A DISSERTATION FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Identification of functional elements involved in RNA splicing, domestication, and agronomic performance in mungbean genome

BY

DANI SATYAWAN

FEBRUARY 2017

MAJOR IN CROP SCIENCE AND BIOTECHNOLOGY

DEPARTMENT OF PLANT SCIENCE

Identification of functional elements involved in RNA splicing, domestication, and agronomic performance in mungbean genome

DANI SATYAWAN

DEPARTMENT OF PLANT SCIENCE
THE GRADUATE SCHOOL OF SEOUL NATIONAL UNIVERSITY

GENERAL ABSTRACT

Mungbean (*Vigna radiata*) is an important Asian pulse crop with annual cultivation area of 6 million hectares. The completion of mungbean reference genome provided a major boost for mungbean genomic research. Annotation of the reference genome data was improved by evaluating the prevalence of alternative splicing (AS) in the genome. At least 37.9% of mungbean genes undergo AS but there are indications that AS in mungbean is predominantly stochastic. A large proportion of AS isoforms exists at very low copy, or expressed at much lower level than the default transcript. Conservation in closely related species is also rare, with only 2.8%

of genes share conserved AS between mungbean and adzuki bean, and only 16 soybean genes share conserved AS with mungbean.

Using genotyping by sequencing (GBS) on 276 cultivated and wild mungbean accessions from around the world, it was found that nucleotide diversity among cultivated accessions decreases to 30% of the level found in wild accessions. Linkage disequilibrium also decays at longer interval in cultivated mungbean, where LD blocks are on average 4.6 times longer than wild accessions. Wild and cultivated accessions are clearly separated in phylogenetic and population structure analysis, and there are some correlations between geographic origin and subpopulation membership. Several loci were identified as possible candidate for regions that underwent positive selection during mungbean domestication. The genes in those intervals are enriched with genes associated with growth and reproductive traits.

SNP data obtained from GBS were also used for genome-wide association study (GWAS) to dissect the genetic background of several agronomic traits such as flowering time, maturity time, pod formation time, seed weight, number of seeds per pod, peak harvest, cumulative weekly harvest, final yield, and synchronicity. Phenotyping was performed on 222 cultivated accessions and using two different association methods at least 79 markers were found to be significantly associated with the traits at p-

value <0.0001. Some of the genes that intersect with significant markers share homology with soybean genes and QTL that could explain their role in trait formation, which makes them attractive candidates for follow up studies. The data can be used as a basis for mapping studies or parental selection in mungbean breeding programs.

Keywords: *Vigna radiata*; alternative splicing; genotyping by sequencing; domestication; GWAS; agronomic traits

Student number: 2014-30834

CONTENTS

LIST OF FIGURES

Figure I-6. Number of AS isoforms, according to distance from the regular splice site, for alternative donor and alternative acceptor isoform types.

Figure I-7. Proportion of DNA bases surrounding alternative splice types compared to the regular splice sites. AS isoforms with FPKM >10 were categorized as having high concentration (**H**), while those with FPKM<10 were grouped into the low concentration group (**L**).

Figure I-8. **(A)** AgriGO annotation of the relationships and significance level of enriched GO groups of the genes with conserved AS in mungbean and adzuki bean. Inside the boxes: numbers inside the brackets are the p-value, while the numbers on lower left sides are annotated/total number in query and on the lower right are annotated/total number in background/reference. **(B)** Gene ontology enrichment of genes with AS isoforms that are conserved in mungbean and adzuki bean.

Figure II-1. Distribution of sequence variants along the 11 chromosomes of mungbean. **(A)** Fst values between wild and cultivated accessions across the chromosomes (1-11). The peaks that intersect the dark green area are the 10% highest values, which indicate loci with the most differentiation between wild and cultivated accessions and probably play important roles in domestication. **(B)** Histogram of the number of variants along the chromosomes for wild (orange) and cultivated (black) accessions.

Figure II-2. In an unrooted neighbor-joining tree, cultivated mungbean (black) form a small and tight cluster compared to wild accessions (red) even though there were 233 cultivated accessions compared to just 42 wild accessions. This indicates a high genetic similarity among them and a significant reduction of genetic diversity compared to the wild population.

Figure II-3. Origins of mungbean accessions sequenced in this study, viewed in the context of their phylogenetic relationship. Unk denotes accessions with no known country of origin data. Red branches indicate wild accessions.

Figure II-4. Pattern of population structure calculated using Bayesian clustering in STRUCTURE program. Each column represents a single accession and colors in the columns represent genetic components contributed by each subpopulation when the accessions were divided into 2, 3, 4, and 5 subpopulation (k=2 to k=5). The accessions were ordered according to their phylogenetic relationship in a neighbor-joining tree.

Figure II-5. Average genetic contributions from each of the 3 population structure subgroups in cultivated accessions obtained from different countries. Only countries that contributed more than 3 accessions were included. Geographical origin seems to partly explain the membership of accessions from the same countries in population subgroups, as nearby countries with similar climates are dominated by the same subgroups.

Figure II-6. Plotting the two main components that explain the most variation from principal component analysis (PCA) in x axis and y axis reveals clustering of accessions that is consistent with their geographical origins. Accessions that are separated from other accessions from the same country could be the result of recent germplasm exchange.

Figure II-7. Plot of average linkage disequilibrium (measured as R square between markers in y axis) decay as markers are further separated from each other in the chromosome (x axis).

Figure II-8. Gene ontology enrichment of genes located in segments with Fst outlier found by LOSITAN according to their function in the cell. (A) Genes that showed enrichment in the molecular function groups seem to perform activities commonly associated with signal transduction proteins and transcription factors. (B) Genes belonging to cellular component group are enriched by membrane proteins.

Figure II-9. Gene ontology enrichment of genes located in segments with Fst outlier found by LOSITAN that are involved in biological processes. The genes are enriched for genes involved in developmental growth, meristem growth and maintenance, reproductive structure development (flowers, gametes, fruits, and seeds), organelle organization, and metabolic processes for minerals, protein, and lipids.

Figure III-1. Distribution and correlation among traits in the association panel. Histogram of the trait distribution is presented in the diagonal panels. Lower panels are matrices of scatter plots that show the correlations between traits, while the upper panels show the significance and absolute value of the correlations (r).

Figure III-2. Prolonged productivity of peduncles is one of the reasons of non-synchronous pod maturity in mungbean. New flowers are still formed even when earlier pods from the same peduncles matured already.

Figure III-3. Defining synchronicity in mungbean by the length of productive days (PRODAY) until weekly harvests hit non-productive threshold, which is defined as the period when the pod count is consistently less than 3 per plant when harvested each week. Synchronous plants should have shorter PRODAY compared to non-synchronous plants, so accession 1 here has a higher degree of synchronicity than accession 2. The y-axis is the number of pods that can be harvested each week from 5 plants, while the x-axis denotes the number of days since planting.

Figure III-4. Synchronicity was also defined as the ratio between cumulative yield at optimum harvest time (in this case at week 2 for both accession A and B) and the final yield at the end of the study. The ratio between M and F is higher in accession B, since it produced lesser number of pods after its

peak harvest time. A highly synchronous accession should have an M/F ratio of close to 1 as it stops producing pods after the peak harvest.

Figure III-5. Manhattan plot of p-values from GWAS when wild accessions were included in the analysis. Note the presence of numerous dots forming a horizontal line close to $-\log_{10}(p)=1$, which indicates the presence of population structure. Those dots passed the significance threshold in some traits, and they cannot be removed even after population structure, PCA, and kinship were accounted for.

Figure III-6. Manhattan plots of significant marker-trait association produced by GAPIT. The y-axis is $-\log_{10}$ of p-values of the markers' significance, while x-axis is the chromosome number and base pair positions of the markers.

Figure III-7. Manhattan plots of significant marker-trait association produced by TASSEL. The y-axis is $-\log_{10}$ of p-values of the markers' significance, while x-axis is the chromosome number and base pair positions of the markers.

LIST OF TABLES

Table III-1. List of accessions genotyped and phenotyped for GWAS to map agronomic traits.

Table III-2. Loci showing significant associations with the measured traits ($p<0.0001$). Each locus is defined as the interval between a significant marker and its flanking markers. Some loci at the end of the table are significant for multiple traits. Traits followed by parentheses indicate significance in one of the algorithms only, while those without are significant in both. MAF is minor allele frequency, while R-square estimates the phenotype variation caused by the marker. Effect size estimates the effect on phenotype when the minor allele is present in the accession.

LIST OF ABBREVIATIONS

AS	Alternative splicing
NMD	Nonsense-mediated decay
GBS	Genotyping by sequencing
SNP	Single nucleotide polymorphism
LD	Linkage disequilibrium
GO	Gene ontology
GWAS	Genome-wide association study
DF	Days to flowering
DM	Days to maturity
DPF	Days of pod formation
C17	Cumulative harvest at DPF x 1.7
CUMPOD	Cumulative pod count at the end of study
C17_TOT	Proportion of C17 to total harvest (C17 / CUMPOD)
PRODAY	Productive days

F_YLD	Final yield obtained by single harvest
F_YLD_TOT	Proportion of F_YLD to total (F_YLD/CUMPOD)
S_POD	Seed per pod
SDWT100	Seed weight of 100 seeds

GENERAL INTRODUCTION

Mungbean (*Vigna radiata*) is a pulse crop that is widely planted in Asia with annual plantation area of 6 million hectares worldwide (Nair et al., 2012). Mungbean can be consumed as vegetable sprouts, or cooked as an ingredient for soup, porridge, pancake, noodles, or sweet paste for cake fillings. Nutritionally, mungbean has high protein content and is richer in folate and iron compared to other legumes (Keatinge et al., 2011). Moreover, it can fix atmospheric nitrogen through symbiosis with nitrogen-fixing bacteria, making it an ideal intercropping crop that can improve soil fertility and texture. Despite these attractive qualities, genomic information in mungbean is lacking compared to other legumes like soybean (*Glycine max*) or chickpea (*Cicer arietinum*) (Kang et al., 2014).

Since the blueprint of a living organism is programmed in its genome, a better understanding of the genome will allow us to identify, select, and develop superior crops efficiently. The genome of mungbean had been sequenced recently (Kang et al., 2014), covering approximately 80% of the estimated genome size (579 megabases). The total number of protein-encoding genes is predicted to be around 22,427 and 50.1% of the sequenced genome was shown to contain repetitive sequences. The availability of a good quality reference genome will greatly assist genetic research in mungbean since the reference provides the physical location of

genes and their regulatory elements in high resolution. In other crops, the availability of physical map data and the increasing affordability of high coverage whole genome sequencing and high-throughput marker genotyping had greatly assisted forward genetics studies to pinpoint the exact location of genes and mutations that contribute to desirable phenotypes (Huang et al., 2010; Zhou et al., 2015).

Once the DNA sequence responsible for the desired trait has been identified, then the superior DNA variant can be transferred to elite plant varieties using cross-pollination to further enhance their productivity or quality. Currently, tracking the inheritance of DNA segment with known benefit is more precise and more efficient than traditional breeding (Collard and Mackill, 2008). Plants carrying the desired DNA could be selected in seedling form as there is no need to wait until the trait is expressed. Moreover, since the right environment is not necessary to select the trait, selected plants can be grown virtually anywhere to reduce cost associated with plant maintenance. Further cost saving can be obtained during maintenance stage since most unwanted plants have been eliminated very early in the selection stage, which means that only selected few will need to be maintained to obtain the next plant generation. Using DNA markers that track all the chromosomes in the genome, Neeraja et al. (2007) demonstrated that DNA transfer could be completed in 2-3 generation shorter than traditional breeding.

Nevertheless, a major bottleneck in the application of DNA-based breeding in mungbean exists in the form of our limited knowledge in identifying superior alleles and their molecular function in the cell. In essence, breeders still have to rely on existing natural DNA variation since our capability in modifying and optimizing DNA sequence using genetic engineering is still limited and even viewed negatively by the general public (Priest, 2000). Natural genetic resources with rich variation are therefore essential for successful identification of useful DNA variants. Phenotypic evaluation of germplasm collection is insufficient to probe the underlying genetic diversity, as many alleles produce similar observable phenotypes, and their effects will only be visible when they are combined in hybrid lines. However, creating the crosses and observing the progeny can entail significant amount of time, cost, and efforts, so it will be preferable if useful alleles can be directly identified based on DNA sequence and phenotype data of the original germplasm accessions.

In the past decade, there has been significant progress in the development of statistical and computational tools that can be used to parse big sequence data and analyze the presence of correlation between sequence variation and useful traits, often termed as genome-wide association study or GWAS (Hirschhorn and Daly, 2005). Various algorithms and software packages have been developed for GWAS with the ability to account for multiple gene effects and the presence of non-consequential

correlation that arise from similarities between genetically related individuals (Yu et al., 2006). Several studies had confirmed the accuracy of the prediction software (Andersen et al., 2005). Molecular study on genes carrying the DNA variation with strong correlation to the desired traits has also confirmed the efficacy of the algorithms. Once the underlying alleles for important traits have been identified, the molecular mechanisms of the trait formation can also be elucidated (Kang et al., 2016). As molecular data accumulates, there will come a time when the function of the majority of the genome is known. Combined with the development of efficient DNA editing technology like CRISPR (Cong et al., 2013), this can have a significant impact on plant breeding in the future. Future breeders can simply tweak and optimize the gene network instead of relying on natural variations, which could become more limited in the future due to the loss of natural habitat for wild accessions and climate change.

However, in order to achieve such feat, a high quality annotation of the functional regions in the genome is required. Annotation enables the identification of nucleotide mutation effect on the expression and amino acid composition of the coded protein. One of the factors that needs to be considered when annotating open reading frames is the presence of alternative splicing (AS), which is shown to be present in more than 50% of all expressed genes in many animal and plant species (Chamala et al., 2015). Mutations in genes that have AS can affect the resulting protein in

multiple ways, so this factor needs to be considered in gene function studies. Since AS seems to affect a disproportionately large number of genes and information on AS is limited in mungbean, annotating the AS status of protein coding gene is a priority in order to ensure that the effect mapped mutations on genes can be deduced correctly.

The studies in this manuscript attempted to elucidate the function of various components of the mungbean genome, which will be useful for implementing genomics-assisted breeding in mungbean. First, the annotation of coding regions in the mungbean genome was expanded by adding data on the prevalence of AS in mungbean and catalogue AS isoforms that could be functional in mungbean. Next, we assessed the degree of DNA sequence variation in wild and cultivated mungbean populations in order to evaluate their potential for GWAS, and find the location of genes that are involved in the domestication of mungbean. Domestication genes may reflect the traits that are preferred by farmers and consumers since ancient times, so their identification will allow us to further improve the traits in the future. Finally, loci that contribute to important agronomic traits such as yield components and synchronicity were identified using GWAS. This will provide a starting point for future studies to elucidate the molecular mechanisms of trait formation in mungbean, as well as identifying accessions with superior phenotypes and causal alleles that can be utilized in future mungbean breeding programs.

LITERATURAL REVIEWS

The impacts of next generation sequencing on genetic studies

Next generation sequencing (NGS) is probably the most disruptive technology in genetics and genomics in the previous decade. Using massively parallel sequencing by synthesis it could create unprecedented sequencing output at a much more affordable cost (Schuster, 2007). Development of the sequencing technology further increased the output and bring down the cost, that it slowly became a standard for DNA sequence analysis (Metzker, 2010). Early versions of the machines could only output very short sequence fragments, but the high sequencing capacity allowed higher depth sequencing of the target DNA, so by using bioinformatics analysis it is possible to reassemble the short reads into a longer sequence data with lower error rates (Zerbino and Birney, 2008). Combined with shotgun sequencing strategy and sophisticated sequence assembly algorithms, this had created a race in the assembly of the reference genomes of multiple species. Within 10 years of its introduction, the genome of ~99 species had been sequenced and assembled with the aid of NGS

technology (Kang et al., 2015). Big resequencing projects that sequenced thousands of accessions had also been performed in several major crops in order to catalogue existing sequence variations in crop germplasms (Li et al., 2014; Zhou et al., 2015).

Early reference plant genomes that were constructed exclusively using NGS data usually had significant gaps because of the presence of repetitive elements, which can hinder de novo assembly algorithms and cannot be solved simply by increasing sequencing depth (Butler et al., 2008). New sequencing technology that could output longer fragments and optical mapping has partly solved this problem and improve the quality of the genome assembly (Faino et al., 2015). Nevertheless, incomplete reference genomes still serve as useful templates for genetic studies, as most repetitive regions contains few genes and have little known functions.

Aside from DNA sequencing, NGS can be adapted for other types of analysis as long as the intended analysis can be reflected in DNA sequences used for library construction. Hence RNA can also be sequenced in NGS by first converting it to cDNA, and their expression level can be tracked by simply counting the number of fragments originated from the same genes in the sequencing data (Wang et al., 2009). Methylation status can also be quantified for the whole genome, by bisulfite treatment to convert unmethylated cytosine to uracil and then thymine prior to library

construction (Smith et al., 2009). DNA and RNA segments that interact with proteins can also be identified in high throughput fashion, by crosslinking proteins that bind to DNA, shear the DNA-protein complexes, and purify the protein-bound DNA for library construction. A similar procedure can also be done on RNA-binding proteins (Park, 2009). Through reduced-representation library construction, where only a small amount of the genome is sampled, NGS was also found to be an economical approach to genotype individuals in a population (Elshire et al., 2011). Thus, NGS had been successfully applied to study gene expression level, epigenetic regulation, DNA-protein interaction at whole genome level, and genetic diversity of numerous species.

RNA-seq based transcriptomics as a new tool for geneticists

The use of NGS in transcriptome studies (RNA-seq) significantly reduced the complexity of RNA studies, especially because NGS allows simultaneous assessment of sequence variations as well as expression level of a gene by quantifying the number of reads from transcribed regions. Compared to microarray based analysis, which is the previous gold standard for high throughput transcriptome studies, RNA-seq is not constrained by the number of transcripts spotted on the array. Previous

knowledge about the transcripts is not required and the data can potentially reveal expression of previously unknown genes as well as transcription of non-coding genomic regions (Mortazavi et al., 2008). Transcriptional aberrations such as chimeric transcripts can also be observed using RNA-seq (Maher et al., 2009). Since expression level data is also accompanied by sequence data, RNA-seq can also be used to track the expression level of different alleles in multiallelic genes (Skelly et al., 2011).

Downstream RNA processing, such as alternative splicing (AS), can also be evaluated in RNA-seq data (Chamala et al., 2015). AS variations are not normally reflected in the protein-coding sequences, thus they are easily overlooked as the causal mechanism of a phenotype in a candidate gene search without transcriptome data. AS was also found to regulate transcript quantity rather than creating variations in protein composition (Neu-Yilik et al., 2004), which is similar to post-transcriptional regulation by micro RNA. Several studies had also catalogued the presence of miRNA and other non-coding RNA using NGS (Tarazona et al., 2011), and the genes that regulate them was elucidated using eQTL approach. In such studies, the miRNA level is used as phenotype data and associated with SNP variation (Huan et al., 2015).

Accumulation of data on genomic loci that regulate gene expression will play important roles in candidate gene identification during forward genetics

studies. In a study that mapped the locus with important roles in heterosis in rice, it was found that a significant portion of the causal genes undergo modulation of gene expression in overdominant hybrids instead of producing proteins with altered amino acid composition (Huang et al., 2016). This demonstrates the value of good annotation regarding gene expression level regulation in reference genome.

Towards plant translational genomics

As genomic information is accumulating for many crop species, new techniques are constantly being developed to utilize those data for more practical applications such as plant breeding. Tools for selecting the best genotypes are constantly being refined that they are now faster and cheaper than ever. Aside from techniques like genotyping by sequencing (GBS), which was designed to genotype the whole genome (Elshire et al., 2011), methods for detecting and selecting smaller number of high impact markers in a sizeable breeding population have also been developed. Automatic instrument capable of genotyping 24,960 genotypes in 65 x 384 array configuration within 8 hours is now available (Kang et al., 2015), which will streamline marker-assisted selection in crop improvement. Gene editing technique that could avoid the regulatory obstruction faced by transgenic plants is also being developed in many crop species (Luo et al., 2016).

The critical part of genomics-assisted breeding largely lies in identifying the target genes and markers to select them, along with parental lines that carry the favorable alleles. In soybean, efforts to unify existing data generated by various research groups in a single database are underway. Using genomic database like phytozome (https://phytozome.jgi.doe.gov/pz/portal.html) or soykb (Joshi et al., 2012), it is possible to view genetic and physical maps of the genome, as well as additional data such as gene expression level, methylation status, QTL intervals, and functional annotations of most genes. Similar databases also exist for other crops such as Gramene (http://www.gramene.org/) and PGDBj (http://pgdbj.jp/).

Although more information is constantly being added into the databases, for many traits the available information is not sufficiently detailed for immediate application in breeding program. QTL data typically still spans a large stretch in the chromosomes, and selecting for such large intervals increase the risk of linkage drag. Researchers working on minor crops like mungbean will also need to build the genomic information from scratch before any practical applications can be obtained from genomics, although comparative genomics strategy can be applied for well-conserved traits. The studies in this manuscript illustrate the follow up steps after the construction of a reference genome, where additional annotation and functional characterization of the genomic components are progressively added so

that the genomic information in mungbean can have beneficial applications for mungbean improvement.

REFERENCES

Andersen, J.R., Schrag, T., Melchinger, A.E., Zein, I. and Lübberstedt, T. (2005) Validation of Dwarf8 polymorphisms associated with flowering time in elite European inbred lines of maize (Zea mays L.). *Theoretical and Applied Genetics* **111**, 206-217.

Butler, J., MacCallum, I., Kleber, M., Shlyakhter, I.A., Belmonte, M.K., Lander, E.S., Nusbaum, C. and Jaffe, D.B. (2008) ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome research* **18**, 810-820.

Chamala, S., Feng, G., Chavarro, C. and Barbazuk, W.B. (2015) Genome-wide identification of evolutionarily conserved alternative splicing events in flowering plants. *Frontiers in bioengineering and biotechnology* **3**, 33.

Collard, B.C. and Mackill, D.J. (2008) Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Philosophical Transactions of the Royal Society B: Biological Sciences* **363**, 557-572.

Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W. and Marraffini, L.A. (2013) Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819-823.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S. and Mitchell, S.E. (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS one* **6**, e19379.

Faino, L., Seidl, M.F., Datema, E., van den Berg, G.C., Janssen, A., Wittenberg, A.H. and Thomma, B.P. (2015) Single-molecule real-time sequencing combined with optical mapping yields completely finished fungal genome. *MBio* **6**, e00936-00915.

Hirschhorn, J.N. and Daly, M.J. (2005) Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics* **6**, 95-108.

Huan, T., Rong, J., Liu, C., Zhang, X., Tanriverdi, K., Joehanes, R., Chen, B.H., Murabito, J.M., Yao, C. and Courchesne, P. (2015) Genome-wide identification of microRNA expression quantitative trait loci. *Nature communications* **6**.

Huang, X., Wei, X., Sang, T., Zhao, Q., Feng, Q., Zhao, Y., Li, C., Zhu, C., Lu, T. and Zhang, Z. (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. *Nature genetics* **42**, 961-967.

Huang, X., Yang, S., Gong, J., Zhao, Q., Feng, Q., Zhan, Q., Zhao, Y., Li, W., Cheng, B. and Xia, J. (2016) Genomic architecture of heterosis for yield traits in rice. *Nature* **537**, 629-633.

Joshi, T., Patil, K., Fitzpatrick, M.R., Franklin, L.D., Yao, Q., Cook, J.R., Wang, Z., Libault, M., Brechenmacher, L. and Valliyodan, B. (2012) Soybean Knowledge Base (SoyKB): a web resource for soybean translational genomics. *BMC genomics* **13**, 1.

Kang, H., Wang, Y., Peng, S., Zhang, Y., Xiao, Y., Wang, D., Qu, S., Li, Z., Yan, S. and Wang, Z. (2016) Dissection of the genetic architecture of rice resistance to the blast fungus *Magnaporthe oryzae*. *Molecular plant pathology*.

Kang, Y.J., Kim, S.K., Kim, M.Y., Lestari, P., Kim, K.H., Ha, B.K., Jun, T.H., Hwang, W.J., Lee, T., Lee, J., Shim, S., Yoon, M.Y., Jang, Y.E., Han, K.S., Taeprayoon, P., Yoon, N., Somta, P., Tanya, P., Kim, K.S., Gwag, J.G., Moon, J.K., Lee, Y.H., Park, B.S., Bombarely, A., Doyle, J.J., Jackson, S.A., Schafleitner, R., Srinives, P., Varshney, R.K. and Lee, S.H. (2014) Genome sequence of mungbean and insights into evolution within *Vigna* species. *Nature communications* **5**, 5443.

Kang, Y.J., Lee, T., Lee, J., Shim, S., Jeong, H., Satyawan, D., Kim, M.Y. and Lee, S.H. (2015) Translational genomics for plant breeding with the genome sequence explosion. *Plant biotechnology journal*.

Keatinge, J., Easdown, W., Yang, R., Chadha, M. and Shanmugasundaram, S. (2011) Overcoming chronic malnutrition in a future warming world:

the key importance of mungbean and vegetable soybean. *Euphytica* **180**, 129-141.

Li, J.-Y., Wang, J. and Zeigler, R.S. (2014) The 3,000 rice genomes project: new opportunities and challenges for future rice research. *GigaScience* **3**, 1.

Luo, M., Gilbert, B. and Ayliffe, M. (2016) Applications of CRISPR/Cas9 technology for targeted mutagenesis, gene replacement and stacking of genes in higher plants. *Plant cell reports*, 1-12.

Maher, C.A., Palanisamy, N., Brenner, J.C., Cao, X., Kalyana-Sundaram, S., Luo, S., Khrebtukova, I., Barrette, T.R., Grasso, C. and Yu, J. (2009) Chimeric transcript discovery by paired-end transcriptome sequencing. *Proceedings of the National Academy of Sciences* **106**, 12353-12358.

Metzker, M.L. (2010) Sequencing technologies—the next generation. *Nature Reviews Genetics* **11**, 31-46.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. and Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods* **5**, 621-628.

Nair, R., Schafleitner, R., Kenyon, L., Srinivasan, R., Easdown, W., Ebert, A. and Hanson, P. (2012) Genetic improvement of mungbean. *SABRAO Journal of Breeding and Genetics* **44**, 177-190.

Neeraja, C., Maghirang-Rodriguez, R., Pamplona, A., Heuer, S., Collard, B., Septiningsih, E., Vergara, G., Sanchez, D., Xu, K. and Ismail, A. (2007) A marker-assisted backcross approach for developing submergence-tolerant rice cultivars. *Theoretical and Applied Genetics* **115**, 767-776.

Neu-Yilik, G., Gehring, N.H., Hentze, M.W. and Kulozik, A.E. (2004) Nonsense-mediated mRNA decay: from vacuum cleaner to Swiss army knife. *Genome biology* **5**, 218.

Park, P.J. (2009) ChIP–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics* **10**, 669-680.

Priest, S.H. (2000) US public opinion divided over biotechnology? *Nature biotechnology* **18**, 939-942.

Schuster, S.C. (2007) Next-generation sequencing transforms today's biology. *Nature* **200**, 16-18.

Skelly, D.A., Johansson, M., Madeoy, J., Wakefield, J. and Akey, J.M. (2011) A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. *Genome research* **21**, 1728-1737.

Smith, Z.D., Gu, H., Bock, C., Gnirke, A. and Meissner, A. (2009) High-throughput bisulfite sequencing in mammalian genomes. *Methods* **48**, 226-232.

Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A. and Conesa, A. (2011) Differential expression in RNA-seq: a matter of depth. *Genome research* **21**, 2213-2223.

Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10**, 57-63.

Yu, J., Pressoir, G., Briggs, W.H., Bi, I.V., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M. and Holland, J.B. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature genetics* **38**, 203-208.

Zerbino, D.R. and Birney, E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research* **18**, 821-829.

Zhou, Z., Jiang, Y., Wang, Z., Gou, Z., Lyu, J., Li, W., Yu, Y., Shu, L., Zhao, Y., Ma, Y., Fang, C., Shen, Y., Liu, T., Li, C., Li, Q., Wu, M., Wang, M., Wu, Y., Dong, Y., Wan, W., Wang, X., Ding, Z., Gao, Y., Xiang, H., Zhu, B., Lee, S.H., Wang, W. and Tian, Z. (2015) Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nature biotechnology* **33**, 408-414.

CHAPTER I

Genome-wide characterization of RNA splicing in mungbean

ABSTRACT

Alternative splicing (AS) can produce multiple mature mRNAs from the same primary transcript, thereby generating diverse proteins and phenotypes from the same gene. To assess the prevalence of AS in mungbean (*Vigna radiata)*, we analyzed whole-genome RNA sequencing data from root, leaf, flower, and pod tissues, and found that at least 37.9% of mungbean genes are subjected to AS. The number of AS transcripts exhibited a strong correlation with exon number, and thus resembled a uniform probabilistic event rather than a specific regulatory function. The proportion of frameshift splicing was close to the expected frequency of random splicing. However, alternative donor and acceptor AS events tended to occur at multiples of three nucleotides (i.e., the codon length) from the main splice site. Genes with high exon number and expression level, which should have the most AS if splicing is purely stochastic, exhibited less AS, implying the existence of negative selection against excessive random AS.

Functional AS is probably rare: a large proportion of AS isoforms exist at very low copy per cell on average, or are expressed at much lower levels than default transcripts. Conserved AS was only detected in 629 genes (2.8% of all genes in the genome) when compared to *Vigna angularis*, and in 16 genes in more distant species like soybean. These observations highlight the challenges of finding and cataloging candidates for experimentally proven AS isoforms in a crop genome.

Keywords: alternative splicing, mungbean (*Vigna radiata*), RNA sequencing, stochastic process, evolutionary conservation

INTRODUCTION

Alternative splicing (AS) is the differential splicing of introns from pre-mRNA to yield several distinct mature mRNAs (isoforms) from a single gene. In general, four fundamental types of differential splicing can alter the coding region: intron retention, exon skipping, alternative donor, and alternative acceptor (Breitbart et al., 1987). In intron retention, intron sequences that are normally spliced out are retained in the mature mRNA, producing a longer transcript with extra coding sequences. By contrast, in exon skipping, some exon segments are spliced out from the final transcript to yield shorter mRNA molecules. The location of the splicing reaction can also change at only one of the splice sites; this situation is referred to as ‘alternative donor’ if the change occurs at the 5’ end of the intron, and ‘alternative acceptor’ if the change occurs at the 3’ end of the intron.

The mature mRNAs produced by AS can harbor additional bases or lack some exon sequences, resulting in alteration of amino acid composition, physical characteristics, or chemical function of the encoded proteins. Thus, AS can increase the number of protein types and phenotypes produced by a small number of genes. Inclusion of additional sequences and missense splicing can also introduce premature stop codons into the transcripts, making them vulnerable to degradation by the nonsense-mediated decay

(NMD) pathway (Neu-Yilik et al., 2004) and decreasing the quantity of those particular transcripts in the cell. Several lines of evidence show that cells actually utilize this pathway to modulate and fine-tune the number of RNA molecules for a particular gene under certain conditions (Filichkin and Mockler, 2012; Kawashima et al., 2014).

Consequently, AS could explain the complexity paradox, i.e., the observation that the genomes of certain complex organisms harbor a smaller number of coding regions than those of some simpler organisms (Graveley, 2001). Several experimentally proven AS isoforms produce multiple proteins with distinct characteristics and function from the same coding region (Inoue et al., 1990; Lah et al., 2014; Ullrich et al., 1995), potentially explaining how a single gene could perform multiple functions in the cell. The advent of next-generation sequencing (NGS), which can generate large quantities of transcriptome data faster and more cheaply than previous methods, has aided in the identification of AS in many different organisms. Software and script packages such as ASTALAVISTA (Foissac and Sammeth, 2007) and ASprofile (Florea et al., 2013), have been developed to rapidly identify splicing variants by examining variations in exon-intron boundaries in genome-wide alignment data generated using NGS. The results are quite surprising: in some cases, AS occurred in more than half of the annotated genes (Marquez et al., 2012; Pan et al., 2008; Shen et al., 2014). If all AS produces functionally divergent proteins, then

this process regulates the bulk of transcript generation and protein synthesis in the cell. Hence, proper annotation of the occurrence of AS in the genome is very important as a reference for functional genomic studies.

Several studies have attempted to catalogue the global occurrence of AS in the genomes of several plants, including soybean (Shen et al., 2014), Arabidopsis (Filichkin et al., 2010; Marquez et al., 2012), and maize (Thatcher et al., 2014), by utilizing mRNA sequences obtained from different tissue types under diverse environmental conditions to capture as many transcript types as possible. Nevertheless, although those studies revealed that a large number of plant genes undergo AS, very little experimental evidence of functional AS proteins is available (Severing et al., 2009). Several groups have suggested that the scarcity of demonstrably functional AS isoforms could be due to the random nature of AS itself (Hon et al., 2013; Melamud and Moult, 2009a; Zhang et al., 2009), implying that most AS isoforms have no function because they are merely the byproducts of erroneous splicing. Consequently, it is unlikely that all AS events are part of a distinct layer of gene regulation. That said, because AS isoforms that confer selective advantage could be retained by progeny with stronger AS signals for those isoforms, functional AS could still evolve and be retained by natural selection.

Because advantageous AS isoforms have a higher probability of being retained over the course of evolution, it should be possible to identify them in comparative studies of related species. Mungbean (*Vigna radiata*) and its close relatives in the *Vigna* genus, like adzuki bean (*Vigna angularis*), are good candidates for such studies. Their genome sequences have recently been published (Kang et al., 2014; Kang et al., 2015), enabling transcript alignment and facilitating identification of AS isoforms. They are also related to soybean, whose genome is already well characterized, and for which comprehensive data regarding AS is available. Because mungbean and adzuki bean are widely planted for food consumption (annual plantation area of 6 million and 840,000 hectares, respectively), any practical applications that could be derived from genomic studies in these plants will have considerable economic impact (Nair et al., 2012; Rubatzky and Yamaguchi, 1997).

We performed global transcriptome analysis to identify and catalogue AS events that occur in mungbean. To infer the characteristics of AS regulation in this species, we tested for stochastic AS in the RNA population. To identify AS events with the strongest likelihood of being functional, for the purpose of subsequent in-depth studies, we investigated AS conservation in adzuki bean and soybean. The resultant whole-genome annotation of AS isoforms represents a valuable contribution to the annotation of mungbean genome.

MATERIALS AND METHODS

Plant materials and RNA sequencing

RNA sequence data was obtained from Kang et al. (2014); in that study, a pure line mungbean plant from cultivar VC1973A (developed by AVRDC) was used as the source material for RNA extraction. The plants were planted in a greenhouse; following sowing, tissues were harvested from root after 2 weeks, leaf after 1 month, flower after 2 months, and whole pods after 2.5 months.

Sequence alignment, transcript assembly, and AS identification

The cleaned sequence data were aligned to the mungbean reference genome (Kang et al., 2014) using TopHat (Trapnell et al., 2009) with default settings. The resultant gapped alignment data in binary alignment format were then used as input for Cufflinks (Trapnell et al., 2012) under default settings to assemble the transcripts and identify splicing junctions from the alignment data. For AS detection and annotation, the assembled transcriptome files (in .gtf format) were submitted to ASTALAVISTA (Foissac and Sammeth, 2007) web interface (http://genome.crg.es/astalavista/). AS events were also annotated with

ASprofile (Florea et al., 2013), which also uses Cufflinks output as input data. The resulting AS annotations were checked at random by visual examination of the AS genome coordinates in the original binary alignment (.bam) files using Integrated Genome Viewer (Robinson et al., 2011).

Isoform quantitation

The FPKM (fragments per kilobase of transcript per million mapped reads) value for each AS isoform was provided by ASprofile, based on Cufflinks estimation, after assembly of each transcript. When the FPKM value was not available for a chromosomal segment of interest, transcript quantity was estimated based on alignment coverage on that segment, calculated using the coverageBed command in Bedtools (Quinlan and Hall, 2010). The number of aligned fragments were then multiplied by fragment length, and divided by the number of bases in the segment of interest, to yield the average coverage per base value. Because each tissue had different sequence coverage, coverage per base was only compared within a tissue.

Statistical analysis

Basic arithmetical analysis, such as calculations of sums and means, was performed in Microsoft Excel. Calculation of descriptive statistics was performed in the R statistical package, using the “describeBy” command in the psych library. Calculations of Pearson’s correlations and the corresponding significance values were also carried out in R using the “cor.test” command.

Sequence junction analysis

To visualize the presence of sequence conservation near splice sites, coordinates of splice sites along with 10 bases upstream and 10 bases downstream from those sites were input into Bedtools using the fastaFromBed command to obtain the DNA sequences between those coordinates. Sequences with FPKM value >10 were put in the high group, while the rest were put in the low group. The sequences were then used as input for weblogo (http://weblogo.threeplusone.com/) to visualize the proportion of bases commonly found surrounding the splice sites.

Comparative analysis

The sequence of 50 bp of exonic region surrounding AS sites and intron sequences from intron-retention AS events were obtained using Bedtools from mungbean, adzuki bean (*Vigna angularis*), and soybean (*Glycine max*). Splice sites were detected in adzuki bean and soybean using ASprofile with the same settings as used for AS detection in mungbean. The RNA sequences used for AS detection were obtained from sequences provided by Kang et al (2015) for adzuki bean and Shen et al (2014) for soybean. Sequences that share similarities were detected using local BLAST+ search (Camacho et al., 2009), with mungbean sequences used as the local sequence database. The resulting matches were filtered using the following criteria: sequences at the splicing junctions must have exact match while sequences further away are allowed to have gaps and mismatches, intron sequences in intron retention type have at least 80% similarities, and the length difference of retained introns and skipped exons between two species must not exceed 30 nucleotides or introduce frameshift. Full sequences of proteins containing conserved AS were then identified, and their homologs in soybean were identified using blastp in the BLAST+ package. Matching soybean gene ID with the highest e-values were then submitted to agriGO (http://bioinfo.cau.edu.cn/agriGO/analysis.php) to obtain the gene ontology classification of those genes.

RESULTS

Characteristics of AS types in mungbean

The number of AS events (hereafter, AS number) of each type were detected in silico based on alignment of RNAseq data to the mungbean reference genome. Shotgun sequencing generated, on average, 38.6 million 100 bp reads per sample (Table I-1), close to 10 times the size of the mungbean genome. The total length of annotated transcribed regions is 104 million bases; therefore, the sequence alignment produced roughly 37× sequencing coverage for all open reading frames. However, because most of the sequenced RNAs are derived from mature RNAs whose introns have been spliced out, the sequencing depth in exonic regions was 101× on average.

Table I-1. Sequencing data and mapping rates in each tissue, as summarized by samtools flagstat

	Flower	Leaf	Pod	Root
QC-passed reads	37290562	37056646	39815968	40400677
Duplicates	0	0	0	0
Mapped reads	37290562	37056646	39815968	40400677
Paired in sequencing	37290562	37056646	39815968	40400677
Read 1	19191255	19258791	20579470	20869091
Read 2	18099307	17797855	19236498	19531586
Properly paired	25385936	25073574	25308664	24801220
With itself and mate mapped	32163296	31513876	35505358	35710818
Singleton	5127266	5542770	4310610	4689859
With mate mapped to different chromosome	220204	225244	359740	500378
With mate mapped to different chromosome (mapQ>=5)	117032	126848	191558	294748

Table I-2. Number of AS isoform types in each tissue, based on isoform detection with ASTALAVISTA and ASprofile.

Detection Method	Tissue	Intron Retention	Exon Skipping	Alternative Donor	Alternative Acceptor	Affected Genes
ASTALAVISTA	Flower	3620	659	1141	2101	4414
	Leaf	3512	647	1107	2060	4256
	Pod	4076	589	1093	2132	4541
	Root	5419	772	1320	2401	5429
ASprofile	Flower	4256	6494	3373	3059	8051
	Leaf	4173	6185	3278	2897	8152
	Pod	4582	6394	3455	3103	7931
	Root	5874	6724	3885	3336	7549

The number of AS events detected varied with the software pipeline: ASprofile annotated more AS events than ASTALAVISTA (Table I-2). A closer inspection of the binary alignment (BAM) files using a genome browser revealed that the higher number of AS events detected by ASprofile was due to reporting of new exons not found in the mungbean genome annotation, as well as increased sensitivity in detecting rare splicing junctions. Depending on the AS types, ASTALAVISTA did not detect 85.3 to 90.1 percent of AS events detected by ASprofile. However, ASprofile also failed to detect 56.5 to 85 percent of AS events detected by ASTALAVISTA. Neither pipeline is clearly superior to the other as they both missed AS events detected by the other pipeline, but ASprofile output was used for further analysis due to its increased sensitivity and better annotation system.

ASprofile estimated that 44.6% of mungbean genes are subjected to AS, whereas ASTALAVISTA estimated this proportion as 37.9%. Both figures are lower than the proportions reported for Arabidopsis (Marquez et al., 2012), maize (Thatcher et al., 2014), rice (Lu et al., 2010), and soybean (Shen et al., 2014), but fairly similar to those of closely related legumes like Medicago and Phaseolus (Chamala et al., 2015).

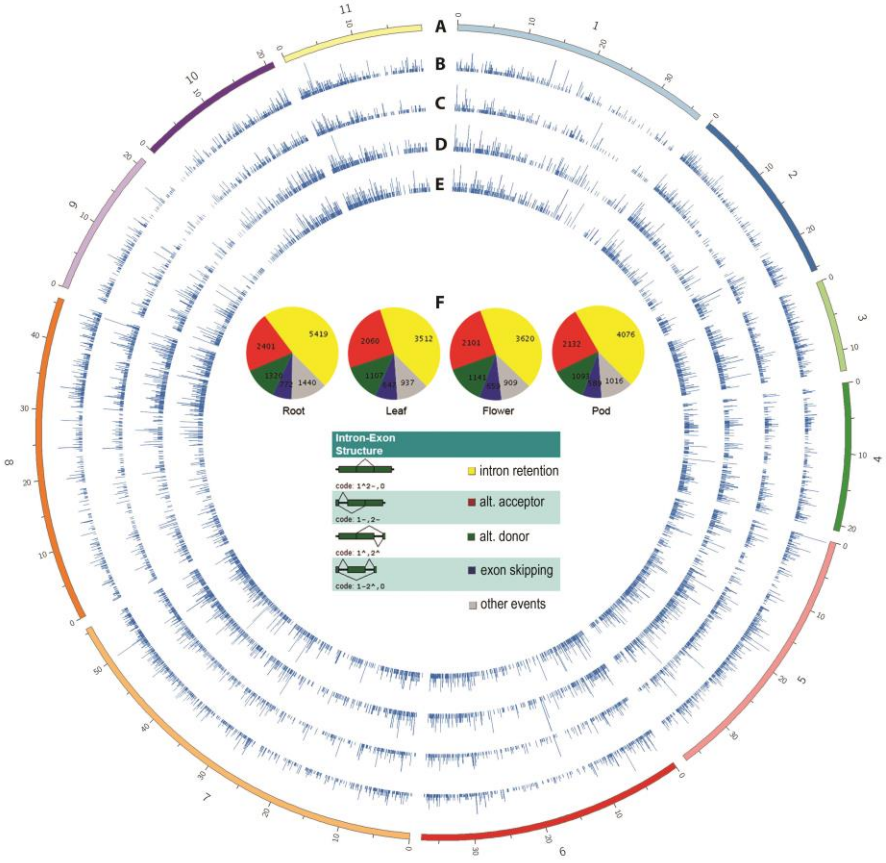


Figure I-1. Types and chromosomal distribution of AS in four mungbean tissues. From outer ring to inner rings: **(A)** size of chromosome (in megabases); **(B)** histogram of AS number across chromosomes in root, **(C)** leaf, **(D)** flower, and **(E)** pod tissues. **(F)** Proportions of each type of AS across the four tissues, as classified by ASTALAVISTA.

The distribution of AS was generally similar across tissues (Figure I-1), although tissue-specific AS isoforms were detected, and various tissues yielded different numbers of AS isoforms. Among all tissues, roots had the highest number of AS events, as well as the highest AS number per gene (Table I-2). One potential reason for this is that roots express the largest number of tissue-specific genes, and these genes tend to be highly expressed (Table I-3), potentially aiding in detection of AS isoforms in RNA from roots. About 2.3% of tissue-specific AS events were the consequence of tissue-specific gene expression, and their absence in other tissues is caused by the lack of expression of those genes; however, the remaining AS isoforms are tissue-specific even though the originating transcripts are expressed at significant levels in more than one tissue type. In many cases, we found that the absence of AS isoforms in other tissues was not merely caused by low expression levels and under-representation in RNA samples from these tissues.

Table I-3. AS and expression levels of genes expressed in only one tissue type. Expression level is calculated using transcript per million (TPM) method.

	Flower	Leaf	Pod	Root
Number of expressed genes	18421	16561	17626	18199
Number of tissue-specific genes	509	117	132	703
Cumulated expression (TPM) of tissue-specific genes	21008	1753	499	22816
Average AS per gene	1.43	1.46	1.45	1.54

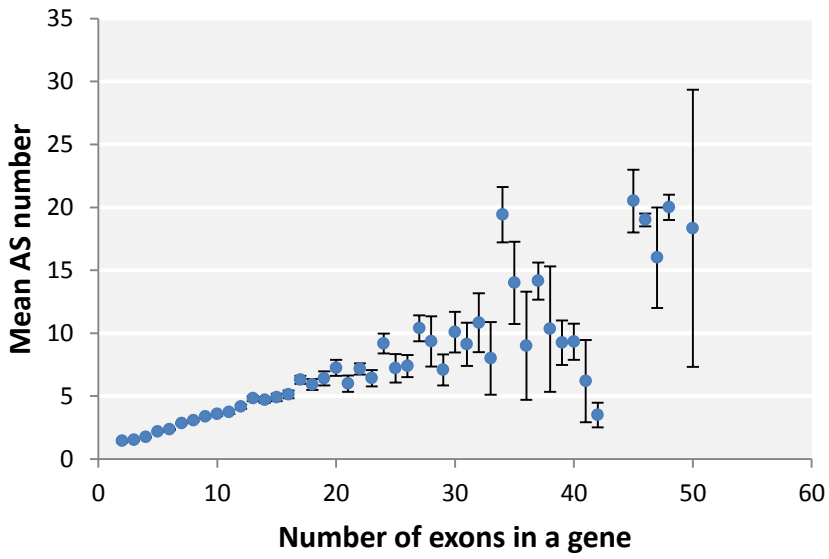
The number of transcripts representing a particular isoform is difficult to quantify accurately without long-read sequence data, because some genes have multiple AS events and some isoforms may combine with others from the same gene to generate a distinct transcript structure. Because long-read RNAseq data were not available for this study, we simply assumed that such combinations were non-existent, and then estimated the quantity of each AS isoform based on the number of sequence fragments that aligned to the splice junction that underwent AS. Based on this assumption, a significant portion of detected AS isoforms (20.54%) had FPKM values lower than 1, i.e., their concentration is very low in an average cell. Moreover, a considerable proportion of AS isoforms (24.4%) were expressed at levels < 10% of those of the more abundant constitutive splice forms.

Mungbean AS exhibits signs of stochastic splicing

The prevalence of AS isoforms with low concentration in our mungbean AS data raises the possibility that a significant number of mungbean AS could be the result of random errors with little effect on the protein composition of the cell. To determine whether stochastic splicing is prevalent in mungbean, we investigated the correlation between the presence of AS and several aspects of the plant's genomic features that

may increase the probability of random splicing errors. We found that mean AS number was strongly correlated with the number of exons in a gene with a Pearson r-value of 0.879 and p-value of 4.09e-14 (Figure I-2A). This is consistent with the random splicing error model: the higher the exon number, the larger the number of splicing junctions, and the greater the chance of error associated with splicing of those junctions. However, an obvious consequence of probabilistic splicing error is that genes with a large number of introns will have a higher probability of accumulating useless splicing errors, which could be dangerous to the cell. To determine whether mungbean has evolved a mechanism to reduce the likelihood of such errors, we plotted the average number of AS per exon for genes with different exon numbers. The plot reveals a clear trend toward fewer AS events per exon as the number of exons increases (r=-0.649, p=7.755e-05), although the pattern is less clear for genes containing more than 25 exons (Figure I-3A).

A.



B.

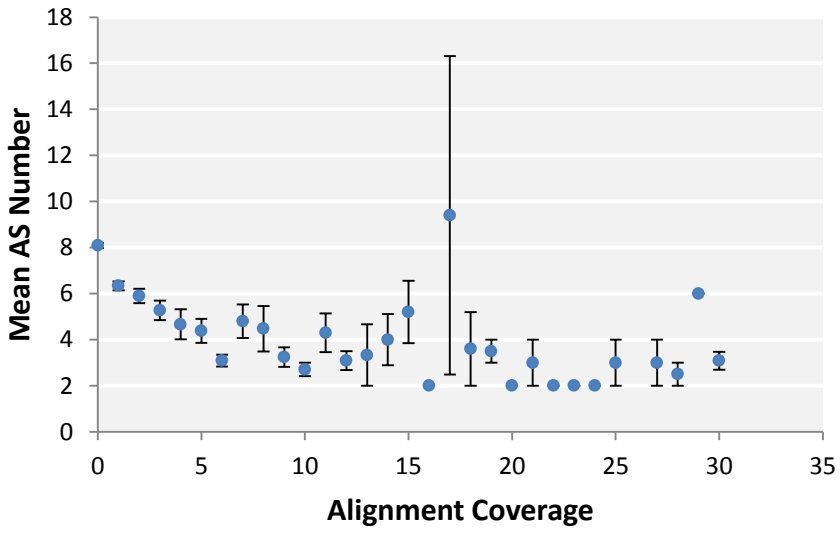


Figure I-2. Correlation of mean AS number with number of exons in a gene **(A)** and gene expression level estimated from alignment coverage **(B).**

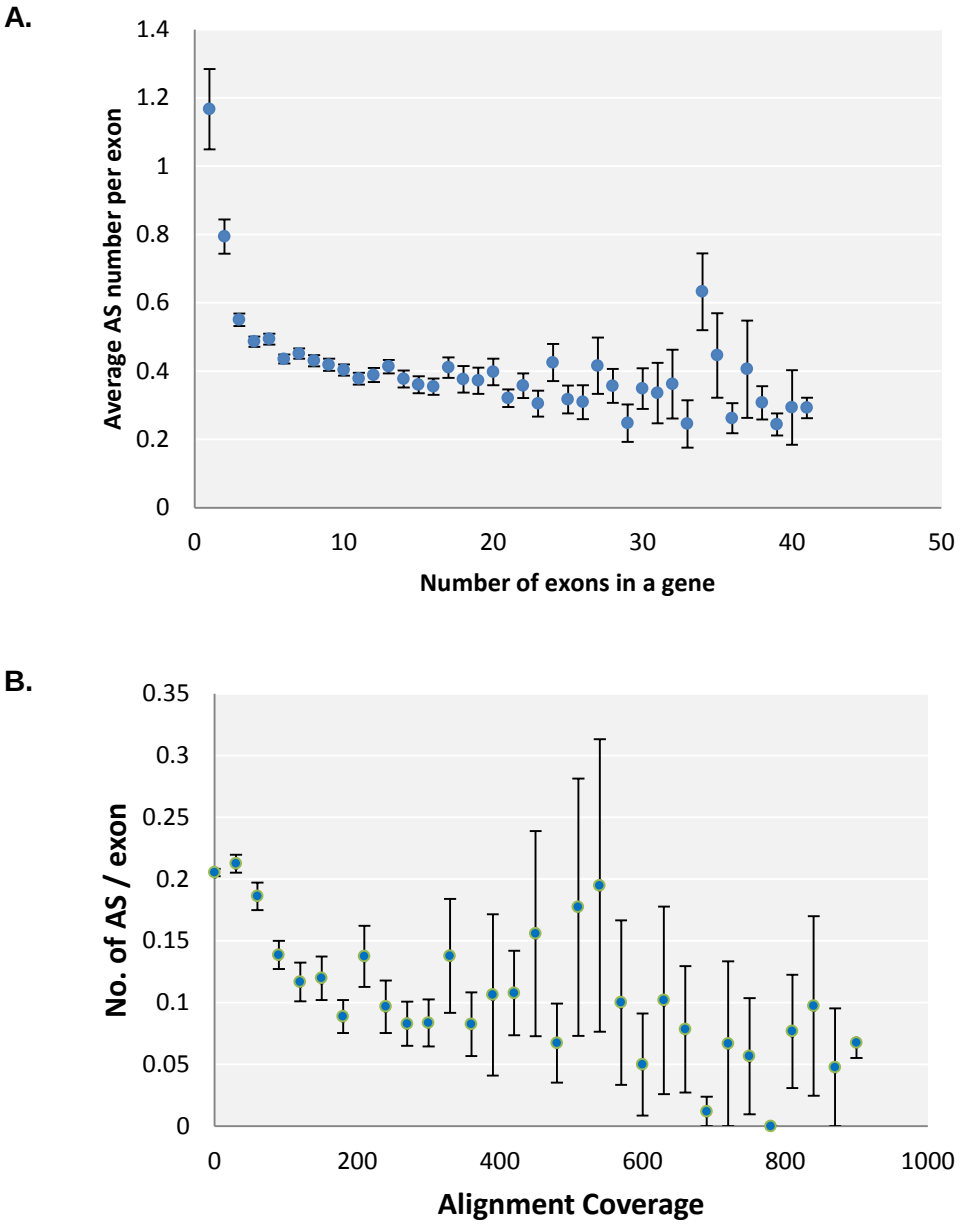


Figure I-3. Comparison of the average number of AS events per exon, calculated by dividing the number of AS events in a gene with the number of exons in that gene, for genes containing different numbers of exons **(A)** and genes expressed at different levels **(B)**, as determined by alignment coverage per base.

Erroneous splicing is also more disadvantageous for highly expressed genes, because in such cases, it would create a large amount of mis-spliced mRNA, which in turn is more likely to be translated into a large quantity of non-functional protein. Consistent with this, we observed a trend toward fewer AS events per exon in highly expressed genes (Figure I-3B), although the correlation was weak (r=-0.057 and p-value=2.2e-16). A correlation plot of AS number vs expression level revealed a negative correlation (r=-0.463 and p-value=0.001) between the two variables (Figure I-2B). This observation contrasts with findings in other plants such as soybean, in which highly expressed genes also usually have higher numbers of AS events (Shen et al., 2014). One reason for this could be that, in mungbean, the average number of exons is lower among highly expressed genes (Figure I-4), and exon number correlates more strongly to AS number than expression level. Moreover, the average number of AS was not higher in highly expressed genes than in genes carrying the same exon numbers expressed at lower levels (Figure I-5).

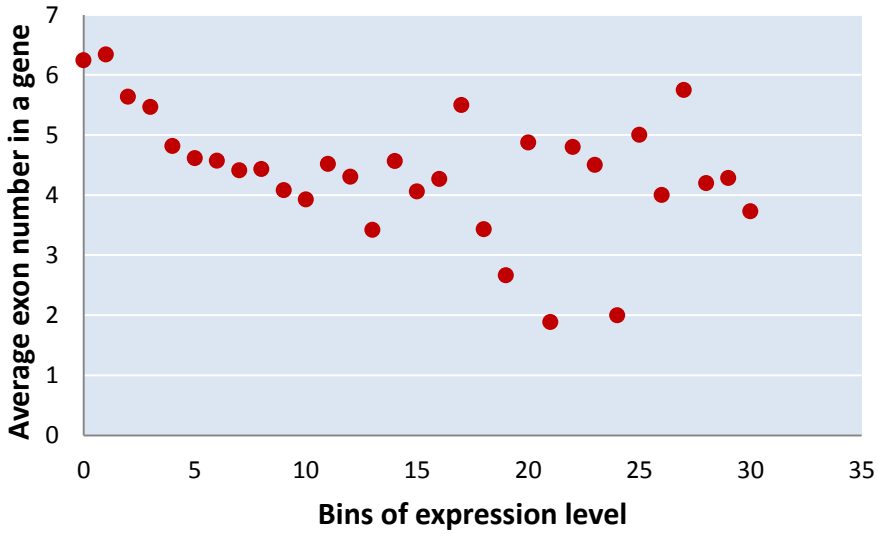


Figure I-4. Average number of exons in a gene, with genes grouped by expression level.

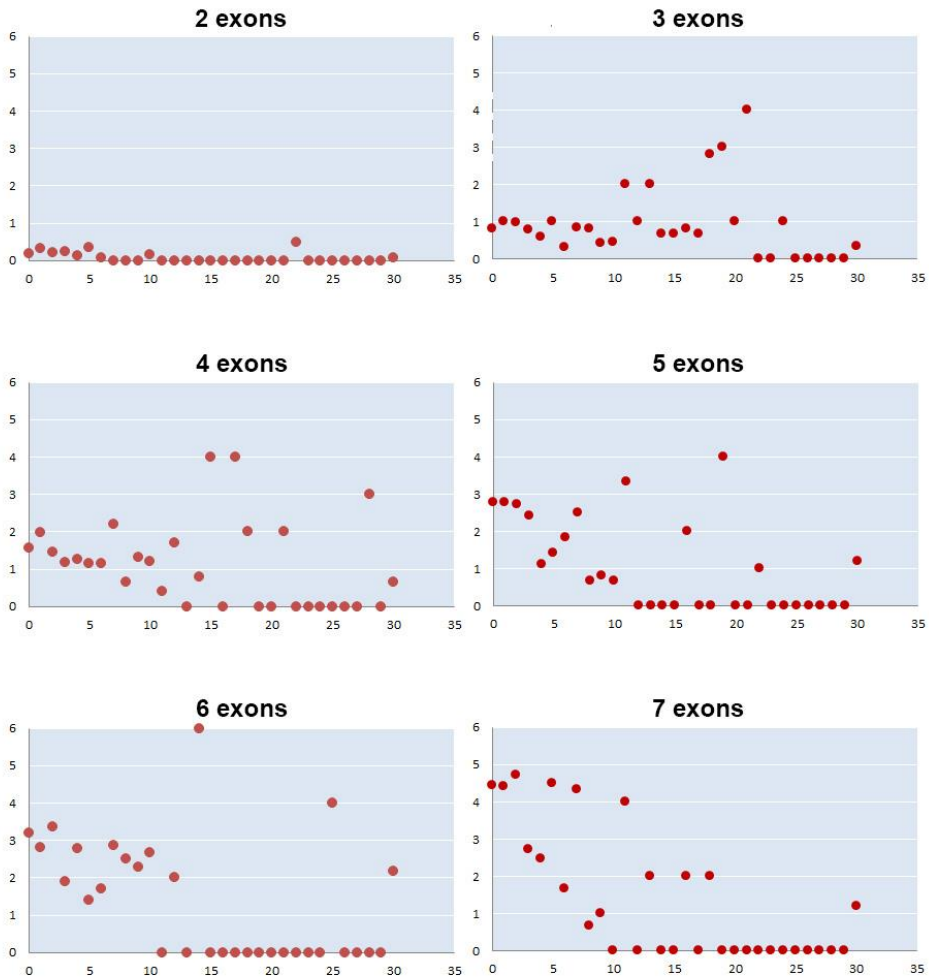


Figure I-5. Average number of AS events (y axis) in genes grouped by expression levels (x axis), depending on the number of exons found within the genes.

Another important effect of AS on the final transcript sequence is the creation of frameshift mutations, which could significantly alter the amino acid composition downstream of the splice site. Splicing error can introduce or eliminate n+1, n+2, and n+3 nucleotides to the mature mRNA, where n is a multiple of three nucleotides and only the n+3 variant will preserve the downstream codons. Assuming that all three variants are equally likely to occur, random splicing error should introduce frameshift 67% of the time. Frameshift mutations have the highest probability of rendering the resulting protein non-functional; therefore, we were curious to see whether this phenomenon would be repressed in mungbean AS. Frameshift formation is close to 67% for AS events of the exon skipping and intron retention types (Table I-4). However, the creation of frameshift was lower than expected in the alternative donor and alternative acceptor types of AS: plotting the number of AS events that occurred several bases from regular splicing sites revealed a preference for multiples of 3 (i.e., the length of a codon) in these types of AS (Figure I-6). This could be partially explained by the common occurrence of NAGNAG motifs near the 3' end of introns (Shi et al., 2014). Because plant splice sites are normally located at AG bases at the 3' ends of introns, such motifs would direct spliceosomes that miss their target to alternative targets located at distances that are integral multiples of codon lengths, thereby preventing the formation of frameshifted mRNA. Curiously,

this effect was still visible at positions as far as 90 bases from the regular splice sites, distances at which NAGNAG motifs are unlikely to persist.

Table I-4. Proportion of AS isoforms carrying frameshift codons in the mature mRNA, classified according to AS types.

Type of AS	Root	Leaf	Pod	Flower
Alternative Donor	34.3%	33.9%	32.6%	29.9%
Alternative Acceptor	46.3%	42.1%	42.3%	41.7%
Exon Skipping	72.0%	72.5%	71.2%	72.2%
Intron Retention	61.6%	63.2%	62.9%	62.6%

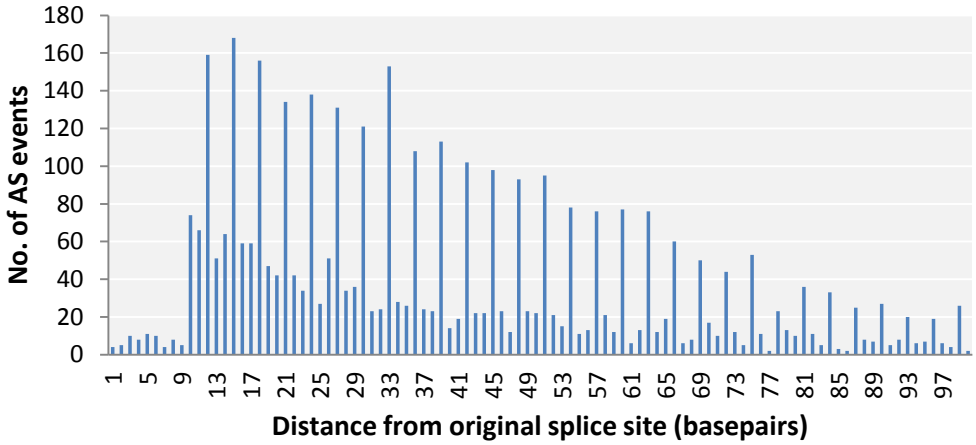


Figure I-6. Number of AS isoforms, according to distance from the regular splice site, for alternative donor and alternative acceptor isoform types.

Because frameshift mutation is very common in the more abundant intron retention and exon skipping types, it is unlikely that this is mostly caused by natural selection against frameshift splicing; therefore, other mechanisms could be responsible for these effects.

The role of sequence variation and the extent of AS conservation

The presence of motifs like NAGNAG at the 3' ends of introns raises the question of whether AS sites occur only at canonical splice sites or utilize other bases as well. We surveyed all splice junctions of the default and AS isoforms and categorized each AS site as high or low, using FPKM value of 10 as a threshold. At the 3' ends of introns, all isoforms utilize the AG splice site; by contrast, at the 5' end, a majority of splice sites occur at GU bases but a small fraction also occur at GC bases (Figure I-7). There were no obvious sequence patterns around the two bases that could explain why some AS sites are more frequently spliced than others. Hence, although the pattern of AS appears random, it is still constrained by the availability of bases that can be used as splice sites. However, because the required motifs at each end are only two bases long, and the abundance of these dinucleotides in the genome is relatively high, it is not surprising that AS is so prevalent in many organisms.

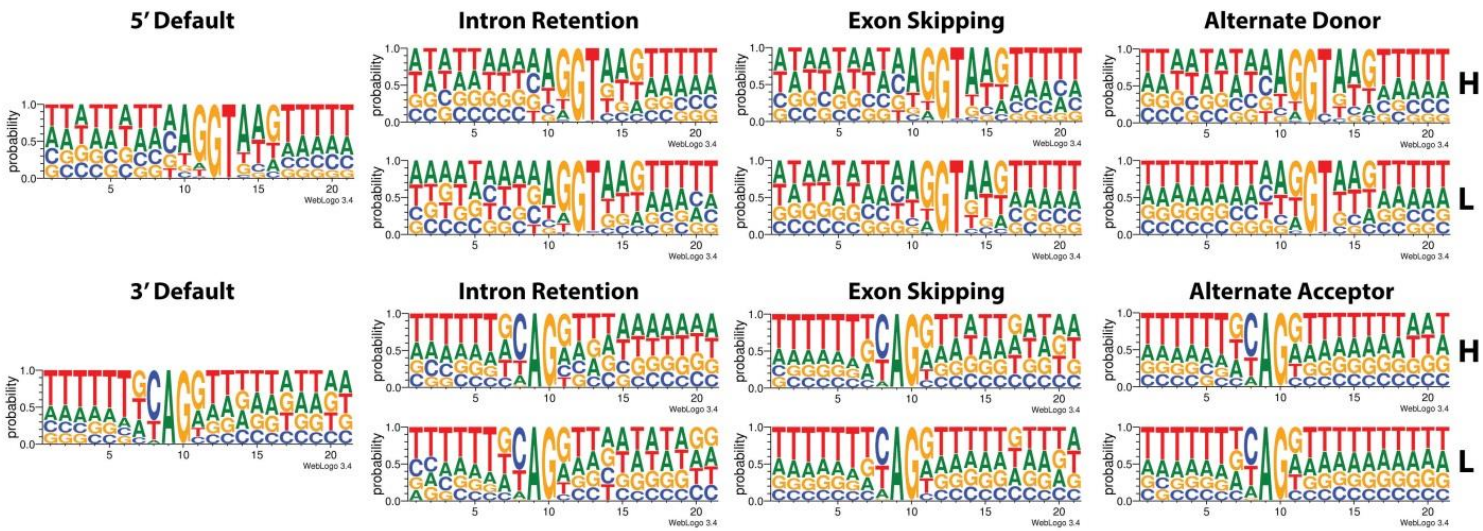


Figure I-7. Proportion of DNA bases surrounding alternative splice types compared to the regular splice sites. AS isoforms with FPKM >10 were categorized as having high concentration (**H**), while those with FPKM<10 were grouped into the low concentration group (**L**).

Nevertheless, functional AS isoforms have been detected in the past (Inoue et al., 1990; Ullrich et al., 1995); hence, it is not unlikely that mungbean also harbors some functional isoforms in its transcriptome. We tried to identify candidate AS isoforms for more in-depth study of their function. However, given that a significant portion of AS in mungbean may not have any function at all, we paid extra attention to AS sites that are conserved in other species. Conservation among species does not necessarily imply function, but it at least indicates that the isoforms in question do not impose negative selection pressure on the plant over evolutionary timescales. To this end, we compared transcript sequences from mungbean to those from adzuki bean (*Vigna angularis*), a closely related species in the *Vigna* genus. BLAST analysis of exon sequences surrounding AS junctions identified 3600 AS sites with high sequence similarities in both species, which is comparable to the results obtained by Chamala et al. (2015), who identified more than 5000 conserved AS between common bean and soybean using a similar method. However, a closer examination also revealed that the exact splice sites are rarely conserved in both species. By applying the strict criteria that both splice sites must be located at the exact same position and the differences in nucleotide length between the two species must not introduce frameshift, we reduced the set of candidates to 629 genes, comprising 859 conserved AS events.

Gene Ontology (GO) analysis of the genes carrying conserved AS identified 488 GO groups with significant enrichment for cellular components only (Figure I-8A), while other GO groups are not significantly different from background level (Figure I-8B). However, the number of conserved AS isoforms dwindles even further when the comparison is made between more distantly related species. A comparison of AS events between mungbean and soybean (*Glycine max*) yielded only 16 conserved AS isoforms retained at the exact same base position in both species (Table I-5). All but one AS junction was also conserved in adzuki bean, although two of them will create frameshift mutations in adzuki bean. Based on this observation, we conclude that AS events that confer selective advantages, and are thus retained over evolutionary timescales, are very rare. However, other groups have observed that when the conservation criteria are relaxed to ignore the exact splicing position and focus on exon sequence conservation among the same AS types, the number of conserved AS events increases considerably, and such events can be identified even in species outside the angiosperms (Chamala et al., 2015). While it is possible that such approach could identify conserved AS with similar function, it would be inadequate to identify possible inclusion of frameshift caused by differences in nucleotide length among species, which can create a very different protein if the isoforms are translated.

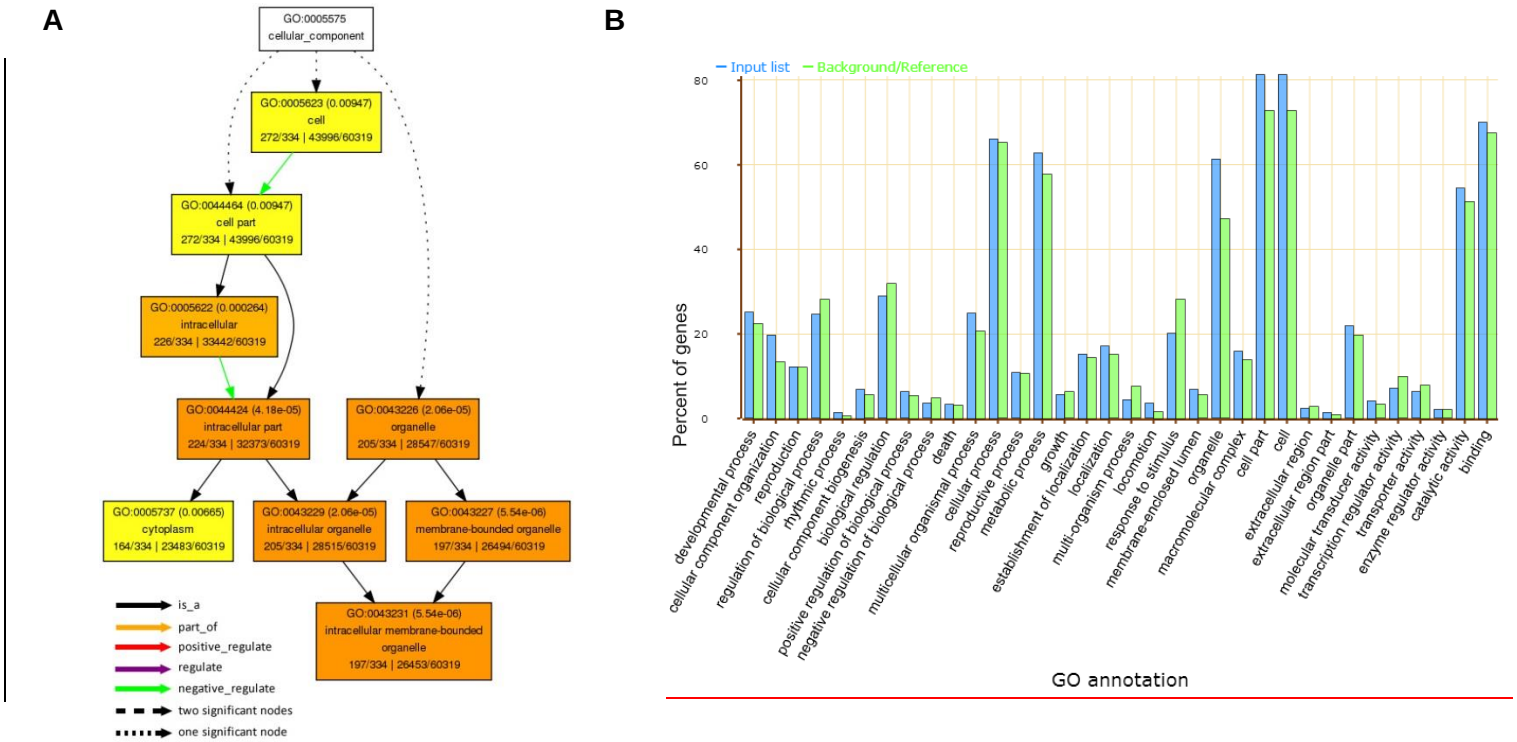


Figure I-8. **(A)** AgriGO annotation of the relationships and significance level of enriched GO groups of the genes with conserved AS in mungbean and adzuki bean. Inside the boxes: numbers inside the brackets are the p-value, while the numbers on lower left sides are annotated/total number in query and on the lower right are annotated/total number in background/reference. **(B)** Gene ontology enrichment of genes with AS isoforms that are conserved in mungbean and adzuki bean.

Table I-5. Types and locations of conserved AS in mungbean and soybean, identified using blast alignments of exons and introns surrounding AS sites

Splicing type	query chromosom	query base start	query base end	query type	subject chromosome	subject base start	subject base end	subject type	protein
Alternative Acceptor	Chr09	50001778	50002006	exon 3' end	scaffold-134	776745	776973	exon 3' end	myosin-2 heavy chain, non muscle, putative, expressed
Alternative Acceptor	Chr09	50001778	50002006	exon 5' end	scaffold-134	776745	776973	exon 5' end	myosin-2 heavy chain, non muscle, putative, expressed
Alternative Acceptor	Chr09	50001800	50002006	exon 5' end	scaffold-134	776767	776973	exon 5' end	myosin-2 heavy chain, non muscle, putative, expressed
Exon Skipping	Chr03	1736837	1737085	exon 3' end	scaffold-292	169049	169294	exon 3' end	RNA recognition motif containing protein, expressed
Exon Skipping	Chr03	1736837	1737085	exon 5' end	scaffold-292	169049	169294	exon 5' end	RNA recognition motif containing protein, expressed
Intron Retention	Chr07	19260385	19260472	intron	scaffold-304	103541	103628	intron	unknown
Intron Retention	Chr07	19260385	19260472	exon 3' end	scaffold-304	103541	103628	exon 3' end	unknown
Intron Retention	Chr07	19260385	19260472	exon 5' end	scaffold-304	103541	103613	exon 5' end	unknown
Alternative Acceptor	Chr16	3935708	3935828	exon 3' end	scaffold-7	2990119	2990239	exon 3' end	HNH endonuclease domain-containing protein, putative, expressed
Alternative Acceptor	Chr16	3935708	3935828	exon 5' end	scaffold-7	2990119	2990239	exon 5' end	HNH endonuclease domain-containing protein, putative, expressed
Alternative Acceptor	Chr16	3935774	3935828	exon 5' end	scaffold-7	2990185	2990239	exon 5' end	HNH endonuclease domain-containing protein, putative, expressed
Alternative Donor	Chr12	1610622	1610731	exon 3' end	Vr02	1835909	1836018	exon 3' end	zinc finger family protein, putative, expressed
Alternative Donor	Chr12	1610622	1610731	exon 5' end	Vr02	1835909	1836018	exon 5' end	zinc finger family protein, putative, expressed
Alternative Donor	Chr12	1610651	1610731	exon 5' end	Vr02	1835938	1836018	exon 5' end	zinc finger family protein, putative, expressed

Intron Retention	Chr18	1462230	1462355	intron	Vr02	23089041	23089166	intron	SAP domain containing protein, expressed
Intron Retention	Chr18	1462230	1462355	exon 3' end	Vr02	23089041	23089166	exon 3' end	SAP domain containing protein, expressed
Intron Retention	Chr18	1462230	1462355	exon 5' end	Vr02	23089041	23089166	exon 5' end	SAP domain containing protein, expressed
Alternative Acceptor	Chr03	43026078	43026296	exon 3' end	Vr03	8702981	8703199	exon 3' end	G-patch domain containing protein, expressed
Alternative Acceptor	Chr03	43026078	43026296	exon 5' end	Vr03	8702981	8703199	exon 5' end	G-patch domain containing protein, expressed
Alternative Acceptor	Chr03	43026153	43026296	exon 5' end	Vr03	8703056	8703199	exon 5' end	G-patch domain containing protein, expressed
Intron Retention	Chr13	42595381	42595462	intron	Vr04	5370098	5370179	intron	CCR4-NOT transcription factor, putative, expressed
Intron Retention	Chr13	42595381	42595462	exon 3' end	Vr04	5370098	5370179	exon 3' end	CCR4-NOT transcription factor, putative, expressed
Intron Retention	Chr13	42595381	42595462	exon 5' end	Vr04	5370098	5370179	exon 5' end	CCR4-NOT transcription factor, putative, expressed
Intron Retention	Chr08	13899972	13900091	intron	Vr05	30683549	30683668	intron	zinc ion binding protein, putative, expressed
Intron Retention	Chr08	13899972	13900091	exon 3' end	Vr05	30683549	30683668	exon 3' end	zinc ion binding protein, putative, expressed
Intron Retention	Chr08	13899972	13900091	exon 5' end	Vr05	30683549	30683668	exon 5' end	zinc ion binding protein, putative, expressed
Alternative Donor	Chr07	40088673	40088971	exon 5' end	Vr07	9471799	9472042	exon 5' end	unknown
Alternative Donor	Chr07	40088673	40088916	exon 3' end	Vr07	9471799	9472042	exon 3' end	unknown
Alternative Donor	Chr07	40088673	40088971	exon 3' end	Vr07	9471799	9472097	exon 3' end	unknown
Exon Skipping	Chr17	10006847	10007099	exon 3' end	Vr07	34496540	34496791	exon 3' end	unknown
Exon Skipping	Chr17	10006847	10007099	exon 5' end	Vr07	34496540	34496791	exon 5' end	unknown
Exon Skipping	Chr17	7672264	7672319	exon 3' end	Vr07	35662171	35662226	exon 3' end	nuclear protein ZAP-related, putative, expressed
Exon	Chr17	7672264	7672319	exon 5'	Vr07	35662171	35662226	exon 5'	nuclear protein ZAP-related, putative,

Skipping				end				end	expressed
Exon Skipping	Chr07	44528131	44528192	exon 3' end	Vr07	50694954	50695015	exon 3' end	non-lysosomal glucosylceramidase, putative, expressed
Exon Skipping	Chr07	44528131	44528192	exon 5' end	Vr07	50694954	50695015	exon 5' end	non-lysosomal glucosylceramidase, putative, expressed
Exon Skipping	Chr10	48204025	48204091	exon 3' end	Vr08	41923821	41923887	exon 3' end	Inositol 1, 3, 4-trisphosphate 5/6-kinase, putative, expressed
Exon Skipping	Chr10	48204025	48204091	exon 5' end	Vr08	41923821	41923887	exon 5' end	Inositol 1, 3, 4-trisphosphate 5/6-kinase, putative, expressed
Exon Skipping	Chr10	51347373	51347433	exon 5' end	Vr08	45455121	45455181	exon 5' end	unknown
Exon Skipping	Chr10	51347373	51347433	exon 3' end	Vr08	45455121	45455181	exon 3' end	unknown
Exon Skipping	Chr20	47701043	47701103	exon 3' end	Vr08	45455121	45455181	exon 3' end	unknown
Exon Skipping	Chr20	47701043	47701103	exon 5' end	Vr08	45455121	45455181	exon 5' end	unknown

DISCUSSION

Based on the findings in this study, we conclude that the noise hypothesis fits the pattern of AS events in mungbean; consequently, a large proportion of AS isoforms in mungbean probably have no function. However, the noisy splicing model does not exclude the possibility that useful and functional AS isoforms could emerge among the resultant non-functional isoforms. As suggested by our observations regarding genes with high exon numbers or expression levels, natural selection will act on genes that produce AS isoforms at concentrations that could be dangerous to the cell, possibly by favoring the propagation of individuals with stronger splicing signals at the correct bases, or those that lack the bases that are used as AS sites. Similarly, AS isoforms that confer selective advantage will be retained or even strengthened, so that the spliceosome will regularly cut at the alternative site as well as the regular site. The presence of numerous low-abundance isoforms is probably not significantly harmful to mungbean cells, but it could be useful as a source of variants upon which natural selection can act.

The arguments in support of noisy splicing have been outlined by several groups (Hon et al., 2013; Pickrell et al., 2010; Zhang et al., 2009). Melamud and Moult (2009b) noted that while some tissue-specific AS

isoforms are conserved across species, these represent a relatively small fraction of AS events. A large proportion of AS isoforms also carry premature stop codons, which make them vulnerable to NMD. Even if these isoforms are translated, most of the alternative protein structures are predicted to be non-viable. The number of detected AS isoforms also tends to increase in genes with more introns or genes expressed at higher levels, in line with the view of AS as a probabilistic event. The more introns to be processed, the greater the probability of noisy splicing. Similarly, higher levels of gene expression increase the likelihood of a splicing error among the pool of processed transcripts. This could explain why RNAseq data typically allow the detection of more AS isoforms, because the sequenced libraries usually have a very high sequencing depth. Thus, splicing errors that are not normally found in average cells become visible; this is compounded by the use of PCR during library preparation, which could amplify uncommon transcripts to a more easily detectable level.

Protein studies provide another line of evidence supporting noisy splicing. In human cells, the observed protein diversity revealed by high-throughput mass spectrophotometry is much lower than that predicted from AS studies of transcriptome data. Abascal et al. (2015) found that most human proteins exist as single dominant isoform, and detected only 282 AS isoforms among 12,716 genes at the protein level; this is nowhere near the prediction of 95% based on transcriptome data (Pan et al., 2008). However,

the absence of protein products of a given AS isoform may not necessarily mean that the isoform serves no function; in some cases, degradation of AS isoforms through the NMD pathway serves to regulate the concentration of transcripts in the cell (Drechsel et al., 2013; Fu et al., 2009). Nevertheless, this lack of representation at the protein level undermines the idea that AS increases the protein diversity generated by a given number of genes.

On the other hand, several compelling arguments support the notion that AS plays an important regulatory role in the cell. Barbazuk et al. (2008) presented several lines of evidence for the functional importance of AS: the predominance of AS in some gene families, versus its absence in others; the existence of AS events that correlate with specific tissue and developmental cues; incorporation of AS products into ribosomes; and conservation of some AS events among distantly related species. Because noisy splicing has probably existed since the emergence of introns in eukaryotes, it is likely that a large number of useful isoforms have evolved from it, resulting in the phenomena detailed above. However, based on the observed low level of conservation among species, the contribution of AS to protein variation and regulation of gene expression does not appear to be significant.

The actual proportion of functional AS isoforms in mungbean obviously cannot be precisely determined without experimental tests. However, in

general the number of AS isoforms is lower in plants than in animals (Kim et al., 2008). It will be interesting to speculate whether this is due to the different research focus in plants and animals, or instead to intrinsic differences in the splicing mechanisms in the two kingdoms. One issue that should be investigated is the effect of genome expansion (e.g., polyploidization) on AS. An organism whose genome cannot tolerate significant expansion will benefit greatly by the ability to increase the functionality of its existing genome via AS. Hence, to generate additional transcript diversity, it may be advantageous to maintain a less-specific splicing machinery. However, polyploid plants can easily obtain new gene variants by allowing duplicated genes to evolve independently; this strategy is potentially safer because it does not disrupt the function of the original gene. In allotetraploid soybean, duplicated genes undergo less AS (Shen et al., 2014); this is curious because in such cases, the penalty for incorrect splicing would be less severe because a backup copy is present elsewhere in the genome.

Because examples of functional AS in plants are rare, it would be prudent to assume that a significant portion of AS in plants has no distinct function; thus, more evidence is required to support the claim that there is extensive functional diversity generated by AS in plants. Thus, detected AS events should be treated like genetic marker data, which are useful to identify and catalogue, but strong experimental evidence such as QTL

mapping and transgene expression are necessary to assert that a given sequence variation causes a particular phenotype. Additionally, approaches like comparative genomics, which are used to select candidate markers that are likeliest to alter a phenotype, could also be applied to finding AS isoforms that encode a novel function. We believe that our comparative AS detection data could be used as a starting point to perform more in-depth studies of the phenotypic diversity generated by AS.

REFERENCES

Abascal, F., Ezkurdia, I., Rodriguez-Rivas, J., Rodriguez, J.M., del Pozo, A., Vazquez, J., Valencia, A. and Tress, M.L. (2015) Alternatively Spliced Homologous Exons Have Ancient Origins and Are Highly Expressed at the Protein Level. PLoS computational biology 11, e1004325.

Barbazuk, W.B., Fu, Y. and McGinnis, K.M. (2008) Genome-wide analyses of alternative splicing in plants: opportunities and challenges. Genome research 18, 1381-1392.

Breitbart, R.E., Andreadis, A. and Nadal-Ginard, B. (1987) Alternative splicing: a ubiquitous mechanism for the generation of multiple protein isoforms from single genes. Annual review of biochemistry 56, 467-495.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T.L. (2009) BLAST+: architecture and applications. BMC bioinformatics 10, 1.

Chamala, S., Feng, G., Chavarro, C. and Barbazuk, W.B. (2015) Genome-wide identification of evolutionarily conserved alternative splicing events in flowering plants. Frontiers in bioengineering and biotechnology 3, 33.

Drechsel, G., Kahles, A., Kesarwani, A.K., Stauffer, E., Behr, J., Drewe, P., Rätsch, G. and Wachter, A. (2013) Nonsense-mediated decay of alternative precursor mRNA splicing variants is a major determinant of the *Arabidopsis* steady state transcriptome. The Plant cell 25, 3726-3742.

Filichkin, S.A. and Mockler, T.C. (2012) Unproductive alternative splicing and nonsense mRNAs: A widespread phenomenon among plant circadian clock genes. Biology Direct 7, 20.

Filichkin, S.A., Priest, H.D., Givan, S.A., Shen, R., Bryant, D.W., Fox, S.E., Wong, W.K. and Mockler, T.C. (2010) Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*. Genome research 20, 45-58.

Florea, L., Song, L. and Salzberg, S.L. (2013) Thousands of exon skipping events differentiate among splicing patterns in sixteen human tissues. F1000Research 2.

Foissac, S. and Sammeth, M. (2007) ASTALAVISTA: dynamic and flexible analysis of alternative splicing events in custom gene datasets. Nucleic acids research 35, W297-W299.

Fu, Y., Bannach, O., Chen, H., Teune, J.H., Schmitz, A., Steger, G., Xiong, L. and Barbazuk, W.B. (2009) Alternative splicing of anciently

exonized 5S rRNA regulates plant transcription factor TFIIIA. Genome research 19, 913-921.

Graveley, B.R. (2001) Alternative splicing: increasing diversity in the proteomic world. Trends in Genetics 17, 100-107.

Hon, C.C., Weber, C., Sismeiro, O., Proux, C., Koutero, M., Deloger, M., Das, S., Agrahari, M., Dillies, M.A., Jagla, B., Coppee, J.Y., Bhattacharya, A. and Guillen, N. (2013) Quantification of stochastic noise of splicing and polyadenylation in *Entamoeba histolytica*. Nucleic Acids Res 41, 1936-1952.

Inoue, K., Hoshijima, K., Sakamoto, H. and Shimura, Y. (1990) Binding of the Drosophila sex-lethal gene product to the alternative splice site of transformer primary transcript.

Kang, Y.J., Kim, S.K., Kim, M.Y., Lestari, P., Kim, K.H., Ha, B.K., Jun, T.H., Hwang, W.J., Lee, T., Lee, J., Shim, S., Yoon, M.Y., Jang, Y.E., Han, K.S., Taeprayoon, P., Yoon, N., Somta, P., Tanya, P., Kim, K.S., Gwag, J.G., Moon, J.K., Lee, Y.H., Park, B.S., Bombarely, A., Doyle, J.J., Jackson, S.A., Schafleitner, R., Srinives, P., Varshney, R.K. and Lee, S.H. (2014) Genome sequence of mungbean and insights into evolution within *Vigna* species. Nature communications 5, 5443.

Kang, Y.J., Satyawan, D., Shim, S., Lee, T., Lee, J., Hwang, W.J., Kim, S.K., Lestari, P., Laosatit, K. and Kim, K.H. (2015) Draft genome sequence of adzuki bean, *Vigna angularis*. Scientific reports 5.

Kawashima, T., Douglass, S., Gabunilas, J., Pellegrini, M. and Chanfreau, G.F. (2014) Widespread use of non-productive alternative splice sites in *Saccharomyces cerevisiae*. PLoS genetics 10, e1004249.

Kim, E., Goren, A. and Ast, G. (2008) Alternative splicing: current perspectives. BioEssays : news and reviews in molecular, cellular and developmental biology 30, 38-47.

Lah, G.J.-e., Li, J.S.S. and Millard, S.S. (2014) Cell-specific alternative splicing of Drosophila Dscam2 is crucial for proper neuronal wiring. Neuron 83, 1376-1388.

Lu, T., Lu, G., Fan, D., Zhu, C., Li, W., Zhao, Q., Feng, Q., Zhao, Y., Guo, Y., Huang, X. and Han, B. (2010) Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq. Genome research 20, 1238-1249.

Marquez, Y., Brown, J.W., Simpson, C., Barta, A. and Kalyna, M. (2012) Transcriptome survey reveals increased complexity of the alternative splicing landscape in *Arabidopsis*. Genome research 22, 1184-1195.

Melamud, E. and Moult, J. (2009a) Stochastic noise in splicing machinery. Nucleic Acids Res 37, 4873-4886.

Melamud, E. and Moult, J. (2009b) Structural implication of splicing stochastics. Nucleic Acids Res 37, 4862-4872.

Nair, R., Schafleitner, R., Kenyon, L., Srinivasan, R., Easdown, W., Ebert, A. and Hanson, P. (2012) Genetic improvement of mungbean. SABRAO Journal of Breeding and Genetics 44, 177-190.

Neu-Yilik, G., Gehring, N.H., Hentze, M.W. and Kulozik, A.E. (2004) Nonsense-mediated mRNA decay: from vacuum cleaner to Swiss army knife. Genome biology 5, 218.

Pan, Q., Shai, O., Lee, L.J., Frey, B.J. and Blencowe, B.J. (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nature genetics 40, 1413-1415.

Pickrell, J.K., Pai, A.A., Gilad, Y. and Pritchard, J.K. (2010) Noisy splicing drives mRNA isoform diversity in human cells. PLoS genetics 6, e1001236.

Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841-842.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G. and Mesirov, J.P. (2011) Integrative genomics viewer. Nature biotechnology 29, 24-26.

Rubatzky, V.E. and Yamaguchi, M. (1997) Peas, beans, and other vegetable legumes. In: World Vegetables pp. 474-531. Springer.

Severing, E.I., van Dijk, A.D., Stiekema, W.J. and van Ham, R.C. (2009) Comparative analysis indicates that alternative splicing in plants has a limited role in functional expansion of the proteome. BMC genomics 10, 154.

Shen, Y., Zhou, Z., Wang, Z., Li, W., Fang, C., Wu, M., Ma, Y., Liu, T., Kong, L.A., Peng, D.L. and Tian, Z. (2014) Global Dissection of Alternative Splicing in Paleopolyploid Soybean. The Plant cell.

Shi, Y., Sha, G. and Sun, X. (2014) Genome-wide study of NAGNAG alternative splicing in *Arabidopsis*. Planta 239, 127-138.

Thatcher, S.R., Zhou, W., Leonard, A., Wang, B.B., Beatty, M., Zastrow-Hayes, G., Zhao, X., Baumgarten, A. and Li, B. (2014) Genome-wide analysis of alternative splicing in *Zea mays*: landscape and genetic regulation. The Plant cell 26, 3472-3487.

Trapnell, C., Pachter, L. and Salzberg, S.L. (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25, 1105-1111.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L. and Pachter, L. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature protocols 7, 562-578.

Ullrich, B., Ushkaryov, Y.A. and Südhof, T.C. (1995) Cartography of neurexins: more than 1000 isoforms generated by alternative splicing and expressed in distinct subsets of neurons. Neuron 14, 497-507.

Zhang, Z., Xin, D., Wang, P., Zhou, L., Hu, L., Kong, X. and Hurst, L.D. (2009) Noisy splicing, more than expression regulation, explains why some exons are subject to nonsense-mediated mRNA decay. BMC biology 7, 23.

CHAPTER II

Assessment of mungbean genetic diversity and domestication based on genotyping by sequencing

ABSTRACT

Mungbean (*Vigna radiata*) is an important vegetable and source of carbohydrate and protein in Asia, but limited data is available regarding their genetic characteristics. We performed genotyping by sequencing on 276 diverse mungbean germplasm from more than 20 countries to assess their genetic diversity, population structure, linkage disequilibrium, and identify chromosome segments under selection during domestication. At least 4% of the genome was sequenced in each accession, with an average sequencing depth of 8.6 times. A total of 18,722 variants were identified and 18,213 of them were single nucleotide polymorphisms. Nucleotide diversity in cultivated accessions is reduced to 30% of the level found in wild accessions. Linkage disequilibrium decays to background level after 18,600 bases among the wild accessions, compared to 86,000 bases among cultivated accessions. Wild and cultivated accessions are clearly separated in phylogenetic and population structure analysis, and correlation between

geographical origin and phylogenetic clustering was also observed. Several chromosome segments with reduced diversity among cultivated accessions were identified, which could be the result of selective sweep during mungbean domestication. The genes in those intervals are enriched with genes associated with growth and reproductive traits.

Keywords: Genotyping by sequencing; Mungbean; genetic diversity; domestication

INTRODUCTION

Traditional plant breeding can be viewed as an effort to accumulate useful alleles from various parental lines into a new plant variety. Since superior alleles and the individuals that carry them are often difficult to identify, one way to improve the chances in obtaining them is by utilizing germplasm pool with high genetic diversity. Before the advent of next generation sequencing and high throughput genotyping system, quantification of genetic diversity of a population is limited by the number of polymorphic genetic markers available for the organism. This introduces bias since the sampled sequence may have different allelic diversity compared to the rest of the genome (Moragues et al., 2010). As the cost of genome sequencing becomes more affordable, the proportion of the genome that can be assessed for the presence of variation increases considerably, to the extent that resequencing the whole genome of hundreds or thousands of crop accessions is feasible for some research groups (Li et al., 2014; Zhou et al., 2015). Sequence variation data can also be viewed in the context of their physical location in the genome if a reference genome sequence is also available. This has greatly expanded the number of analysis that can be performed on sequence data.

The availability of high density DNA sequence variation data along with information about their position relative to coding regions and regulatory elements has enabled researchers to deduce the evolutionary history of an organism (Consortium, 2012), assess the diversity and population structure of breeding materials (Valliyodan et al., 2016), and map loci that are responsible for domestication and the emergence of agronomical traits commonly selected by modern breeders (Huang et al., 2010). Chromosomal segments that were selected during domestication reflect the preference of farmers, breeders, and consumers since ancient times, so understanding the molecular mechanisms that create such traits is important for further improvement in the future as well as for domesticating wild plants that possess potential commercial values. Such loci typically have lower allelic diversity in cultivated accessions compared to wild accessions and can be identified by scanning the whole genome if sufficiently dense DNA variation data exist along the genome (Meyer and Purugganan, 2013).

Another approach that is commonly used to identify useful alleles in a germplasm collection is genome-wide association study (GWAS), where the association between DNA sequence variations and the presence of desired traits in a population is statistically evaluated to identify chromosomal regions that could harbor the causal genes or mutations that alter the traits (Hirschhorn and Daly, 2005). The approach had been shown to be a sound method, as it could correctly identify genes that had been known to controls

certain traits (Andersen et al., 2005). Effective GWAS has several prerequisites: a population that has high diversity in both genotype and phenotype of interest, DNA markers at sufficient density that cover most or all linkage disequilibrium blocks, data on confounding cofactors such as population structure and kinship, and precise phenotypic data (Zhu et al., 2008). Phenotypic diversity is important since there is obviously little value in trying to map drought tolerance genes if the population does not exhibit variation in drought tolerance. High marker density is necessary because insufficient number of markers with uneven coverage of the genome means that some part of the genome will not be assessed in the association analysis. Genetic kinship and population structure data is associated with the risk of obtaining false association when calculating marker-trait association. Many alleles may appear to be significantly associated with the trait but they could exist simply because accessions that are related to each other share more similar alleles in a large part of their genome, due to a common recent ancestor. Imprecise phenotyping will obviously reduce the accuracy of the calculation of association.

In this study, we applied genotyping by sequencing (GBS) on 42 wild mungbeans and 233 cultivated mungbeans from more than 20 countries, along with a *Vigna glabrescens* accession as an outgroup. Using the allelic variation data generated from GBS, we assessed their genetic diversity, population structure, pattern of linkage disequilibrium, and footprint of

domestication process. The resulting data can be used to assess the effect of selective breeding on allelic diversity in the mungbean genome and determine whether GWAS is a feasible approach to identify the causal loci for various agronomic traits in this population.

MATERIALS AND METHODS

Sequencing and variant calling

Genotyping by sequencing was performed according to Elshire et al. (2011), where DNA samples extracted from mungbean leaves were digested with Ape KI enzyme and ligated to Illumina sequencing adapters containing unique sequence barcode for each sample and pooled for multiplexed sequencing in Illumina HiSeq 2000. The resulting sequencing reads were filtered based on their barcodes to separate reads originated from different accessions. Following removal of adapter sequences from each read, reads were aligned to mungbean reference genome using BWA package (Li and Durbin, 2009) and sequence variants were identified using Samtool's (Li et al., 2009) mpileup command. Variants were filtered and annotated using vcftools (Danecek et al., 2011) and SnpEff package (Cingolani et al., 2012), and only variants with quality score higher than 30 with sequencing depth of more than 2 reads per site were kept for further analysis. Circular plot for variant number across chromosomes was drawn in Circos (Krzywinski et al., 2009).

Phylogenetic and population structure analysis

For the construction of phylogenetic tree, SNP variant data were inputted to DarWin 6 (Perrier and Jacquemoud-Collet, 2006) to generate a dissimilarity matrix, which was then used to create a neighbor-joining tree. The tree was exported to Figtree v1.4.2 (Rambaut, 2007) to draw the tree and also converted to newick format required by the tree drawing program iTOL (Letunic and Bork, 2007), which was used to stack charts and text on top of phylogenetic trees.

Population structure analysis was performed using a random sample of 1000 SNP in STRUCTURE 2.3.4 program (Pritchard et al., 2000). The number of clusters were set from k=2 to k=7, with 50,000 burn in and 100,000 MCMC reps. Other parameters were set to default settings. To observe the dominant cluster in each country of origin, population structure data for each accession was grouped according country, and the q-values were averaged. The data was then charted onto google map using PhyloGeoViz application (http://phylogeoviz.org/). Principal component analysis were performed using GenAlEx 6.5 (Peakall and Smouse, 2006) in Microsoft Excel, and PCA plot for the two most significant PC was also plotted in Excel.

Linkage Disequilibrium Profiling

Linkage disequilibrium was calculated based on r2 statistics (Hill and Robertson, 1968) implemented in vcftools (Danecek et al., 2011). LD calculation was performed separately for wild and cultivated mungbeans. The r2 between markers were then binned and averaged for each 10,000 bp increments, and charted in Microsoft Excel using logarithmic trend lines. LD decay cutoff was set at 0.2, and the exact base pair position for the cutoff value was calculated from the trend line equation.

Identification of selective sweep regions for domestication

Fixation index (Fst) between wild and cultivated accessions were calculated according to Hudson et al. (1992) as implemented in PopGenome (Pfeifer et al., 2014) in sliding windows of 400,000 bp. SNPs that undergo positive selection were identified using an approach developed by Beaumont and Nichols (1996) and implemented in LOSITAN (Antao et al., 2008) with the following settings: 100,000 simulations, 0.99 confidence interval, 0.01 false discovery rate, infinite alleles mutation model, “Neutral” mean Fst, and force mean Fst. The procedure was run 10 times, and loci that were consistently identified as undergoing positive selection were selected. Gene ontology analysis was performed by submitting Arabidopsis

homologs of the genes inside the loci identified by LOSITAN to AgriGO (Du et al., 2010)

RESULTS

Profiles and distribution of sequence variants

In total, 276 accessions were sequenced, comprising 42 wild and 233 cultivated mungbean accessions. The sequencing was performed using Genotyping by sequencing (GBS) protocols in Illumina HiSeq 2000 sequencer. Since GBS only sequence a small subset of the genome located between two *ApeK*I restriction sites that are in close proximity to each other, a global sequencing coverage calculation is meaningless since a large portion of the genome will not be sequenced. In this case, more than 96% of the genome was not sequenced, even in mungbean accession with the highest sequence coverage. As for the remaining 4% that were sequenced, on average those regions were sequenced 13 times in the accession with the highest sequence coverage, while in the accession with the lowest sequence coverage the average depth was 4.3 times.

The sequence data were then aligned to mungbean reference genome sequence (Kang et al., 2014) using BWA aligner, and sequence variations in each accession were identified and filtered to obtain variants that are supported by at least 2 times sequencing depth in all accessions and have quality score of at least 30. A summary of the variant calling statistics can be

found in Table II-1, which lists the number and types of the variants and the location of the discovered variants relative to coding regions along with their potential impacts on the proteins coded by the coding regions.

The distribution of the variant across the genome is relatively even, and there are only few regions in the genome where a sequence variant was not found in a long stretch of the chromosome (Figure II-1B). There is also observable reduction in the number of variants that could be found in each chromosome interval among cultivated accessions compared to the wild ones, and this trend could be observed throughout the genome. The average nucleotide diversity ($\boldsymbol{\pi}$) values calculated in 100 kilobase window among wild accessions is 7.63E-06 and it decreased to 2.33E-06 in cultivated accessions. These numbers may underestimate the real value since it was assumed that the whole genome was resequenced instead of a mere 4% covered by GBS. Nevertheless, it still serves a useful comparison to quantify extent of sequence diversity reduction in cultivated mungbean.

Table II-1. Properties of the variants obtained from genotyping by sequencing of 276 wild and cultivated mungbean

GENOME TOTAL LENGTH	464,815,385
GENOME EFFECTIVE LENGTH	435,725,587
SEQUENCING READ NUMBER	
MINIMUM	660387
MAXIMUM	14258096
MEDIAN	2628462
NUMBER OF ALIGNED READS	
MINIMUM	613257 (14.86%)
MAXIMUM	4593382 (95.70%)
MEDIAN	2114011 (91.93%)
VARIANT RATE	1 variant every 23,273 bases
NUMBER OF VARIANTS	18,722
SNP	18,213
INS	184
DEL	325
LOCATION RELATIVE TO GENES	
DOWNSTREAM	6,989
EXON	11,541
INTERGENIC	3,654
INTRON	3,210
SPLICE SITE ACCEPTOR	8
SPLICE SITE DONOR	19
SPLICE SITE REGION	227
TRANSCRIPT	6
UPSTREAM	4,758
UTR 3 PRIME	412
UTR 5 PRIME	534

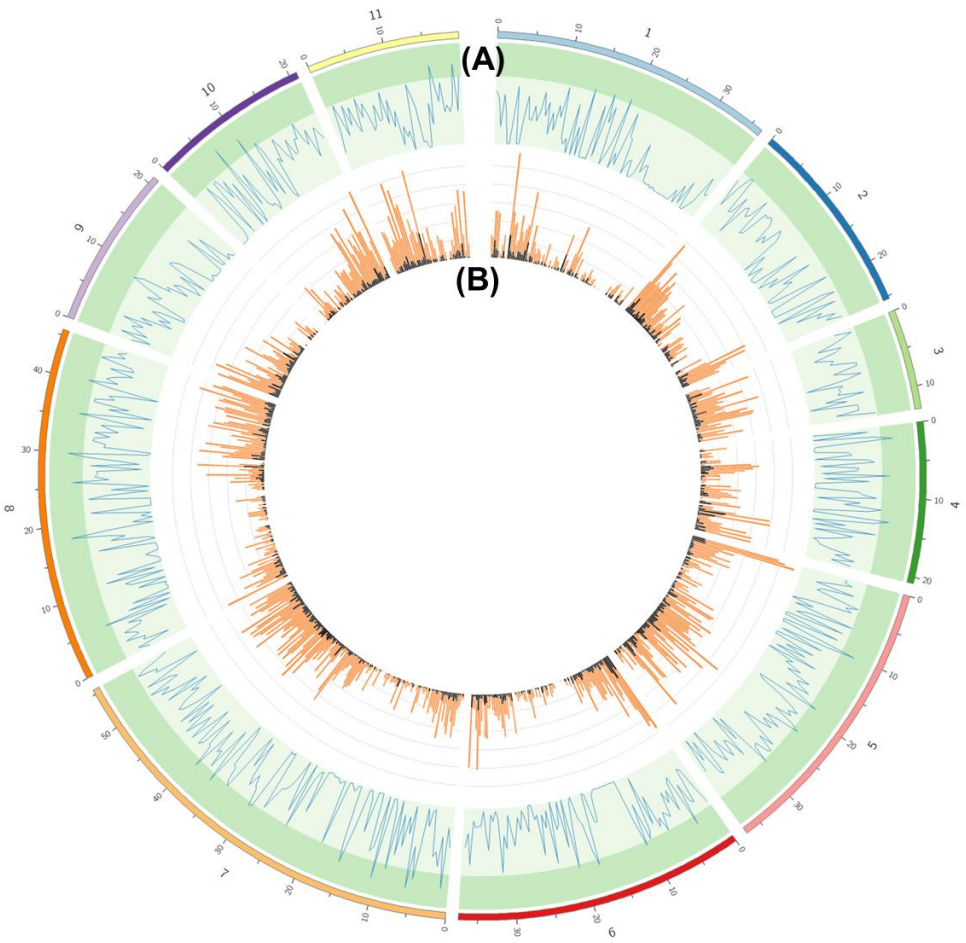


Figure II-1. Distribution of sequence variants along the 11 chromosomes of mungbean. **(A)** Fst values between wild and cultivated accessions across the chromosomes (1-11). The peaks that intersect the dark green area are the 10% highest values, which indicate loci with the most differentiation between wild and cultivated accessions and probably play important roles in domestication. **(B)** Histogram of the number of variants along the chromosomes for wild (orange) and cultivated (black) accessions.

Phylogenetic relationship and population structure

Phylogenetic analysis was conducted to assess the relatedness of each accession to each other, and estimate the divergence and genetic diversity among the accessions. Pervasive population structure usually is also observable in a phylogenetic tree, which could be useful during accession selection for GWAS. As observed in nucleotide diversity calculation, there are observable differences in terms of similarity and spread among cultivated and wild accessions. Cultivated accessions show little genetic divergence to each other compared to the wild accessions (Figure II-2). The cultivar that shows most resemblance to wild accession is cultivar JP231223 from India (Figure II-3), and the cluster which is closest to the wild accessions is dominated by accessions from India and its neighbors (Pakistan, Myanmar, and China). This supports the hypothesis that mungbean was domesticated in India (Fuller, 2007). The presence of two Australian accessions is interesting, as mungbean cultivation in that country only began relatively recently.

To elucidate the presence of population structure and trace the ancestry of each accession, a Bayesian inference analysis was performed in STRUCTURE using 1000 randomly selected SNPs. Figure II-4 shows the pattern of admixture in each accession, viewed in the context of their phylogenetic relationship. A calculation based on ln likelihood indicated that

the optimum model is obtained when the population is divided into 3 populations when both the cultivated and wild accessions were included, and also into 3 groups when only the cultivated accessions were assessed. There are some correlations between population group membership and geographical origin (Figure II-5), as genetic background of one subgroup seems to predominate in accessions originated from nearby areas.

Figure II-2. In an unrooted neighbor-joining tree, cultivated mungbean (black) form a small and tight cluster compared to wild accessions (red) even though there were 233 cultivated accessions compared to just 42 wild accessions. This indicates a high genetic similarity among them and a significant reduction of genetic diversity compared to the wild population.

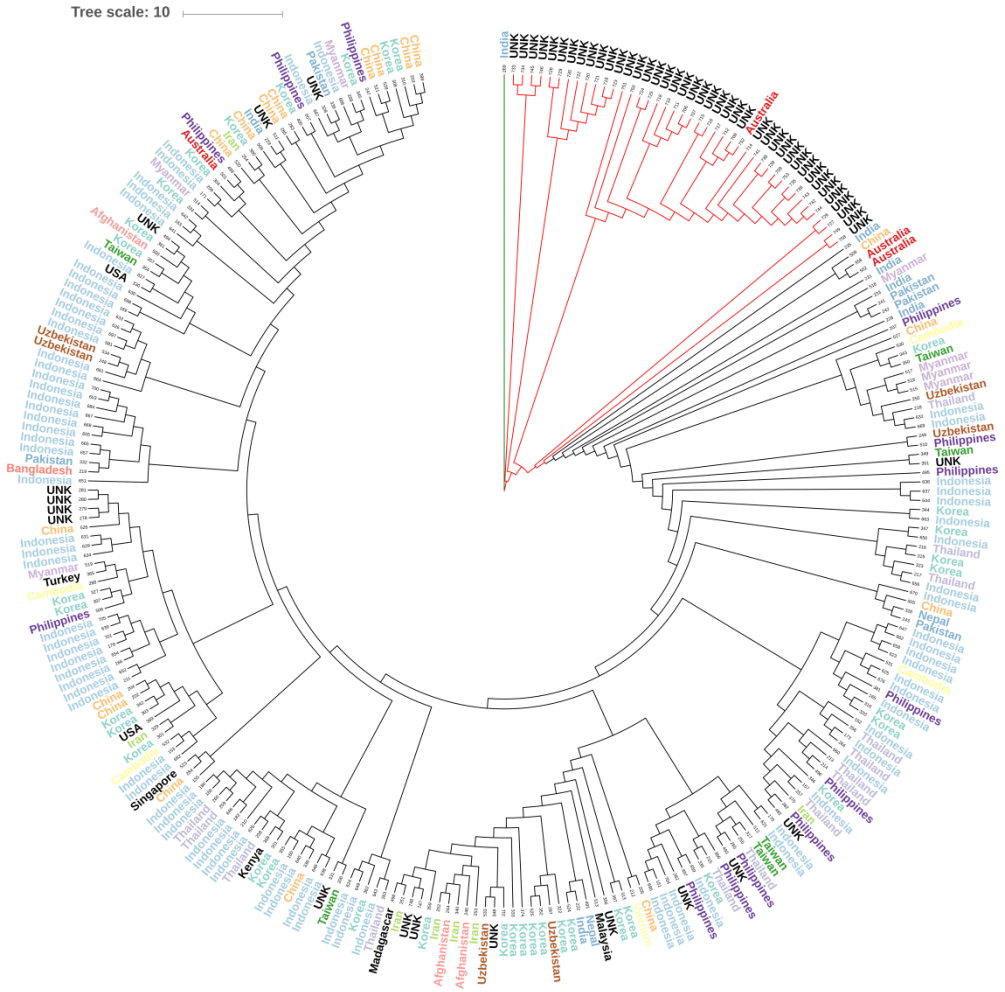


Figure II-3. Origins of mungbean accessions sequenced in this study, viewed in the context of their phylogenetic relationship. Unk denotes accessions with no known country of origin data. Red branches indicate wild accessions.

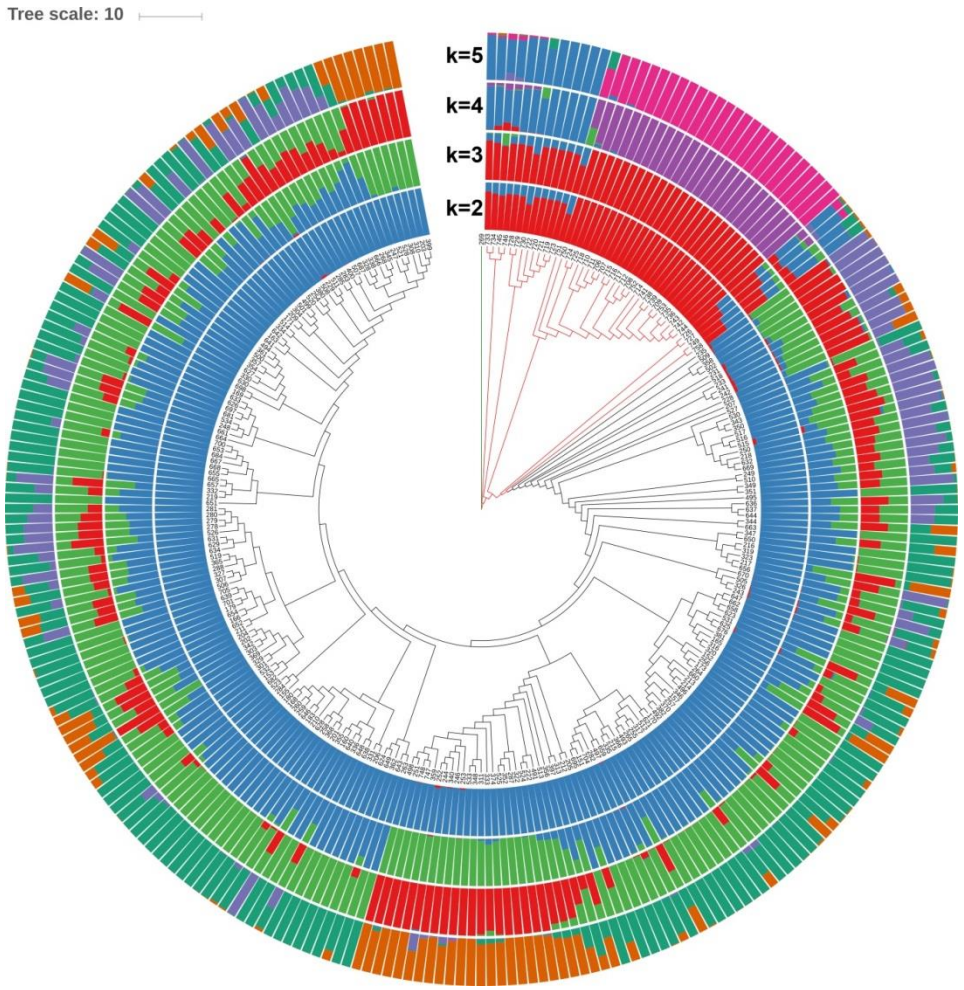


Figure II-4. Pattern of population structure calculated using Bayesian clustering in STRUCTURE program. Each column represents a single accession and colors in the columns represent genetic components contributed by each subpopulation when the accessions were divided into 2, 3, 4, and 5 subpopulation (k=2 to k=5). The accessions were ordered according to their phylogenetic relationship in a neighbor-joining tree.

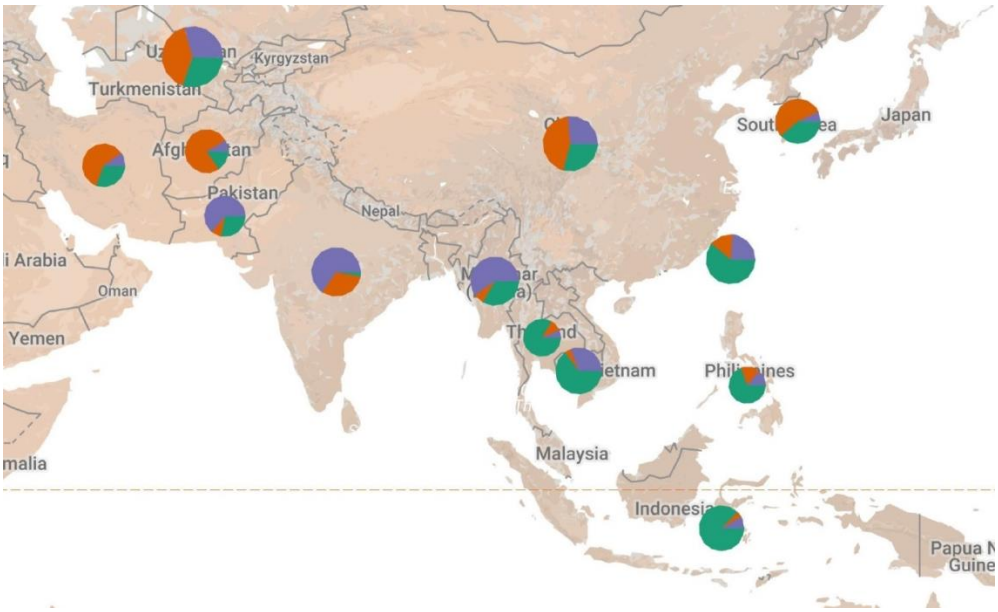


Figure II-5. Average genetic contributions from each of the 3 population structure subgroups in cultivated accessions obtained from different countries. Only countries that contributed more than 3 accessions were included. Geographical origin seems to partly explain the membership of accessions from the same countries in population subgroups, as nearby countries with similar climates are dominated by the same subgroups.

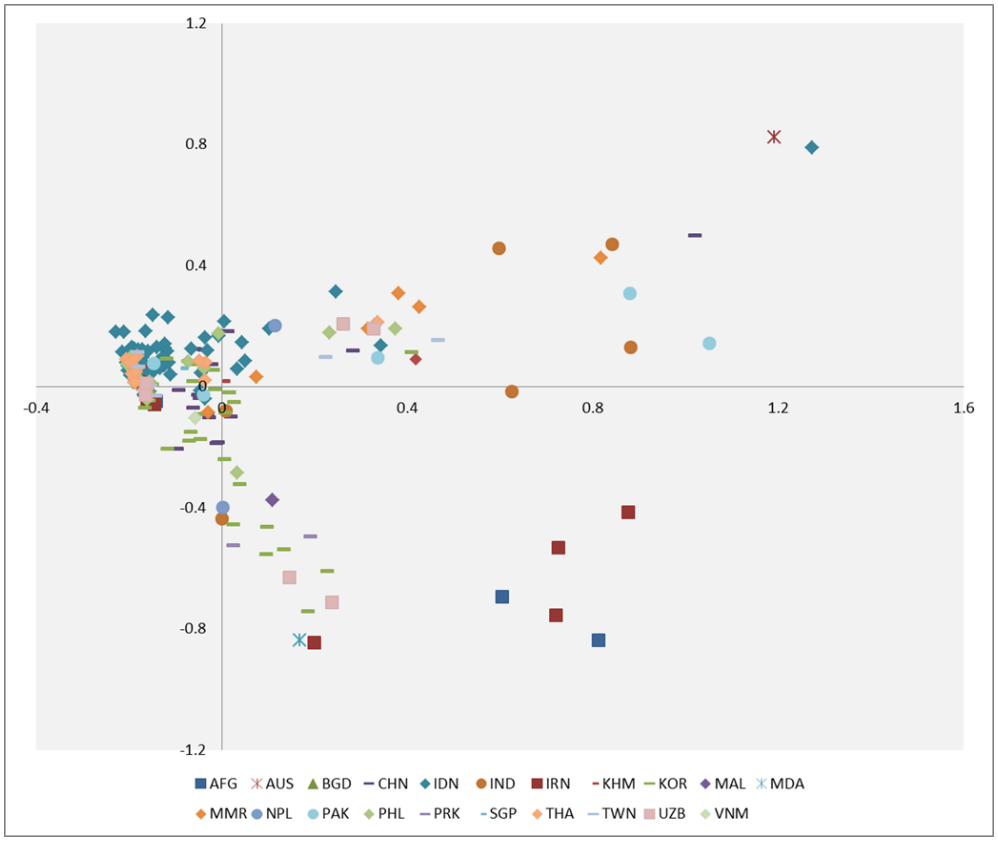


Figure II-6. Plotting the two main components that explain the most variation from principal component analysis (PCA) in x axis and y axis reveals clustering of accessions that is consistent with their geographical origins. Accessions that are separated from other accessions from the same country could be the result of recent germplasm exchange.

Principal component analysis (PCA) seems to confirm this trend, as accessions from certain areas often cluster together when the components that explain the largest variations were plotted together (Figure II-6). Together, this could be used to infer which accessions are developed at a certain geographical locale and which are the result of recent introduction. Frequent germplasm exchange is a prominent feature in modern cultivar development, and without genetic data it is often difficult to distinguish introduced lines from native lines, especially if some degree of hybridization has been performed.

Linkage disequilibrium

The extent of linkage disequilibrium (LD) across the genome is important to quantify, especially if GWAS is planned in the future. The size of LD blocks influence the design of GWAS experiment, especially in deciding the minimum number of markers required in the experiment. Obviously higher marker number is better, but it can be prohibitively costly. On the other hand, if the number of markers is too low, many LD blocks may not be sampled and their association to the trait of interest will not be obtainable.

Figure 5 shows the differences in average distance of LD decay in wild and cultivated mungbean. There are several ways to define the average

size of LD block in a population, such as by setting a threshold value, and the distance where the average r^2 drops below the threshold is defined as the end of an LD block. Caldwell et al. (2006) used 0.2 as the threshold value, and using this value, the LD block size in cultivated mungbean is 86,000 bp compared to 18,600 bp in wild mungbean. Consequently, GWAS study in cultivated mungbean should utilize at least 5,400 markers, or 24,990 markers if wild mungbeans are also included.

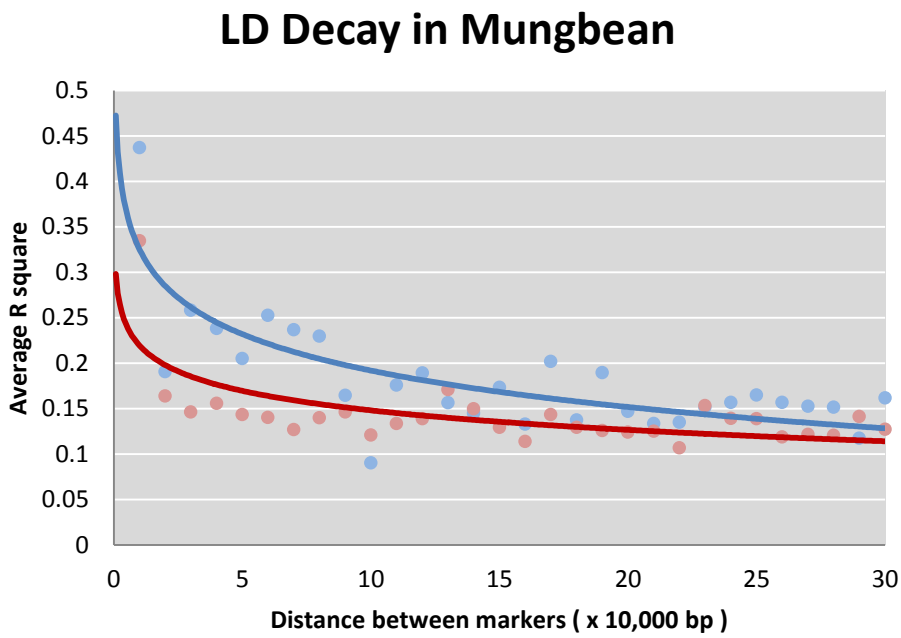


Figure II-7. Plot of average linkage disequilibrium (measured as R square between markers in y axis) decay as markers are further separated from each other in the chromosome (x axis).

Regions undergoing selective sweep in domesticated mungbean

The availability of sequence variation data in populations comprising a large number of wild and cultivated accessions makes it possible to observe chromosome segments that are under selection during domestication. Selected regions generally show less diversity since only certain alleles were desirable to ancient farmers that selected them, and further evolutions in such loci are more limited due to lower probability of obtaining variants with no negative effects on the selected trait and bottleneck effect caused by the small number of ancestral lines being selected.

Several methods can be used to detect such regions, and the simplest approach is by comparing the nucleotide diversity in each locus between cultivated and wild populations. Comparison of fixation index (Fst) across chromosomes can also be used to identify loci that shows the most differentiation between two populations (Figure II-1A), although the direction of the selection (ie. whether the fixation is necessary for survival in the wild or selection for domestication) can be difficult to discern. Such simple approaches are also prone to false positives generated by genetic drift and founder effect. Due to the limited number of variants that could be identified using GBS, Fst calculation could only be performed in a very wide sliding window interval, which in this case was 400,000 bp. Such a wide interval means that candidate genes undergoing selective sweep are very difficult to

determine, and selective sweep occurring at a narrow segment could be masked by surrounding segments.

Here, we utilized an approach developed by Beaumont and Nichols (1996) and implemented in LOSITAN, which is equipped to deal with very large number of variants and have low false positives and false negatives in comparison to other algorithms (Narum and Hess, 2011), and identified 135 intervals that are under positive selection (Table II-2). Fifteen intervals also overlap with 41 intervals that represent the top 5% intervals with the highest Fst values in Popgenome. Overall, the 135 intervals cover more than 7 million bases (approximately 1.5% of the genome) and intersect with 470 genes. Gene ontology analysis of those genes showed enrichment in DNA-binding protein, hydrolase, and phosphorus transferase activity, which could describe the activity of transcription factors and signal transduction proteins (Figure II-8A). Location-wise, the proteins under selection are enriched for membrane proteins, especially in cell-to-cell junction (Figure II-8B). In terms of biological processes that are controlled by the genes, there is GO enrichment for developmental growth, meristem growth and maintenance, reproductive structure development (flower, gametes, fruit, and seeds), organelle organization, and metabolic processes for minerals, protein, and lipids (Figure II-9). All of those processes are what is expected to be up-regulated in plants that are selected for more rigorous growth and reproductive capacity during domestication.

Table II-2. Chromosome intervals identified as having outlier value of Fst, which is indicative of positive selection and the number of genes located in those intervals. Intervals in bold are intervals that were also identified as the top 5% highest Fst values in simple pairwise Fst calculation in Popgenome.

Chromosome	Interval Start	Interval End	Interval Size	No. of Genes in Interval
Vr01	10052975	10052990	15	0
Vr01	10407884	10448159	40275	4
Vr01	12635230	12747292	112062	4
Vr01	**17049153**	**17049196**	**43**	**1**
Vr01	3236258	3236293	35	1
Vr01	33322461	33439703	117242	9
Vr01	4585760	4585870	110	1
Vr01	4657173	4657203	30	1
Vr01	4657255	4661323	4068	1
Vr01	4700974	4700990	16	1
Vr01	4905513	4947448	41935	5
Vr01	6884091	6898584	14493	5
Vr01	7954130	7954189	59	1
Vr01	8463077	8480783	17706	3
Vr01	8895037	8895062	25	0
Vr01	8895046	8946680	51634	3
Vr01	9073066	9080219	7153	2
Vr01	9080219	9080236	17	0
Vr01	9080236	9080269	33	0
Vr01	9080251	9080334	83	0
Vr01	9080384	9080389	5	0
Vr01	9080390	9080407	17	0
Vr01	9080476	9080503	27	0
Vr02	10661243	11233540	572297	23
Vr02	11619797	11619809	12	1
Vr02	1179113	1179143	30	0
Vr02	1361688	1398023	36335	4
Vr02	**1848700**	**1879167**	**30467**	**4**
Vr02	23396049	23410094	14045	2

Vr02	24643057	24645350	2293	1
Vr02	3099823	3099847	24	0
Vr02	530702	570802	40100	7
Vr02	5333402	5358675	25273	2
Vr02	5572700	5572713	13	1
Vr02	6404761	6404802	41	1
Vr03	10830039	10830845	806	1
Vr03	10830927	10867702	36775	5
Vr03	12633656	12685384	51728	5
Vr03	12809815	12881890	72075	6
Vr03	7603018	7603168	150	1
Vr04	**13571230**	**13571280**	**50**	**0**
Vr04	**13571268**	**13571291**	**23**	**0**
Vr04	14630805	14666590	35785	3
Vr04	15691636	15691655	19	1
Vr04	20687326	20718434	31108	6
Vr04	5067798	5070025	2227	2
Vr04	614827	741061	126234	5
Vr05	15906599	15906607	8	1
Vr05	21156965	21156970	5	1
Vr05	2285694	2411344	125650	12
Vr05	23687567	23767100	79533	8
Vr05	238150	240347	2197	1
Vr05	24392377	24394334	1957	1
Vr05	24394464	24436471	42007	4
Vr05	24436530	24485993	49463	7
Vr05	24513651	24536366	22715	4
Vr05	26522548	26569523	46975	5
Vr05	33236355	33282084	45729	5
Vr05	33721405	33721438	33	1
Vr05	867008	867035	27	1
Vr06	14648477	15693810	1045333	18
Vr06	2605848	2654293	48445	4
Vr06	**2949727**	**3005013**	**55286**	**4**
Vr06	30160843	30236085	75242	6
Vr06	31389477	31391903	2426	1
Vr06	33507032	33507059	27	0

Vr06	36481	39911	3430	0
Vr06	3941684	3993453	51769	5
Vr06	456134	456188	54	1
Vr06	9259771	9290065	30294	2
Vr07	16954267	17207781	253514	10
Vr07	**23821298**	**24143747**	**322449**	**10**
Vr07	37236850	37236873	23	1
Vr07	38428099	38462789	34690	7
Vr07	41775466	41854773	79307	7
Vr07	49062266	49083452	21186	4
Vr07	49517938	49522118	4180	2
Vr07	**5039553**	**5146253**	**106700**	**9**
Vr07	51383881	51383966	85	1
Vr07	51383942	51383978	36	1
Vr07	51547944	51590733	42789	5
Vr07	8026771	8111508	84737	3
Vr07	8111613	8271306	159693	7
Vr07	**947186**	**1045679**	**98493**	**9**
Vr08	24682436	24685872	3436	1
Vr08	**25769281**	**25770969**	**1688**	**1**
Vr08	29304501	29305356	855	1
Vr08	31361393	31445006	83613	9
Vr08	32951657	32969629	17972	5
Vr08	3439794	3586283	146489	9
Vr08	36941887	36941921	34	1
Vr08	37056457	37202229	145772	9
Vr08	37217547	37236631	19084	1
Vr08	37619077	37629079	10002	2
Vr08	37945235	37994101	48866	6
Vr08	37994101	37999001	4900	1
Vr08	38442376	38613139	170763	8
Vr08	38626643	38631171	4528	1
Vr08	**40057373**	**40082072**	**24699**	**2**
Vr08	42794590	42836128	41538	6
Vr08	43971132	43971189	57	1
Vr08	43971143	44024160	53017	5
Vr08	44988909	44988966	57	1

Vr08	7160880	7171196	10316	1
Vr09	10796963	10796987	24	1
Vr09	10796975	10796996	21	1
Vr09	11705933	11745316	39383	2
Vr09	1614202	1614223	21	1
Vr09	1792890	1811871	18981	2
Vr09	23488	44400	20912	3
Vr09	3841783	4027889	186106	7
Vr09	4117386	4150087	32701	2
Vr09	4346255	4346348	93	1
Vr09	4633397	4886875	253478	13
Vr09	4886875	4886895	20	1
Vr09	5039588	5039642	54	1
Vr09	**8447953**	**8447984**	**31**	**0**
Vr10	**11365140**	**11371928**	**6788**	**0**
Vr10	**11371923**	**11392825**	**20902**	**2**
Vr10	**11392934**	**11495025**	**102091**	**7**
Vr10	15830053	15847459	17406	2
Vr10	9062514	9124687	62173	4
Vr11	10936774	11828609	891835	31
Vr11	**15455469**	**15542812**	**87343**	**2**
Vr11	15926205	15948963	22758	2
Vr11	16169086	16194601	25515	3
Vr11	2351772	2358277	6505	1
Vr11	3050290	3076023	25733	6
Vr11	4054145	4054206	61	1
Vr11	5387499	5387512	13	1
Vr11	6095263	6095299	36	1
Vr11	6799763	6850633	50870	3
Vr11	7183270	7183314	44	1
Vr11	7287362	7325124	37762	5
Vr11	8625943	8625975	32	1
Grand Total			**7118858**	**470**

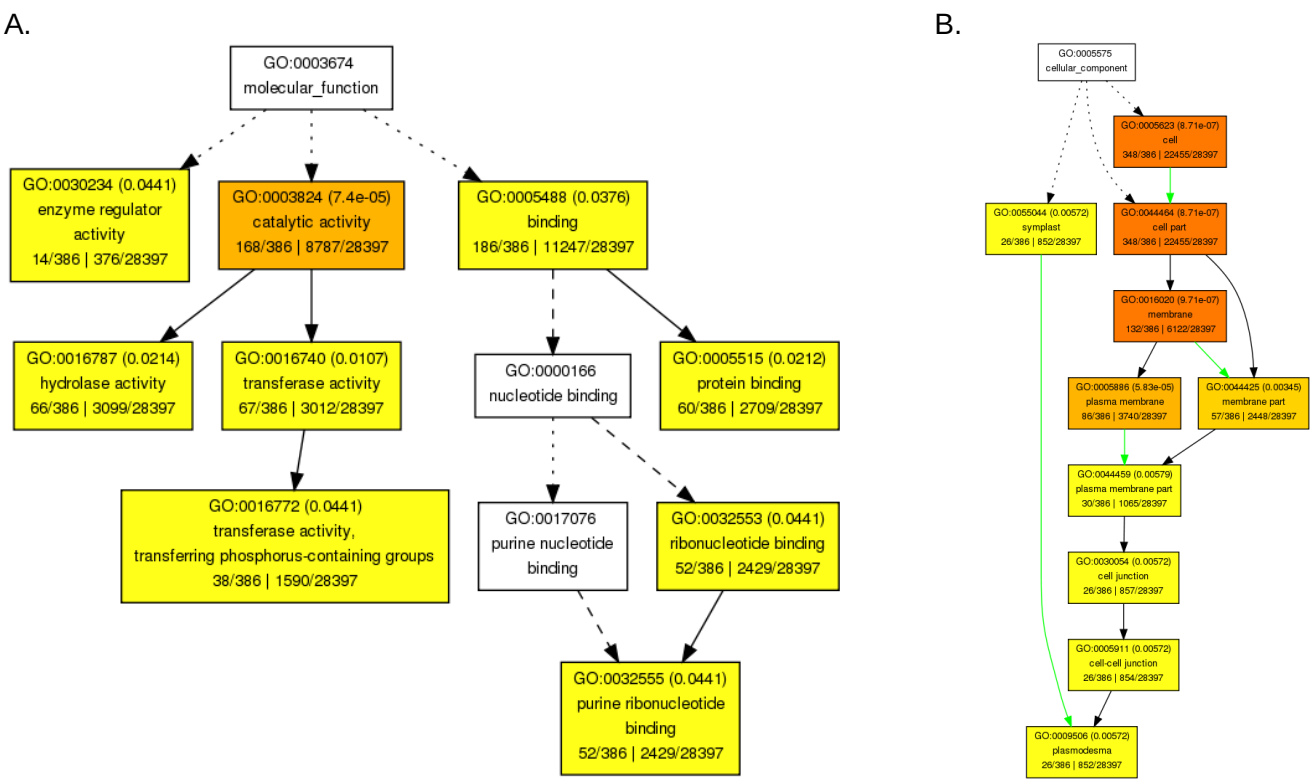


Figure II-8. Gene ontology enrichment of genes located in segments with Fst outlier found by LOSITAN according to their function in the cell. (A) Genes that showed enrichment in the molecular function groups seem to perform activities commonly associated with signal transduction proteins and transcription factors. (B) Genes belonging to cellular component group are enriched by membrane proteins.

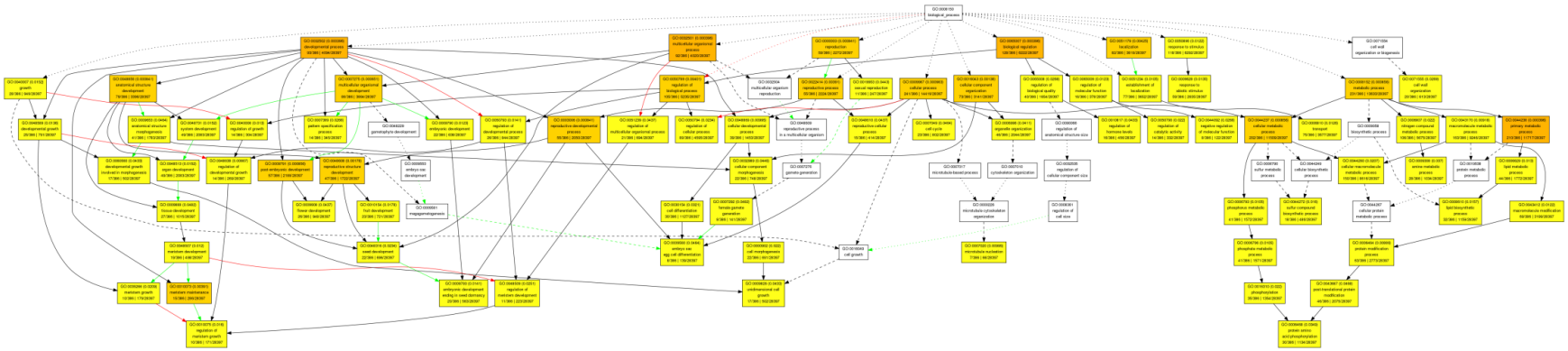


Figure II-9. Gene ontology enrichment of genes located in segments with Fst outlier found by LOSITAN that are involved in biological processes. The genes are enriched for genes involved in developmental growth, meristem growth and maintenance, reproductive structure development (flowers, gametes, fruits, and seeds), organelle organization, and metabolic processes for minerals, protein, and lipids.

DISCUSSION

Domestication in mungbean resulted in significant reduction in genetic diversity. Nucleotide diversity level in cultivated accessions is only 30% of the diversity level observed in wild mungbean. This reduction is worse compared to soybean, where diversity in landraces is approximately 47% of the value observed in wild accessions and even elite cultivars have 35% of the level of diversity of wild accessions (Zhou et al., 2015). Linkage disequilibrium also decays over a longer distance in cultivated mungbean accessions, which on average is 4.6 times longer than in wild accessions. In soybean, LD is 3 times longer in landraces and 4.9 times longer in elite cultivars (Zhou et al., 2015). Longer distance of LD decay means that smaller number of markers is required for GWAS, at the expense of resolution. Lack of resolution can make candidate gene identification more difficult since significant markers will cover longer intervals in the chromosomes, which are more likely to contain more genes.

Such problem was also observed in the identification of loci that were selected during mungbean domestication. Genome-wide scan of fixation index across the chromosomes could only be performed in wide intervals due to the lack of SNPs. Consequently, even though the regions with the most fixations could be identified, the intervals are too wide to pinpoint the

candidate genes that could be selected during domestication. The use of Beaumont and Nichols algorithm could narrow down some of the intervals, but the majority still spans segments with multiple genes in it. Even when the interval is narrow enough to contain only a single gene, deducing the mechanism of how such genes contribute to domestication based on available information in public database is not always easy. For example, the interval in chromosome Vr01 between bases 3,236,528 and 3,236,293 intersects with Vradi01g01790. This gene is most similar to AT5G39680.1 gene in Arabidopsis, which has these gene ontology annotations: zinc ion binding for molecular function, mitochondrion for its cellular location, and involvement in embryo development ending in seed dormancy. It is expressed in flower, plant embryo (at cotyledonary stage), seed, shoot apex, guard cell, and petals (Schmid et al., 2005). A similar search in soybean shows that its homolog is also involved in embryo development, vascular tissue histogenesis, chloroplast RNA processing, and leaf development. It is also located in a QTL region for Aluminum tolerance, and the highest expression is in flower and apical meristem.

From the descriptions of the gene homologs, we can deduce on how such gene can make positive contribution to plant growth. Nonetheless, why those processes are heavily differentiated between wild and cultivated mungbean population is difficult to explain without further experimental works like creating near isogenic lines for the selective sweep segments.

Another approach that could be taken for candidate genes with even less obvious function is by cross-referencing the domestication sweep data with the results of other mapping studies like QTL analysis or GWAS for specific traits that may contribute to domestication. This would possibly narrow down the function to a more specific question, such as what role in seed weight instead of broad domestication.

This study illustrated the benefits that can be obtained from higher resolution genomic study in a germplasm collection. Information collected regarding nucleotide diversity, phylogenetic relationship, and phenotypic performance of each accession can be utilized to manage germplasm collection. Large germplasm collection captures most of the existing natural genetic variation, but can also be expensive to maintain and breeders often have difficulties in identifying accessions that can be useful for their breeding program. Some genebanks develop core collections for many different crops, which contain smaller subset of the whole collection, yet maintain the maximum genotypic and phenotypic diversity that can be obtained from them (Upadhyaya and Ortiz, 2001). The data generated in this study can be used to generate such collection for mungbean, and assist breeders in choosing parental lines with desired level of genetic diversity, depending on the aim of the cultivar development.

The identification of loci that are selected during mungbean domestication also offer many benefit for future breeders. Domestication traits reflect the qualities that are preferred by farmers and consumers, so future breeders wishing to further fine-tune those traits will be assisted by the identification of the genes and mechanisms that form those traits. Progress in gene editing technology could someday allow breeders to tinker with target genes instead of finding and incorporating beneficial natural variations. Such approach may become a necessity in a world where rapid climate change and habitat destruction may reduce the chance for naturally adaptive crops to evolve.

Understanding the process of domestication will also enable us to repeat the domestication process for new potential crops in shorter duration. This is currently being attempted in potatoes, since existing potato cultivars are tetraploids and more difficult to improve due to their genetic characteristics. Breeding progress can be accelerated if a new diploid variety can be developed from diploid wild potatoes (Jansky and Peloquin, 2006). Mungbean belongs to the genus *Vigna*, which comprise many wild species with interesting and potentially useful agronomic traits. Several species with unique traits like pod size that reach 60 cm in length, ability to grow in salt water and rocky lands, have been identified (Tomooka et al., 2002). If those useful traits cannot be transferred to mungbean due to incompatibility, it may be better to domesticate those species instead. Arguably,

domestication data from mungbean will be more useful for those species than data from unrelated model species.

REFERENCES

Andersen, J.R., Schrag, T., Melchinger, A.E., Zein, I. and Lübberstedt, T. (2005) Validation of Dwarf8 polymorphisms associated with flowering time in elite European inbred lines of maize (*Zea mays* L.). Theoretical and Applied Genetics 111, 206-217.

Antao, T., Lopes, A., Lopes, R.J., Beja-Pereira, A. and Luikart, G. (2008) LOSITAN: a workbench to detect molecular adaptation based on a F ST-outlier method. BMC bioinformatics 9, 1.

Beaumont, M.A. and Nichols, R.A. (1996) Evaluating loci for use in the genetic analysis of population structure. Proceedings of the Royal Society of London B: Biological Sciences 263, 1619-1626.

Bradbury, P.J., Zhang, Z., Kroon, D.E., Casstevens, T.M., Ramdoss, Y. and Buckler, E.S. (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23, 2633-2635.

Butler, J., MacCallum, I., Kleber, M., Shlyakhter, I.A., Belmonte, M.K., Lander, E.S., Nusbaum, C. and Jaffe, D.B. (2008) ALLPATHS: de novo assembly of whole-genome shotgun microreads. Genome research 18, 810-820.

Caldwell, K.S., Russell, J., Langridge, P. and Powell, W. (2006) Extreme population-dependent linkage disequilibrium detected in an inbreeding plant species, *Hordeum vulgare*. Genetics 172, 557-567.

Chamala, S., Feng, G., Chavarro, C. and Barbazuk, W.B. (2015) Genome-wide identification of evolutionarily conserved alternative splicing events in flowering plants. Frontiers in bioengineering and biotechnology 3, 33.

Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X. and Ruden, D.M. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly 6, 80-92.

Collard, B.C. and Mackill, D.J. (2008) Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Philosophical Transactions of the Royal Society B: Biological Sciences 363, 557-572.

Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W. and Marraffini, L.A. (2013) Multiplex genome engineering using CRISPR/Cas systems. Science 339, 819-823.

Consortium, T.G. (2012) The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485, 635-641.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T. and Sherry, S.T. (2011) The variant call format and VCFtools. Bioinformatics 27, 2156-2158.

Du, Z., Zhou, X., Ling, Y., Zhang, Z. and Su, Z. (2010) agriGO: a GO analysis toolkit for the agricultural community. Nucleic acids research, gkq310.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S. and Mitchell, S.E. (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PloS one 6, e19379.

Faino, L., Seidl, M.F., Datema, E., van den Berg, G.C., Janssen, A., Wittenberg, A.H. and Thomma, B.P. (2015) Single-molecule real-time sequencing combined with optical mapping yields completely finished fungal genome. MBio 6, e00936-00915.

Fuller, D. Q. (2007). Contrasting patterns in crop domestication and domestication rates: recent archaeobotanical insights from the Old World. Annals of Botany, 100(5), 903-924.

Hill, W. and Robertson, A. (1968) Linkage disequilibrium in finite populations. Theoretical and Applied Genetics 38, 226-231.

Hirschhorn, J.N. and Daly, M.J. (2005) Genome-wide association studies for common diseases and complex traits. Nature Reviews Genetics 6, 95-108.

Huan, T., Rong, J., Liu, C., Zhang, X., Tanriverdi, K., Joehanes, R., Chen, B.H., Murabito, J.M., Yao, C. and Courchesne, P. (2015) Genome-wide identification of microRNA expression quantitative trait loci. Nature communications 6.

Huang, X., Wei, X., Sang, T., Zhao, Q., Feng, Q., Zhao, Y., Li, C., Zhu, C., Lu, T. and Zhang, Z. (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. Nature genetics 42, 961-967.

Huang, X., Yang, S., Gong, J., Zhao, Q., Feng, Q., Zhan, Q., Zhao, Y., Li, W., Cheng, B. and Xia, J. (2016) Genomic architecture of heterosis for yield traits in rice. Nature 537, 629-633.

Hudson, R.R., Slatkin, M. and Maddison, W. (1992) Estimation of levels of gene flow from DNA sequence data. Genetics 132, 583-589.

Jansky, S.H. and Peloquin, S.J. (2006) Advantages of wild diploid *Solanum* species over cultivated diploid relatives in potato breeding programs. Genetic Resources and Crop Evolution 53, 669-674.

Joshi, T., Patil, K., Fitzpatrick, M.R., Franklin, L.D., Yao, Q., Cook, J.R., Wang, Z., Libault, M., Brechenmacher, L. and Valliyodan, B. (2012) Soybean Knowledge Base (SoyKB): a web resource for soybean translational genomics. BMC genomics 13, 1.

Kang, H., Wang, Y., Peng, S., Zhang, Y., Xiao, Y., Wang, D., Qu, S., Li, Z., Yan, S. and Wang, Z. (2016) Dissection of the genetic architecture of rice resistance to the blast fungus *Magnaporthe oryzae*. Molecular plant pathology.

Kang, Y.J., Kim, S.K., Kim, M.Y., Lestari, P., Kim, K.H., Ha, B.K., Jun, T.H., Hwang, W.J., Lee, T., Lee, J., Shim, S., Yoon, M.Y., Jang, Y.E., Han, K.S., Taeprayoon, P., Yoon, N., Somta, P., Tanya, P., Kim, K.S., Gwag, J.G., Moon, J.K., Lee, Y.H., Park, B.S., Bombarely, A., Doyle, J.J., Jackson, S.A., Schafleitner, R., Srinives, P., Varshney, R.K. and Lee, S.H. (2014) Genome sequence of mungbean and insights into evolution within *Vigna* species. Nature communications 5, 5443.

Kang, Y.J., Lee, T., Lee, J., Shim, S., Jeong, H., Satyawan, D., Kim, M.Y. and Lee, S.H. (2015) Translational genomics for plant breeding with the genome sequence explosion. Plant biotechnology journal.

Keatinge, J., Easdown, W., Yang, R., Chadha, M. and Shanmugasundaram, S. (2011) Overcoming chronic malnutrition in a future warming world: the key importance of mungbean and vegetable soybean. Euphytica 180, 129-141.

Khattak, G., Haq, M., Ashraf, M., Tahir, G. and Marwat, E. (2001) Detection of epistasis, and estimation of additive and dominance components of genetic variation for synchrony in pod maturity in mungbean (*Vigna radiata* (L.) Wilczek). Field Crops Research 72, 211-219.

Korte, A. and Farlow, A. (2013) The advantages and limitations of trait analysis with GWAS: a review. Plant methods 9, 1.

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

Letunic, I. and Bork, P. (2007) Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics 23, 127-128.

Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078-2079.

Li, J.-Y., Wang, J. and Zeigler, R.S. (2014) The 3,000 rice genomes project: new opportunities and challenges for future rice research. GigaScience 3, 1.

Lipka, A.E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P.J., Gore, M.A., Buckler, E.S. and Zhang, Z. (2012) GAPIT: genome association and prediction integrated tool. Bioinformatics 28, 2397-2399.

Luo, M., Gilbert, B. and Ayliffe, M. (2016) Applications of CRISPR/Cas9 technology for targeted mutagenesis, gene replacement and stacking of genes in higher plants. Plant cell reports, 1-12.

Maher, C.A., Palanisamy, N., Brenner, J.C., Cao, X., Kalyana-Sundaram, S., Luo, S., Khrebtukova, I., Barrette, T.R., Grasso, C. and Yu, J. (2009) Chimeric transcript discovery by paired-end transcriptome sequencing. Proceedings of the National Academy of Sciences 106, 12353-12358.

Metzker, M.L. (2010) Sequencing technologies—the next generation. Nature Reviews Genetics 11, 31-46.

Meyer, R.S. and Purugganan, M.D. (2013) Evolution of crop species: genetics of domestication and diversification. Nature Reviews Genetics 14, 840-852.

Moragues, M., Comadran, J., Waugh, R., Milne, I., Flavell, A. and Russell, J.R. (2010) Effects of ascertainment bias and marker number on estimations of barley diversity from high-throughput SNP genotype data. Theoretical and Applied Genetics 120, 1525-1534.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. and Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature methods 5, 621-628.

Nair, R., Schafleitner, R., Kenyon, L., Srinivasan, R., Easdown, W., Ebert, A. and Hanson, P. (2012) Genetic improvement of mungbean. SABRAO Journal of Breeding and Genetics 44, 177-190.

Nakaya, A. and Isobe, S.N. (2012) Will genomic selection be a practical method for plant breeding? Annals of botany 110, 1303-1316.

Narum, S.R. and Hess, J.E. (2011) Comparison of FST outlier tests for SNP loci under selection. Molecular Ecology Resources 11, 184-194.

Neeraja, C., Maghirang-Rodriguez, R., Pamplona, A., Heuer, S., Collard, B., Septiningsih, E., Vergara, G., Sanchez, D., Xu, K. and Ismail, A. (2007) A marker-assisted backcross approach for developing submergence-tolerant rice cultivars. Theoretical and Applied Genetics 115, 767-776.

Neu-Yilik, G., Gehring, N.H., Hentze, M.W. and Kulozik, A.E. (2004) Nonsense-mediated mRNA decay: from vacuum cleaner to Swiss army knife. Genome biology 5, 218.

Park, P.J. (2009) ChIP–seq: advantages and challenges of a maturing technology. Nature Reviews Genetics 10, 669-680.

Peakall, R. and Smouse, P.E. (2006) GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. Molecular ecology notes 6, 288-295.

Perrier, X. and Jacquemoud-Collet, J. (2006) DARwin software.

Pfeifer, B., Wittelsbürger, U., Onsins, S.E.R. and Lercher, M.J. (2014) PopGenome: an efficient Swiss army knife for population genomic analyses in R. Molecular biology and evolution, msu136.

Priest, S.H. (2000) US public opinion divided over biotechnology? Nature biotechnology 18, 939-942.

Pritchard, J.K., Stephens, M. and Donnelly, P. (2000) Inference of population structure using multilocus genotype data. Genetics 155, 945-959.

Rambaut, A. (2007) FigTree, a graphical viewer of phylogenetic trees. See http://tree. bio. ed. ac. uk/software/figtree.

Schmid, M., Davison, T.S., Henz, S.R., Pape, U.J., Demar, M., Vingron, M., Schölkopf, B., Weigel, D. and Lohmann, J.U. (2005) A gene expression map of *Arabidopsis thaliana* development. Nature genetics 37, 501-506.

Schuster, S.C. (2007) Next-generation sequencing transforms today's biology. Nature 200, 16-18.

Skelly, D.A., Johansson, M., Madeoy, J., Wakefield, J. and Akey, J.M. (2011) A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. Genome research 21, 1728-1737.

Smith, Z.D., Gu, H., Bock, C., Gnirke, A. and Meissner, A. (2009) High-throughput bisulfite sequencing in mammalian genomes. Methods 48, 226-232.

Spindel, J., Begum, H., Akdemir, D., Virk, P., Collard, B., Redoña, E., Atlin, G., Jannink, J.-L. and McCouch, S.R. (2015) Genomic selection and association mapping in rice (*Oryza sativa*): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. PLoS genetics 11, e1004982.

Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A. and Conesa, A. (2011) Differential expression in RNA-seq: a matter of depth. Genome research 21, 2213-2223.

Tomooka, N., Vaughan, D.A., Moss, H. and Maxted, N. (2002) Introduction. In: The Asian *Vigna* pp. 1-7. Springer.

Upadhyaya, H. and Ortiz, R. (2001) A mini core subset for capturing diversity and promoting utilization of chickpea genetic resources in crop improvement. Theoretical and Applied Genetics 102, 1292-1298.

Valliyodan, B., Qiu, D., Patil, G., Zeng, P., Huang, J., Dai, L., Chen, C., Li, Y., Joshi, T. and Song, L. (2016) Landscape of genomic diversity and trait discovery in soybean. Scientific reports 6.

Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics 10, 57-63.

Yu, J., Pressoir, G., Briggs, W.H., Bi, I.V., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M. and Holland, J.B. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature genetics 38, 203-208.

Zerbino, D.R. and Birney, E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome research 18, 821-829.

Zhou, Z., Jiang, Y., Wang, Z., Gou, Z., Lyu, J., Li, W., Yu, Y., Shu, L., Zhao, Y., Ma, Y., Fang, C., Shen, Y., Liu, T., Li, C., Li, Q., Wu, M., Wang, M., Wu, Y., Dong, Y., Wan, W., Wang, X., Ding, Z., Gao, Y., Xiang, H., Zhu, B., Lee, S.H., Wang, W. and Tian, Z. (2015) Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nature biotechnology 33, 408-414.

Zhu, C., Gore, M., Buckler, E.S. and Yu, J. (2008) Status and prospects of association mapping in plants. The Plant Genome 1, 5-20.

CHAPTER III

Genome-wide association study to identify loci associated with agronomic traits

ABSTRACT

Dissecting the genetic basis of important agronomic traits will assist future breeding programs in mungbean, an important vegetable and protein source in Asia. We performed genome-wide association study in mungbean using 7551 SNP markers developed using genotyping by sequencing on 222 cultivated mungbean accessions from all over the world. The traits being evaluated were flowering time, maturity time, pod formation time, seed weight, number of seeds per pod, peak harvest, cumulative weekly harvest, final yield, and synchronicity. Using two different association methods, 77 and 79 markers were found to be significantly associated with some of the traits at p-value <0.0001 respectively. Some of the genes that intersect with significant markers share homology with soybean genes and QTL that could explain their role in trait formation, which makes them attractive candidates for follow up studies. The data can be used as a basis for mapping studies or parental selection in mungbean breeding programs.

Keywords: GWAS; genotyping by sequencing; Mungbean; agronomic traits

INTRODUCTION

Dissecting the genetic basis of valuable traits in crops is a challenging task, since traditionally it involves tracking genotypic and phenotypic variation in a segregating population created from hybridization of genetically diverse parents. Such approach is usually constrained by the number of alleles present in the parental lines, the time required to synthesize the population, and the amount of recombination occurred in the population to obtain sufficient resolution and pinpoint the targeted causal genes (Korte and Farlow, 2013). Those drawbacks can be addressed in genome-wide association studies, since it utilizes more genetically diverse accessions that generally do not need to be crossed to each other. Linkage disequilibrium also tends to be less of a problem since more recombination events have occurred during the evolution of the accessions. This increases the resolution that can be obtained in association mapping, provided that sufficient number of markers can be obtained to cover the whole genome. Due to the declining cost of whole genome sequencing, larger number of DNA markers can now be identified at each gene or intervals that are within the linkage disequilibrium distance from each other.

The recent publication of mungbean reference genome (Kang et al., 2014) means that this approach is now applicable for this crop in order to

dissect the genetic basis of agronomic traits. Traits that contribute to yield such as seed weight and pod number is obviously important since all mungbean improvement programs ultimately must produce varieties with economically viable yield. Another trait that is still a major problem in mungbean cultivation is synchronicity. Flowering does not occur at uniform time in mungbean, but is typically spread out over a long period (Khattak et al., 2001). If harvesting is only performed once, a large portion of the yield potential could be wasted, as the peak early harvest period often only accounts for 50% of the total yield that could be obtained from a cultivar. But maximizing yield by delaying harvest can also cause yield loss as mature pods may fall off or lost by pest and pathogen attacks. Preventing yield loss through multiple harvests also has its own challenges as it represent additional cost to the selling price and each harvest must also be done carefully to prevent damaging the plants, which could complicate the harvesting procedure and prevent the usage of mechanical tools.

Synchronous plants are the ideal solution, but the complexity of the genetic basis of the trait in mungbean is currently unknown. If the trait is controlled by a few loci with large effects, then breeding for this trait should be easier to accomplish. The interaction of the trait with other agronomic traits is also unclear. If persistent negative correlation to yield is found, then incorporating the trait into a new cultivar will be more complicated since we

need to consider the balance between the cost saving of single harvest and the loss of yield incurred by the addition of synchronicity trait.

To answer these questions, we planted 234 accessions of wild and cultivated mungbeans in June 2016, and scored agronomic traits such as flowering time, maturity time, seed weight, and pod numbers. Synchronicity was deduced by performing weekly harvest of mature pods for eight weeks on each accession. The genetic basis of those traits was then dissected by using association analysis with thousands of SNP data obtained using genotyping by sequencing.

MATERIALS AND METHODS

Plant Materials

The association panel consisted of 232 mungbean accessions originated from 23 countries (Table III-1). Based on genotype analysis and plant morphology, 222 accessions can be categorized as cultivated varieties, while the remaining are wild accessions. The accessions were planted at Seoul National University experimental farm in Suwon, Korea at June 22, 2016. Each accession was planted in a row containing 10 plants with 15 cm space between them, and the distance between rows was set at 70 cm.

Table III-1. List of accessions genotyped and phenotyped for GWAS to map agronomic traits.

GBS Code	Accession Name	Genebank Code	Country of Origin	Field Code	Wild/ Cultivar
491	네팔 Pokhara 수집	IT109327	NPL	18_1_1	Cultivated
265	NATIVE COLLECTION	IT201188	PHL	18_1_2	Cultivated
494	Acc. 7861	801374	UNK	18_1_3	Cultivated
495	Acc. 363	807483	PHL	18_1_4	Cultivated
496	EG MG-13	807492	PHL	18_1_5	Cultivated
497	Dull Yellow	807498	PHL	18_1_6	Cultivated
268	K001419	K001419	MMR	18_1_7	Cultivated
498	RUS-NYW-2000-201	K005410	MDA	18_1_8	Cultivated
499	EG-MG-60	K024059	PHL	18_1_10	Cultivated
502	V01471	K024060	IDN	18_1_11	Cultivated
500	V01673	K024062	AFG	18_1_12	Cultivated
501	HERNITAGERS	K024067	AUS	18_1_13	Cultivated
505	ML-9	K130624	CHN	18_1_14	Cultivated
506	CES-J-24	K130625	PHL	18_1_15	Cultivated
507	FG-MG-1743	K130627	PHL	18_1_16	Cultivated
508	VC-2307A	K130631	CHN	18_1_17	Cultivated
509	R-288-8	K130632	CHN	18_1_18	Cultivated
510	CES-87	K130634	PHL	18_1_19	Cultivated
511	P-3-A-40	K130639	UNK	18_1_20	Cultivated
277	Pe Nauk	K130643	UNK	18_1_21	Cultivated
278	P-69-319	K130644	UNK	18_1_22	Cultivated
279	P-4-44	K130646	UNK	18_1_23	Cultivated
512	NCM-1	K130648	TWN	18_1_24	Cultivated
280	Local	K130649	UNK	18_1_25	Cultivated
281	Lokal	K130650	UNK	18_1_26	Cultivated
282	Local	K130652	UNK	18_1_27	Cultivated
513	MYS-PYJ-2007-55	K131569	UNK	18_1_28	Cultivated
283	KJA17	K136410	CHN	18_1_29	Cultivated
284	K163475	K163475	CHN	18_1_30	Cultivated
514	CN 72	K165772	MMR	18_1_31	Cultivated
515	VC 6141-54	K165773	MMR	18_1_32	Cultivated
516	VC 6368-46-40	K165775	MMR	18_1_33	Cultivated

517	VC 6173-B-10	K165776	MMR	18_1_34	Cultivated
518	VC 6173C	K165781	MMR	18_1_35	Cultivated
519	VC 12-3-4A	K165782	MMR	18_1_36	Cultivated
520	CHN-2010-18	K166126	CHN	18_1_37	Cultivated
521	CHN-2010-19	K166127	CHN	18_1_38	Cultivated
285	CHN-2010-20	K166128	CHN	18_1_39	Cultivated
523	YV 542	K169726	SGP	18_1_40	Cultivated
524	VIR 6559	K173279	PRK	18_1_41	Cultivated
525	VIR 6560	K173280	PRK	18_1_42	Cultivated
526	CHN-북방농업연구소-2011-11	K175297	CHN	18_1_43	Cultivated
527	CHN-북방농업연구소-2011-12	K175298	CHN	18_1_44	Cultivated
528	CHN-PMW-2011-4	K175546	CHN	18_1_45	Cultivated
530	KHM-LWJ-2011-9	K176372	KHM	18_1_46	Cultivated
531	KHM-LWJ-2011-10	K176373	KOR	18_1_47	Cultivated
532	KHM-LWJ-2011-12	K176375	KHM	18_1_48	Cultivated
287	UZB-Shahrisabz-2011-21	K187555	UZB	18_1_49	Cultivated
533	UZB-Shahrisabz-2011-33	K187567	UZB	18_1_50	Cultivated
534	UZB-Khazarbag-2011-60	K187592	UZB	18_1_51	Cultivated
288	KHM-농사연-2012-4	K191035	KHM	18_1_52	Cultivated
621	Arta Moseng		IDN	18_1_53	Cultivated
183	Si Walik		IDN	18_1_54	Cultivated
623	Lok Madura		IDN	18_1_55	Cultivated
624	Arta Item		IDN	18_1_56	Cultivated
625	Arta ijo		IDN	18_1_57	Cultivated
626	Manyar		IDN	18_1_58	Cultivated
627	Bhakti		IDN	18_1_59	Cultivated
628	No 129		IDN	18_1_60	Cultivated
629	Nuri		IDN	18_2_1	Cultivated
630	Kenari		IDN	18_2_2	Cultivated
631	Betet		IDN	18_2_3	Cultivated
632	Gelatik		IDN	18_2_4	Cultivated
633	Parkit		IDN	18_2_5	Cultivated
634	Merpati		IDN	18_2_6	Cultivated
636	Camar		IDN	18_2_7	Cultivated
637	Merak		IDN	18_2_8	Cultivated
638	Calon haji 1a		IDN	18_2_9	Cultivated
639	Samsek-a		IDN	18_2_10	Cultivated

642	Calon haji ongko	IDN	18_2_12	Cultivated
643	Plastik	IDN	18_2_13	Cultivated
644	Lok Ps Jailolo	IDN	18_2_14	Cultivated
646	FOrewehal	IDN	18_2_15	Cultivated
647	Fore Lotu	IDN	18_2_16	Cultivated
648	Bue bura	IDN	18_2_17	Cultivated
649	Lok Kab Borong A	IDN	18_2_18	Cultivated
650	Lok Kota Kumbah A	IDN	18_2_19	Cultivated
651	Nilon	IDN	18_2_20	Cultivated
652	Lok Mutoha M-1	IDN	18_2_21	Cultivated
653	Lok Abuki	IDN	18_2_22	Cultivated
654	Lok Majenang A	IDN	18_2_23	Cultivated
655	Lok Pangalengan	IDN	18_2_24	Cultivated
656	Lok Tarogong	IDN	18_2_25	Cultivated
657	Mentik hitam	IDN	18_2_26	Cultivated
658	PB-1 (benggolo Puti)	IDN	18_2_27	Cultivated
659	Lok Kudus	IDN	18_2_28	Cultivated
660	Lok Ngawi	IDN	18_2_29	Cultivated
661	Lok Pemeungpeuk	IDN	18_2_30	Cultivated
662	Lok Bungbulang	IDN	18_2_31	Cultivated
663	Lok Kupang	IDN	18_2_32	Cultivated
664	Butek Surade	IDN	18_2_33	Cultivated
665	Fore Belu	IDN	18_2_34	Cultivated
666	Perkutut	IDN	18_2_35	Cultivated
667	Tecer Hitam	IDN	18_2_36	Cultivated
668	Lok Sampang 1	IDN	18_2_37	Cultivated
669	Lima-1	IDN	18_2_38	Cultivated
670	Lok Jerowaru	IDN	18_2_39	Cultivated
247	Bohabe yellow mongo	PHL	18_2_40	Cultivated
248	Zilola	UZB	18_2_41	Cultivated
249	Durdona	UZB	18_2_42	Cultivated
250	Turon	UZB	18_2_43	Cultivated
152	Kh.50 hari (L.insana)	IDN	18_2_44	Cultivated
153	Lokal Landa Baru	IDN	18_2_45	Cultivated
676	Lokal Mutaha K2	IDN	18_2_46	Cultivated
155	Lokal Garut M	IDN	18_2_47	Cultivated
157	Mentik Coklat	IDN	18_2_48	Cultivated

681	Lok. Pasar Welahan	IDN	18_2_49	Cultivated
682	Lokal Sidamulih	IDN	18_2_50	Cultivated
160	Kuyak	IDN	18_2_51	Cultivated
684	RR-2	IDN	18_2_52	Cultivated
687	Lok. Garut	IDN	18_2_53	Cultivated
165	Lok Puda 1	IDN	18_2_54	Cultivated
166	Lok Galis	IDN	18_2_55	Cultivated
168	Arta Koneng	IDN	18_2_56	Cultivated
170	Lok. Pasanggaran	IDN	18_2_57	Cultivated
171	Lok. Jonggat	IDN	18_2_58	Cultivated
695	Fore Modok	IDN	18_2_59	Cultivated
173	ArthaZatim	IDN	18_2_60	Cultivated
697	Fue Nutu	IDN	19_1_1	Cultivated
698	Foe Nutu	IDN	19_1_2	Cultivated
700	Kambe Morowisa	IDN	19_1_3	Cultivated
701	Kambe Kulita Kokana	IDN	19_1_4	Cultivated
179	Kabe Mor	IDN	19_1_5	Cultivated
704	Lok. Bajawa Ngada	IDN	19_1_6	Cultivated
705	Lok. Ps. Embai Golewa Barat	IDN	19_1_7	Cultivated
201	JP229109	KOR	19_1_8	Cultivated
202	JP229144	CHN	19_1_9	Cultivated
203	JP229145	CHN	19_1_10	Cultivated
204	JP229215	CHN	19_1_11	Cultivated
205	JP229216	CHN	19_1_12	Cultivated
206	JP99049	TWN	19_1_13	Cultivated
207	JP231194	PHL	19_1_14	Cultivated
211	JP229233	IDN	19_1_15	Cultivated
212	JP78939	VNM	19_1_16	Cultivated
213	JP229096	THA	19_1_17	Cultivated
214	JP229097	THA	19_1_18	Cultivated
215	JP229098	THA	19_1_19	Cultivated
216	JP229099	THA	19_1_20	Cultivated
217	JP231216	THA	19_1_21	Cultivated
218	JP231220	THA	19_1_22	Cultivated
219	JP229130	BGD	19_1_23	Cultivated
222	JP229163	IND	19_1_24	Cultivated
228	JP229177	IND	19_1_25	Cultivated

229	JP229175		IND	19_1_26	Cultivated
231	JP229211		IND	19_1_27	Cultivated
233	JP229190		IND	19_1_28	Cultivated
235	JP231223		IND	19_1_29	Cultivated
241	JP103138-1		PAK	19_1_30	Cultivated
242	JP103138-2		PAK	19_1_31	Cultivated
243	JP99066		PAK	19_1_32	Cultivated
244	JP229241		AFG	19_1_33	Cultivated
246	JP31324		AFG	19_1_34	Cultivated
251	JP229254		IRN	19_1_35	Cultivated
252	JP229257		IRN	19_1_36	Cultivated
253	JP229263		IRN	19_1_37	Cultivated
254	JP31331		IRN	19_1_38	Cultivated
255	CN900001		THA	19_1_39	Cultivated
256	CN900002		THA	19_1_40	Cultivated
257	CN900004		THA	19_1_41	Cultivated
258	CN900005		THA	19_1_42	Cultivated
259	CN900007		THA	19_1_43	Cultivated
260	CN900008		THA	19_1_44	Cultivated
263	CN900014		THA	19_1_45	Cultivated
264	CN900015		THA	19_1_46	Cultivated
752	JP229290		AUS	19_1_47	Wild
301	충북재래 2 호	26018	KOR	19_1_48	Cultivated
303	경기재래 17 호	26062	KOR	19_1_49	Cultivated
304	수원 2 호	26252	KOR	19_1_50	Cultivated
307	경북예천-1985-2800	102800	KOR	19_1_51	Cultivated
308	경남창녕-1985-2855	102855	KOR	19_1_52	Cultivated
310	경기양평-1985-3400	103400	KOR	19_1_53	Cultivated
311	경기양평-1985-3418	103418	KOR	19_1_54	Cultivated
313	전북김제-1985-3835	103835	KOR	19_1_55	Cultivated
316	경기평택-1985-4183	104183	KOR	19_1_56	Cultivated
317	대만수집	104747	TWN	19_1_57	Cultivated
319	경남고성-1985-5626	105626	KOR	19_1_58	Cultivated
320	반월수집	105639	KOR	19_1_59	Cultivated
322	고림녹두	111041	KOR	19_1_60	Cultivated
323	경기화성-1985-12822	112822	KOR	19_2_1	Cultivated
326	Siraha Local-2	136322	NPL	19_2_2	Cultivated

327	경북성주-1986-24373	138114	KOR	19_2_3	Cultivated
328	VC3566-B-2-1-3	145301	UNK	19_2_4	Cultivated
329	V1153	154078	IRN	19_2_5	Cultivated
330	V3686	154080	USA	19_2_6	Cultivated
331	Yellowgram	154085	UNK	19_2_7	Cultivated
332	Pakistan	154087	PAK	19_2_8	Cultivated
333	전북정읍-1989-5463	162743	KOR	19_2_9	Cultivated
334	충북보은-1989-5499	162779	KOR	19_2_10	Cultivated
335	전북임실-1989-5600	162880	KOR	19_2_11	Cultivated
336	Vo1301	163175	CHN	19_2_12	Cultivated
338	Vo3484	163234	PAK	19_2_13	Cultivated
340	Vo5551	163280	IRN	19_2_14	Cultivated
342	MBLS90-19	168064	KOR	19_2_15	Cultivated
343	MBLS90-30	168075	KOR	19_2_16	Cultivated
344	경북청송-1992-2658	175816	KOR	19_2_17	Cultivated
345	경북상주-1993-2473	180833	KOR	19_2_18	Cultivated
346	경북금릉-1993-2475	180835	KOR	19_2_19	Cultivated
347	전북고창-1993-3516	181876	KOR	19_2_20	Cultivated
348	92 유자 466	182212	UNK	19_2_21	Cultivated
349	VC1089A	182225	TWN	19_2_22	Cultivated
350	VC2768B	182247	TWN	19_2_23	Cultivated
351	VC3523	182255	UNK	19_2_24	Cultivated
352	수원 19 호	182273	KOR	19_2_25	Cultivated
354	VC3890B	182296	TWN	19_2_26	Cultivated
355	교모 5	183235	KOR	19_2_27	Cultivated
356	도문-1	183263	UNK	19_2_28	Cultivated
357	92 예천수집	183264	KOR	19_2_29	Cultivated
358	Celera	183789	AUS	19_2_30	Cultivated
359	전남담양-1994-3231	185570	KOR	19_2_31	Cultivated
361	청녹두	185575	KOR	19_2_32	Cultivated
362	전북남원-1994-3237	185576	KOR	19_2_33	Cultivated
365	V01122B-G	189472	TKY	19_2_34	Cultivated
366	V01946A-Y	189516	PHL	19_2_35	Cultivated
368	V03538B-G	189552	KNY	19_2_36	Cultivated
369	V03720B-G	189553	USA	19_2_37	Cultivated
370	V03827A-G	189554	IRN	19_2_38	Cultivated
374	강원정선-1995-2895	191127	KOR	19_2_39	Cultivated

380	Original Tashkent 1978		199273	KOR	19_2_40	Cultivated
381	ACC6		201172	PHL	19_2_41	Cultivated
382	ACC11		201177	PHL	19_2_42	Cultivated
397	경남통영-2000-28		212101	KOR	19_2_43	Cultivated
399	CHN-LJR-2000-35		212108	CHN	19_2_44	Cultivated
400	장안녹두		216796	KOR	19_2_45	Cultivated
718	W116	W116		UNK	19_2_46	Wild
726	W147	W147		UNK	19_2_47	Wild
730	W162	W162		UNK	19_2_48	Wild
734	W169	W169		UNK	19_2_49	Wild
738	W176	W176		UNK	19_2_50	Wild
742	W190	W190		UNK	19_2_51	Wild
743	W191	W191		UNK	19_2_52	Wild
744	W192	W192		UNK	19_2_53	Wild
745	W203	W203		UNK	19_2_54	Wild

Phenotypic Data Collection

The traits scored were: days to flowering (DF), which is the number of days from planting when 50% of the plants produced open flowers; days to maturity (DM), which is the time when 50% of the plants produce mature pods; and the number of pods that could be harvested from 5 plants every week starting from day 65 after planting. Additional traits were then calculated from those data: length of pod formation (DPF), which is the time required for a mature pod to develop from flower and measured by subtracting DM with DF; the weight of 100 seeds (SDWT100); cumulative total yield (CUMPOD), where the total number of pods harvested each week was calculated at the end of the experiment at day 121; and the average number of seeds in a pod (S_POD), which was calculated by dividing the total weight of a harvest with SDWT100 to estimate the number of seeds obtained in a harvest, and divide that number with the number of pods in that particular harvest.

The optimum harvest day was calculated by finding the time when each accession produce maximum yield, subtract the day with flowering day, and compare it to the time needed to develop mature pods. This was done because each accession had different flowering time and pod production period, so the number needs to be normalized to developmental stages to make it more comparable. It was found that on average, maximum harvest

occur at DPFx1.7. Total harvest at this time was measured for all accessions (C17) so that if breeders decide to harvest only once and ignore synchronicity, they can simply select for traits with the highest C17 harvest. The proportion of C17 to CUMPOD was also calculated (C17_TOT), as synchronous plants should produce very little additional pods after C17 and the value of the ratio should be close to 1. Smaller ratio means that flowering is asynchronous and the plant still produces a large number of pods after the optimum harvest time. Another measurement for synchronicity was also formulated, by measuring the length of time the plants could produce a large number of pods (PRODAY). Synchronous plants should have short PRODAY, as production should stops after peak harvest.

The remaining 5 plants were harvested only once, at the end of the experiment at day 114 for early flowering accessions, and day 121 for late flowering ones. The number of pods obtained at this time is considered as final yield (F_YLD), to emulate a scenario where farmers choose to delay harvest to accommodate non-synchronous varieties. To see if there are differences between weekly harvest and final yield, we also calculated the ratio between the two (F_YLD_TOT), by dividing final yield with CUMPOD. This number should represent the amount of yield loss that could happen if harvest is delayed, and whether weekly harvest actually stimulate additional pod formation.

Genotyping by Sequencing

The genotype data was collected using genotyping by sequencing approach (Elshire et al., 2011), where total genomic DNA was digested with Ape KI enzyme and ligated to sequencing adapters carrying unique barcodes for each accession. The resulting library was sequenced using Illumina HiSeq 2000, and the resulting sequence data were aligned to mungbean reference genome using BWA package (Li and Durbin, 2009). Sequence variation were identified from the alignment data using SAMTOOLS package (Li et al., 2009), and filtered to obtain single nucleotide polymorphism (SNP) with at least one sequencing depth in each accession. This resulted in the identification of 9309 SNPs that are distributed evenly in the genome. Eliminating variants in the unmapped scaffolds to focus on the 11 chromosomes reduced the SNP number to 7551. The number is still larger than the minimum of 5600 SNPs required if average LD block in mungbean is 86 kilobases.

Genome-Wide Association Analysis

Association analysis was performed using two software packages: Genome Association and Prediction Integrated Tool (GAPIT) (Lipka et al., 2012) and Trait Analysis by aSSociation, Evolution and Linkage (TASSEL) (Bradbury et al., 2007). GAPIT utilizes efficient mixed-model association (EMMA) association test in R environment and set to consider three principal components, while the association test in TASSEL was set to use compressed mixed linear model (MLM) with optimum level of compression and P3D variance component estimation and also population structure and kinship data as covariates.

RESULTS

Phenotypes of the Association Panel

Figure III-1 summarizes the range, distribution, and correlation among the 11 phenotypes scored in this study. In general, mungbean cultivars in the association panel produced more pods when harvested multiple times compared to just single harvest after 114 days of planting. Each plant could produce between 2 to 171 pods when harvested multiple times (CUMPOD), compared to 0-147 pods when harvested only once (F_YLD) after 114 days. The ratio between single harvest compared to multiple harvests (F_YLD_TOT) varied between 0 to 3.4, but the average is only 0.49. There are some indications that early harvesting induces additional flower and pod formation, which could explain the additional yield in multiple harvests. However, plants with zero ratios could also have a large number of fallen pods prior to harvesting or extensive pod shattering, so their actual yield appeared lower unless they are harvested early. Plants that mature later also tend to have higher ratios, since the final harvest time represented a relatively early stage of pod production in their life cycle.

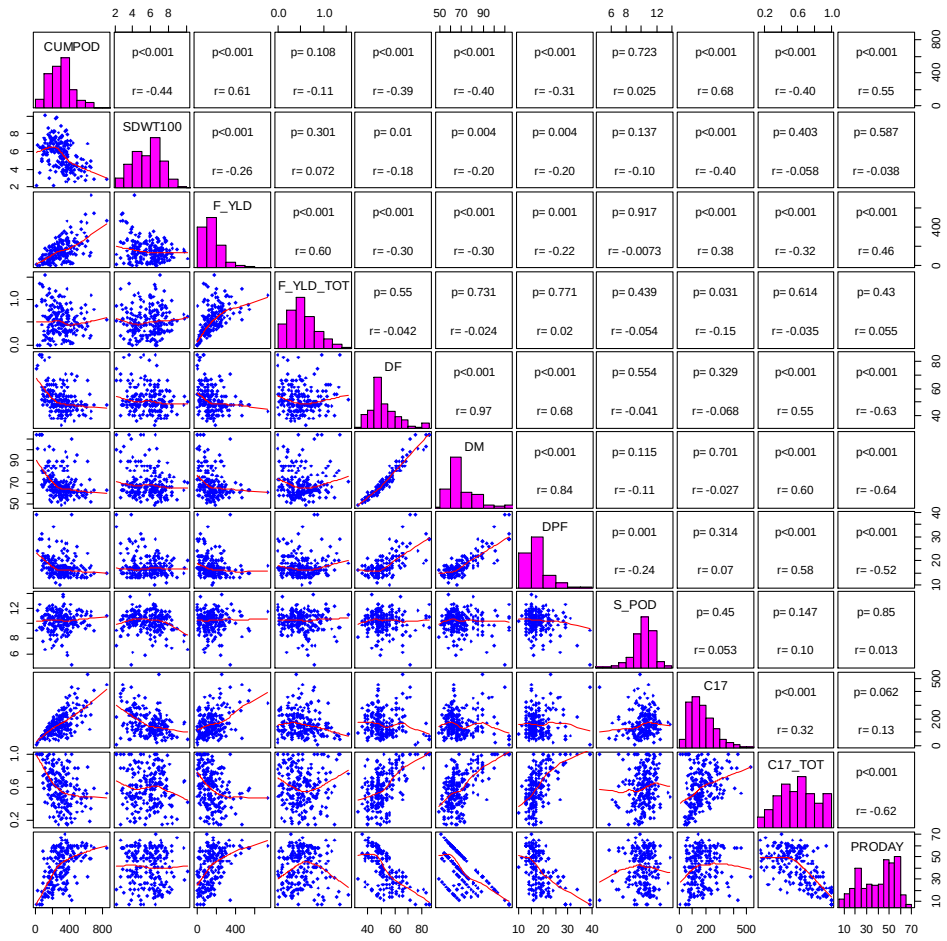


Figure III-1. Distribution and correlation among traits in the association panel. Histogram of the trait distribution is presented in the diagonal panels. Lower panels are matrices of scatter plots that show the correlations between traits, while the upper panels show the significance and absolute value of the correlations (r).

Weight of 100 seeds (SDWT100) varies from 2.2 to 10.1 gram, with a median of 5.6 gram. Seed weight has weak but significant negative correlations to final yield in single and multiple harvest treatment. Average number of seeds in a pod (S_POD) ranges from 4.5 to 13.7 seeds per pod, with a median value of 10.4. There is no significant correlation between the number of seeds in a pod and seed weight, so accessions with long pods that also bear heavy seeds exist in the association panel.

Flowering time (DF) has a very strong correlation with maturity time (DM), as well as the time it takes to develop pods following fertilization (DPF). After planting, more than half of the accessions flower before 50 days and the pods will form and mature 17 days later. Some accessions have very long flowering and maturity time that by day 121 no flower or mature pods can be obtained from them. These accessions are not suitable for planting in Korea as ambient temperature would have started to drop to temperatures that are not conducive to growth and reproduction.

Another trait being measured in this study is synchronicity. In most cases, synchronicity was not found to be the result of apical meristem indeterminacy, since vegetative growth usually stops when the first flower emerges. In some cases, non-synchronicity is caused by prolonged productivity of peduncles, where new flowers keep emerging even after the first batches of pods are matured (Figure III-2). New peduncles can also

emerge after the first batch of pods is harvested. In this study, two approaches used to measure synchronicity; one is by measuring the length of time the accessions can maintain economically viable pod production after flowering (PRODAY), and the other is by measuring the proportion of peak production relative to cumulative yield obtained by the end of the study (C17_TOT). In this study, the threshold for economical harvesting was arbitrarily set at 3 pods per plant, and the time when the productivity of an accession dropped below this and never recovered was defined as the beginning of non-productive time (figure III-3). PRODAY varies from 7 days to 70 days in the panel, although the very short ones are usually caused by late flowering time. Nevertheless, the shorter the productive time, the closer it is to the definition of ideal synchronous crop.

Figure III-2. Prolonged productivity of peduncles is one of the reasons of non-synchronous pod maturity in mungbean. New flowers are still formed even when earlier pods from the same peduncles matured already.

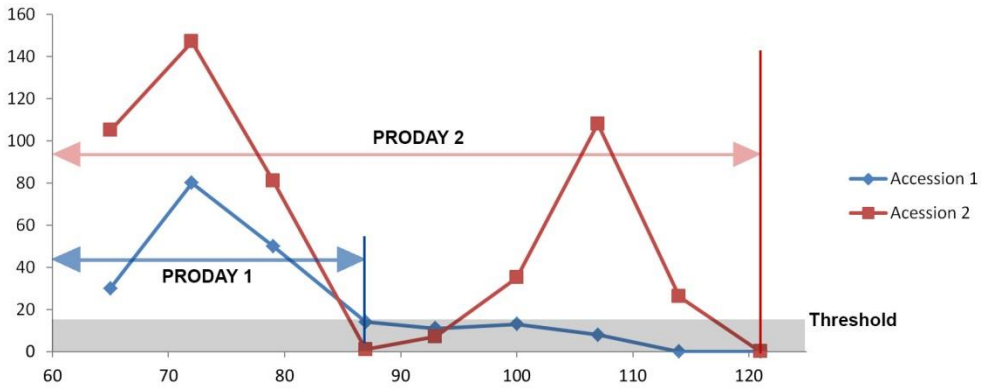


Figure III-3. Defining synchronicity in mungbean by the length of productive days (PRODAY) until weekly harvests hit non-productive threshold, which is defined as the period when the pod count is consistently less than 3 per plant when harvested each week. Synchronous plants should have shorter PRODAY compared to non-synchronous plants, so accession 1 here has a higher degree of synchronicity than accession 2. The y-axis is the number of pods that can be harvested each week from 5 plants, while the x-axis denotes the number of days since planting.

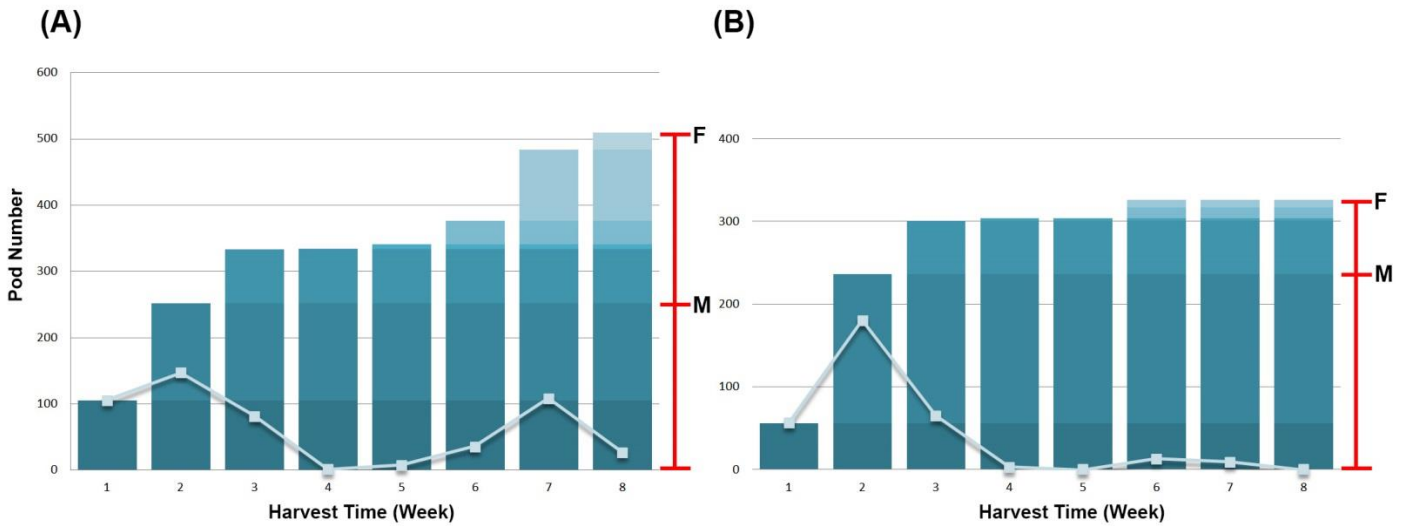


Figure III-4. Synchronicity was also defined as the ratio between cumulative yield at optimum harvest time (in this case at week 2 for both accession A and B) and the final yield at the end of the study. The ratio between M and F is higher in accession B, since it produced lesser number of pods after its peak harvest time. A highly synchronous accession should have an M/F ratio of close to 1 as it stops producing pods after the peak harvest.

Aside from shorter productive time, synchronous plants should also produce no more pods after harvest, so the ratio between harvested pods at optimum harvest time and after 114 days should be closer to 1. On average, most accession produce peak harvest at 1.7 times DPF (pod-forming days), and cumulative harvest produced until this day was termed as C17. So a low ratio between harvests at this time to final cumulative harvest indicates non-synchronous behavior (Figure III-4). As with PRODAY, this parameter is confounded by late flowering time as accessions that flowered late would not have a chance to produce more pods after 1.7 DPF due to time constraint of this study. However, for its intent and purposes, such accessions could also be termed as synchronous accessions as further harvesting will not be possible due to the low temperature commonly found after October in Korea. Breeders can also ignore the ratio (thus regarding subsequent yield as unimportant) and simply select for accessions with the highest C17 yield, as one accession can produce as much as 105 pods per plant in this period.

GWAS

Originally, GWAS was carried out in all 232 accessions since both phenotype and genotype data were available for all accessions. However, inclusion of wild accessions seems to introduce a lot of spurious, false-

positive associations (Figure III-5). In some traits, the false positives even pass the significance thresholds, and they could not be controlled by using either population structure q-values or PCA as covariates, or kinship matrix as random effects. Since the effect of population structure from wild accessions was too strong, they are subsequently removed from GWAS to include only 222 of the cultivated accessions. No obvious patterns of spurious associations were observed in subsequent analyses.

GAPIT identified 79 markers that are associated with various traits with a p-value of <0.0001, while the mixed linear model in TASSEL identified 77 markers at the same p-value threshold (Figure III-6, Figure III-7, and Table III-2). The largest number of markers that show significant association with a trait is 49 markers found to be associated with F_YLD_TOT trait by GAPIT. Some of those markers are located very close to each other, at less than the average distance of LD decay in cultivated mungbean (approximately 86 kb), so it is possible that they pinpoint to a common mutation in that LD interval simply because they are located in the same LD block. For traits that measure synchronicity, PRODAY had no significant association with any markers at p<0.0001, while C17_TOT produced one significant association with a marker at chromosome Vr07 at base 48142032 according to GAPIT, although TASSEL calculated less significance level at that locus.

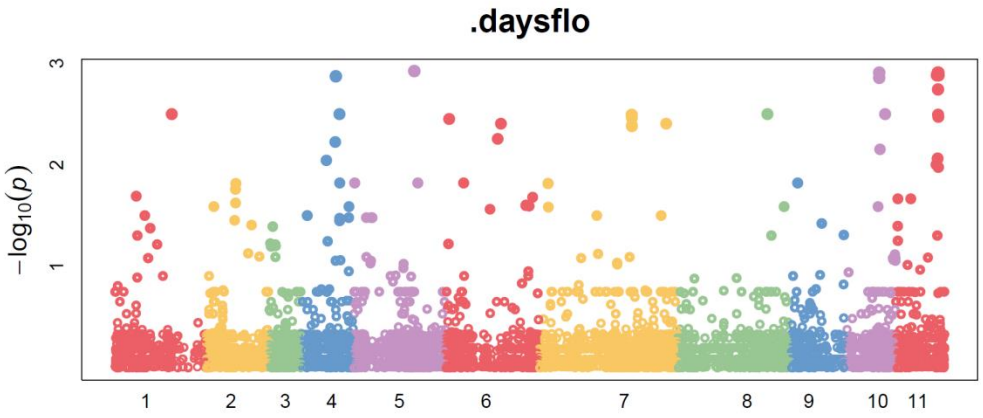


Figure III-5. Manhattan plot of p-values from GWAS when wild accessions were included in the analysis. Note the presence of numerous dots forming a horizontal line close to $-\log_{10}(p)=1$, which indicates the presence of population structure. Those dots passed the significance threshold in some traits, and they cannot be removed even after population structure, PCA, and kinship were accounted for.

Table III-2. Loci showing significant associations with the measured traits (p<0.0001). Each locus is defined as the interval between a significant marker and its flanking markers. Some loci at the end of the table are significant for multiple traits. Traits followed by parentheses indicate significance in one of the algorithms only, while those without are significant in both. MAF is minor allele frequency, while R-square estimates the phenotype variation caused by the marker. Effect size estimates the effect on phenotype when the minor allele is present in the accession.

Chromosome	Locus Start	Locus End	Trait	p-value	MAF	R-square	Effect Size	Genes in Locus
Vr04	19744338	19763241	C17	7.77E-05	0.0185	0.090	-121.89	2
Vr07	36267994	36420654	C17 (TASSEL)	6.15E-05	0.0139	0.076	110.76	6
Vr07	48141848	48142082	C17_TOT (GAPIT)	5.66E-05	0.0787	0.068	0.18	1
Vr04	8368961	8368993	DF	8.19E-05	0.0046	0.075	-18.99	0
Vr04	13695949	13727978	DF	3.45E-05	0.0324	0.084	-10.17	5
Vr11	16370812	16386388	DF	8.19E-05	0.0046	0.075	18.99	2
Vr11	16430954	16496088	DF	8.19E-05	0.0046	0.075	18.99	6
Vr11	16522593	16620002	DF	8.19E-05	0.0046	0.075	-18.99	9
Vr11	16755523	16867008	DF	8.19E-05	0.0046	0.075	18.99	9
Vr11	16906945	16977433	DF	8.19E-05	0.0046	0.075	-18.99	5
Vr11	17003409	17056019	DF	8.19E-05	0.0046	0.075	-18.99	6
Vr07	51691455	51984231	DM (GAPIT)	1.91E-05	0.0139	0.060	-19.33	9
Vr02	21274950	21359567	DPF	9.95E-05	0.0231	0.073	-5.97	6
Vr06	33720789	33895879	DPF	1.22E-05	0.0139	0.094	7.31	17

Vr08	23354046	23705559	DPF	2.12E-05	0.0139	0.088	7.22	5
Vr02	6590385	6722940	DPF (TASSEL)	6.96E-05	0.0093	0.077	-5.09	4
Vr02	16757125	16910016	DPF (TASSEL)	9.47E-05	0.0185	0.074	-3.90	5
Vr05	15661112	15698117	S_POD	4.63E-07	0.0162	0.150	2.11	6
Vr07	24541951	24725397	F_YLD	1.04E-06	0.0045	0.096	280.34	10
Vr09	9590146	9780159	F_YLD	8.36E-05	0.0294	0.061	85.15	2
Vr07	9054437	9581944	F_YLD (TASSEL)	8.23E-05	0.1364	0.089	-0.18	26
Vr04	10920642	11483322	F_YLD_TOT	3.04E-08	0.0091	0.007	-0.88	30
Vr04	12559278	13046958	F_YLD_TOT	7.50E-05	0.0227	0.004	-0.41	16
Vr04	14473574	14606067	F_YLD_TOT	7.31E-07	0.0091	0.005	0.77	8
Vr04	14606172	14912361	F_YLD_TOT	7.31E-07	0.0091	0.001	0.77	20
Vr04	18206296	18256075	F_YLD_TOT	2.72E-06	0.0182	0.008	0.54	4
Vr04	18423582	18508177	F_YLD_TOT	7.98E-08	0.0091	0.001	0.84	10
Vr06	955275	955328	F_YLD_TOT	6.67E-12	0.0045	0.003	1.48	0
Vr06	22114556	22271065	F_YLD_TOT	9.04E-08	0.0045	0.000	1.09	6
Vr07	24143717	24541951	F_YLD_TOT	1.58E-05	0.0091	0.003	0.62	14
Vr07	37226320	37236867	F_YLD_TOT	6.67E-12	0.0045	0.007	-1.48	1
Vr07	49293268	49435202	F_YLD_TOT	8.31E-06	0.0114	0.007	0.66	13
Vr07	51010945	51011151	F_YLD_TOT	9.04E-08	0.0045	0.001	1.09	0
Vr07	52191658	52289499	F_YLD_TOT	9.04E-08	0.0045	0.005	-1.09	4
Vr07	55474838	55492391	F_YLD_TOT	9.04E-08	0.0045	0.009	-1.09	2
Vr08	13328607	13522025	F_YLD_TOT	7.50E-05	0.0227	0.006	0.41	11
Vr08	23354046	23705559	F_YLD_TOT	9.28E-07	0.0136	0.018	0.66	5
Vr08	24018207	24683019	F_YLD_TOT	5.22E-05	0.0182	0.000	-0.45	22

Vr08	24685867	25511119	F_YLD_TOT	3.71E-06	0.0227	0.008	0.50	35
Vr08	32308863	32426640	F_YLD_TOT	7.31E-07	0.0091	0.004	-0.77	8
Vr09	9837612	9837662	F_YLD_TOT	3.04E-08	0.0091	0.007	0.88	0
Vr10	7306049	7557433	F_YLD_TOT	9.38E-08	0.0091	0.001	-0.84	13
Vr10	20089638	20234877	F_YLD_TOT	1.70E-07	0.0182	0.000	0.59	11
Vr11	15542755	15715699	F_YLD_TOT	6.67E-12	0.0045	0.006	-1.48	8
Vr04	19198263	19289531	F_YLD_TOT (TASSEL)	7.80E-06	0.1145	0.014	-0.05	6
Vr02	9228764	9501886	SDWT100	8.07E-05	0.0579	0.094	0.79	16
Vr06	931489	955327	SDWT100	6.71E-05	0.1111	0.096	-0.72	1
Vr01	20697235	20724712	SDWT100 (GAPIT)	3.56E-05	0.0139	0.036	1.88	3
Vr06	955328	1013977	SDWT100 (GAPIT)	8.69E-05	0.1273	0.033	-0.67	9
Vr04	18873105	18912117	SDWT100 (TASSEL)	6.89E-05	0.1296	0.095	-0.65	6
Vr10	4563374	5412928	DF (TASSEL)	8.98E-05	0.0440	0.091	7.32	29
			DM (GAPIT)	7.99E-05	0.0440	0.051	9.73	
			DPF (GAPIT)	4.42E-05	0.0440	0.056	3.20	
Vr05	16811658	16987571	C17 (TASSEL)	8.01E-05	0.0139	0.074	116.84	17
Vr09	20982995	20984803	CUMPOD	7.82E-05	0.0833	0.090	-95.15	0
Vr02	24139639	24341716	DF	9.86E-06	0.3542	0.069	4.73	15
			DM (GAPIT)	8.71E-05	0.3542	0.050	5.31	

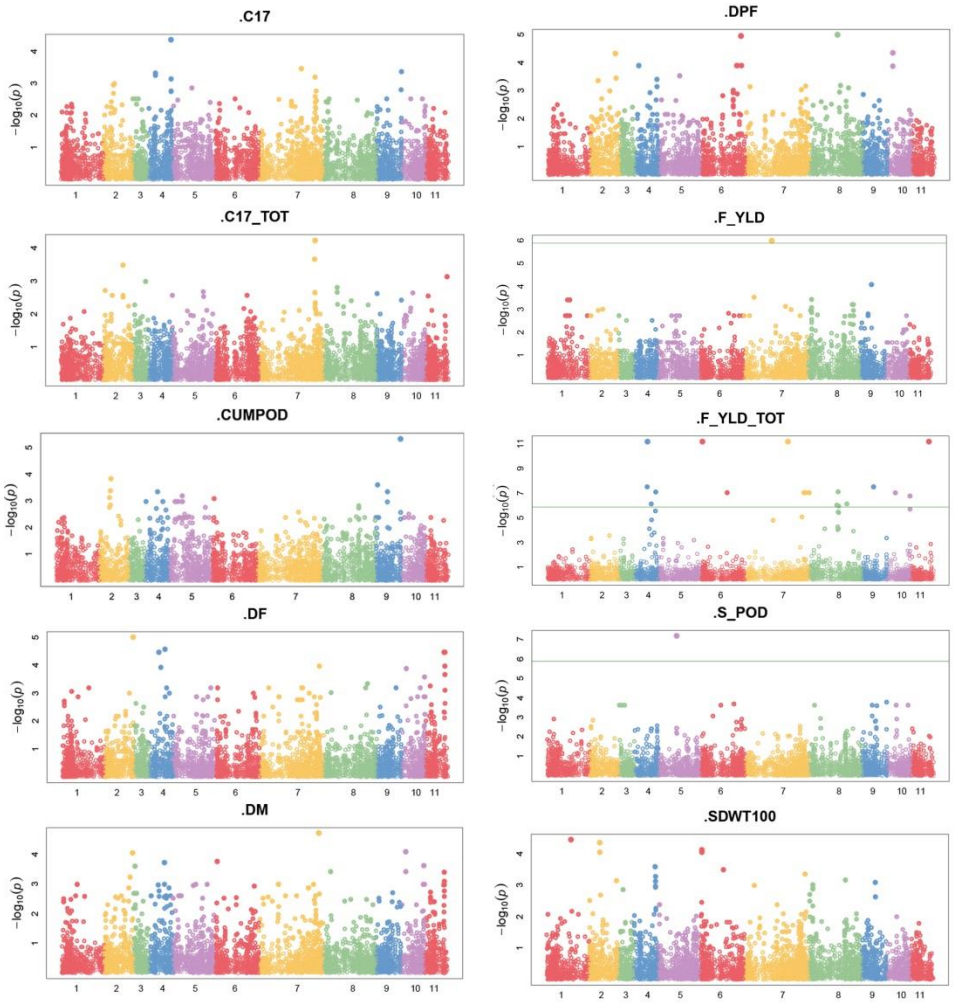


Figure III-6. Manhattan plots of significant marker-trait association produced by GAPIT. The y-axis is $-\log_{10}$ of p-values of the markers' significance, while x-axis is the chromosome number and base pair positions of the markers.

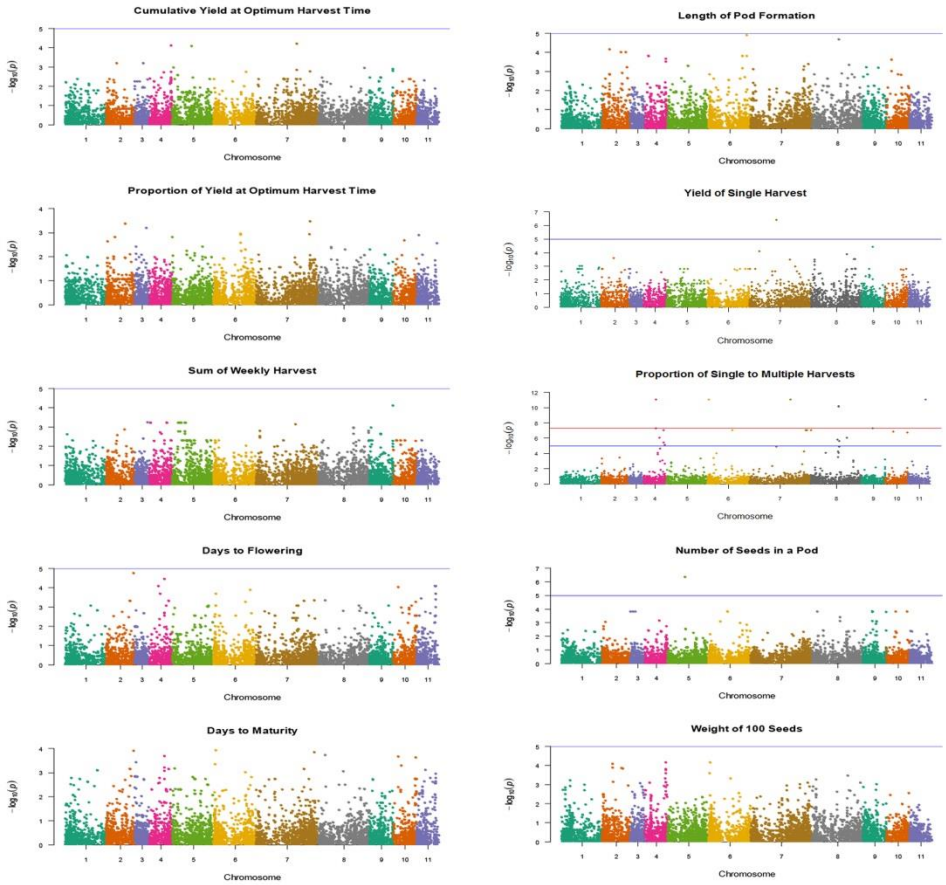


Figure III-7. Manhattan plots of significant marker-trait association produced by TASSEL. The y-axis is $-\log_{10}$ of p-values of the markers' significance, while x-axis is the chromosome number and base pair positions of the markers.

The predicted effect of minor alleles located in significant loci on the phenotype can also be seen in Table III-2. Some of the allelic effects are exactly the same for some neighboring alleles, indicating that they are originated from the same individuals with distinct phenotypes compared to the rest of the population. Significant association that is originated from very few individuals (as indicated by very low minor allele frequency) is generally more prone to error, especially to imprecise phenotyping or non-obvious population structure. Therefore, extra scrutiny and further testing will be needed to confirm function of those types of markers. The r-square values of the significant markers are generally small, the highest was observed only on a marker for S_POD at 0.15.

Candidate Gene Identification

Assuming that the mutation that changes the phenotype is located in the interval between the significant markers and their adjacent markers, all of the coding regions located upstream and downstream of the markers were identified (Table III-2). To understand the genetic basis and cellular processes that could be responsible for the trait formation, comparative genomics approach was explored as related plants should largely follow the same regulatory pattern. The protein sequences of the genes surrounding the markers were compared to soybean proteins using the BLAST tool in

soykb database (Joshi et al., 2012). The characteristics of the genes that showed the highest similarity to the mungbean genes were explored further to deduce if the genes could explain emergence of the trait being investigated.

One of the notable results obtained using this approach is the soybean gene most similar to nearest gene to CUMPOD locus at chromosome Vr09 (Vradi09g09940) is a gene located at chromosome Gm07 between base 6906819 and 6907736 (Glyma07g08310.1). This gene is located in a QTL interval that is involved in multiple traits like yield, pod, and insect. The gene ontology annotation for this gene is response to stress (GO:0006950). This information seems to support the mungbean GWAS result that this locus is influencing cumulative pod production.

Another potentially interesting SNP is located at chromosome Vr02 at base 9,377,515 and was identified as significant by both TASSEL and GAPIT for SDWT100 trait. The SNP is located inside Vradi02g07660 gene and this gene is most similar to Glyma12g10390.1 gene in soybean, which is annotated as a disease resistance protein. The protein has no GO annotation yet, but its domain is classified as similar to nucleoporin and dirigent protein. RNAseq data from this gene showed that its expression is exclusive to seeds and green pods, and peaked at 25 days after flowering. This makes it a fitting candidate for a gene that can cause a difference in

seed weight. The fact that its function is currently unknown makes it an attractive target for follow up study since it is likely to perform a novel function.

Multiple SNPs were also flagged as significant for flowering time by both GAPIT and TASSEL at the upstream region of Vradi11g11860 gene at chromosome Vr11. Promoters and enhancers are often found in upstream region, and they regulate the timing and strength of gene expression. The gene shows a high similarity to Glyma19g05420, a gene with no annotation but contain PHD-finger and Zinc finger domains. Its GO annotation encompasses many functions, including regulation of transcription, histone acetylation, apoptosis, and flower development. It is possible that this gene is involved in regulation of transcription or methylation, although unfortunately no expression data is available for this gene in soybean.

The genes that co-locate with significant markers are not always annotated with functions or expression pattern that shows obvious correlation with the trait that they are associated with. One example is a SNP at Vr11 chromosome which intersects with Vradi11g11520 gene and was detected as significantly associated with flowering time. The most similar soybean gene, Glyma20g01400, has neither functional nor GO annotation. The protein contains a GRAM domain, which indicates that it is a membrane-bound protein, and its RNA expression pattern is confined to

root tips along with low expression in flower. It is expected for genes that regulate flowering to be expressed in meristem or leaves, and not in the flower itself nor in distant tissues like root tips, so it is difficult to infer the function that such gene could provide in regulating the timing of flowering based on such data.

DISCUSSION

Although GWAS had identified several loci that showed strong association with agronomic traits in this mungbean population, it is obvious that further investigations will need to be carried out to pinpoint the exact genes that explain the variation in phenotypes. For a start, although the number of markers used in this study probably have covered most of the LD blocks in the population, their non-uniform distribution in the chromosomes means that some LD blocks might not be represented in the association study. Therefore, it is prudent to assume that some of the causal genes were probably not captured in the association analysis if they are located in those non-represented LD blocks. Some significant markers are also located far from their neighboring markers and the interval between the flanking markers contains numerous genes, where each of those genes could also be the candidates for the causal gene. Similarly, for significant markers that are located in a large LD block, the association may actually extend to the whole LD block instead of nucleotides in close proximity of the marker. Consequently, even when the significant marker is located inside a gene, that gene is not necessarily the causal gene and the actual gene is located somewhere else inside that large LD block. More genes should be considered for candidate gene search under such circumstances to account for those uncertainties.

Some traits are also very complex and could arise from numerous confounding factors. For example, the final harvest for plants that are only harvested once after 114 days (F_YLD) is influenced by many factors. The presence of shattering gene will drastically reduce F_YLD. Genes that weaken the attachment of mature pods to the pedicle will also promote yield loss and reduce F_YLD. On the other hand, genes that promote late-flowering may result in high F_YLD as late-flowering accessions suffer no yield loss and the plant is practically in its peak productive time at day 114. Consequently, F_YLD data actually represented many different gene actions, so candidate gene selections based on existing gene annotation will need to take those factors into consideration. Alternatively, each potential confounding factor can be regarded as component traits and scored individually in future GWAS. For example, seed shattering and pod abscission can be phenotyped separately. This should reduce the complexity of the measured traits and assist in candidate gene identification.

The above examples show that dissecting the molecular mechanism of trait formation remains a complex process, even in this era when well-annotated reference genomes are available. Further experiments like fine mapping or gene knock outs are still necessary to pinpoint the exact nucleotides that cause the phenotypic variations. However, for breeding purposes the marker interval produced from this study should be tight enough for marker assisted selection. Provided that the significant markers

show a large effect on the trait, it is possible to apply GWAS result directly into breeding program by using high-performing accessions carrying the significant markers as parents in marker-assisted selection. Some groups have also started exploring the possibility of incorporating markers with smaller effects into the models used in genomic selection (Spindel et al., 2015). In genomic selections, the contributions of numerous markers to a phenotypic trait in a training population are calculated to make an equation that can be used to predict the phenotype of an individual with known genotypes. Genomic selection is sometimes criticized as a black box method, since no knowledge in the mechanism of gene action is necessary to apply it in plant breeding, and some argue that incorporating loci with known effects into the prediction model will increase the accuracy of phenotypic prediction (Nakaya and Isobe, 2012).

Another aspect that became apparent following this study is that the current algorithms used in GWAS still cannot cope well with strong population structure. Wild accessions were also included in this study, but their inclusion in the association analysis created patterns that indicate the presence of spurious association that could not be eliminated even when population structure and kinship data were used as covariates. Eliminating the wild accessions produced cleaner manhattan plots with no obvious spurious association. This means that although wild accessions can potentially provide a lot of useful alleles due to their broader genetic

diversity, those alleles cannot be easily mapped using GWAS in mungbean. A more traditional mapping using bi-parental crosses is probably more effective in mining the useful alleles carried by wild accessions. This also means that mapping domestication-related traits cannot be effectively accomplished using GWAS in mungbean.

In conclusion, GWAS in cultivated mungbean had identified numerous loci that are significantly associated with various traits. Although further confirmation studies will need to be performed to dissect the molecular mechanism that contributes to each trait, this study has reduced the pool of genes that need to be examined for more in-depth study. Candidate markers and accessions were also identified, which could be used in future mapping studies or breeding program.

REFERENCES

Bradbury, P.J., Zhang, Z., Kroon, D.E., Casstevens, T.M., Ramdoss, Y. and Buckler, E.S. (2007) TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**, 2633-2635.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S. and Mitchell, S.E. (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS one* **6**, e19379.

Joshi, T., Patil, K., Fitzpatrick, M.R., Franklin, L.D., Yao, Q., Cook, J.R., Wang, Z., Libault, M., Brechenmacher, L. and Valliyodan, B. (2012) Soybean Knowledge Base (SoyKB): a web resource for soybean translational genomics. *BMC genomics* **13**, 1.

Kang, Y.J., Kim, S.K., Kim, M.Y., Lestari, P., Kim, K.H., Ha, B.K., Jun, T.H., Hwang, W.J., Lee, T., Lee, J., Shim, S., Yoon, M.Y., Jang, Y.E., Han, K.S., Taeprayoon, P., Yoon, N., Somta, P., Tanya, P., Kim, K.S., Gwag, J.G., Moon, J.K., Lee, Y.H., Park, B.S., Bombarely, A., Doyle, J.J., Jackson, S.A., Schafleitner, R., Srinives, P., Varshney, R.K. and Lee, S.H. (2014) Genome sequence of mungbean and insights into evolution within *Vigna* species. *Nature communications* **5**, 5443.

Khattak, G., Haq, M., Ashraf, M., Tahir, G. and Marwat, E. (2001) Detection of epistasis, and estimation of additive and dominance components of genetic variation for synchrony in pod maturity in mungbean (*Vigna radiata* (L.) Wilczek). *Field Crops Research* **72**, 211-219.

Korte, A. and Farlow, A. (2013) The advantages and limitations of trait analysis with GWAS: a review. *Plant methods* **9**, 1.

Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079.

Lipka, A.E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P.J., Gore, M.A., Buckler, E.S. and Zhang, Z. (2012) GAPIT: genome association and prediction integrated tool. *Bioinformatics* **28**, 2397-2399.

Nakaya, A. and Isobe, S.N. (2012) Will genomic selection be a practical method for plant breeding? *Annals of botany* **110**, 1303-1316.

Spindel, J., Begum, H., Akdemir, D., Virk, P., Collard, B., Redoña, E., Atlin, G., Jannink, J.-L. and McCouch, S.R. (2015) Genomic selection and association mapping in rice (Oryza sativa): effect of trait genetic architecture, training population composition, marker number and

statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. *PLoS genetics* **11**, e1004982.

국문초록

아시아 지역의 재배 및 수요가 높은 작물인 녹두는, reference genome 의 해독 이후 유전체 연구가 활발히 진행 되고 있다. 그 중 alternative splicing (AS)에 대한 다양한 정보도 존재 하는데, 최소 37.9%의 AS가 녹두에 존재하며, AS isoform의 대부분은 낮은 copy로 존재하거나, 발현의 양이 낮다. 종과 관련된 보존된 부분 또한 드물다. 그리고 녹두와 팥 사이의 공유되는 유전자들 중 고정된 AS 는 2.8%로 드물며. 오직 16개의 공통적인 AS가 대두 유전자와 녹두 사이에서 확인 되었다.

전체 276 재배종 및 야생종 녹두를 GBS를 이용하여, 염기서열 다양성이 야생종에 비하여 재배종에서 30% 감소함을 확인 할 수 있었다. 긴 간격의 Linkage disequilibrium decay를 LD block이 야생종 보다 평균 4.6배 길어짐으로서 확인하였다. 재배종과 야생종은 phylogenetic 및 population structure 분석을 통해서도 확연하게 분리되었으며, 지리적인 기원지와 소그룹 사이의 상관관계를 알 수 있다. 녹두의 순화 과정에서의 positive selection이 진행되었으리라 가능성을 갖는 여러 loci들을 규명하였고, 이러한 간격들의 유전자들은 녹두의 생장 및 생식과 관련된 유전자들이 두드러졌다.

GBS를 통해 얻은 SNP data는 또한 개화시, 성숙기, 꼬투리 형성기, 종자 무

게, 꼬투리 당립수, 최고 수확기, 주당 수확 누적수, 최종 수확량 그리고 동시 등숙성과 같은 농업적 형질에 대해 genome-wide association study (GWAS)를 이용하였다. 전체 222 재배종과 최소 79개 마커를 이용한 두 가지 연관 조사 방법을 통해 p-value 0.0001 이하의 연관된 형질을 규명하였다. 유의미한 마커들과 교차하는 유전자들은 대두의 유전자 및 QTL과 동일성을 보이고, 이는 형질 형성의 역할을 설명할 수 있으며 차후 연구에의 의미 있는 후보로서 진행 될 수 있다. 이러한 데이터는 녹두 유전체 분석에 있어서 부모의 선발 및 유전적 지도 작성에 대한 의미를 갖는다.

주요어: 녹두 (*Vigna radiata*), alternative splicing. genotyping by sequencing (GBS) 순화, GWAS, 농업적 형질

학번: 2014-30834