THESIS FOR DEGREE OF MASTER OF SCIENCE

Comparative Sequence Analysis of Mungbean DNA Mismatch Repair Genes

BY

ANDARI RISLIAWATI

FEBRUARY, 2016

MAJOR IN CROP SCIENCE AND BIOTECHNOLOGY

DEPARTMENT OF PLANT SCIENCE

THE GRADUATE SCHOOL OF SEOUL NATIONAL UNIVERSITY

Comparative Sequence Analysis of Mungbean DNA Mismatch Repair Genes

UNDER THE DIRECTION OF DR. SUK-HA LEE
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF SEOUL NATIONAL UNIVERSITY

BY
ANDARI RISLIAWATI

MAJOR IN CROP SCIENCE AND BIOTECHNOLOGY
DEPARTMENT OF PLANT SCIENCE

NOVEMBER, 2015

APPROVED AS A QUALIFIED THESIS OF ANDARI RISLIAWATI
FOR THE DEGREE OF MASTER
BY THE COMMITTEE MEMBERS

FEBRUARY, 2016

CHAIRMAN _____
Tae-Jin Yang, Ph.D.

VICE-CHAIRMAN _____
Suk-Ha Lee, Ph.D.

MEMBER _____
Hak Soo Seo, Ph.D.

Comparative Sequence Analysis of Mungbean DNA Mismatch Repair Genes

ANDARI RISLIAWATI

ABSTRACT

Mungbean (*Vigna radiata* (L.) R. Wilczek) is a high-protein grain legume that could improve soil condition through nitrogen fixation. It is grown widely in southern Asia and also a promising drought-tolerant crop due to its adaptation to dry environment. However, world mungbean production has been stagnant and its breeding progress is hampered by low genetic diversity. Some efforts have been done to overcome this problem, such as germplasm exploration and induced-mutation. However, none gave any satisfactory result. This could be caused by strong activity of *MSH* genes which has been reported to preserve genomic integrity in other species. To explore more about *MSH* genes in mungbean, we sequenced the *MSH* genes of 12 mungbean germplasm from 8 countries and compared the sequence to 70 *MSH* genes sequence from 14 species of eudicots clade.

We identified the location of *MSH* paralogs that involved in DNA mismatch repair, i.e. *MSH3* in chromosome 8, *MSH2* in chromosome 7, *MSH6* in chromosome 3, and *MSH7* in chromosome 1. All five conserved domains exist in *MSH2*, *MSH3*, and *MSH6* paralogs, whereas *MSH7* lacks one domain. Compare to other species, mungbean lost the Walker A motif at *MSH2* and partial of HTH subdomain at *MSH3*. We also identified 3

synonymous SNPs, 4 non-synonymous SNPs, and 1 deletion among mungbean germplasm at neighbored-motifs of domain I and domain V at *MSH2*, *MSH3* and *MSH6*. Two non-synonymous SNPs at *MSH6* and one deletion at *MSH2* are predicted having different protein function compare to the mungbean reference. This prediction can be tested further through genome editing technology to support the mutagenesis experiment in creating breeding materials that high-acceptable to mutation exposure. Mungbean accessions carrying the SNPs can be evaluated as well for its responsiveness to mutation and based on that, SNP markers can be designed to screen appropriate germplasm carrying favorable allele. Therefore the identification of *MSH* genes in mungbean may contribute to the mungbean genetic potential improvement toward induced-mutation.

Keywords: Mungbean, *MSH*, Domain, Motifs, SNPs, Protein function prediction

Student number: 2014-22124

CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Mungbean (*Vigna radiata* (L.) Wilczek) is an important grain legume and grown widely in developing countries, particularly in Southern Asia countries (Shanmugasundaram et al., 2009a). Mungbean is not only cultivated for its high protein seed content but also can be utilized as fodder that contributes to soil fertility (Senaratne et al., 1995; Shanmugasundaram et al., 2009b). Compare to other crops, mungbean can adapt with severe environment such as water limitation. Therefore, mungbean is considered as one promising drought-tolerant crop in the future (Aslam et al., 2013).

Despite its importance, world mungbean production has been stagnant due to some agronomic shortage such as low yield and susceptibility to diseases and insects (Shanmugasundaram et al., 2009a). Plant breeding as a knowledge and art to improve the heritable genetic of a particular trait in a plant, is used to cope these shortages (Allard, 1999). Unfortunately mungbean germplasm lacks of diversity, which is necessary for successful plant breeding research (Lestari et al., 2014; Sangiri et al., 2007).

The genetic diversity of mungbean germplasm can be increased through induced-mutation using a chemical mutagenic agent like

ethylmethane sulphonate (EMS) or a physical mutagenic agent like Gamma-ray Irradiation (Harten, 1998). However, the sudden changes in DNA due to mutagenesis can be unfavorable because they occur randomly within the genome and reduce the seed fertility of the next generation (Tah, 2006). In the assembly of mungbean mutant cultivar resistant to Yellow Mosaic Virus (YMV), approximately only 10 percent of mutants found among 2500-3000 plants grown in every generation, but none of them showed resistance to YMV. However after re-mutation of the 3rd generation (M3), YMV mutant was obtained and still need to be homogenized until 6th generation (M6). Thus, practically mutation breeding in mungbean is laborious and time consuming because large number of population is needed for the starting point of the research (Reddy, 2009).

The unfavorable mutation effect at the molecular level that can damage and change the normal DNA sequence of an organism may occur due to the failure of DNA repair mechanism. There are several DNA repair mechanisms in a cell of living organism. One of them is known as mismatch repair (MMR). This mechanism produces a protein that corrects DNA mismatches during DNA replication, homologous recombination (HR) or as a result of DNA damage caused by mutagenic agent. The understanding of MMR system is based on the *in-vitro* reconstitution of purified MMR

protein of prokaryotic organism, *Escherichia coli,* in which three genes involved, namely *MutS*, *MutL*, and *MutH* genes. In eukaryotic organism these genes have homologs and not all homologs have role in the MMR system. As reported in *Arabidopsis* plant, only four *MutS* homolog (*MSH*) involved in MMR system and worked as heterodimers, i.e. *MSH2-MSH3 (MutSβ), MSH2-MSH6 (MutSα),* and *MSH2-MSH7 (MutSY*). Each of them has particular mechanism in recognizing the DNA mismatch within the genome (Culligan et al., 2000; Schofield and Hsieh, 2003; Kunkel and Erie, 2005; Iyer et al., 2006).

Considering the importance of MMR in DNA repair mechanism and its relation to mutation activity, some researchers have reported the effect of MMR disruption. According to Schofield and Hsieh (2003), the gene deficient in MMR could lead to the increases of spontaneous mutation due to the frequent exhibit of microsatellite instability at mono- and di-nucleotide repeats. Study on tomato and arabidopsis showed that the crop has complete homologs of *MSH* genes and suppression of these genes also increased the HR which leads the acceleration of wild cultivar introgression. In case of tomato, this crop also has low genetic diversity in nature (Li et al., 2005; Tam et al., 2009; Tam et al., 2011). Another study of MMR

inactivation on human cancer genome caused a large scale regional mutation rate variation as well (Supek and Lehner, 2015).

As we proposed earlier, mungbean has problems in low genetic diversity and low mutation variation as those in tomato. By correlating these facts, we assume that MMR mechanisms, particularly the presence of *MSH* genes possibly correlated with the mutation behavior in mungbean. Therefore in this study we hypothesized that the *MSH* genes exist in mungbean and in a complete homologs form. To verify this hypothesis, we characterized the mungbean *MSH* genes by identifying its location within the mungbean genome. Since the whole genome mungbean reference is available, we used comparative sequence analysis and include several germplasm from various countries to explore any DNA variation of the genes among accessions as well as the prediction of protein alteration among germplasm. Result of this study may facilitate further work such as gene manipulation or mutant detection in the early stage of mungbean plant. Thus, is expected to improve the breeding efficiency of mungbean.

LITERATURE REVIEW

DNA Repair Mechanisms in Plant

During the lifetime, DNA of any organism including plant can be damaged by spontaneous cleavage of chemical bonds in DNA, by environmental agents such as ultraviolet and ionizing radiation, and by reaction with genotoxic chemical that are by-products of normal cellular metabolism or occur in the environment. This damage can cause a mutation, a change in the normal DNA sequence. The mutation if left uncorrected and accumulate within the cell, may cause no longer function of the cell and unable to produce viable offspring. Thus the prevention of DNA sequence errors in all types of cells is important for survival and several cellular mechanisms for repairing damaged DNA and correcting sequence errors have evolved.

The first line of defense in preventing mutations is the proofreading activity of DNA polymerase. In prokaryotes for instance, during their DNA replication, 1 incorrect nucleotide per 10^4 polymerized nucleotides may occur. To correct this error, DNA polymerase through the exonuclease activity pause the replication and transfers the 3'end of the growing chain to its exonuclease site where the incorrect mispaired base is removed. Then 3'

end is transferred back to the polymerase site, where this region is copied correctly (Lodish et al., 2013).

In addition to proofreading activity, cells have other repair systems for preventing mutations, i.e. base excision repairs (BER), nucleotide excision repair (NER), double strand break repair (DSBR), and mismatch repair (MMR). The BER is mainly caused by chemical mutagenic agent such as ethyl methane sulfonate (EMS) which cause base modification of a mutated G-A. The repair pathway is initiated by removal of the damaged base by a DNA glycosylase enzyme which results in cleaving of AP site (3' side of the abasic) by AP endonuclease. This cleaved site then becomes the substrate for the SSB repair pathway through short-patch or long-patch repair mechanism (Bray and West, 2005).

In contrast to BER, NER can detect modifications indirectly by conformational changes to the DNA duplex rather than relying on the recognition of specific DNA damage products. NER targets the damaged strand and removes a 24-32 base oligonucleotide containing the damaged product. DNA synthesis and ligation completes the repair process (Sancar et al., 2004).

Meanwhile the MMR complements the activity of DNA polymerase proofreading activity in order to maintain the genomic integrity. MMR may

also have an important role in recognizing mismatches at sites of recombination between DNA sequences, thereby reducing the rate of occurrence of recombination events which might lead to inappropriate chromosome rearrangements of interspecies hybridization (Wu et al., 2003). MMR in prokaryote is performed by MutHLS system, whereby *Mut*S homodimers recognize and bind to insertion/deletion loops (1-4 bp) and repair mismatch. In the presence of ATP, *MutS* recruits MutL (an ATPase) and activates *MutH* (methylation sensitive endonuclease) that cleaves the transiently unmethylated DNA strand, targeting MMR to newly synthesized DNA strand. Prokaryotes have two homologs namely *MutS1*, work for MMR which is described before, and *MutS*2 which involves in meiotic crossing over and chromosome segregation. In eukaryotes, homologs of *MutS* and *MutL* have both found, but not *MutH*. Homologs of *MutS* in eukaryote namely *MSH1* to *MSH7*, with *MSH7* is being unique to plant. Whereas homologs of *MutL* in eukaryote, namely *MLH1*, *MLH2* or *hPMS1*, *MLH3*, and *PMS1* or *hPMS2*. Heterodimers of *MSH* protein in eukaryote provide substrate specificity, i.e. *MutSα* (*MSH2-MSH6*) which repairs base-base mismatch, *MutSβ* (*MSH2-MSH3*) which repairs +1 insertion/deletion loops (IDLs) and larger loops of 2-8 bp, and *MutSϒ* (*MSH2-MSH7*) which repairs G/T mismatch. While *MSH1* is required for mitochondrial stability

and *MSH4-MSH5* function in meiosis and involve in resolution of Holliday junctions during meiosis (Obmolova et al., 2000; Schofield and Hsieh, 2003; and Kunkel and Erie, 2005).

Unlike NER, BER, and MMR which repair the error of single strand DNA, the DSBR repair the double-strand breaks in DNA (dsDNA). These are particularly severe lesions because incorrect rejoining of dsDNA can lead to gross chromosomal rearrangements that can affect the functioning of genes. The DSBR is mainly caused by the activity of nuclease such as *HindII*, *EcoRI*, and *FokI*. This enzyme may capable to cleave phosphodiester bonds between the nucleotide subunits of nucleic acids. Two systems have evolved in DSBR, i.e. homologous recombination (HR) and non-homologous end-joining (NHEJ). HR uses an identical or very similar DNA sequence as a template for the repair of a DSB, while NHEJ recombines DNA largely independent of the sequence (Bray and West, 2005; Lodish et al., 2013).

Comparative Sequence and Phylogenetic Analysis

It is well known that plant genomes tend to be large and complex, thus made very diverse in growth habit and environmental adaptation. Despite this diversity, plant geneticists have found that plants exhibit

extensive conservation of both gene content and gene order. On the other hand, the advent of DNA marker and sequencing technology not only facilitated the rapid generation of detailed plant genetic maps but also allowed map comparisons among species. The comparison between closely related species indicated extensive collinearity of genetic maps. Many comparative studies also show that within the limits of sequence divergence that permit cross-hybridization, the large majority of plant genes have close homologs within most other plant genomes. This means that different plant species often use homologous genes for very similar functions. This becomes the basis of the comparative sequence analyses which commonly applied in the reverse genetic approach (Bennetzen, 2000).

In relation with the phylogenetic analyses, comparative method is applied to gain insight the historical relationships of lineages based on evolutionary hypotheses. Moreover, it is known that differences and similarities among species are the basis of phylogenetic analyses thus made the comparative sequence and phylogenetic study are closely related each other. However, building hypotheses about the evolutionary history of species is a challenging task, as it requires knowledge about the underlying methodology and an ability to flexibly manipulate data in diverse formats. Although most practitioners are not experts in phylogenetic, the appropriate

handling of phylogenetic information is crucial for making evolutionary inferences in comparative study.

Single Nucleotide Polymorphism (SNP)

A single nucleotide polymorphism (SNP) is a variation in a single nucleotide which may occur at some specific position in the genome, where each variation is present to some appreciable degree within a population. SNPs may fall within the coding sequences of genes, non-coding regions of genes, or in the intergenic regions. SNPs in the coding region are of two types, synonymous and nonsynonymous SNPs. Synonymous SNPs do not affect the protein sequence while nonsynonymous SNPs change the amino acid sequence of protein. The nonsynonymous SNPs are of two types, i.e. missense and nonsense. SNPs that are not in protein-coding regions may still affect gene splicing, transcription factor binding, messenger RNA degradation, or the sequence of non-coding RNA. Gene expression affected by this type of SNP is referred to as an *eSNP* (expression SNP) and may be upstream or downstream from the gene.

There are several methods applied for discovery and identification of new SNPs, i.e. (1) locus specific-PCR amplification, (2) alignment among available genomic sequences, (3) whole genome shotgun sequences, (4)

overlapping regions in BACs and PACs, and (5) reduced representation shotgun (RRS). The first and the second methods can be used only for genomic regions with known sequences since prior sequence information is necessary. In the third method, several fold coverage of the whole genome is required before SNPs can be detected by alignment of sequences belonging to the same locus. The fourth one is the common methods for SNPs detection by a mismatch that have been used for genome sequencing. Whereas the last method is used when the genomic sequences may not be available or it may not be desirable to use the available genomic sequences for the discovery of SNPs. This approach uses subsets of genome, each containing manageable number of loci to permit resampling (Gupta et al., 2001).

In Silico Prediction of Protein Function Changes

Many genetic variations are SNPs which can be in the form of synonymous and non-synonymous SNPs. Non-synonymous SNPs are neutral if the resulting point-mutated protein is not functionally visible from the wild type and non-neutral otherwise. The *in silico* prediction of the effect from non-synonymous SNPs are developed recently which have given a great contribution to the efficiency of genomic study. SNAP2 and

PROVEAN are some example of the *in silico* prediction beside other popular tool such as SIFT and Polyphen-2.

SNAP (Screening for non-acceptable polymorphisms) is based on neural network that predicts the functional effects of mutations by distinguishing between effect and neutral variants of non-synonymous SNPs. The most important input signal for the prediction is the evolutionary information taken from an automatically generated multiple sequence alignment. Structural features such as predicted secondary structure and solvent accessibility are considered as well. If available also annotation (i.e. known functional residues, pattern, regions) of the sequence or close homologs are pulled in. In a cross-validation over 100,000 experimentally annotated variants, SNAP2 reached sustained two-state accuracy (effect/neutral) of 82% (at an AUC of 0.9) (Bromberg and Rost, 2007).

Contrast with other *in silico* program, PROVEAN (Protein Variation Effect Analyzer) not only predicts the effect of amino acid substitution but also an insertion and deletion as well. In PROVEAN, a delta alignment score is computed for each supporting sequence. The scores are then averaged within and across clusters to generate the final PROVEAN score. If the PROVEAN score is equal to or below a predefined threshold (e.g. -2.5), the protein variant is predicted to have a "deleterious" effect. If the

PROVEAN score is above the threshold, the variant is predicted to have a "neutral" effect. This score based on the reference and variant versions of a protein query sequence with respect to sequence homologs collected from the NCBI NR protein database through BLAST. Compare to SIFT and Polyphen-2, the prediction results by PROVEAN is in agreement and shared by all about 78.5% (15,618/19,898) of disease-associated variants and 46.8% (16,244/34,701) common variants (Choi et. al., 2012; Choi and Chan, 2015).

MATERIALS AND METHODS

Identification of Mungbean *MSH* Location

NCBI database search was performed to find previous identified and potential *MSH* family genes in the model plant, *Arabidopsis thaliana*. We used "*MSH1, MSH2, MSH3, MSH4, MSH5, MSH6*, and *MSH7*" as a query to search the protein/amino acid sequences of *MSH* homologs that involved in MMR. We aligned the longest sequence and *RefSeq* type from arabidopsis *MSH* homologs (*AtMSH*) to the mungbean whole-genome reference which was assembled by Van et al. (2013) through the SNU's Crop Genomics Laboratory homepage (http://plantgenomics.snu.ac.kr/sequenceserver). The most matched sequence/ GeneID was selected as the gene sequence for each *MSH* homologs (*VrMSH*).

We also identified the domain within *MSH* genes by analyzing the *VrMSH* genes into the integrated protein signature databases (InterPro) database (http://www.ebi.ac.uk/interpro/). The domain identification is needed in further analysis.

Phylogenetic and Multiple Alignment Analysis of *MSH* Genes among Species

We performed the NCBI database search to obtain *MSH* protein sequence of 14 species which cover several order. These species were distributed randomly under eudicots clade and analyzed together with *MSH* mungbean sequences in phylogenetic and multiple alignments analysis (Table 1).

The phylogenetic analysis was performed by MEGA6 software using Neighbor-Joining (NJ) method which is supported by bootstrap 1000 replications The distance matrices for specific groups of *MSH* protein sequences were computed based on the Jones-Taylor-Thornton (J-T-T) model (Felsenstein, 1985; Saitou and Nei, 1987; Jones et al., 1992; Tamura et al., 2013). Whereas the multiple alignment and synteny analysis was performed by MEGA6 software as well using ClustalW program with default values for gap opening (10), extension (0.2) penalties and the GONNET 250 protein similarity matrix.

Table 1. List of *MSH* protein sequences used in phylogenetic and multiple alignment analysis

Crop/Clade	*MSH1*	*MSH2*	*MSH3*	*MSH4*	*MSH5*	*MSH6*	*MSH7*
	NCBI ID (prot. length)	NCBI ID (prot. length)	NCBI ID (prot. length)	NCBI ID (prot. length)	NCBI ID (prot. length)	NCBI ID (prot. length)	NCBI ID (prot. length)
Medicago/ Legumes	-	-	-	-	-	-	gi357500449 (1160 aa)
Chickpea/ Legumes	gi502122529 (1141 aa)	gi502151706 (942 aa)	-	-	gi502099189 (809 aa)	-	gi502163835 (1098 aa)
Soybean/ Legumes	gi356575134 (1134 aa)	gi356563103 (942 aa)	-	-	gi571506599 (812 aa)	-	gi571478271 (1079 aa)
Strawberry/ Fabids	gi764569327 (1141 aa)	gi470126534 (942 aa)	gi70144922 (1106 aa)	gi470130586 (792 aa)	gi470119462 (809 aa)	gi764592252 (1252 aa)	gi764505215 (1075 aa)
Cucumber/ Fabids	gi778678067 (1152 aa)	gi778656285 (942 aa)	gi778679553 (1110 aa)	gi778708277 (789 aa)	gi449463733 (807 aa)	gi449436747 (1307 aa)	gi449443325 (1095 aa)
Jatropha/ Malphigiales	gi802633693 (1146 aa)	gi802797191 (936 aa)	gi802769131 (1105 aa)	gi802582003 (792 aa)	gi802588379 (807 aa)	gi802689824 (1304 aa)	gi802627380 (1108 aa)

Poplar/ Malphigiales	-	gi222858604 (944 aa)	-	-	-	-	-
Arabidopsis/ Malvids	gi75297828 (1118 aa)	gi42565226 (937 aa)	gi30686920 (1081 aa)	gi79476962 (792 aa)	gi186510260 (807 aa)	gi332656719 (1324 aa)	gi12643849 (1109 aa)
Cacao/ Malvids	-	gi508773672 (967 aa)	gi508775913 (1115 aa)	-	gi508720408 (818 aa)	-	-
Brassica rapa/ Malvids	gi685328741 (1122)	gi685289267 (937)	gi685293250 (1098 aa)	gi685343968 (792 aa)	gi685310043 (807 aa)	gi685287036 (1337 aa)	gi685263673 (1101 aa)
Eucalyptus/ Myrtales	gi702441803 (1152 aa)	gi702259274 (942 aa)	-	gi702336056 (790 aa)	gi702363375 (807 aa)	gi702500679 (1318 aa)	gi702305220 (1083 aa)
Grape/ Vitales	gi225433289 (1144 aa)	gi731426269 (945 aa)	gi731423415 (1111 aa)	-	gi731432937 (872 aa)	gi225437545 (1297 aa)	gi731406967 (1105 aa)
Tomato/ Asterids	gi460404638 (1137 aa)	gi350538025 (943 aa)	gi723679590 (1119 aa)	gi723719921 (792 aa)	gi723735564 (834 aa)	-	gi723713547 (1082 aa)
Potato/ Asterids	gi565347746 (1137 aa)	gi565376482 (943 aa)	-	-	gi565343547 (831 aa)	-	gi565348531 (1078 aa)

Variation Identification of *MSH* Genes among Mungbean Accessions

Twelve accessions from the mungbean germplasm collection (Table 2) were selected based on the genetic diversity analysis from the previous study (Sangiri et al., 2007; Lestari et al., 2014). Germplasm from region which have higher diversity content was selected more than other regions. Thus, the genetic diversity of chosen germplasm will be as similar as possible to the actual genetic diversity of entire collection. We used CTAB methods to extract the DNA from the young leaves of the chosen germplasm (Gelvin and Schilperoort, 1995). The DNA quality and quantity was observed and measured by agarose gel electrophoresis and NanoDrop platform, respectively.

Table 2. List of mungbean germplasm used in the study

ID	Name	Country of Origin	ID	Name	Country of Origin
V1	JP2291819	India	V7	Tecer Hijau	Indonesia
V2	JP229177	India	V8	Utang Wewe	Indonesia
V3	JP229193	India	V9	JP78939	Vietnam
V4	JP229130	Bangladesh	V10	JP229096	Thailand
V5	JP81649	Srilanka	V11	Sunhwanokdu	South Korea
V6	JP99066	Pakistan	V12	Gyonggijere5	South Korea

Since the domain I and V of *MSH* genes play important role in the MMR, we developed primers using *Primer3* software which flank the important motifs within those domains. The location of the targeted domains and motifs were obtained from previous analysis. We used these primers to amplify the DNA of 12 chosen mungbean germplasm. PCR conditions were: one cycle of 94^0C for 5 min; then 35 cycles of 94^0C denaturation for 30s, 60^0C for 30s-45s, and 72^0C for 30s; with a final extension cycle of 72^0C for 5 min. PCR products were visualized by agarose gel electrophoresis and were sequenced by NICEM sequencing facility (http://nicem.snu.ac.kr). Sequence files then were manually edited and aligned using MEGA6 software.

Afterward we aligned the *MSH* genes sequences of 12 mungbean accessions and identified any occurrence of the single nucleotide polymorphisms (SNPs) which lay within the coding regions of the genes. We focused the observation around the important motifs within the domain I and V of the mungbean *MSH2*, *MSH3*, *MSH6*, and *MSH7* because only these homologs related to the MMR. Then based on the SNPs found, we computationally predicted the effect of non-synonymous SNPs to the alteration of protein function.

The prediction was performed by SNAP2 program which can be accessed online at https://rostlab.org/services/snap/. SNAP2 is based on neural network that predicts the functional efects of mutations by distinguishing between effect and neutral variants of non-synonymous SNPs (Bromberg and Rost, 2007). Since the SNAP2 only predict the amino acid substitution (AAS) we perfomed another *in silico* prediction to predict the effect of insert and deletions (InDels), i.e the PROVEAN which can be accessed at http://provean.jcvi.org/seq_submit.php. PROVEAN (Protein Variation Effect Analyzer) is a software tool which predicts wether an AAS or InDels has an impact on the biological function of protein. This computation is comparable to popular tools such as SIFT (Sorting Tolerant from Intolerant) or PolyPhen-2 (Choi and Chan, 2015).

.

RESULTS

The Characteristics of Mungbean *MSH* Genes

The alignment of Arabidopsis *MSH* proteins to the whole genome sequence of Korean mungbean cultivar (Suhnwanokdu) resulting in seven most matched gene ID within the mungbean genome. Four of them were detected matched to more than one *MSH* homologs. However, we defined the gene for each homologs based on the highest E-value in corresponding homologs. Therefore the location of *MSH* homologs genes in the mungbean genome are *MSH1* and *MSH3* in chromosome 8, *MSH2* in chromosome 7, *MSH4* in chromosome 11, *MSH5* in chromosome 6, *MSH6* in chromosome 3, and *MSH7* in chromosome 1. Analysis of these protein sequences using BLASTp into NCBI database shows that all homologs of mungbean *MSH* that involve in mismatch repair are most similar to soybean *MSH* protein. The levels of identity are 74% for *MSH1*, 89% for *MSH2*, 89% for *MSH3*, 80% for *MSH5*, 98% for *MSH6*, and 96% for *MSH7*.

For the four homologs of *MSH* genes that related to the MMR, analysis of their *MSH* protein sequences into InterPro database indicates that the sequences are likely to be functional homologs of the DNA mismatch repair proteins. Multiple significant hits from Pfam, SMART, Superfamily,

and PANTHER database were detected and showing that the sequences contain the conserved domains and motifs recognizable for *MutS*/*MSH* protein. Based on the Pfam database, we identify five domains in *MSH2*, *MSH3*, and *MSH6*, while *MSH7* only has four domains. The length of these genes are 8211 – 9322 base pairs with the *MSH2* as the shortest and *MSH3* as the longest genes. However, *MSH6* has longer protein sequence than *MSH3* although its nucleotide is shorter (Figure 1).

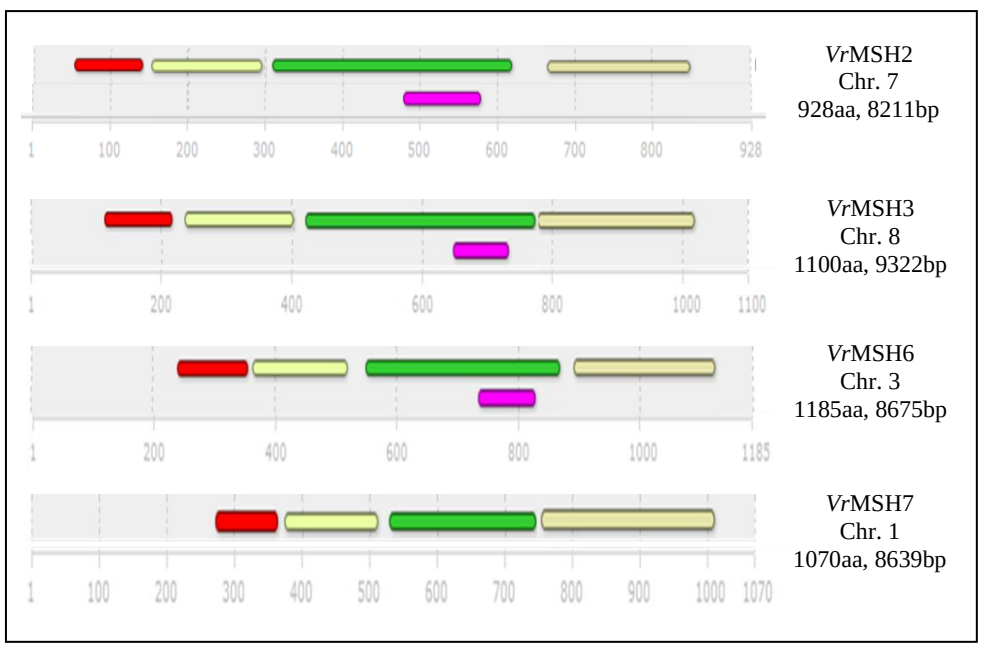


Figure 1. Domain structure of *VrMSH* genes. Color line: red=domain I, yellow=domain II, green=domain III, purple=domain IV, light brown=domain V

Phylogeny and Comparison of *MSH* Protein among Crop Species

The evolutionary history among *MSH* genes in several species of eukaryotes was inferred using the Neighbor-Joining method with 1000 replication of bootstrap analysis and involved 77 amino acids sequences of full length *MSH* protein from 15 species. The phylogenetic tree shows clearly separation of seven homologs of *MSH* genes from *MSH1* to *MSH7* (Figure 2). *MSH1* homolog is the deepest branch within the cluster which is supported with high bootstrap value of 100. Following this, three main groups are identified. The first consists of *MSH6* and *MSH 7* (99% bootstrap value); the second group consists of *MSH2* and *MSH5* (50% bootstrap value); and the third group consists of *MSH3* and *MSH4* (56% bootstrap value).

Meanwhile the mungbean *MSH* genes also resolve clearly within their respective protein groups (Figure 2). Mungbean *MSH2* and *MSH7* are sister to other legumes *MSH2* and *MSH7* (soybean, chickpea, and medicago), all with strongly supported bootstrap values (100%). Contrary with that, mungbean *MSH3* and *MSH6* are grouped separately with other legumes. The pattern of phylogenetic tree can be used as alignment basis to identify protein sequences variation between mungbean *MSH* and their orthologues in closest species.

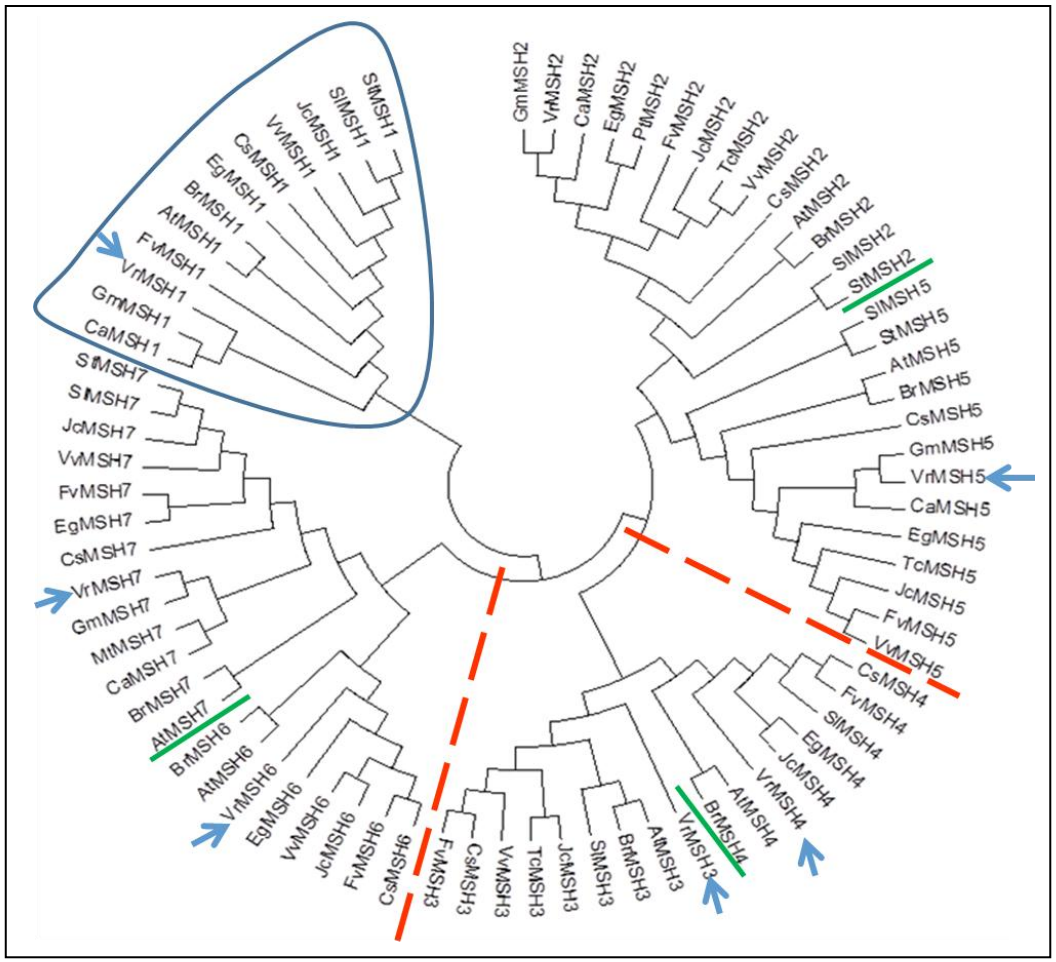
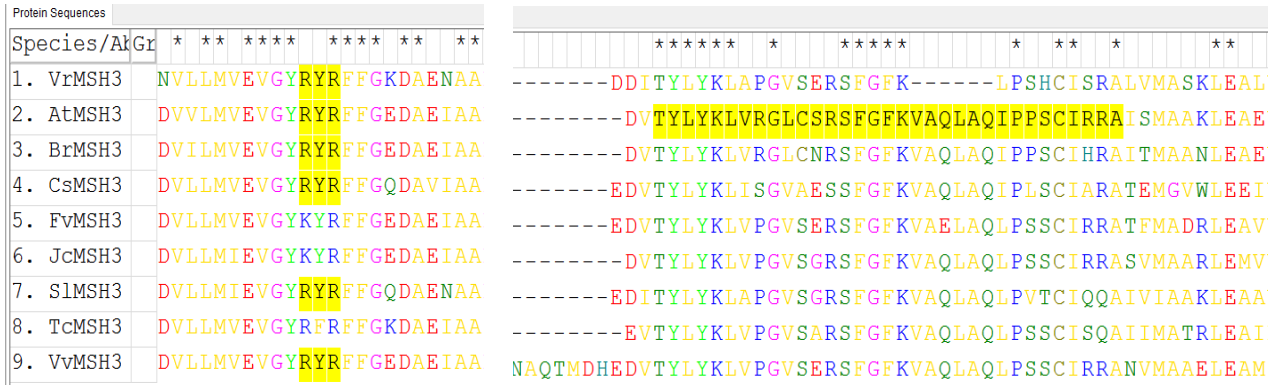


Figure 2. Evolutionary relationships of *MSH* genes in eukaryotes.

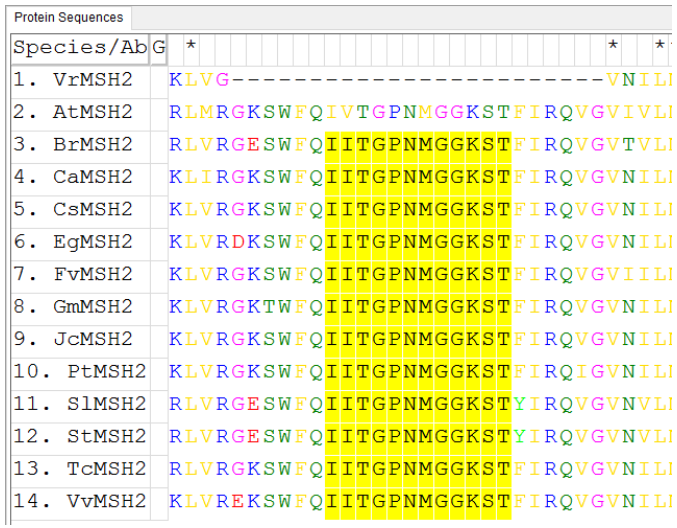
Species abbreviation:
At=*Arabidopsis thaliana*, Br=*Brassica rapa*, Ca=*Cicer arietinum*, Cs=*Cucumis sativus*, Eg=*Eucalyptus grandis*, Fv=*Fragaria vesca*, Gm=*Glycine max*, Jc=*Jatropha curcas*, Mt=*Medicago truncatula*, Pt=*Populus trichocarpa*, Tc=*Theobroma cacao*, Sl=*Solanum lycopersicum*, St=*Solanum tuberosum*, Vr=*Vigna radiata*, Vv=*Vitis vinifera*

We also conducted multiple alignments of *MSH* proteins for *MSH2*, *MSH3*, *MSH6*, and *MSH7*. From the alignment we identified amino acids variation among species, especially in the neighbor-motif within domain I and domain V. In domain I, we identify FYE motif both in *MSH6* and *MSH7* for all species. Another recognition motif of MFE in *MSH2* is identified as well in all species. However, for *MSH3* we detect varies of recognition motifs which include RYR in mungbean, arabidopsis, brassica, chickpea, tomato, and grape; KYR in strawberry and jatropha; and RFR only in cacao (Figure 3a).

Five important motifs that involve in ATP hydrolysis are well known at domain V of *MSH* genes, i.e. Walker A, Motif C, Walker B, Motif D, and HTH subdomain. In our multiple alignment analysis, Walker A motif is absence in mungbean *MSH2* contrast with other species (Figure 3b). Mungbean *MSH*3 also loses its six residues within the HTH subdomain, although the specific motif of YGA still remains (Figure 3a). The HTH subdomain has YGA residues in which each residue is separated by 4 and 23 residues respectively.



(a)



(b)

Figure 3.

(a) Multiple alignment analysis of domain I and HTH subdomain domain V at *MSH3* protein sequences from nine species

(b) Multiple alignment analysis of Walker A domain V at *MSH2* protein sequences from 14 species.

Same color of amino acid denotes same biochemical group. Dashes denote gaps. Important motifs are denoted by yellow boxes.

MSH Genes Variation among Mungbean Accessions

To identify the variation of *MSH* protein sequences among mungbean accessions, we developed primers in domain I and V of the genes since they are play importance role in MMR (Table 3). Based on these primers, we found some single nucleotide polymorphisms (SNPs) within the coding region of the genes. Total of seven SNPs are identified at *MSH2*, *MSH3*, and *MSH6*. Four of them are categorized as non-synonymous SNPs. These non-synonymous SNPs are found in mungbean accessions from India, Srilanka, Pakistan, and Indonesia. We also found another one deletion at *MSH2* from Indonesia mungbean accession (Table 4).

Since only non-synonymous SNPs that can alter the composition of amino acids and probably the protein function as well, we performed *in silico* prediction to predict this changes effect. The prediction was carried out by the SNAP2 program which has score range from -100 to -50 (neutral effect), >-50 to 50 (weak effect), and >50 to 100 (strong effect) to the protein function changes. From 4 non-synonymous SNPs found, 2 of them have neutral effect and the other two are predicted changing the protein function compare to the reference (Figure 4).

Table 3. Motifs, locations, and primers designed for Domain I and V of *MSH* homologs gene related to MMR

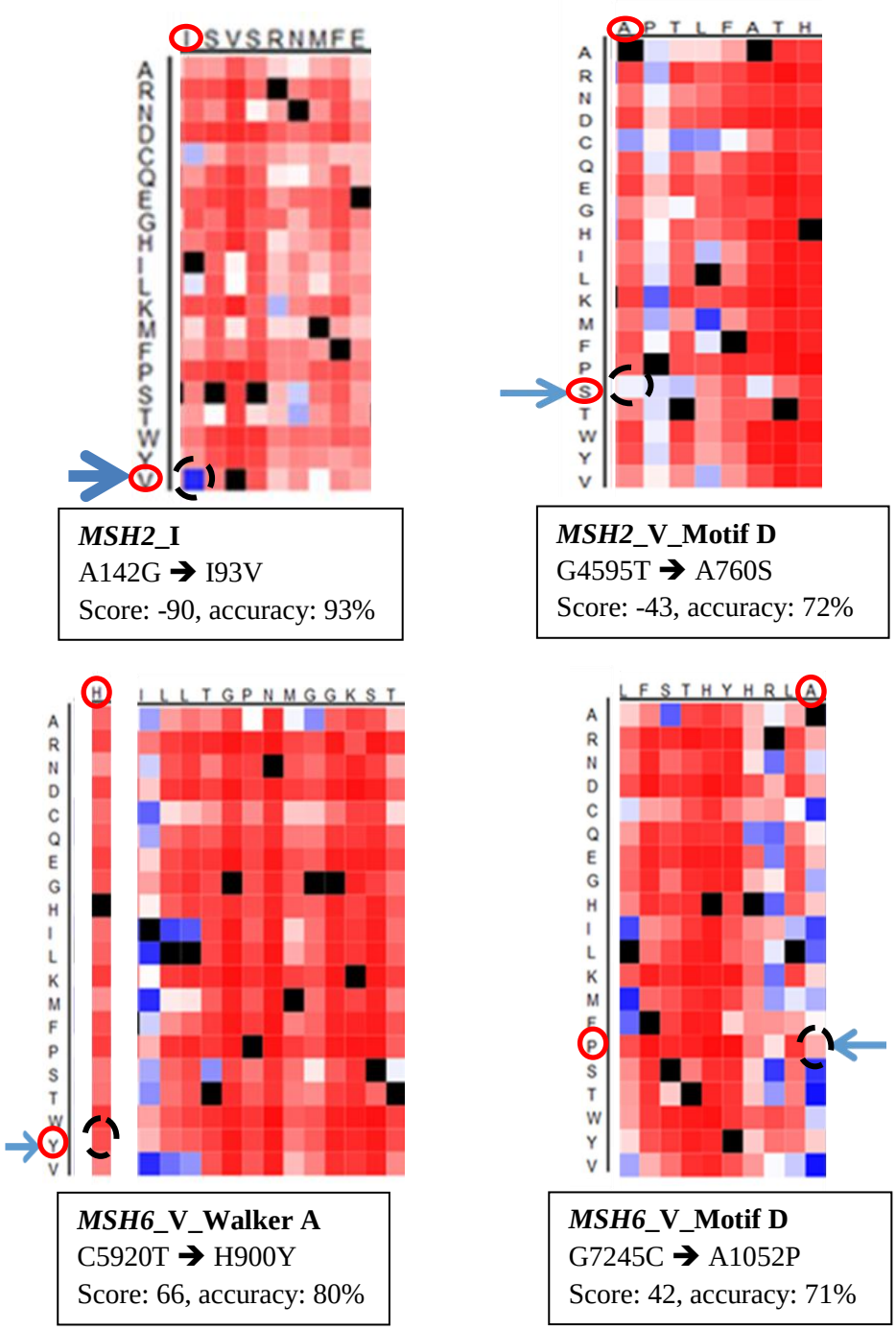
Homolog_Domain	Motif	Forward primer	Reverse primer
*MSH2*_I	MFE	ATGGCGACAATGCAACTTTC	GCGTTCCACTTTTGACCAGT
*MSH2*_V	Motif C, Walker B	ATTCTCCCCAGCTACGTGGT	GAAGAAGCCATTGTACAGGTCA
*MSH2*_V	Motif D	CAATGGTGGCATTGGTGTAA	CAAGGGCTAAAGCAGTCAGC
*MSH3*_I	RYR	CAGGAACCTTCTTCCCCTTC	GTGGGCGTAAATGCCTAAGA
*MSH3*_V	Walker A	ATCTGAATGCCCCACTTTCA	GACACCGCATTGGATCTACC
*MSH3*_V	Motif C, Walker B	CTGCACGTCCTGGATAGGAT	CAAGCTGGCAATCTTTGGAT
*MSH3*_V	Motif D	ATGAGCTTGGGAGAGGAACA	CTGGGCAACCTTAAATCCAA
*MSH6*_I	FYE	CCACAATGAGGTTGGTCTCC	GTCCATTCCTTCCAACCAAA
*MSH6*_V	Walker A	GCCAGAATCACAGTCAAGCA	ATGAAGGACCAACATGCACA
*MSH6*_V	Motif C, Walker B	ACCTTCCGCACAAAATGTTC	GAATGGGGGCCAAAGATAAT
*MSH6*_V	Motif D	GGAATACCTTGGGATCGTTG	GGGACTGCAACTTCTGATGG
*MSH7*_I	FYE	ATGCCGCAATTAATGGTCAA	CATCATCAATCCCACTTTCAGA
*MSH7*_V	Walker A	TGACACTGGAGGAACTGTGC	GAGAAGAAACCTGGGCCATA
*MSH7*_V	Motif C, Walker B	CACGACTTGGAGCCAAAGAT	AATGGCGTAGCCATCAAAAG
*MSH7*_V	Motif D	TTTGGTCCCGAGCATTTTTA	CATTGTAACGCGTGGATGAG

Table 4. SNPs and InDels found around important motif of Domain I and V

Type	*MSH_* domain	Genotype	SNP position (3'..5')		Motif
			DNA	Amino acid	
Synonymous SNPs	*MSH3*_I	V1,V3,V6,V8	C420T*	A140A	RYR
	*MSH6*_V	V1	G7039A	G1023G	Walker B
	*MSH6*_V	V5	C7241T	R1050R	Motif D
Non-synonymous SNPs	*MSH2*_I	V1,V3,V6,V8	A142G**	I93V	MFE
	*MSH2*_V	V7	G4595T	A760S	Motif D
	*MSH6*_V	V5	C5920T	H900Y	Walker A
	*MSH6*_V	V2	G7245C	A1052P	Motif D
Deletion	*MSH2*_I	V7	A1525del	M80del	1bp deletion

*C420T ➔ A140A: SNP is found at nucleotide position of 420 where Cytosine (C) is replaced by Thymine (T) and do not change the resulting amino acid at position 140

** A142G ➔ I93V: SNP is found at nucleotide position of 142 where Adenosine (A) is replaced by Guanosine (G), thus resulting in amino acid changes at position 93 from Isoleucine (I) to Valine (V).



***MSH2*_I**
A142G ➔ I93V
Score: -90, accuracy: 93%

***MSH2*_V_Motif D**
G4595T ➔ A760S
Score: -43, accuracy: 72%

***MSH6*_V_Walker A**
C5920T ➔ H900Y
Score: 66, accuracy: 80%

***MSH6*_V_Motif D**
G7245C ➔ A1052P
Score: 42, accuracy: 71%

Figure 4. Prediction of protein function changes of non-synonymous SNPs at *MSH2* and *MSH6* from SNAP2 program

Meanwhile for the one base deletion at domain I *MSH2*, we predicted the effect through PROVEAN web server. We include not only the deletion effect but the AAS as well that are found in *MSH2*. The result shows that the base deletion causes deleterious effects. The AAS at *MSH6* mungbean homologs also shows similar result with those analyzed through SNAP2 program (Table 5).

Table 5. PROVEAN result for amino acid substitution (AAS) and deletion at mungbean *MSH2* and *MSH6*

Homologs	Variant	PROVEAN score	Prediction (cutoff= -2.5)
MSH2	M80del	-10.087	Deleterious
	I93V	0.828	Neutral
	A760S	-1.763	Neutral
MSH6	H900Y	-5.650	Deleterious
	A1052P	-3.026	Deleterious

DISCUSSION

The Characteristics of Mungbean *MSH* Genes

The main objective of our study is to identify and characterize the *MSH* homologs of mungbean. We utilized the availability of mungbean whole genome sequence and *MutS/MSH* genes bioinformatics resources to determine the location of mungbean *MSH* homologs. Those open-access resources allowed us to perform a faster and better characterization of mungbean *MSH* genes.

We identify two homologs located in the chromosome 8, i.e. *MSH1* and *MSH3*. The distance among two homologs is around 31Mb. Meanwhile *MSH2*, *MSH4*, *MSH5*, *MSH6*, and *MSH7* are located in chromosome 7, 11, 6, 3, and 1, respectively. As reported in rye, *MSH2* was mapped to chromosome 1R, *MSH3* was mapped to chromosome 2R and *MSH6* to chromosome 5R. In tomato, *MSH2* and *MSH7* were located in chromosome 6 and 7, respectively. Meanwhile in wheat, *MSH2*, *MSH3*, and *MSH6* were completely detected in three genomes of wheat (A, B, and D genomes). *MSH2* was detected on chromosome 1A, 1B, and 1D; *MSH3* on chromosomes 2A, 2B, and 2D; and *MSH6* on chromosome 5 and 3 of genomes A, B, and D (Korzun et al., 1999; Tam et al., 2009). Accordingly,

in mungbean no homologs that form a heterodimer contact are located in the same chromosome. As has been stated previously, *MSH* proteins functions as heterodimers with distinct mismatch specificities. *MSH2-MSH3* or *MSHβ* recognizes insertion/deletion loops and larger loops of 2-8bp. *MSH2-MSH6* or *MSHα* recognizes base-base mismatch, while *MSH2-MSH7* or *MSHϒ* recognizes a G/T mismatch (Culligan and Hays. 2000; Wu *et. al.* 2003).

Among the four *MSH* homologs that related to MMR in mungbean, only *MSH7* is absence of domain IV which is also supported in other study (Tam et al. 2009). Based on clamp-sliding model, domain IV has function in DNA binding together with domain I. When domain I bind specifically to the mismatch site, domain IV forms a jaws and bind non-specifically to the dsDNA (Tachiki et al., 1998). Therefore, this domain often called as clamp domain and thus make the *MSH7* is being unique for plant. It is presumed that the deletion of clamp domain in *MSH7* causing the reduction of *MSH7* protein-kinking efficiency and therefore bind less well to heteroduplex DNA and to an extra looped out of the nucleotide (Wu et al., 2003).

Based on the domain location within the mungbean *MSH* genes, we found that domain III is the longest domain and containing domain IV as well. This can be explained since domain III act as the core domain for *MSH* genes, whereby connected directly to the three domains, i.e. domain II,

domain IV, and domain V by peptide bonds (Obmolova et al. 2000). On the other hand, the location of domain I is not same in all mungbean *MSH* paralogs. Domain I of *MSH6* and *MSH7* are located slightly far from the transcription start site (TSS) about 240bp and 274bp, respectively. This similar pattern may cause their grouping in same cluster in the phylogenetic tree of *MSH* genes.

Phylogeny and Comparison of *MSH* protein among Crop Species

We built the phylogenetic tree of *MSH* protein sequences among species using their full length of protein. As reported by Culligan et al. (2000), the use of only the C-terminal regions in phylogenetic analysis resulted in tree instabilities. This instability makes the critical identification of relationship among *MSH* homologs being difficult. Our result shows general agreement with other studies especially in term of the most distant of *MSH1* from other homologs (Culligan *et. al*. 2000; Tam *et. al*. 2009). This support the theory of *MSH1* as the eukaryotic precursor which was transferred from *MutS* gene evolution of prokaryotes through the mitochondrial endosymbiotic events. The gene was duplicated at the nucleus and one gene was targeted back the protein to the mitochondrion while the others give rise to nuclear mismatch repair genes. Therefore, *MSH* gene

families are monophyletic which appear to share common ancestor (Culligan *et. al*. 2000).

In our study after the speciation of *MSH1*, we determine three major groups which consist of two homologs for each group. This is slightly different with those in tomato whereby *MSH5* was grouped separately apart from other homologs (Tam et al., 2009). The terminal branch also shows different pattern with those in tomato. In our study, *MSH2* and *MSH7* cluster have longer terminal branch length compare to *MSH3* and *MSH6*. This terminal branch length denotes how far the changes between orthologs. The longer branch of *MSH7* cluster can be understood since this homologs is being unique for plant, thus high variation among orthologs is possible. However, a long branch of *MSH2* does not support its biochemical function as core dimer in the complex protein network, which should be restricted for permissible changes.

From five domains of *MSH* genes, two domains play an important role, i.e. domain I and domain V. From domain I, the information of mismatch site recognition is transferred to the domain V through domain II and III (Obmolova et al., 2000). Therefore domain I is called as mismatch recognition domain and has specific motif called as FYE motif which is conserved for *MSH1*, *MSH6*, and *MSH7*; vary for *MSH3* and missing for

MSH4 and *MSH5*. The absence of these aromatic residues in *MSH4* and *MSH5* consistent with the evolution of *MSH* functional diversification which cause both homologs do not have role in mismatch repair (Culligan et al., 2000). Accordingly, our study shows consistent results in term of the conserved FYE motifs for *MSH1*, *MSH6*, and *MSH7* and varies motif for *MSH3.* In our study, three different mismatch recognition motifs for *MSH3* are identified, i.e. RYR, KYR, and RFR. Although vary, these motifs have similar pattern in term of containing two positively charged residues and one aromatic residue.

Once the mismatch site information is received by domain V, it is known that with the presence of ATP, *MSH* recruits *MLH* that could lead to the activation of *MutH*. The *MutH* can induce double strand break at the GATC site and let the DNA polymerases to correct the DNA mismatch (Schofield and Hsieh, 2003; Kunkel and Erie, 2005; Iyer et al., 2006). Therefore domain V of *MSH* genes majorly comprised of ATP binding site which contain five important motifs (Table 6).

Regarding to our study, mungbean *MSH* homologs contain all of these motifs, except for Walker A motif which is missing in mungbean *MSH2* compared to other species. Walker A functions in forming a loop that binds to the alpha and beta phosphates of di- and tri- nucleotide (Culligan

and Hays 2000; Tam *et. al*. 2009). However, since *MSH2* always become the partner of other homologs when they form a heterodimer in MMR system, the absence of Walker A might be expected. We assume that the need to form a binding-loop during the *MSH* heterodimer contact to the hetero-duplex DNA is done solely by Walker A region of *MSH2* heterodimer partner, i.e. *MSH3*, *MSH6*, or *MSH7*. However this early assumption needs to be tested further.

Meanwhile some part of HTH subdomain of *MSH3* also missing. However, the loss of six residues of HTH subdomain in *MSH3* apparently does not affect the function of the HTH for dimerization process since the YGA motif is remain conserved.

Table 6. Motifs in Domain V and their role in ATP hydrolisis

Motif	Consensus	Function
Walker A	GxxGxGKST	Form a loop that binds to the alpha and beta phosphates of di- and tri- nucleotide
Motif C	STF	Involved in ATP hydrolysis as a gamma phosphate sensor as a signal to the membrane spanning domain
Walker B	4 alliphatic+2 negatively charged + invariant D	Coordinate MG+ ion or polarize the attacking water molecule in ATP hydrolysis and act as switch region
Motif D	Invariant H	Polarizing the attacking water molecule during ATP hydrolysis
HTH subdomain	Y, G, A	Dimerization interface

MSH Genes Variation among Mungbean Accessions

SNP is a DNA sequence variation which occurs abundantly within the genome in which a single nucleotide differs among individual in a population. SNPs may fall within the coding, non-coding or intergenic regions within the genome. In the coding region, SNPs can be found in two types, i.e. synonymous SNPs and non-synonymous SNPs. Most of the

attention of the researcher is to the non-synonymous SNPs because it changes the amino acid sequence of protein. However, the change does not always resulting in protein function alternation. Since the domain I and domain V of *MSH* genes play the major role for its protein function, we observed the occurrence of non-synonymous SNPs in those domains, especially around the important motifs of the domain.

Based on the multiple alignment of partial sequence of *MSH2*, *MSH3*, *MSH6* and *MSH7* of 12 mungbean accessions, we found three synonymous SNPs, four non-synonymous SNPs, and one base deletion. The four non-synonymous SNPs were laid around the important motifs of domain I at *MSH2* and domain V at *MSH2* and *MSH6*. Based on SNAP2 program, two non-synonymous SNPs at *MSH6* are predicted to be affecting the protein function with the accuracy of 80% for *MSH6* Walker A motif and 71% for *MSH6* Motif D.

According to Brombreg and Rost (2007), SNAP2 program predict each substitution of amino acids independently and show every possible substitution at each position of a protein in a heatmap representation (Figure 4). Dark red indicates a high score (score>50, strong signal for effect), white indicates weak signals (-50<score<50), and blue a low score (score<-50, strong signal for neutral/no effect. While black marks the corresponding

wildtype residues. This signal is quantified in a score value which in line with the accuracy rate of prediction. Therefore a higher score result will also present the higher accuracy of analysis.

Meanwhile based on the PROVEAN software, the deletion at mungbean *MSH2* shows deleterious effect. The amino acid substitution (AAS) also shows the similar result with those run by SNAP2 program. Comparable to other *in silico* program, PROVEAN can generate predictions not only for single AAS but also for multiple AAS, insertions, and deletions using the same underlying scoring scheme. The score are obtained based on alignment approach. This approach correlates with the deleteriousness of a sequence variation (Choi et al., 2012; Choi and Chan, 2015). The combination used of multiple tools to predict the effect of AAS and InDels may increase the chance of identifying functional variants that had been missed by other tools.

However, although may never be accurate enough to replace the biological experiments, *in silico* predictions such as SNAP2 and PROVEAN can speed up the selection of potential non-synonymous SNPs that is predicted to be affecting the protein function and may ease the further work. Any information obtained from computational method to detect the presence of SNPs, insertions, and deletions and their effect to the protein function can

be used further in crop improvement programs, especially in term of the associations among SNPs and the traits of economic value.

Related to our study, the information about mungbean *MSH* genes characteristic can be utilized further to gain insight into the association between the genes and the mutation rate in mungbean. Mutagenesis experiments based on the information of the lack of Walker A at *MSH2* and partial HTH subdomain at *MSH3* using CRISPR or other genome editing technologies can be used to create mungbean genotype that high-acceptable to mutation exposure. Meanwhile any mungbean genotypes carrying the affected-non-synonymous SNPs can be evaluate as well for its responsiveness to mutation. Then based on this information, series of SNP markers can be developed to screen the mungbean germplsm collection that naturally brings the mutant gene of *MSH*. Eventually this may help the plant breeders in creating a better plant for the future.

REFERENCE

Allard RW (1999). Principles of Plant Breeding. 2nd ed. New York: John Willey and Sons.

Aslam M, Maqbool MA, Zaman QU, Latif MZ, and Ahmad RM (2013). Responses of mungbean genotypes to drought stress at early growth stages. *International Journal of Basic and Applied Sciences* 13(05):22-27.

Bennetzen JL (2000). Comparative sequence analysis of plant nuclear genomes: microlinearity and its many exceptions. *The Plant Cell* 12:1021-1029.

Bray CM and West CE (2005). DNA repair mechanisms in plants: crucial sensors and effectors for the maintenance of genome integrity. *New Phytologist* 168:511-528.

Bromberg Y and Rost B (2007). SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Research* 35(11): 3823-3835.

Choi Y and Chan AP (2015). PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* 31(16):2745-2747.

Choi Y, Sims GE, Murphy S, Miller JR, and Chan AP (2012). Predicting the functional effect of amino acid substitutions and indels. *PloS ONE* 7(10):e46688

Culligan KM and Hays J (2000). Arabidopsis MutS homologs-*AtMSH2*, *AtMSH3*, *AtMSH6*, and a novel *AtMSH7*-form three distinct protein heterodimers with different specificities for mismatch DNA. *The Plant Cell* 12:991-1002.

Culligan KM, Meyer-Gauen G, Lyons-Weiler J, and Hays J (2000). Evolutionary origin, diversification and specialization of eukaryotic MutS homolog mismatch repair proteins. *Nucleic Acids Research* 28(2):463-471.

Felsenstein J (1985). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* 39:783-791.

Gelvin SB and Schilperoort RA (1995). Plant Molecular Biology Manual. Norwell MA: Kluwer Academic Publisher,

Gupta PK, Roy JK, and Prasad M (2001). Single nucleotide polymorphisms: A new paradigm for molecular marker technology and DNA polymorphism detection with emphasis on their use in plants. *Current Science* 80(4):524-535.

Harten AM (1998). Mutation Breeding: Theory and Practical Applications. England: Cambridge University Press.

Iyer RR, Pluciennik A, Burdett V, and Modrich PL (2006). DNA mismatch repair: functions and mechanisms. *Chem. Rev.* 106:302-323.

Jones DT, Taylor WR, and Thornton JM (1992). The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences* 8:275-282.

Korzun V, Borner A, Siebert R, Malyshew S, Hilpert M, Kunze R, and Puchta H (1999). Chromosomal location and genetic mapping of the mismatch repair gene homologs *MSH2*, *MSH3*, and *MSH6* in rye and wheat. *Genome* 42:1255-1257.

Kunkel TA and Erie DA (2005). DNA mismatch repair. *Annu. Rev. Biochem.* 74:681-710.

Lestari P, Kim SK, Reflinur, Kang YJ, Nurwita D, and Lee SH (2014). Genetic diversity of mungbean (*Vigna radiata* L.) germplasm in Indonesia. *Plant Genetic Resources: Characterization and Utilization* 12(S1): S91-S94.

Li L, Jean M, and Belzile F (2006). The impact of sequence divergence and DNA mismatch repair on homeologous recombination in Arabidopsis. *The Plant Journal* 45:908-916.

Lodish H, Berk A, Kaiser CA, Krieger M, Bretscher A, Ploegh H, Amon A, and Scott MP (2013). Molecular Cell Biology. 7th ed. New York: WH Freeman and Co. 1154 p.

Obmolova G, Ban C, Hsieh P, and Yang W (2000). Crystal structures of mismatch repair protein MutS and its complex with a substrate DNA. *Nature article* 407: 703-710.

Reddy KS (2009). A new mutant for yellow mosaic virus resistance in mungbean (Vigna radiata (L.) Wilczek) variety SML-668 by recurrent gamma-ray irradiation. In: Q.Y. Shu, ed., Induced Plant Mutations in the Genomics Era. Food and Agriculture Organization of the United Nations, Rome, pp. 361-362.

Saitou N. and Nei M (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4:406-425.

Sancar A, Lindsey-Boltz LA, Unsal-Kacmaz K, and Linn S (2004). Molecular mechanisms of mammalian DNA repair and the DNA damage checkpoints. *Annual Review of Biochemistry* 73:39-85.

Sangiri C, Kaga A, Tomooka N, Vaughan D, and Srinives P (2007). Genetic diversity of the mungbean (*Vigna radiata*, Leguminosae) genepool on the basis of microsatellite analysis. *Australian Journal of Botany* 55:837-847.

Schofield MJ and Hsieh P (2003). DNA mismatch repair: molecular mechanisms and biological function. *Annu. Rev. Microbiol.* 57:579-608.

Senaratne R, Liyanage NDL, and Soper RJ (1995). Nitrogen fixation of and N transfer from cowpea, mungbean and groundnut when intercropped with maize. *Fertilizer Research* 40:41-48.

Shanmugasundaram S, Keatinge JDH, and Hughes J (2009a). Counting on beans: mungbean improvement in Asia. In: D.J. Spielman and R. Pandya-Lorch, eds., Millions Fed: Proven Successes in Agricultural Development. IFPRI, Washington DC.

Shanmugasundaram S, Keatinge JDH, Hughes J (2009b). The mungbean transformation: diversifying crops, defeating malnutrition. *IFPRI Discussion Paper 922*. International Food Policy Research Institute, Washington DC.

Supek F and Lehner B (2015). Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature Letter:*1-4.

Tachiki H, Kato R, Masui R, Hasegawa K, Itakura H, Fukuyama K, and Kuramitsu S (1998). Domain organization and functional analysis of *Thermus thermophilus* MutS protein. *Nucleid Acids Research* 26(18):4153-4159.

Tah PR (2006). Induced macromutation in mungbean [*Vigna radiata* (L.) Wilczek]. *International Journal of Botany* 2(3):219-228.

Tam SM, Samipak S, Britt A, and Chetelat RT (2009). Characterization and comparative sequence analysis of the DNA mismatch repair *MSH2* and *MSH7* genes from tomato. *Genetica* 137:341-354.

Tam SM, Hays JB, and Chetelat RT (2011). Effects of suppressing the DNA mismatch repair system on homeologous recombination in tomato. *Theor. Appl. Genet.* 123:1445-1458.

Tamura K, Stecher G, Peterson D, Filipski A, and Kumar S (2012). Mega6: Molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* 30(12):2725-2729.

Van K, Kang YJ, Han KS, Lee YH, Gwag JG, Moon JK, and Lee SH (2013). Genome-wide SNP discovery in mungbean by Illumina HiSeq. *Theor. Appl. Genet* 126:2017-2027.

Wu SY, Culligan K, Lamers M, and Hays J (2003). Dissimilar mispair-recognition spectra of Arabidopsis DNA-mismatch-repair proteins *MSH2-MSH6* and *MSH2-MSH7*. *Nucleic Acids Research* 31(20):6027-6034.

ABSTRACT IN KOREAN

녹두 DNA 불일치복구 유전자의 비교서열분석

ANDARI RISLIAWATI

초록

녹두(*Vigna radiata* (L.) R. Wilczek)는 질소고정을 통하여 토양 상태를 향상시킬 수 있는 고단백질 콩과 작물이다. 녹두는 남아시아 폭넓게 재배되고, 건조한 환경에서의 적응력이 뛰어난 가뭄 내성 작물이다. 그러나 전세계적으로 녹두의 생산량 증가는 정체되어 있고, 품종 개발은 이들의 낮은 유전적 다양성 때문에 제한되고 있다. 이러한 문제들을 극복하기 위해 유전자원 탐색(germplasm exploration), 돌연변이 유도와 같은 노력들이 있었지만, 만족스러운 결과는 얻지 못했다. 이는 *MSH* 유전자의 강한 발현으로 인하여 일어날 수 있다. *MSH* 유전자는 선행연구를 통해서 유전체의 온전성을 보존하기 위해 발현된다고 알려진 바 있다. 녹두에서의 *MSH* 유전자에 대해 더 탐색하기 위해, 우리는 8개의 나라에서 수집된 12개의 녹두 자원(종자)에서 *MSH* 유전자의 염기서열을 분석하였고, 쌍떡잎식물 14개 식물에서 70개의 *MSH* 유전자 서열과 비교하였다.

우리는 염색체 8번의 *MSH3*, 염색체 7번의 *MSH2*, 염색체 3번의 *MSH6*, 염색체 1번의 *MSH7*에서 DNA 불일치 복구와 관련된 *MSH* 유전자 상동유전자의 위치를 확인했다. *MSH2, MSH3, MSH6* 상동유전자에서 5개의 보존된 도메인이 모두 존재하는 반면, *MSH7*에서는 한 개의 도메인이 부족했다. 다른 종과 비교했을 때, 녹두는 *MSH2*에 있는 워커 A 모티프(Walker A motif)와 일부 HTH 서브도메인를 잃었다. 우리는 또한 단일염기 다형성(SNP)을 확인하였고, 녹두 유전자원(germplasm)의 도메인 I와 도메인 V사이의 아웃모티브에서 염기 결실을 확인하였다. 참고한 녹두와 비교했을 때, 염기 결실뿐만 아니라 두 개의 비동의

단일염기에서도 다른 단백질 기능을 예측되었다. 단일염기 다형성을 가지고 있는 녹두를 통해 돌연변이에 대한 민감성을 평가할 수 있고, 단일염기 다형성 마커는 원하는 대립유전자를 가지고 있는 적절한 유전자원을 가릴 수 있도록 설계할 수 있다. 그러므로, 녹두의 *MSH* 유전자 확인 연구는 돌연변이 유도를 통한 녹두 육종에 기여를 할 수 있을 것으로 생각되며, 이를 위해서는 추후, 유전체 삽입 등의 실험을 통해 기능적 연구가 더 필요할 것이다.

핵심단어: 녹두, *MSH*, 도메인(Domain), 모티프(Motifs), 단일염기 다형성(SNPs), 단백질 기능 예측

학번: 2014-22124

THESIS FOR DEGREE OF MASTER OF SCIENCE

Comparative Sequence Analysis of Mungbean DNA Mismatch Repair Genes

BY

ANDARI RISLIAWATI

FEBRUARY, 2016

MAJOR IN CROP SCIENCE AND BIOTECHNOLOGY

DEPARTMENT OF PLANT SCIENCE

THE GRADUATE SCHOOL OF SEOUL NATIONAL UNIVERSITY

Comparative Sequence Analysis of Mungbean DNA Mismatch Repair Genes

UNDER THE DIRECTION OF DR. SUK-HA LEE
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF SEOUL NATIONAL UNIVERSITY

BY
ANDARI RISLIAWATI

MAJOR IN CROP SCIENCE AND BIOTECHNOLOGY
DEPARTMENT OF PLANT SCIENCE

NOVEMBER, 2015

APPROVED AS A QUALIFIED THESIS OF ANDARI RISLIAWATI
FOR THE DEGREE OF MASTER
BY THE COMMITTEE MEMBERS

FEBRUARY, 2016

CHAIRMAN _____
Tae-Jin Yang, Ph.D.

VICE-CHAIRMAN _____
Suk-Ha Lee, Ph.D.

MEMBER _____
Hak Soo Seo, Ph.D.

Comparative Sequence Analysis of Mungbean DNA Mismatch Repair Genes

ANDARI RISLIAWATI

ABSTRACT

Mungbean (*Vigna radiata* (L.) R. Wilczek) is a high-protein grain legume that could improve soil condition through nitrogen fixation. It is grown widely in southern Asia and also a promising drought-tolerant crop due to its adaptation to dry environment. However, world mungbean production has been stagnant and its breeding progress is hampered by low genetic diversity. Some efforts have been done to overcome this problem, such as germplasm exploration and induced-mutation. However, none gave any satisfactory result. This could be caused by strong activity of *MSH* genes which has been reported to preserve genomic integrity in other species. To explore more about *MSH* genes in mungbean, we sequenced the *MSH* genes of 12 mungbean germplasm from 8 countries and compared the sequence to 70 *MSH* genes sequence from 14 species of eudicots clade.

We identified the location of *MSH* paralogs that involved in DNA mismatch repair, i.e. *MSH3* in chromosome 8, *MSH2* in chromosome 7, *MSH6* in chromosome 3, and *MSH7* in chromosome 1. All five conserved domains exist in *MSH2*, *MSH3*, and *MSH6* paralogs, whereas *MSH7* lacks one domain. Compare to other species, mungbean lost the Walker A motif at *MSH2* and partial of HTH subdomain at *MSH3*. We also identified 3

synonymous SNPs, 4 non-synonymous SNPs, and 1 deletion among mungbean germplasm at neighbored-motifs of domain I and domain V at *MSH2*, *MSH3* and *MSH6*. Two non-synonymous SNPs at *MSH6* and one deletion at *MSH2* are predicted having different protein function compare to the mungbean reference. This prediction can be tested further through genome editing technology to support the mutagenesis experiment in creating breeding materials that high-acceptable to mutation exposure. Mungbean accessions carrying the SNPs can be evaluated as well for its responsiveness to mutation and based on that, SNP markers can be designed to screen appropriate germplasm carrying favorable allele. Therefore the identification of *MSH* genes in mungbean may contribute to the mungbean genetic potential improvement toward induced-mutation.

Keywords: Mungbean, *MSH*, Domain, Motifs, SNPs, Protein function prediction

Student number: 2014-22124

CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Mungbean (*Vigna radiata* (L.) Wilczek) is an important grain legume and grown widely in developing countries, particularly in Southern Asia countries (Shanmugasundaram et al., 2009a). Mungbean is not only cultivated for its high protein seed content but also can be utilized as fodder that contributes to soil fertility (Senaratne et al., 1995; Shanmugasundaram et al., 2009b). Compare to other crops, mungbean can adapt with severe environment such as water limitation. Therefore, mungbean is considered as one promising drought-tolerant crop in the future (Aslam et al., 2013).

Despite its importance, world mungbean production has been stagnant due to some agronomic shortage such as low yield and susceptibility to diseases and insects (Shanmugasundaram et al., 2009a). Plant breeding as a knowledge and art to improve the heritable genetic of a particular trait in a plant, is used to cope these shortages (Allard, 1999). Unfortunately mungbean germplasm lacks of diversity, which is necessary for successful plant breeding research (Lestari et al., 2014; Sangiri et al., 2007).

The genetic diversity of mungbean germplasm can be increased through induced-mutation using a chemical mutagenic agent like

ethylmethane sulphonate (EMS) or a physical mutagenic agent like Gamma-ray Irradiation (Harten, 1998). However, the sudden changes in DNA due to mutagenesis can be unfavorable because they occur randomly within the genome and reduce the seed fertility of the next generation (Tah, 2006). In the assembly of mungbean mutant cultivar resistant to Yellow Mosaic Virus (YMV), approximately only 10 percent of mutants found among 2500-3000 plants grown in every generation, but none of them showed resistance to YMV. However after re-mutation of the 3rd generation (M3), YMV mutant was obtained and still need to be homogenized until 6th generation (M6). Thus, practically mutation breeding in mungbean is laborious and time consuming because large number of population is needed for the starting point of the research (Reddy, 2009).

The unfavorable mutation effect at the molecular level that can damage and change the normal DNA sequence of an organism may occur due to the failure of DNA repair mechanism. There are several DNA repair mechanisms in a cell of living organism. One of them is known as mismatch repair (MMR). This mechanism produces a protein that corrects DNA mismatches during DNA replication, homologous recombination (HR) or as a result of DNA damage caused by mutagenic agent. The understanding of MMR system is based on the *in-vitro* reconstitution of purified MMR

protein of prokaryotic organism, *Escherichia coli,* in which three genes involved, namely *MutS*, *MutL*, and *MutH* genes. In eukaryotic organism these genes have homologs and not all homologs have role in the MMR system. As reported in *Arabidopsis* plant, only four *MutS* homolog (*MSH*) involved in MMR system and worked as heterodimers, i.e. *MSH2-MSH3 (MutSβ), MSH2-MSH6 (MutSα),* and *MSH2-MSH7 (MutSY*). Each of them has particular mechanism in recognizing the DNA mismatch within the genome (Culligan et al., 2000; Schofield and Hsieh, 2003; Kunkel and Erie, 2005; Iyer et al., 2006).

Considering the importance of MMR in DNA repair mechanism and its relation to mutation activity, some researchers have reported the effect of MMR disruption. According to Schofield and Hsieh (2003), the gene deficient in MMR could lead to the increases of spontaneous mutation due to the frequent exhibit of microsatellite instability at mono- and di-nucleotide repeats. Study on tomato and arabidopsis showed that the crop has complete homologs of *MSH* genes and suppression of these genes also increased the HR which leads the acceleration of wild cultivar introgression. In case of tomato, this crop also has low genetic diversity in nature (Li et al., 2005; Tam et al., 2009; Tam et al., 2011). Another study of MMR

inactivation on human cancer genome caused a large scale regional mutation rate variation as well (Supek and Lehner, 2015).

As we proposed earlier, mungbean has problems in low genetic diversity and low mutation variation as those in tomato. By correlating these facts, we assume that MMR mechanisms, particularly the presence of *MSH* genes possibly correlated with the mutation behavior in mungbean. Therefore in this study we hypothesized that the *MSH* genes exist in mungbean and in a complete homologs form. To verify this hypothesis, we characterized the mungbean *MSH* genes by identifying its location within the mungbean genome. Since the whole genome mungbean reference is available, we used comparative sequence analysis and include several germplasm from various countries to explore any DNA variation of the genes among accessions as well as the prediction of protein alteration among germplasm. Result of this study may facilitate further work such as gene manipulation or mutant detection in the early stage of mungbean plant. Thus, is expected to improve the breeding efficiency of mungbean.

LITERATURE REVIEW

DNA Repair Mechanisms in Plant

During the lifetime, DNA of any organism including plant can be damaged by spontaneous cleavage of chemical bonds in DNA, by environmental agents such as ultraviolet and ionizing radiation, and by reaction with genotoxic chemical that are by-products of normal cellular metabolism or occur in the environment. This damage can cause a mutation, a change in the normal DNA sequence. The mutation if left uncorrected and accumulate within the cell, may cause no longer function of the cell and unable to produce viable offspring. Thus the prevention of DNA sequence errors in all types of cells is important for survival and several cellular mechanisms for repairing damaged DNA and correcting sequence errors have evolved.

The first line of defense in preventing mutations is the proofreading activity of DNA polymerase. In prokaryotes for instance, during their DNA replication, 1 incorrect nucleotide per 10^4 polymerized nucleotides may occur. To correct this error, DNA polymerase through the exonuclease activity pause the replication and transfers the 3'end of the growing chain to its exonuclease site where the incorrect mispaired base is removed. Then 3'

end is transferred back to the polymerase site, where this region is copied correctly (Lodish et al., 2013).

In addition to proofreading activity, cells have other repair systems for preventing mutations, i.e. base excision repairs (BER), nucleotide excision repair (NER), double strand break repair (DSBR), and mismatch repair (MMR). The BER is mainly caused by chemical mutagenic agent such as ethyl methane sulfonate (EMS) which cause base modification of a mutated G-A. The repair pathway is initiated by removal of the damaged base by a DNA glycosylase enzyme which results in cleaving of AP site (3' side of the abasic) by AP endonuclease. This cleaved site then becomes the substrate for the SSB repair pathway through short-patch or long-patch repair mechanism (Bray and West, 2005).

In contrast to BER, NER can detect modifications indirectly by conformational changes to the DNA duplex rather than relying on the recognition of specific DNA damage products. NER targets the damaged strand and removes a 24-32 base oligonucleotide containing the damaged product. DNA synthesis and ligation completes the repair process (Sancar et al., 2004).

Meanwhile the MMR complements the activity of DNA polymerase proofreading activity in order to maintain the genomic integrity. MMR may

also have an important role in recognizing mismatches at sites of recombination between DNA sequences, thereby reducing the rate of occurrence of recombination events which might lead to inappropriate chromosome rearrangements of interspecies hybridization (Wu et al., 2003). MMR in prokaryote is performed by MutHLS system, whereby *Mut*S homodimers recognize and bind to insertion/deletion loops (1-4 bp) and repair mismatch. In the presence of ATP, *MutS* recruits MutL (an ATPase) and activates *MutH* (methylation sensitive endonuclease) that cleaves the transiently unmethylated DNA strand, targeting MMR to newly synthesized DNA strand. Prokaryotes have two homologs namely *MutS1*, work for MMR which is described before, and *MutS*2 which involves in meiotic crossing over and chromosome segregation. In eukaryotes, homologs of *MutS* and *MutL* have both found, but not *MutH*. Homologs of *MutS* in eukaryote namely *MSH1* to *MSH7*, with *MSH7* is being unique to plant. Whereas homologs of *MutL* in eukaryote, namely *MLH1*, *MLH2* or *hPMS1*, *MLH3*, and *PMS1* or *hPMS2*. Heterodimers of *MSH* protein in eukaryote provide substrate specificity, i.e. *MutSα* (*MSH2-MSH6*) which repairs base-base mismatch, *MutSβ* (*MSH2-MSH3*) which repairs +1 insertion/deletion loops (IDLs) and larger loops of 2-8 bp, and *MutSϒ* (*MSH2-MSH7*) which repairs G/T mismatch. While *MSH1* is required for mitochondrial stability

and *MSH4-MSH5* function in meiosis and involve in resolution of Holliday junctions during meiosis (Obmolova et al., 2000; Schofield and Hsieh, 2003; and Kunkel and Erie, 2005).

Unlike NER, BER, and MMR which repair the error of single strand DNA, the DSBR repair the double-strand breaks in DNA (dsDNA). These are particularly severe lesions because incorrect rejoining of dsDNA can lead to gross chromosomal rearrangements that can affect the functioning of genes. The DSBR is mainly caused by the activity of nuclease such as *HindII*, *EcoRI*, and *FokI*. This enzyme may capable to cleave phosphodiester bonds between the nucleotide subunits of nucleic acids. Two systems have evolved in DSBR, i.e. homologous recombination (HR) and non-homologous end-joining (NHEJ). HR uses an identical or very similar DNA sequence as a template for the repair of a DSB, while NHEJ recombines DNA largely independent of the sequence (Bray and West, 2005; Lodish et al., 2013).

Comparative Sequence and Phylogenetic Analysis

It is well known that plant genomes tend to be large and complex, thus made very diverse in growth habit and environmental adaptation. Despite this diversity, plant geneticists have found that plants exhibit

extensive conservation of both gene content and gene order. On the other hand, the advent of DNA marker and sequencing technology not only facilitated the rapid generation of detailed plant genetic maps but also allowed map comparisons among species. The comparison between closely related species indicated extensive collinearity of genetic maps. Many comparative studies also show that within the limits of sequence divergence that permit cross-hybridization, the large majority of plant genes have close homologs within most other plant genomes. This means that different plant species often use homologous genes for very similar functions. This becomes the basis of the comparative sequence analyses which commonly applied in the reverse genetic approach (Bennetzen, 2000).

In relation with the phylogenetic analyses, comparative method is applied to gain insight the historical relationships of lineages based on evolutionary hypotheses. Moreover, it is known that differences and similarities among species are the basis of phylogenetic analyses thus made the comparative sequence and phylogenetic study are closely related each other. However, building hypotheses about the evolutionary history of species is a challenging task, as it requires knowledge about the underlying methodology and an ability to flexibly manipulate data in diverse formats. Although most practitioners are not experts in phylogenetic, the appropriate

handling of phylogenetic information is crucial for making evolutionary inferences in comparative study.

Single Nucleotide Polymorphism (SNP)

A single nucleotide polymorphism (SNP) is a variation in a single nucleotide which may occur at some specific position in the genome, where each variation is present to some appreciable degree within a population. SNPs may fall within the coding sequences of genes, non-coding regions of genes, or in the intergenic regions. SNPs in the coding region are of two types, synonymous and nonsynonymous SNPs. Synonymous SNPs do not affect the protein sequence while nonsynonymous SNPs change the amino acid sequence of protein. The nonsynonymous SNPs are of two types, i.e. missense and nonsense. SNPs that are not in protein-coding regions may still affect gene splicing, transcription factor binding, messenger RNA degradation, or the sequence of non-coding RNA. Gene expression affected by this type of SNP is referred to as an *eSNP* (expression SNP) and may be upstream or downstream from the gene.

There are several methods applied for discovery and identification of new SNPs, i.e. (1) locus specific-PCR amplification, (2) alignment among available genomic sequences, (3) whole genome shotgun sequences, (4)

overlapping regions in BACs and PACs, and (5) reduced representation shotgun (RRS). The first and the second methods can be used only for genomic regions with known sequences since prior sequence information is necessary. In the third method, several fold coverage of the whole genome is required before SNPs can be detected by alignment of sequences belonging to the same locus. The fourth one is the common methods for SNPs detection by a mismatch that have been used for genome sequencing. Whereas the last method is used when the genomic sequences may not be available or it may not be desirable to use the available genomic sequences for the discovery of SNPs. This approach uses subsets of genome, each containing manageable number of loci to permit resampling (Gupta et al., 2001).

In Silico Prediction of Protein Function Changes

Many genetic variations are SNPs which can be in the form of synonymous and non-synonymous SNPs. Non-synonymous SNPs are neutral if the resulting point-mutated protein is not functionally visible from the wild type and non-neutral otherwise. The *in silico* prediction of the effect from non-synonymous SNPs are developed recently which have given a great contribution to the efficiency of genomic study. SNAP2 and

PROVEAN are some example of the *in silico* prediction beside other popular tool such as SIFT and Polyphen-2.

SNAP (Screening for non-acceptable polymorphisms) is based on neural network that predicts the functional effects of mutations by distinguishing between effect and neutral variants of non-synonymous SNPs. The most important input signal for the prediction is the evolutionary information taken from an automatically generated multiple sequence alignment. Structural features such as predicted secondary structure and solvent accessibility are considered as well. If available also annotation (i.e. known functional residues, pattern, regions) of the sequence or close homologs are pulled in. In a cross-validation over 100,000 experimentally annotated variants, SNAP2 reached sustained two-state accuracy (effect/neutral) of 82% (at an AUC of 0.9) (Bromberg and Rost, 2007).

Contrast with other *in silico* program, PROVEAN (Protein Variation Effect Analyzer) not only predicts the effect of amino acid substitution but also an insertion and deletion as well. In PROVEAN, a delta alignment score is computed for each supporting sequence. The scores are then averaged within and across clusters to generate the final PROVEAN score. If the PROVEAN score is equal to or below a predefined threshold (e.g. -2.5), the protein variant is predicted to have a "deleterious" effect. If the

PROVEAN score is above the threshold, the variant is predicted to have a "neutral" effect. This score based on the reference and variant versions of a protein query sequence with respect to sequence homologs collected from the NCBI NR protein database through BLAST. Compare to SIFT and Polyphen-2, the prediction results by PROVEAN is in agreement and shared by all about 78.5% (15,618/19,898) of disease-associated variants and 46.8% (16,244/34,701) common variants (Choi et. al., 2012; Choi and Chan, 2015).

MATERIALS AND METHODS

Identification of Mungbean *MSH* Location

NCBI database search was performed to find previous identified and potential *MSH* family genes in the model plant, *Arabidopsis thaliana*. We used "*MSH1, MSH2, MSH3, MSH4, MSH5, MSH6*, and *MSH7*" as a query to search the protein/amino acid sequences of *MSH* homologs that involved in MMR. We aligned the longest sequence and *RefSeq* type from arabidopsis *MSH* homologs (*AtMSH*) to the mungbean whole-genome reference which was assembled by Van et al. (2013) through the SNU's Crop Genomics Laboratory homepage (http://plantgenomics.snu.ac.kr/sequenceserver). The most matched sequence/ GeneID was selected as the gene sequence for each *MSH* homologs (*VrMSH*).

We also identified the domain within *MSH* genes by analyzing the *VrMSH* genes into the integrated protein signature databases (InterPro) database (http://www.ebi.ac.uk/interpro/). The domain identification is needed in further analysis.

Phylogenetic and Multiple Alignment Analysis of *MSH* Genes among Species

We performed the NCBI database search to obtain *MSH* protein sequence of 14 species which cover several order. These species were distributed randomly under eudicots clade and analyzed together with *MSH* mungbean sequences in phylogenetic and multiple alignments analysis (Table 1).

The phylogenetic analysis was performed by MEGA6 software using Neighbor-Joining (NJ) method which is supported by bootstrap 1000 replications The distance matrices for specific groups of *MSH* protein sequences were computed based on the Jones-Taylor-Thornton (J-T-T) model (Felsenstein, 1985; Saitou and Nei, 1987; Jones et al., 1992; Tamura et al., 2013). Whereas the multiple alignment and synteny analysis was performed by MEGA6 software as well using ClustalW program with default values for gap opening (10), extension (0.2) penalties and the GONNET 250 protein similarity matrix.

Table 1. List of *MSH* protein sequences used in phylogenetic and multiple alignment analysis

Crop/Clade	*MSH1*	*MSH2*	*MSH3*	*MSH4*	*MSH5*	*MSH6*	*MSH7*
	NCBI ID (prot. length)	NCBI ID (prot. length)	NCBI ID (prot. length)	NCBI ID (prot. length)	NCBI ID (prot. length)	NCBI ID (prot. length)	NCBI ID (prot. length)
Medicago/ Legumes	-	-	-	-	-	-	gi357500449 (1160 aa)
Chickpea/ Legumes	gi502122529 (1141 aa)	gi502151706 (942 aa)	-	-	gi502099189 (809 aa)	-	gi502163835 (1098 aa)
Soybean/ Legumes	gi356575134 (1134 aa)	gi356563103 (942 aa)	-	-	gi571506599 (812 aa)	-	gi571478271 (1079 aa)
Strawberry/ Fabids	gi764569327 (1141 aa)	gi470126534 (942 aa)	gi70144922 (1106 aa)	gi470130586 (792 aa)	gi470119462 (809 aa)	gi764592252 (1252 aa)	gi764505215 (1075 aa)
Cucumber/ Fabids	gi778678067 (1152 aa)	gi778656285 (942 aa)	gi778679553 (1110 aa)	gi778708277 (789 aa)	gi449463733 (807 aa)	gi449436747 (1307 aa)	gi449443325 (1095 aa)
Jatropha/ Malphigiales	gi802633693 (1146 aa)	gi802797191 (936 aa)	gi802769131 (1105 aa)	gi802582003 (792 aa)	gi802588379 (807 aa)	gi802689824 (1304 aa)	gi802627380 (1108 aa)

Poplar/ Malphigiales	-	gi222858604 (944 aa)	-	-	-	-	-
Arabidopsis/ Malvids	gi75297828 (1118 aa)	gi42565226 (937 aa)	gi30686920 (1081 aa)	gi79476962 (792 aa)	gi186510260 (807 aa)	gi332656719 (1324 aa)	gi12643849 (1109 aa)
Cacao/ Malvids	-	gi508773672 (967 aa)	gi508775913 (1115 aa)	-	gi508720408 (818 aa)	-	-
Brassica rapa/ Malvids	gi685328741 (1122)	gi685289267 (937)	gi685293250 (1098 aa)	gi685343968 (792 aa)	gi685310043 (807 aa)	gi685287036 (1337 aa)	gi685263673 (1101 aa)
Eucalyptus/ Myrtales	gi702441803 (1152 aa)	gi702259274 (942 aa)	-	gi702336056 (790 aa)	gi702363375 (807 aa)	gi702500679 (1318 aa)	gi702305220 (1083 aa)
Grape/ Vitales	gi225433289 (1144 aa)	gi731426269 (945 aa)	gi731423415 (1111 aa)	-	gi731432937 (872 aa)	gi225437545 (1297 aa)	gi731406967 (1105 aa)
Tomato/ Asterids	gi460404638 (1137 aa)	gi350538025 (943 aa)	gi723679590 (1119 aa)	gi723719921 (792 aa)	gi723735564 (834 aa)	-	gi723713547 (1082 aa)
Potato/ Asterids	gi565347746 (1137 aa)	gi565376482 (943 aa)	-	-	gi565343547 (831 aa)	-	gi565348531 (1078 aa)

Variation Identification of *MSH* Genes among Mungbean Accessions

Twelve accessions from the mungbean germplasm collection (Table 2) were selected based on the genetic diversity analysis from the previous study (Sangiri et al., 2007; Lestari et al., 2014). Germplasm from region which have higher diversity content was selected more than other regions. Thus, the genetic diversity of chosen germplasm will be as similar as possible to the actual genetic diversity of entire collection. We used CTAB methods to extract the DNA from the young leaves of the chosen germplasm (Gelvin and Schilperoort, 1995). The DNA quality and quantity was observed and measured by agarose gel electrophoresis and NanoDrop platform, respectively.

Table 2. List of mungbean germplasm used in the study

ID	Name	Country of Origin	ID	Name	Country of Origin
V1	JP2291819	India	V7	Tecer Hijau	Indonesia
V2	JP229177	India	V8	Utang Wewe	Indonesia
V3	JP229193	India	V9	JP78939	Vietnam
V4	JP229130	Bangladesh	V10	JP229096	Thailand
V5	JP81649	Srilanka	V11	Sunhwanokdu	South Korea
V6	JP99066	Pakistan	V12	Gyonggijere5	South Korea

Since the domain I and V of *MSH* genes play important role in the MMR, we developed primers using *Primer3* software which flank the important motifs within those domains. The location of the targeted domains and motifs were obtained from previous analysis. We used these primers to amplify the DNA of 12 chosen mungbean germplasm. PCR conditions were: one cycle of 94^0C for 5 min; then 35 cycles of 94^0C denaturation for 30s, 60^0C for 30s-45s, and 72^0C for 30s; with a final extension cycle of 72^0C for 5 min. PCR products were visualized by agarose gel electrophoresis and were sequenced by NICEM sequencing facility (http://nicem.snu.ac.kr). Sequence files then were manually edited and aligned using MEGA6 software.

Afterward we aligned the *MSH* genes sequences of 12 mungbean accessions and identified any occurrence of the single nucleotide polymorphisms (SNPs) which lay within the coding regions of the genes. We focused the observation around the important motifs within the domain I and V of the mungbean *MSH2*, *MSH3*, *MSH6*, and *MSH7* because only these homologs related to the MMR. Then based on the SNPs found, we computationally predicted the effect of non-synonymous SNPs to the alteration of protein function.

The prediction was performed by SNAP2 program which can be accessed online at https://rostlab.org/services/snap/. SNAP2 is based on neural network that predicts the functional efects of mutations by distinguishing between effect and neutral variants of non-synonymous SNPs (Bromberg and Rost, 2007). Since the SNAP2 only predict the amino acid substitution (AAS) we perfomed another *in silico* prediction to predict the effect of insert and deletions (InDels), i.e the PROVEAN which can be accessed at http://provean.jcvi.org/seq_submit.php. PROVEAN (Protein Variation Effect Analyzer) is a software tool which predicts wether an AAS or InDels has an impact on the biological function of protein. This computation is comparable to popular tools such as SIFT (Sorting Tolerant from Intolerant) or PolyPhen-2 (Choi and Chan, 2015).

.

RESULTS

The Characteristics of Mungbean *MSH* Genes

The alignment of Arabidopsis *MSH* proteins to the whole genome sequence of Korean mungbean cultivar (Suhnwanokdu) resulting in seven most matched gene ID within the mungbean genome. Four of them were detected matched to more than one *MSH* homologs. However, we defined the gene for each homologs based on the highest E-value in corresponding homologs. Therefore the location of *MSH* homologs genes in the mungbean genome are *MSH1* and *MSH3* in chromosome 8, *MSH2* in chromosome 7, *MSH4* in chromosome 11, *MSH5* in chromosome 6, *MSH6* in chromosome 3, and *MSH7* in chromosome 1. Analysis of these protein sequences using BLASTp into NCBI database shows that all homologs of mungbean *MSH* that involve in mismatch repair are most similar to soybean *MSH* protein. The levels of identity are 74% for *MSH1*, 89% for *MSH2*, 89% for *MSH3*, 80% for *MSH5*, 98% for *MSH6*, and 96% for *MSH7*.

For the four homologs of *MSH* genes that related to the MMR, analysis of their *MSH* protein sequences into InterPro database indicates that the sequences are likely to be functional homologs of the DNA mismatch repair proteins. Multiple significant hits from Pfam, SMART, Superfamily,

and PANTHER database were detected and showing that the sequences contain the conserved domains and motifs recognizable for *MutS*/*MSH* protein. Based on the Pfam database, we identify five domains in *MSH2*, *MSH3*, and *MSH6*, while *MSH7* only has four domains. The length of these genes are 8211 – 9322 base pairs with the *MSH2* as the shortest and *MSH3* as the longest genes. However, *MSH6* has longer protein sequence than *MSH3* although its nucleotide is shorter (Figure 1).

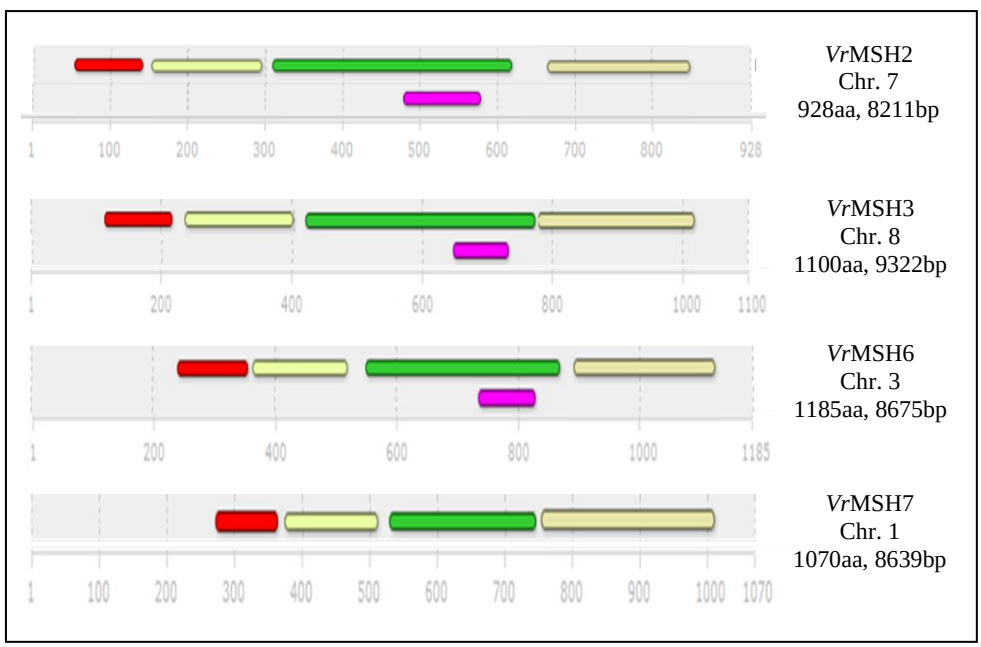


Figure 1. Domain structure of *VrMSH* genes. Color line: red=domain I, yellow=domain II, green=domain III, purple=domain IV, light brown=domain V

Phylogeny and Comparison of *MSH* Protein among Crop Species

The evolutionary history among *MSH* genes in several species of eukaryotes was inferred using the Neighbor-Joining method with 1000 replication of bootstrap analysis and involved 77 amino acids sequences of full length *MSH* protein from 15 species. The phylogenetic tree shows clearly separation of seven homologs of *MSH* genes from *MSH1* to *MSH7* (Figure 2). *MSH1* homolog is the deepest branch within the cluster which is supported with high bootstrap value of 100. Following this, three main groups are identified. The first consists of *MSH6* and *MSH 7* (99% bootstrap value); the second group consists of *MSH2* and *MSH5* (50% bootstrap value); and the third group consists of *MSH3* and *MSH4* (56% bootstrap value).

Meanwhile the mungbean *MSH* genes also resolve clearly within their respective protein groups (Figure 2). Mungbean *MSH2* and *MSH7* are sister to other legumes *MSH2* and *MSH7* (soybean, chickpea, and medicago), all with strongly supported bootstrap values (100%). Contrary with that, mungbean *MSH3* and *MSH6* are grouped separately with other legumes. The pattern of phylogenetic tree can be used as alignment basis to identify protein sequences variation between mungbean *MSH* and their orthologues in closest species.

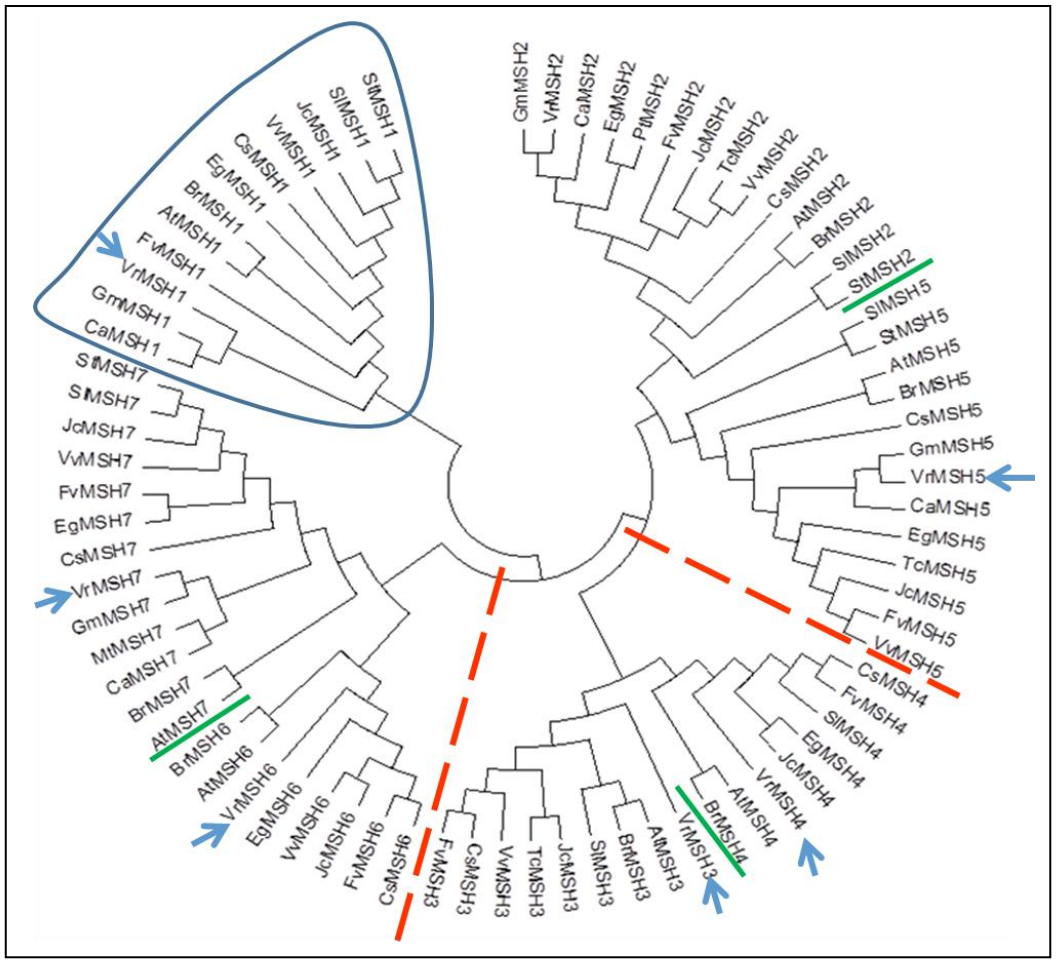
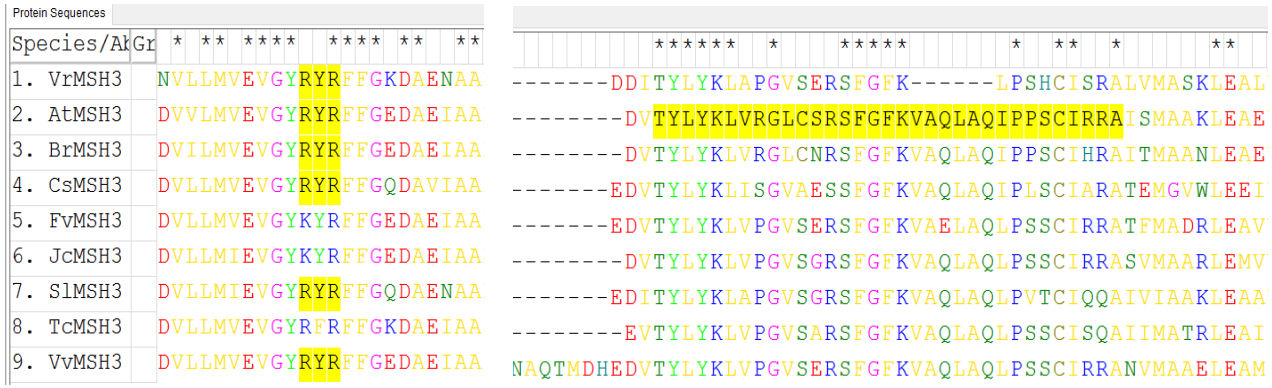


Figure 2. Evolutionary relationships of *MSH* genes in eukaryotes.

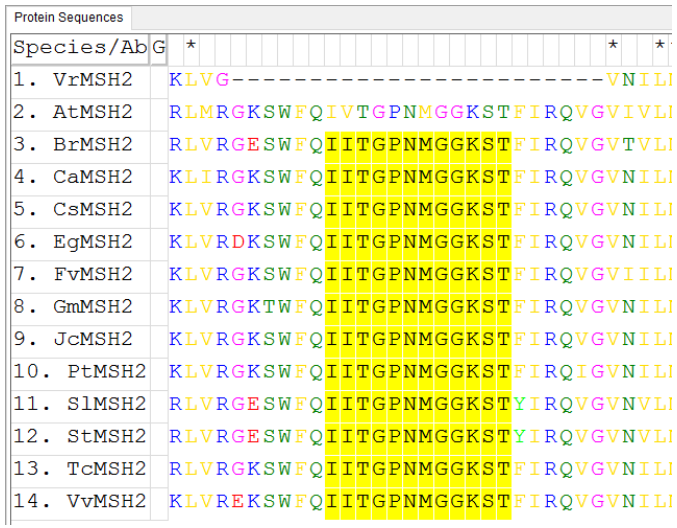
Species abbreviation:
At=*Arabidopsis thaliana*, Br=*Brassica rapa*, Ca=*Cicer arietinum*, Cs=*Cucumis sativus*, Eg=*Eucalyptus grandis*, Fv=*Fragaria vesca*, Gm=*Glycine max*, Jc=*Jatropha curcas*, Mt=*Medicago truncatula*, Pt=*Populus trichocarpa*, Tc=*Theobroma cacao*, Sl=*Solanum lycopersicum*, St=*Solanum tuberosum*, Vr=*Vigna radiata*, Vv=*Vitis vinifera*

We also conducted multiple alignments of *MSH* proteins for *MSH2*, *MSH3*, *MSH6*, and *MSH7*. From the alignment we identified amino acids variation among species, especially in the neighbor-motif within domain I and domain V. In domain I, we identify FYE motif both in *MSH6* and *MSH7* for all species. Another recognition motif of MFE in *MSH2* is identified as well in all species. However, for *MSH3* we detect varies of recognition motifs which include RYR in mungbean, arabidopsis, brassica, chickpea, tomato, and grape; KYR in strawberry and jatropha; and RFR only in cacao (Figure 3a).

Five important motifs that involve in ATP hydrolysis are well known at domain V of *MSH* genes, i.e. Walker A, Motif C, Walker B, Motif D, and HTH subdomain. In our multiple alignment analysis, Walker A motif is absence in mungbean *MSH2* contrast with other species (Figure 3b). Mungbean *MSH*3 also loses its six residues within the HTH subdomain, although the specific motif of YGA still remains (Figure 3a). The HTH subdomain has YGA residues in which each residue is separated by 4 and 23 residues respectively.



(a)



(b)

Figure 3.

(a) Multiple alignment analysis of domain I and HTH subdomain domain V at *MSH3* protein sequences from nine species

(b) Multiple alignment analysis of Walker A domain V at *MSH2* protein sequences from 14 species.

Same color of amino acid denotes same biochemical group. Dashes denote gaps. Important motifs are denoted by yellow boxes.

MSH Genes Variation among Mungbean Accessions

To identify the variation of *MSH* protein sequences among mungbean accessions, we developed primers in domain I and V of the genes since they are play importance role in MMR (Table 3). Based on these primers, we found some single nucleotide polymorphisms (SNPs) within the coding region of the genes. Total of seven SNPs are identified at *MSH2*, *MSH3*, and *MSH6*. Four of them are categorized as non-synonymous SNPs. These non-synonymous SNPs are found in mungbean accessions from India, Srilanka, Pakistan, and Indonesia. We also found another one deletion at *MSH2* from Indonesia mungbean accession (Table 4).

Since only non-synonymous SNPs that can alter the composition of amino acids and probably the protein function as well, we performed *in silico* prediction to predict this changes effect. The prediction was carried out by the SNAP2 program which has score range from -100 to -50 (neutral effect), >-50 to 50 (weak effect), and >50 to 100 (strong effect) to the protein function changes. From 4 non-synonymous SNPs found, 2 of them have neutral effect and the other two are predicted changing the protein function compare to the reference (Figure 4).

Table 3. Motifs, locations, and primers designed for Domain I and V of *MSH* homologs gene related to MMR

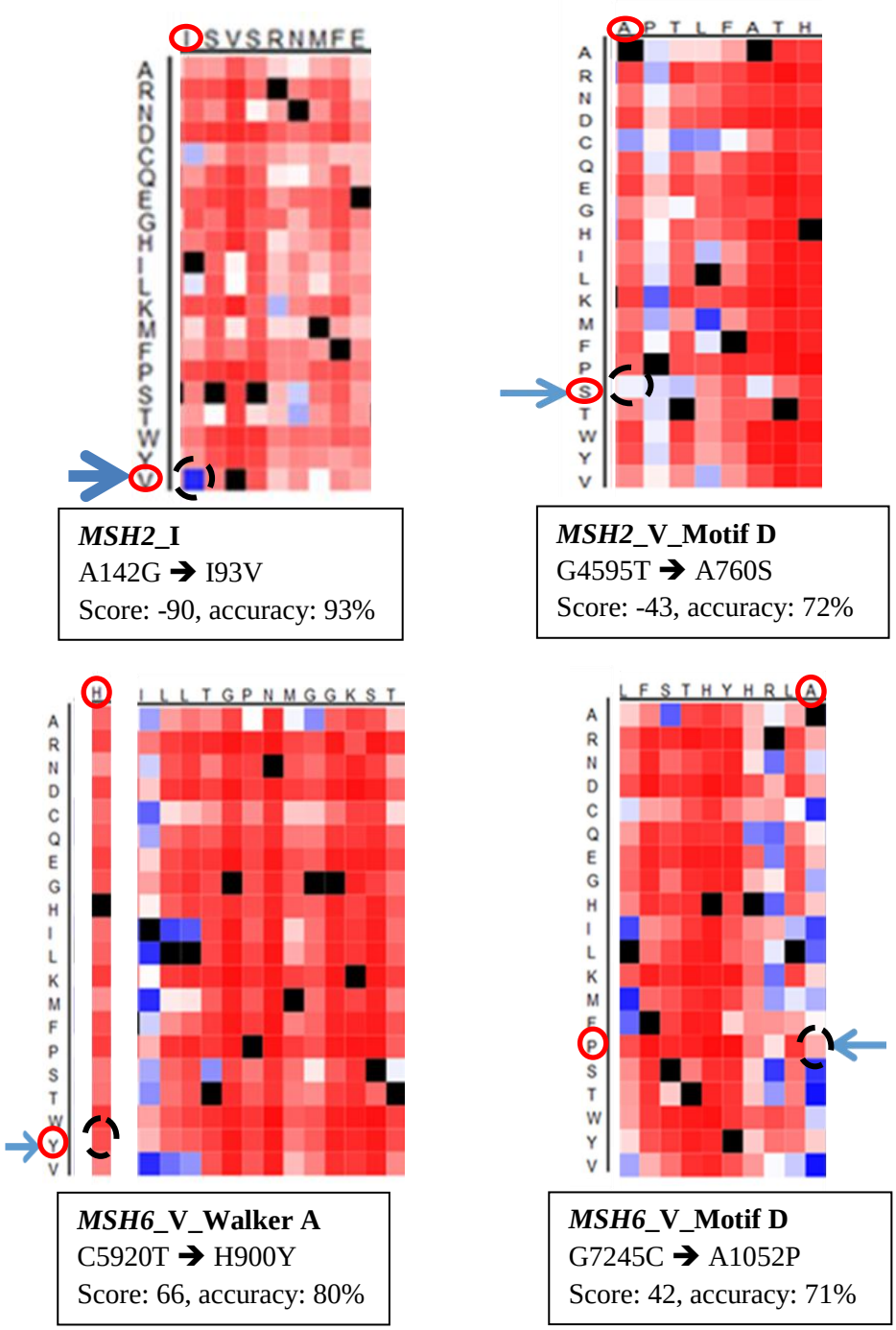
Homolog_Domain	Motif	Forward primer	Reverse primer
*MSH2*_I	MFE	ATGGCGACAATGCAACTTTC	GCGTTCCACTTTTGACCAGT
*MSH2*_V	Motif C, Walker B	ATTCTCCCCAGCTACGTGGT	GAAGAAGCCATTGTACAGGTCA
*MSH2*_V	Motif D	CAATGGTGGCATTGGTGTAA	CAAGGGCTAAAGCAGTCAGC
*MSH3*_I	RYR	CAGGAACCTTCTTCCCCTTC	GTGGGCGTAAATGCCTAAGA
*MSH3*_V	Walker A	ATCTGAATGCCCCACTTTCA	GACACCGCATTGGATCTACC
*MSH3*_V	Motif C, Walker B	CTGCACGTCCTGGATAGGAT	CAAGCTGGCAATCTTTGGAT
*MSH3*_V	Motif D	ATGAGCTTGGGAGAGGAACA	CTGGGCAACCTTAAATCCAA
*MSH6*_I	FYE	CCACAATGAGGTTGGTCTCC	GTCCATTCCTTCCAACCAAA
*MSH6*_V	Walker A	GCCAGAATCACAGTCAAGCA	ATGAAGGACCAACATGCACA
*MSH6*_V	Motif C, Walker B	ACCTTCCGCACAAAATGTTC	GAATGGGGGCCAAAGATAAT
*MSH6*_V	Motif D	GGAATACCTTGGGATCGTTG	GGGACTGCAACTTCTGATGG
*MSH7*_I	FYE	ATGCCGCAATTAATGGTCAA	CATCATCAATCCCACTTTCAGA
*MSH7*_V	Walker A	TGACACTGGAGGAACTGTGC	GAGAAGAAACCTGGGCCATA
*MSH7*_V	Motif C, Walker B	CACGACTTGGAGCCAAAGAT	AATGGCGTAGCCATCAAAAG
*MSH7*_V	Motif D	TTTGGTCCCGAGCATTTTTA	CATTGTAACGCGTGGATGAG

Table 4. SNPs and InDels found around important motif of Domain I and V

Type	*MSH_* domain	Genotype	SNP position (3'..5')		Motif
			DNA	Amino acid	
Synonymous SNPs	*MSH3*_I	V1,V3,V6,V8	C420T*	A140A	RYR
	*MSH6*_V	V1	G7039A	G1023G	Walker B
	*MSH6*_V	V5	C7241T	R1050R	Motif D
Non-synonymous SNPs	*MSH2*_I	V1,V3,V6,V8	A142G**	I93V	MFE
	*MSH2*_V	V7	G4595T	A760S	Motif D
	*MSH6*_V	V5	C5920T	H900Y	Walker A
	*MSH6*_V	V2	G7245C	A1052P	Motif D
Deletion	*MSH2*_I	V7	A1525del	M80del	1bp deletion

*C420T ➔ A140A: SNP is found at nucleotide position of 420 where Cytosine (C) is replaced by Thymine (T) and do not change the resulting amino acid at position 140

** A142G ➔ I93V: SNP is found at nucleotide position of 142 where Adenosine (A) is replaced by Guanosine (G), thus resulting in amino acid changes at position 93 from Isoleucine (I) to Valine (V).



***MSH2*_I**
A142G ➔ I93V
Score: -90, accuracy: 93%

***MSH2*_V_Motif D**
G4595T ➔ A760S
Score: -43, accuracy: 72%

***MSH6*_V_Walker A**
C5920T ➔ H900Y
Score: 66, accuracy: 80%

***MSH6*_V_Motif D**
G7245C ➔ A1052P
Score: 42, accuracy: 71%

Figure 4. Prediction of protein function changes of non-synonymous SNPs at *MSH2* and *MSH6* from SNAP2 program

Meanwhile for the one base deletion at domain I *MSH2*, we predicted the effect through PROVEAN web server. We include not only the deletion effect but the AAS as well that are found in *MSH2*. The result shows that the base deletion causes deleterious effects. The AAS at *MSH6* mungbean homologs also shows similar result with those analyzed through SNAP2 program (Table 5).

Table 5. PROVEAN result for amino acid substitution (AAS) and deletion at mungbean *MSH2* and *MSH6*

Homologs	Variant	PROVEAN score	Prediction (cutoff= -2.5)
MSH2	M80del	-10.087	Deleterious
	I93V	0.828	Neutral
	A760S	-1.763	Neutral
MSH6	H900Y	-5.650	Deleterious
	A1052P	-3.026	Deleterious

DISCUSSION

The Characteristics of Mungbean *MSH* Genes

The main objective of our study is to identify and characterize the *MSH* homologs of mungbean. We utilized the availability of mungbean whole genome sequence and *MutS/MSH* genes bioinformatics resources to determine the location of mungbean *MSH* homologs. Those open-access resources allowed us to perform a faster and better characterization of mungbean *MSH* genes.

We identify two homologs located in the chromosome 8, i.e. *MSH1* and *MSH3*. The distance among two homologs is around 31Mb. Meanwhile *MSH2*, *MSH4*, *MSH5*, *MSH6*, and *MSH7* are located in chromosome 7, 11, 6, 3, and 1, respectively. As reported in rye, *MSH2* was mapped to chromosome 1R, *MSH3* was mapped to chromosome 2R and *MSH6* to chromosome 5R. In tomato, *MSH2* and *MSH7* were located in chromosome 6 and 7, respectively. Meanwhile in wheat, *MSH2*, *MSH3*, and *MSH6* were completely detected in three genomes of wheat (A, B, and D genomes). *MSH2* was detected on chromosome 1A, 1B, and 1D; *MSH3* on chromosomes 2A, 2B, and 2D; and *MSH6* on chromosome 5 and 3 of genomes A, B, and D (Korzun et al., 1999; Tam et al., 2009). Accordingly,

in mungbean no homologs that form a heterodimer contact are located in the same chromosome. As has been stated previously, *MSH* proteins functions as heterodimers with distinct mismatch specificities. *MSH2-MSH3* or *MSHβ* recognizes insertion/deletion loops and larger loops of 2-8bp. *MSH2-MSH6* or *MSHα* recognizes base-base mismatch, while *MSH2-MSH7* or *MSHϒ* recognizes a G/T mismatch (Culligan and Hays. 2000; Wu *et. al.* 2003).

Among the four *MSH* homologs that related to MMR in mungbean, only *MSH7* is absence of domain IV which is also supported in other study (Tam et al. 2009). Based on clamp-sliding model, domain IV has function in DNA binding together with domain I. When domain I bind specifically to the mismatch site, domain IV forms a jaws and bind non-specifically to the dsDNA (Tachiki et al., 1998). Therefore, this domain often called as clamp domain and thus make the *MSH7* is being unique for plant. It is presumed that the deletion of clamp domain in *MSH7* causing the reduction of *MSH7* protein-kinking efficiency and therefore bind less well to heteroduplex DNA and to an extra looped out of the nucleotide (Wu et al., 2003).

Based on the domain location within the mungbean *MSH* genes, we found that domain III is the longest domain and containing domain IV as well. This can be explained since domain III act as the core domain for *MSH* genes, whereby connected directly to the three domains, i.e. domain II,

domain IV, and domain V by peptide bonds (Obmolova et al. 2000). On the other hand, the location of domain I is not same in all mungbean *MSH* paralogs. Domain I of *MSH6* and *MSH7* are located slightly far from the transcription start site (TSS) about 240bp and 274bp, respectively. This similar pattern may cause their grouping in same cluster in the phylogenetic tree of *MSH* genes.

Phylogeny and Comparison of *MSH* protein among Crop Species

We built the phylogenetic tree of *MSH* protein sequences among species using their full length of protein. As reported by Culligan et al. (2000), the use of only the C-terminal regions in phylogenetic analysis resulted in tree instabilities. This instability makes the critical identification of relationship among *MSH* homologs being difficult. Our result shows general agreement with other studies especially in term of the most distant of *MSH1* from other homologs (Culligan *et. al*. 2000; Tam *et. al*. 2009). This support the theory of *MSH1* as the eukaryotic precursor which was transferred from *MutS* gene evolution of prokaryotes through the mitochondrial endosymbiotic events. The gene was duplicated at the nucleus and one gene was targeted back the protein to the mitochondrion while the others give rise to nuclear mismatch repair genes. Therefore, *MSH* gene

families are monophyletic which appear to share common ancestor (Culligan *et. al*. 2000).

In our study after the speciation of *MSH1*, we determine three major groups which consist of two homologs for each group. This is slightly different with those in tomato whereby *MSH5* was grouped separately apart from other homologs (Tam et al., 2009). The terminal branch also shows different pattern with those in tomato. In our study, *MSH2* and *MSH7* cluster have longer terminal branch length compare to *MSH3* and *MSH6*. This terminal branch length denotes how far the changes between orthologs. The longer branch of *MSH7* cluster can be understood since this homologs is being unique for plant, thus high variation among orthologs is possible. However, a long branch of *MSH2* does not support its biochemical function as core dimer in the complex protein network, which should be restricted for permissible changes.

From five domains of *MSH* genes, two domains play an important role, i.e. domain I and domain V. From domain I, the information of mismatch site recognition is transferred to the domain V through domain II and III (Obmolova et al., 2000). Therefore domain I is called as mismatch recognition domain and has specific motif called as FYE motif which is conserved for *MSH1*, *MSH6*, and *MSH7*; vary for *MSH3* and missing for

MSH4 and *MSH5*. The absence of these aromatic residues in *MSH4* and *MSH5* consistent with the evolution of *MSH* functional diversification which cause both homologs do not have role in mismatch repair (Culligan et al., 2000). Accordingly, our study shows consistent results in term of the conserved FYE motifs for *MSH1*, *MSH6*, and *MSH7* and varies motif for *MSH3.* In our study, three different mismatch recognition motifs for *MSH3* are identified, i.e. RYR, KYR, and RFR. Although vary, these motifs have similar pattern in term of containing two positively charged residues and one aromatic residue.

Once the mismatch site information is received by domain V, it is known that with the presence of ATP, *MSH* recruits *MLH* that could lead to the activation of *MutH*. The *MutH* can induce double strand break at the GATC site and let the DNA polymerases to correct the DNA mismatch (Schofield and Hsieh, 2003; Kunkel and Erie, 2005; Iyer et al., 2006). Therefore domain V of *MSH* genes majorly comprised of ATP binding site which contain five important motifs (Table 6).

Regarding to our study, mungbean *MSH* homologs contain all of these motifs, except for Walker A motif which is missing in mungbean *MSH2* compared to other species. Walker A functions in forming a loop that binds to the alpha and beta phosphates of di- and tri- nucleotide (Culligan

and Hays 2000; Tam *et. al*. 2009). However, since *MSH2* always become the partner of other homologs when they form a heterodimer in MMR system, the absence of Walker A might be expected. We assume that the need to form a binding-loop during the *MSH* heterodimer contact to the hetero-duplex DNA is done solely by Walker A region of *MSH2* heterodimer partner, i.e. *MSH3*, *MSH6*, or *MSH7*. However this early assumption needs to be tested further.

Meanwhile some part of HTH subdomain of *MSH3* also missing. However, the loss of six residues of HTH subdomain in *MSH3* apparently does not affect the function of the HTH for dimerization process since the YGA motif is remain conserved.

Table 6. Motifs in Domain V and their role in ATP hydrolisis

Motif	Consensus	Function
Walker A	GxxGxGKST	Form a loop that binds to the alpha and beta phosphates of di- and tri- nucleotide
Motif C	STF	Involved in ATP hydrolysis as a gamma phosphate sensor as a signal to the membrane spanning domain
Walker B	4 alliphatic+2 negatively charged + invariant D	Coordinate MG+ ion or polarize the attacking water molecule in ATP hydrolysis and act as switch region
Motif D	Invariant H	Polarizing the attacking water molecule during ATP hydrolysis
HTH subdomain	Y, G, A	Dimerization interface

MSH Genes Variation among Mungbean Accessions

SNP is a DNA sequence variation which occurs abundantly within the genome in which a single nucleotide differs among individual in a population. SNPs may fall within the coding, non-coding or intergenic regions within the genome. In the coding region, SNPs can be found in two types, i.e. synonymous SNPs and non-synonymous SNPs. Most of the

attention of the researcher is to the non-synonymous SNPs because it changes the amino acid sequence of protein. However, the change does not always resulting in protein function alternation. Since the domain I and domain V of *MSH* genes play the major role for its protein function, we observed the occurrence of non-synonymous SNPs in those domains, especially around the important motifs of the domain.

Based on the multiple alignment of partial sequence of *MSH2*, *MSH3*, *MSH6* and *MSH7* of 12 mungbean accessions, we found three synonymous SNPs, four non-synonymous SNPs, and one base deletion. The four non-synonymous SNPs were laid around the important motifs of domain I at *MSH2* and domain V at *MSH2* and *MSH6*. Based on SNAP2 program, two non-synonymous SNPs at *MSH6* are predicted to be affecting the protein function with the accuracy of 80% for *MSH6* Walker A motif and 71% for *MSH6* Motif D.

According to Brombreg and Rost (2007), SNAP2 program predict each substitution of amino acids independently and show every possible substitution at each position of a protein in a heatmap representation (Figure 4). Dark red indicates a high score (score>50, strong signal for effect), white indicates weak signals (-50<score<50), and blue a low score (score<-50, strong signal for neutral/no effect. While black marks the corresponding

wildtype residues. This signal is quantified in a score value which in line with the accuracy rate of prediction. Therefore a higher score result will also present the higher accuracy of analysis.

Meanwhile based on the PROVEAN software, the deletion at mungbean *MSH2* shows deleterious effect. The amino acid substitution (AAS) also shows the similar result with those run by SNAP2 program. Comparable to other *in silico* program, PROVEAN can generate predictions not only for single AAS but also for multiple AAS, insertions, and deletions using the same underlying scoring scheme. The score are obtained based on alignment approach. This approach correlates with the deleteriousness of a sequence variation (Choi et al., 2012; Choi and Chan, 2015). The combination used of multiple tools to predict the effect of AAS and InDels may increase the chance of identifying functional variants that had been missed by other tools.

However, although may never be accurate enough to replace the biological experiments, *in silico* predictions such as SNAP2 and PROVEAN can speed up the selection of potential non-synonymous SNPs that is predicted to be affecting the protein function and may ease the further work. Any information obtained from computational method to detect the presence of SNPs, insertions, and deletions and their effect to the protein function can

be used further in crop improvement programs, especially in term of the associations among SNPs and the traits of economic value.

Related to our study, the information about mungbean *MSH* genes characteristic can be utilized further to gain insight into the association between the genes and the mutation rate in mungbean. Mutagenesis experiments based on the information of the lack of Walker A at *MSH2* and partial HTH subdomain at *MSH3* using CRISPR or other genome editing technologies can be used to create mungbean genotype that high-acceptable to mutation exposure. Meanwhile any mungbean genotypes carrying the affected-non-synonymous SNPs can be evaluate as well for its responsiveness to mutation. Then based on this information, series of SNP markers can be developed to screen the mungbean germplsm collection that naturally brings the mutant gene of *MSH*. Eventually this may help the plant breeders in creating a better plant for the future.

REFERENCE

Allard RW (1999). Principles of Plant Breeding. 2nd ed. New York: John Willey and Sons.

Aslam M, Maqbool MA, Zaman QU, Latif MZ, and Ahmad RM (2013). Responses of mungbean genotypes to drought stress at early growth stages. *International Journal of Basic and Applied Sciences* 13(05):22-27.

Bennetzen JL (2000). Comparative sequence analysis of plant nuclear genomes: microlinearity and its many exceptions. *The Plant Cell* 12:1021-1029.

Bray CM and West CE (2005). DNA repair mechanisms in plants: crucial sensors and effectors for the maintenance of genome integrity. *New Phytologist* 168:511-528.

Bromberg Y and Rost B (2007). SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Research* 35(11): 3823-3835.

Choi Y and Chan AP (2015). PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* 31(16):2745-2747.

Choi Y, Sims GE, Murphy S, Miller JR, and Chan AP (2012). Predicting the functional effect of amino acid substitutions and indels. *PloS ONE* 7(10):e46688

Culligan KM and Hays J (2000). Arabidopsis MutS homologs-*AtMSH2*, *AtMSH3*, *AtMSH6*, and a novel *AtMSH7*-form three distinct protein heterodimers with different specificities for mismatch DNA. *The Plant Cell* 12:991-1002.

Culligan KM, Meyer-Gauen G, Lyons-Weiler J, and Hays J (2000). Evolutionary origin, diversification and specialization of eukaryotic MutS homolog mismatch repair proteins. *Nucleic Acids Research* 28(2):463-471.

Felsenstein J (1985). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* 39:783-791.

Gelvin SB and Schilperoort RA (1995). Plant Molecular Biology Manual. Norwell MA: Kluwer Academic Publisher,

Gupta PK, Roy JK, and Prasad M (2001). Single nucleotide polymorphisms: A new paradigm for molecular marker technology and DNA polymorphism detection with emphasis on their use in plants. *Current Science* 80(4):524-535.

Harten AM (1998). Mutation Breeding: Theory and Practical Applications. England: Cambridge University Press.

Iyer RR, Pluciennik A, Burdett V, and Modrich PL (2006). DNA mismatch repair: functions and mechanisms. *Chem. Rev.* 106:302-323.

Jones DT, Taylor WR, and Thornton JM (1992). The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences* 8:275-282.

Korzun V, Borner A, Siebert R, Malyshew S, Hilpert M, Kunze R, and Puchta H (1999). Chromosomal location and genetic mapping of the mismatch repair gene homologs *MSH2*, *MSH3*, and *MSH6* in rye and wheat. *Genome* 42:1255-1257.

Kunkel TA and Erie DA (2005). DNA mismatch repair. *Annu. Rev. Biochem.* 74:681-710.

Lestari P, Kim SK, Reflinur, Kang YJ, Nurwita D, and Lee SH (2014). Genetic diversity of mungbean (*Vigna radiata* L.) germplasm in Indonesia. *Plant Genetic Resources: Characterization and Utilization* 12(S1): S91-S94.

Li L, Jean M, and Belzile F (2006). The impact of sequence divergence and DNA mismatch repair on homeologous recombination in Arabidopsis. *The Plant Journal* 45:908-916.

Lodish H, Berk A, Kaiser CA, Krieger M, Bretscher A, Ploegh H, Amon A, and Scott MP (2013). Molecular Cell Biology. 7th ed. New York: WH Freeman and Co. 1154 p.

Obmolova G, Ban C, Hsieh P, and Yang W (2000). Crystal structures of mismatch repair protein MutS and its complex with a substrate DNA. *Nature article* 407: 703-710.

Reddy KS (2009). A new mutant for yellow mosaic virus resistance in mungbean (Vigna radiata (L.) Wilczek) variety SML-668 by recurrent gamma-ray irradiation. In: Q.Y. Shu, ed., Induced Plant Mutations in the Genomics Era. Food and Agriculture Organization of the United Nations, Rome, pp. 361-362.

Saitou N. and Nei M (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4:406-425.

Sancar A, Lindsey-Boltz LA, Unsal-Kacmaz K, and Linn S (2004). Molecular mechanisms of mammalian DNA repair and the DNA damage checkpoints. *Annual Review of Biochemistry* 73:39-85.

Sangiri C, Kaga A, Tomooka N, Vaughan D, and Srinives P (2007). Genetic diversity of the mungbean (*Vigna radiata*, Leguminosae) genepool on the basis of microsatellite analysis. *Australian Journal of Botany* 55:837-847.

Schofield MJ and Hsieh P (2003). DNA mismatch repair: molecular mechanisms and biological function. *Annu. Rev. Microbiol.* 57:579-608.

Senaratne R, Liyanage NDL, and Soper RJ (1995). Nitrogen fixation of and N transfer from cowpea, mungbean and groundnut when intercropped with maize. *Fertilizer Research* 40:41-48.

Shanmugasundaram S, Keatinge JDH, and Hughes J (2009a). Counting on beans: mungbean improvement in Asia. In: D.J. Spielman and R. Pandya-Lorch, eds., Millions Fed: Proven Successes in Agricultural Development. IFPRI, Washington DC.

Shanmugasundaram S, Keatinge JDH, Hughes J (2009b). The mungbean transformation: diversifying crops, defeating malnutrition. *IFPRI Discussion Paper 922*. International Food Policy Research Institute, Washington DC.

Supek F and Lehner B (2015). Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature Letter:*1-4.

Tachiki H, Kato R, Masui R, Hasegawa K, Itakura H, Fukuyama K, and Kuramitsu S (1998). Domain organization and functional analysis of *Thermus thermophilus* MutS protein. *Nucleid Acids Research* 26(18):4153-4159.

Tah PR (2006). Induced macromutation in mungbean [*Vigna radiata* (L.) Wilczek]. *International Journal of Botany* 2(3):219-228.

Tam SM, Samipak S, Britt A, and Chetelat RT (2009). Characterization and comparative sequence analysis of the DNA mismatch repair *MSH2* and *MSH7* genes from tomato. *Genetica* 137:341-354.

Tam SM, Hays JB, and Chetelat RT (2011). Effects of suppressing the DNA mismatch repair system on homeologous recombination in tomato. *Theor. Appl. Genet.* 123:1445-1458.

Tamura K, Stecher G, Peterson D, Filipski A, and Kumar S (2012). Mega6: Molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* 30(12):2725-2729.

Van K, Kang YJ, Han KS, Lee YH, Gwag JG, Moon JK, and Lee SH (2013). Genome-wide SNP discovery in mungbean by Illumina HiSeq. *Theor. Appl. Genet* 126:2017-2027.

Wu SY, Culligan K, Lamers M, and Hays J (2003). Dissimilar mispair-recognition spectra of Arabidopsis DNA-mismatch-repair proteins *MSH2-MSH6* and *MSH2-MSH7*. *Nucleic Acids Research* 31(20):6027-6034.

ABSTRACT IN KOREAN

녹두 DNA 불일치복구 유전자의 비교서열분석

ANDARI RISLIAWATI

초록

녹두(*Vigna radiata* (L.) R. Wilczek)는 질소고정을 통하여 토양 상태를 향상시킬 수 있는 고단백질 콩과 작물이다. 녹두는 남아시아 폭넓게 재배되고, 건조한 환경에서의 적응력이 뛰어난 가뭄 내성 작물이다. 그러나 전세계적으로 녹두의 생산량 증가는 정체되어 있고, 품종 개발은 이들의 낮은 유전적 다양성 때문에 제한되고 있다. 이러한 문제들을 극복하기 위해 유전자원 탐색(germplasm exploration), 돌연변이 유도와 같은 노력들이 있었지만, 만족스러운 결과는 얻지 못했다. 이는 *MSH* 유전자의 강한 발현으로 인하여 일어날 수 있다. *MSH* 유전자는 선행연구를 통해서 유전체의 온전성을 보존하기 위해 발현된다고 알려진 바 있다. 녹두에서의 *MSH* 유전자에 대해 더 탐색하기 위해, 우리는 8개의 나라에서 수집된 12개의 녹두 자원(종자)에서 *MSH* 유전자의 염기서열을 분석하였고, 쌍떡잎식물 14개 식물에서 70개의 *MSH* 유전자 서열과 비교하였다.

우리는 염색체 8번의 *MSH3*, 염색체 7번의 *MSH2*, 염색체 3번의 *MSH6*, 염색체 1번의 *MSH7*에서 DNA 불일치 복구와 관련된 *MSH* 유전자 상동유전자의 위치를 확인했다. *MSH2, MSH3, MSH6* 상동유전자에서 5개의 보존된 도메인이 모두 존재하는 반면, *MSH7*에서는 한 개의 도메인이 부족했다. 다른 종과 비교했을 때, 녹두는 *MSH2*에 있는 워커 A 모티프(Walker A motif)와 일부 HTH 서브도메인를 잃었다. 우리는 또한 단일염기 다형성(SNP)을 확인하였고, 녹두 유전자원(germplasm)의 도메인 I와 도메인 V사이의 아웃모티브에서 염기 결실을 확인하였다. 참고한 녹두와 비교했을 때, 염기 결실뿐만 아니라 두 개의 비동의

단일염기에서도 다른 단백질 기능을 예측되었다. 단일염기 다형성을 가지고 있는 녹두를 통해 돌연변이에 대한 민감성을 평가할 수 있고, 단일염기 다형성 마커는 원하는 대립유전자를 가지고 있는 적절한 유전자원을 가릴 수 있도록 설계할 수 있다. 그러므로, 녹두의 *MSH* 유전자 확인 연구는 돌연변이 유도를 통한 녹두 육종에 기여를 할 수 있을 것으로 생각되며, 이를 위해서는 추후, 유전체 삽입 등의 실험을 통해 기능적 연구가 더 필요할 것이다.

핵심단어: 녹두, *MSH*, 도메인(Domain), 모티프(Motifs), 단일염기 다형성(SNPs), 단백질 기능 예측

학번: 2014-22124