A DISSERTATION FOR DEGREE OF MASTERS OF SCIENCE

# Genome Divergence between *Vigna angularis* & *Vigna nakashimae*

FEBURARY, 2014

BY

KHUSHBOO RASTOGI

MAJOR IN CROP SCIENCE AND BIOTECHNOLOGY

DEPARTMENT OF PLANT SCIENCE

THE GRADUATE SCHOOL OF SEOUL NATIONAL UNIVERSITY

# Genome Divergence between
# *Vigna angularis* & *Vigna nakashimae*

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

BY
KHUSHBOO RASTOGI

MAJOR IN CROP SCIENCE AND BIOTECHNOLOGY
DEPARTMENT OF PLANT SCIENCE

FEBURARY, 2014

APPROVED AS A QUALIFIED DISSERTATION OF KHUSHBOO RASTOGI
FOR THE DEGREE OF MASTER
BY THE COMMITTEE MEMBERS

FEBURARY, 2014

CHAIRMAN

Seo Hak-Soo, Ph.D.

VICE-CHAIRMAN

Lee Suk-Ha, Ph.D.

MEMBER

Paek Nam-Chon, Ph.D.

# Genome Divergence between

# *Vigna angularis* & *Vigna nakashimae*

KHUSHBOO RASTOGI

## ABSTRACT

The Azuki bean, *Vigna angularis (2n=2x=22),* is one of the important legum e crop grown in the world. Comparing the genome of domesticated *(Vigna a ngularis)* and undomesticated *(Vigna nakashimae)* forms of Azuki bean can f acilitate crop improvement. We used the Next-generation massively parallel DNA sequencing technologies which provide ultrahigh throughput at a subst antially lower unit data cost. However, the data generated is very short read l ength sequences and constructing de novo assembly from it is extremely chal lenging. Here, we describe a novel method for finding genome divergence be tween the domesticated and the wild variety of Azuki bean. *We* de novo sequ enced and assembled the *Vigna angularis* and *Vigna nakashimae* genome, ac hieving an N50 contig size of approximately 11 and 7 kilo base pairs (kbp) re spectively. The genome size of Azuki bean is estimated to be around 545 me ga-bases (Mb). We predicted 45,985 protein-coding genes in *Vigna angulari s*, 70% more than that of Arabidopsis, approximately similar to poplar and so ybean. For *Vigna nakashimae* we predicted 38,965 protein-coding genes whi ch is 40% more than Arabidopsis and approximately similar to potato. *Vigna*

*angularis diverged from Vigna nakashimae* approximately 1.9 million years ago. The data obtained from structural translocations *and* gene categories fro m the evolutionary relationship between *Vigna angularis* and *Vigna nakashi mae* suggest that these genes can be a probable cause for the domestication o f Azuki bean. The development of this de novo short read assembly method c reates new opportunities for building reference genome and carrying out acc urate analyses of unexplored genomes in a cost effective manner as well as o vercomes the limitations of the re-sequencing method for discovering structu ral translocations.

Keyword: *Vigna angularis, Vigna nakashimae,* structural translocations, *dom estication,* genome divergence, Next-generation sequencing.

Student number: 2012-22611

# CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# INTRODUCTION

Azuki bean has been an important legume and ceremonial food in East Asia for thousands of years (Yamaguchi 1992; Lee 2013). Despite of its agricultural and biological importance, knowledge about its genetics and genome is very limited. We therefore, sequenced and assembled the genome for both wild (*Vigna nakashimae*) and domestic (*Vigna angularis*) Azuki bean.

Azuki bean is a member of the *Phaseoleae* tribe. This tribe includes many important legume crop species such as soybean (*Glycine max*), cowpea (*Vigna ungiculata*), common bean (*Phaseolus vulgaris*), mung bean (*Vigna radiata*), alfalfa (*Medicago sativa*), chickpea (*Cicer arietinum*), clover (*Trifolium spp.*), pea (*Pisum sativum*), lentil (*Lens culinaris*), barrel medic (*Medicago truncatula*) and lotus (*Lotus japonicus*). The *Medicago truncatula* and *Lotus japonicus* have emerged as important model species for understanding legume genomics (Cannon, May et al. 2009). *Vigna angularis* var. *nipponensis* is considered as the conspecific wild progenitor of *V.angularis* (Kaga, Isemura et al. 2008). Apart from *nipponensis, Vigna nakashimae* is also considered as the wild form of *Vigna angularis* (Somta, Kaga et al. 2006; Yoon, Lee et al. 2007). Azuki bean has 11 pairs of chromosomes (2n = 22) and is a self-fertile annual vine. *V.nakashimae* and *V.angularis* shows several morphological and physiological differences: wild redbean has a straggling and climbing architecture with many lateral branches, produce black wrinkled seed and have black pod which shatters easily. Whereas the domesticated red bean has erect architecture, with straw coloured pods and round red seed. These differences are collectively called

as the domestication syndrome, occur due to structural variations or single nucleotide polymorphism or may be due to selection pressure over several thousands of years for adaptation, to cultivated environments and human nutritional requirements and preferences (Kaga, Isemura et al. 2008).

Previously the genome of the crop plants like rice (Matsumoto, Wu et al. 2005), soybean (Schmutz, Cannon et al. 2010), sorghum (Paterson, Bowers et al. 2009), maize (Schnable, Ware et al. 2009), grape (Jaillon, Aury et al. 2007) poplar (Tuskan, Difazio et al. 2006) and eucalyptus (Myburg, Grattapaglia et al. 2011) have been derived using Sanger sequencing technology. With the recent development of NGS technologies, there is a continuous decline in the sequencing cost. This unbridled and explosive innovation has led to the opening of new opportunities for genome-wide association studies (GWAS), population genomics, characterization of rare polymorphisms, and personal clinical genomics (Vezzi, Narzisi et al. 2012). NGS technology provides us better opportunities for understanding the complexity of the existing genomes and the strong relationships between the genotype and evolution. In a nutshell, NGS technology provides us a powerful tool for the discovery of the new genes and in the identification of genetic variations by investigating the functional and evolutionary divergence among close or distinct relatives of species. This can be beneficial in determining the cause for domestication.

For the Azuki bean genome, we carried out a novel method to explore genome divergence or genetic diversity between the wild and the cultivated red bean. Here, we *de novo* sequenced and assembled the *Vigna angularis* and *Vigna nakashimae* genome. The assembled genomes were then compared to identify structural variations. The analysis of the structural variations has helps to understand as to how the genomes have shaped

themselves in the course of the evolution. It can also help to find the species-species specific genes which could lead to morphological and physiological variation.

In this study, we focused on estimating the genome size and predicting the gene count for both wild and domesticated red bean. Apart from this we also tried to find syntenic regions and evolutionary history. In succinct, this analysis explains domestication, evolutionary and genetic history of Azuki bean and its divergence from its wild form.

# LITERATURE REVIEW

## Next Generation Sequencing Technology

In a matter of just few years sequencing technologies have undergone an unbridled and explosive innovation. The genomics community is highly benefited with the expansion in the NGS technologies as, now we can *de novo* sequence for a number of species (Li, Fan et al. 2009; Kuczynski, Costello et al. 2010), detect methylated regions in genome (Wu, Yi et al. 2011), find DNA sequence variation within a species (Elshire, Glaubitz et al. 2011; Siu, Zhu et al. 2011) and even can perform gene expression profiling (Varshney, Hiremath et al. 2009; Wall, Leebens-Mack et al. 2009; Hiremath, Farmer et al. 2011).

Prior to the advent of next-generation sequencing (NGS) technology, Sanger sequencing has been the unrivaled approach for characterizing the genomes for numerous organisms including model plants and crops species like rice (Matsumoto, Wu et al. 2005), soybean (Schmutz, Cannon et al. 2010), sorghum (Paterson, Bowers et al. 2009), maize (Schnable, Ware et al. 2009), grape (Jaillon, Aury et al. 2007) poplar (Tuskan, Difazio et al. 2006) and eucalyptus (Myburg, Grattapaglia et al. 2011). However, with the advancement and continuously evolving nature of NGS technologies, there is a continuous decline in the sequencing cost and an increase in the sequence read lengths.

There is an exponential increase in the sequence throughput from the different sequencing platform (Table 1). The storing and the management of

the data generated is very challenging task. Apart for this, the primary, secondary and tertiary analysis solutions like quality control, base calling, de novo assembly, alignment to a reference genome, variant calling, Chip-Seq, transcriptome analysis are necessary to make sense of the larger volumes of sequence data. To overcome this number of tools/software packages have been developed in last few years. Some of these tools are listed in Table 1 (Thudi, Li et al. 2012).

NGS technology provides a powerful tool for the discovery of the genes and in the identification of genetic variations in close or distinct relatives of species, which can be a cause of domestication. NGS also helps in determining the genetic basis for the phenotypic differences between the species by comparing the whole genome. Therefore, the plant breeders use more of the NGS technology to introduce diversity on the variety (Reif, Zhang et al. 2005) without altering the plant performance and product quality.

**Table 1. Some important tools for analysis of NGS data**

| Tool/Program/Assemblers | |
|---|---|
| ***De novo* alignment** | |
| ABySS | SSAKE |
| EULER-SR | MIRA2 |
| SOAP*denovo* | |
| Velvet | VCAKE |
| **Alignment to a reference genome** | |
| Bowtie | SeqMap |
| Exonerate | SHRiMP |
| GenomeMapper | Slider |
| GMAP | SOAP |
| MAQ | SSAHA |
| PASS | Vmatch |
| RMAP | Zoom |
| **SNP/Indel Discovery** | |
| ssahaSNP | PolyBayesShort |
| PyroBayes | Alpheus |
| **Transcriptomics** | |
| G-Mo.R-Se | QPalma |
| MapNext | TopHat |
| **Genome annotation/genome browser/alignment viewer/assembly database** | |
| EagleView | SAM |
| LookSeq | XMatchView |
| MapView | |
| **Miscelleneous** | |
| CNV-Seq | PeakSeq |
| FindPeaks | SISSRs |
| MACS | |

## *De Novo* Assembly

*De novo* whole-genome sequence assembly (WGSA) is done when there is no prior information or knowledge of the underlying genome's structure. For doing *de novo* assembly we reconstruct the genome sequence from a large number of short sequences (i.e., reads). The basic principle behind the assemblers is that if two reads share a sufficiently long subsequence (a prefix "matching" a suffix) then they can be assumed to have originated from the same locations in the genome (Vezzi, Narzisi et al. 2012). Thus performing *de novo* assembly for the genome can be extremely challenging and herculean task, if we come across complicated situation like: large scale duplications, contaminations from the vector, repeated sequences, genome polymorphism etc. These hurdles can be conquered up to a certain extent that is with limited success only when reads are available with high-coverage, pre-processing with repeat-maskers, k-mer based error corrections or gap-filling attempts are done using various heuristics (Vezzi, Narzisi et al. 2012). Improved base-callers (using a Bayesian priors on the genome's base distributions) and novel assembly methods (combining short- and long-range information in a dovetail fashion) have been more effective in improving base-accuracy and in resolving repeat boundaries (Menges, Narzisi et al. 2011; Narzisi and Mishra 2011).

## Structural variation

Structural variants are the major contributors for the variations in the genome and are considered as they help in better understanding as to how the genome has modified itself throughout evolutionary history. Structural variation comprises of balanced and unbalanced forms of variations. Unbalanced forms of variations include deletion, duplications, and insertions – that change the number of copies of a segment of the genome while the balanced forms include inversions and translocations – that do not alter the copy number of the genome (Raphael 2012; Weischenfeldt, Symmons et al. 2013).

The basic mechanisms behind the cause of these SVs are homologous recombination, non-replicative non-homologous repair and replication based mechanisms (Gu, Zhang et al. 2008; Hastings, Lupski et al. 2009; Yang, Luquette et al. 2013). Structural variations vary widely in size in complexity and are therefore categorized as small scale and large scale structural variations. Large scale structural variations can be visualized directly through cytogenetic techniques whereas detecting small scale structural variations is more tedious and hard than single nucleotide polymorphisms (Feuk, Carson et al. 2006).

There are three approaches for detecting SVs from next generation sequencing data. The first one is based on re-sequencing. In this approach, sequence reads of the closely related individual genome are aligned to the reference genome. This method is very sensitive and sometime notoriously difficult as it can lead to false identification of structural variations. The error occurs when there are repetitive sequences near break point, if the genome

has multiple stated and complex architectures, and if recurrent variants exits at same locus (Kim, Lee et al. 2010; Raphael 2012).

The second approach is similar to the first one except that here we use the unmapped reads of the individual instead of the actual reads obtained after sequencing to find the structural variations (Chekanov, Boulygina et al. 2010).

The third approach is for the *de novo* assembled genomes. In this method, the overlapping reads are assembled to construct the genome of the individual. The assembled genome of one species or variety are then compared to the another species or variety to identify all types of variants. This method is best for finding all types of variants if the genomes are assembled properly (Feuk, Carson et al. 2006; Raphael 2012).

**Synonymous Substitutions, Non-Synonymous Substitutions and Evolutionary Time**

One of the powerful tool, for appraising the mechanism of evolutiona ry divergence in DNA sequences, is to estimate the number of non-synonym ous (amino-acid replacing) and synonymous (silent) substitution termed as K a and Ks, respectively. (IMURA ; Gillespie 1991; WenHsiung 1997; Nekrute nko, Makova et al. 2002; Pal, Papp et al. 2006). Ka reflects the non-synony mous substitutions per non-synonymous site whereas Ks reflects the synony mous substitutions per synonymous site. The Ka/Ks ratio (denoted as ω) is w idely used as an estimator of selective strength for DNA sequence evolution, with $\omega > 1$ indicating positive selection, $\omega < 1$ indicating purifying (negative) selection, and ω close to or equal to 1 indicating neutral mutation.

When the nucleotide substitution between the two species is negligibl e that is not more than one nucleotide difference, it can be estimated manuall y by counting the silent and the amino acid altering nucleotide. However, if t he frequency is more than, we have to use statistical methods (Zhang and Yu 2006).

In general the rate of synonymous substitution is much higher than th at of non-synonymous substitution and is similar for many different genes. K s represent the neutral evolutionary rate that is it is assumed to be same for al l the sites. Therefore, synonymous substitutions can be used as a molecular c lock for dating the evolutionary time of closely related species (Kafatos, Efst ratiadis et al. 1977; Kimura 1977; Miyata and Yasunaga 1980; Miyata, Yasu naga et al. 1980; Perler, Efstratiadis et al. 1980; Nei and Gojobori 1986).

# MATERIALS AND METHODS

## *De Novo* Sequencing

For sequencing we used two accession of Azuki bean *Vigna angularis* var. IT213134 and *Vigna nakashimae var.* IT178530 respectively for this research work. These two accessions of Azuki bean were sequenced using Illumina Hi-seq 2000. The benefits of using Illumina Hi-seq 2000 are it produces unprecedented output with high accuracy, breakthrough user experience and unmatched cost effectiveness.

## *De Novo* Assembly

The wild and the cultivar accessions of Azuki bean were assembled into contigs using ABySS genome assembler (Simpson, Wong et al. 2009) ( http://www.bcgsc.ca/platform/bioinfo/software/abyss) version 1.3.3 with k-mer frequency of 84 and quality threshold value as 20. ABySS is a *de novo,* parallel, paired-end sequence assembler that is designed for short reads. The single-processor version is useful for assembling genomes up to 100 Mbp in size. The parallel version is implemented using MPI and is capable of assembling larger genomes. The output of ABySS is a set of contigs assembled from short reads.

## Ab initio gene prediction

We used the ab initio method to predicted gene from the contigs usin g geneid version 1.4.4 (Guigó, Knudsen et al. 1992) (http://genome.crg.es/so ftware/geneid/). Geneid predict genes in anonymous genomic sequences desi gned with hierarchical structure. The prediction of gene was based on the crit eria that the minimum length of a gene should be 200Bp. The reason for cons idering this criterion is that a gene should be long enough to perform its resp ective function.

## Homology search

The homologies between wild and cultivar Azuki bean were predicte d using blastp (Altschul, Gish et al. 1990) (http://blast.ncbi.nlm.nih.gov/Blast. cgi) with an E-value threshold of 1e-100. In protein-protein blast, protein seq uence is given as query which returns the most similar protein sequence from the protein database that is specified by the user. This helps us to find the fu nction of a newly sequenced gene, predict new members of gene families, an d explore evolutionary relationships.

**Synteny and Evolutionary relationship**

To evaluate the synteny relationship between the wild and cultivar Az uki bean MCScanX (Wang, Tang et al. 2012) (http://chibba.pgml.uga.edu/m cscan2/) was used. MCScan (Multiple Collinearity Scan) identifies collinear blocks in genomes or sub-genomes and then conduct multi-alignments of col linear blocks using collinear genes as anchors (Tang, Bowers et al. 2008; Tan g, Wang et al. 2008) whereas MCScanX is a diverse tool for the evolutionary analyses of synteny and collinearity, aiding efforts to construct gene families using collinearity information, infer gene duplication modes and enrichment s, characterize collinear genes with nucleotide substitution rates, detect collin ear tandem arrays, perform statistical analyses of duplication depths and colli near orthologs, and analyse collinearity within gene families. MCScanX ena bles rapid and convenient conversion of synteny and collinearity information into evolutionary insights (Wang, Tang et al. 2012).

The degree of evolutionary change is generally estimated from the no n-synonymous (amino-acid replacing) and synonymous (silent) substitution r ates among protein-coding sequences, termed as Ka and Ks, respectively (IM URA ; Gillespie 1991; WenHsiung 1997; Nekrutenko, Makova et al. 2002; P al, Papp et al. 2006). Ka reflects the non-synonymous substitutions per non-s ynonymous site whereas Ks reflects the synonymous substitutions per synon ymous site. The Ka/Ks ratio (denoted as ω) is widely used as an estimator of selective strength for DNA sequence evolution, with $\omega > 1$ indicating positiv e selection, $\omega < 1$ indicating purifying (negative) selection, and ω close to or equal to 1 indicating neutral mutation.

## Structural Translocations

To visualize the structural translocations between *Vigna angularis* and *Vigna nakashimae* we used Circos (Krzywinski, Schein et al. 2009). Circos facilitates the identification and analysis of similarities and differences arising from comparisons of genomes. It is effective in displaying variation in genome structure and, generally, any other kind of positional relationships between genomic intervals. Circos uses a circular ideogram layout to facilitate the display of relationships between pairs of positions by the use of ribbons, which encode the position, size, and orientation of related genomic elements.

## Divergence Time

The divergence time is estimated if we know the number of nucleotide substitutions between two species (K) and the substitution rate $\nu$ , the divergence time is calculated as $K/(2\nu)$ (Haubold and Wiehe 2001). The nucleotide substitution value is basically the synonymous substitutions which occur in the nucleotide sequence. The Ks value is approximately constant throughout the genome as in case of synonymous substitution nucleotide gets changed but the protein formed remains same (silent mutation) and thereby performs similar function as before, whereas in case of non-synonymous substitution nucleotide changes leading to the formation of a new protein and thereby a new function is performed by the protein sequence. If this change is beneficial it is retained otherwise the less

effective are discarded. So divergence time can't be calculated using the non-synonymous substitution (Kafatos, Efstratiadis et al. 1977; Kimura 1977; Miyata and Yasunaga 1980; Miyata, Yasunaga et al. 1980; Perler, Efstratiadis et al. 1980; Nei and Gojobori 1986).

**Specific Gene Categories**

Specific genes which were retained and which had diverged between *Vigna angularis* and *Vigna nakashimae* were retrieved using MapMan tool (Thimm, Bläsing et al. 2004; Usadel, Nagel et al. 2005; Urbanczyk-Wochniak, Usadel et al. 2006; Usadel, Poree et al. 2009). MapMan is a user-driven tool that displays large datasets (e.g. gene expression data from Arabidopsis Affymetrix arrays) onto diagrams of metabolic pathways or other processes.

# RESULTS

## Overview of *De Novo* Sequencing and Assembly

We used Illumina Hi-seq 2000 technology for sequencing the two acc ession of Azuki bean IT213134 (*Vigna angularis*) and IT178530 (*Vigna nak ashimae*). The paired-end sequencing libraries were constructed with an inser t size of 300bp. In total, we generated ~ 15-Gb and ~ 16-Gb of usable sequen ce of *Vigna angularis* and *Vigna nakashimae* respectively. The sequencing d epth was around 34.49X and 32.40X respectively (Table 2).

We assembled the short reads using the ABySS - a *De Novo*, parallel, paired-end sequence assembler (Simpson, Wong et al. 2009). ABySS uses th e de Bruijn graph algorithm (Pevzner, Tang et al. 2001) and applies a stepwis e strategy to make it feasible to assemble the red bean genome. The algorith m is sensitive to sequencing errors, so we excluded the data generated from p oor libraries, filtered low-quality reads, and used the high-quality reads for d e novo assembly. We obtained ~0.68 million and ~1.2 million contigs for *Vig na angularis* and *Vigna nakashimae* respectively (Table 3). The N50 value o btained was ~11,000 and ~7500 respectively.

**Table 2. Sequencing status of 2 deep sequenced accession**

| Accession | Origin | Total number of reads | Sequencing depth (X) | Genome Size (Mb) |
|---|---|---|---|---|
| IT213134 (Vigna angularis) | South Korea | 186,177,263 | ~ 34.49 | ~ 545 |
| IT178530 (Vigna nakashimae) | South Korea | 174,886,064 | ~ 32.40 | ~545 |

**Table 3. ABySS Genome Assembly results**

| Accession | Total contig Number | Largest contig and size (bp) | Smallest contig size (bp) | N 50 | Number of contig longer than N50 |
|---|---|---|---|---|---|
| IT213134 (Vigna angularis) | 678,525 | 22388167 / 119,694 | 84 | 11,253 | 12,515 |
| IT178530 (Vigna nakashimae) | 1,170,289 | 3033294 / 96,540 | 84 | 7,451 | 17,708 |

**Estimating Genome Size**

To obtain the best estimate of the genome size, we used the coverage depth and the total data generated from the assembled contigs. The genome size is calculated form the formula:

**Genome size = total data generated / coverage depth**

In order to estimate the coverage depth, we calculated the mean k-mer cover age. For this we chopped our all reads into k-mer of length 24. We then coun ted the frequency with which each 24-mer represented in my data is found a mong all of the reads generated and created the frequency histogram. For the non-repetitive regions of the genome, this histogram produced an asymptote peak near 1 because of rare sequencing errors and an actual peak around 129. This peak value (or peak depth) is the mean k-mer coverage for my data. We then calculated the actual coverage of the genome using the formula:

$$N = M * L / (L - K + 1)$$

where N is the actual coverage of the genome, M is the mean k-mer coverage, L is the mean read length and K is the k-mer size. Genome size was estimate d to be approximated around 545 Mb from the ratio of total data and coverag e depth (Li, Fan et al. 2009) (Table 2).

## Ab initio gene prediction and Homology search

To predict the number of genes in the red bean genome, we used the ab-inito method. We predicted 67,835 genes in *Vigna angularis* and 58,888 genes in *Vigna nakashimae* consisting of complete and partial genes (Table 4). The number of complete genes was estimated to be 45,985 and 38,965 for the domesticated and undomesticated Azuki bean.

Orthologous gene pairs approximately around 0.14 million were first identified based on an all-against-all BLASTp search with an E-value cutoff ≤ 1e-100 between the wild and the domesticated Azuki bean (Altschul, Madd en et al. 1997).

## Syntenic Regions

To better understand the evolutionary history and species divergence,syntenic regions between *Vigna angularis* and *Vigna nakashimae* were identified using the MCscanX software. The output of all-against-all BLAST for the genome wide set of Azuki bean proteins and the formatted annotation filewere used as the input for MCScanX in order to identify the genome wide syntenic regions. The expected number of occurrences of a pair of collinear blocks is estimated by the formula described in (Wang, Wang et al. 2012; Wang, Tang et al. 2012). In total, 1,395 syntenic regions, containing 2,790 gene pairs were classified between *Vigna angularis* and *Vigna nakashimae*. These data can be freely accessed and visualized (Table 5). Moreover, non-synonymous (Ka) and synonymous (Ks) substitution rates of orthologous gene pairs were also calculated and provided (Table 5). Of

these 1,395 syntenic regions, we took only 948 syntenic pairs into consideration for further finding the evolutionary relationship between the wild and the domesticated Azuki bean. 447 syntenic pairs were discarded as the Ks value obtained for these pairs was equal to zero which is not possible (Table 5).

**Evolutionary Relationship**

The Non-synonymous (Ka, non-synonymous amino acid substitution s per non-synonymous site) and Synonymous (Ks, synonymous amino acid s ubstitutions per synonymous site) substitution rates are the two major parame ters for inferring the evolutionary features of a given genome (Albalat, Marfa ny et al. 1994). The Ka and the Ks substitution rates and Ka/Ks ratios are usu ally calculated to determine how the newly formed genes evolved in the cou rse of evolution (Zhang, Li et al. 2006). For this study, we first estimated the Ka, Ks & Ka/Ks for the each syntenic gene pair and then calculated the distr ibution frequency for Ka/Ks (Figure 1). Of the 731 gene pairs which had Ka /Ks value less than one, we obtained two peaks depicting negative selection whereas on contrariety we retrieved one peak depicting positive selection fro m 207 gene pairs for which the Ka/Ks value was more than one (Figure 1). The 1$^{st}$ and the 2$^{nd}$ peak were characterised by a low Ka/Ks value indicating t hat most of the genes in this category have retained their original functions d uring evolution. However, the 3$^{rd}$ peak was characterised by a high Ka/Ks va lue which indicate that genes within these cluster have usually evolved more quickly than others and are more likely to produce new functions during the l ong evolutionary history of red bean.

**Table 4. Ab-initio Gene Prediction result**

| Accession | Total gene count | Complete gene | 5' Partial | 3' Partial | Middle Partial |
|---|---|---|---|---|---|
| IT213134 (Vigna angularis) | 67,835 | 45,985 | 7,971 | 10,377 | 3,502 |
| IT178530 (Vigna nakashimae) | 58,888 | 38,965 | 7,126 | 9,949 | 2,848 |

**Table 5. Syntenic region between *Vigna angularis* and *Vigna nakashimae***

| **Total Syntenic regions between *V.angularis* & *V.nakashime*** | **Ks value equal to zero** | **Ka/Ks value more than one** | **Ka/Ks value less than one** | **Ka/Ks value equal to one** |
|---|---|---|---|---|
| 1,395 | 447 | 207 | 731 | 10 |

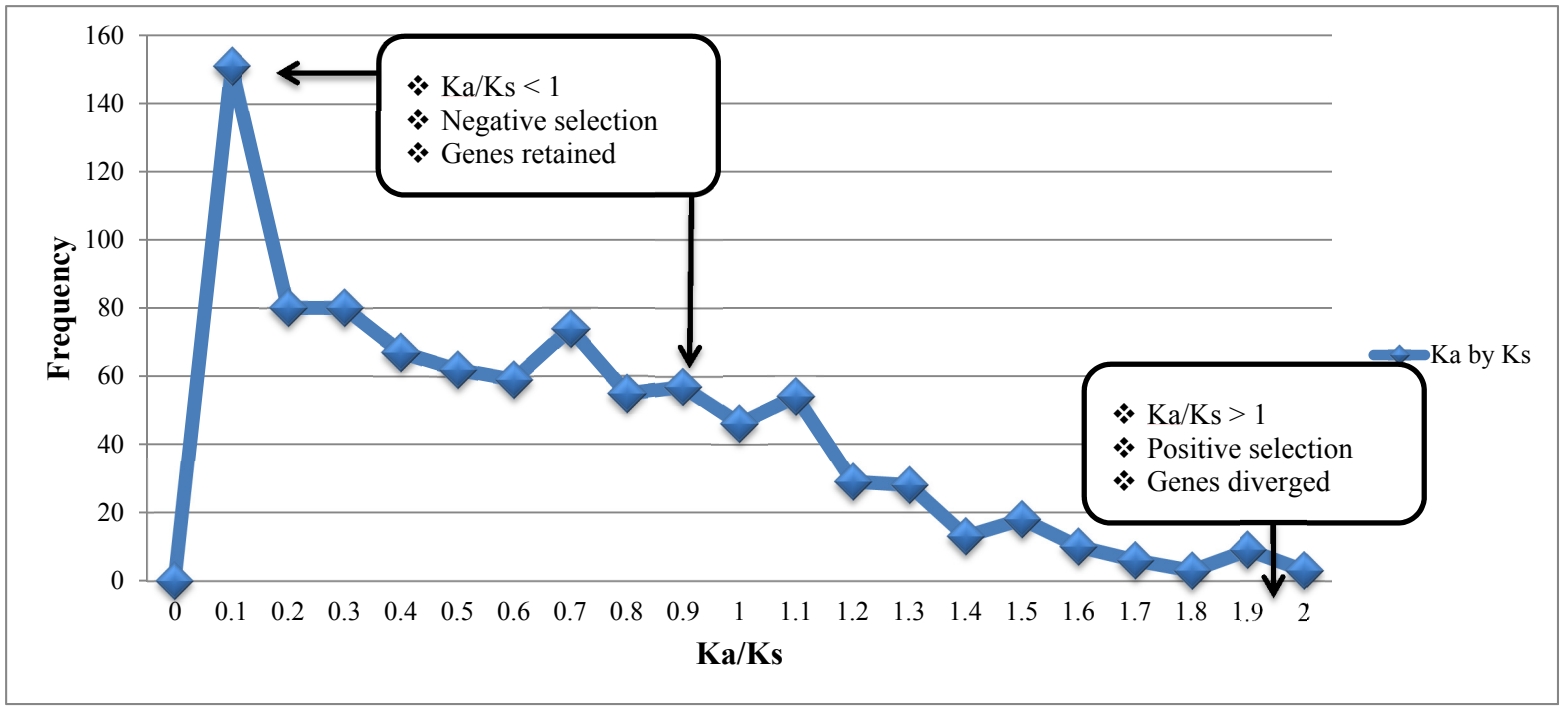


**Figure 1. Distribution of Ka/Ks for the gene pair in order to find the evolutionary changes between *Vigna angularis &Vigna nakashimae*. Ka indicates non-synonymous substitutions per non-synonymous site of a gene pair; Ks indicate synonymous substitutions per synonymous site of a gene pair; Ka/Ks is the ratio of Ka to Ks.**

## Divergence Time

In order to estimate the divergence time between *Vigna angularis* and *Vigna nakashimae,* we used the peak obtained at 0.02 from the distribution o f synonymous substitution rate (Ks) (Haubold and Wiehe 2001) (Figure 2). T he substitution rate $\nu$ was assumed to be 5.1 X $10^{-9}$ synonymous substitutio ns per site every one billion year (Lynch and Conery 2000; Kim, Lee et al. 2 010; Schmutz, Cannon et al. 2010). The divergence time between the wild an d the domesticated Azuki bean was predicted approximately around 1.9 milli on year ago (Mya).

## Structural Translocations and Synteny Pairs

The syntenic regions obtained between the wild and the domesticated Azuki bean are visualised using Circos (Krzywinski, Schein et al. 2009). From these 948 synteny pairs only the particular contigs which had a minimum of at least 9 syntenic relationships between the wild and domesticated Azuki bean were selected. That means we had in total of 158 syntenic relationships to be visualised (Table 6). These 158 syntenic pairs (316 links) between *Vigna angularis* and *Vigna nakashimae*, when visualised revealed synteny relationship as well structural translocations between the two species. Of these 158 syntenic pairs, 79 pairs around 50% depicted structural translocations and another 79 pair around 50% depicted synteny relationship between the *Vigna angularis* and *Vigna nakashimae* (Figure 3).

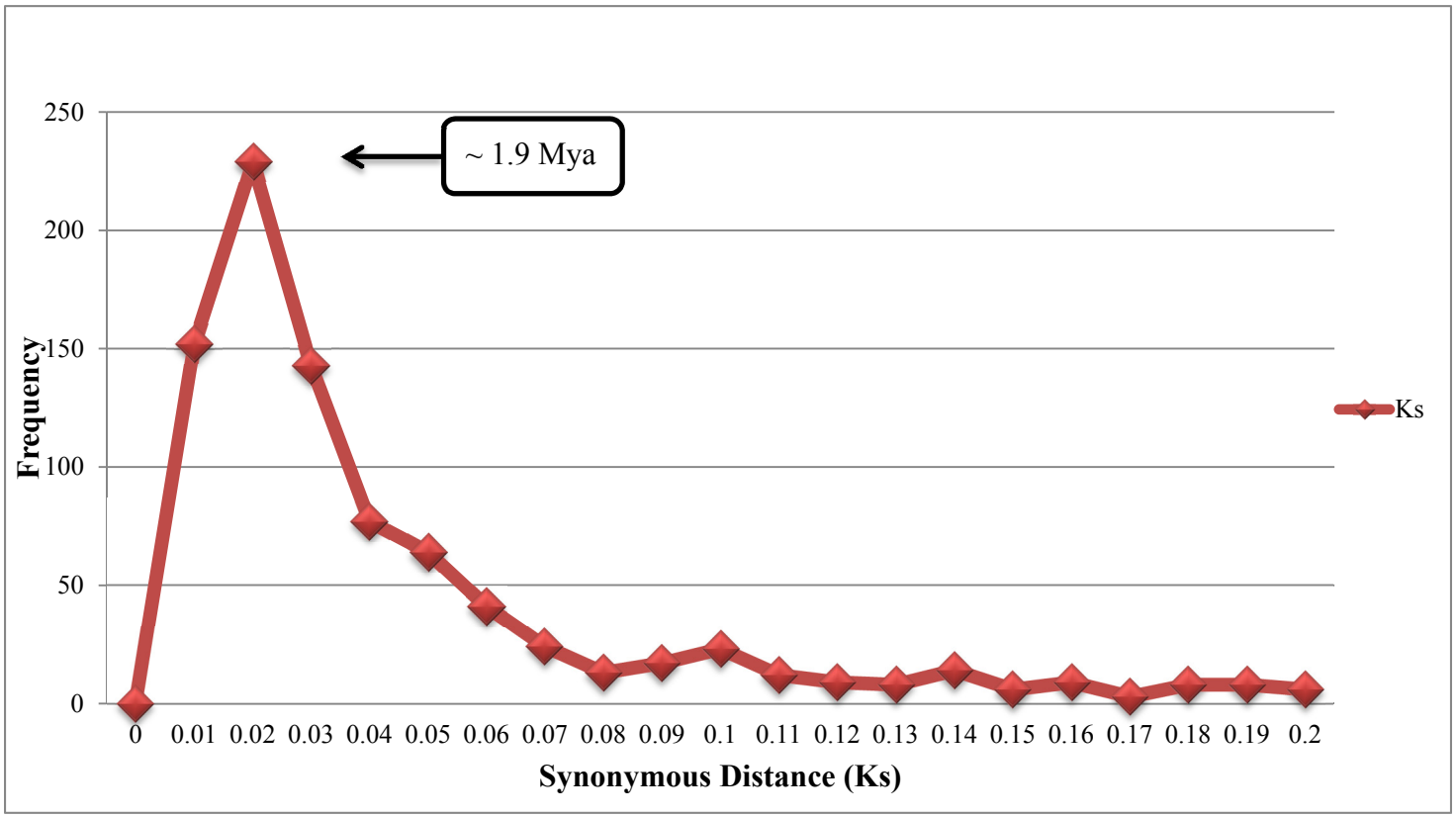


**Figure 2. Frequency distribution of synonymous substitution rate (Ks).**

**Table 6. Structural Translocations and Synteny Pairs between *Vigna angularis* and *Vigna nakashimae***

| Synteny pair | *Vigna angularis* | *Vigna nakashimae* |
|---|---|---|
| 0- 1: | 2257042_8 | 2930711_13 |
| 0- 2: | 2257042_9 | 2930711_12 |
| 0- 3: | 2257042_10 | 2930711_11 |
| 0- 4: | 2257042_11 | 2930711_10 |
| 0- 5: | 2257042_13 | 2930711_8 |
| 0- 6: | 2257042_14 | 2930711_7 |
| 0- 7: | 2257042_15 | 2930711_6 |
| 0- 8: | 2257042_16 | 2930711_5 |
| 0- 9: | 2257042_17 | 2930711_4 |
| 0- 10: | 2257042_18 | 2930711_3 |
| 0- 11: | 2257042_19 | 2930711_2 |
| 28- 1: | 2307231_6 | 2961559_3 |
| 28- 2: | 2307231_7 | 2961559_4 |
| 28- 3: | 2307231_8 | 2961559_5 |
| 28- 4: | 2307231_9 | 2961559_6 |
| 28- 5: | 2307231_10 | 2961559_7 |
| 28- 6: | 2307231_12 | 2961559_9 |
| 28- 7: | 2307231_13 | 2961559_10 |
| 28- 8: | 2307231_14 | 2961559_11 |
| 28- 9: | 2307231_15 | 2961559_12 |
| 28- 10: | 2307231_16 | 2961559_13 |
| 29- 1: | 2307231_3 | 3011189_9 |
| 29- 2: | 2307231_4 | 3011189_10 |
| 29- 3: | 2307231_6 | 3011189_12 |
| 29- 4: | 2307231_7 | 3011189_13 |
| 29- 5: | 2307231_8 | 3011189_14 |
| 29- 6: | 2307231_9 | 3011189_15 |
| 29- 7: | 2307231_10 | 3011189_16 |
| 29- 8: | 2307231_12 | 3011189_18 |
| 29- 9: | 2307231_13 | 3011189_19 |
| 29- 10: | 2307231_14 | 3011189_20 |
| 29- 11: | 2307231_15 | 3011189_21 |

| | | |
|---|---|---|
| 29- 12: | 2307231_16 | 3011189_22 |
| 62- 1: | 2350148_6 | 3035756_50 |
| 62- 2: | 2350148_7 | 3035756_49 |
| 62- 3: | 2350148_9 | 3035756_47 |
| 62- 4: | 2350148_11 | 3035756_45 |
| 62- 5: | 2350148_13 | 3035756_44 |
| 62- 6: | 2350148_14 | 3035756_43 |
| 62- 7: | 2350148_16 | 3035756_42 |
| 62- 8: | 2350148_17 | 3035756_41 |
| 62- 9: | 2350148_18 | 3035756_40 |
| 62- 10: | 2350148_19 | 3035756_39 |
| 62- 11: | 2350148_21 | 3035756_36 |
| 62- 12: | 2350148_22 | 3035756_35 |
| 62- 13: | 2350148_23 | 3035756_34 |
| 75- 1: | 2354480_2 | 2963198_11 |
| 75- 2: | 2354480_3 | 2963198_12 |
| 75- 3: | 2354480_4 | 2963198_13 |
| 75- 4: | 2354480_5 | 2963198_14 |
| 75- 5: | 2354480_6 | 2963198_15 |
| 75- 6: | 2354480_7 | 2963198_16 |
| 75- 7: | 2354480_8 | 2963198_17 |
| 75- 8: | 2354480_9 | 2963198_18 |
| 75- 9: | 2354480_10 | 2963198_19 |
| 77- 1: | 2354480_16 | 3035756_2 |
| 77- 2: | 2354480_18 | 3035756_4 |
| 77- 3: | 2354480_19 | 3035756_5 |
| 77- 4: | 2354480_20 | 3035756_6 |
| 77- 5: | 2354480_22 | 3035756_8 |
| 77- 6: | 2354480_23 | 3035756_9 |
| 77- 7: | 2354480_24 | 3035756_10 |
| 77- 8: | 2354480_25 | 3035756_11 |
| 77- 9: | 2354480_26 | 3035756_12 |
| 77- 10: | 2354480_27 | 3035756_13 |
| 85- 1: | 2355361_41 | 2963198_18 |
| 85- 2: | 2355361_42 | 2963198_17 |
| 85- 3: | 2355361_43 | 2963198_16 |
| 85- 4: | 2355361_44 | 2963198_15 |

| | | |
|---|---|---|
| 85- 5: | 2355361_45 | 2963198_14 |
| 85- 6: | 2355361_46 | 2963198_13 |
| 85- 7: | 2355361_47 | 2963198_12 |
| 85- 8: | 2355361_48 | 2963198_11 |
| 85- 9: | 2355361_49 | 2963198_10 |
| 87- 1: | 2355361_5 | 3035756_31 |
| 87- 2: | 2355361_6 | 3035756_30 |
| 87- 3: | 2355361_7 | 3035756_29 |
| 87- 4: | 2355361_8 | 3035756_28 |
| 87- 5: | 2355361_10 | 3035756_26 |
| 87- 6: | 2355361_11 | 3035756_25 |
| 87- 7: | 2355361_12 | 3035756_24 |
| 87- 8: | 2355361_13 | 3035756_23 |
| 87- 9: | 2355361_14 | 3035756_22 |
| 87- 10: | 2355361_15 | 3035756_21 |
| 87- 11: | 2355361_16 | 3035756_20 |
| 87- 12: | 2355361_17 | 3035756_19 |
| 87- 13: | 2355361_18 | 3035756_18 |
| 87- 14: | 2355361_20 | 3035756_16 |
| 87- 15: | 2355361_21 | 3035756_15 |
| 87- 16: | 2355361_23 | 3035756_13 |
| 87- 17: | 2355361_24 | 3035756_12 |
| 87- 18: | 2355361_25 | 3035756_11 |
| 87- 19: | 2355361_26 | 3035756_10 |
| 87- 20: | 2355361_27 | 3035756_9 |
| 87- 21: | 2355361_28 | 3035756_8 |
| 87- 22: | 2355361_30 | 3035756_6 |
| 87- 23: | 2355361_31 | 3035756_5 |
| 87- 24: | 2355361_32 | 3035756_4 |
| 87- 25: | 2355361_34 | 3035756_2 |
| 87- 26: | 2355361_35 | 3035756_1 |
| 105- 1: | 2360297_2 | 2963699_2 |
| 105- 2: | 2360297_3 | 2963699_3 |
| 105- 3: | 2360297_4 | 2963699_4 |
| 105- 4: | 2360297_5 | 2963699_5 |
| 105- 5: | 2360297_6 | 2963699_6 |
| 105- 6: | 2360297_7 | 2963699_8 |

| | | |
|---|---|---|
| 105- 7: | 2360297_9 | 2963699_9 |
| 105- 8: | 2360297_10 | 2963699_10 |
| 105- 9: | 2360297_12 | 2963699_11 |
| 105- 10: | 2360297_13 | 2963699_12 |
| 105- 11: | 2360297_14 | 2963699_13 |
| 105- 12: | 2360297_15 | 2963699_14 |
| 105- 13: | 2360297_17 | 2963699_15 |
| 105- 14: | 2360297_19 | 2963699_16 |
| 106- 1: | 2360297_2 | 3001113_14 |
| 106- 2: | 2360297_3 | 3001113_13 |
| 106- 3: | 2360297_4 | 3001113_12 |
| 106- 4: | 2360297_5 | 3001113_11 |
| 106- 5: | 2360297_6 | 3001113_10 |
| 106- 6: | 2360297_7 | 3001113_8 |
| 106- 7: | 2360297_9 | 3001113_7 |
| 106- 8: | 2360297_10 | 3001113_6 |
| 106- 9: | 2360297_12 | 3001113_5 |
| 106- 10: | 2360297_13 | 3001113_4 |
| 106- 11: | 2360297_14 | 3001113_3 |
| 106- 12: | 2360297_15 | 3001113_2 |
| 112- 1: | 2363111_6 | 3033783_13 |
| 112- 2: | 2363111_7 | 3033783_12 |
| 112- 3: | 2363111_9 | 3033783_10 |
| 112- 4: | 2363111_10 | 3033783_9 |
| 112- 5: | 2363111_11 | 3033783_8 |
| 112- 6: | 2363111_12 | 3033783_7 |
| 112- 7: | 2363111_13 | 3033783_6 |
| 112- 8: | 2363111_14 | 3033783_5 |
| 112- 9: | 2363111_15 | 3033783_4 |
| 112- 10: | 2363111_17 | 3033783_2 |
| 112- 11: | 2363111_19 | 3033783_1 |
| 127- 1: | 2372033_3 | 2943734_2 |
| 127- 2: | 2372033_4 | 2943734_3 |
| 127- 3: | 2372033_5 | 2943734_4 |
| 127- 4: | 2372033_6 | 2943734_5 |
| 127- 5: | 2372033_8 | 2943734_7 |
| 127- 6: | 2372033_9 | 2943734_8 |

| | | |
|---|---|---|
| 127- 7: | 2372033_10 | 2943734_9 |
| 127- 8: | 2372033_11 | 2943734_10 |
| 127- 9: | 2372033_12 | 2943734_11 |
| 127- 10: | 2372033_13 | 2943734_12 |
| 192- 1: | 2386304_3 | 3028385_3 |
| 192- 2: | 2386304_4 | 3028385_4 |
| 192- 3: | 2386304_6 | 3028385_6 |
| 192- 4: | 2386304_7 | 3028385_7 |
| 192- 5: | 2386304_8 | 3028385_8 |
| 192- 6: | 2386304_9 | 3028385_9 |
| 192- 7: | 2386304_10 | 3028385_10 |
| 192- 8: | 2386304_11 | 3028385_11 |
| 192- 9: | 2386304_13 | 3028385_15 |
| 192- 10: | 2386304_14 | 3028385_16 |
| 192- 11: | 2386304_15 | 3028385_17 |

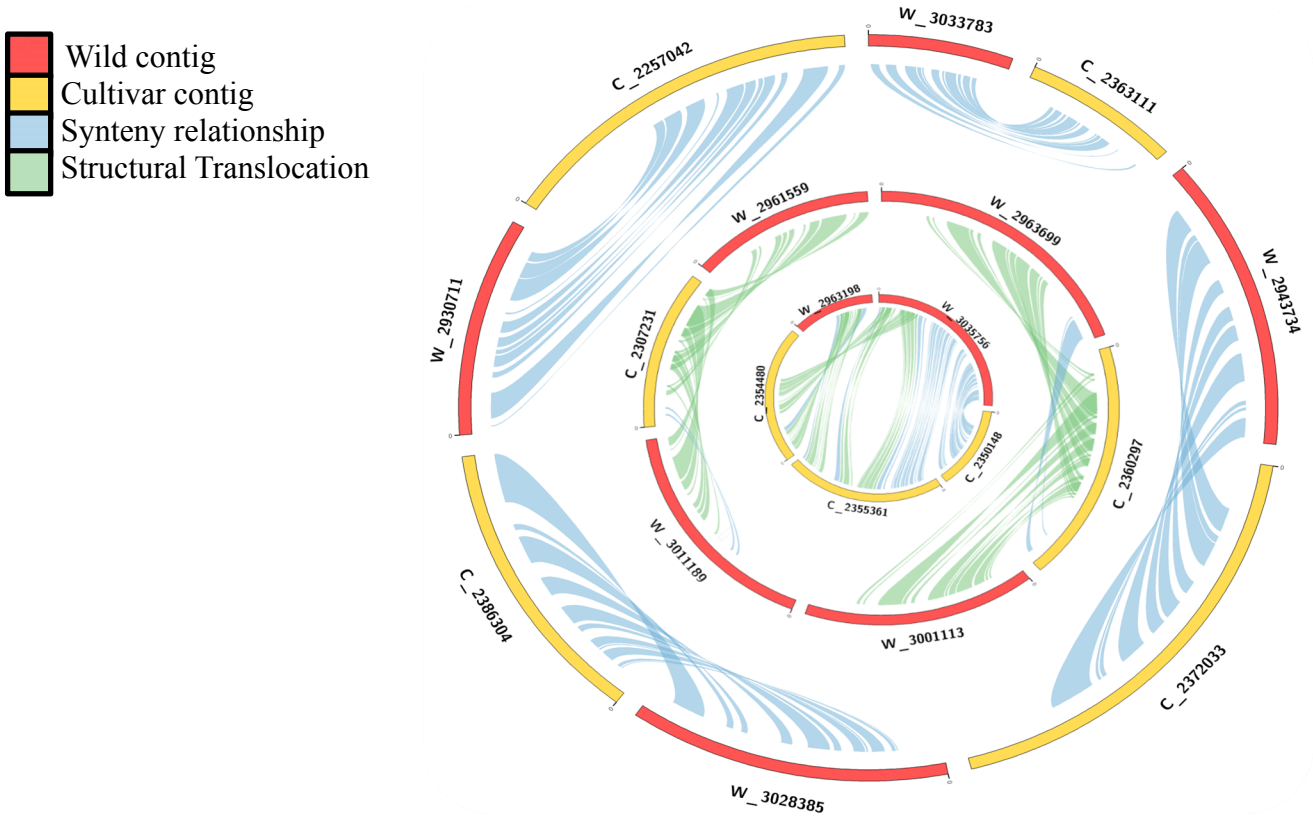


**Figure 3. Synteny pairs and Structural Translocation between *Vigna angularis* and *Vigna nakashimae.***

**Specific Gene Categories**

From the results of the evolutionary relationships between the *Vigna angularis* and *Vigna nakashimae* we tried to retrieve the specific genes which lead to negative and positive selection between the two species (Table 7, 8, 9). The genes set which was retained in the course of evolution (i.e. the genes have undergone silent mutation and thereby have retained their original function) are 109 (Table 7, 8) whereas the gene set which diverged and led to the formation of new functions or a different species during the evolutionary history of red bean were only 20 (Table 9). These 20 genes are the main point to focus, as these genes could be the cause for domestication of Azuki bean.

**Table 7. Peak 1: Specific gene category - Retained**

| BINCODE | NAME |
|---|---|
| '1.1.1.1' | 'PS.lightreaction.photosystem II.LHC-II' |
| '1.1.1.2' | 'PS.lightreaction.photosystem II.PSII polypeptide subunits' |
| '1.2.2' | 'PS.photorespiration.glycolate oxydase' |
| '1.3.13' | 'PS.calvin cycle.rubisco interacting' |
| '2.1.1.1' | 'major CHO metabolism.synthesis.sucrose.SPS' |
| '2.2.1.3.1' | 'major CHO metabolism.degradation.sucrose.invertases.neutral' |
| '4.1.14' | 'glycolysis.cytosolic branch.pyruvate kinase (PK)' |
| '8.1.5' | 'TCA / org transformation.TCA.2-oxoglutarate dehydrogenase' |
| '9.9' | 'mitochondrial electron transport / ATP synthesis.F1-ATPase' |
| '10.1.21' | 'cell wall.precursor synthesis.phosphomannomutase' |
| '10.5.3' | 'cell wall.cell wall proteins.LRR' |
| '10.6.3' | 'cell wall.degradation.pectate lyases and polygalacturonases' |
| '11.1.5' | 'lipid metabolism.FA synthesis and FA elongation.beta hydroxyacyl ACP dehydratase' |
| '11.1.9' | 'lipid metabolism.FA synthesis and FA elongation.long chain fatty acid CoA ligase' |
| '11.1.12' | 'lipid metabolism.FA synthesis and FA elongation.ACP protein' |
| '11.8.6' | 'lipid metabolism.'exotics' (steroids, squalene etc).cycloartenol synthase' |
| '11.9.2.1' | 'lipid metabolism.lipid degradation.lipases.triacylglycerol lipase' |
| '11.9.3.2' | 'lipid metabolism.lipid degradation.lysophospholipases.carboxylesterase' |
| '13.1.3.1.1' | 'amino acid metabolism.synthesis.aspartate family.asparagine.asparagine synthetase' |
| '13.1.4.4.1' | 'amino acid metabolism.synthesis.branched chain group.leucine specific.2-isopropylmalate synthase' |
| '13.1.7.3' | 'amino acid metabolism.synthesis.histidine.phosphoribosyl-AMP cyclohydrolase' |
| '13.99' | 'amino acid metabolism.misc' |
| '15.2' | 'metal handling.binding, chelation and storage' |
| '15.3' | 'metal handling.regulation' |
| '16.1.4' | 'secondary metabolism.isoprenoids.carotenoids' |

| | |
|---|---|
| '16.2.1.4' | 'secondary metabolism.phenylpropanoids.lignin biosynthesis.HCT' |
| '16.7' | 'secondary metabolism.wax' |
| '17.1.3' | 'hormone metabolism.abscisic acid.induced-regulated-responsive-activated' |
| '17.2.3' | 'hormone metabolism.auxin.induced-regulated-responsive-activated' |
| '17.4.1' | 'hormone metabolism.cytokinin.synthesis-degradation' |
| '19.1' | 'tetrapyrrole synthesis.glu-tRNA synthetase' |
| '19.12' | 'tetrapyrrole synthesis.magnesium-protoporphyrin IX monomethyl ester (oxidative) cyclase' |
| '23.4.2' | 'nucleotide metabolism.phosphotransfer and pyrophosphatases.guanylate kinase' |
| '25.5' | 'C1-metabolism.Methylenetetrahydrofolate dehydrogenase & Methenyltetrahydrofolate cyclohydrolase' |
| '26.7' | 'misc.oxidases - copper, flavone etc' |
| '26.11' | 'misc.alcohol dehydrogenases' |
| '26.19' | 'misc.plastocyanin-like' |
| '26.24' | 'misc.GCN5-related N-acetyltransferase' |
| '26.28' | 'misc.GDSL-motif lipase' |
| '27.1' | 'RNA.processing' |
| '27.1.2' | 'RNA.processing.RNA helicase' |
| '27.1.19' | 'RNA.processing.ribonucleases' |
| '27.3' | 'RNA.regulation of transcription' |
| '27.3.6' | 'RNA.regulation of transcription.bHLH,Basic Helix-Loop-Helix family' |
| '27.3.22' | 'RNA.regulation of transcription.HB,Homeobox transcription factor family' |
| '27.3.23' | 'RNA.regulation of transcription.HSF,Heat-shock transcription factor family' |
| '27.3.25' | 'RNA.regulation of transcription.MYB domain transcription factor family' |
| '27.3.35' | 'RNA.regulation of transcription.bZIP transcription factor family' |
| '27.3.40' | 'RNA.regulation of transcription.Aux/IAA family' |
| '27.3.49' | 'RNA.regulation of transcription.GeBP like' |
| '29.1.20' | 'protein.aa activation.phenylalanine-tRNA ligase' |

| | |
|---|---|
| '29.2.1.1.3.2.34' | 'protein.synthesis.ribosomal protein.prokaryotic.unknown organellar.50S subunit.L34' |
| '29.2.1.2.1.24' | 'protein.synthesis.ribosomal protein.eukaryotic.40S subunit.S24' |
| '29.2.2.3.4' | 'protein.synthesis.ribosome biogenesis.Pre-rRNA processing and modifications.WD-repeat proteins' |
| '29.4.1.58' | 'protein.postranslational modification.kinase.receptor like cytoplasmatic kinase VIII' |
| '29.5.11.1' | 'protein.degradation.ubiquitin.ubiquitin' |
| '29.5.11.20' | 'protein.degradation.ubiquitin.proteasom' |
| '29.8' | 'protein.assembly and cofactor ligation' |
| '30.1' | 'signalling.in sugar and nutrient physiology' |
| '30.2.10' | 'signalling.receptor kinases.leucine rich repeat X' |
| '30.11' | 'signalling.light' |
| '31.2' | 'cell.division' |
| '31.2.5' | 'cell.division.plastid' |
| '31.3' | 'cell.cycle' |
| '34.1' | 'transport.p- and v-ATPases' |
| '34.2' | 'transport.sugars' |
| '34.12' | 'transport.metal' |
| '34.19.4' | 'transport.Major Intrinsic Proteins.SIP' |
| '34.21' | 'transport.calcium' |
| '35.1.21' | 'not assigned.no ontology.epsin N-terminal homology (ENTH) domain-containing protein' |
| '35.1.23' | 'not assigned.no ontology.aconitase C-terminal domain-containing protein' |
| '35.1.26' | 'not assigned.no ontology.DC1 domain containing protein' |
| '35.1.41' | 'not assigned.no ontology.hydroxyproline rich proteins' |

## Table 8. Peak 2: Specific gene category - Retained

| BINCODE | NAME |
|---|---|
| '1.2.6' | 'PS.photorespiration.hydroxypyruvate reductase' |
| '6.5' | 'gluconeogenesis / glyoxylate cycle.pyruvate dikinase' |
| '11.8.1' | 'lipid metabolism.'exotics' (steroids, squalene etc).sphingolipids' |
| '11.9.3.5' | 'lipid metabolism.lipid degradation.lysophospholipases.phosphoinositide phospholipase C' |
| '18.2.2' | 'Co-factor and vitamine metabolism.thiamine.hydroxymethylpyrimidine kinase' |
| '19' | 'tetrapyrrole synthesis' |
| '20.2' | 'stress.abiotic' |
| '21.2.2' | 'redox.ascorbate and glutathione.glutathione' |
| '26.2' | 'misc.UDP glucosyl and glucoronyl transferases' |
| '26.4.1' | 'misc.beta 1,3 glucan hydrolases.glucan endo-1,3-beta-glucosidase' |
| '26.10' | 'misc.cytochrome P450' |
| '26.21' | 'misc.protease inhibitor/seed storage/lipid transfer protein (LTP) family protein' |
| '27.1.1' | 'RNA.processing.splicing' |
| '27.3.4' | 'RNA.regulation of transcription.ARF, Auxin Response Factor family' |
| '27.3.5' | 'RNA.regulation of transcription.ARR' |
| '27.3.50' | 'RNA.regulation of transcription.General Transcription' |
| '27.3.55' | 'RNA.regulation of transcription.HDA' |
| '27.4' | 'RNA.RNA binding' |
| '28.1' | 'DNA.synthesis/chromatin structure' |
| '29.2.1.1.2.1.29' | 'protein.synthesis.ribosomal protein.prokaryotic.mitochondrion.30S subunit.S29' |
| '29.2.1.2.1.16' | 'protein.synthesis.ribosomal protein.eukaryotic.40S subunit.S16' |
| '29.2.1.2.1.29' | 'protein.synthesis.ribosomal protein.eukaryotic.40S subunit.S29' |
| '29.2.1.2.2.34' | 'protein.synthesis.ribosomal protein.eukaryotic.60S subunit.L34' |
| '29.2.2' | 'protein.synthesis.ribosome biogenesis' |
| '29.2.3.1' | 'protein.synthesis.initiation.deoxyhypusine synthase' |
| '29.2.6' | 'protein.synthesis.ribosomal RNA' |
| '29.2.99' | 'protein.synthesis.misc' |
| '29.4.1.59' | 'protein.postranslational modification.kinase.receptor like cytoplasmatic kinase IX' |

| | |
|---|---|
| '29.5.1' | 'protein.degradation.subtilases' |
| '29.5.9' | 'protein.degradation.AAA type' |
| '30.2.1' | 'signalling.receptor kinases.leucine rich repeat I' |
| '30.2.3' | 'signalling.receptor kinases.leucine rich repeat III' |
| '34.14' | 'transport.unspecified cations' |
| '34.15' | 'transport.potassium' |
| '34.18' | 'transport.unspecified anions' |
| '35.1.12' | 'not assigned.no ontology.pumilio/Puf RNA-binding domain-containing protein' |

## Table 9. Peak 3: Specific gene category - Diverged

| BINCODE | NAME |
|---|---|
| '1.2.7' | 'PS.photorespiration.glycerate kinase' |
| '2.2.1.3.2' | 'major CHO metabolism.degradation.sucrose.invertases.cell wall' |
| '11.9.3.4' | 'lipid metabolism.lipid degradation.lysophospholipases.phospholipase A2' |
| '16.2.1.3' | 'secondary metabolism.phenylpropanoids.lignin biosynthesis.4CL' |
| '16.2.1.9' | 'secondary metabolism.phenylpropanoids.lignin biosynthesis.COMT' |
| '17.2.2' | 'hormone metabolism.auxin.signal transduction' |
| '21.4' | 'redox.glutaredoxins' |
| '26.12' | 'misc.peroxidases' |
| '26.13' | 'misc.acid and other phosphatases' |
| '27.3.14' | 'RNA.regulation of transcription.CCAAT box binding factor family, HAP2' |
| '27.3.39' | 'RNA.regulation of transcription.AtSR Transcription Factor family' |
| '27.3.66' | 'RNA.regulation of transcription.Psudo ARR transcription factor family' |
| '27.3.68' | 'RNA.regulation of transcription.PWWP domain protein' |
| '27.3.71' | 'RNA.regulation of transcription.SNF7' |
| '29.3.2' | 'protein.targeting.mitochondria' |
| '29.5.11.4.1' | 'protein.degradation.ubiquitin.E3.HECT' |
| '30.2.9' | 'signalling.receptor kinases.leucine rich repeat IX' |
| '30.6' | 'signalling.MAP kinases' |
| '34.6' | 'transport.sulphate' |
| '34.99' | 'transport.misc' |

# DISSCUSSION

We used a novel method to predict the structural variations between the wild and domesticated Azuki bean. In this method we *de novo* sequenced and assembled the genomes of *Vigna angularis* and *Vigna nakashimae*. The assembled genomes were then compared to identify the structural variations. According to (Feuk, Carson et al. 2006; Raphael 2012) this method is best for finding all types of variants in the genomes and doesn't require a reference genome also . This is applicable only if the genomes are assembled properly. Any mis-assembly or the errors arising due to repetitive sequences can lead to false prediction of the variants. The biggest advantage of using this method is that it helps in overcoming the limitations of the other two methods used to predict the structural variations. The re-sequencing method is good for finding the single nucleotide polymorphism and the insertion deletions occurring in the genome but not for finding large scale structural variations. This method produces error when there are repetitive sequences near break point or if the genome has multiple stated and complex architectures, and if recurrent variants exits at same locus (Kim, Lee et al. 2010; Raphael 2012). The second method is also very effective as it uses the unmapped reads to find the structural variations. But the major problem using this method is that it can't be used if the reference genome is not available for the particular individual.

The *de novo* assembly generated approximately 0.68 million and 1.2 million reads for *Vigna angularis* and *Vigna nakashimae*. The N50 value obtained is within the equivalent range but we found a big difference between the total number of contigs and the numbers of contigs longer than

N50 value (Table 3). The numbers of contigs longer than N50 value obtained were less may be due to short read length, quality of reads and heterozygosity of the accessions. We estimated the genome size of Azuki bean to be approximately around 545 Mb (Table 2). This is in accordance with the genome size of the other members of the Vigna species like Mungbean and black gram. The genome size of *Vigna radiata* and *Vigna mungo* is estimated to be 579 Mb (Somta and Srinives 2007).

The total number of genes predicted in *Vigna angularis* and *Vigna nakashimae* is 67,835 and 58,888 respectively (Table 4) of these approximately around 33% were partial genes. That is 67% of the gene predicted were complete. We predicted 45,985 protein-coding genes in *Vigna angularis* which is 70% more than that of Arabidopsis and approximately similar to poplar and soybean (Arabidopsis 2000; Tuskan, Difazio et al. 2006; Schmutz, Cannon et al. 2010). In case of *Vigna nakashimae* we predicted 38,965 protein-coding genes which is 40% more than Arabidopsis and approximately similar to potato (Arabidopsis 2000; Xu, Pan et al. 2011).

While identifying the synteny relationship between the wild and domesticated Azuki bean, 447 synteny pairs had Ks value of zero (Table 5). This means the non-synonymous changes are more while synonymous changes are less and that is not possible. The reason for behind this abnormal result is that the genes predicted were based on ab-inito method not on transcriptome basis so may be the genes predicted were pseudo genes.

The Ka, Ks and Ka/Ks analysis in this study led to the inference of how the *Vigna angularis* has evolved from *Vigna nakashimae*. In 1$^{st}$ and 2$^{nd}$ peak (Figure 1) the Ka/Ks < 1 that means the alleles of *Vigna angularis* and *Vigna nakashimae* locus share very similar genes. This indicates that the

genes present were very essential for the normal growth and development of the plant and hence even the selection pressure couldn't change them. While in the 3rd peak the alleles had different genes. This indicates that the genes have evolved to produce new function and this led the *Vigna angularis* and *Vigna nakashimae* to diverge from each other approximately around 1.9 Mya (Figure 2). This selection pressure led to speciation which generally occurs due to geographical isolation or reduced gene flow etc. We can also say this divergence led to the domestication of *Vigna angularis*. Domestication is a complex evolutionary process involving human interactions which leads to morphological and physiological changes in plant and animal species, that distinguishes the domesticated taxa from their wild ancestors (Hancock 2005; Purugganan and Fuller 2009). Evolutionary biologists, however, tend to view domestication as a special class of species diversification, distinct from species divergence through natural selection in the wild (Purugganan and Fuller 2009).

The genes categories which led the *Vigna angularis* to diverge from *Vigna nakashimae* were retrieved (Table 9). In accordance with previous studies (Fang, Wang et al. 2012; Chakrabarti, Zhang et al. 2013) these gene categories can be a major factor which led to the domestication of *Vigna angularis.* According to (Fang, Wang et al. 2012) the EOD3 encodes the Arabidopsis cytochrome P450/CYP78A6 and is expressed in most plant organs. Overexpression of this EOD3 dramatically increases the seed size of wild-type plants, whereas eod3-ko loss-of-function mutants form small seeds. In the second case study by (Chakrabarti, Zhang et al. 2013); SlKLUH a P450 enzyme belonging to CYP78A subfamily gene controls fruit mass by increased cell layers and delayed fruit ripening. It also modulates plant architecture by regulating number and length of the side shoots. A SNP

identified in the promoter of SlKLUH has led to increased fruit size which led to the domestication of tomato.

Structural translocations identified (Figure 3) between the wild and domesticated Azuki bean can also be a major factor for the two species to diverge from each other. Numerous studies have been conducted on maize (Iltis 2000; Schnable, Ware et al. 2009), soybean (Schmutz, Cannon et al. 2010), cotton (Desai, Chee et al. 2006; Wang, Wang et al. 2012) etc., which shown that structural translocations can be a factor leading to domestication of the species.

In summary, using *de novo* method, we sequenced and assembled the *Vigna angularis* and *Vigna nakashimae* genomes. The structural translocations detected and the gene categories obtained in this study can be an underlying cause for the domestication of *Vigna angularis* or may account for genetic changes in the genome. Our results suggest that the *Vigna angularis* and *Vigna nakashimae* diverged from each other approximately around 1.9 Mya. Regardless of the genome-wide comparison done for the wild and domesticated forms, it is necessary to sequence more of Azuki bean genome and validate the structural translocations obtained. This approach could be widely used to introduce diversity in the variety without altering the plant performance or its product quality or it could help in understanding the underlying processes of domestication.

# REFERENCES


Albalat, R., G. Marfany, et al. (1994). "Analysis of nucleotide substitutions and amino acid conservation in theDrosophila Adh genomic region." Genetica **94**(1): 27-36.

Altschul, S. F., W. Gish, et al. (1990). "Basic local alignment search tool." Journal of Molecular Biology **215**(3): 403-410.

Altschul, S. F., T. L. Madden, et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res **25**(17): 3389-3402.

Arabidopsis, G. I. (2000). "Analysis of the genome sequence of the flowering plant Arabidopsis thaliana." Nature **408**(6814): 796.

Cannon, S. B., G. D. May, et al. (2009). "Three sequenced legume genomes and many crop species: rich opportunities for translational genomics." Plant Physiol **151**(3): 970-977.

Chakrabarti, M., N. Zhang, et al. (2013). "A cytochrome P450 regulates a domestication trait in cultivated tomato." Proceedings of the National Academy of Sciences **110**(42): 17125-17130.

Chekanov, N., E. Boulygina, et al. (2010). "Individual Genome of the Russian Male: SNP Calling and a de novo Assembly of Unmapped Reads." Acta naturae **2**(3): 122.

Desai, A., P. W. Chee, et al. (2006). "Chromosome structural changes in diploid and tetraploid A genomes of Gossypium." Genome **49**(4): 336-345.

Elshire, R. J., J. C. Glaubitz, et al. (2011). "A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species." PLoS One **6**(5): e19379.

Fang, W., Z. Wang, et al. (2012). "Maternal control of seed size by EOD3/CYP78A6 in Arabidopsis thaliana." The Plant Journal **70**(6): 929-939.

Feuk, L., A. R. Carson, et al. (2006). "Structural variation in the human genome." Nature Reviews Genetics **7**(2): 85-97.

Gillespie, J. H. (1991). The causes of molecular evolution, Oxford University Press.

Gu, W., F. Zhang, et al. (2008). "Mechanisms for human genomic rearrangements." Pathogenetics **1**(1): 4.

Guigó, R., S. Knudsen, et al. (1992). "Prediction of gene structure." Journal of Molecular Biology **226**(1): 141-157.

Hancock, J. F. (2005). "Contributions of domesticated plant studies to our understanding of plant evolution." Ann Bot **96**(6): 953-963.

Hastings, P., J. R. Lupski, et al. (2009). "Mechanisms of change in gene copy number." Nature Reviews Genetics **10**(8): 551-564.

Haubold, B. and T. Wiehe (2001). "Statistics of divergence times." Mol Biol Evol **18**(7): 1157-1160.

Hiremath, P. J., A. Farmer, et al. (2011). "Large-scale transcriptome analysis in chickpea (Cicer arietinum L.), an orphan legume crop of the semi-arid tropics of Asia and Africa." Plant biotechnology journal **9**(8): 922-931.

Iltis, H. H. (2000). "Homeotic sexual translocations and the origin of maize (Zea mays, Poaceae): A new look at an old problem." Economic Botany **54**(1): 7-42.

IMURA, K. M., 1983 The Neutral Theory of Molecular Evolution, Cambridge University Press, Cambridge, UK.

Jaillon, O., J.-M. Aury, et al. (2007). "The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla." Nature **449**(7161): 463-467.

Kafatos, F. C., A. Efstratiadis, et al. (1977). "Molecular evolution of human and rabbit beta-globin mRNAs." Proceedings of the National Academy of Sciences **74**(12): 5618-5622.

Kaga, A., T. Isemura, et al. (2008). "The genetics of domestication of the azuki bean (Vigna angularis)." Genetics **178**(2): 1013-1036.

Kim, M. Y., S. Lee, et al. (2010). "Whole-genome sequencing and intensive analysis of the undomesticated soybean (Glycine soja Sieb. and Zucc.) genome." Proceedings of the National Academy of Sciences **107**(51): 22032-22037.

Kimura, M. (1977). "Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution."

Krzywinski, M., J. Schein, et al. (2009). "Circos: an information aesthetic for comparative genomics." Genome Res **19**(9): 1639-1645.

Kuczynski, J., E. K. Costello, et al. (2010). "Direct sequencing of the human microbiome readily reveals community differences." Genome Biol **11**(5): 210.

Lee, G.-A. (2013). "Archaeological perspectives on the origins of azuki (Vigna angularis)." The Holocene **23**(3): 453-459.

Li, R., W. Fan, et al. (2009). "The sequence and de novo assembly of the giant panda genome." Nature **463**(7279): 311-317.

Lynch, M. and J. S. Conery (2000). "The evolutionary fate and consequences of duplicate genes." Science **290**(5494): 1151-1155.

Matsumoto, T., J. Wu, et al. (2005). "The map-based sequence of the rice genome." Nature **436**(7052): 793-800.

Menges, F., G. Narzisi, et al. (2011). "TotalReCaller: improved accuracy and performance via integrated alignment and base-calling." Bioinformatics **27**(17): 2330-2337.

Miyata, T. and T. Yasunaga (1980). "Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application." Journal of Molecular Evolution **16**(1): 23-36.

Miyata, T., T. Yasunaga, et al. (1980). "Nucleotide sequence divergence and functional constraint in mRNA evolution." Proceedings of the National Academy of Sciences **77**(12): 7328-7332.

Myburg, A., D. Grattapaglia, et al. (2011). The Eucalyptus grandis Genome Project: Genome and transcriptome resources for comparative analysis of woody plant biology. BMC Proceedings, BioMed Central Ltd.

Narzisi, G. and B. Mishra (2011). "Scoring-and-unfolding trimmed tree assembler: concepts, constructs and comparisons." Bioinformatics **27**(2): 153-160.

Nei, M. and T. Gojobori (1986). "Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions." Mol Biol Evol **3**(5): 418-426.

Nekrutenko, A., K. D. Makova, et al. (2002). "The K(A)/K(S) ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study." Genome Res **12**(1): 198-202.

Pal, C., B. Papp, et al. (2006). "An integrated view of protein evolution." Nat Rev Genet **7**(5): 337-348.

Paterson, A. H., J. E. Bowers, et al. (2009). "The Sorghum bicolor genome and the diversification of grasses." Nature **457**(7229): 551-556.

Perler, F., A. Efstratiadis, et al. (1980). "The evolution of genes: the chicken preproinsulin gene." Cell **20**(2): 555-566.

Pevzner, P. A., H. Tang, et al. (2001). "An Eulerian path approach to DNA fragment assembly." Proceedings of the National Academy of Sciences **98**(17): 9748-9753.

Purugganan, M. D. and D. Q. Fuller (2009). "The nature of selection during plant domestication." Nature **457**(7231): 843-848.

Raphael, B. J. (2012). "Structural Variation and Medical Genomics." PLoS computational biology **8**(12): e1002821.

Reif, J. C., P. Zhang, et al. (2005). "Wheat genetic diversity trends during domestication and breeding." Theoretical and Applied Genetics **110**(5): 859-864.

Schmutz, J., S. B. Cannon, et al. (2010). "Genome sequence of the palaeopolyploid soybean." Nature **463**(7278): 178-183.

Schnable, P. S., D. Ware, et al. (2009). "The B73 maize genome: complexity, diversity, and dynamics." Science **326**(5956): 1112-1115.

Simpson, J. T., K. Wong, et al. (2009). "ABySS: a parallel assembler for short read sequence data." Genome Res **19**(6): 1117-1123.

Siu, H., Y. Zhu, et al. (2011). "Implication of next-generation sequencing on association studies." BMC Genomics **12**(1): 322.

Somta, P., A. Kaga, et al. (2006). "Development of an interspecific Vigna linkage map between Vigna umbellata (Thunb.) Ohwi & Ohashi and V. nakashimae (Ohwi) Ohwi & Ohashi and its use in analysis of bruchid resistance and comparative genomics." Plant Breeding **125**(1): 77-84.

Somta, P. and P. Srinives (2007). "Genome research in mungbean (Vigna radiata (L.) Wilczek) and blackgram (V. mungo (L.) Hepper)." ScienceAsia **33**(Suppl 1): 69-74.

Tang, H., J. E. Bowers, et al. (2008). "Synteny and collinearity in plant genomes." Science **320**(5875): 486-488.

Tang, H., X. Wang, et al. (2008). "Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps." Genome Res **18**(12): 1944-1954.

Thimm, O., O. Bläsing, et al. (2004). "mapman: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes." The Plant Journal **37**(6): 914-939.

Thudi, M., Y. Li, et al. (2012). "Current state-of-art of sequencing technologies for plant genomics research." Briefings in functional genomics **11**(1): 3-11.

Tuskan, G. A., S. Difazio, et al. (2006). "The genome of black cottonwood, Populus trichocarpa (Torr. & Gray)." Science **313**(5793): 1596-1604.

Urbanczyk-Wochniak, E., B. Usadel, et al. (2006). "Conversion of MapMan to allow the analysis of transcript data from Solanaceous species: effects of genetic and environmental alterations in energy metabolism in the leaf." Plant molecular biology **60**(5): 773-792.

Usadel, B., A. Nagel, et al. (2005). "Extension of the visualization tool MapMan to allow statistical analysis of arrays, display of corresponding genes, and comparison with known responses." Plant Physiol **138**(3): 1195-1204.

Usadel, B., F. Poree, et al. (2009). "A guide to using MapMan to visualize and compare Omics data in plants: a case study in the crop species, Maize." Plant, cell & environment **32**(9): 1211-1229.

Varshney, R., P. Hiremath, et al. (2009). "A comprehensive resource of drought-and salinity-responsive ESTs for gene discovery and marker development in chickpea (Cicer arietinum L.)." BMC Genomics **10**(1): 523.

Vezzi, F., G. Narzisi, et al. (2012). "Feature-by-feature–evaluating de novo sequence assembly." PLoS One **7**(2): e31002.

Wall, P. K., J. Leebens-Mack, et al. (2009). "Comparison of next generation sequencing technologies for transcriptome characterization." BMC Genomics **10**(1): 347.

Wang, J., J. Wang, et al. (2012). "Isolation and partial characterization of an R2R3MYB transcription factor from the bamboo species Fargesia fungosa." Plant Molecular Biology Reporter **30**(1): 131-138.

Wang, K., Z. Wang, et al. (2012). "The draft genome of a diploid cotton Gossypium raimondii." Nat Genet.

Wang, Y., H. Tang, et al. (2012). "MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity." Nucleic Acids Res **40**(7): e49.

Weischenfeldt, J., O. Symmons, et al. (2013). "Phenotypic impact of genomic structural variation: insights from and for human disease." Nature Reviews Genetics **14**(2): 125-138.

WenHsiung, L. (1997). Molecular evolution, Sinauer Associates Incorporated.

Wu, G., N. Yi, et al. (2011). "Statistical quantification of methylation levels by next-generation sequencing." PLoS One **6**(6): e21034.

Xu, X., S. Pan, et al. (2011). "Genome sequence and analysis of the tuber crop potato." Nature **475**(7355): 189-195.

Yamaguchi, H. (1992). "Wild and weed azuki beans in Japan." Economic Botany **46**(4): 384-394.

Yang, L., L. J. Luquette, et al. (2013). "Diverse Mechanisms of Somatic Structural Variations in Human Cancer Genomes." Cell **153**(4): 919-929.

Yoon, M., J. Lee, et al. (2007). "Genetic relationships among cultivated and wild Vigna angularis (Willd.) Ohwi et Ohashi and relatives from Korea based on AFLP markers." Genetic Resources and Crop Evolution **54**(4): 875-883.

Zhang, Z., J. Li, et al. (2006). "KaKs_Calculator: calculating Ka and Ks through model selection and model averaging." Genomics, Proteomics & Bioinformatics **4**(4): 259-263.

Zhang, Z. and J. Yu (2006). "Evaluation of six methods for estimating synonymous and nonsynonymous substitution rates." Genomics, Proteomics & Bioinformatics **4**(3): 173-181.

## 초록

팥은 세계적으로 중요한 콩과 작물이며 야생팥과 재배팥을 비교하는 비교 유전체학적 관점으로 접근하면 팥의 가사화 과정을 이해하여 더욱 효율적인 육종을 할 수 있을 것이다. 한편 차세대 염기서열 분석방법이 등장함에 따라 낮은 가격으로 방대한 양의 염기서열 자료를 생산할 수 있게 되었지만 이를 통해 생산된 염기서열 도막들은 너무 짧아 *de novo* assembly를 하기 힘들다는 단점이 있다. 이에 우리는 야생종과 재배종 사이의 유전체 변이를 찾을 수 있는 새로운 방법을 소개하고자 한다. 먼저, 재배팥(*Vigna angularis*)과 야생팥(*Vigna nakashimae*)의 염기서열을 밝혀내고, 이를 assemble하였다. 그 결과 재배팥은 11 Kb, 야생팥은 7 Kb의 N50 크기를 갖는 contig를 생산할 수 있었고, 팥의 유전체 크기를 525 Mb로 예측할 수 있었다. 이 contig를 이용하여 유전자를 예측한 결과, 재배팥에서는 단백질을 만들어내는 유전자 45,985개를 예측해낼 수 있었는데, 이는 애기장대보다 70% 더 많은 수치이며 콩과 포플라와 비슷한 수치이다. 야생팥에서는 38,965개의 단백질 생산 유전자를 예측하였으며 이는 애기장대보다 40% 많은 수치이며 감자와 비슷한 수치이다. 야생팥과

재배팥의 비교를 통해서 얻은 전좌된 염기서열과 유전자 자료를 얻을 수 있었는데, 이들이 팥의 가사화 과정에 기여했을 것이라 추측할 수 있었다. 우리가 제시한 새로운 방법은 효율적인 비용을 통해 잘 연구되지 않은 작물의 유전체를 연구하는 데에 큰 도움이 될 것으로 생각되며, re-sequencing 방법으로는 어려움이 있었던 종간 유전체 비교 분석을 가능하게 할 수 있을 것이다.

주요어: 재배팥(*Vigna angularis*), 야생팥(*Vigna nakashimae*), 전좌, 가사화, 유전체 변이, 차세대 염기서열 분석

# *ACKNOWLEDGEMENT*


*"The only source of knowledge is experience"*

*- Albert Einstein*

*"Trust in the world with all your heart and lean not on your own understanding; in all your ways acknowledge Him and He will make your paths straight".*

*- Proverbs 3: 5-6.*

*My deepest gratitude goes to Almighty God who has blessed me abundantly to complete my research work with success. I give him all glory and honour for all that he had done for me. With great reverence and pleasure, I obediently consider myself highly privileged to have* ***Dr. Suk Ha Lee****, Professor, Department of Plant Science, College of Agriculture and Life Science, SNU as chairman of the advisory committee. With a deep sense of lifelong indebtedness, I express my heartfelt gratitude to my professor for valuable guidance, care, constant encouragement and help throughout the course of this study and in bringing out this thesis.*

*I extend my thanks to members of the advisory committee,* ***Dr. Seo Hak Soo****, Associate Professor, and* ***Dr. Paek Nam Chon,*** *Professor, Department of Plant Science, College of Agriculture and Life Science, SNU for their valuable suggestions throughout the course of my study.*

*I hold in high regard the efforts of* ***all my teachers*** *for enriching my overall knowledge and help rendered throughout the course of study.*

*With much pleasure and respect, I extend my countless gratitude to my Lab members* ***Dr. Lee, Dr. Kim, Dr. Van, Dr. Puji, Hyunju onni, Young onni, Sue onni, Min Young onni, Ahra, Myo Yeon onni, Su Yeon, WonJoo oppa, Kwang Soo oppa, Yang Jae oppa, Sang Nae oppa, Jay oppa, Tae Young, Puntaree, Kularb*** *for their wholehearted help, critical suggestions, cordial team work and care rendered throughout research programme.*

*My heart is joyous to express the feelings with thanks to **all my friends** for their boundless affection, deep concern, spontaneous help and sustained support rendered to me during my research work.*

*On a personal note I wish to place my thanks to the giant pillars of my life, my beloved **father** and affectionate **mother** for their motivations and prayers to keep me in high spirits to pursue the programme successfully. I also express my deep sense of gratitude to my beloved **brother** for their inordinate help and constant encouragement throughout my study period and I dedicate this humble piece of work to them.*

*Above all, I bow my head before **God Almighty** who gave me everything to pursue this endeavor to completion.*

*With colourful memories, I dedicate this sculpture to the hearts that have really taken pains to bring me to the present position.*

***(Khushboo Rastogi)***