A THESIS FOR THE DEGREE OF MASTER SCIENCE

Genome-wide association study for seed size and color in mungbean (*Vigna radiata* (L.) Wilczek)

BY

SIPHANH DAOVONGDEUAN

August, 2017

MAJOR IN CROP SCIENCE AND BIOTECHNOLOGY

DEPARTMENT OF PLANT SCIENCE

THE GRADUATE SCHOOL OF SEOUL NATIONAL UNIVERSITY

Genome-wide association study for seed size and color in mungbean (*Vigna radiata* (L.) Wilczek)

SIPHANH DAOVONGDEUAN

ABSTRACT

Seed size is one of the most important traits that significantly affect the whole production, and seed color is greatly affected the price, preference and acceptability for consumers. Therefore, identifying and understanding genes involved in the control of seed size and color are crucial and necessary for mungbean improvement. However, many studies of specific-controlled genes of seed size and color have not been conducted. The objectives of this study were to investigate phenotypic variations in seed size and color among mungbean accessions and to identify loci associated with seed size and color in cultivated mungbean accessions using genome-wide association study (GWAS). Phenotypic data were collected from seed as seed size and color. Principle component analysis (PCA) plots were constructed based on phenotypic data using 209 cultivated and 9 wild

mungbean accessions from more than 23 countries. The cultivated and wild mungbean accessions in both seed size and color are clearly separated. Manhattan plots were constructed with a total 18,171 of SNP markers developed using genotyping by sequencing (GBS) on 209 cultivated mungbean accessions. Using GWAS to perform general line model (GLM). A single significant marker associated with red and green seed color components was identified on Chr04 at 3370425 bp with p-value of <0.005, while blue seed color component and seed sizes had no significant marker associated with.

Keywords: Mungbean, seed size, seed color, GBS, GWAS, PCA, GLM.

Student number : 2015-22341

CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

The high-yielding mungbean varieties are normally large-seeded and green, whereas the yellow ones are usually small-seeded and relatively low-yielders. Seed size varies in mungbean and significantly affects to its whole production. Seed color and luster are also important for mungbean. For example, the price and acceptability of mungbean in the Philippines are greatly affected by the luster, color and size of the seed (Manguiat et al., 1978). It is indicated that an increase in seed number per plant was a major driver of the increase in yields, however, a change in individual in seed size are constant. Therefore, an increasing in seed size will be the next driver for increasing yield (Ainsworth et al. 2012). Seed coat color was controlled by mono-gene, and green was dominant to yellow, while black was dominant to green (Wang li-xia, 2013).

Bose (1959), Sen and Ghose (1959), and van Rheenen (1965) reported dull, rough seed surface is a monogenic dominant over glossy, smooth seed surface. Sen and Ghosh (1959) identified a second dominant gene that governs dull, rough seed coat. Dull seeds are covered by an inner pod membrane that renders the seed dull; when this membrane is removed the seed coat underneath is shiny (Watt et al., 1977). The pod membrane

may contain brown or black pigment through which the seed coat color may not be apparent.

Variations of seed morphology traits greatly affect seed production, preference and acceptability. The main causes making these traits different are their genetics. Their genotypic traits control each of their phenotypic traits appeared. Therefore, for the future improvement of mungbean varieties, knowing genetic information related to traits is very crucial for breeders, especially important genes controlling those desired traits. This study is on genome-wide association studies for seed size and color in mungbean (*Vigna radiata* (L.) Wilczek), with the objectives are to investigate phenotypic variations in seed size and color among mungbean accessions and to identify loci associated with seed size and color in mungbean accessions using genome-wide association study (GWAS). The study is specifically to investigate the correlation in seed size and color among 218 cultivated and wild mungbean accessions related to phenotypic data and to examine a genome-wide set of genetic variants in different individuals to see variants associated with a trait of 209 accessions of cultivated mungbean by using association analysis of 18,171 of SNP markers generated by genotyping by sequencing (GBS) approach.

LITERATURE REVIEW

Mungbean

Mungbean (*Vigna radiata* (L.) Wilczek) also known as Mungo, green gram and golden gram ranks among the most important vegetable pulse crop that is widely grown throughout south and Southeast Asia, especially underdeveloped and developing countries where it is can be found a low cost of valuable nutrition supplying to its population. Mungbean is a cheap source of many kinds of nutrition as carbohydrates, iron, folate and protein, where can be gained from their diary consumption. Mungbean has been consumed as vegetables or processed into flour, soups, porridge, noodles and ice-cream. The utilization of mungbean for food is different according to regions. For example, the traditional Indian porridge dhal, the Thailand noodle salad Yum Woon Sen, and the Korean sprout side dish Sukjunamul (Kang et al., 2014).

Mungbean is an annual and warm-season, 0.3-1.5 m tall, erect or sub-erect plant, sometimes slightly twining at the tips. It is asynchrony, deep-rooted, much branched with long petioles and has a short life cycle (approximately 60 days). It is mainly cultivated on small farm in south, east and Southeast Asia (Kang et al., 2014, Rachie and Roberts, 1974).

Mungbean belongs to the family of *Fabaceace*, the Leguminosae genus *Vigna*, sub-genus *Ceratotropis* and the tribe *Phaseoleae* (Tomooka et al., 2002,). Mungbean is diploid in nature with 2n=2x=22 and has a small genome size estimated to be 0.60 pg/1C (579 Mbp) which is similar to the other *Vigna* species (Prakit and Peerasak, 2007). Mungbean is a general name of colorful mungbean such as, for the green called green gram, for the yellow called golden gram, and for the black called black gram, black matpe or urd.

It is believed that mungbean probably originated in India (De Candole, 1886; Zhukovsky, 1950; Bailey, 1970) or the Indo-Burmese region (Vavilov, 1951; H.B. Singh et al., 1970; Jain and Mehra, 1980). India is probable center of domestication (Smartt, 1985). Lukoki et al. (1980) proposed that V. radiata var. sublobata (Roxb.) Verde. that occurs wild in India and crosses easily with V. radiata var. radiata is the wild ancestor of mungbean. During the early domestication process, mungbean cultivation migrated to other Asian countries and to Africa. The modern cultivated mungbean is currently distributed throughout southern and eastern Asia, Africa, and Austronesia (Vishnu-Mittre, 1974; Lambrides and Godwin, 2007). The cultivation area of mungbean in the world is around 6 million hectares (Nair et al., 2012), the largest mungbean cultivation area is in Asian

countries are India, China, Myanmar and Indonesia, accounting for 90% of world production (Lambrides and Godwin, 2007).

Genotyping by sequencing (GBS)

Next-generation genotyping, or Genotyping by sequencing, also called GBS, is a method to discover single nucleotide polymorphisms (SNP) in order to perform genotyping studies, such as genome-wide association studies (GWAS). GBS method is simple, quick, extremely specific, highly reproducible, and may reach important regions of the genome that are inaccessible to sequence capture approaches. GBS used for whole genome sequencing and for re-sequencing projects where the genomes of several specimens are sequenced to discover large numbers of single nucleotide polymorphisms (SNPs) for exploring within-species diversity, constructing haplotype maps and performing genome-wide association studies (GWAS). GBS has been demonstrated to be a robust and cost-effective genotyping method capable of producing thousands to millions of SNPs across a wide range of species. GBS approach is suitable for population studies, germplasm characterization, breeding and trait mapping in diverse organisms.

GBS uses restriction enzymes to reduce genome complexity and genotype multiple DNA samples and then to sequence GBS libraries using next generation sequencing technologies. The result of GBS is around 100 bp single-end reads. This method was first described by Elshire et al. (2011) by using high molecular weight DNAs extracted from plant leaves and digested using a specific restriction enzyme (RE) as *ApeKI* previously defined by cutting frequently in the major repetitive fraction of the genome. Barcode adapters are then ligated to sticky ends and PCR amplification is performed. Next-generation sequencing technology is performed resulting in about 100 bp single-end reads. Raw sequence data are filtered and aligned to a reference genome using usually Burrows-Wheeler alignment tool (BWA) or Bowtie 2. The next step is to identify SNPs from aligned tags and score all discovered SNPs for various coverage, depth and genotypic statistics. A large-scale, species-wide SNP production has been run; it is possible to quickly call known SNPs in newly sequenced samples.

GBS method that apply reduced representation approach, the filtering criteria can include (1) a minimum read depth (often ≥ 3 per genotype), (2) > 90% nucleotide within a genotype having identical call at a given position (< 10% sequencing error), and (3) a read depth ≤ mean of the sequence depth over the entire mapping assembly (Kumar *et al*., 2012).

Advantages of genotyping by sequencing are to sequence predetermined areas of genetic variation over many samples, to provide a low cost per sample for certain applications, to reduce ascertainment bias compared to arrays, to identify variants other than SNPs, including small insertions, deletions, and microsatellites, to enable comparative analyses across samples in the absence of a reference genome and to informs genetic mapping, screening backcross lines, purity testing, constructing haplotype maps, association mapping, and genomic selection for plant studies.

GBS method has been widely used in many plants for several purposes, such as genomic selection study (Poland and Rife, 2012); genetic diversity study (Wong *et al*., 2015) and constructing high-density genetic map (Poland *et al*., 2012; Li *et al*., 2015).

Single Nucleotide Polymorphisms (SNPs)

Single nucleotide polymorphisms, frequently called SNPs (snips), are the most common type of genetic variation among individuals. Each SNP represents a difference in a single DNA building block, called a nucleotide (A, T, C, and G). For example, a SNP may replace the nucleotide cytosine (C) with the nucleotide thymine (T) in a certain stretch of DNA. SNPs act as chromosomal tags to specific regions of DNA, and these regions can be scanned for variations. Most commonly, these variations are found in the DNA between genes. Variation can be classified as SNP if more than 1% of a population does not carry the same nucleotide at a specific position in the DNA sequence.

Single nucleotides may be changed (substitution), removed (deletions) or added (insertion) to a polynucleotide sequence. Single nucleotide polymorphisms may fall within coding sequences of genes, non-coding regions of genes, or in the intergenic regions between genes. SNPs within a coding sequence will not necessarily change the amino acid sequence of the protein that is produced, due to degeneracy of the genetic code.

SNPs are genetic markers of choice for both linkage and association mapping and for population structure and evolution analysis.

They are virtually unlimited, evenly distributed along the genome, bi-allelic and co-dominant.

The applications of SNPs include Association studies that can determine whether a genetic variant is associated with a disease or trait. A tag SNP is a representative single-nucleotide polymorphism (SNP) in a region of the genome with high linkage disequilibrium (the non-random association of alleles at two or more loci). Tag SNPs are useful in whole-genome SNP association studies in which hundreds of thousands of SNPs across the entire genome are genotyped, haplotype mapping are sets of alleles or DNA sequences can be clustered so that a single SNP can identify many linked SNPs, linkage disequilibrium (LD) is a term used in population genetics, indicates non-random association of alleles at two or more loci, not necessarily on the same chromosome. It refers to the phenomenon that SNP allele or DNA sequence which is close together in the genome tends to be inherited together. LD is affected by two parameters as distance between the SNPs (the larger the distance the lower the LD) and recombination rate (the lower the recombination rate the higher the LD)

Genome-Wide Association Studies (GWAS)

A genome-wide association study (GWA study, or GWAS), also known as whole genome association study (WGA study, or WGAS), is an examination of a genome-wide set of genetic variants in different individuals to see if any variant is associated with a trait. GWASs typically focus on associations between single-nucleotide polymorphisms (SNPs) and traits. GWASs investigate the entire genome and are the gene-specific candidate-driven studies. GWASs identify SNPs and other variants in DNA associated with a trait, but they cannot on their own specify which genes are causal. GWAS focus on SNPs, the single nucleotide sites that differ between individuals and the purpose is to determine alleles that correlate to different diseases and traits (Stranger B, Stahl E, Raj T., 2011).

The first successful GWAS was published in 2005 by investigating patients with age-related macular degeneration (Klein RJ; Zeiss C; Chew EY; Tsai JY; et al., 2005). GWAS has been used in many kinds of plants such as Arabidopsis thaliana (Shindo C, Bernasconi G, Hardtke CS, 2007), barley (Inka Gawenda, 2015), maize (Yingjie Xiao, 2016), rice (Filippo Biscarini, 2016). GWAS in plants (in Arabidopsis thaliana, rice, and maize) have explained a much greater proportion of the phenotypic variation than that explained by human GWAS studies. In plants at least, the assumption

that common genetic variation explains common phenotypic variation holds. In plants, rare variation can become sufficiently common in large families or populations to be identifiable by GWAS. For example, GWAS have identified SNPs and population structure that can explain up to 45% of the phenotypic variation in flowering time.

Genome-wide association approach (GWAS) overcomes several limitations of traditional gene mapping by (i) providing higher resolution, often to the gene level, and (ii) using samples from previously well-studied populations in which commonly occurring genetic variations can be associated with phenotypic variation (Benjamin Brachi, Geoffrey P Morris and Justin O Borevitz., 2011).

MATERIALS AND METHODS

Plant Materials

234 mungbean accessions derived from gene bank, originated from 23 countries, unknown sources and wild accessions (Table 2, 3). The categorization of mungbean accessions based on genetic analysis and morphology consists of 225 cultivated varieties and 9 wild accessions. This season could harvest totally 218 accessions made up of 200 green mungbean, 12 black mungbean and 6 yellow mungbean (Table 1). The mungbean accessions were planted in Seoul National University farm in Suwon, Korea. They were planted totally 10 holes for a row for an accession, with 2 seeds a hole. The space between holes is 15 cm and between rows is 70 cm in distance.

Table 1. Mungbean accessions in colors and lusters harvested in 2017

Accession colors	luster		Total
	dull	shiny	
Green	103	97	200
Yellow	1	5	6
Black	10	2	12
Total	114	104	**218**

Table 2. Mungbean accessions in countries

No.	Countries	Accession 2017
1	Afghanistan	3
2	Australia	2
3	Bangladesh	1
4	Cambodia	3
5	China	17
6	India	7
7	Indonesia	65
8	Iran	7
9	Kenya	1
10	Myanmar	7
11	Madagascar	1
12	Nepal	2
13	North Korea	2
14	Pakistan	5
15	Singapore	1
16	South Korea	35
17	Taiwan	6
18	Thailand	14
19	The Philippines	12
20	Turkey	1
21	Unknown	11
22	America	2
23	Uzbekistan	5
24	Vietnam	1
25	Wild	7
Total		218

Table 3. Mungbean accessions in continents

No.	Continents	Accession
1	Africa	2
2	America	2
3	Asia	194
4	Australia	2
5	Unknown	11
6	Wild	7
Total		218

Table 4. List of mungbean accessions gained SNPs used for GWAS

GBS	accession name	gene bank code	origin	Field	wild/cultivated	luster	color
491	Pokhara	IT109327	NPL	18_1_1	cultivated	S	G
265	NATIVE COLLECTION	IT201188/701320	PHL	18_1_2	cultivated	D	G
494	Acc. 7861	801374	UNK	18_1_3	cultivated	D	G
495	Acc. 363	807483	PHL	18_1_4	cultivated	S	G
496	EG MG-13	807492	PHL	18_1_5	cultivated	S	G
497	Dull Yellow	807498	PHL	18_1_6	cultivated	S	Y
268	K001419	K001419	MMR	18_1_7	cultivated	S	G
498	RUS-NYW-2000-201	K005410	MDA	18_1_8	cultivated	S	G
269	V01160	K024055	IND	18_1_9	cultivated	S	G
499	EG-MG-60	K024059	PHL	18_1_10	cultivated	D	G
502	V01471	K024060	IDN	18_1_11	cultivated	S	G
500	V01673	K024062	AFG	18_1_12	cultivated	D	G
501	HERNITAGERS	K024067	AUS	18_1_13	cultivated	S	G
505	ML-9	K130624	CHN	18_1_14	cultivated	S	G
506	CES-J-24	K130625	PHL	18_1_15	cultivated	S	G

507	FG-MG-1743	K130627	PHL	18_1_16	cultivated	S	G
508	VC-2307A	K130631	CHN	18_1_17	cultivated	S	G
509	R-288-8	K130632	CHN	18_1_18	cultivated	D	G
510	CES-87	K130634	PHL	18_1_19	cultivated	S	G
511	P-3-A-40	K130639	UNK	18_1_20	cultivated	S	G
277	Pe Nauk	K130643	UNK	18_1_21	cultivated	D	G
278	P-69-319	K130644	UNK	18_1_22	cultivated		
279	P-4-44	K130646	UNK	18_1_23	cultivated	S	G
512	NCM-1	K130648	TWN	18_1_24	cultivated	S	G
280	Local	K130649	UNK	18_1_25	cultivated	D	G
281	Lokal	K130650	UNK	18_1_26	cultivated		
282	Local	K130652	UNK	18_1_27	cultivated	D	G
513	MYS-PYJ-2007-55	K131569	UNK	18_1_28	cultivated	S	G
283	KJA17	K136410	CHN	18_1_29	cultivated	S	G
284	K163475	K163475	CHN	18_1_30	cultivated	S	G
514	CN 72	K165772	MMR	18_1_31	cultivated	S	G
515	VC 6141-54	K165773	MMR	18_1_32	cultivated	S	G
516	VC 6368-46-40	K165775	MMR	18_1_33	cultivated	S	G
517	VC 6173-B-10	K165776	MMR	18_1_34	cultivated	S	G
518	VC 6173C	K165781	MMR	18_1_35	cultivated	D	G
519	VC 12-3-4A	K165782	MMR	18_1_36	cultivated	D	G
520	CHN-2010-18	K166126	CHN	18_1_37	cultivated	S	G
521	CHN-2010-19	K166127	CHN	18_1_38	cultivated	S	G
285	CHN-2010-20	K166128	CHN	18_1_39	cultivated	S	G
523	YV 542	K169726	SGP	18_1_40	cultivated	S	G
524	VIR 6559	K173279	PRK	18_1_41	cultivated	S	G
525	VIR 6560	K173280	PRK	18_1_42	cultivated	D	G
526	CHN- - 2011-11	K175297	CHN	18_1_43	cultivated	S	G
527	CHN- - 2011-12	K175298	CHN	18_1_44	cultivated	S	G
528	CHN-PMW-2011-4	K175546	CHN	18_1_45	cultivated	S	G
530	KHM-LWJ-2011-9	K176372	KHM	18_1_46	cultivated	S	G
531	KHM-LWJ-2011-10	K176373	KOR	18_1_47	cultivated	D	G
532	KHM-LWJ-2011-12	K176375	KHM	18_1_48	cultivated	S	G
287	UZB-Shahrisabz-2011-21	K187555	UZB	18_1_49	cultivated	S	G

533	UZB-Shahrisabz-2011-33	K187567	UZB	18_1_50	cultivated	S	G
534	UZB-Khazarbag-2011-60	K187592	UZB	18_1_51	cultivated	S	G
288	KHM- -2012-4	K191035	KHM	18_1_52	cultivated	S	G
621	Arta Moseng	Puji 1	IDN	18_1_53	cultivated	D	G
183	Si Walik	Puji 2	IDN	18_1_54	cultivated	D	G
623	Lok Madura	Puji 3	IDN	18_1_55	cultivated	D	G
624	Arta Item	Puji 4	IDN	18_1_56	cultivated	D	G
625	Arta ijo	Puji 5	IDN	18_1_57	cultivated	D	G
626	Manyar	Puji 6	IDN	18_1_58	cultivated	S	G
627	Bhakti	Puji 7	IDN	18_1_59	cultivated	S	G
628	No 129	Puji 8	IDN	18_1_60	cultivated	S	G
629	Nuri	Puji 9	IDN	18_2_1	cultivated	D	G
630	Kenari	Puji 10	IDN	18_2_2	cultivated	D	G
631	Betet	Puji 11	IDN	18_2_3	cultivated	S	G
632	Gelatik	Puji 12	IDN	18_2_4	cultivated	D	G
633	Parkit	Puji 13	IDN	18_2_5	cultivated	S	G
634	Merpati	Puji 14	IDN	18_2_6	cultivated	S	G
636	Camar	Puji 16	IDN	18_2_7	cultivated	S	G
637	Merak	Puji 17	IDN	18_2_8	cultivated	S	G
638	Calon haji 1a	Puji 18	IDN	18_2_9	cultivated	D	G
639	Samsek-a	Puji 19	IDN	18_2_10	cultivated	D	G
640	Antap ongko 1a	Puji 20	IDN	18_2_11	cultivated	S	R
642	Calon haji ongko	Puji 22	IDN	18_2_12	cultivated	D	G
643	Plastik	Puji 23	IDN	18_2_13	cultivated	S	G
644	Lok Ps Jailolo	Puji 24	IDN	18_2_14	cultivated	D	G
646	FOrewehal	Puji 26	IDN	18_2_15	cultivated	D	G
647	Fore Lotu	Puji 27	IDN	18_2_16	cultivated	D	G
648	Bue bura	Puji 28	IDN	18_2_17	cultivated	D	G
649	Lok Kab Borong A	Puji 29	IDN	18_2_18	cultivated	S	G
650	Lok Kota Kumbah A	Puji 30	IDN	18_2_19	cultivated	S	G
651	Nilon	Puji 31	IDN	18_2_20	cultivated	S	G
652	Lok Mutoha M-1	Puji 32	IDN	18_2_21	cultivated	S	G
653	Lok Abuki	Puji 33	IDN	18_2_22	cultivated	D	G
654	Lok Majenang A	Puji 34	IDN	18_2_23	cultivated	D	G
655	Lok Pangalengan	Puji 35	IDN	18_2_24	cultivated	D	G
656	Lok Tarogong	Puji 36	IDN	18_2_25	cultivated	D	G
657	Mentik hitam	Puji 37	IDN	18_2_26	cultivated	D	G

658	PB-1 (benggolo Puti)	Puji 38	IDN	18_2_27	cultivated	S	G
659	Lok Kudus	Puji 39	IDN	18_2_28	cultivated	D	G
660	Lok Ngawi	Puji 40	IDN	18_2_29	cultivated	D	G
661	Lok Pemeungpeuk	Puji 41	IDN	18_2_30	cultivated	S	G
662	Lok Bungbulang	Puji 42	IDN	18_2_31	cultivated	D	G
663	Lok Kupang	Puji 43	IDN	18_2_32	cultivated	S	G
664	Butek Surade	Puji 44	IDN	18_2_33	cultivated	S	G
665	Fore Belu	Puji 45	IDN	18_2_34	cultivated		
666	Perkutut	Puji 46	IDN	18_2_35	cultivated	S	G
667	Tecer Hitam	Puji 47	IDN	18_2_36	cultivated	D	B
668	Lok Sampang 1	Puji 48	IDN	18_2_37	cultivated	D	G
669	Lima-1	Puji 49	IDN	18_2_38	cultivated	D	G
670	Lok Jerowaru	Puji 50	IDN	18_2_39	cultivated	D	G
247	Bohabe yellow mongo		PHL	18_2_40	cultivated	S	Y
248	Zilola		UZB	18_2_41	cultivated	S	G
249	Durdona		UZB	18_2_42	cultivated	S	G
250	Turon		UZB	18_2_43	cultivated	S	G
152	Kh.50 hari (L.insana)	Puji 54	IDN	18_2_44	cultivated	S	G
153	Lokal Landa Baru	Puji 55	IDN	18_2_45	cultivated	D	G
676	Lokal Mutaha K2	Puji 56	IDN	18_2_46	cultivated	D	G
155	Lokal Garut M	Puji 58	IDN	18_2_47	cultivated	S	G
157	Mentik Coklat	Puji 60	IDN	18_2_48	cultivated	D	G
681	Lok. Pasar Welahan	Puji 61	IDN	18_2_49	cultivated	D	G
682	Lokal Sidamulih	Puji 62	IDN	18_2_50	cultivated	D	G
160	Kuyak	Puji 63	IDN	18_2_51	cultivated	S	G
684	RR-2	Puji 64	IDN	18_2_52	cultivated	S	G
687	Lok. Garut	Puji 67	IDN	18_2_53	cultivated	S	G
165	Lok Puda 1	Puji 68	IDN	18_2_54	cultivated		
166	Lok Galis	Puji 69	IDN	18_2_55	cultivated		
168	Arta Koneng	Puji 71	IDN	18_2_56	cultivated	D	Y
170	Lok. Pasanggaran	Puji 73	IDN	18_2_57	cultivated	D	G
171	Lok. Jonggat	Puji 74	IDN	18_2_58	cultivated	D	G
695	Fore Modok	Puji 75	IDN	18_2_59	cultivated		
173	ArthaZatim?	Puji 76	IDN	18_2_60	cultivated	D	G
697	Fue Nutu	Puji 77	IDN	19_1_1	cultivated	D	G
698	Foe Nutu	Puji 78	IDN	19_1_2	cultivated	S	G
700	Kambe Morowisa	Puji 80	IDN	19_1_3	cultivated	D	G
701	Kambe Kulita Kokana	Puji 81	IDN	19_1_4	cultivated	D	G

179	Kabe Mor	Puji 82	IDN	19_1_5	cultivated	D	G
704	Lok. Bajawa Ngada	Puji 84	IDN	19_1_6	cultivated	D	G
705	Lok. Ps. Embai Golewa Barat	Puji 85	IDN	19_1_7	cultivated		
201	JP229109		KOR	19_1_8	cultivated	S	G
202	JP229144		CHN	19_1_9	cultivated	S	G
203	JP229145		CHN	19_1_10	cultivated	S	G
204	JP229215		CHN	19_1_11	cultivated	D	G
205	JP229216		CHN	19_1_12	cultivated	D	G
206	JP99049		TWN	19_1_13	cultivated	S	G
207	JP231194		PHL	19_1_14	cultivated		
211	JP229233		IDN	19_1_15	cultivated		
212	JP78939		VNM	19_1_16	cultivated	D	G
213	JP229096		THA	19_1_17	cultivated	D	G
214	JP229097		THA	19_1_18	cultivated	S	G
215	JP229098		THA	19_1_19	cultivated	S	G
216	JP229099		THA	19_1_20	cultivated	S	G
217	JP231216		THA	19_1_21	cultivated	S	G
218	JP231220		THA	19_1_22	cultivated	D	G
219	JP229130		BGD	19_1_23	cultivated	D	G
222	JP229163		IND	19_1_24	cultivated	S	G
228	JP229177		IND	19_1_25	cultivated	S	B
229	JP229175		IND	19_1_26	cultivated	D	B
231	JP229211		IND	19_1_27	cultivated	S	G
233	JP229190		IND	19_1_28	cultivated	S	G
235	JP231223		IND	19_1_29	cultivated	D	G
241	JP103138-1		PAK	19_1_30	cultivated	S	G
242	JP103138-2		PAK	19_1_31	cultivated	S	B
243	JP99066		PAK	19_1_32	cultivated	D	G
244	JP229241		AFG	19_1_33	cultivated	S	G
246	JP31324		AFG	19_1_34	cultivated	S	G
251	JP229254		IRN	19_1_35	cultivated	S	G
252	JP229257		IRN	19_1_36	cultivated	S	G
253	JP229263		IRN	19_1_37	cultivated	S	G
254	JP31331		IRN	19_1_38	cultivated	D	G
255	CN900001		THA	19_1_39	cultivated	S	G
256	CN900002		THA	19_1_40	cultivated	S	G
257	CN900004		THA	19_1_41	cultivated	D	G
258	CN900005		THA	19_1_42	cultivated	S	G
259	CN900007		THA	19_1_43	cultivated	S	G

260	CN900008		THA	19_1_44	cultivated	S	G
263	CN900014		THA	19_1_45	cultivated	S	G
264	CN900015		THA	19_1_46	cultivated	S	G
752	JP229290		AUS	19_1_47	cultivated		
301	2	26018	KOR	19_1_48	cultivated	D	G
303	17	26062	KOR	19_1_49	cultivated	D	G
304	2	26252	KOR	19_1_50	cultivated	S	G
307	-1985-2800	102800	KOR	19_1_51	cultivated	D	G
308	-1985-2855	102855	KOR	19_1_52	cultivated	D	G
310	-1985-3400	103400	KOR	19_1_53	cultivated	D	G
311	-1985-3418	103418	KOR	19_1_54	cultivated	D	G
313	-1985-3835	103835	KOR	19_1_55	cultivated	D	G
316	-1985-4183	104183	KOR	19_1_56	cultivated	D	G
317		104747	TWN	19_1_57	cultivated	S	G
319	-1985-5626	105626	KOR	19_1_58	cultivated	D	G
320		105639	KOR	19_1_59	cultivated	D	G
322		111041	KOR	19_1_60	cultivated	D	G
323	-1985-12822	112822	KOR	19_2_1	cultivated	D	G
326	Siraha Local-2	136322	NPL	19_2_2	cultivated	D	G
327	-1986-24373	138114	KOR	19_2_3	cultivated	D	G
328	VC3566-B-2-1-3	145301	UNK	19_2_4	cultivated	S	G
329	V1153	154078	IRN	19_2_5	cultivated	D	G
330	V3686	154080	USA	19_2_6	cultivated	D	G
331	Yellowgram	154085	UNK	19_2_7	cultivated	S	Y
332	Pakistan	154087	PAK	19_2_8	cultivated	D	G
333	-1989-5463	162743	KOR	19_2_9	cultivated	D	G
334	-1989-5499	162779	KOR	19_2_10	cultivated	S	G
335	-1989-5600	162880	KOR	19_2_11	cultivated	D	G
336	Vo1301	163175	CHN	19_2_12	cultivated	D	G

338	Vo3484	163234	PAK	19_2_13	cultivated	D	G
340	Vo5551	163280	IRN	19_2_14	cultivated	S	G
342	MBLS90-19	168064	KOR	19_2_15	cultivated	D	G
343	MBLS90-30	168075	KOR	19_2_16	cultivated	D	G
344	-1992-2658	175816	KOR	19_2_17	cultivated	D	G
345	-1993-2473	180833	KOR	19_2_18	cultivated	D	G
346	-1993-2475	180835	KOR	19_2_19	cultivated	D	G
347	-1993-3516	181876	KOR	19_2_20	cultivated	D	G
348	92 466	182212	UNK	19_2_21	cultivated	S	G
349	VC1089A	182225	TWN	19_2_22	cultivated	D	G
350	VC2768B	182247	TWN	19_2_23	cultivated	S	G
351	VC3523	182255	UNK	19_2_24	cultivated	D	G
352	19	182273	KOR	19_2_25	cultivated	S	G
354	VC3890B	182296	TWN	19_2_26	cultivated	S	G
355	5	183235	KOR	19_2_27	cultivated	D	G
356	-1	183263	UNK	19_2_28	cultivated	D	G
357	92	183264	KOR	19_2_29	cultivated	D	G
358	Celera	183789	AUS	19_2_30	cultivated	S	G
359	-1994-3231	185570	KOR	19_2_31	cultivated	D	B
361		185575	KOR	19_2_32	cultivated	D	G
362	-1994-3237	185576	KOR	19_2_33	cultivated	D	G
365	V01122B-G	189472	TKY	19_2_34	cultivated	D	G
366	V01946A-Y	189516	PHL	19_2_35	cultivated	S	Y
368	V03538B-G	189552	KNY	19_2_36	cultivated	D	G
369	V03720B-G	189553	USA	19_2_37	cultivated	D	G
370	V03827A-G	189554	IRN	19_2_38	cultivated	S	G
374	-1995-2895	191127	KOR	19_2_39	cultivated	D	G
380	Original Tashkent 1978	199273	KOR	19_2_40	cultivated	D	G
381	ACC6	201172	PHL	19_2_41	cultivated	D	G
382	ACC11	201177	PHL	19_2_42	cultivated	S	Y
397	-2000-28	212101	KOR	19_2_43	cultivated	D	G
399	CHN-LJR-2000-35	212108	CHN	19_2_44	cultivated	S	G

400		216796	KOR	19_2_45	cultivated	D	G
718	W116			19_2_46	wild		
726	W147			19_2_47	wild	D	B
730	W162			19_2_48	wild	D	B
734	W169			19_2_49	wild	D	B
738	W176			19_2_50	wild	D	B
742	W190			19_2_51	wild	D	B
743	W191			19_2_52	wild	D	B
744	W192			19_2_53	wild	D	B
745	W203			19_2_54	wild		

Phenotypic data collection

Mungbean seeds harvested from this season have been used for phenotyping as seed size and color. Ten seeds from each accession were randomized as representatives of its population and used them for phenotypic data collection. Seeds will be scanned in the EPSON scanner after seed positions were made. Seed positions have been set into 3 positions for measurement as of SL, SH and SW (Figure 1.) and for seed color has three color components as R, G, and B (Red, Green and Blue), then measured the seed size and color by using ImageJ 1.50i program (Figure 2 and Figure 3). Average value of ten seeds will be calculated as median both seed size and seed color.

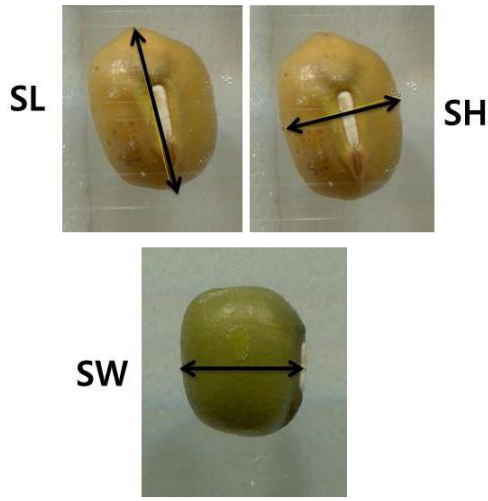

Figure 1. Three positions for seed size measurement

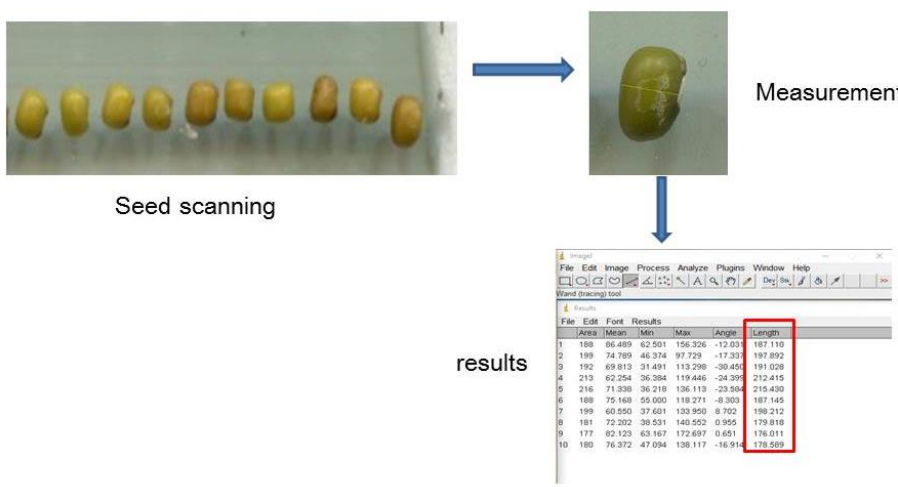

Figure 2. Seed size measurement

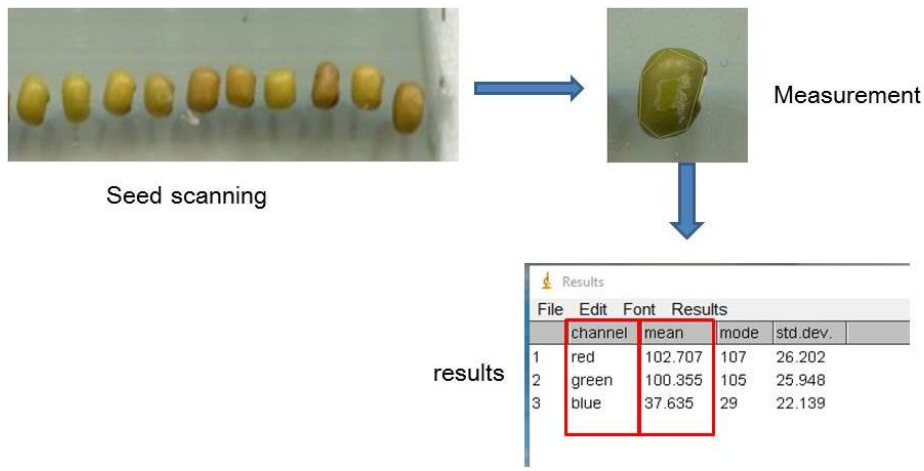

Figure 3. Seed color measurement

GBS approach and SNPs detection

GBS approach (Elshire et al., 2011) was done by digesting genomic DNA using APeKI restriction enzyme. Digested DNA fragments were ligated to sequencing adapters for repairing between primer sites and end sites, which carrying unique barcodes of each accession of mungbean. The ligated fragments were PCR amplified for specific primers. The library was sequenced by using Illumna HiSeq 2000 and sequencing data were aligned to mungbean reference genome by using BWA package (Li and Durbin, 2009). Variation of sequencing data were identified from the alignment data by using Sequence Alignment Map (SAM tools package) (Li et al., 2009).

SAM tools were used to generate a pileup of read bases using the alignments to a reference sequence, and BCF tools were used to call variant from the output of the SAM tools mpileup command. The variant calling results were stored in Variation Calling Format (VCF) file. The filtration obtained single nucleotide polymorphism (SNP) with at least one sequencing depth in each accession.

Genome-Wide Association Studies (GWAS)

Association analysis was performed using Traits Analysis by aSSociation, Evolution, and Linkage (TASSEL) (Bradbury et al., 2007). General linear model (GLM) was used to associate between markers and mean phenotypic values, in order to identify a significant marker that could be associated with a measured trait. A total of 18,171 SNP markers of 209 cultivated mungbean accessions were used in the performance. Minimum allele frequency of 0.005 was set during filtering stage to construct Manhattan plots as results.

RESULTS

Variation of seed size and color in natural population

Mungbean from each accession was selected for ten seeds to measure as seed size and seed color. Seed size was measured as SL, SH and SW, ranging from 2.71-6.37 mm, 1.89-4.57 mm and 2.06-4.53 mm, with the means of 4.89±0.69 (Mean±SD), 3.76±0.45 and 3.73±0.44, respectively. Seed color was defined as R, G and B, ranging from 33.70-157.52, 34.60-128.51 and 22.04-63.31, with the means of 99.33±16.69, 92.76±14.73 and 40.74±8.27, respectively (Table 5).

Table 5. Descriptive statistics of six seed traits for 218 accessions. SD indicates standard deviation, SL indicates seed length, SW indicates seed width, SH indicates seed height, R indicates red, G indicates green and B indicates blue.

Traits	Range	Mean±SD
SL	2.71-6.37 (mm)	4.89±0.69
SH	1.89-4.57 (mm)	3.76±0.45
SW	2.06-4.53 (mm)	3.73±0.44
R	33.70-157.52	99.33±16.69
G	34.60-128.51	92.76±14.73
B	22.04-63.31	40.74±8.27

Phenotypic structure of seed size and color variation

A total 218 mungbean accessions were examined the simple correlation among three seed size (Table.6) and three seed color traits (Table.8). In the correlation matrix of seed size and seed color traits, in overall; all traits are positively correlated with each other within each group.

Principal component analysis (PCA) was performed to identify the major sources of variation in the seed size and color traits among mungbean accessions. PC1 and PC2 of seed size explained 97.07% (PC1 92.05%, PC2 5.02% and PC3 2.93%) (Fig 4). The remaining variances spread over one additional component vectors and explained only a small part of the total variations. PC1 and PC2 of seed color explained 98.13% of the variations for the three traits (PC1 79.04%, PC2 19.09% and PC3 1.87%) (Fig 5). For seed size, the highest value is the correlation between SH and SW is 0.912 (Table.6), this means that, PC1 mainly described the variation in SH and SW with higher values of 0.967 and 0.963, however PC2 captures primarily the differences in SH and SW with value of 0.131 and 0.181 (Table 7). In seed color, the highest value is the correlation between R and G is 0.941 (Table.8). PC1 mainly described the variation in SH and SW with higher values of 0.939 and 0.960, however PC2 captures primarily the variation in B with value of 0.658 (Table 9).

Table 6. Correlation matrix for seed size. SL indicates seed length, SW indicates seed width, SH indicates seed height.

Traits	SL	SH	SW
SL		0.870	0.860
SH			0.912
SW			

Table 7. Correlations between variables. SL indicates seed length, SW indicates seed width, SH indicates seed height.

Traits	PC1	PC2	PC3
SL	0.948	-0.317	-0.022
SH	0.967	0.131	0.219
SW	0.963	0.181	-0.198

Table.8. Correlation matrix for seed color. R indicates red, G indicates green and B indicates blue.

Traits	R	G	B
R		0.941	0.511
G			0.575
B			

Table 9. Correlations between variables. SL indicates seed length, SW indicates seed width, SH indicates seed height.

traits	PC1	PC2	PC3
R	0.939	-0.303	-0.162
G	0.960	-0.220	0.172
B	0.753	0.658	-0.017

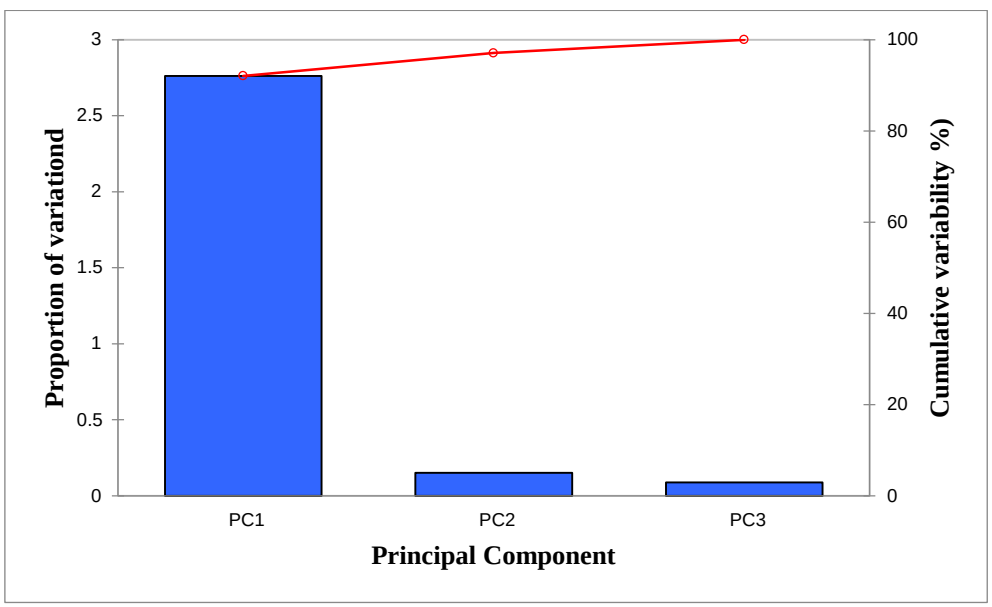

Figure 4. Principal component analysis of three seed size traits. Proportion of variation (eigenvalue) is the location of association between the components and the original variables explained by each individual principal component, while cumulative variability is the percentage of components contained in variation of original variables.

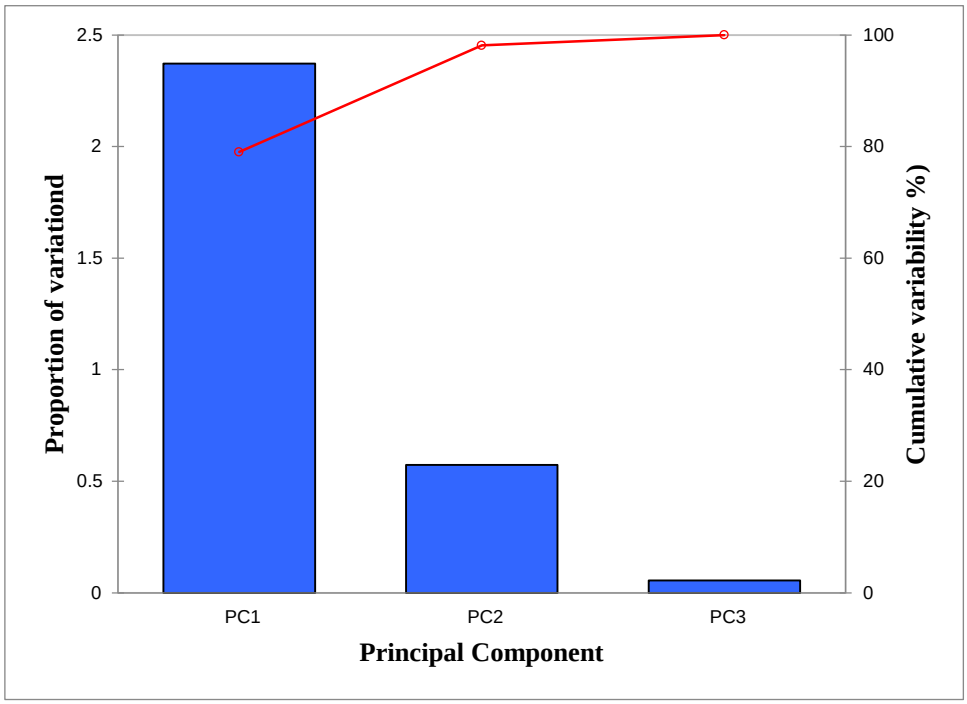

Figure 5. Principal component analysis of three seed color traits. Proportion of variation (eigenvalue) is the location of association between the components and the original variables explained by each individual principal component, while cumulative variability is the percentage of components contained in variation of original variables.

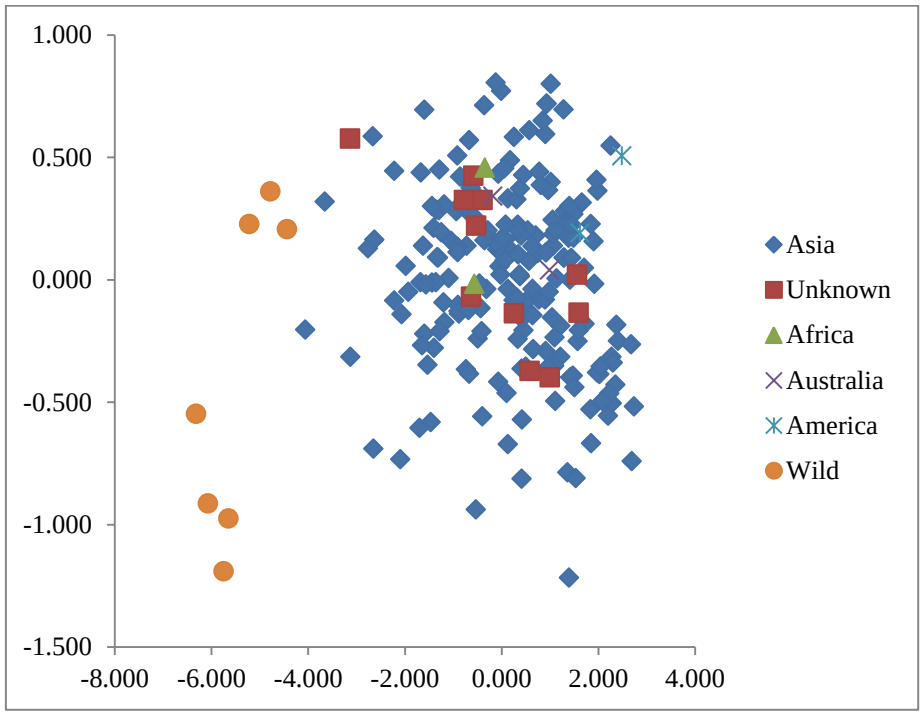

Figure 5. PCA scatter plots showing the distribution of mungbean accessions seed size.

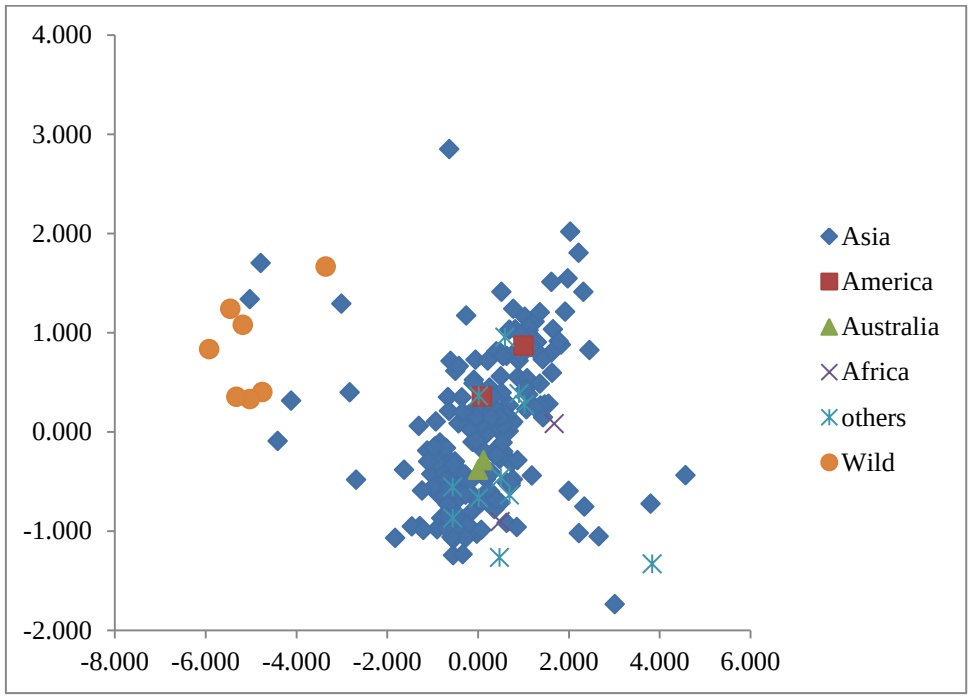

Figure 5. PCA scatter plots showing the distribution of mungbean accessions in seed color

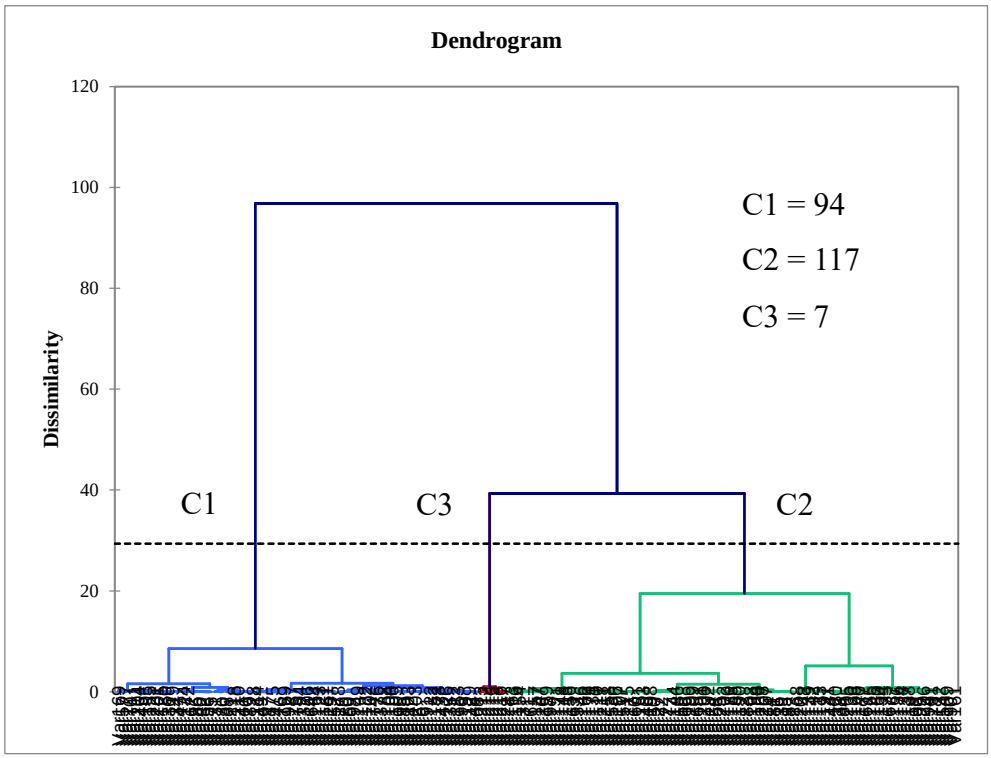

Fig 6. A dendrogram based on seed size. C indicates cluster. The diagram is grouped into three clusters: C1, C2 and C3 consisting of 94, 117 and 7 accessions, respectively. C1 is the biggest seed size and C3 is the smallest. All wild accessions are included in C3.

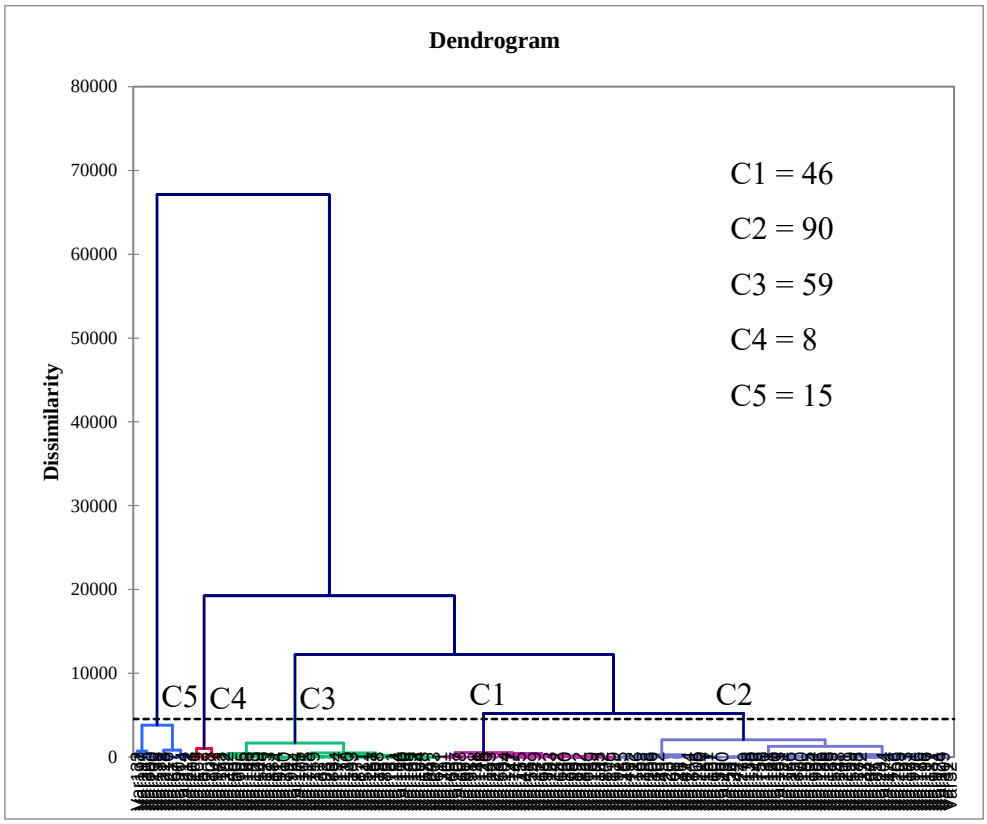

Fig 7. A dendrogram based on seed color. C indicates cluster. The diagram is grouped into five clusters: C1, C2, C3, C4 and C5 consisting of 46, 90, 59, 8 and 15 accessions, respectively. All the green mungbean accessions were clustered into C1, C2 and C3 depending on their brightness of the color. C4 consists of mostly yellow mungbean accessions. C5 consists of mostly black mungbean accessions.

Dendrograms based on seed size and color were grouped into clusters with three clusters for seed size (Fig 6.) and five clusters for seed color (Fig 7.). Each accession that is close to each other was grouped into the same cluster. The three clusters of seed size are obviously classified with cluster one as the biggest to cluster three as the smallest accessions. All wild mungbean accessions were grouped in the cluster three. Five clusters of seed color were classified based on color of mungbeans as green, yellow and black, and the brightness of the color. In cluster one to cluster three consist of mostly green mungbean accessions, due to brightness and darkness of color as indicator of classification, while cluster four are mostly yellow accessions and some green accessions that are close to the yellow mungbean, whereas cluster five mainly are black accessions and green accessions that have darker seed color.

Genome-wide association study

Identification of a significant marker associated with a trait for General linear model (GLM) in TASSEL is needed to pass the significant threshold of –log10 of p-value at 7. As the analysis results, General linear model in TASSEL identified a SNP marker associated with two traits as red and green on Chr04, locus position at 3374025 bp at p-value<0.005

(minimum allele frequency), while other traits as blue, seed length, seed height and seed width had no significance associated with any SNP markers (Fig 8). However, a locus position at 3374025 bp is in the new mungbean reference genome and its annotation is in the process. Therefore, to identify the candidate genes, we used the previous mungbean reference genome to identify which this locus position matches to the position between 17,682,006 to 17,685,676 at chromosome 4.

Table 10. A significant marker associated with measured traits (P-value<0.005). Observation shows two genotypes as 206 accessions with genotype CC and 3 accessions with genotype TT, Allele_Estimate indicates the differences between two homozygotes.

Trait	Chromosome	Position	P-value	Observation		Allele_Estimate
R	Chr04	3374025	1.52E-10	206	C	52.95
				3	T	
G	Chr04	3374025	1.76E-09	206	C	42.31
				3	T	

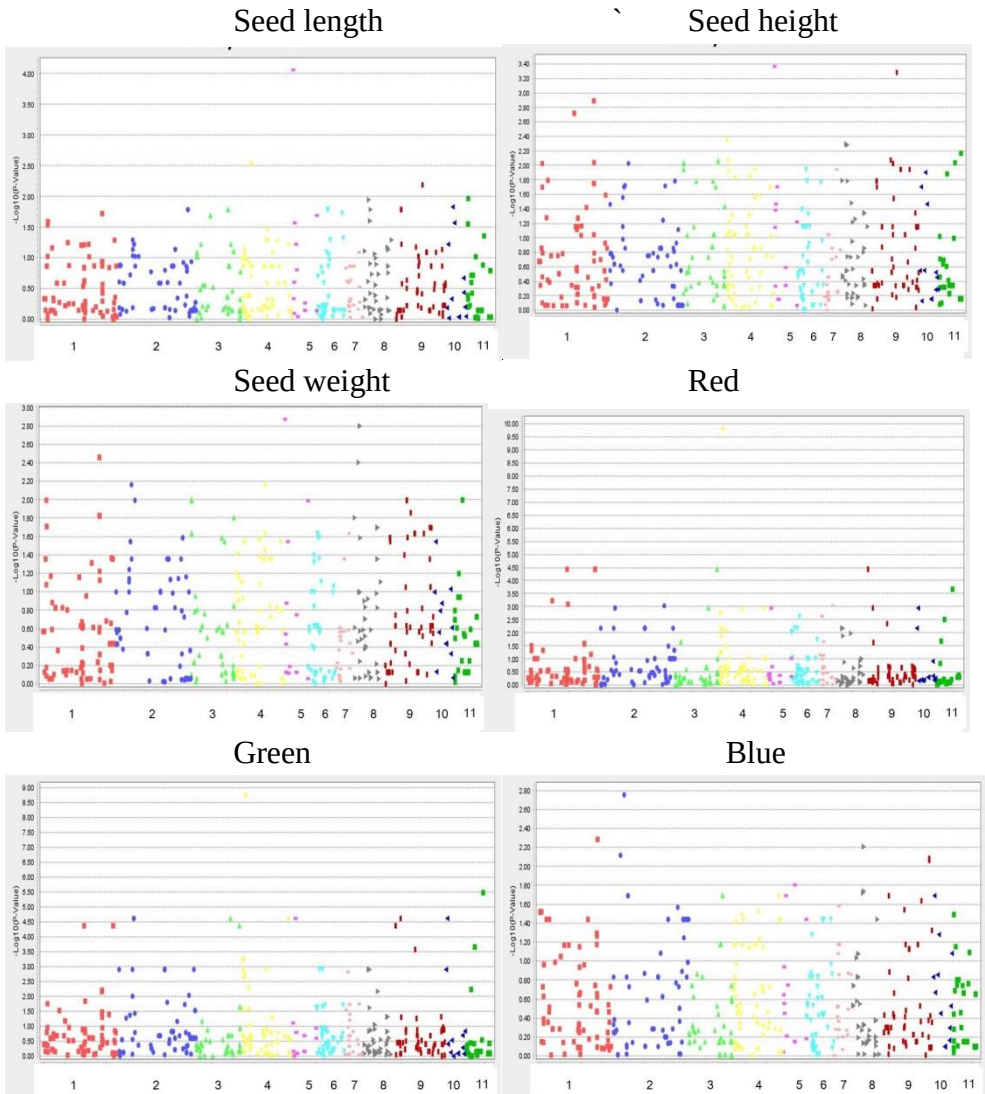

Fig. 8. Manhattan plots of p-value from GWAS produced by TASSEL. The y-axis is –log10 of p-values and x-axis is the chromosome number. A significant marker associated with red and green color was identified on chr04 at 3370425 bp with p-value <0.005.

Candidate gene identification

A significant SNP marker associated with seed traits as red and green identified on Chr04 at position 3370425 bp. By this position, it is in the new mungbean reference genome, however, the annotation of new mungbean reference genome has not been finished. Therefore, when we compared to previous mungbean reference genome is matched to the position between 17,682,006 to 17,685,676 (3,671bp) at chromosome 4 which is located in a gene of Vradi04g08950. The gene ontology annotation for this gene is related to several functions as pectinesterase activity (GO:0030599), cell wall modification (GO:0042545) and cell wall (GO:0005618). To compare the gene sequence of mungbean to soybean, using https://phytozome.jgi.doe. gov/pz/portal.html#!search?show=BLAST for glymax cine. The responded homologous genes of soybean were Gylyma.19g025500, Gylyma.09g234800, Gylyma.18g262500, Gylyma.13g06900 and their functions mostly are the same as mungbean gene of Vradi04g08950. One of the notable genes obtained is Gylyma.18g262500, this gene is located in a QTL interval that is involved in multiple traits as seed, pod and yield (https://www.soybase.org/).

DISCUSSIONS

Phenotypic structure of seed size and color variation

In correlation analysis, both seed size and color were found positively correlated. To obtain deeper insight into the relationship, PCA was performed to identify the major sources of variation in seed morphology. For seed size, PC1 captured SH and SW, while PC2 primarily described the differences in SH and SW. whereas seed color of PC1 mainly captured R and G, while PC2 primarily described the differences in B.

Seed size and color were grouped into clusters. Three clusters of seed size had the biggest to the smallest, while all wild mungbean accessions were grouped into the smallest cluster. Green, yellow and black accessions were measured to obtain three color components as red, green and blue; these three color components grouped three colored mungbean accessions into five clusters. Yellow and black mungbean accessions were separated into their own clusters, while some green mungbean accessions were included in the cluster of yellow and black. The color of mungbean depends on composition of pigmentation. Anthocyanins, chlorophylls and carotenoids are the major pigments that impart various shades of color to

plant parts. The intensity of pigmentation increases with the increase in concentration of anthocyanins and chlorophyll (Swain 1965).

Genome-wide association study

Since a significant marker associated with two traits was identified by GWAS as a color component like red and green color at a locus position of new mungbean reference genome. However, the annotation of new mungbean reference genome has not been finished. Therefore, to search for the gene candidates controlling the traits appeared, it is needed to compare to the annotation of previous mungbean reference genome and the location of nearest soybean gene candidates in the QTL interval that involved traits. Furthermore, further investigation will need to be carried out after new mungbean reference genome has been finished, to find accurate and nearest candidate genes to the position of new mungbean reference genome.

As the results, a significant marker associated with two traits was identified at the same position. The alleles possibly affect both spectrum or more than one pigment. The seed color is affected by not just one component, but at least two. Plant pigments mostly consist of chlorophyll, carotenoids, anthocyanins and betalains. The green is possibly from chlorophyll-like pigment, while the black is probably anthocyanin or its

derivative. It's possible that yellow comes from another pigment as well. Therefore, if it is scored just based on green color will miss the anthocyanin component, and if it focuses on black you will miss the chlorophyll side, and those two can interact or mask each other. Seed coat color was controlled by mono-gene and the seed luster had a complicated genetic model and there was no relationship between seed coat color and luster. The inheritance of anthocyanin coloration on different parts of plant was also agreeable to mono-gene (Wang li-xia, 2013). Whereas the traits that had no significant markers associated with, as blue color component and seed size like seed length, seed height and seed width. They are considered to be determined by several factors with small effects as rare alleles are in the association with these traits. Further study related to this result is needed to know the major color pigment component in the color of these mungbean as green, yellow and black and their luster, in order to know more details and gene controlled these traits for coping with future mungbean improvement.

In conclusion, although a significant marker associated with two traits was identified, it is still needed to investigate again, in order to identify candidate genes in the new mungbean reference genome after finished annotation and further study need to be conducted for the confirmation of these results.

REFERENCES

Benjamin Brachi, Geoffrey P Morris and Justin O Borevitz. 2011. Genome-wide association studies in plants: the missing heritability is in the field Brachi et al. Genome Biology 2011, 12:232.

Claudia Perea, Juan Fernando De La Hoz, Daniel Felipe Cruz, Juan David Lobaton, Paulo Izquierdo, Juan Camilo Quintero, Bodo Raatz and Jorge Duitama. 2016. Bioinformatic analysis of genotype by sequencing (GBS) data with NGSEP. BMC Genomics 201617 (Suppl 5): 498. DOI: 10.1186/s 12864-016-2827-7.

Elshire, Robert J.; Glaubitz, Jeffrey C.; Sun, Qi; Poland, Jesse A.; Kawamoto, Ken; Buckler, Edward S.; Mitchell, Sharon E. 2011. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLOS ONE*. 6 (5): e19379. doi:10.1371/journal.pone. 0019379.ISSN1932-6203. PMC3087801. PMID21573248.

Filippo Biscarini, Paolo Cozzi, Laura Casella, Paolo Riccardi, Alessandra Vattari, Gabriele Orasen , Rosaria Perrini, Gianni Tacconi, Alessandro Tondelli, Chiara Biselli, Luigi Cattivelli, Jennifer Spindel, Susan McCouch, Pamela Abbruscato, Giampiero Valé, Pietro Piffanelli, Raffaella Greco. 2016. Genome-Wide Association Study for Traits

Related to Plant and Grain Morphology and Root Architecture in Temperate Rice Accessions. PLoSONE11(5): e0155425.doi:10.1371/ journal. pone.0155425.

Gilles Boutet, Susete Alves Carvalho, Matthieu Falque, Pierre Peterlongo, Emeline Lhuillier, Olivier Bouchez, Clément Lavaud, Marie-Laure Pilet-Nayel, Nathalie Rivière and Alain Baranger. 2016. SNP discovery and genetic mapping using genotyping by sequencing of whole genome genomic DNA from a pea RIL population. BMC Genomics 201617:121. DOI: 10.1186/s12864-016-2447-2.

Inka Gawenda, Patrick Thorwarth, Toren Ginther, Frank Ordon and Karl J. Schmid. 2014. Genome-wide association studies in elite varieties of German winter barley using single-marker and haplotype-based methods. doi:10.1111/pbr.12237. Plant Breeding, 134, 28 –39 (2015).

Jiangfeng; Zhao, Xiaoqing; Laroche, André; Lu, Zhen-Xiang; Liu, HongKui; Li, Ziqin. 2014. Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. Plant Genetics and Genomics. 5: 484. doi:10.3389/fpls.2014.00484. PMC 4179701. PMID 25324846.

Klein RJ; Zeiss C; Chew EY; Tsai JY; et al. 2005. Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science*. 308

(5720): 385–9. doi: 10.1126/science.1109557. PMC 1512523. PMID 15761122.

P.H. Munguiat, M.R Lantican, A.L Carpena and I.G. catedral. 1978. Inheritance of seed color, size and luster in mungbean. The philippines journal of crop science.

Shindo C, Bernasconi G, Hardtke CS. 2007. Natural genetic variation in Arabidopsis: tools, traits and prospects for evolutionary ecology. Ann Bot 2007, 99:1043-1054.

Stranger B, Stahl E, Raj T. 2011. Progress and Promise of Genome-Wide Association Studies for Human Complex Trait Genetics. Genetics. 2011 Feb 187(2): 367-383.

Sue K. Kim, Ramakrishnan M. Nair, Jayern Lee, and Suk-Ha Lee. 2015. Genomic resources in mungbean for future breeding programs. Front Plant Sci. 2015; 6: 626. doi:10.3389/fpls. 2015.00626.

Yingjie Xiao, Haijun Liu, Liuji Wu, Marilyn Warburton and Jianbing Yan. 2016. Genome-wide Association Studies in Maize: Praise and Stargaze. Mol. Plant. 10, 359-374. Cellpress partner journal.

Wang Lixia, Cheng Xuzhen, Wang Suhua, Liu Yan. 2013. Inheritance of Several Traits in Mungbean (*Vigna radiata*), Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing 100081, China.