020). Finally, enhancing legume production and usage as a sustainable and inexpensive meat alternative is being seen as a feasible avenue. However, while capable of fixing their own nitrogen, legumes yield much less grain per acre than

cereals do. In Ontario, for example, corn produced on average a grain yield of approximately 4.16 metric tons/acre while soybean yielded only 1.38 metric tons/acre, during the 2020 season (http://www.omafra.gov.on.ca/english/stats/crops/index.html). The lower productivity coupled with the rapidly growing population means that further, sustainable enhancement of legume production is also an important goal (Liu et al., 2020). Furthermore, the common bean (*Phaseolus vulgaris*) is a weak nitrogen fixer compared to other legumes and is routinely supplemented with industrial nitrogen to enhance yield. These two examples clearly indicate that further enhancement of legume production is also an important goal toward sustainable food production (Liu et al., 2020).

9. Rationale and Objectives

L. japonicus HAR1 is a regulator of nitrogen-fixing root nodule symbiosis and a phosphate acquiring VAM. *HAR1* has also been shown to regulate root system architecture (Wopereis et. al., 2001; J. Therrien and K. Szczyglowski, unpublished data). There is also evidence, based on the work with the *SUNN* locus, the *M. truncatula* orthologue of *HAR1*, pointing to its role in a resource partitioning mechanism that regulates the mobilization of soil-derived nitrogen from roots to shoots (Lagunas et. al., 2019). Given these attributes, *HAR1* and its orthologues in crop species may be important targets for legume crop improvement. I hypothesised that generating a range of *cis*-regulatory *HAR1* alleles using a CRISPR/Cas system would lead to the identification of synthetic variants with enhanced capacity for SNF and likely also other useful properties with regard to the plant nutrient efficiencies. As the CRISPR/Cas-based approach had to first be implemented in the Szczyglowski laboratory, the following objectives were set for my MSc thesis work:

1) Design guide RNAs targeting the *HAR1* promoter and develop the corresponding CRISPR/Cas9 and CRISPR/Cas12a multiplex expression cassettes.

2) Test the expression, maturation, and genome editing efficiencies of the corresponding CRISPR cassettes *in planta* using an *Agrobacterium rhizogenes*-induced hairy root system.

Hairy roots are induced by *A. rhizogenes*, which transfers a T-DNA segment of its root inducing (Ri) plasmid to plant cells causing a callus-like structure to form at the infection site due to alterations to a hormonal status of the infected cells (Beach & Gresshoff, 1988). Hairy roots subsequently start to emerge from the callus-like structure. This effect has been adopted as a short-cut method to generate transgenic roots (hairy roots) on non-transgenic shoots, in order to study biological processes (Stiller et. al., 1997). Generation of hairy roots is relatively rapid, hence allowing for the swift determination of CRISPR cassette functionality in genome editing.

The long term goal of my thesis work was to contribute to the basic knowledge required for the enhancement of symbiotic nitrogen fixation and other nutrient acquisition-related traits in legume crops.

Chapter 2: Materials and Methods

1. Plant material, bacterial strains, and growth conditions.

All experiments were conducted using wild type *Lotus japonicus* ecotype Gifu plants. Seeds were lightly sanded to scarify the protective coating. They were then sterilized by incubation in 5mL of 70% ethanol/0.1% SDS solution, decanting after 1 minute, and performing an additional incubation in 5mL of 20% bleach/0.1% SDS solution, also for 1 minute. The seeds were then washed 10 times using 5mL of sterile Milli-Q H_2O, vortexing briefly with each wash. Seeds were imbibed overnight in 5mL of sterile water at room temperature. The seeds were subsequently germinated on Gamborg's ½ B5 media (Sigma-Aldrich, Oakville, ON, Canada) solidified with 0.8% Phytagel (Sigma-Aldrich, Oakville, ON, Canada). Germination was carried out for 5 days in the dark at room temperature, in order to generate etiolated seedlings, to be used for *Agrobacterium rhizogenes* transformation (see below).

Escherichia coli strains *DH5α* and TOP10 (Thermo Fisher Scientific, Waltham, USA) were used for all cloning experiments. Both strains were cultured on either solid or liquid LB media (liquid cultures were shaken at 220rpm) at 37°C. *A. rhizogenes* strain *1193* was used to generate transgenic hairy roots and was grown on LB media (liquid cultures were shaken at 220rpm) containing 100µg/mL rifampicin and 100µg/mL L-histidine at 28°C. Antibiotics used for the selection of plasmids included: kanamycin at 25µg/mL for any construct in the pCAMBIA1600 vector backbone; kanamycin at 50µg/mL for any construct in the pGEL031, pGEL032 or pCRII-Blunt-TOPO vector backbones; and ampicillin at 100µg/mL for any construct in the pBluescript SK (+) or pUC57 vector backbones.

2. Designing single guide RNAs (sgRNAs) targeting the *Lotus japonicus HAR1* promoter.

Putative CRISPR/Cas9 target sites, with the Cas9 protospacer adjacent motif (PAM), NGG (where "N" is any nucleotide), at the 3' end were identified in the *HAR1* promoter region using the CRISPR-P v2.0 software (http://crispr.hzau.edu.cn/CRISPR2/). Additionally, the same software was used to identify putative CRISPR/Cas12a target sites in the *HAR1* promoter region based on the Cas12a PAM, TTTV (where "V" is any nucleotide other than a T). The 4kb region upstream from the predicted *HAR1* translation initiation ATG codon was arbitrarily defined as the promoter (*Appendix A*). Target sites that have the least likelihood to share sequence identity with off-target regions in the *L. japonicus* genome were subject to analysis of tertiary structure (as predicted by the CRISPR-P v2.0 software – see results). The target sites with the optimal single guide RNA (sgRNA) tertiary structure were selected to be incorporated into the multiplex CRISPR/Cas9 cassette. CRISPR/Cas12a mediated genome editing requires a much shorter scaffold portion of the sgRNA, called the Cas12a direct repeat (DR). The CRISPR/Cas12a expression cassettes had crisprRNAs (crRNAs) incorporated with the shorter scaffold flanking both the N and C terminals.

3. Developing corresponding CRISPR/Cas9 multiplex sgRNA expression system (double promoter system).

The sgRNAs selected based on the *in-silico* analysis described above were used to design the dual promoter CRISPR/Cas9-multiplex expression system (*CRISPR-HAR1p-1*). sgRNAs were assembled in a cassette that was driven by the *L. japonicus* small nuclear RNA (snRNA) *U6* promoter (*LjU6p*), while the Cas9 protein was expressed under the constitutive 2x CaMV 35S promoter (2x35S). The cassette was designed to begin and end with *BbsI* restriction sites (i.e. type

IIS restriction enzyme target site), to match the cloning sites of the intermediary vector, pBluescript SK(+) (Wang et. al., 2016). The intermediary vector carried the *L. japonicus U6* promoter and RNA pol III terminator with two *BbsI* restriction sites between them (Wang et al., 2016). The sequence of a *L. japonicus glycine* pre-tRNA (trna76 – chromosome 6: 38328889-38328959, *Lotus japonicus* Gifu v1.2; https://lotus.au.dk/), was selected to precede each of the individual sgRNAs. The sequence of the cassette was designed as follows: *BbsI* - pre-tRNA – sgRNA1 – pre-tRNA – sgRNA2…. pre-tRNA – sgRNA9 – *BbsI; Appendix A*) and was synthesized by Bio Basic Inc. (Markham, ON, Canada). The synthetic fragment arrived cloned into the pUC57 vector (Bio Basic Inc.; Markham, ON, Canada). It was subsequently cloned into the pBluscript SK(+) vector, containing the *U6* snRNA promoter and RNA pol III terminator (see above), using a *BbsI* restriction enzyme digestion (20U per reaction), as follows: 4µg of pUC57 carrying the sgRNA cassette, and 4µg of the intermediary cloning vector were digested with *BbsI*, in individual reactions. The reactions were performed in 1x CutSmart buffer at 37˚C for 1 hour (New England Biolabs Ltd., Whitby, Canada). The digestion products were separated using a 1% agarose gel (run at 5V for 12 hours, followed by 120V for 30 minutes). The desired bands of 4.1kb (intermediary vector) and 1.5kb (the sgRNA cassette) were isolated from the gel using the QIAquick Gel Extraction Kit (QIAGEN, Toronto, Canada), following the manufacturer's protocol. The intermediary vector (50ng) and the sgRNA cassette (150ng) were then combined in a ligation reaction using T4 DNA ligase (2000U), incubated in 1x T4 DNA ligase buffer at 16˚C overnight (New England Biolabs Ltd., Whitby, Canada). 1µL of the resulting ligation product was mixed with 20µL of *E. coli* DH5α electrocompetent cells in a prechilled 1mm electroporation cuvette. A brief electric current (1.8kV, with an expected time constant of ~5.0msec) was applied to the cuvette using the *E. coli* 1 (Ec1) setting on the Bio-Rad Laboratories MicroPulser Electroporator

(Mississauga, ON, Canada). 1mL of SOC medium (2% tryptone, 0.5% yeast extract, 10 mM NaCl, 2.5 mM KCl, 10 mM MgCl2, 10 mM MgSO4, and 20 mM glucose; Thermo Fisher Scientific, Waltham, USA) was added to the cells immediately after the electric shock was applied. The cells were allowed to recover for 1 hour at 37°C, with shaking at 220rpm. The cells (10µL) were then plated on LB media containing ampicillin for selection of the plasmid. Following overnight incubation at 37°C, individual colonies were selected and grown in a liquid cultures, that were used for plasmid DNA isolation (see section 4 for details). The resulting plasmid was then double digested with *KpnI*/*XbaI* (10U of each enzyme per reaction), in NEBuffer 2.1 at 37°C for 1 hour (New England Biolabs Ltd., Whitby, Canada). The digestion released the entire expression unit, including the *U6* promoter, sgRNA multiplexed cassette and terminator, to be cloned into the pCAMBIA1600 destination vector. The vector contained the *Streptococcus pyrogenes Cas9* (*SpCas9)*, under the control of the 2x35S promoter, and the hygromycin and kanamycin resistance genes as the plant and bacterial selection markers, respectively (Wang et al., 2016). The expression unit was ligated into the pCAMBIA1600 vector using T4 DNA ligase as described above, to generate the *CRISPR-HAR1p-1* vector. The integrity of the *CRISPR-HAR1p-1* vector was confirmed using an *EcoRI*/*KpnI*/*XbaI* triple restriction enzyme digestion. The *EcoRI* (10U) digestion was conducted in 1x CutSmart buffer at 37°C for 1 hour, then the DNA was purified using phenol chloroform extraction, and double digested using *KpnI*/*XbaI* (10U of each enzyme), by incubating in NEBuffer 2.1 at 37°C for 1 hour (New England Biolabs Ltd., Whitby, Canada). The product of the triple digest was separated on a 0.8% agarose gel, for evaluation of fragment sizes.

4. Developing a single transcriptional unit (STU) construct for CRISPR/Cas9.

The first eight sgRNAs (including their preceding glycine pre-tRNAs) were amplified as a single fragment (i.e. 8x sgRNA cassette) from the *CRISPR-HAR1p-1* expression system using GXL PrimeSTAR polymerase (Takara Bio, Kusatsu, Japan; cycling regime: 10 minutes at 98°C, 25 cycles of 98°C for 10 seconds, 63.5°C for 15 seconds, and 68°C for 1 minute, followed by 10 minutes at 68°C) and primers (*HAR1-gRNA-F-BsaI* and *HAR1-gRNA-R-BsaI; Appendix B*) designed to introduce 5' overhangs that contain the *BsaI* cloning sites matching those within the pGEL031 destination vector (Tang et. al., 2019; Addgene plasmid # 137900; *Appendix C*). The destination vector contains the *SpCas9* gene driven by the *Zea mays* UBIQUITIN promoter (*ZmUBQ*p), the hygromycin resistance gene as a plant selective marker, and a kanamycin resistance gene as a bacterial selection marker. The destination vector also contains the *ccdB* (suicide) gene, to select against the presence of the un-recombined vector, and two *BsaI* cloning sites that allow easy cassette incorporation. The amplified 8x sgRNA cassette was cloned into the pGEL031 vector using Golden Gate assembly (Engler et. al., 2008). Briefly, 5µL of crude PCR product was mixed with *BsaI* (30U), T4 DNA ligase (1000U) and pGEL031 (~1µg) in 1x T4 DNA ligase buffer (New England Biolabs Ltd., Whitby, Canada), in a total volume of 20µL. Using a PCR machine, the reaction was cycled 30 times under the following conditions: 37°C for 5 minutes, 16°C for 5 minutes, followed by 1 hold period at 60°C for 5 minutes. 2µL of the reaction product, was directly transformed into chemically competent *E. coli* TOP10 cells (Thermo Fisher Scientific, Waltham, USA) using heat shock procedure at 42°C for 30 seconds. Transformed cells were allowed to recover in 250µL of SOC media for 1 hour at 37°C, with shaking at 220rpm and were grown on medium containing kanamycin to select for cells containing the recombined vector, called hereafter *CRISPR-HAR1p-2*.

The integrity of the recombined binary vector was confirmed using a *HpaI*/*HindIII* (10U of each enzyme) double digestion, incubated in 1x CutSmart buffer at 37˚C for 1 hour. The product was separated on a 0.8% agarose gel to confirm correct fragment sizes.

5. Developing single transcriptional unit constructs for CRISPR/Cas12a.

The Cas12a crRNAs selected based on the *in-silico* analysis described above, were used to design a single transcriptional unit (STU) system for genome editing using CRISPR/Cas12a. The cassette was designed to contain *BsaI* restriction sites matching those within the pGEL032 destination vector (Tang et. al., 2019; Addgene plasmid # 137901; *Appendix C*). pGEL032 contains the *Lachnospiraceae bacterium Cas12a* gene (*LbCas12a*) driven by the maize ubiquitin (*ZmUBQ*) promoter, while its remaining features are the same as pGEL031 (*Appendix C*). The original pGEL032 destination vector was modified to have the *LbCas12a* gene and crRNA expression cassette driven by either the *LjUBQp* or the 2x35S promoter, respectively. The *LjUBQp* and the 2x35S promoter were both amplified using primers designed to introduce *AscI* and *SbfI* restriction sites that would allow promoter replacement in the destination vector. The primers used were AscI_UBQ-F + UBQ_SbfI-R, and AscI_2x35S-F + 2x35S_SbfI-R, respectively (see *Appendix B*). Both the pGEL032-*LjUBQp* (pGEL032a) and pGEL032-2x35S (pGEL032b) variations of the destination vector were used to generate STU systems. The cassette was originally designed to contain ten Cas12a target sites, however this sequence proved to be too repetitive to be chemically synthesized in a single fragment, so it was re-designed in five smaller fragments, each containing 2 gRNAs. These fragments were designed to contain *BsaI* restriction sites flanking both ends. The sticky ends resulting from *BsaI* digestion were complimentary with either each other or the pGEL032 destination vector. The five Golden Gate assembly fragments were designed

as follows: *BsaI* – crRNA1 – crRNA2 – *BsaI*; *BsaI* – crRNA3 – crRNA4 – *BsaI* ……. *BsaI* – crRNA9 – crRNA10 – *BsaI*. Two variations were ordered for each fragment. The first variation contained *BsaI* restriction sites complimentary to the pGEL032 vector, regardless of the promoter driving cassette expression, and served as a duplex cassette that targeted two genomic sites simultaneously. The second variation of each of the synthetic fragments contained *BsaI* restriction sites complimentary to each other, allowing the incorporation of all 10 crRNAs into a multiplex cassette targeting all ten genomic target sites simultaneously. *BsaI* restriction sites were included to allow for incorporation into the vectors (*Appendix A*).

The synthetic fragments arrived from Bio Basic Inc. (Markham, ON, Canada) cloned individual into the pUC57 vector, and were subsequently cloned into the pGEL032 destination vector using the same Golden Gate assembly and transformation procedures described above. The completed duplex vectors were named *CRISPR-HAR1p-*$_{LjUBQp}$*1,2, CRISPR-HAR1p-*$_{LjUBQp}$*3,4, CRISPR-HAR1p-*$_{LjUBQp}$*5,6, CRISPR-HAR1p-*$_{LjUBQp}$*7,8, CRISPR-HAR1p-*$_{LjUBQp}$*9,10, CRISPR-HAR1p-*$_{2x35S}$*1,2, CRISPR-HAR1p-*$_{2x35S}$*3,4, CRISPR-HAR1p-*$_{2x35S}$*5,6, CRISPR-HAR1p-*$_{2x35S}$*7,8,* and *CRISPR-HAR1p-*$_{2x35S}$*9,10*. The multiplex expression cassette, containing all ten target sites, in pGEL032 driven by the *LjUBQp* and 2x35S promoters were named *CRISPR-HAR1p-*$_{LjUBQp}$*10sgRNA* and *CRISPR-HAR1p-*$_{2x35S}$*10sgRNA,* respectively. Integrity of each recombined binary vector was confirmed using a *XbaI*/*BglII* (10U of each enzyme) double digestion. The double digestion was conducted in 1x NEBuffer 3.1 (New England Biolabs Ltd., Whitby, Canada) at 37˚C for 1 hour. The double digest was separated using a 0.8% agarose gel.

6. Transforming *Agrobacterium rhizogenes 1193.*

A. rhizogenes strain *1193* was separately transformed with the *CRISPR-HAR1p-1* and *CRISPR-HAR1p-2* vectors, by electroporation using the Bio-Rad Laboratories MicroPulser Electroporator (Mississauga, ON, Canada) on the *Agrobacterium* (Agr) setting. 20µL of *A. rhizogenes 1193* electro-competent cells (generated by growing cells to an OD_{600} = 0.5, performing ten washes in cold, sterile water, then suspending the bacteria in sterile 10% glycerol) were transferred to a pre-chilled 1mm electroporation cuvette, along with 1µL of either *CRISPR-HAR1p-1* or *CRISPR-HAR1p-2* (~200ng/µL in 10mM Tris-HCl buffer, 0.1mM EDTA, pH 8.0). The cuvettes were subjected to a brief electric current (2.2kV, with an expected time constant of ~5.00msec), after which 1mL SOC (Thermo Fisher Scientific, Waltham, USA) was immediately added to the cells. The cells were given 1 hour to recover at 28˚C, while shaking at 220rpm. The transformed cells (10µL of a 1:10 dilution of the original recovery culture + 40µL of SOC) were then plated on LB agar plates containing 100µg/mL rifampicin and 100µg/mL L-histidine, along with the appropriate selective antibiotic (see section 1) and grown for 2 days at 28˚C. Individual colonies were grown in selective liquid LB media for 2 days at 28˚C with shaking (220 rpm). The binary vectors were isolated using the High-Speed Plasmid Mini Kit (Geneaid, New Taipei City, Taiwan). The resultant vectors (~200ng in 1µL of 10mM Tris-HCl buffer, 0.1mM EDTA, pH 8.0) were transformed into the electro-competent *E. coli DH5α* strain (Thermo Fisher Scientific, Waltham, USA) using the same electroporator as described above and the *E. coli 1* (Ec1) setting (1.8kV, with an expected time constant of ~5.00msec). Cells were grown out as described above and plasmid DNA was re-isolated using the same High-Speed Plasmid Mini Kit (Geneaid, New Taipei City, Taiwan). The vectors were sent to the Robarts' Research Facility at the University of

Western Ontario (London, ON, Canada) for confirmatory DNA sequencing (see *Appendix B* for primers).

7. Preparing transgenic hairy root cultures.

L. japonicus seeds (50 seeds per construct) were germinated as described above. Elongated hypocotyls were poked three times with the 28 gauge needle of a sterile U-100 syringe (Becton Dickinson and Company, Frankin Lakes, NJ, USA) to lightly wound the plant. *A. rhizogenes 1193* carrying the appropriate binary vector (either *CRISPR-HAR1p-1* or *CRISPR-HAR1p-2*) was streaked over the wounds. The plants were returned to the same media, incubated at room temperature in the dark for 24 hours, then transferred to a growth cabinet held at 23˚C with a 16h light/8h dark cycle. Callus-like structures had formed at inoculation sites within ten days. The true roots were removed at the hypocotyls, which were then transferred, along with the callus-like structures and developing shoots, to a hairy root growth medium comprised of liquid Gamborg's ½ B5 media with cefotaxime 300µg/mL (Sigma-Aldrich, Oakville, ON, Canada) to select against *Agrobacterium*. The transformed hypocotyls were returned to the same incubator as the previous step. Hairy roots developed on the plants after approximately ten days. Once hairy roots grew to a length of approximately 3cm they were removed at the shoot/root junction and transferred individually to petri dishes (100 x 25mm) containing 50mL of hairy root culture media, which was comprised of liquid B5 media with hygromycin (15µg/mL) and cefotaxime (300µg/mL). Plates were sealed with surgical tape and incubated for four weeks at room temperature with continuous shaking at 60rpm under a 16h light/8h dark cycle. The hairy roots were transferred to fresh media every four weeks until completion of the experiment.

8. Genotyping hairy roots for the presence of the transgene.

DNA was isolated from hairy roots using the CTAB method (Doyle and Doyle, 1987). Briefly, 25mg of fresh tissue was milled in 2x CTAB buffer (2% CTAB, 100mM Tris pH 8.0, 0.1mM EDTA, pH 8.0, 1.4M NaCl, 1% PVP 40,000). A phenol-chloroform (1:1) extraction was subsequently preformed to remove proteins, cell membrane, cell wall, and the contents of the cytosol. The DNA is not soluble in the phenol-chloroform, so centrifugation at maximum speed in a microcentrifuge for 2 minutes was used to separate phases. The aqueous phase was transferred to a clean tube, and DNA was recovered through precipitation by adding 3/4 volume of isopropanol in the presence of 1/10 volume of 3M sodium acetate, pH 5.2. The DNA was pelleted by centrifugation (max speed) and then washed once with 250µL of 70% ethanol. After removing the ethanol, the DNA was air dried for an hour. The DNA was re-suspended in 10mM Tris-HCl, 0.1mM EDTA, pH 8.0. The *CRISPR-HAR1p-1* and *CRISPR-HAR1p-2* T-DNA fragments, were predicted to be integrated into the genomic DNA of the transgenic hairy roots. A Polymerase Chain Reaction (PCR) was used to amplify diagnostic fragments of the *CRISPR-HAR1p-1* and *CRISPR-HAR1p-2* transgenes, using the GXL PrimeSTAR polymerase (Takara Bio, Kusatsu, Japan) and the following cycling parameters: 10 minutes at 98˚C, 25 cycles of 98˚C for 10 seconds, 55˚C for 15 seconds, and 68˚C for 1 minute, followed by 10 minutes at 68˚C. The forward primers, *LjU6-cassette-F* for *CRISPR-HAR1p-1* and *STU-C9-tDNA-F* for *CRISPR-HAR1p-2*, were unique to their respective cassette, while the reverse primer, LjHAR1-cassette-mid-R, was common to both cassettes (*Appendix B*).

9. Confirming cassette expression using circular reverse transcription (cRT) PCR.

The post-transcriptional maturation of the polycistronic sgRNA cassettes was tested using the hairy root system. Total RNA was extracted from 75mg of tissue from 3 transgenic hairy roots using the *mirVana* miRNA Isolation Kit (Thermo Fisher Scientific, Waltham, USA), following the manufacturer's protocol for total RNA isolation. RNA purity and concentration were assessed using the QIAxpert spectrophotometer (QIAGEN, Toronto, Canada). The RNA (6µg per hairy root sample) was treated with Turbo DNaseI (6U; Thermo Fisher Scientific, Waltham, USA) incubated in 1x Turbo DNase buffer at 37°C for 1 hour, to remove any potential DNA contamination. 3µg of the resulting total RNA was circularized by incubating for 2 hours at 25°C with T4 RNA Ligase 1 (30U) in 1x T4 RNA ligase buffer (New England Biolabs Ltd., Whitby, Canada) containing 60U of RNase inhibitor to prevent RNA degradation. Each 60µL circularization reaction was diluted with 40µL of sterile water. Proteins were removed by addition of 100µL of phenol-chloroform-isoamyl alcohol (25:24:1 v/v), followed by vortexing, and centrifuging at maximum speed for 5 minutes to separate phases. The aqueous phase (approximately 100µL) was transferred to a clean tube, and the circularized RNA was precipitated by adding $1/10^{th}$ volume of 3M sodium acetate (10µL), 5µL of glycogen (20mg/mL), and 2.5 volumes (250µL) of anhydrous ethanol. After mixing by vortexing, samples were centrifuged at the maximum speed for 30 minutes. The supernatant was decanted, and the resulting RNA pellet was washed with 250µL of 70% ethanol. The RNA was then air dried for 3 minutes at room temperature and re-suspended in 30µL of sterile, nuclease free water. Reverse transcription (RT) reactions, each containing approximately 200 ng of circularized RNA as template, were conducted using sgRNA specific primers (*Appendix B*) and the iScript Select cDNA Synthesis Kit (Bio-Rad Laboratories, Mississauga, ON, Canada), following the manufacturer's protocol. PCR amplification of sgRNAs was conducted on the 1^{st}

cDNA strand, using 2µL of crude RT reaction, Phusion Taq Polymerase (Thermo Fisher Scientific, Waltham, USA) and the following cycling conditions: 10 minutes at 98˚C, 35 cycles of 98˚C for 30 seconds, 55˚C for 30 seconds, and 72˚C for 1 minute, followed by 10 minutes at 72˚C (see *Appendix B* for primers). Amplification products were cloned into the pCRII-Blunt-TOPO vector from the Zero Blunt TOPO PCR Cloning Kit (Thermo Fisher Scientific, Waltham, USA), following the manufacturer's protocol. Sanger sequencing of the amplification products (*Appendix B* for primer sequences) was used to determine how efficiently the *CRISPR-HAR1p 1* and *CRISPR-HAR1p 2* cassettes are being processed. Sequencing was contracted by the Robarts Research Institute facility at University of Western Ontario (London, ON, Canada).

10. Identification of mutations in target region.

Sanger sequencing was used to determine the *HAR1* promoter sequence. Genomic DNA was extracted from transgenic hairy roots using the CTAB protocol described above (see section 6). The *L. japonicus HAR1* promoter region was amplified in two overlapping fragments using the GXL PrimeSTAR polymerase (Takara Bio, Kusatsu, Japan) and primer pairs *HAR1p-F* and *HAR1p-mid-R*, and *HAR1p-mid-F* with *HAR1p-R* (*Appendix B*). The following cycling parameters were used: 10 minutes at 98˚C, 25 cycles of 98˚C for 10 seconds, 60˚C for 15 seconds, and 68˚C for 2.5 minutes, followed by 10 minutes at 68˚C. The amplicons were cloned into the pCRII-Blunt-TOPO vector using the Zero Blunt TOPO PCR Cloning Kit (Thermo Fisher Scientific, Waltham, USA), following the manufacturer's protocol. Sanger sequencing of the cloned fragments was performed at Eurofins Genomics (Louisville, USA – see *Appendix B* for primer sequences).

11. Identification of *HAR1* orthologues in other legumes.

Orthologues of the *L. japonicus HAR1* gene were identified using the Genome Context Viewer tool (https://legumeinfo.org/gcv2/instructions) on the Legume Information System (LIS; https://legumeinfo.org/). The accession number for the *HAR1* locus, Lj3g3v3375780 (MG20 v3.0 genome annotation; https://lotus.au.dk/view/transcript/Lj3g3v3375780.1), was provided to the tool search bar to query all the legume genomes associated with the LIS. Any database sequences that shared sequence identity with the query gene were selected as potential orthologues. Subsequently, genomic regions surrounding the query gene (i.e. *HAR1*) and the genes identified as potential orthologues were analyzed for further instances of shared sequence identity. Surrounding genes that shared sequence identity with each other were selected, and when a sufficient number of them was found, a microsynteny map was generated. The parameters used to generate the output were as follows: the number of neighbouring genes surrounding the query gene that were analyzed was 20 ("Neighbors = 20" under the "Micro-Synteny" tab); the minimum number of genes that must match between the query track and search tracks was 8 ("Match = 8" under the "Micro-Synteny Alignment" tab); and the rest of the settings were at default. Information on the algorithms used by the program can be found on the associated websites (https://github.com/legumeinfo/gcv/wiki/User-Help & https://legumeinfo.org/gcv2/instructions).

12. Analysis of natural variation at the *NARK* locus in different soybean varieties.

The location and identity of single nucleotide polymorphisms (SNPs) in 350 different soybean varieties is provided as a master dataset at http://soykb.org/public_data.php. In order to retrieve the SNPs for the *Glycine max NARK* locus, an orthologue of *L. japonicus HAR1* gene,

from the master dataset, I developed a python script, **retrieveSNPs.py**. Each chromosome has a unique master dataset for SNPs, and the correct chromosome had to be specified in the initial steps (see *Appendix D* for script parameters). Subsequently, the python script **setup_SNP.py** was developed and used to create an intermediate file which contained the entire *NARK* sequence of the soybean Williams 82 reference genome, associated to its genomic address (see *Appendix D* for script parameters). The output files from the **retrieveSNPs.py** script and the **setup_SNP.py** script were merged using R-studio (see *Appendix D* for R-code). To allow for the counting of polymorphic lines from the merged file it was necessary to remove nucleotides if they matched the reference sequence at their given position. To this end I developed and utilized a python script, **removeRedundantSNPs.py**. The output file from the **removeRedundantSNPs.py** script was utilized to conduct statistical and graphical analysis of SNP frequencies. All statistical analyses were carried out in Excel (Version 2016) (see *Appendix D* for excel formulae). Graphical analysis was conducted using R-studio package "ggplot" (see *Appendix D* for R-code). See *Appendix D* for the pipeline scripts used to conduct this analysis.

Chapter 3: Results

1. Selection of a target locus for improved nitrogen fixation.

Root nodules of legumes provide the optimal environment for symbiotic rhizobia to convert atmospheric N_2 into NH_4^+. Altering the number of nodules that the plant sustains could alter the level of source nitrogen available to the host. Therefore, the genes that act as regulators of nodule formation were considered as viable targets to alter the nodulation phenotype. The *HAR1* gene from the model legume *L. japonicus* was chosen as the sole target for this thesis work. HAR1 is the central, systemic regulator of symbiotic plant nutrition, including nitrogen-fixing root nodule formation and phosphate-acquiring mycorrhiza symbiosis. HAR1 also responds to abiotic cues related to plant nutrition, such as nitrate and phosphate, and acts locally and systemically to regulate root system architecture, which makes this receptor an interesting target for improvement of nutrient-related traits in legume plants (Figure 6).

The *L. japonicus HAR1* gene functions in a dose-dependent manner. Plants that are homozygous for the loss of function allele, *har1-1*, develop five times more nodules than wild-type, the mutant phenotype referred to as hypernodulation (Wopereis et al., 2000; Krusell et al., 2002). However, heterozygous *HAR1/har1-1* plants exhibit an intermediate level of nodulation (Figure 7). This indicates that HAR1 acts in a dosage-dependent manner to regulate nodulation. Importantly, unlike the *har1-1* homozygote, which suffers from significantly restricted shoot and root growth, the development of heterozygote plants is similar to wild-type (M. Pampuch and K. Szczyglowski, unpublished data), suggesting that perhaps further fine-tuning of the HAR1 receptor levels might lead to improved plant productivity.

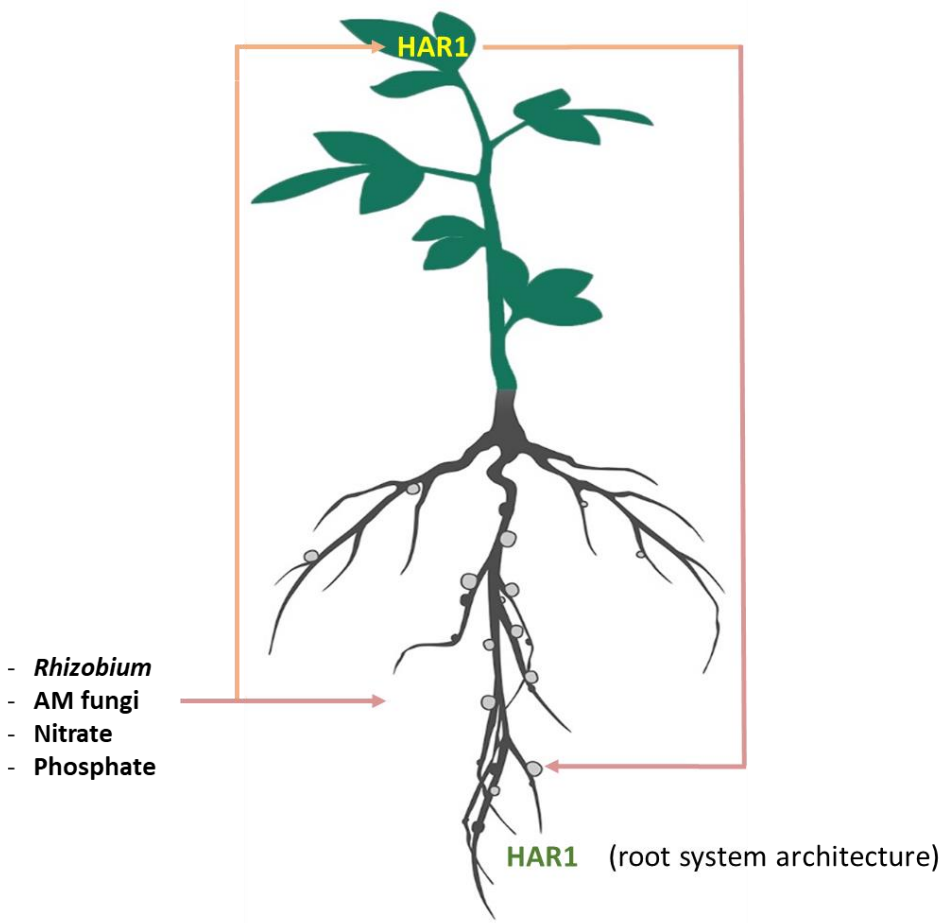


Figure 5. The *Lotus japonicus HAR1* gene is the central regulator of the symbiotic relationship between *Lotus japonicus* and a nitrogen-fixing bacterium, *Mesorhizobium loti*. HAR1 partakes in both local and systemic regulation. In shoots, it is responsible for controlling the extent of root nodule formation, while in roots and shoots it mediates the root system architecture. HAR1 responds also to arbuscular mycorrhiza (AM) fungi and nutrient related cues, such as levels of nitrate and phosphate.

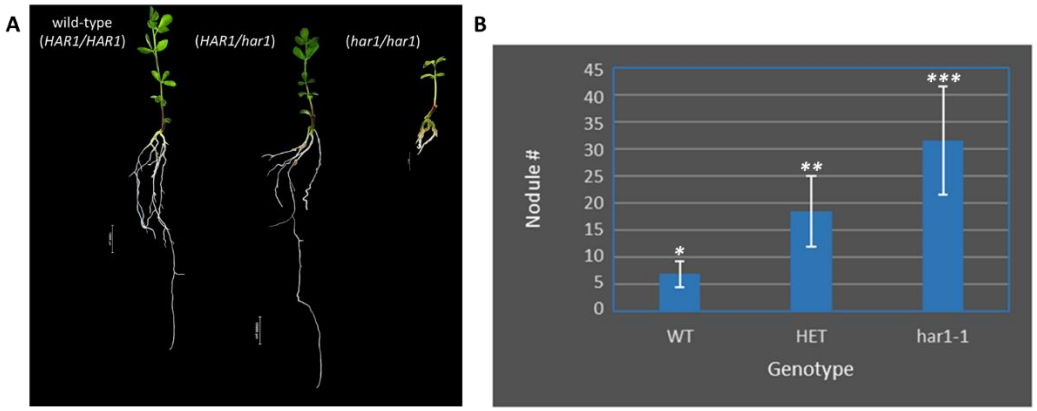


Figure 6. The dose-dependent impact of *HAR1* on symbiotic nodule formation. (A) The phenotypic variation between the wild type, heterozygote, and homozygote *har1-1* mutant with respect to plant growth. **(B)** A graphical representation of the average nodule counts per plant for a given genotype, 21 days after *rhizobial* inoculation. The wild type has fewer nodules than the null mutant, while the heterozygote has an intermediate number of nodules. Error bars represent standard deviation. Significance between groups is denoted by the stars above each bar (ANOVA, $p<0.05$) (Mark Pampuch and Krzysztof Szczyglowski, Unpublished Data).

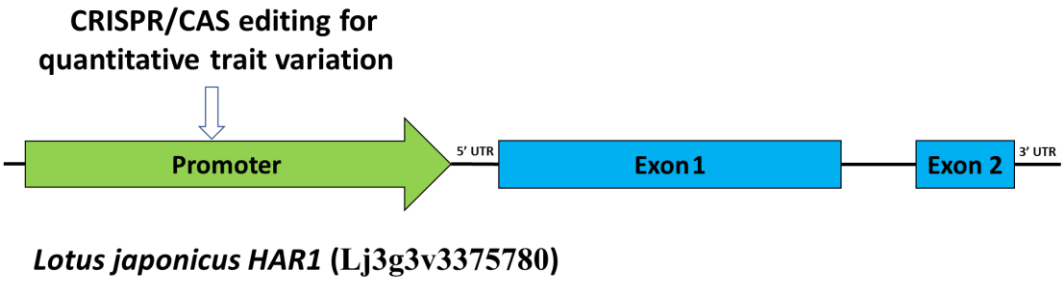


Figure 7. The *Lotus japonicus HAR1* gene structure. Different gene elements, including the promoter region, 5' UTR, exons, intron, and 3' UTR are shown.

To test this assumption, the promoter region of *HAR1* was selected as the primary target for CRISPR/Cas mediated editing (See *Appendix E* for a flowchart showing the experimental design), with the main, long-term goal of generating a range of quantitative trait variation at the locus. The 4kb region upstream from the predicted *HAR1* gene translation initiation codon (ATG) was arbitrarily selected for this purpose (Figure 8; see also *Appendix A*).

2. Orthologues of *L. japonicus HAR1* are present in other legume species.

L. japonicus HAR1 has orthologues in other legume species, and these have been shown, at least in some cases, to have a similar role in regulating symbiotic and non-symbiotic plant nutrition traits (Mirzaei et. al., 2017; Huault et. al., 2014). By identifying orthologues in other important legume crop species, such as *Phaseolus vulgaris* (common bean), and *Vigna unguiculate* (cowpea), it should be possible to fine tune these important traits directly in the crops. The context viewer of the Legume Information System (https://legumeinfo.org/gcv2/instructions) was used to generate a microsynteny map for six different legume species using the *L. japonicus HAR1* locus as the query (Figure 9). The orthology was assumed when at least eight neighbouring genes shared significant (a minimum alignment score of 30) sequence identity (Cleary and Farmer, 2018) with the *L. japonicus HAR1* region. Indeed, such evidence for orthology was found for all six legume species analysed (Figure 9).

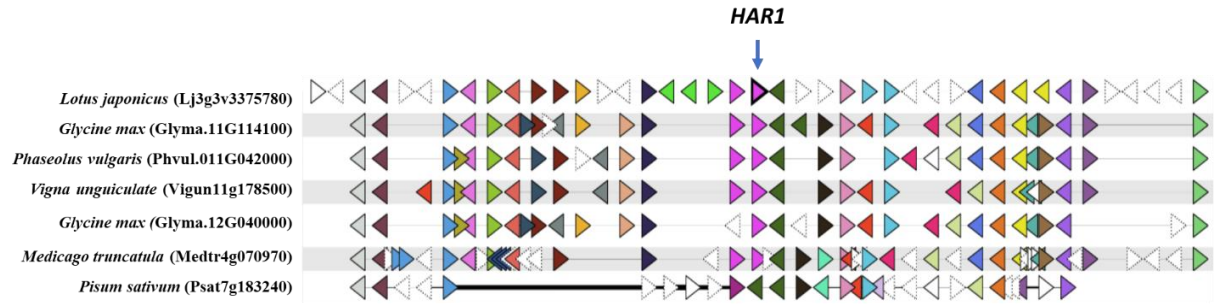


Figure 8. A microsynteny analysis identifies potential orthologues of the *Lotus japonicus HAR1* gene in important crop or model legume species. Each lane represents a genomic sequence predicted to be syntenic with the *L. japonicus HAR1* region. The triangles denote genes, with the bolded triangle in the top lane representing the query gene (i.e. *HAR1*). The same color triangles refer to presumed orthologous positions in the analyzed legume species. The accession numbers of *HAR1* and its orthologues are provided in parenthesis. The analysis was carried out using the legume information system context viewer web application (https://legumeinfo.org/gcv2/instructions).

3. The selection of Cas9 sgRNA targets sites and construction of the multiplex sgRNA expression cassette.

The use of CRISPR technologies allows for a site-specific genome editing, and in the case of the Cas9 endonuclease the target sites must contain a NGG PAM site adjacent to the 3' end of the gRNA target sequence. If the PAM is present and there is complementarity between the sgRNA and the genomic DNA, cleavage should occur. Potential target sites for CRISPR/Cas9 editing were identified in the *L. japonicus HAR1* promoter using the CRISPR-P v2.0 software. As any given sgRNA can potentially target multiple regions in the genome, the identified sgRNAs were

analyzed for additional target sites. The identified sgRNAs that had additional recognition sites in the *L. japonicus* genome, outside the *HAR1* promoter region, were not used in the subsequent experiments. For any given target site to be considered as a true off-target site it had to contain two or less mismatches with a given sgRNA. Thus, only the sgRNAs that contained no predicted true off-target sites were selected for further use. These sgRNAs had a minimum score of "3 MM" in the off-target analysis, indicating that these regions had more than two mismatches with the additional target region(s), as provided by the software.

The sgRNAs that had no true off-target sites were subjected to further analysis to determine whether they would have characteristics of sgRNAs that are associated with efficient DNA editing *in planta*. To be classified in the potentially efficient category, a given sgRNA must have been able to fold *in silico* into the appropriate secondary structure by forming three intact stem loops, called stem loop repeat/anti-repeat (RAR), stem loop 2, and stem loop 3 (Figure 10) (Liang et al. 2016). A fourth stem loop, stem loop 1, may be present, but is not required (Figure 10). Furthermore, such guide sequence was also required to have a GC content between 30% - 80% and to be free from six base-pair long self-pairing events. Furthermore, the guide portion of the sgRNA should not have more than seven consecutive or 12 total base pairing events between the guide and the rest of the sgRNA sequence (Figure 10).

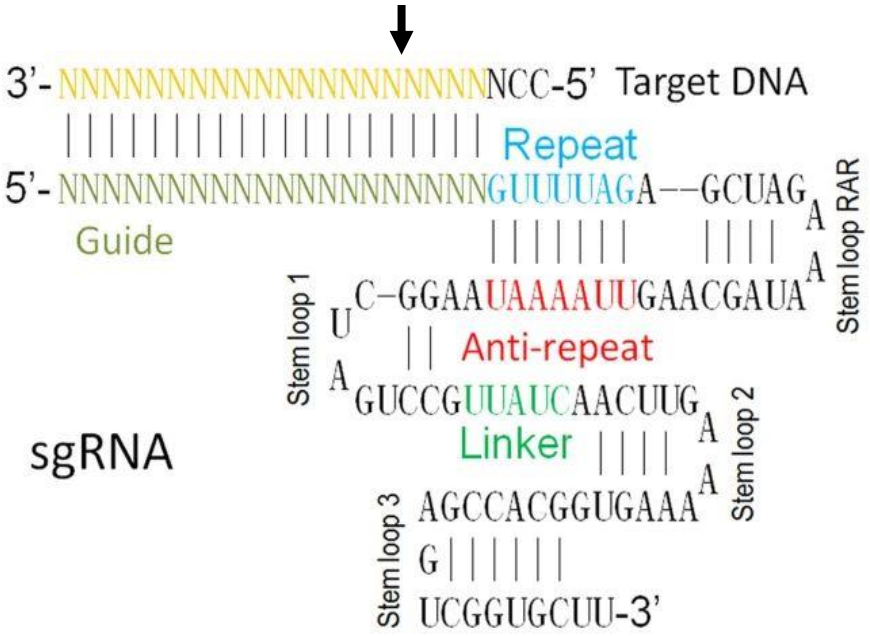


Figure 9. A typical secondary structure of mature sgRNA. The guide portion of the sgRNA is represented by green nucleotide symbol (N), while the tracrRNA is represented by the subsequent blue, red, green, and black sequences. A hypothetical target genomic sequence is shown along with the NCC protospacer adjacent motif (PAM), typical for Cas9. The arrow indicates a predicted double stranded break site. The figure was modified from Liang et al. 2016.

Based on these parameters nine unique Cas9 target sites were selected within the 4kb promoter region of the *L. japonicus HAR1* gene (Figure 11; *Appendix A*). The nine sgRNAs selected based on the *in-silico* analysis described above were incorporated into the sequence of the multiplex sgRNA expression cassette using commercial DNA synthesis (Bio Basic Inc., Markham, Ontario, Canada). The expression of the sgRNAs as a multiplex system provided three main benefits compared to expressing the sgRNAs under individual promoters. Firstly, the use of a multiplex system allowed for sgRNAs to be expressed under a single promoter, cutting down on the time required for cloning and plant transformation experiments. Secondly, the use of a multiplex sgRNA expression system allowed for different sgRNAs to generate DNA edits in the same cell, leading to small and large deletions (Rodríguez-Leal et. al., 2017). Lastly, it was shown that using a multiplex sgRNA expression system resulted in greater DNA editing efficiency of individual sgRNAs compared to expressing them under individual promoters (Xie et. al., 2015).

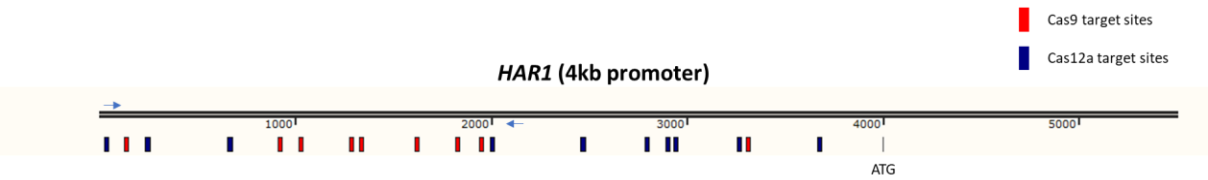


Figure 10. The locations of predicted Cas9 and Cas12a target sites in the promoter of the *HAR1* gene. The promoter region represents the 4kb long sequence upstream of the *HAR* translation initiation (ATG) codon. Cas9 target sites are marked by red boxes and Cas12a target sites are marked by blue boxes. The blue arrows indicate the position of the primers used to check for mutations in Figure 23 and Figure 24.

The cassette was designed to contain the nine selected sgRNAs, each preceded by the sequence of the *L. japonicus* glycine pre-tRNA. Such a configuration was shown to guide post-transcriptional processing using the endogenous tRNA maturation cellular machinery (Xie et. al., 2015) (see *Appendix A* for cassette sequence). This mechanism uses endogenous RNase P and RNase Z processing enzymes to cleave the tRNA sequences at both the 5' leader (AACAAA) and the 3' trailer (CA) sequences, respectively (Figure 12A). These sequences naturally occur in pre-tRNAs and were included in the multiplexed construct to facilitate processing. The post-transcriptional processing of the pre-tRNA sequences by RNase P and RNase Z released individual sgRNAs from the polycistronic transcript (Xie et. al., 2015). As a result, nine mature gRNA were predicted to be formed.

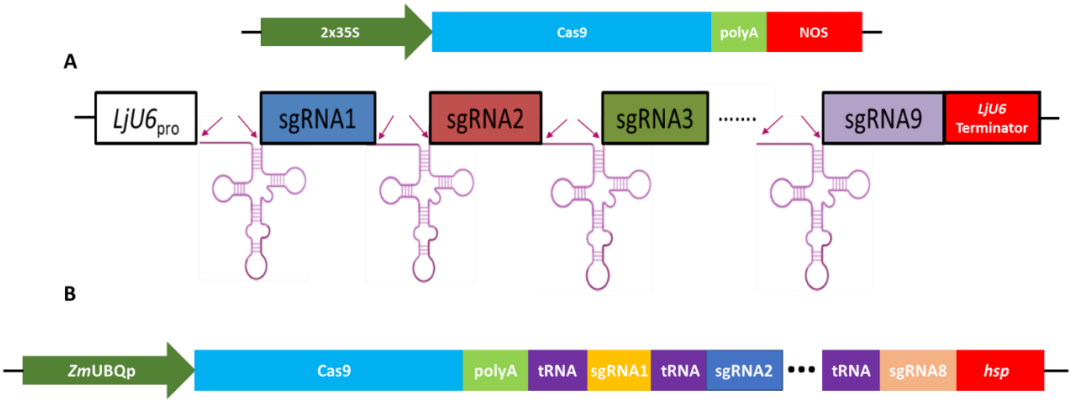


Figure 11. The structure of the Cas9 sgRNA expression cassettes. (A) The dual promoter expression cassette contains an array of sgRNAs driven by the *Lotus japonicus U6* promoter. The *L. japonicus* glycine-pre-tRNA separate each sgRNA and are to be used at the processing stage (arrows; see main text for more details). The construct contains a *L. japonicus U6* terminator. The *Cas9* gene is driven by the 2x35S promoter and has a NOS terminator. **(B)** The structure of the single transcriptional unit (STU) expression cassette. The entire construct, including the *Cas9* gene and the expression cassette, is driven by a single promoter. The STU construct contains the heat shock protein (*hsp*) terminator.

The expression cassette was synthesised at Bio Basic Inc. cloned into the pUC57 vector backbone. The cassette was isolated from the vector using a *BbsI* restriction enzyme digestion and the isolated fragment was successfully recombined into the intermediary, modified pBluescript SK (+) vector, containing the *L. japonicus U6* promoter and terminator. The cloning was confirmed using a *KpnI/XbaI* restriction enzyme double digestion (data not shown). The completed vector was subsequently used to generate the binary vector for *A. rhizogenes* mediated transformation.

4. The binary vector for editing *in-planta* using a dual promoter expression system was constructed.

A. rhizogenes mediated transformation delivers the transfer (T)-DNA portion of its root-inducing (Ri) plasmid into a host plant, which stimulates the formation of hairy roots. To achieve delivery of the desired gene or gene construct to the host plant, *A. rhizogenes* is often appended with an additional, so called binary vector, containing the desired sequences within its T-DNA region. This is co-transferred to plant cells with the original, Ri plasmid derived T-DNA (Kereszt et. al., 2007).

In order to test the genome editing capabilities of the sgRNAs multiplex cassette, a binary vector that contained all the necessary sequences for CRISPR/Cas9 mediated genome editing was developed. The Cas9 enzyme and the multiplex sgRNA expression cassette were expressed under distinct, 2x35S and *LjU6* promoters, respectively (see Material and Methods), thus the system was called the dual promoter expression system. The entire dual promoter expression system, as well as a plant selection marker (see the Material and Methods section), were included in the transfer region of the binary vector.

The pCAMBIA1300 vector backbone, containing 2x35S-*Sp*Cas9-NOS (kind gift from Prof. Duanmu, Huazhong Agricultural University, Wuhan, China), was used as the binary vector. The pBluescript SK (+) vector containing the multiplex sgRNA expression cassette was digested using *KpnI* and *XbaI* to release the cassette and the resulting fragment was ligated into the binary vector to give rise to the final construct, called *CRISPR-HAR1p-1*.

The integrity of the *CRISPR-HAR1p-1* vector was verified using an *EcoRI*/*KpnI*/*XbaI* triple restriction enzyme digestion. Upon confirming the expected DNA banding pattern (Figure 13), the vector was also subjected to sequencing. The resultant sequence was identical to the expected sequence (data not shown). Hence, the *CRISPR-HAR1p-1* vector was used in subsequent experiments.

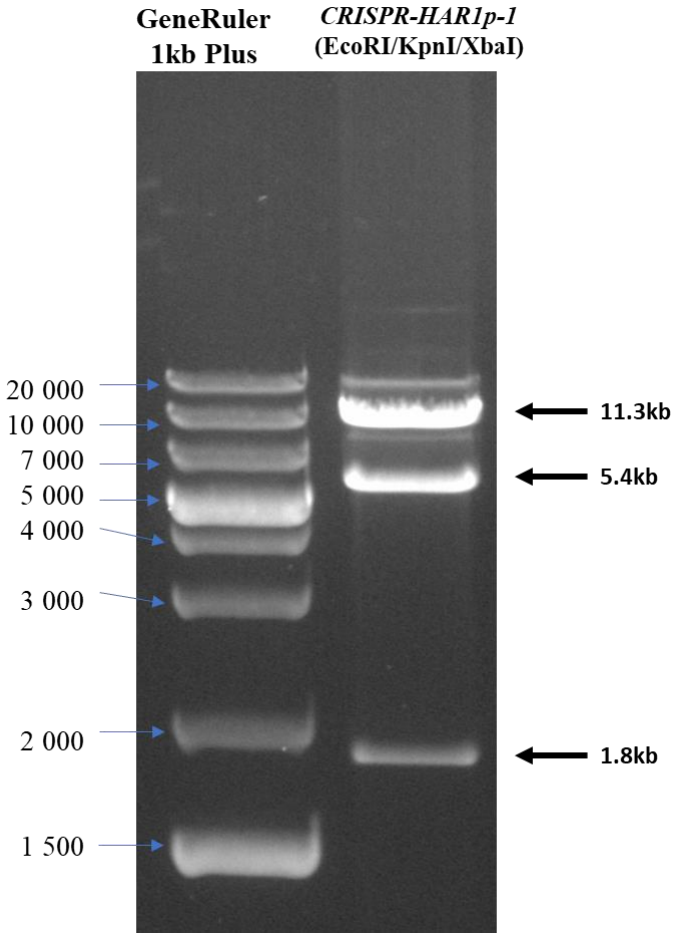


Figure 12. Confirmatory, restriction enzyme analysis of the *CRISPR-HAR1p-1* vector. An image of the ethidium bromide stained agarose gel, showing the predicted restriction fragment pattern of the *CRISPR-HAR1p-1* vector digested with *EcoRI*, *KpnI*, and *XbaI*. The expected DNA bands were predicted to be 11.3, 6 and 1.8 kb in size (arrows).

5. The binary vector for editing *in-planta* using a Cas9 single transcriptional unit (STU) system was constructed.

While work with *CRISPR-HAR1p-1* was ongoing, a new, apparently improved genome editing system, called a single transcriptional unit (STU) system, was reported (Tang et. al., 2019). I decided to adopt this system for use in the *L. japonicus HAR1* editing experiment because STU was shown to be more effective at inducing genomic mutations as compared to the dual promoter system. I had surmised that by employing both systems the likelihood of discovering successful gene edits would be increased.

The first eight sgRNAs, including their respective glycine pre-tRNAs, were successfully amplified using the *CRISPR-HAR1p-1* vector as the DNA template. The resulting multiplex sgRNA expression cassette, containing eight sgRNAs, was then introduced into the pGEL032 binary vector (see *Appendix C* for the vector map) using the Golden Gate assembly (Engler et. al., 2008). The Golden Gate assembly method allowed for the combined restriction digestion and ligation in a single reaction (see Materials and Methods). As a result, the *CRISPR-HAR1p-2* vector was developed (Figure 12B).

The integrity of the recombined binary vector was confirmed using a *HpaI*/*HindIII* double digestion (Figure 14). The vector was also sequenced, and no mutations were detected (data not shown). The *CRISPR-HAR1p-2* vector was used in subsequent experiments.

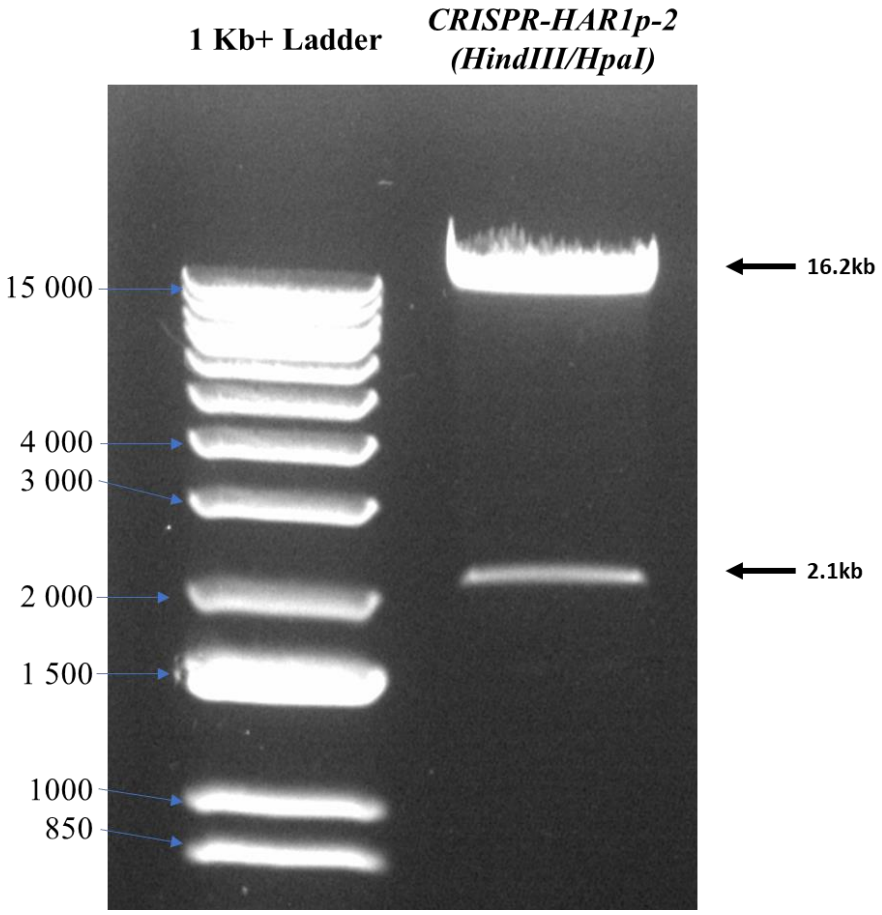


Figure 13. Confirmatory, restriction enzyme analysis of the *CRISPR-HAR1p-2* vector. An image of the ethidium bromide stained agarose gel, showing the banding pattern resulting from a double digestion using the *HpaI* and *HindIII* restriction enzymes. The expected bands were predicted to be 16.2kb and 2.1kb in size (arrows).

6. The CRISPR/Cas12a multiplex sgRNA expression system and the binary vector for editing *in-planta* using a Cas12a single transcriptional unit (STU) system.

In order to increase the number of potential target sites for the *HAR1* promoter editing experiments, and thus increase the likelihood of obtaining a broader range of mutations, additional constructs were designed to preform genome editing using the Cas12a enzyme. The Cas12a enzyme uses a different PAM sequence than Cas9 (that is TTTV instead of NGG), which was predicted to significantly extend the range of selected targets within the AT-rich *HAR1* promoter. The CRISPR-P v2.0 software, used to identify Cas9 target sites, had identified potential Cas12a target sites in the *L. japonicus HAR1* promoter. As with Cas9, the gRNAs used by Cas12a may target multiple genomic regions due to sequence identity. As such the constructs using Cas12a omitted any potential target sites that had true off-targets (see above). To narrow down the target sites which would be used, the predicted editing efficiency of each of the associated crRNAs was assessed. That is, the CRISPR-P V2.0 software predicted the likelihood of editing occurring given the sequence of each of the target sites. For each target site, the software provided an "on-score" between 0 and 1 that represented the likelihood of editing occurring at the site. A higher on-score means there is a greater likelihood for editing to occur, as predicted by the software. To be chosen, a target site had to contain an on-score of at least 0.6 or greater. Ten different target sites were eventually selected within the *L. japonicus HAR1* promoter for genome editing with Cas12a (Figure 11, *Appendix A*).

The binary vector, pGEL032 (see *Appendix C* for the vector map), contains the Cas12a STU expression system driven by the *Zea mays UBIQUITIN* (*ZmUBQ*) promoter. To my knowledge, the effectiveness of generating DNA edits in *L. japonicus* using this promoter does not appear to be published at the time of writing. To this end two alternative binary vectors,

carrying the *L. japonicus LjUBQ* and 2x35S promoters, were developed by replacing the original *ZmUBQ* promoter (see Material and Methods). The latter promoter apparently allows for high levels of expression and is also known to be active in the phloem tissue of dicotyledonous plants (Benyon et. al., 2013), where the *HAR1* gene is normally expressed. The expression cassettes were designed to include *BsaI* recognition sequences complimentary to those in the pGEL032 binary vector.

The structure of the mature crRNA used in Cas12a mediated genome editing contains a short direct repeat, instead of tracrRNA sequence, flanking it at the 3' end. These direct repeats allow for processing of the transcribed RNA by the Cas12a enzyme, as well as allosterically activating the Cas12a enzyme (Tang et. al., 2017). Five cassettes, encompassing all ten selected Cas12a target sites, were synthesized, each designed to drive expression of two gRNAs. They were constructed to be introduced directly into the pGEL032 binary vectors using Golden Gate assembly. The five constructs were individually assembled into the pGEL032a (2x35S promoter variant) and pGEL032b (*LjUBQ* promoter variant) binary vectors, resulting in ten unique duplex guide RNA expression cassettes (Figure 15A). The completed vectors were herein called *CRISPR-HAR1p-$_{2x35S}$1,2, CRISPR-HAR1p-$_{2x35S}$3,4, CRISPR-HAR1p-$_{2x35S}$5,6, CRISPR-HAR1p-$_{2x35S}$7,8,* and *CRISPR-HAR1p-$_{2x35S}$9,10, CRISPR-HAR1p-$_{LjUBQp}$1,2, CRISPR-HAR1p-$_{LjUBQp}$3,4, CRISPR-HAR1p-$_{LjUBQp}$5,6, CRISPR-HAR1p-$_{LjUBQp}$7,8,* and *CRISPR-HAR1p-$_{LjUBQp}$9-10,* respectively (Figure 15A).

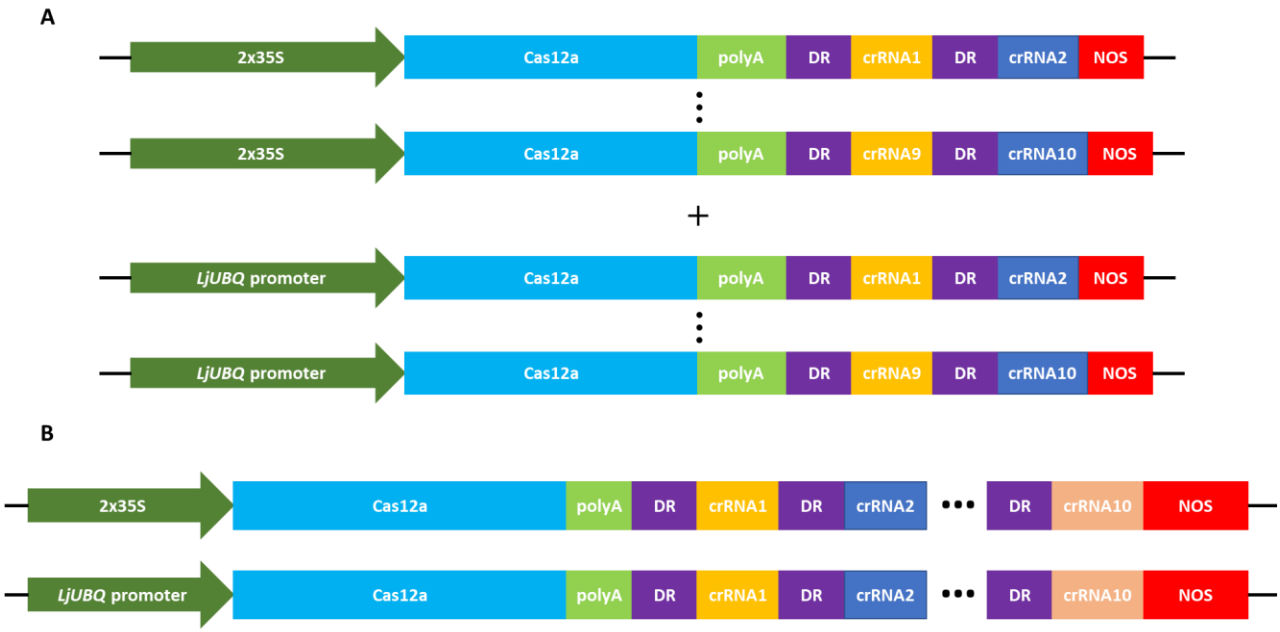


Figure 14. The structure of the Cas12a crRNA STU expression cassettes. (A) A schematic representation of *CRISPR-HAR1p-*$_{2x35S}$*1,2*, *CRISPR-HAR1p-*$_{2x35S}$*9,10*, *CRISPR-HAR1p-*$_{LjUBQp}$*1,2*, and *CRISPR-HAR1p-*$_{LjUBQp}$*9,10*. The cassettes start with the *LbCas12a* gene, followed by the polyA tail. The crRNAs, flanked by direct repeats (DR) on both 5' and 3' ends, are directly adjacent to the polyA tail. All constructs contain the NOS terminator. **(B)** The structure of the multiplex Cas12a expression cassettes containing all ten crRNAs.

In a separate experiment, two multiplex crRNA expression cassettes containing all ten crRNAs (the same as those used in the duplex constructs) were developed (Figure 15B) in the pGEL032a and pGEL032b plasmids using Golden Gate assembly, respectively. The completed vectors were herein called *CRISPR-HAR1p-*$_{2x35S}$*multiplex* and *CRISPR-HAR1p-*$_{LjUBQp}$*multiplex.* The integrity of all Cas12a STU expression vectors still has to be confirmed using a *BglII*/*XbaI* double digestion, and sequencing.

7. *Transforming CRISPR-HAR1p-1* and *CRISPR-HAR1p-2* vectors into *A. rhizogenes*.

In order to test the editing activity of the *CRISPR-HAR1p-1* and *CRISPR-HAR1p-2* vectors, the hairy root approach was employed (Guillon et. al., 2008). *A. rhizogenes* strain 1193 was individually transformed with the *CRISPR-HAR1p-1* and *CRISPR-HAR1p-2* vectors. Both the *CRISPR-HAR1p-1* and *CRISPR-HAR1p-2* vectors were re-isolated from *A. rhizogenes* and transformed back to *E.coli* to check for their integrity using the same diagnostic restriction enzyme digestions as described above. The two vectors were confirmed to have correct sequence (data not shown) and thus the transformed *A. rhizogenes* strains were deemed ready for use in hairy root transformation experiments.

8. Generating hairy roots and developing immortal cultures.

L. japonicus seedlings with elongated hypocotyls were used for plant transformation (see Material and Methods). The seedlings were inoculated with *A. rhizogenes* strain 1193 carrying either *CRISPR-HAR1p-1* or *CRISPR-HAR1p-2* vectors. Transgenic hairy roots formed on the non-transgenic shoots. These roots were then used to test the processing and genome editing capabilities of the sgRNA expression cassettes.

Individual hairy roots were removed from the shoots and were used to establish individual immortal root tissue cultures (Figure 16). They were cultured in the presence of hygromycin B to select for hairy roots that carried the T-DNA of the binary, *CRISPR-HAR1p-1* or *CRISPR-HAR1p-2*, vectors. The roots which did not contain the appropriate T-DNA did not grow and were quickly killed by hygromycin B, appearing bleached. Approximately 50% of the hairy roots survived and were cultured indefinitely, as long as they were needed for downstream applications.

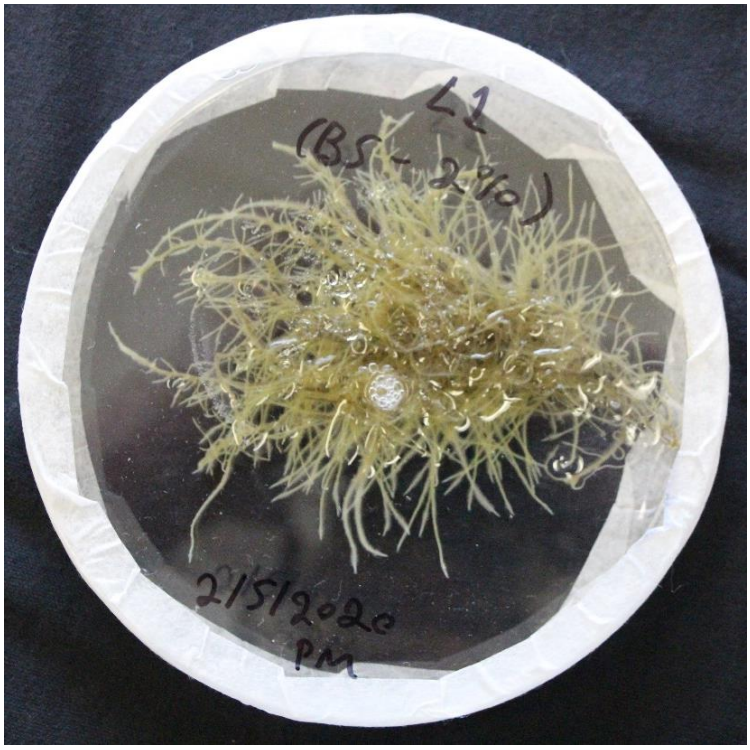


Figure 15. An example of an immortal *L. japonicus* hairy root culture. The hairy roots were able to grow indefinitely in a petri dish (i.e. they are immortal) and could be easily propagated (for more details see main text).

9. Transgenic hairy roots were genotyped for the presence of the dual promoter and STU systems.

Once the hairy roots were formed, each individual primary root was considered a unique transformation event. The T-DNA found in each transformation event may differ from the rest of the events in the same experiment by the site of genomic integration. In some cases, the T-DNA may be inserted without the full sequence being incorporated into the host's genome. This means that some of the roots that are selected for by hygromycin B may not contain the sgRNA cassette

or contain only a portion of the cassette. Confirmation of the presence of the sgRNA cassettes in the genomic DNA of hygromycin B resistant hairy roots was completed using PCR genotyping.

Genomic DNA was isolated from 23 independent hairy root cultures, as generated using *A. rhizogenes* carrying the T-DNA from the *CRISPR-HAR1p-1* vector. PCR based genotyping was conducted using the genomic DNAs as templates, targeting the *LjU6* promoter and the fifth sgRNA from the inserted multiplex sgRNA cassette. The presence of the transgene was confirmed in 22 of the 23 cultures, as indicated by the presence of the 950bp DNA fragment (Figure 17). It is possible that lack of a DNA band in one sample (lane 13) is the result of a pipetting error. As this sample had survived hygromycin B selection, it may contain the T-DNA. The PCR was not repeated because a sufficient number of samples were confirmed positive for the presence of the T-DNA, allowing continuation of experiments.

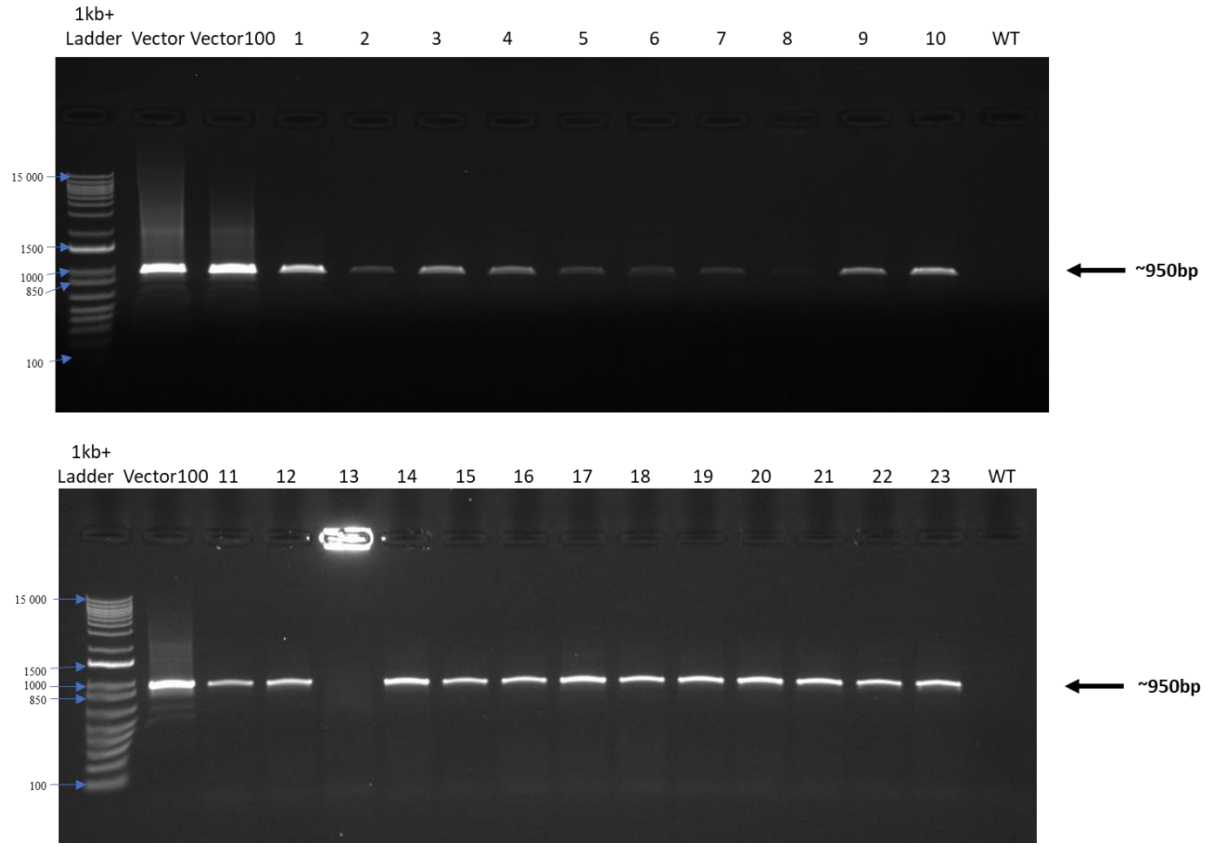


Figure 16. PCR-based genotyping of *L. japonicus* transgenic hairy roots for the presence of the *CRISPR-HAR1p-1* T-DNA. An image of the ethidium bromide stained agarose gel, showing the DNA band of approximately 950bp in size (arrow) indicates presence of the sgRNA cassette. The *CRISPR-HAR1p-1* binary vector was used as a positive control (where the lane marked 'Vector' reflects use of 200ng of the vector as the PCR template and the lanes marked 'Vector100' only 2 ng of the same vector. The wild-type *L. japonicus* (Gifu) DNA was used as a negative control.

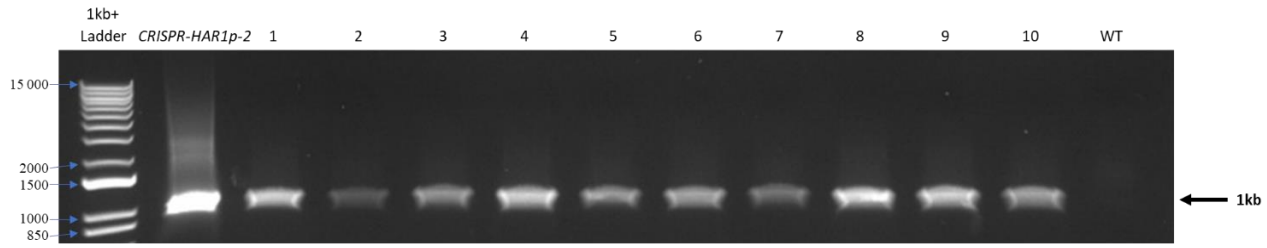


Figure 17. PCR-based genotyping of *L. japonicus* transgenic hairy roots for the presence of the *CRISPR-HAR1p-2* T-DNA. An image of the ethidium bromide stained agarose gel, showing the DNA band of approximately 1kb in size (arrow) indicates presence of the sgRNA cassette. The *CRISPR-HAR1p-2* binary vector was used as a positive control and wild-type *L. japonicus* (Gifu) DNA was used as a negative control.

The same genotyping scheme was applied to ten hairy root cultures derived from the *CRISPR-HAR1p-2* vector transformation experiment. PCR was conducted using a different forward primer (*Appendix B*) than the genotyping conducted on the hairy roots derived from the *CRISPR-HAR1p-1* transformation experiment, with the same reverse primer. In this case the forward primer targeted the 3' end of the Cas9 gene. The expected DNA fragment size was 1kb for this experiment, and the presence of the transgene was confirmed in all ten cases (Figure 18).

10. Expression and maturation of the mRNA encoding the *CRISPR-HAR1p-1* derived multiplex sgRNA cassette.

Having confirmed that the T-DNA was successfully incorporated into the genome of the hairy roots, the next step was to check whether the transcript was present and was being processed.

For a multiplex sgRNA construct to be functional in genome editing *in planta*, it needs to be expressed and processed. The transcription of the construct gives rise to a polycistronic mRNA that must be processed to produce individual sgRNA molecules. To test the expression and processing of the multiplex cassette, the RNA content of the samples was analyzed using the circular (c)RT-PCR approach.

The multiplex sgRNA expression cassette in the *CRISPR-HAR1p-1* vector was designed so that the polycistronic RNA would be processed by the endogenous tRNA maturation machinery (Xie et. al., 2015). To verify that this was indeed happening, total RNA derived from the cultured hairy roots was analyzed. If the multiplex sgRNA cassette was being transcribed, it would have to be detectable using the cRT-PCR. The advantage of cRT-PCR in comparison to a regular RT-PCR procedure is that the former allows for detection of both unprocessed and mature sgRNA-species, hence addressing simultaneously both the transcription and processing of multiplex gRNA constructs. If the cassette is transcribed and post-transcriptional processing occurs, individual sgRNAs of 100 nucleotides in length should be detected (Figure 19).

By designing primers specific to individual sgRNAs, it was possible to conduct cRT-PCR to verify that both transcription and processing were occurring. This procedure made it possible also to distinguish between the presence of unprocessed transcript, the result of transcription activity of the transgene, and mature sgRNAs, the outcome of successful transcription and processing (Figure 19).

Three transgenic hairy root cultures derived from the *CRISPR-HAR1p-1* transformation experiment were used in this analysis. The total RNA isolated from hairy root cultures was circularized using T4 RNA ligase. The cRT step was individually conducted using the primers to recognize four different sgRNA species, namely sgRNA1, sgRNA4, sgRNA5, and sgRNA8. The

resulting single strand products were subsequently used as templates for PCR amplification using the primer targeting the tracer portion that was common to all eight gRNAs in the *CRISPR-HAR1p-1* sgRNA multiplex cassette. The resulting PCR products were resolved on agarose gels to visualise the DNA bands and were also cloned into the pCR-Blunt II-TOPO™ vector and sent for sequencing to confirm their identities (Figures 20 and 21). Using sgRNA1 as an example, four different cDNA products were detected. These included a fully processed sgRNA1, an unprocessed fragment encompassing sgRNA1, a portion of sgRNA2 and the intervening $tRNA_{Gly}$ sequence, ligation artefacts and finally, a non-specific PCR product. This outcome varied depending on the targeted sgRNA (Figures 20 and 21) (see Discussion for further details).

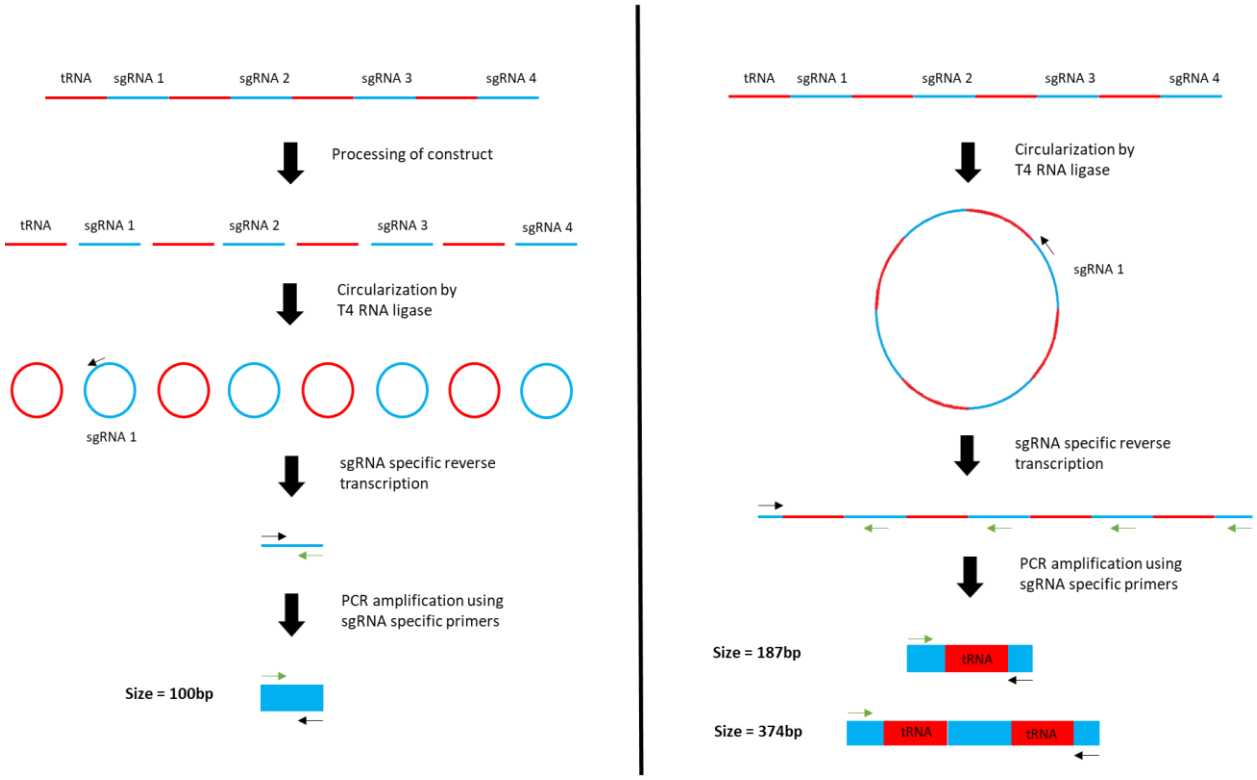


Figure 18. A schematic of different outcomes for the cRT-PCR experiments. The left panel represents the workflow and outcome if the multiplexed sgRNA cassette is fully processed and only mature sgRNAs, 100bp-long products, are detected. The right panel represents the workflow and the outcome if the cassette is not being processed. The entire primary transcript is circularized, and the presence of one or more sgRNAs in a single PCR product, together with the tRNA sequence(s) is detected.

A

sgRNA1

PCR Product	Size (bp)	Number of clones containing PCR product	Schematic of the PCR Product
Mature sgRNA 1	100	4	
Unprocessed sgRNA 1	187	3	tRNA
Ligation Artifact	287	4	tRNA
Non-Specific Amplification	183	1	

B

sgRNA4

PCR Product	Size (bp)	Number of clones containing PCR product	Schematic of the PCR Product
Mature sgRNA 4	100	1	
Unprocessed sgRNA 4	187	2	tRNA
Non-Specific Amplification	50 - 185	3	

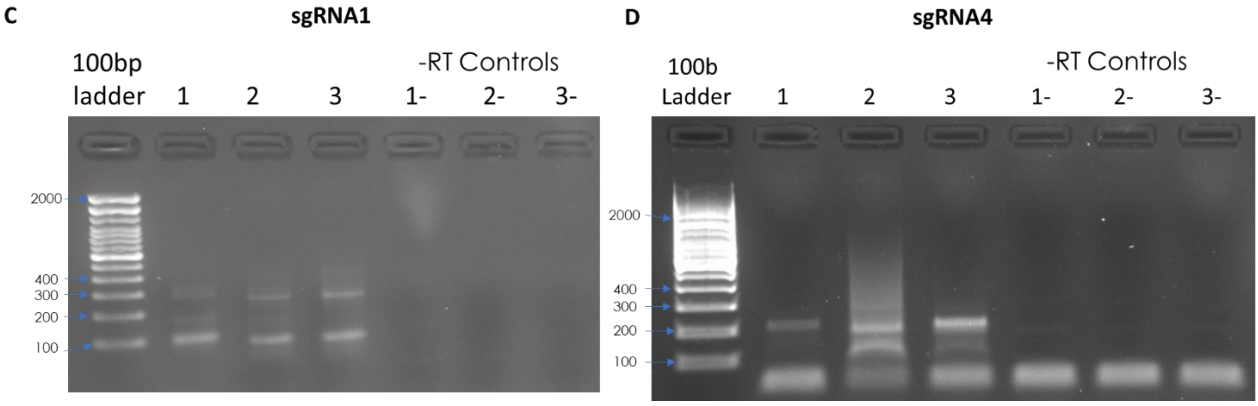


Figure 19. Outcomes of the cRT-PCR experiments conducted on the *CRISPR-HAR1p-1* hairy roots, using primers specific to sgRNA1 and sgRNA4, respectively. (A) Shows the frequency of different PCR products and their structures using the sgRNA1 specific primer in the reverse-transcription (RT) step. **(B)** Shows the results of a similar experiment using the primer specific to sgRNA4 in the RT step. **(C and D)** The images of the ethidium bromide stained agarose gels corresponding to A and B above, respectively. 1, 2 and 3 refer to three independent hairy root cultures.

sgRNA5

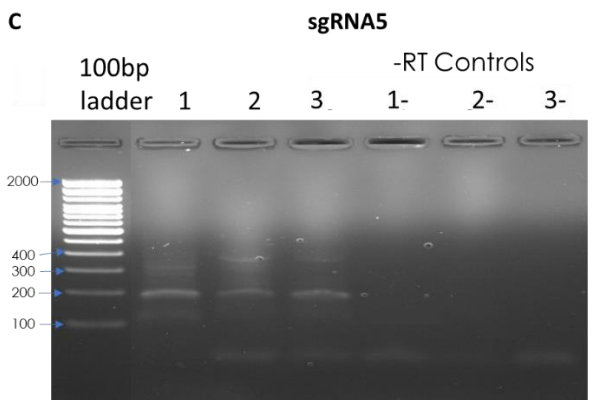
A

PCR Product	Size (bp)	Number of clones containing PCR product	Schematic of the PCR Product
Mature sgRNA 5	100	4	
Unprocessed sgRNA 5	187	4	tRNA

sgRNA8

B

PCR Product	Size (bp)	Number of clones containing PCR product	Schematic of the PCR Product
Mature sgRNA 8	100	2	
Non-Specific Amplification	50 - 185	4	



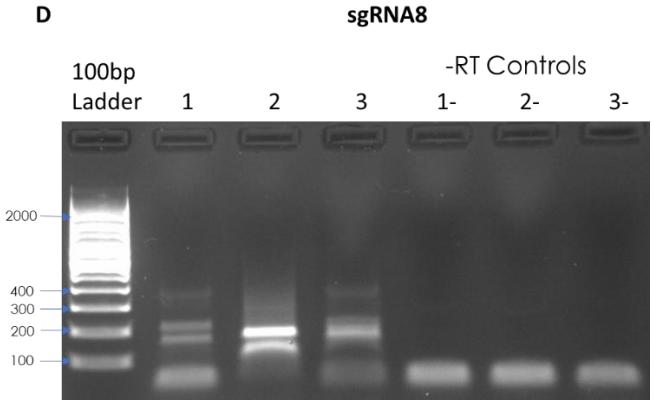


Figure 20. Outcomes of the cRT-PCR experiments conducted on the *CRISPR-HAR1p-1* hairy roots, using primers specific to sgRNA5 and sgRNA8, respectively. (A) Shows the frequency of different PCR products and their structures using the sgRNA5 specific primer in the reverse-transcription (RT) step. **(B)** Shows the results of a similar experiment using the primer specific to sgRNA8 in the RT step. **(C and D)** The images of the ethidium bromide stained agarose gels corresponding to A and B above, respectively. 1, 2 and 3 refer to three independent hairy root cultures.

11. Expression and maturation of the mRNA encoding the *CRISPR-HAR1p-2* derived multiplex sgRNA cassette.

The *CRISPR-HAR1p-2* experiment had the same requirement for post-transcriptional processing of the corresponding polycistronic transcript as with *CRISPR-HAR1p-1*. A similar cRT-PCR protocol was applied using total RNA isolated from three transgenic hairy root cultures that were derived from the *CRISPR-HAR1p-2* transformation experiment. In this case, the cRT step was conducted by targeting three sgRNAs: sgRNA1, sgRNA4 and sgRNA8.

The fragments generated using the sgRNA1 specific primer were sequenced and showed instances of both processed and unprocessed sgRNAs (Figure 22A and D), in addition to non-specific amplification products. Fragments generated by using the primers specific to sgRNA4 and sgRNA8, respectively, were also sent for sequencing but most of them failed, possibly due to poor template quality (Figure 22B, C, E and F). The goal was to have the sequence of at least ten individual fragments for each of the sgRNA that were being targeted; however, the experiment could not be repeated due to time constraints imposed by the COVID-19 pandemic-related shut downs. As a result, the maturation efficiencies of the sgRNA cassette derived from the *CRISPR-HAR1p-2* vector were defined based only on the analysis of sgRNA1, which was at 50%.

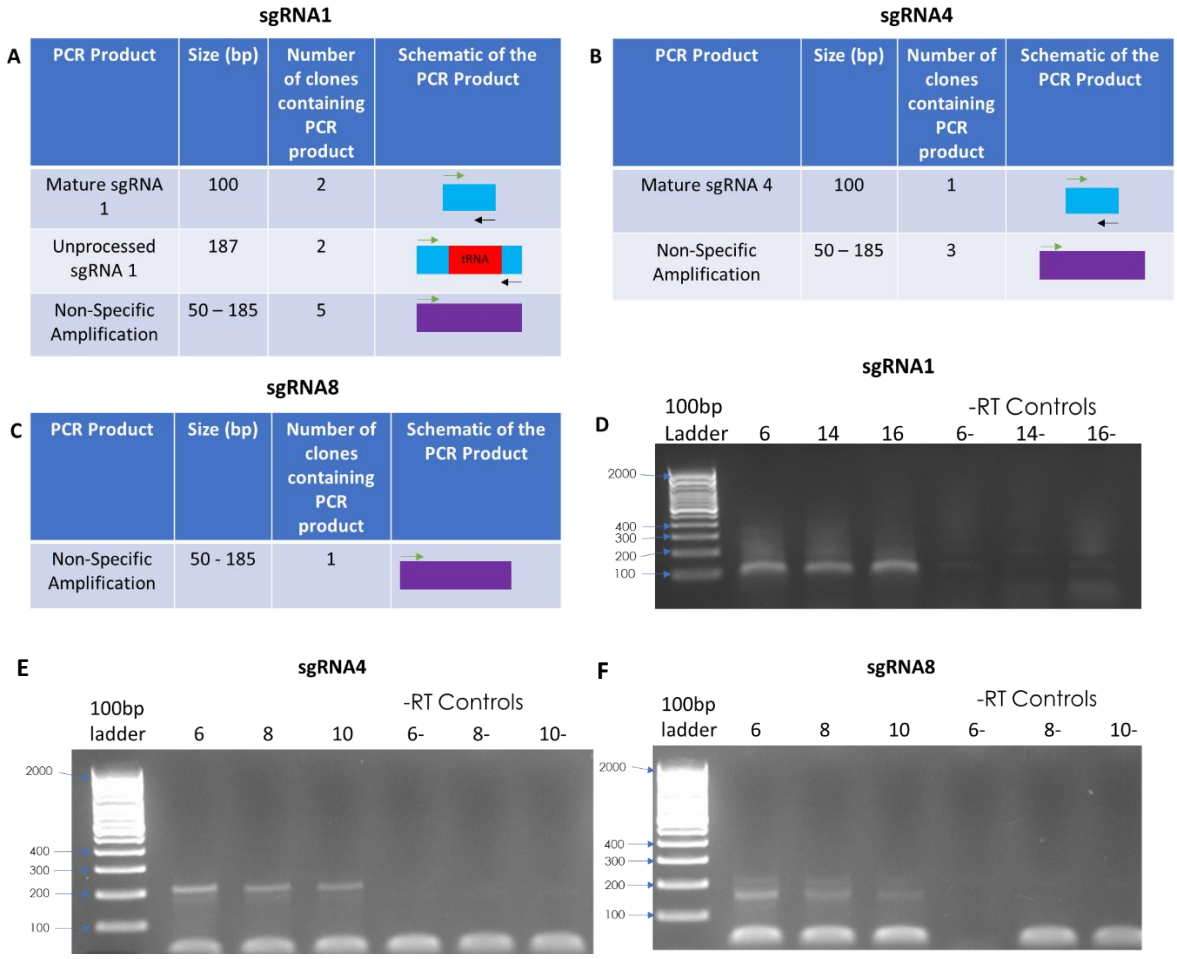


Figure 21. Outcomes of the cRT-PCR experiments conducted on the *CRISPR-HAR1p-2* hairy roots, using primers specific to sgRNA1, sgRNA4, and sgRNA8, respectively. (A) Shows the frequency of different PCR products and their structures using the sgRNA1 specific primer in the reverse-transcription (RT) step. **(B)** Shows the results of a similar experiment using the primer specific to sgRNA4 in the RT step. **(C)** shows the results of a similar experiment using the primer specific to sgRNA8 in the RT step. **(D, E, F)** The images of the ethidium bromide stained agarose gels corresponding to A, B, and C above, respectively. The line numbers correspond to specific hairy root cultures.

12. The *CRISPR-HAR1p-1* and *CRISPR-HAR1p-2* transgenic hairy root genomic DNAs were analyzed for signs of gene editing at the target locus.

Having confirmed the expression and at least partial maturation of sgRNA species in the roots derived from the *CRISPR-HAR1p-1* and *CRISPR-HAR1p-2* transformation experiments, it was also imperative to determine editing efficiency of the cassettes. Each individual sgRNA present in the cassette has the potential to induce a mutation, SNP or small deletion, at its target site. However, since there were several sgRNAs utilized simultaneously, there was also an expectation of possibly generating large deletions when at least two guides generated a double stranded break. To determine whether editing had occurred, the *HAR1* promoter region was PCR amplified from a selection of transgenic hairy roots that were analyzed for the presence of the CRISPR cassettes, and the resulting DNA fragments were evaluated by gel electrophoresis and sequencing.

The gene specific primers, *HAR1p-F* and *HAR1p-mid-R*, encompassing 2.1kb of the 4kb *HAR1* promoter region, were used. The 2.1kb fragment generated by this PCR amplification contains eight of the nine target sites. The size of DNA fragments generated by the PCR amplification in all experimental samples were the same as the wild type control (Figure 23). The resulting DNA fragments were cloned into the pCR-Blunt II-TOPO™ vector and were sequenced. No mutations, including SNPs, were detected (data not shown). To check for mutations at the remaining target site, PCR amplification using the primers *HAR1p-mid-F* and *HAR1p-R* was also conducted. These primers cover a genomic region that encompassed the final target site and overlaps with the first PCR amplification region, resulting in a wild type fragment of 1.6kb. No mutations were detected by PCR amplification of this *HAR1* promoter region in the *CRISPR-HAR1p-1* transgenic hairy roots (data not shown).

The editing efficiency of the *CRISPR-HAR1p-2* derived multiplex sgRNA cassette was also tested using the same PCR procedure as above. Upon PCR amplification of genomic DNA samples with *HAR1p-F* and *HAR1p-mid-R*, all of the transformed hairy root cultures tested yielded DNA fragments that appeared to be the same size as the wild type (non-transgenic) control (Figure 24). PCR amplification using the primers *HAR1p-mid-F* and *HAR1p-R* resulted in DNA fragments that looked wild type in all the samples tested, indicating that no major deletions were present (data not shown). Due to time limitation, these samples were not sequenced.

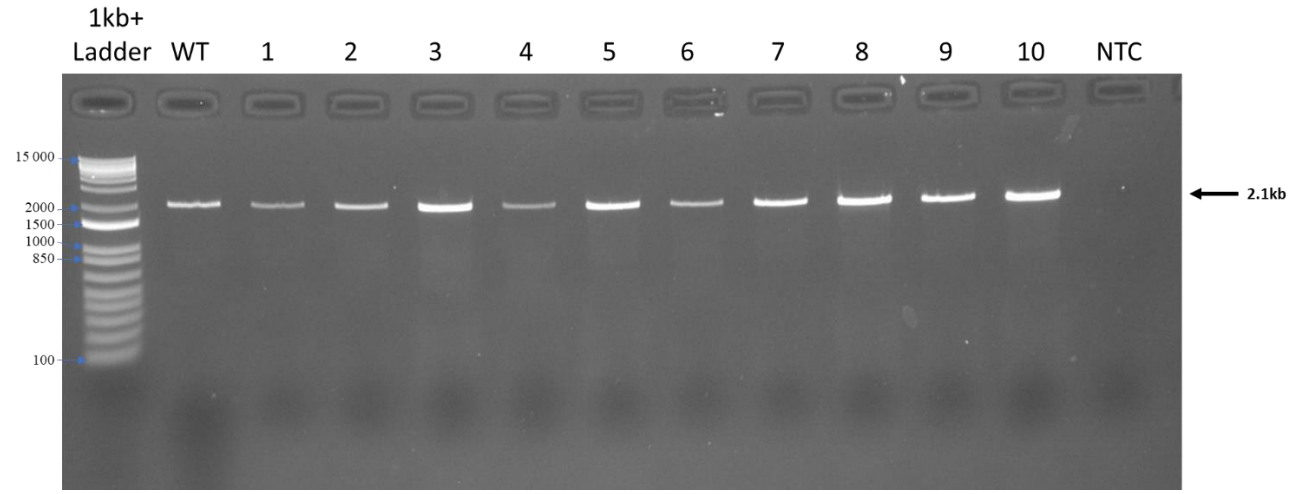


Figure 22. The check for genome modifications in the *HAR1* promoter region from ten (1-10) different hairy roots carrying the *CRISPR-HAR1p-1* sgRNA cassette. An image of the ethidium bromide stained agarose gel, showing the PCR amplification products of the *HAR1* promoter from different samples. Wild-type (WT) *Lotus japonicus* roots were used as a positive control. The amplicons were sequenced to detect any mutations. A negative, no template control (NTC), was used. Expected size of the wild-type PCR product is 2.1 kb.

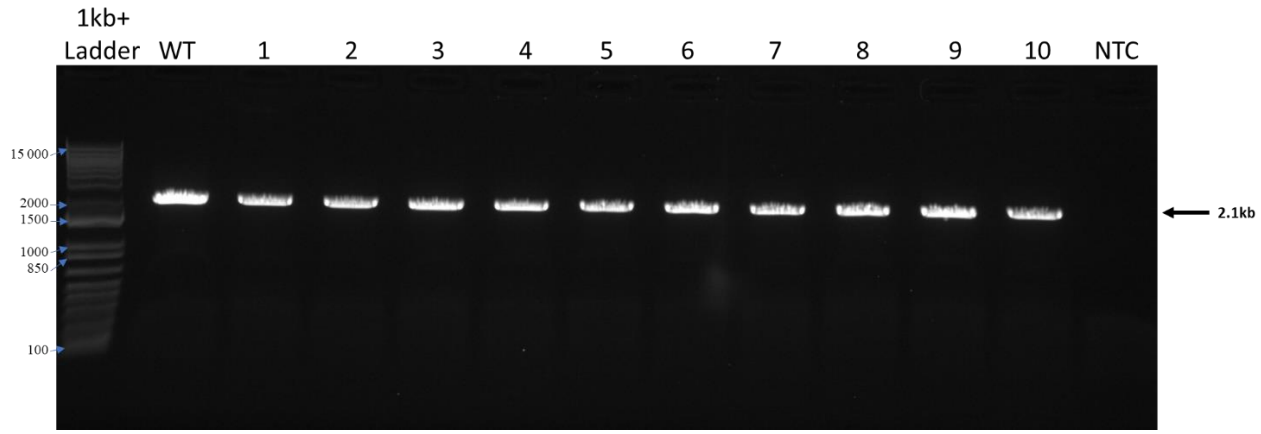


Figure 23. The check for genome modifications in the *HAR1* promoter region from ten (1-10) different hairy roots carrying the *CRISPR-HAR1p-2* sgRNA cassette. An image of the ethidium bromide stained agarose gel, showing the PCR amplification products of the *HAR1* promoter from different samples. Wild-type (WT) *Lotus japonicus* roots were used as a positive control. The amplicons were sequenced to detect any mutations. A negative, no template control (NTC), was used. Expected size of the wild-type PCR product is 2.1 kb.

13. Sequence analysis of the *L. japonicus HAR1* promoter region for conservation with its orthologues and for *cis*-acting regulatory elements.

Alterations to the *HAR1* promoter sequence, as intended in the long term by this work, are expected to impact the quantitative aspect of the gene functioning. In this context, *in-silico* identification of conserved regions of the promoter across species and the prediction of *cis*-acting elements was deemed necessary. The *G. max NARK* (Kinkema & Gresshoff, 2008), *G. max CLV1A* (Mirzaei et. al., 2017), and *M. truncatula SUNN* (Schnabel et. al., 2005) genes were identified as orthologues to the *L. japonicus HAR1* gene (Figure 9). The promoter regions of the three

orthologues were analyzed for DNA sequence conservation with the *L. japonicus HAR1* promoter region (Figure 25). Two regions that show at least 70% identity in the overlapping 100bp sliding windows (see Material and Methods) were identified. These regions, highlighted in red in Figure 25, are present in all three predicted orthologues, but not in the *L. japonicus Symbiosis Receptor Kinase (SYMRK)* gene (Stracke et. al., 2002), the sequence of which was used as the outgroup for the comparison.

To predict what transcription factor binding sites (*cis*-acting regulatory elements) were present within the promoter region of the *L. japonicus HAR1* gene, its sequence was analyzed in silico using the PlantCARE software (http://bioinformatics.psb.ugent.be/webtools/plantcare/html/) (Figure 26). Several different cis-acting elements were predicted, including seven elements involved in light response. The *cis*-acting elements that were predicted to be present in the two conserved regions described above included a TC rich repeat, an MBS motif, an ABRE motif, a CAT box, and a GT1 motif (Figure 25 and Figure 26). The TATA-box was omitted from the analysis results as it was too common and overwhelmed the other predicted elements.

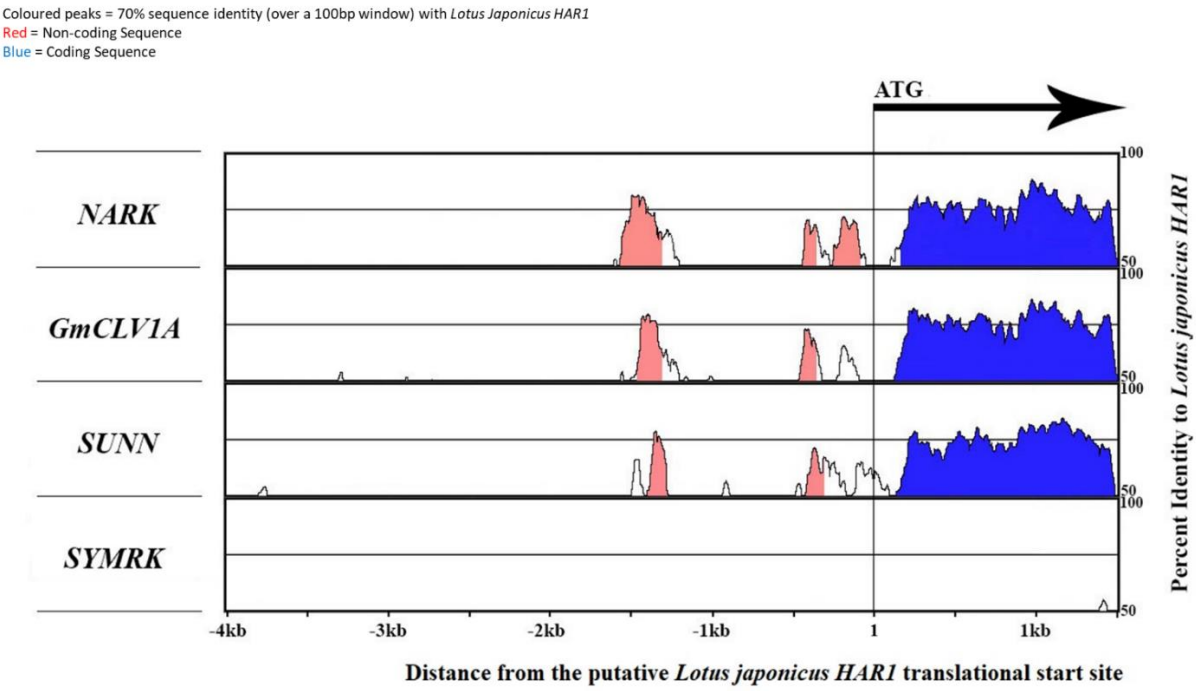


Figure 24. **A graph showing regions of sequence conservation between *L. japonicus HAR1* and its orthologous loci in *Glycine max* and *Medicago truncatula.*** *NARK* and *GmCLV1A*: *NODULATION RECEPTOR KINASE* and its paralogue, respectively, in *G. max*. *SUNN*: *M. truncatula SUPER NUMERIC NODULES.* The *Lotus japonicus SYMBIOSIS RECEPTOR KINASE* (SYMRK) gene was used as an outgroup. The x-axis represents the position corresponding to the *HAR1* locus, where A in the predicted ATG initiation start codon is 1. The peaks represent sequences, that share at least 70% nucleotide sequence identity with *HAR1*, over a 100bp window. Peaks that are coloured red and blue represent non-coding and coding sequences, respectively. White peaks represent regions that contained more than 50% sequence identity with *HAR1*, but not more than 70%. The analysis was performed using the mVista web program (https://genome.lbl.gov/vista/mvista/submit.shtml).

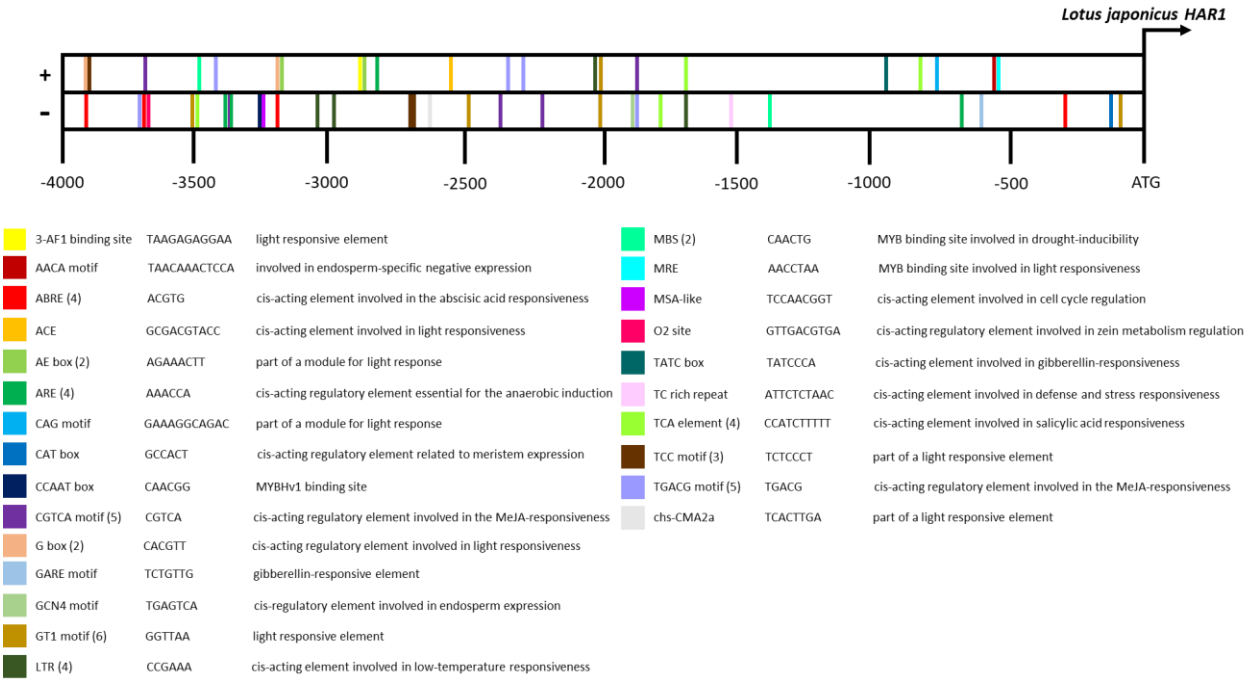


Figure 25. The *cis*-regulatory elements found within the *Lotus japonicus HAR1* promoter. The plus and the minus strands are both depicted. The *cis*-regulatory elements are shown in the strand where they exist. Different coloured regions represent different regulatory elements. To generate this figure the sequence of the *HAR1* promoter was analyzed using the PlantCARE software: http://bioinformatics.psb.ugent.be/webtools/plantcare/html/.

14. Natural variation in the *Glycine max NARK* locus was evaluated across 350 varieties of cultivated soybean.

Similar to the identification of conserved sequences and prediction of *cis*-acting elements, it was also of interest to determine natural variation. This analysis was accomplished using the *G. max NARK* locus, a soybean orthologue of *L. japonicus HAR1*, through a publicly accessible resource that includes 350 soybean accession lines for which genome-wide polymorphism is known (https://soykb.org/public_data.php). The polymorphism, which included all SNPs, was evaluated using the *G. max* Williams 82 *NARK* gene region (Chromosome 12: 2881838 - 2890668; Glyma12G040000; *Glycine max* Wm82.a2.v1) and the resulting data were summarized in a histogram (Figure 27). A simple statistical analysis was used to determine the frequency and location (coding vs. non-coding regions of the gene) of SNPs. In this analysis the coding region was defined as all exons while the non-coding regions included the promoter, the intron, and both the 5' and 3' UTRs. Any soybean variety was considered polymorphic if it had any nucleotide difference from the reference sequence. The number of varieties carrying polymorphisms at every given position as compared to the reference were also calculated. The total tally of polymorphic lines at each given position was used to generate the histogram shown in Figure 27. Polymorphic positions across the entire locus were added and then divided by the total number of positions analyzed, which generated a polymorphism coefficient, representing the number of polymorphisms per base pair. Similar calculations were also carried out for coding and non-coding regions. Dividing the number 1 by the resulting coefficients gave the average number of positions that were polymorphic within the respective regions. On average, 1 in every 86 positions was found to be polymorphic when the whole *GmNARK* locus was analyzed (Table 1). In contrast,

SNPs in the coding regions of *GmNARK* appeared with an average frequency of 1 in every 72 positions and 1 in every 94 positions in the non-coding regions (Table 1).

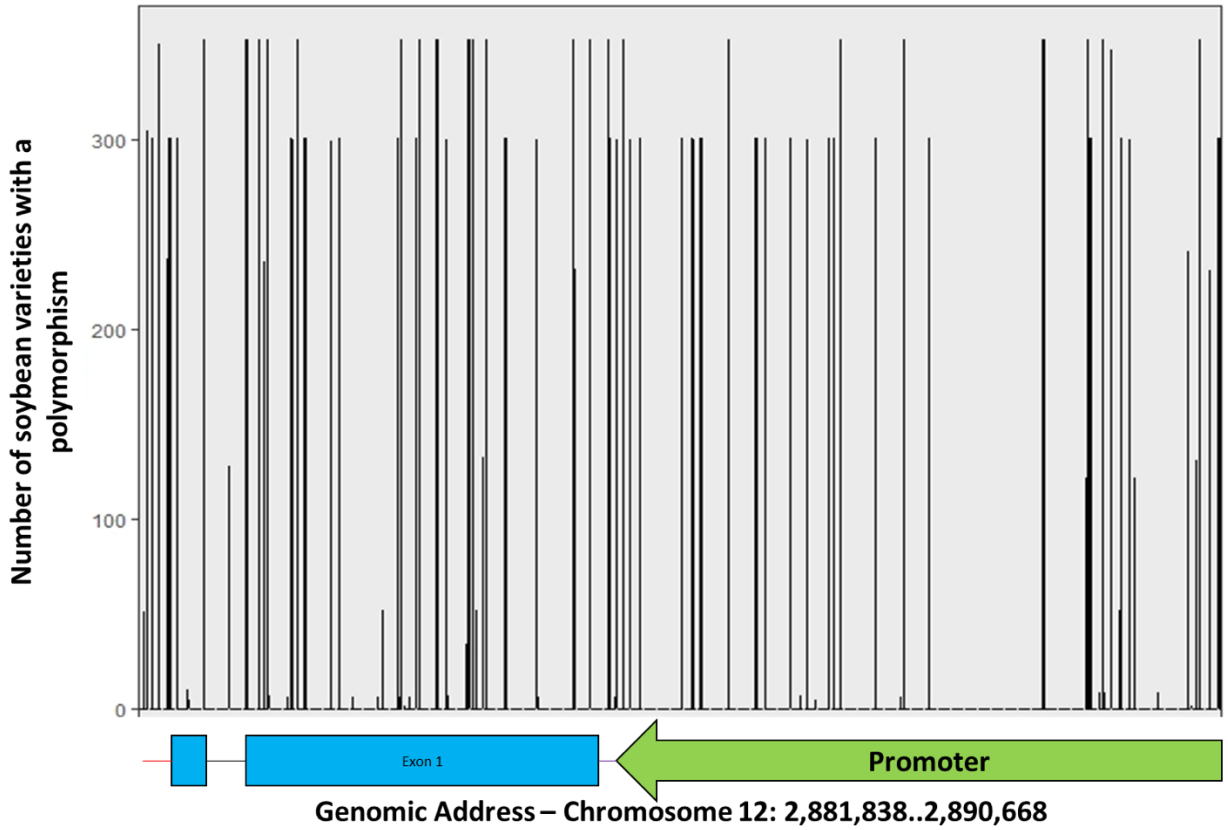


Figure 26. A histogram showing the natural SNP variation at the *Glycine max NARK* locus. The *NARK* sequences from 350 different soybean varieties were used to generate the graph using the soykb.org website and a bioinformatics pipeline developed for this thesis (see main text for more detail). Each position on the x-axis represents individual nucleotide positions within the locus, and the bars correspond to the number of varieties that carry an SNP at a given position, as compared to the Williams 82 reference sequence (*Glycine max* Wm82.a2.v1).

Region	Ratio of Polymorphic Positions in *GmNARK*	Ratio of Polymorphic Positions in *GmGBP1* (Zhao et. al., 2018)
Whole Locus	1/86	1/106
Coding	1/72	1/153
Non-Coding	1/94	1/75

Table 1. The natural variation that exists in the *Glycine max NARK* locus compared to the natural variation in the *Glycine max GBP1* locus. The ratios of polymorphic positions are shown in the second column. The last column shows the ratios for the *GmGBP1* gene obtained from literature (see Discussion).

Chapter 4: Discussion

1. A brief summary of the obtained results.

The primary goal of my thesis work was to employ CRISPR/Cas technologies to induce quantitative trait variation at the *L. japonicus HAR1* locus. *HAR1* has a dose-dependent role in regulating the extent of nodulation, an important homeostatic process in the plant nitrogen economy. It also regulates several other aspects of the plant N and P nutrition and hence represents an interesting target for improvement of nutrient uptake/use efficiency in legumes and possibly also non-legume plants.

To my knowledge, targeting regulatory gene regions to obtain a quantitative trait variation has yet to be demonstrated in any legume species. In order to advance toward this goal, I have developed several gene editing constructs using publicly available CRISPR/Cas9 and CRISPR/Cas12a genome editing modules (Liu et. al., 2017). The on-line protocol for the selection of target sites was applied to the *HAR1* locus and both the double promoter and single transcription unit CRISPR/Cas constructs were developed based on these selected targets (for further discussion of this aspect of the work see below). The binary vector pGEL032, used at the plant transformation step, has been modified in order to expand on the repertoire of promoter sequences driving the expression of the multiplex gRNA cassettes and the *Cas12a* gene. This was done to guide and possibly enhance the editing efficiency by creating a better alignment with the *L. japonicus* genetic background (i.e. by using the *L. japonicus* ubiquitin promoter sequence) and also by taking into consideration the tissue (phloem) specificity of *HAR1* gene expression. A methodology to clone gRNA cassettes into three different destination vectors is described in this thesis, encompassing

the use of Golden Gate cloning, a protocol that allows for several DNA fragments to be joined in a single reaction (Engler et. al., 2008).

The hairy root system was used to evaluate a selection of the gRNA-CRISPR/Cas9 constructs by testing their expression and editing efficiencies. While hairy roots have routinely been used to perform various molecular analyses in the Szczyglowski laboratory (Hossain et. al., 2012; Shrestha et. al., 2021), none of my predecessors had established long-term, axenic cultures. I describe in the thesis the culturing conditions that facilitate the growth of *L. japonicus* hairy roots as long-term, immortal tissue cultures (see Materials and Methods – Section 5; and Figure 16). These cultures were established for the purpose of rapidly testing the expression, maturation and editing efficiency of the CRISPR/Cas constructs, before their use in fully transgenic plants.

Admittedly, the work has not been finalized and follow-up experiments are needed. Nonetheless, my thesis work has established a set of protocols for CRISPR/Cas-based genome editing that should be useful in testing and subsequent implementation of successful, multiplex editing of the *L. japonicus* genome. The current toolbox provides a strong foundation for future genome editing experiments by members of the Szczyglowski laboratory, something which was not available prior to this work.

2. The *L. japonicus* gene editing toolbox: the current state of affairs.

There are fourteen different binary vectors which I developed specifically to edit the *L. japonicus HAR1* promoter region. Two of these vectors utilize *Cas9* while the remaining twelve used the *Cas12*a gene. The two *Cas9* containing vectors, *CRISPR-HAR1p-1* and *CRISPR-HAR1p-2*, did not result in detectable gene editing (Figure 23 and Figure 24), while the *Cas12a* vectors are yet to be tested.

The *Cas9*-based multiplex gRNA cassettes were expressed in hairy roots but their maturation, as evaluated by sequencing analysis of the corresponding cRT-PCR products, was deemed less than optimal (Figure 20, Figure 21, and Figure 22). This was because the efficient processing of multiplex gRNA transcripts was expected to be associated with total lack or near total lack of detectable, unprocessed gRNA species (Xie et. al., 2015), which was not observed in my experiments. While the reason for this remains unknown, several factors have been considered. The Cas9 system used was meant to rely on the endogenous *L. japonicus* tRNA processing apparatus to release individual sgRNA species from the multiplex (polycistronic) transcript. Although there is no a priori reason why this system would be less efficient, this has never been tested in *L. japonicus*. The pre-$tRNA_{Gly}$ sequence (trna76 – chromosome 6: 38328889-38328959, *Lotus japonicus* Gifu v1.2; https://lotus.au.dk/) selected for use in the constructs seems typical for $tRNA_{Gly}$ that recognize the GCC codon, of which there are predicted to be 24 in the *L. japonicus* Gifu genome. Alignment of their primary sequences showed that tRNA76 has a polymorphic nucleotide at the 42nd position (C→T), which is otherwise highly conserved (Figure 28**)**. Whether this particular polymorphism could interfere with the efficiency of the tRNA processing, as mediated by RNases Z and P, remains to be seen (see Perspectives and next steps).

Consideration has also been given to the promoter sequences that were used. Even though there is little evidence to suggest that CRISPR/Cas machinery would be unable to access heterochromatic regions in the genome, it appears that the condensed chromatin may reduce the editing efficiency to as low as 2.8% (Feng, et. al., 2016). At rates this low, detecting gene edits would be unlikely without having a rather large experimental population. I have only tested 20 independent hairy roots. Therefore, if *HAR1* is located in a hetrochromatic domain in hairy root cells, or the hypocotyl cells which give rise to hairy roots, an editing efficiency of 5% would be

necessary in order to successfully identify a single change at the *HAR1* locus. Furthermore, as previously mentioned, *HAR1* is known to be expressed in shoot phloem cells (Nontachaiyapoom et. al., 2007). Assuming that its expression in roots is in the same cellular domain, it was decided that testing a phloem-specific promoter is warranted and may generate an interesting insight. Hence, the corresponding binary plasmid, pGEL032-AtS13, is being prepared for testing with the pGEL032-LjUBQ and pGEL032-2x35S vectors. All downstream protocols, including cRT-PCR and locus specific genotyping have been successfully employed and are now standard procedures in the Szczyglowski laboratory.

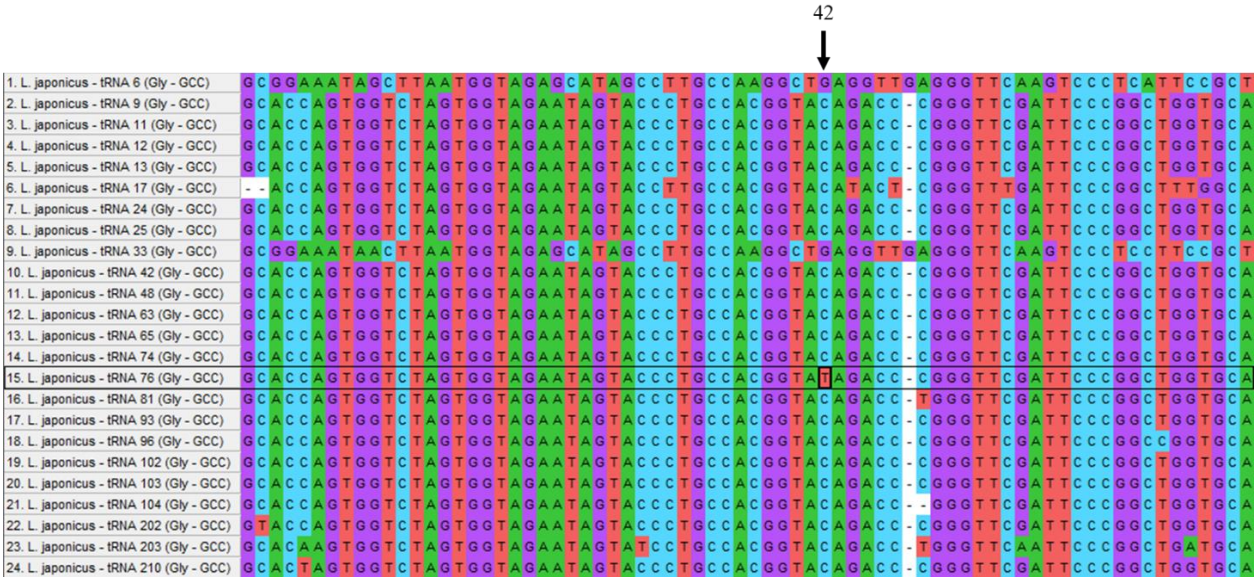


Figure 27. The alignment of *L. japonicus* $tRNA_{Gly}$ sequences that recognize the GCC codon. tRNA 76 (box; line 15) was used in the *CRISPR-HAR1p-1* and *CRISPR-HAR1p-2* vectors. It has a sequence that appears to be typical of the family of tRNA, with the exception of position 42, which has a T instead of a C (arrow).

Considering that multiplexed constructs encompassing several independent gRNAs for both the Cas9 and Cas12a predicted target sequences were used, it is unlikely that the observed

lack of editing at the *HAR1* locus was due to poor choice of gRNAs. Multiplexing was shown to be more efficient at generating genomic edits compared to using individual sgRNAs (Xie et. al., 2014). Thus, I was not only expecting that the efficiency of the editing would be increased, but also that this could lead to a broader range of *HAR1* variants, as different combinations of gRNAs are likely to mediate DNA cleavage in the same cells (Rodriguez-Leal et. al., 2017). These expectations were not met in the initial experiments and I presume that, as outlined above, issues other than the selection of gRNAs, are more likely to be responsible for this negative outcome (see Perspectives and next steps**)**.

3. Considerations on dual versus single transcriptional unit systems.

The *CRISPR-HAR1p-1* vector was designed to conduct multiplex genome editing using a two promoter system. The 2x35S promoter drove the expression of the Cas9 protein while the *LjU6* promoter was used to express the multiplex sgRNA cassette (Figure 12). There is evidence that the editing efficiencies are further improved by expressing both the Cas protein and the sgRNA cassette under a single RNA PolII promoter (Tang et. al., 2018). The *CRISPR-HAR1p-2* vector was designed to test this phenomenon in *L. japonicus*. The use of a STU construct ensures that a given Cas protein is expressed together with the sgRNAs, in effect eliminating the possibility that gene edits did not occur due to absence of the Cas protein.

4. Selection of target sites at the *L. japonicus HAR1* promoter.

Cas9 has been used extensively for editing plant genomes (Puchta, 2017). However, this is restricted to sites that contain the NGG PAM sequence, which poses a challenge when regulatory gene regions are to be targeted, such as promoters, which are often AT rich. Indeed, the 4kb *L. japonicus HAR1* promoter has a GC content of only 39%, and this was reflected by the fact that the number of high quality Cas9 target sites in the region was relatively low. To increase the number of viable targets it was decided to use the Cas12a protein, which recognizes the PAM sequence of TTTV. The Cas12a PAM is more likely to occur in AT-rich regions compared to the Cas9 PAM, allowing for an increase in target sites that can be selected (Wolter & Puchta, 2019). In addition to improving target site selection, the Cas12a protein has the ability to self-process multiplex gRNA transcripts without the need for intervening tRNA sequences (Tang et. al., 2018). To this end, I developed twelve vectors to target the *HAR1* promoter using Cas12a. Selecting Cas12a sites more than doubled the total number of possible targets in the *HAR1* promoter, significantly improving the likelihood of generating a broad range of quantitative variation at the locus.

5. Hairy roots provide a shortcut to checking the editing capabilities of developed constructs.

The major hurdle that is associated with targeted genome modifications is the difficulty associated with generating stably-transformed transgenic plants. Many plants, including most legume crops, are recalcitrant with respect to regenerative procedures, which creates a significant bottleneck. For some, like common bean, a transformation and regeneration protocol has only recently been established (Song et. al., 2020). *L. japonicus* is easily transformable and stably-

transformed transgenic plants can be regenerated via a tissue culture dependent process on a relatively large scale (Stiller et. al., 1997). However, it takes on average six months to obtain T_0 seedlings (Okuma et. al., 2020). In contrast, transgenic hairy roots can be generated in less than two months, which allows for the rapid evaluation of editing efficiencies of untested gRNAs. If the editing efficiencies are too low, it is a clear indication that it is necessary to make modifications to the construct or select new targets before attempting to generate stably-transformed transgenic plants. In addition to the relative ease with which hairy root cultures can be developed and propagated, they can be used to regenerate transgenic plants (Stiller et. al., 1997).

6. Desired phenotypic outcomes from editing the *L. japonicus HAR1* locus.

Even though editing for quantitative variation at the *HAR1* locus has yet to be achieved, the rationale behind the effort has been clear. Different nutrient acquisition-associated traits are expected to be affected by modulating the HAR1 protein dosage. These include root system architecture and uptake and assimilation of soil nutrients, including nitrogen-fixing and VAM symbioses. It will be interesting to see whether any of the prospective quantitative modifications will be beneficial. A lack of HAR1 function is clearly detrimental to the plant performance (Figure 7; see also Wopereis et al., 2000). However, heterozygous *HAR1*/*har1-1 L. japonicus* plants, which produce ca. 50% of the wild-type *HAR1* mRNA level and develop approximately 50% more nodules, grow and produce seeds similar to wild-type plants (Mark Pampuch and K. Szczyglowski, unpublished data). Whether this enhanced nodulation pattern could be beneficial under stress conditions, such as drought, excessive cold or a short growth season, has yet to be determined. However, a more exciting idea is to generate a continuum of quantitative variation at the *HAR1*

locus and then search for beneficial alleles. A relevant question in this context is whether more subtle changes to the *HAR1* mRNA level will be advantageous and if other genomic modifications could enhance the effect.

Legumes, and other plants, utilize xylem to transport nutrients from roots to sink tissues, such as leaves and seeds. The strength of sink tissues dictates the level of demand for nutrients. This, in turn impacts the formation and performance of source tissues and these sink-source relationships orchestrate the overall plant economy (Tegeder & Masclaux-Daubresse, 2017; Smith et al., 2018). Interestingly, it has been shown that enhancing transport of ureides (the main nitrogen fixation and assimilation components in tropical legumes) from nodules, the source tissue, to shoot (the sink tissue) led to increased soybean productivity (Carter and Tegeder, 2016). Thus, parallel editing for enhanced nodule formation and ureide transport, with the latter appearing to mimic increased shoot demand for nitrogen, may be a fruitful avenue for increasing legume productivity.

Use of synthetic alleles, will undoubtably be met with scorn and disapproval from a portion of the general public. These concerns could be partially alleviated by exploring the relevant natural variation, once desirable synthetic alleles are characterized, although such an approach may be limited in terms of its potential success. Nonetheless, with this in mind, both the sequence conservation and natural variation that exist at the soybean *NARK* locus, the orthologue of *L. japonicus HAR1*, were evaluated.

7. Natural variation at the *NARK* locus.

Orthologues of *L. japonicus HAR1* have been identified in legume crops (Figure 9). Two regions in the *HAR1* promoter were characterized as being conserved, showing more than 70%

sequence identity with the *G. max* and *M. truncatula* orthologues (Figure 25). These regions may encompass important regulatory modules in the promoters. *In silico* analysis showed that several predicted *cis*-regulatory elements, including MBS, ABRE, GT1 and CAT box motifs, are present in these conserved promoter regions. It will be interesting to see whether any phenotypic variants that may arrive as the result of editing experiments will be due to modifications within these conserved regions and/or predicted *cis*-regulatory elements. Furthermore, by analyzing the phenotypic effects of any sequence alterations to the *cis*-acting regulatory elements that are generated in future genome editing experiments, it will be possible to determine which of those are actually functionally relevant to *HAR1*.

Levels of polymorphism were evaluated across the entire *Glycine max NARK* locus (Figure 27). This natural variation is representative of 350 different varieties of cultivated soybean. Surprisingly, the results of the statistical analysis indicate that there is a greater level of variation in the *NARK* coding regions than in the non-coding. In the coding region, SNPs appeared with a frequency of 1 in every 72 positions as compared to 1 in every 94 positions in the non-coding regions (Table 1). These values were compared to the published data for the soybean *GmGBP1* gene, encoding an orthologue of the *Arabidopsis thaliana* SNW/SKI-interacting protein (*SKIP*) (Zhao et. al., 2018). The *GmGBP1* gene had an average of 1 in every 106 positions that were polymorphic. On average, 1 in every 153 positions were found to be polymorphic in the coding regions of *GmGBP1*. SNPs appeared in the non-coding regions of *GmGBP1* with the frequency of 1 in every 75 positions (Table 1).

This pipeline allows for the comparison of any synthetic variation to the variation which

already exists in cultivated soybeans. This pipeline also provides the ability to design a wide range of genetic tools (i.e. primers or gRNAs) that can be used in different soybean varieties simultaneously.

8. Perspectives and next steps.

Current experiments, albeit unsuccessful at generating synthetic variation at the *HAR1* locus, established a solid foundation which was necessary to perform targeted genome modifications in *L. japonicus* and other legume species. Tools have been developed to help to identify the problems that resulted in the negative genome editing outcome. Testing the Cas12a-containing constructs, which eliminate the need for tRNA based processing, is expected to be informative. The self-processing properties of Cas12a should assure efficient production of mature gRNAs, thus possibly addressing two problems, (1) an inefficient processing of Cas9-dependent multiplex gRNA transcripts and (2) lack of editing due to a low concentration of mature gRNAs. Having a number of binary vectors containing only two gRNAs should also be helpful in this context as it is currently unclear whether a transcript containing a relatively large number of gRNAs (e.g. 8 to 10) will be effective in *L. japonicus*. An alternative approach could be to use a mixture of *Agrobacterium* strains, each containing a binary vector with a small number of gRNAs.

Using constructs with different promoters, as prepared by this thesis work, should also be instructive. Particularly interesting in this context will be testing of the phloem-specific promoter, which should show whether utilization of tissue specific expression is necessary. Also, the number of independent hairy roots tested for evidence of editing at the *HAR1* locus will have to be significantly increased to make sure that negative outcomes are not due to lower than expected

efficiency. If this is the case, subsequent experiments could be refined by focusing the effort on improving the process.

Finally, it was originally envisaged that *L. japonicus* plants carrying any *HAR1* promoter edits would be directly regenerated from axenic hairy root cultures. Although this remains a possibility, a new method allowing for direct generation of edited plants using ectopically induced meristems, without the need for tissue culture, has emerged (Maher et. al., 2019). However, the feasibility of the method for *L. japonicus* has yet to be tested. If successful, this would provide a significant shortcut for generating a large number of genome edited plants. Hairy roots are good for testing the editing efficiency but whole plants are necessary to evaluate phenotypic impacts.

Although my thesis work focused entirely on the model plant, parallel work has already been initiated to replicate this effort in crop legume species, such as soybean, common bean and/or alfalfa. The identification of presumed *HAR1* orthologs (Figure 9) constituted the first step toward this goal. Each of these species, as well as other legume crops, could benefit from an enhanced nitrogen uptake efficiency, primarily to support high yields and environmental performance. The same line of research may guide improvement of input (nitrogen)-thirsty cereals such as wheat and maize. Together, these advances should help lead to crop improvements which will provide the necessary food supply in a sustainable manner.

9. Limitations.

My thesis work was significantly impacted by the COVID-19 related shutdowns. I had no access to my laboratory at AAFC for a duration of six months, starting at the beginning of April 2020. Experimentation was stalled and several of my hairy root cultures were lost during this time, requiring protocols to be restarted once access was re-established. Furthermore, the access was

never fully restored during the second year of my funding period. Work had to be planned and completed in six hour shifts, impairing my ability to progress data collection at the necessary rate. This prevented me from testing the constructs that I had prepared, as it was necessary to troubleshoot the initially negative editing results. Follow-up work is clearly needed to fill this experimental gap.

References

Ågren, I., Wetterstedt, M., & Billberger, K. (2012). Nutrient limitation on terrestrial plant growth – modeling the interaction between nitrogen and phosphorus. *New Phytologist* **194**, 953-960.

Araya, T., Miyamoto, M., Wibowo, J., Suzuki, A., Kojima, S., Tsuchiya, Y., Sawa, S., Fukuda, H., Wirén, N., & Takahashi, H. (2014). CLE-CLAVATA1 peptide-receptor signaling module regulates the expansion of plant root systems in a nitrogen-dependent manner. *Proceedings of the National Academy of Sciences* **111**, 2029-2034.

Bailey-Serres, J., Parker, J., Ainsworth, E., Oldroyd, G., & Schroeder, J. (2019). Genetic strategies for improving crop yields. *Nature* **575**, 109–118.

Bauer, A., Edgeworth, M., Edwards, L., Ellis, E., Gibbard, P., & Merritts, D. (2021). *Nature* **597**, 332.

Beach, K., & Gresshoff, P. (1988). Characterization and culture of *Agrobacterium rhizogenes* transformed roots of forage legumes. *Plant Science* **57**, 73-81.

Benyon, S., Stover, E., Bowman, K. D., Niedz, R., Shatters, R. G., Zale, J., & Belknap, W. (2013). GUS expression driven by constitutive and phloem-specific promoters in citrus hybrid US-802. *In Vitro Cellular & Developmental Biology Plant* **49**, 255–265.

Bloch, S., Clark, R., Gottlieb, S., Kent Wood, L., Shah, N., Mak,S., Lorigan, J., Johnson, J., Davis-Richardson, A., Williams, L., McKellar, M., Soriano, D., Petersen, M., Horton, A., Smith, O., Wu, L., Tung, E., Broglie, R., Tamsir, A., & Temme, K. (2020). Biological nitrogen fixation in maize: optimizing nitrogenase expression in a root-associated diazotroph,. *Journal of Experimental Botany* **71**, 4591–4603.

Breitburg, D., Levin, L., Oschlies, A., Grégoire, M., Chavez, F., Conley, D., Garçon, V., Gilbert, D., Gutiérrez, D., Isensee, K., Jacinto, G., Limburg, K., Montes, I., Naqvi, S., Pitcher, C., Rabalais, N., Roman, M., Rose, K,. Seibel, B., Telszewski, M., Yasuhara, M., & Zhang, J. (2018). Declining oxygen in the global ocean and coastal waters. *Science* **5**, No. 6371.

Brundrett, C. (2002). Coevolution of roots and mycorrhizas of land plants. *New Phytologist* **154**, 275-304.

Bücking, H., & Kafle, A. (2015). Role of Arbuscular Mycorrhizal Fungi in the Nitrogen Uptake of Plants: Current Knowledge and Research Gaps. *Agronomy* **5**, 587-612.

Burgess, S., Muirhead, J. & Bowring, S. (2017). Initial pulse of Siberian Traps sills as the trigger of the end-Permian mass extinction. *Nature Communications* **8**, 164.

Cafaro, P. (2015). Three ways to think about the sixth mass extinction. *Biological Conservation* **192**, 387-393.

Canfield, D., Glazer, A., & Falkowski, P. (2010). The Evolution and Future of Earth's Nitrogen Cycle. *Science* **330**, No. 6001.

Carter, A., & Tegeder, M. (2016). Increasing Nitrogen Fixation and Seed Development in Soybean Requires Complex Adjustments of Nodule Nitrogen Metabolism and Partitioning Processes. *Current Biology* **26**, 2044-2051.

Chen, K., & Gao, C. (2013). TALENs: Customizable Molecular DNA Scissors for Genome Engineering of Plants. *Journal of Genetics and Genomics* **40**, 271-279.

Collier, R., & Tegeder, M. (2012). Soybean ureide transporters play a critical role in nodule development, function and nitrogen export. *The Plant Journal* **72**, 355-367.

Crespi, M., & Gálvez, S. (2000). Molecular Mechanisms in Root Nodule Development. *Journal of Plant Growth Regulation* **19**, 155–166.

Crews, T., & Peoples, M. (2004). Legume versus fertilizer sources of nitrogen: ecological tradeoffs and human needs. *Agriculture, Ecosystems & Environment* **102**, 279-297.

Crutzen, P. (2002). Geology of mankind. *Nature* **415**, 23.

Doudna, J., & Charpentier, E. (2014). The new frontier of genome engineering with CRISPR-Cas9. *Science* **346**, No. 6213.

Du, Q., Yang, J., Sadiq, S., Yang, R., Yu, J., & Li, W. (2021). Comparative transcriptome analysis of different nitrogen responses in low-nitrogen sensitive and tolerant maize genotypes. *Journal of Integrative Agriculture* **20**, 2043-2055.

Egli, D. (2008). Comparison of Corn and Soybean Yields in the United States: Historical Trends and Future Prospects. *Agronomy Journal* **100**, S-79-S-88.

Engler, C., Kandzia, R., & Marillonnet, S. (2008). A One Pot, One Step, Precision Cloning Method with High Throughput Capability. *PLOS one* **3**, e3647.

Feng, C., Yuan, J., Wang, R., Liu, Y., Birchler, J., & Ha, F. (2016). Efficient Targeted Genome Modification in Maize Using CRISPR/Cas9 System. *Journal of Genetics and Genomics* **43**, 37-43.

Ferguson, J., Indrasumunar, A., Hayashi, S., Lin, M.-H., Lin, Y.-H., Reid, E. & Gresshoff, P. (2010). Molecular Analysis of Legume Nodule Development and Autoregulation. *Journal of Integrative Plant Biology* **52**, 61-76.

Feynman, J., & Ruzmaikin, A. (2007). Climate stability and the development of agricultural societies. *Climatic Change* **84**, 295–311.

Fitzpatrick, C., Copeland, J., Wang, P., Guttman, D., Kotanen, P., & Johnson, M. (2018). Assembly and ecological function of the root microbiome across angiosperm plant species. *Proceedings of the National Academy of Sciences* **115**, E1157-E1165.

Folke, C., Polasky, S., Rockström, J., Galaz, V., Westley, F., Lamont, M., Scheffer, M., Osterblom, H., Carpenter, S., Chapin, S., Seto, K., Weber, E., Crona, B., Daily, G., Dasgupta, P., Gaffney, O., Gordon, L., Hoff, H., Levin, S., Lubchenco, J., Steffen, W., & Walker, B. (2021). Our future in the Anthropocene biosphere. *Ambio* **50,** 834–869.

Forde, B. (2014) Nitrogen signalling pathways shaping root system architecture: an update. *Current Opinion in Plant Biology* **21**, 30-36.

Franche, C., Lindström, K. & Elmerich, C. (2009) Nitrogen-fixing bacteria associated with leguminous and non-leguminous plants. *Plant Soil* **321**, 35–59.

French, E., Kaplan, I., Iyer-Pascuzzi, A., Nakatsu, C., & Enders, L. (2021). Emerging strategies for precision microbiome management in diverse agroecosystems. *Nature Plants* **7**, 256–267.

Geddes, B., Paramasivan, P., Joffrin, A., Thompson, A., Christensen, K., Jorrinn, B., Brett, P., Conway, S., Oldroyd, G., & Poole, P. (2019). Engineering transkingdom signalling in plants to control gene expression in rhizosphere bacteria. *Nature Communications* **10**, 3430.

Giehl, R., Gruber, B., & von Wirén, N. (2014). It's time to make changes: modulation of root system architecture by nutrient signals. *Journal of Experimental Botany* **65**, 769–778.

Giehl, R., Gruber, B., & von Wirén, N. (2014). Root Nutrient Foraging. *Plant Physiology* **166**, 509–517.

Gomiero, T., Pimentel, D., & Paoletti, M. (2011). Environmental Impact of Different Agricultural Management Practices: Conventional vs. Organic Agriculture. *Critical Reviews in Plant Science* **30**, 95-124.

Griesmann, M., Chang, Y., Liu, X., Song, Y., Haberer, G., Crook, M., Billaut-Penneteau, B., Lauressergues, D., Keller, J., Imanashi, L, Roswanjaya, Y., Kohlen, W., Pujic, P., Battenberg, K., Alloisio, N., Liang, Y, Hilhorst, H., Salgado, M., Hocher, V., Gherbi, H., Svistoonoof, S., Doyle, J., He, S., Xu, Y., Xu, S., Gao, Q., Fang, X., Fu, Y., Normand, P., Berry, A., Wall, L., Ane, J., Pawloski, K., Xu, X., Yang, H., Spanngal, M., Mayer, K., Wong, G., Parniske, M., Delaux, P., & Cheng, S. (2018). Phylogenomics reveals multiple losses of nitrogen-fixing root nodule symbiosis. *Science* **361**, No. 6398.

Guillon S., Trémouillaux-Guiller J., Kumar Pati P., & Gantet P. (2008). Hairy Roots: a Powerful Tool for Plant Biotechnological Advances. *Springer Books*.

Gupta, A. (2004). Origin of agriculture and domestication of plants and animals linked to early Holocene climate amelioration. *Current Science* **87**, 54–59.

Ha, D., Lee, J., Lee, N., Kim, D., Ko, J., & Kim, Y. (2020). Highly efficient and safe genome editing by CRISPR-Cas12a using CRISPR RNA with a ribosyl-2′-O-methylated uridinylate-rich 3′-overhang in mouse zygotes. *Experimental & Molecular Medicine* **52**, 1823–1830.

Haber, F., (1920). The Synthesis of Ammonia from Its Elements. *Nobel Lecture*.

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

Heden, P. (2003). The genes of the Green Revolution. *Trends in Genetics* **19**, 5-9.

Helgason, T., & Fitter, A. (2005). The ecology and evolution of the arbuscular mycorrhizal fungi. *Mycologist* **19**, 96-101.

Hiei, Y., Ishida, Y., Komari, T. (2014). Progress of cereal transformation technology mediated by *Agrobacterium tumefaciens*. *Frontiers in Plant Science* **5**, 628.

Hodge, A. (2004). The plastic plant: root responses to heterogeneous supplies of nutrients. *New Phytologist* **162**, 9-24.

Hongfu, Y., Kexin, Z., Jinnan, T., Zunyi, Y., & Shunbao, W. (2001). The Global Stratotype Section and Point (GSSP) of the Permian-Triassic Boundary. *Episodes* **24**, 102-114.

Hossain, S., Liao, J., James, E. K., Sato, S., Tabata, S., Jurkiewicz, A., Madsen, L. H., Stougaard, J., Ross, L., & Szczyglowski, K. (2012). *Lotus japonicus ARPC1* is required for rhizobial infection. *Plant Physiology* **160**, 917–928.

Hossain, S., Shrestha, A., Zhong, S., Miri, M., Austin, R., Sato, S., Ross, L., Huebert, T. Tromas, A. Torres-Jerez, I., Tang, Y., Udvardi, M., Murray, J., & Szczyglowski, K. (2016). Lotus japonicus NF-YA1 Plays an Essential Role During Nodule Differentiation and Targets Members of the SHI/STY Gene Family. *Molecular Plant-Microbe Interactions* **29**, 950-964.

Huault, E., Laffont, C., Wen, J., Mysore, K., Ratet, P., Duc, G., & Frugier, F. (2014) Local and Systemic Regulation of Plant Root System Architecture and Symbiotic Nodulation by a Receptor-Like Kinase. *PLOS Genetics* **10**, e1004891.

Hublin, J., Ben-Ncer, A., Bailey, S., Freidline, S., Neubauer, S., Skinner, M., Bergmann, I., Cabec, A., Benazzi, S., Harvati, K., & Gunz, P. (2017). New fossils from Jebel Irhoud, Morocco and the pan-African origin of *Homo sapiens*. *Nature* **546**, 289–292.

Isidra-Arellano, M., Pozas-Rodríguez, E., del Rocío Reyero-Saavedra, M., Arroyo-Canales, J., Ferrer-Orgaz, S., del Socorro Sánchez-Correa, M., Cardenas, L., Covarrubias, A. & Valdés-López, O. (2020). Inhibition of legume nodulation by Pi deficiency is dependent on the autoregulation of nodulation (AON) pathway. *Plant Journal* **103**, 1125-1139.

Ji, J., Zhang, C., Sun, Z., Wang, L., Duanmu, D., & Fan, Q. (2019). Genome Editing in Cowpea *Vigna unguiculata* Using CRISPR-Cas9. *International Journal of Molecular Sciences* **20**, 2471.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J., & Charpentier, E. (2012). A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science* **337**, No. 6096.

Johnson, N. (1993). Can Fertilization of Soil Select Less Mutualistic Mycorrhizae? *Ecological Applications* 3, 749-757.

Kaneko, T., Nakamura, Y.,Sato, S., Asamizu, E., Kato, T., Sasamoto, S., Watanabe, A., Idesawa, K., Ishikawa, A., Kawashima, K., Kimura, T., Kishida, Y., Kiyokawa, C., Kohara, M., Matsumoto, M., Matsuno, A., Mochizuki, Y., Nakayama, S., Nakazaki, N., Shimpo, S., Sugimoto, M., Takeuchi, C., Yamada, M., & Tabata, S. (2000). Complete Genome Structure of the Nitrogen-fixing Symbiotic Bacterium *Mesorhizobium loti*. *DNA Research* ***7***, 331–338.

Kelly, L., Giljohann, K., Duane, A., Aquilue, N., Archibald, S., Batllori, E., Bennett, A., Buckland, S., Canelles, Q., Clarke, M., Fortin, M., Hermoso, V., Herrando, S., Keane, R., Lake, F., McCarthy, M., Moran-Ordonez, A., Parr, C., Pausas, J., Penman, T., Regos, A., Rumpff, L., Santos, J., Smith, A., Syphard, A., Tingley, M., & Brotons, L. (2020). Fire and biodiversity in the Anthropocene. *Science* **370**, No. 6519.

Kenrick, P., & Crane, P. (1997). The origin and early evolution of plants on land. *Nature* **389**, 33–39.

Kereszt, A., Li, D., Indrasumunar, A., Nguyen, C., Nontachaiyapoom, S., Kinkema, M., & Gresshoff, P. (2007). Agrobacterium rhizogenes-mediated transformation of soybean to study root biology. *Nature Protocols* **2**, 948–952.

Kiba, T., & Krapp, A. (2016). Plant Nitrogen Acquisition Under Low Availability: Regulation of Uptake and Root Architecture. *Plant and Cell Physiology* **57**, 707–714.

Kinkema, M., & Gresshoff, P. (2008). Investigation of Downstream Signals of the Soybean Autoregulation of Nodulation Receptor Kinase GmNARK. *Molecular Plant-Microbe Interactions* **21**, 1337-1348.

Koonin, E., & Makarova, K. (2019). Origins and evolution of CRISPR-Cas systems. *Philosophical Transactions of the Royal Society – Biological Sciences* **374**, 20180087.

Krusell, L., Madsen, L., Sato, S., Aubert, G., Genua, A., Szczyglowski, K., Duc, G., Kaneko, T., Tabata, S., de Bruijn, F., Pajuelo, E., Sandal, N., Stougaard, J. (2002). Shoot control of root development and nodulation is mediated by a receptor-like kinase. *Nature* **420**, 422–426.

Lagunas, B., Achom, M., Bonyadi-Pour, R., Pardal, A., Richmond, B., Sergaki, C., Vázquez, S., Schäfer, P., Ott, S., Hammond, J., Gifford, M. (2018). Regulation of Resource Partitioning Coordinates Nitrogen and Rhizobia Responses and Autoregulation of Nodulation in *Medicago truncatula*. *Molecular Plant* **12**, 833-846.

Leff, J., Jones, S., Prober, S., Barberán, A., Borer, E., Firn, J., Harpole, W., Hobbie, S., Hofmockel, K., Knops, J., McCulley, R., La Pierre, K., Risch, A., Seabloom, E., Schütz, M., Steenbock, C., Stevens, C., & Fierer, N. (2015). Consistent responses of soil microbial communities to elevated nutrient inputs in grasslands across the globe. *Proceedings of the National Academy of Sciences* **112**, 10967-10972.

Lewis, S., & Maslin, M. (2015). Defining the Anthropocene. *Nature* **519,** 171–180.

Li, S., Tian, Y., Wu, K., Wu, K., Ye, Y., Yu, J., Zhang, J., Liu, Q., Hu, M., Li, H., Tong, Y., Harberd, N., & Fu, X. (2018). Modulating plant growth–metabolism coordination for sustainable agriculture. *Nature* **560**, 595–600.

Liu, H., Ding, Y., Zhou, Y., Jin, W., Xie, K., & Chen, L. (2017). CRISPR-P 2.0: An Improved CRISPR-Cas9 Tool for Genome Editing in Plants. *Molecular Plant* **10**, 530-532.

Liu, K., Niu, Y., Konishi, M., Wu, Y., Du, H., Chung, H., Li, L., Boudsocq, M., McCormack, M., Maekawa, S., Ishida, T., Zhang, C., Shokat, K., Yanagisawa, S. & Sheen, J. (2017). Discovery of nitrate–CPK–NLP signalling in central nutrient–growth networks. *Nature* **545**, 311–316.

Liu, S., Zhang, M., Feng, F., & Tian, Z. (2020). Toward a ‘‘Green Revolution’’ for Soybean. *Molecular Plant* **13**, 688–697.

Lu, Y., Tian, Y., Shen, R., Yao, Q., Wang, M., Chen, M., Dong, J., Zhang, T., Li, F., Lei, M., & Zhu, J. (2020). Targeted, efficient sequence insertion and replacement in rice. *Nature Biotechnology* **38**, 1402–1407.

Magori, S., Oka-Kira, E., Shibata, S., Umehara, Y., Kouchi, H., Hase, Y., Tanaka, A., Sato, S., Tabata, S., & Kawaguchi, M. (2009). TOO MUCH LOVE, a Root Regulator Associated with the Long-Distance Control of Nodulation in *Lotus japonicus*. *Molecular Plant-Microbe Interactions* **22**, 259-268.

Maher, M., Nasti, R., Vollbrecht, M., Starker, C., Clark, M., &Voytas, D. (2019). Plant gene editing through de novo induction of meristems. *Nature Biotechnology* **38**, 84–89.

Mirzaei, S., Batley, J., El-Mellouki, T., Liu, S., Meksem, K., Ferguson, B., & Gresshoff, P. (2017). Neodiversification of homeologous CLAVATA1-like receptor kinase genes in soybean leads to distinct developmental outcomes. *Sci Rep* **7**, 8878.

Morris, J., Puttick, M., Clark, J., Edwards, D., Kenrick, P., Pressel, S., Wellman, C., Yang, Z., Schneider, H., & Donoghue, P. (2018). The timescale of early land plant evolution. *Proceedings of the National Academy of Sciences* **115**, E2274-E2283.

Mousa, W., Shearer, C., Limay-Rios, V., Ettinger, C., Eisen, J., & Raizada, M. (2016). Root-hair endophyte stacking in finger millet creates a physicochemical barrier to trap the fungal pathogen Fusarium graminearum. *Nature Microbiology* **1**, 16167.

Mus, F., Crook, M., Garcia, K., Costas, A., Geddes, B., Kouri, E., Paramasivan, P., Ryu, M., Oldroyd, G., Poole, P., Udvardi, M., Voigt, C., Ane, J., & Peters, J. (2016). Symbiotic Nitrogen Fixation and the Challenges to Its Extension to Nonlegumes. *American Society of Microbiology* **82**, 13.

Nitika, S., Mehak, S., Aman, K., Devpriya, D., Jasneet, S., & Parveen, C. (2021). Biochemical and Genetic Approaches Improving Nitrogen Use Efficiency in Cereal Crops: A Review. *Frontiers in Plant Science* **12**, 757.

Nontachaiyapoom, S., Scott, P., Men, A., Kinkema, M., Schenk, P., & Gresshoff, P. (2007). Promoters of Orthologous *Glycine max* and *Lotus japonicus* Nodulation Autoregulation Genes Interchangeably Drive Phloem-Specific Expression in Transgenic Plants. *Molecular Plant-Microbe Interactions* **20**, 769-780.

Noxon, J. (1976). Atmospheric nitrogen fixation by lightning. *Geophysical Research Letters* **3**, 463-465.

Ogden, D., & Sleep., N. (2012). *Proceedings of the National Academy of Sciences* **109**, 59-62.

Oka-Kira, E., & Kawaguchi, M. (2006). Long-distance signaling to control root nodule number. *Current Opinion in Plant Biology* **9**, 496-502.

Okamoto, S., & Kawaguchi, M. (2015). Shoot HAR1 mediates nitrate inhibition of nodulation in *Lotus japonicus*. *Plant Signaling and Behavior* **10**, e1000138.

Okamoto, S., Shinohara, H., Mori, T. Matsubayashi, Y., & Kawaguchi, M. (2013). Root-derived CLE glycopeptides control nodulation by direct binding to HAR1 receptor kinase. *Nature Communications* **4**, 2191.

Okumo, N., Soyano, T., Suzaki, T., & Kawaguchi, M. (2020). MIR2111-5 locus and shoot-accumulated mature miR2111 systemically enhance nodulation depending on HAR1 in Lotus japonicus. *Nature Communications* **11**, 5192.

Ott, T., van Dongen, J., Günther, C., Krusell, L., Desbrosses, G., Vigeolas, H., Bock, V., Czechowski, T., Geigenberger, P., & Udvardi, M. (2005). Symbiotic leghemoglobins are crucial for nitrogen fixation in legume root nodules but not for general plant growth and development. *Current Biology* 15, 531-535.

Pan, Y., Cassman, N., de Hollander, M., Mendes, L., Korevaar, H., Geerts, R., van Veen, J., & Kuramae, E. (2014). Impact of long-term N, P, K, and NPK fertilization on the composition and potential functions of the bacterial community in grassland soil. *FEMS Microbiology Ecology* **90**, 195–205.

Parniske, M. (2018). Uptake of bacteria into living plant cells, the unifying and distinct feature of the nitrogen-fixing root nodule symbiosis. *Current Opinions in Plant Biology* **44**, 164-174.

Peng, J., Richards, D., Hartley, N., Murphy, G., Devos, K., Flintham, J., Beales, J., Fish, L., Worland, A., Pelica, F., Sudhakar, D., Christou, P., Snape, J., Gale, M., & Harberd, N. (1999). 'Green revolution' genes encode mutant gibberellin response modulators. *Nature* **400**, 256–261.

Pirozynski, K., & Malloch, D. (1975). The Origin Of Land Plants: A Matter Of Mycotrophism. *BioSystems* **6**, 153-164.

Puchta, H. (2017). Applying CRISPR/Cas for genome engineering in plants: the best is yet to come. *Current Opinion in Plant Biology* **36**, 1-8.

Rees, D., & Howard, J. (2000). Nitrogenase: standing at the crossroads. *Current Opinion in Chemical Biology* **4**, 559-566.

Reid, D., Ferguson, B., Hayashi, S., Lin, Y., & Gresshoff, P. (2011). Molecular mechanisms controlling legume autoregulation of nodulation. *Annals of Botany* **108**, 789–795.

Rockström, J., Steffen, W., Noone, K., Persson, Å., Chapin, F. S., Lambin, E., Lenton, T. M., Scheffer, M., Folke, C., Schellnhuber, H. J., Nykvist, B., de Wit, C. A., Hughes, T., van der Leeuw, S., Rodhe, H., Sörlin, S., Snyder, P. K., Costanza, R., Svedin, U., Falkenmark, M., Karlberg, L., Corell, R., Fabry, V., Hansen, J., Walker, B., Liverman, D., Richardson, K., & Foley, J. (2009). Planetary Boundaries: Exploring the Safe Operating Space for Humanity. *Ecology and Society* **14**(2).

Rodríguez-Leal, D., Lemmon, Z., Man, J., Bartlett, M., Lippman, Z. (2017). Engineering Quantitative Trait Variation for Crop Improvement by Genome Editing. *Cell* **171**, 470-480.

Sakadevan, K., & Nguyen, M. (2017). Chapter Four - Livestock Production and Its Impact on Nutrient Pollution and Greenhouse Gas Emissions. *Advances in Agronomy* **141**, 147-184.

Schaarschmidt, S., Gresshoff, P., & Hause, B. (2013). Analyzing the soybean transcriptome during autoregulation of mycorrhization identifies the transcription factors GmNF-YA1a/b as positive regulators of arbuscular mycorrhization. *Genome Biology* **14**, R62.

Scheben, A., & Edwards, D. (2018). Towards a more predictable plant breeding pipeline with CRISPR/Cas-induced allelic series to optimize quantitative and qualitative traits. *Current Opinion in Plant Biology* **45**, 218-225.

Schnabel, E., Journet, E., de Carvalho-Niebel, F., Duc, G., Frugoli, J. (2005). The *Medicago truncatula SUNN* gene encodes a CLV1-like leucine-rich repeat receptor kinase that regulates nodule number and root length. *Plant Molecular Biology* **58**, 809-822.

Schneider, H., & Lynch, J. (2020). Should Root Plasticity Be a Crop Breeding Target?. *Frontiers in Plant Science* **11**, 546.

Sellner, K., Doucette, G., & Kirkpatrick, G. (2003). Harmful algal blooms: causes, impacts and detection. *Journal of Industrial Microbiology and Biotechnology* **30**, 383–406.

Shan, Q., Wang, Y., Li, J., Zhang, Y., Chen, K., Liang, Z., Zhang, K., Liu, J., Xi, J., Qiu, J., & Gao, C. (2013). Targeted genome modification of crop plants using a CRISPR-Cas system. *Nature Biotechnology* **31**, 686–688.

Shrestha, A., Zhong, S., Therrien, J., Huebert, T., Sato, S., Mun, T., Andersen, S. U., Stougaard, J., Lepage, A., Niebel, A., Ross, L., & Szczyglowski, K. (2021). *Lotus japonicus Nuclear Factor YA1*, a nodule emergence stage-specific regulator of auxin signalling. *The New phytologist* **229**, 1535–1552.

Shi, J., & Lai, J. (2015). Patterns of genomic changes with crop domestication and breeding. *Current Opinion in Plant Biology* **24**, 47-53.

Smale, M. (1997). The Green Revolution and wheat genetic diversity: Some unfounded assumptions. *World Development* **25**, 1257-1269.

Smith, M., Rao, I., & Merchant, A., (2018). Source-Sink Relationships in Crop Plants and Their Influence on Yield Development and Nutritional Quality. *Frontiers in Plant Science* **9**, 1889.

Smith, V., & Schindler, D. (2009). Eutrophication science: where do we go from here?. *Trends in Ecology & Evolution* **24**, 201-207.

Song, G., Han, X., Wiersma, A., Zong, X., Awale, H., & Kelly, J. (2020) Induction of competent cells for *Agrobacterium tumefaciens*-mediated stable transformation of common bean (*Phaseolus vulgaris L.*). *PLOS ONE* **15**, e0229909.

Soumare, A., Diedhiou, A., Thuita, M., Hafidi, M., Ouhdouch, Y., Gopalakrishnan, S., & Kouisni, L. (2020). Exploiting Biological Nitrogen Fixation: A Route Towards a Sustainable Agriculture. *Plants* **9**, 1011.

Soyano, T., Kouchi, H., Hirota, A., & Hayashi, M. (2013). NODULE INCEPTION Directly Targets NF-Y Subunit Genes to Regulate Essential Processes of Root Nodule Development in *Lotus japonicus*. *PLOS Genetics* **9**, e1003352.

Steffen, W., Richardson, K., Rockstrom, J., Cornell, S., Fetzer, I., Bennett, E., Biggs, R., Carpenter, S., de Vries, W., de Wit, C., Folke, C., Gerten, D., Heinke, J., Mace, G., Persson, L., Ramanathan, V., Reyers, B., & Sorlin, S. (2015). Planetary boundaries: Guiding human development on a changing planet. *Science* **347**, No. 6223.

Stiller, J., Martirani, L., Tuppale, S., Chian, R., Chiurazzi, M., & Gresshoff, P. (1997). High frequency transformation and regeneration of transgenic plants in the model legume *Lotus japonicus*. *Journal of Experimental Botany* **48**, 1357–1365.

Strullu-Derrien, C., Selosse, M., Kenrick, P., & Martin, F. (2018), The origin and evolution of mycorrhizal symbioses: from palaeomycology to phylogenomics. *New Phytol*ogist **220**, 1012-1030.

Suzaki, T., Yoro, E., & Kawaguchi, M. (2015). Chapter Three - Leguminous Plants: Inventors of Root Nodules to Accommodate Symbiotic Bacteria. *International Review of Cell and Molecular Biology* **316**, 111-158.

Szczyglowski, K., & Stougaard, J. (2008). Lotus genome: pod of gold for legume research. *Trends in Plant Science* **13**, 515-517.

Takahara, M., Magori, S., Soyano, T., Okamoto, S., Yoshida, C., Yano, K., Sato, S., Tabata, S., Yamaguchi, K., Shigenobu, S., Takeda, N., Suzaki, T., & Kawaguchi, M. (2013). TOO MUCH LOVE, a Novel Kelch Repeat-Containing F-box Protein, Functions in the Long-Distance Regulation of the Legume–Rhizobium Symbiosis. *Plant and Cell Physiology* **54**, 433–447.

Tamura, K., Stecher, G., & Kumar, S. (2021). MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Molecular Biology and Evolution* **38**, 3022–3027.

Tang, X., Ren, Q., Yang, L., Bao, Y., Zhong, Z., He, Y., Liu, S., Qi, C., Liu, B., Wang, Y., Sretenovic, S., Zhang, Y., Zheng, X., Zhang, T., Qi, Y., & Zhang, Y. (2019). Single transcript unit CRISPR 2.0 systems for robust Cas9 and Cas12a mediated plant genome editing. *Plant biotechnology journal* **17**, 1431–1445.

Tanksley, S., & Nelson, J. (1995). Advanced backcross QTL analysis: a method for the simultaneous discovery and transfer of valuable QTLs from unadapted germplasm into elite breeding lines. *Theoretical Applied Genetics* **92**, 191-203.

Taylor, T. (1995). Fossil arbuscular mycorrhizae from the Early Devonian. *Mycologia* **87**, 560-573.

Tegeder, M. & Masclaux-Daubresse, C. (2018). Source and sink mechanisms of nitrogen transport and use. *New Phytologist* **217**, 35-53.

Trivedi, P., Leach, J., Tringe, S., Sa, T., & Singh, B. (2020). Plant–microbiome interactions: from community assembly to plant health. *Nature Reviews Microbiology* **18**, 607–621.

Tollefson, J. (2019). Humans are driving one million species to extinction. *Nature* **569**, 171.

Tsikou, D., Yan, Z., Holt, D., Abel, N., Reid, D., Madsen, L., Bhasin, H., Sexauer, M., Stougaard, J., & Markhmann, K. (2018). Systemic control of legume susceptibility to rhizobial infection by a mobile microRNA. *Science* **362**, No. 6411.

Udvardi, M., & Poole, P. (2013). Transport and metabolism in legume-rhizobia symbioses. *Annual Reviews of Plant Biology* **64**, 781-805.

Unkovich, M., & Pate, J. (2000). An appraisal of recent field measurements of symbiotic N2 fixation by annual legumes. *Field Crops Research* **65**, 211-228.

Vance C.P. (2008) Carbon and Nitrogen Metabolism in Legume Nodules. In: Dilworth M.J., James E.K., Sprent J.I., Newton W.E. (eds) Nitrogen-fixing Leguminous Symbioses. Nitrogen Fixation: Origins, Applications, and Research Progress **7**. *Springer, Dordrecht*.

Vandenkoornhuyse, P., Quaiser, A., Duhamel, M., Le Van, A., & Dufresne, A. (2015). The importance of the microbiome of the plant holobiont. *New Phytologist* **206**, 1196-1206.

Vicente, J., & Dean, D. (2017). Keeping the nitrogen-fixation dream alive. *Proceedings of the National Academy of Sciences* **114**, 3009-3011.

Vidal, A., Alvarez, J., Araus, V., Riveras, E., Brooks, M., Krouk, G., Ruffel, S., Lejay, L., Crawford, N., Coruzzi, G., & Gutiérrez, R. (2020). Nitrate in 2020: Thirty Years from Transport to Signaling Networks. *The Plant Cell* **32**, 2094–2119.

Walker, M., Johnsen, S., Rasmussen, S., Popp, T., Steffensen, J., Gibbard, P., Hoek, W., Lowe, J., Andrews, J., Björck, S., Cwynar, L., Hughen, K., Kershaw, P., Kromer, B., Litt, T., Lowe, D., Nakagawa, T., Newnham, R. and Schwander, J. (2009). Formal definition and dating of the GSSP (Global Stratotype Section and Point) for the base of the Holocene using the Greenland NGRIP ice core and selected auxiliary records. *Journal of Quaternary Science* **24**, 3-17.

Walker, M., Johnsen, S., Rasmussen, O., Steffensen, P., Popp, T., Gibbard, P., Hoek, W., Lowe, J., Andrews, J., Björck, S., Cwynar, L., Hughen, K., Kershaw, P., Kromer, B., Litt, T., Lowe, J., Nakagawa, T., Newnham, R., Schwander, J. (2008). The Global Stratotype Section and Point (GSSP) for the base of the Holocene Series/Epoch (Quaternary System/Period) in the NGRIP ice core. *Episodes* **31**, 264-267.

Wang, L., Rubio, C., Xin, X., Zhang, B., Fan, Q., Wang, Q., Ning, G., Becana, M., & Duanmu, D. (2019). CRISPR/Cas9 knockout of leghemoglobin genes in *Lotus japonicus* uncovers their synergistic roles in symbiotic nitrogen fixation. *New Phytologist* **224**, 818-832.

Wang, L., Tan, Q., Fan, Q., Zhu, H., Hong, Z., Zhang, Z., & Duanmu, D. (2016). Efficient Inactivation of Symbiotic Nitrogen Fixation Related Genes in *Lotus japonicus* Using CRISPR-Cas9. *Frontiers in plant science* **7**, 1333.

Wang, Y., Cheng, Y., Chen, K., & Tsay, Y. (2018). Nitrate Transport, Signaling, and Use Efficiency. *Annual Review of Plant Biology* **69**, 85-122.

Weinthal, D., Tovkach, A., Zeevi, V., & Tzfira, T. (2010). Genome editing in plant cells by zinc finger nucleases. *Trends in Plant Science* **15**, 308-321.

Wolter, F., & Puchta, H. (2019), *In planta* gene targeting can be enhanced by the use of CRISPR/Cas12a. *The Plant Journal* **100**, 1083-1094.

Woods, J., Williams, A., Hughes, J. K., Black, M., & Murphy, R. (2010). Energy and the food system. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **365**, 2991–3006.

Wopereis, J., Pajuelo, E., Dazzo, F.B., Jiang, Q., Gresshoff, P.M., De Bruijn, F.J., Stougaard, J. & Szczyglowski, K. (2001). Short root mutant of Lotus japonicus with a dramatically altered symbiotic phenotype. *The Plant Journal* **23**, 97-114.

Wu, K., Wang, S., Song, W., Zhang, J., Wang, Y., Liu, Q., Yu, J., Ye, Y., Li, S., Chen, J., Zhao, Y., Wang, J., Wu, X., Wang, M., Zhang, Y., Liu, B., Wu, Y., Harberd, N., & Fu, X. (2020). Enhanced sustainable green revolution yield via nitrogen-responsive chromatin modulation in rice. *Science* **367**, No. 6478.

Xie, K., Minkenberg, B., & Yang, Y. (2014). Boosting CRISPR/Cas9 multiplex editing capability with the endogenous tRNA-processing system. *Proceedings of the National Academy of Sciences* **112**, 3570-3575.

Yoshida, H., Hirano, K., Sato, T., Mitsuda, N., Nomoto, M., Maeo, K., Koketsu, E., Mitani, R., Kawamura, M., Ishiguro, S., Tada, Y., Ohme-Takagi, S., Matsuoka, M., Ueguchi-Tanaka, M. (2014). DELLA protein functions as a transcriptional activator through the DNA binding of the INDETERMINATE DOMAIN family proteins. *Proceedings of the National Academy of Sciences* **111**, 7861-7866.

Zhang, J., Liu, YX., Zhang, N., Hu, B., Jin, T., Xu, H., Qin, Y., Yan, P., Zhang, X., Guo, X., Hui, J., Cao, S., Wang, X., Wang, C., Wang, H., Qu, B., Fan, G., Yuan, L., Garrido-Oter, R., Chu, C., & Bai., Y. (2019). NRT1.1B is associated with root microbiota composition and nitrogen use in field-grown rice. *Nature Biotechnology* **37**, 676–684.

Zhao, L., Li, M., Xu, C., Yang, X., Li, D., Zhao, X., Wang, K., Li, Y., Zhang, X., Liu, L., Ding, F., Du, H., Wang, C., Sun, J. and Li, W. (2018). Natural variation in GmGBP1 promoter affects photoperiod control of flowering time and maturity in soybean. *The Plant Journal* **96**, 147-162.

Zhu, S., Vivanco, J., & Manter, D. (2016). Nitrogen fertilizer rate affects root exudation, the rhizosphere microbiome and nitrogen-use-efficiency of maize. *Applied Soil Ecology* **107**, 324-333.

Appendix A: Genomic and cassette sequences

> *L. japonicus HAR1* promoter (yellow highlights are Cas9 target sites, blue highlights are Cas12a target sites, red letters represent PAM sequences, and the green highlight is the ATG start codon)

GACCAATAGATTTAATATTTTTAATTTGAATAGTTTAGATCAATTACCAATAAAAAAATTCGTT
TGTTAAATATATATTGTTTTACTGTTATACCCTTCTCCCTTCCCACGTTTCCCCCACCCCACCA
CCGTGCCTGATAACCCACGGCGGAGTCCAGCATGCCGGCCATGGAGATTGATGTGGTCACGACA
AGCCTTGCCGACATCCTTACTGTGAGGAGGTTGAGGGTAAGCTTCAGGCACTGGACACGAGAGA
GAAAAGGAAGATGTTGAAGAAGACGACCTTCGTCACGTCGACACTGTCGTCGCAGGCGGAACAA
GCACCAAGCCCACCCCACCTCCACCACCACCACCACCCAACCCCTTCTCCATCTGAGATCTGGA
AAGTTTGGATTTGTAAAACAGGGAGCTGGATCTGAAAAGTTTTGATTTAATTTGTCAAAATCAT
AAGTGAATTTCAGAATCTCCACCATGGTTCAGCTTTGAATTAACCCGCAGTAGGAGATTCAGAT
AACAACTGGAAAAAAATGGGGAATGCAAAACACATTATTGAGAAAAGGAAAAAGAAAATTAGAT
GCTTGTATGAAGAAAGTGTTGATTTTTCTATACACTATTGTTTGATTTTTGTTATTTCTTGGTT
TTGACGGTGTGCGACTTTGATTTGAGAAGCCTGTGGTGAAAATACAGATCTGGTTTAATCTGAG
TTCTTTCCCCTTCCCCTTTCTTTGTTCTCCACCGTTGGATTTCTAAACATATATTGCAGATTTT
GATGGTGTGTTATGGTGTTAAAACCCAGATTTAGGGGACCAAAGGGACCAAAATCACGTTAGCA
TTGAAGGTCTCCAAGATGATCATGGAGAAACTTTGTATTGTTGCCTCAACAACATGTGTGGTGA
ATTGCTGAAATTGCTGGGTTTGTGGTAGGGGTGTCCATAGGTCGGGTACGGTCGGTTTCGGGCC
GACGCCGTCGGTTACGGCGGGTGAAAAACACCCAACCATCAGGCCCGCCACCGACCATCCAGCA
GAATCGGGTTTATCGGGTTCGGGTTTTTCGGGTTACATTCGGTTCGGTCGGTGCTGTCGGTTTT
TTATAAAAAAGAAAATCACAAAAAGAGAGGAAGATCCAGCAAATAGAAACCCTAACCCTCAACA
AATCATAGAAAAGACAGAAACTTTTCAAAAACATGCATTCAATTCCAATACATATTACATAGAT
CAAGAATTCTTAGAAGAAAAAAGACCTAATTTTTACACTGAAAATAAACCATCACTGTAACCAG
ATCAAGAGAGGCCATGAACGGTGTCCTCTTCACACCATATTTTGAAGATCGAGCCAGATCAAGA
GAGGGAGATATAGATCGCGAGAGGGAGAGTCAGATCAAGTGAGGTAGAGCCAGAGACCAGCGAC
CTGGGTTGGTGGTGAGAGACGGTGTTCGATGGTGGCGACGTAACCCCCTCCATCGTCGCTGCTC
CAACCAAATGCAATCCATCATCTGGTCCTAGTACGATTCATCGATGGTGGGGACTTAACCCCCT
CCGTCGTGCGAGGGCGACGGCGAACAAAGAAGAAGGAGGCTCGAAACCGTGTGGGTGAGGGTGG
AAGCGCATGAAGAAGGAGGCGACGATGACGCGCGGTGTTTGAGAGGAGGTGGCGGCGCTGGGTG
GTGCTTAGGGTTTGCAGGTTGGTGCTCTGAAGATGTGTGACTTTGTGTGTGTGACGGCTTGATA
GATAGAAAGTGAGGTTTGTAACTTAAAAAGATAAAAGTAGGGTTTCTAATGGACTATCACTATT
GGACTGGATTGGGCCTATTTTTTTTTTTTAAACATAGATACTTCGGTCGGTTCGGTCGGGCTCG
AGCCCAATCCGACACCGACCGAACCAACTCGCCTAAAACCGCATTTTTGCACCCGGTTAACCCG
AAATCCGGCCCGACCGGCCCGTTTGGCCAAAAAAGGCCTCGGTTTAGTCGGGCCCGGTTCGGGC
CGGTTGTACACGGACAGGCCTAGTTTGTGGAGCGGAAAGGTAGAGAGAGAAAAGAAGAAGAGAG
AAAAGTTAAAAGAAAAAATGAGTATTTTTGTAATTTAGATGGTGAATAGTAACTGCAGCACCGT
GCGTCAGTTTGTCAGTTACAAACTGACTCACGGTAGATGCAGCCTCAATTATCTCTTATTCATT
TTCACTCAATTTTTCTTTCTATAGATAGTTATGAGAAATGAGTAATACTTCCATCTCTCACACT
CTCTTTTTTTTCTCTTAAAAAAGGTGGTATACACCCAAATGAGTGTGTGTCTGCACATAATCAT
CTTTTTCCTAAGTAGAATCTCTCAATATTTTCGGTTGCCAACTTGCAAAGGGCTTGATTATAAC
ATAACTGTAGTCCCATACTCCCATATCTATGTAGGGAAGTACAAAAATCAGGCTTAGAGAATAA
AAAGAAAGAACTAGGAGAACTAGGACTGTCTGCCATGTAAAAGACGAAGAAATCAAGAGGGCCA
ATAATAGAAGATTTTTTTTGAGCACTTCGCGGACACTAACTCGTTGTCGCTATCATTAGAGGGG

TTGACAAATGACAGTATTGAACATTTTTTTGAAGATGTGTACCACTCATCCATTTATGAGTGGT
TCCCAGTTGCCAAAACTATGAACGAAATGTTGAAGGAATAACAGGAATCCATTCTAGAAGCTTT
GCTATTTGGGAATCTAGAGTTCTTGACATTGTACCCTGTATTAAGCATTAAATTAAAAAAGGAT
GAAAGAAAATGAGCATCTTTCTTTTCTTTTCTATTATCAATATTCGTATGCCTCATCCAATTAA
AAACATGAAAATTAAAGATAACAAATGACATGAATCAGTTATATACGCAAATCAATATTAAATA
TTTTTTAAAAAATCTATATACTATAACCCATCAGAAAATTGGTGTGGGTTTAGCAGAATTCGCC
ATTTATTATATATAATCAATAAAGTTAATTAAAAGAATTCTGTAAGTTGTTATATAATCTATCT
TCTCATTTTCTCCTTTTTTTTATCATATTATCAAATATATATATATATATATATATATATCC
CATTTTCTTTTATCTTTAGTAAAATTTGTTACACTCCAAGATGTAGTGTACACTCATAACTAAC
TATATCTAATGATTAAAAAAAATTAACACATGCACTGGAAAATAAATTTGGTAGATACCATTTT
TTTTTTTTTTTTGATAAGCTGGTAGGTACCATATAATATGTATAAGAAACATGAAATGAGAGAA
GAAAGAAGTGGTGGGAAAGGAAGACATAAGTGAAAATTTTGATTCAGAAATGATGCACGGTGAC
ACAATCCAATTAACGTATCTATGAATATTATATTAGTGGTTTTTAATCACTGCTAATGTGTATG
TCAGTTAAATTTCAACAGAAACCCGTGTATTTACATATAATCTCATATTTTGATTAGACAATAA
CCTAACAAACTCTATCGTTTGGTGTATTGTGGCGAGCTAGCTAGTCTTCATTCTCTTCTCCTTA
ATTAGGTACAGAATCACTTTCAATTCAATAAAATAATTTACTTTAAAATGCATTGTTGCATGAT
ACACACTTTAATCACATATAGAATACGATTCTGCAATATCTAGTGGATGATCGTGGGAGCAGAG
AATTTATGCTTAGTTTGCCTTTTCCACGAGAGTCACAACTATCCCCAGTTATACACACTCCAAG
TGTTCCACGTAAAACAACTTTTCTCCTTTCATTTTATACTCAACAACTTTTTGTTTCATTCCAA
AGTGAGACTTATACAAAGCTTAAATTAAACCTCTGTTACTGGTTACAAAAAGATTTCATGTGTC
TTCCCACAAAAGCTAAGTCCACGAGAGAGAGTGGACTTTCACTAGTGGCGCCAAGCATATTACA
GTATTATACGCCAGCCACCAAATTATATGTTAACCAATCACATACACTAATTTAGTATAATCAT
GATTATAAATAGACACTCAGCATAACAAATGCATG

> Nine (9) sgRNA cassette (found in *CRISPR-HAR1p-1*; yellow highlights represent *LjtRNA*$_{Gly}$, unmarked sequences are sgRNAs, and red letters represent *BbsI* cloning sites, pink highlight is random sequence to protect the *BbsI* cloning site)

GTTCGGGTCTTCAACAAAGCACCAGTGGTCTAGTGGTAGAATAGTACCCTGCCACGGTATAGAC
CCGGGTTCGATTCCCGGCTGGTGCATTAAACATAGATACTTCGGTGTTTTAGAGCTAGAAATAG
CAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTAAC
AAAGCACCAGTGGTCTAGTGGTAGAATAGTACCCTGCCACGGTATAGACCCGGGTTCGATTCCC
GGCTGGTGCATTTGATTCAGAAATGATGCAGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGG
CTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTAACAAAGCACCAGTGGTC
TAGTGGTAGAATAGTACCCTGCCACGGTATAGACCCGGGTTCGATTCCCGGCTGGTGCAAAGGA
GGCGACGATGACGCGGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAA
CTTGAAAAAGTGGCACCGAGTCGGTGCTTTTAACAAAGCACCAGTGGTCTAGTGGTAGAATAGT
ACCCTGCCACGGTATAGACCCGGGTTCGATTCCCGGCTGGTGCAAGATCAAGAGAGGCCATGAA
GTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCA
CCGAGTCGGTGCTTTTAACAAAGCACCAGTGGTCTAGTGGTAGAATAGTACCCTGCCACGGTAT
AGACCCGGGTTCGATTCCCGGCTGGTGCATCCAGCAGAATCGGGTTTATGTTTTAGAGCTAGAA
ATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTT
TAACAAAGCACCAGTGGTCTAGTGGTAGAATAGTACCCTGCCACGGTATAGACCCGGGTTCGAT
TCCCGGCTGGTGCACGTTTGGCCAAAAAAGGCCTGTTTTAGAGCTAGAAATAGCAAGTTAAAAT
AAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTAACAAAGCACCAGT
GGTCTAGTGGTAGAATAGTACCCTGCCACGGTATAGACCCGGGTTCGATTCCCGGCTGGTGCAC

CGTGCCTGATAACCCACGGGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCGTTA
TCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTAACAAAGCACCAGTGGTCTAGTGGTAGAA
TAGTACCCTGCCACGGTATAGACCCGGGTTCGATTCCCGGCTGGTGCATTTGTGGTAGGGGTGT
CCATGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGT
GGCACCGAGTCGGTGCTTTTAACAAAGCACCAGTGGTCTAGTGGTAGAATAGTACCCTGCCACG
GTATAGACCCGGGTTCGATTCCCGGCTGGTGCAGATCGAGCCAGATCAAGAGAGAAGACCTGTT
T

> Eight (8) sgRNA cassette (found in *CRISPR-HAR1p-2*; yellow highlights represent *LjtRNA*$_{Gly}$ and unmarked sequences are sgRNAs, *BsaI* cloning sites not included)

AACAAAGCACCAGTGGTCTAGTGGTAGAATAGTACCCTGCCACGGTATAGACCCGGGTTCGATT
CCCGGCTGGTGCATTAAACATAGATACTTCGGTGTTTTAGAGCTAGAAATAGCAAGTTAAAATA
AGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTAACAAAGCACCAGTG
GTCTAGTGGTAGAATAGTACCCTGCCACGGTATAGACCCGGGTTCGATTCCCGGCTGGTGCATT
TGATTCAGAAATGATGCAGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCGTTAT
CAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTAACAAAGCACCAGTGGTCTAGTGGTAGAAT
AGTACCCTGCCACGGTATAGACCCGGGTTCGATTCCCGGCTGGTGCAAGGAGGCGACGATGAC
GCGGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTG
GCACCGAGTCGGTGCTTTTAACAAAGCACCAGTGGTCTAGTGGTAGAATAGTACCCTGCCACGG
TATAGACCCGGGTTCGATTCCCGGCTGGTGCAAGATCAAGAGAGGCCATGAAGTTTTAGAGCTA
GAAATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGC
TTTTAACAAAGCACCAGTGGTCTAGTGGTAGAATAGTACCCTGCCACGGTATAGACCCGGGTTC
GATTCCCGGCTGGTGCATCCAGCAGAATCGGGTTTATGTTTTAGAGCTAGAAATAGCAAGTTAA
AATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTAACAAAGCACC
AGTGGTCTAGTGGTAGAATAGTACCCTGCCACGGTATAGACCCGGGTTCGATTCCCGGCTGGTG
CACGTTTGGCCAAAAAAGGCCTGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCG
TTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTAACAAAGCACCAGTGGTCTAGTGGTA
GAATAGTACCCTGCCACGGTATAGACCCGGGTTCGATTCCCGGCTGGTGCACCGTGCCTGATAA
CCCACGGGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAA
AGTGGCACCGAGTCGGTGCTTTTAACAAAGCACCAGTGGTCTAGTGGTAGAATAGTACCCTGCC
ACGGTATAGACCCGGGTTCGATTCCCGGCTGGTGCATTTGTGGTAGGGGTGTCCATGTTTTAGA
GCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCG
GTGCTTT

> Two (2) gRNA cassette (Cas12a gRNA 1 and 2; yellow highlights represent DRs and red letters represent the *BsaI* cloning sites)

GGTCTCGAAAAATTTCTACTAAGTGTAGATCCTTTTCCACGAGAGTCACAACTAATTTCTACT
AAGTGTAGATCCACCACTTCTTTCTTCTCTCATAATTCGAGACC

> Two (2) gRNA cassette (Cas12a gRNA 3 and 4 yellow highlights represent DRs and red letters represent the *BsaI* cloning sites)

GGTCTCGAAAAAATTTCTACTAAGTGTAGATGCAGAATTCGCCATTTATTATATAATTTCTACT
AAGTGTAGATTGATGGGTTATAGTATATAGATTAATTCGAGACC

> Two (2) gRNA cassette (Cas12a gRNA 5 and 6 yellow highlights represent DRs and red letters represent the *BsaI* cloning sites)

GGTCTCGAAAAAATTTCTACTAAGTGTAGATTATTATCAATATTCGTATGCCTCAATTTCTACT
AAGTGTAGATTTCGTCTTTTACATGGCAGACAGTAATTCGAGACC

> Two (2) gRNA cassette (Cas12a gRNA 7 and 8 yellow highlights represent DRs and red letters represent the *BsaI* cloning sites)

GGTCTCGAAAAAATTTCTACTAAGTGTAGATCGCTCCACAAACTAGGCCTGTCCAATTTCTACT
AAGTGTAGATACCACAGGCTTCTCAAATCAAAGAATTCGAGACC

> Two (2) gRNA cassette (Cas12a gRNA 9 and 10 yellow highlights represent DRs and red letters represent the *BsaI* cloning sites)

GGTCTCGAAAAAATTTCTACTAAGTGTAGATTCTCTCGTGTCCAGTGCCTGAAGAATTTCTACT
AAGTGTAGATATTTGAATAGTTTAGATCAATTAAATTCGAGACC

> Ten (10) gRNA cassette (Cas12a multiplex yellow highlights represent DRs, *BsaI* cloning sites not included)

AATTTCTACTAAGTGTAGATCCTTTTCCACGAGAGTCACAACTAATTTCTACTAAGTGTAGATC
CACCACTTCTTTCTTCTCTCATAATTTCTACTAAGTGTAGATGCAGAATTCGCCATTTATTATA
TAATTTCTACTAAGTGTAGATTGATGGGTTATAGTATATAGATTAATTTCTACTAAGTGTAGAT
TATTATCAATATTCGTATGCCTCAATTTCTACTAAGTGTAGATTTCGTCTTTTACATGGCAGAC
AGTAATTTCTACTAAGTGTAGATCGCTCCACAAACTAGGCCTGTCCAATTTCTACTAAGTGTAG
ATACCACAGGCTTCTCAAATCAAAGAATTTCTACTAAGTGTAGATTCTCTCGTGTCCAGTGCCT
GAAGAATTTCTACTAAGTGTAGATATTTGAATAGTTTAGATCAATTA

Appendix B: Primer Sequences

Table 1. Primers used for PCR amplification and sequencing of the target locus.

Primer Name	Sequence (5' → 3')
HAR1p-F	ACTGTTATACCCTTCTCCCTTCCC
HAR1p-mid-R	TGTGCAGACACACACTCATTTGG
HAR1p-mid-F	TGTGGAGCGGAAAGGTAGAGAG
HAR1p-R	ACTCTCGTGGAAAAGGCAAACTAAG

Table 2. Primers used to generate cloning fragment to create the *CRISPR-HAR1p-2* vector.

Primer Name	Sequence (5' → 3')
HAR1-gRNA-F-BsaI	ATGCGGTCTCGAAACATGGACACCCCTACCACAAA
HAR1-gRNA-R-BsaI	GTACGGTCTCTTGCACACGTGTTCGGGTCTTCAACA

Table 3. Primers for PCR genotyping of the *CRISPR-HAR1p-1/2* tDNA.

Primer Name	Sequence (5' → 3')
LjU6-cassette-F	GCTGAGGAGACTTGTGCTAT
STU-C9-tDNA-F	AGAGGAAGGTTTGAGGATCTA
LjHAR-cassette-R	CATAAACCCGATTCTGCTGG

Table 4. Primers for replacing the *ZmUBQp* in pGEL032.

Primer Name	Sequence (5' → 3')
AscI-UBQ-F	GGCGCGCCGGCGCGCCGGAGAGAGGATTTTGAGGAAATAATTAATTG
UBQ-SbfI-R	CCTGCAGGCCTGCAGGCTGTAATCACATCAACAACAG
AscI-2x35S-F	GGCGCGCCGGCGCGCCGGTCAACATGGTGGAGCACGAC
2x35S-SbfI-R	CCTGCAGGCCTGCAGGGGGATCCTCTAGAGTCGAGGTCC

Table 5. Primers used in cRT-PCR.

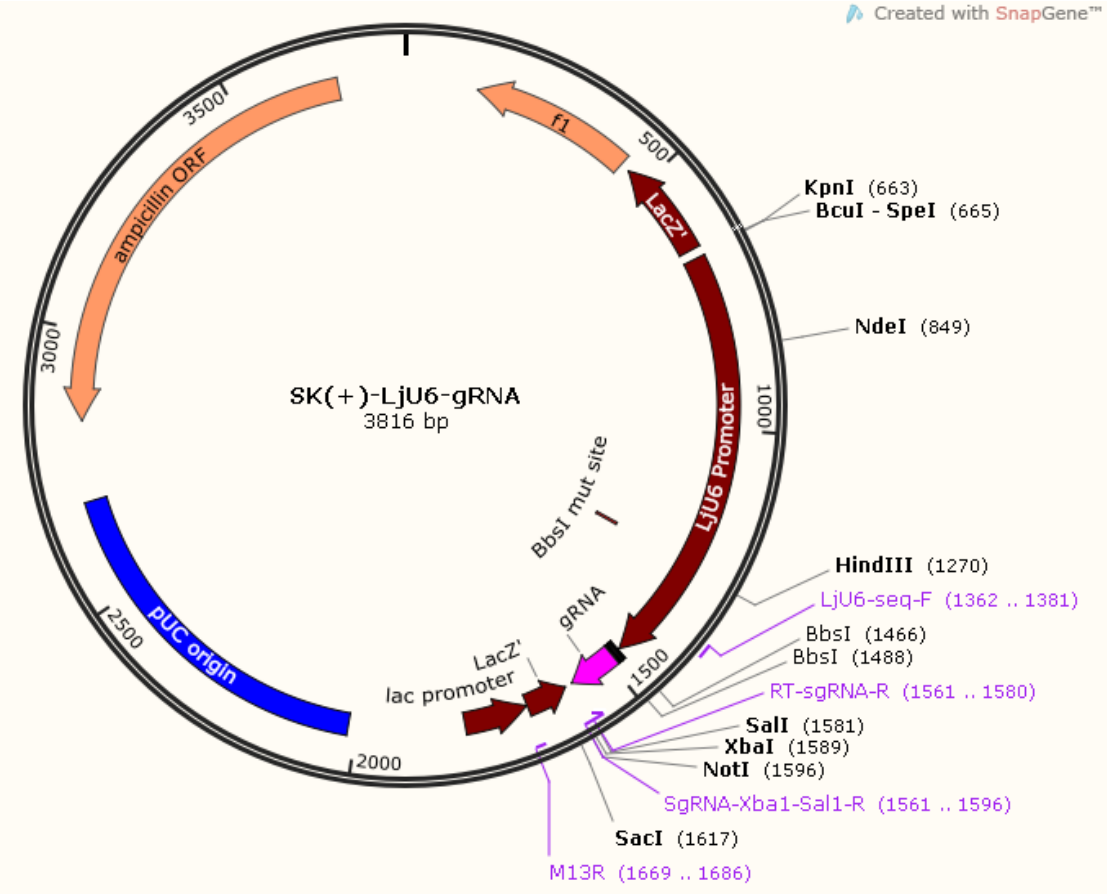
Primer Name	Sequence (5' → 3')
sgRNA-F	GTTTTAGAGCTAGAAATAGC
sgRNA1-R	AAACACCGAAGTATCTATGT
sgRNA4-R	AAACTTCATGGCCTCTCTTG
sgRNA5-R	AAACATAAACCCGATTCTGC
sgRNA8-R	AAACATGGACACCCCTACCA

Table 6. Sequencing primers.

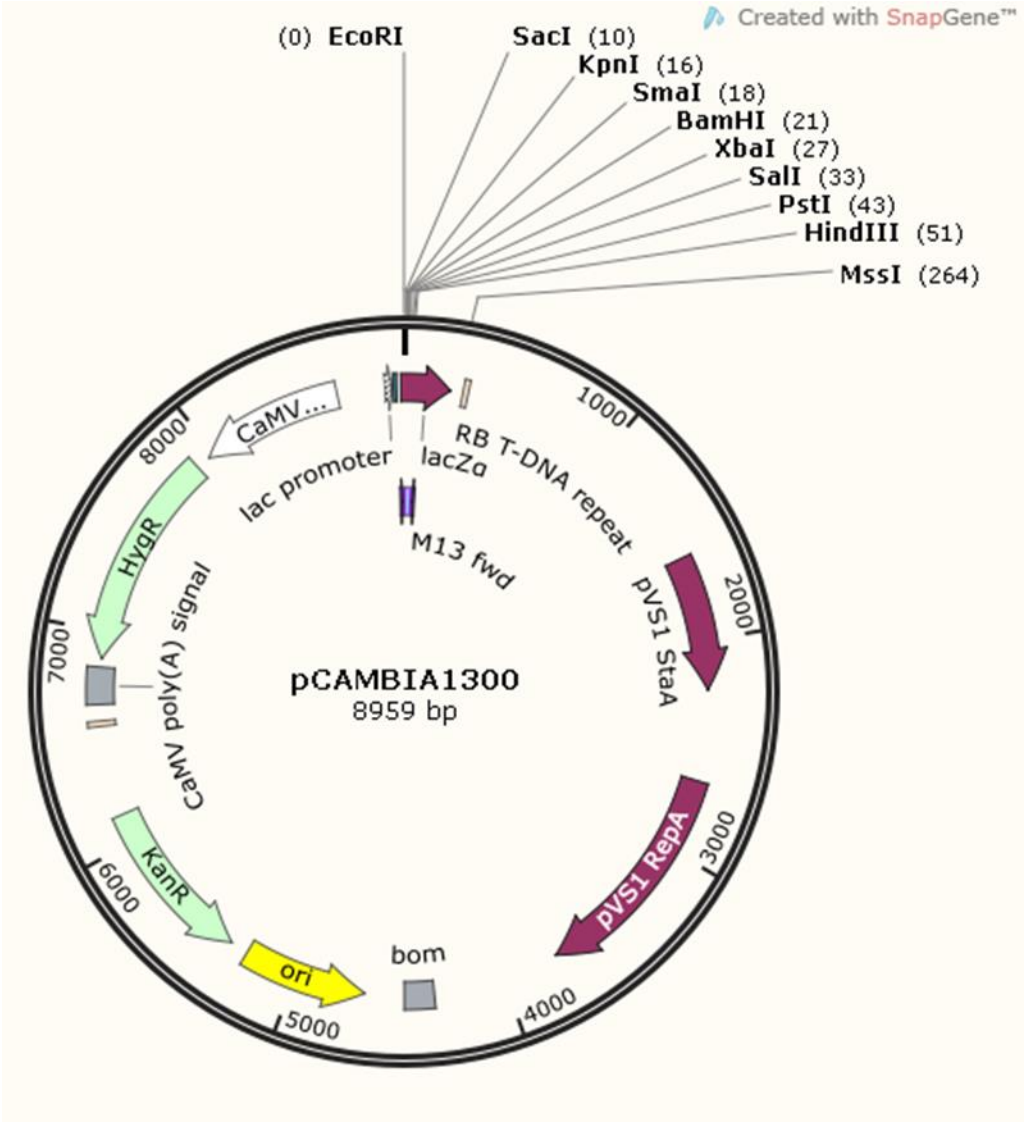
Primer Name	Sequence (5' → 3')
M13 Forward (-20)	GTAAAACGACGGCCAG
M13 Reverse	CAGGAAACAGCTATGAC
HAR1p-Seq5-F1	GTTTTGACGGTGTGCGACT
HAR1p-Seq5-F2	CAGCAAATAGAAACCCTAACCC
HAR1p-Seq5-F3	CGATGGTGGGGACTTAAC
HAR1p-Seq5-F4	GCCTAGTTTGTGGAGCGGAAAG
HAR1p-Seq5-R1	CCAGTCCAATAGTGATAGTC
HAR1p-Seq5-R2	TGGTCTCTGGCTCTACCTCACTTG
HAR1p-Seq5-R3	CCAGCAATTTCAGCAATTCACC
HAR1p-Seq5-R4	CCAGTTGTTATCTGAATCTCCTAC

Appendix C: Vector maps

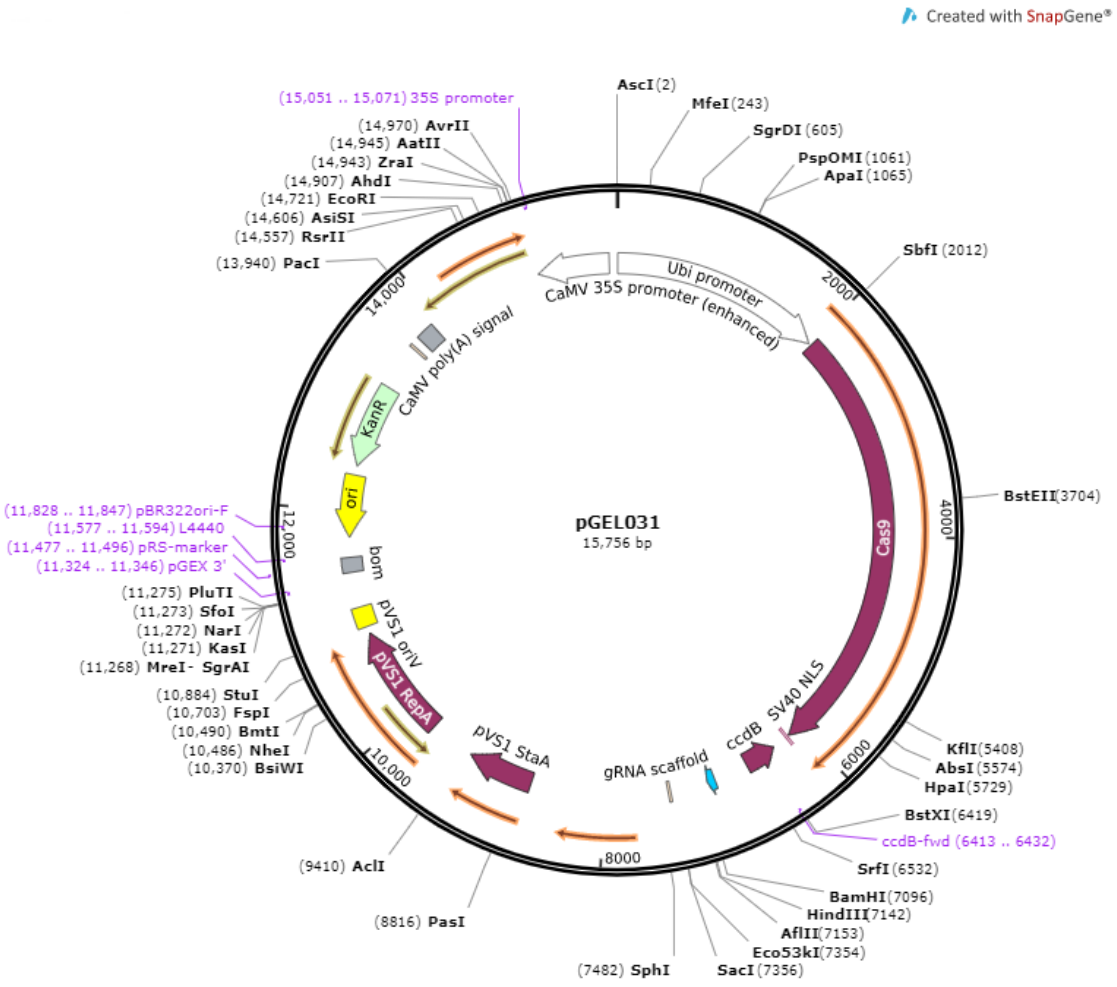
> pBluescript SK (+) (Wang et. al., 2016)



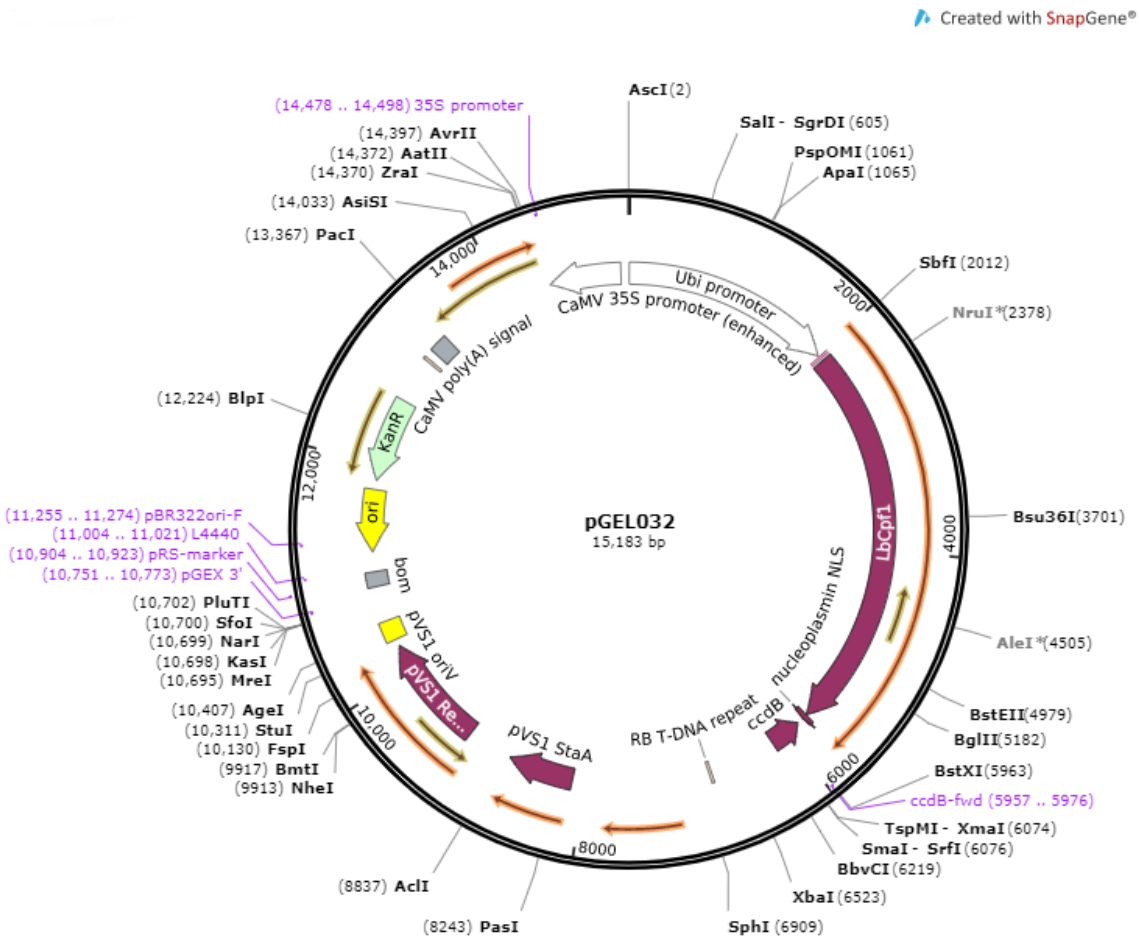
> pCAMBIA1600 (Wang et. al., 2016)



> pGEL031 (Tang et. al., 2019)



> pGEL032 (Tang et. al., 2019)



Appendix D: Bioinformatics pipeline

pipeline.R:

```
# Provides instructions and R-commands used to generate analysis of natural variation using data
# obtained from https://soykb.org/public_data.php


rm(list=ls())

setwd("D:/UWO/MSc/Thesis/NARK natural variation")


library(ggplot2)

library(dplyr)

library(car)

library(tidyr)


# Parse the SNP data

### Run the retrieve_SNPs.py script on Chr12_15x_SNPs.csv using NARK positional
information (2881838..2890668). Put the output file through the csv_transposer.py script (keep
name the same).

### Run the retrieve_SNPs.py script on Chr12_40x_SNPs.csv using NARK positional
information (2881838..2890668). Put the output file through the csv_transposer.py script (keep
name the same).

### Run the setup_SNP.py script using the GmNARK.fasta file and NARK positional
information (2881838..2890668).


# Load in and align the parsed SNP data

GmNARKsetup <- read.csv('GmNARK_setup.csv', header = T)

GmNARK_15 <- read.csv('GmNARK_15x_SNPs.csv', header = F)

GmNARK_40 <- read.csv('GmNARK_40x_SNPs.csv', header = F)

GmNARK <- merge(GmNARK_15, GmNARK_40, by.x = "V2", by.y = "V2", all = T)

GmNARK_Aligned <- merge(GmNARKsetup, GmNARK, by.x = "Position", by.y = "V2", all =
T)

write.csv(GmNARK_Aligned, 'GmNARK_aligned.csv')
```

Count SNPs found in the parsed data

Open the Gm#_aligned.csv files in excel, ctrl F, replace all, find = NA, replace with =

Rename the "ref.x" column to "ref" and rename "Position" to "Pos"

Delete the "ref.y" column and the first column

Save the file and close it

Run the Gm#_aligned.csv files through **remove_redundant_SNPs.py**

Open Gm#_aligned_cleaned.csv in excel

In column after last full one create a column named "Counts"

In next row (first row of positions in locus) enter the formula =COUNTIF(range first variety:last variety, "*")

Drag the formula to capture all the positions in the locus, saveas Gm#_aligned_cleaned_counted.csv

```
GmNARK_cleaned <- read.csv('GmNARK_aligned_cleaned_counted.csv', header = T)

GmNARK_counted <- select(GmNARK_cleaned, Pos, ref, Counts)

write.csv(GmNARK_counted, 'GmNARK.csv')
```

Generate the graphical output

```
GmNARK_graph <- ggplot(data = GmNARK_counted, aes(x = reorder(Pos, desc(Pos)), y = Counts)) +
  geom_bar(stat = 'identity', colour = 'black') +
  theme_bw()
GmNARK_graph
```

Perform statistical analysis in using excel

Open GmNARK.csv in excel

Designate a space for a table with the following columns: Polymorphic Positions, Total Position, Polymorphism Coefficient, Proportion; and rows: Entire Locus, Coding, Non-Coding (you can also analyze individual components of the gene, such as the UTRs or individual exons)

Use the formula =COUNTIF(count cell representing the first position of the locus:count cell representing the last position of the locus, ">0") to fill in the Polymorphic Positions column.

Use the formula =COUNTIF(count cell representing the first position of the locus:count cell representing the last position of the locus, "*") to fill in the Total Positions column.

Use the formula =Polymorphic Positions/Total Positions to fill in the Polymorphism Coefficient column.

Use the formula =1/Polymorphism Coefficient to fill in the Proportion column.

retrieve_SNPs.py

```
"""
Create a csv file containing information on only the relevant positions.
Arguments entered for sys.argv[n]: "ChrXX_##x.csv" "outfile.csv" "locus start position" "locus end position"
"""

import sys
import csv

with open(sys.argv[1],'r') as handle:  # you need the original .tab file saved as a .csv
    SNP_data = csv.reader(handle)
    outfile = csv.writer(open(sys.argv[2], 'w', newline=''))
    lower_bound = int(sys.argv[3])  # the start of the locus of interest
    upper_bound = int(sys.argv[4])  # the end of the locus of interest
    for row in SNP_data:  # write the header row
        list(row)
        if row[0].startswith('#'):
            outfile.writerow(row)
        else:  # retrieve all sequence variants at the given locus
            if row[0].startswith("Chr"):
                if int(row[1]) >= lower_bound and int(row[1]) <= upper_bound:
                    outfile.writerow(row)

print('Done!')
```

setup_SNP.py

```
"""

Take reference Williams82 sequence and create a seqeunce document that aligns the reference sequence to their respective postions.

Arguments entered for sys.argv[n]: "reference sequence.fasta" "outfile.csv" "range - lower limit" "range - upper limit"

"""


import csv

import sys


# import working file and reference sequence

refseq = open(sys.argv[1], 'r')  # must be fasta format

outfile = csv.writer(open(sys.argv[2], 'w', newline=''))

lower_bound = int(sys.argv[3])

upper_bound = int(sys.argv[4]) + 1

locus = range(lower_bound,upper_bound)


# write the positional values for the locus

pos = list(locus)

pos.insert(0,'Position')

outfile.writerow(pos)


# write the refseq file into a comma delimited format containing only sequence

for line in refseq:

    if not line.startswith('>'):

        seq_as_list = list(line)

        outfile.writerow(['ref'] + seq_as_list)


print('Ref Seq Mapped!')
```

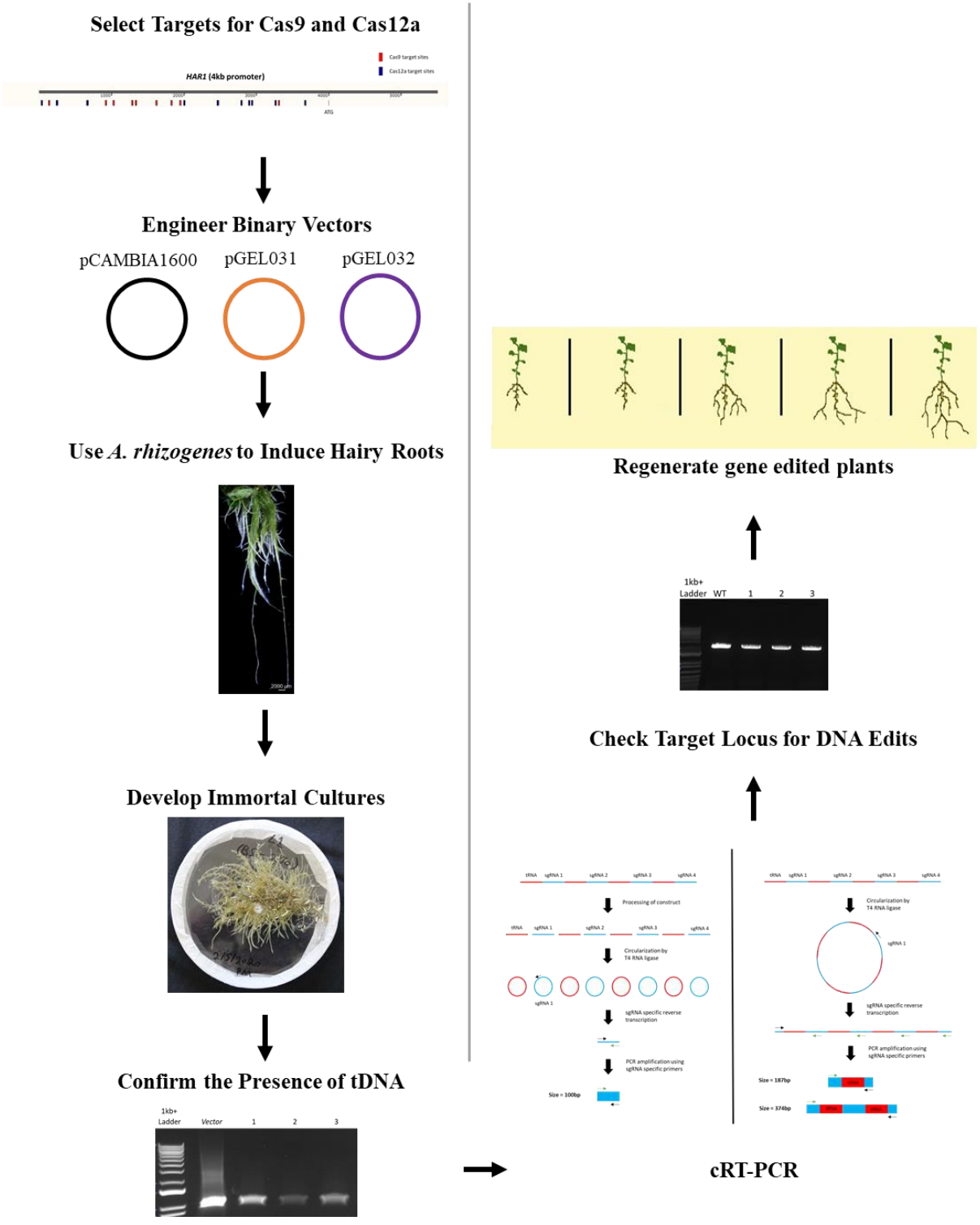
csv_transposer.py

```
"""
Input csv = row1: [x1,x2,x3]; row2: [y1,y2,y3]; row3: [z1,z2,z3]
Output csv = row1: [x1,y1,z1] -> row2: [x2,y2,z2] -> row3: [x3,y3,z3]
No size limit and “#rows =/= #columns” is allowed
Arguments entered for sys.argv[n]: "input.csv" "output.csv"
"""

import sys
import pandas as pd

pd.read_csv(sys.argv[1], header=None, dtype='string').T.to_csv(sys.argv[2], header=False, index=False)

print('Done!')
```

remove_redundant_SNPs.py

```
"""
Removes a countable SNP if it is the same as the reference at all positions.
Arguments entered for sys.argv[n]: "gene_aligned.csv " "gene_aligned_cleaned.csv "
"""

import sys
import csv

infile = csv.reader(open(sys.argv[1], "r"))
outfile = csv.writer(open(sys.argv[2], "w", newline=""))

# locate matching string and replace with a blank
for row in infile:
    if row[0].startswith('Pos'):
        outfile.writerow(row)
        print(len(row))
    else:
        a = row[0:2]
        for i in row[2:]:
            if i == row[1]:
                a.append('')
            else:
                a.append(i)
        print(len(a))
        outfile.writerow(a)

print("done!")
```

Appendix E: Flowchart of the Experimental Design



Petar Miletic – Curriculum Vitae

Department of Biology, Western University

EDUCATION

B.Sc. (Honours), Specialization in Genetics, Department of Biology, Western University, Ontario, Canada. Completed: April 2019
M.Sc. Cellular and Molecular Biology, Department of Biology, Western University, Ontario, Canada. Completed: November 2021

RESEARCH INTERESTS

molecular genetics; targeted genome modifications; CRISPR; crop improvement; bioinformatics

EXPERIENCE

Graduate Research Assistant/Research Affiliate
Agriculture and Agri-Foods Canada and Western University; 2019-2021
Teaching Assistant
Biology 2601: Organismal Physiology; Fall 2019 and Fall 2020
Biology 2581: Genetics; Winter 2020 and Winter 2021
Undergraduate Research Assistant
The University of Western Ontario – Department of Biology; 2018-2019

AWARDS

Western Graduate Research Scholarship (Western University, London, ON); 2019-2021