# EVOLUTION OF FLAVONOID PATHWAY IN LEGUMES

by

Sajedeh Safari

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Bioinformatics and Computational Biology

Charlotte

2016

Approved by:

_____
Dr. Jessica Schlueter

_____
Dr. Jennifer Weller

_____
Dr. Dennis Livesay

_____
Dr. Shannon Schlueter

_____
Dr. Bao-Hua Song

# ABSTRACT

SAJEDEH SAFARI. Evolution of flavonoid pathway in legumes. (Under the direction of Dr. JESSICA SCHLUETER)

Considering the ever growing world population as well as climate change, it has become essential to improve crops in order to reach higher yields, more tolerance to pests and droughts, and better adaptation to different environments. The advancement of next-generation sequencing technology enables us to perform comparative genome analyses for related genomes, specifically soybean, other *Glycine* species, and other related legumes. Although plants and their chemistry have been under study for a very long time, the understanding of the value of their secondary metabolites and their practical application is rather a newer process. One of the most studied pathways in legumes is the flavonoid biosynthetic pathway. Flavonoids play many important roles in plants and their interactions with their environment. Despite all the previous studies on the flavonoid biosynthetic pathway, our knowledge of the connection between the genotype and phenotype of the pathway, and the gene families elaborating the different characteristics of different species, particularly the gene families determining the flower colors and seed coat colors of different species, is still very inadequate. As part of this study I have identified candidate gene families and their putative orthologs for twelve enzymes from the flavonoid pathway among eight legumes and *Arabidopsis thaliana*. The study provided data for looking for candidate genes for flower colors and seed coat colors in sixteen soybean varieties. Many of these pigments can be used as markers to study different biological and evolutionary processes within the legumes. Isoflavone synthase, one of the enzymes of this pathway, was targeted for re-sequencing of ~150,000 bp

regions in seven perennial legumes, where I have annotated these targeted regions, identified orthologous genes and performed evolutionary analyses to understand the genomic dynamics between these species.

# DEDICATION

To my beloved family and all the teachers who introduced me to science.

# ACKNOWLEDGEMENTS

My first and greatest appreciation goes to Dr. Jessica Schlueter for all her invaluable guidance and support during my PhD. Without her knowledge and support my transition into Plant genomics would not be possible. I would also like to thank Dr. Jennifer Weller as well, without her guidance and endless support I would not be able to find my passion in this field and start my PhD. I also greatly appreciate all of the past and present members of the Schlueter group.

I would like to thank the Soymap II project that provided our sequencing data for chapter four. I would also like to acknowledge Dr. Jim Specht and Dr. Randy Shoemaker for their roles in the second chapter of this thesis.

I sincerely thank my dissertation committee members, Dr. Dennis Livesay, Dr. Jennifer Weller, Dr. Shannon Schlueter, and Dr. Bao-Hua Song, for spending their precious time to guide me through this journey.

Finally, I'm forever grateful for the opportunity provided by the Department of Bioinformatics and Genomics, and the Graduate Assistant Support Plan (GASP) for the financial support over the past few years, to allow me to pursue my Ph.D. My appreciation also goes to all of the staff members in the department, and the international students and scholars office for their assistance.

TABLE OF CONTENTS

# CHAPTER 1: INTRODUCTION

The legumes (Fabaceae) include more than 650 genera and 18,000 species, making them the third largest family of flowering plants (Polhill and Raven 1981). They constitute the second most important family of crop plants after Poaceae (grass family) and make up 27% of the world's crop production (Graham and Vance 2003). Legumes, especially soybean (*Glycine max*), are used as animal food and provide almost one-third of human protein intake in addition to being used as edible and industrial oils. They have a unique ability for symbiotic nitrogen fixation (Zhu et al. 2005). Considering the ever growing world population as well as climate change, it has become essential to improve crops in order to reach higher yields, more tolerance to pests and droughts, and better adaptation to different environments. Numerous genetic bottlenecks, including domestication, import to the United States and selective breeding, has caused soybean to lose rare sequence variances as well as greatly reduce its allele frequency (Hyten et al. 2006). In 2010, Schmutz et al. used whole-genome shotgun sequencing to sequence the soybean genome. The advancement of next-generation sequencing technology enables us to perform comparative genome analyses for related genomes, specifically soybean, other *Glycine* species, and other related legumes. This has given us the opportunity to look at associations at the genome level between several closely related species and how the pathways are controlling characteristics of interest.

The significant metabolic capacity of plants allows them to produce a wide range of secondary metabolites to interact with their dynamic environment. Conservation of these primary structures can be seen over a wide range of plants due the basic biosynthetic pathways that might be altered by an array of reactions to generate massive numbers of diverse secondary compounds (Schijlen et al. 2004). Although plants and their chemistry have been under study for a very long time, the understanding of the value of their secondary metabolites and their practical application is a rather new process. Adequate data displaying the occurrence of a variety of secondary metabolites as well as the biosynthesis of these compounds, make it possible to study the dynamic biosynthetic phylogeny (Wink and Waterman 1999).

One of the most studied pathways in legumes is the flavonoid biosynthetic pathway. Flavonoids play many important roles in plants and their interactions with their environment. Flavonoids are secondary metabolites and form a diverse family of aromatic molecules. They are major components that control flower colors and seed coat coloring in legumes. Several of these pigments have been used as markers to study different biological and evolutionary processes within the legumes. For instance, Makoi et al. (2010) studied flavonoids and anthocyanins as markers for effective plant defense in cowpea. Despite all the previous studies on the flavonoid biosynthetic pathway, our knowledge of the connection between the genotype and phenotype of the pathway, and the gene families elaborating the different characteristics of different species, particularly the gene families determining the flower colors and seed coat colors of different species, is still limited. Other characteristics and functions controlled by this pathway include attracting pollinators, defense signaling, UV protection, and signaling during root

nodulation. Better understanding of the evolution of this pathway leads to a better understanding of its functions. Having the full genome sequence of major legumes (soybean, Medicago, Lotus, common bean, pigeon pea, and chickpea) available as well as access to EST sequences of garden pea, and Chamaecrista makes it possible to study flavonoid pathway evolution across various legumes.

## 1.1 Legume Genomics – The Current State

Legumes are divided into three subfamilies: *Mimosoideae*, *Caesalpinioideae*, and *Papilionoideae*. Nearly all important crop legumes, such as soybean (*Glycine max*), peanut (*Arachis hypogaea*), mungbean (*Vigna radiata*), chickpea (*Cicer arietinum*), lentil (*Lens culinaris*), common bean (*Phaseolus vulgaris*), garden pea (*Pisum sativum*), and alfalfa (*Medicago sativa*) belong to the latter, Papilionoideae (figure 1.1). All these crops, excluding peanuts, belong to two of the papilionoid clades, Galegoid, cool season legumes, and Phaseoloid, tropical season legumes (Doyle and Luckow 2003). The morphology of legumes (flowers, seeds and fruits) also vary greatly between the different types and species. Even though the origins of flowers in legumes are not widely understood; most of these legumes are all grouped together because of shared characteristics, for instance the majority of legumes have flower shapes similar to peas (Domoney et al. 2006).

A thorough understanding of genome structural conservation among legume species will tremendously contribute to transferring knowledge between model and crop legumes (Doyle and Luckow 2003). Microsynteny and macrosynteny are well-maintained in grasses (Bennetzen 2000, Devos and Gale 2000), but less so in the legumes (Kevei et

al. 2005). Although macrosyntenic relationships can aid in our understanding of related plant species, there are a lot of exceptions: frequent local genic rearrangements including gene inversion, duplication, translocation, and insertion/deletion (Doyle and Luckow 2003). Regardless of genomic scale divergence (such as gene location and gene neighborhood), the gene content underlying traits is still conserved.

Paramount to understating the conservation of traits among related species is ample genomic information. For the legumes, *Glycine max* (soybean), *Medicago truncatula* (barrel medic), *Lotus japonicus*, *Phaseolus vulgaris* (common bean), *Cicer arietinum* (chickpea), and *Cajanus cajan* (pigeon pea) have fairly complete genome sequences with annotated gene models. As a related legume that sits basal to these sequenced legume, we have EST data from *Chamaecrista fasciculata* (partridge pea). In a historical context, as Mendelian traits were originally described in garden pea, we also have EST sequences from *Pisum sativum.*

Although not a legume, *Arabidopsis thaliana* has long been a model dicot organism to study legumes; it is a diploid plant with 2n = 10 chromosomes and a genome size of ~120 Mb, and around 10% transposable elements. The sequencing was done by an international collaboration in 2000 (The Arabidopsis Genome Initiative 2000). The realization of the need of plant community to develop more genomic resources for crops to address the diversity throughout the plant kingdom, has led to identification and sequencing of plant species other than *Arabidopsis thaliana,* specifically cultivated varieties of legumes. Initially, two model legumes were targeted, *Lotus japonicus* and *Medicago truncatula.* The Miyakogusa Consortium has carried out most of the research activities of *Lotus japonicus* in Japan (VandenBosch and Stacey 2003). As part of this

consortium, the Kazusa DNA Research Institute in Japan has sequenced the ~470 Mb genome (and is continually working to refine the assembly) and also has developed high-density linkage maps of all the six chromosomes as well as generating large number of expressed sequence tags (ESTs) (http://www.kazusa.or.jp/lotus/). Although *Lotus japonicus* and *Medicago truncatula* have two different developmental systems for nodulation, are determinate and indeterminate (respectively), they display several similar characteristics such as: having a small genome, fast generation time, being diploid, and they can be transformed and regenerated with *Agrobacterium tumefaciens* (Fedorova et al. 2002). *Lotus japonicus* genome is one of the few plant genomes rich in Pack-MULE transposons. 30.8% of its genome is transposable elements, out of which 10.4% are class one, and 8.1 % are class two transposons.

The genome of Medicago has high conservation with *Medicago sativa* (alfalfa; a related autotetraploid) and moderate conservation with soybean genome. The ~500 Mbp *Medicago truncatula* genome has been sequenced by a consortium of laboratories in US and Europe in 2011 (Young N. et al. 2011). Retrotransposons constitute 27% of the genome, while DNA transposons are 3.4%. Genetic linkage maps are available for both diploid and tetraploid species however a genome sequencing project has not been planned yet for alfalfa. *Glycine max* (soybean) is a diploidized tetraploid with 2n = 40 chromosomes and with the genome size of about 1115 Mb and it has the most and well developed genomic resources available among legumes. Its genome sequence was published in 2010 (Schmutz et al. 2010) supported by well-developed composite genetic maps, large number of ESTs and RNA-seq studies and a number of BAC libraries.

*Glycine max* genome has 43 % retrotransposons and 70% DNA transposons. More detailed information of these genomes are represented in table 1.1.

The Department of Energy's Joint Genome Institute (JGI) has made the preliminary version of *Phaseolus vulgaris* genome sequence with the size of 637 Mbp as well as gene models available (Schmutz J. et al. 2014). Genome of *Phaseolus vulgaris* is 41% transposons, out of which 35% include Retrotransposons and 5.3% are DNA transposons. Except in pericentromeric regions, *Phaseolus vulgaris* has widespread synteny to G*lycine max.* Some detailed annotation information has been addressed in table 1.1.

The draft genome sequence of diploid *Cicer arietinum* with the size of 738-Mb was made available in 2013 (Varshney et al. 2013). They were able to assemble 70% of the predicted genome length and 27571 genes were predicted, with 210 MB repeat elements. The repeat elements cover around 40.4% of the draft genome, out of which 27.31% are retrotransposons, and 4.55% are DNA transposons. Please refer to table one for more detailed information of the genome.

*Cajanus cajan with* the genome size of 833.07 Mb has been sequenced as well and the draft genome sequence for it was published in 2012 (Varshney et al.). This diploid plant has 48,680 predicted genes and 51.67% of the genome having repetitive sequence. Retrotransposons constitute 37.12% of repetitive sequences, while 8.77% are DNA transposons. See Table 1.1 shows more detailed information on the genome.

*Pisum sativum* (Garden pea) is a self-pollinated, 2n=14, with a genome size of about 4,300 Mb. Because of its big genome size, roughly 4 times bigger than soybean genome and 10 times bigger than *Medicago truncatula*, and a rich repetitive DNA

compared to other sequenced species, the whole genome sequence of garden pea has not been attempted. However, its genetic maps have been compiled by Noel Ellis and it has around 47,500 EST sequences available (http://users.aber.ac.uk/noe2/m_maps/index.htm). *Chamaecrista fasciculata* (partridge pea) has a genome estimated around 740 (Mbp) which is a moderate size genome. It is also an annual plant with a height of around 50 cm and is distributed over eastern North America, making it a great candidate for an emerging legume model species. Partridge pea is phylogenetically unique among legumes: It is among the very few caesalpinioid/mimosoid clade members that nodulate, and is also missing the papilionoid polyploidy event (Singer et al. 2009). Currently only EST sequence data is available for *Chamaecrista fasciculata* (Cannon S. et al. 2010).

## 1.2 The Perennial Glycine

Cultivated soybean (*Glycine soja* subspecies *max*) is an annual plant like its wild progenitor *Glycine soja* subspecies *soja*, and is native to north-eastern Asia. These two species are part of the subgenus *Soja*, one of the two subgenera of the genus *Glycine* (Doyle et al. 2004b), although they are commonly referred to as *Glycine max* (cultivated soybean) and *Glycine soja*. The *Glycine* contains both palaeopolyploids and neopolyploids lending this system to it often being used as a polyploid model (Doyle et al. 2004a). Although the wild varieties of soybean are native to East Asia with the domestication center being in the same region, many members of this genus are mostly native to Australia and Papua New Guinea, this includes at least 26 of perennial species. More interestingly, while some of these species like *Glycine canescens* are

geographically wide spread over Australia, some others like *Glycine aphyonota*, and *Glycine peratosa* are highly localized; making this genus an ideal candidate for studying the diversity of spatial geographic patterns (Gonzalez-Orozco et al. 2012). The annual *Glycines* are estimated to have diverged from the perennials from a common ancestor sometime around 5 MYA (Doyle et al. 2004a, Doyle et al. 2004b). Most species that belong to the legume tribe *Phaseoleae,* have a chromosome number of 2n = 20 or 22 (Goldblatt 1981). Diploid *Glycines* on the other hand, are all 2n = 38 or 40 as a result of the genome duplication which can be detected in the modern soybean genome and is thought to have occurred roughly 9-15 million years ago (Schmutz et al. 2010, Doyle et al. 2004a, Zhu et al. 1994, and Shoemaker et al. 1996). Consequently, the perennial *Glycines* share the same whole genome duplication events observed in *G. max.* In addition, a neopolyploid formation comprised of two of the *Glycine* taxa, *Glycine syndetika* and *Glycine tomentella*, acting as progenitors of allotetraploid *Glycine dolichocarpa* (figure 1.2) has been described. More than 2000 germplasm of perennial Glycines serve as a secondary germplasm pool for *Glycine max* as well as having desirable agronomic traits (Doyle et al. 2004b).

Although perennial *Glycine* species are genetically diverse, traits that are identified and characterized in these species have rarely been used to improve soybean and are limited to genes for resistance to pathogens such as brown spot (Lim and Hymowitz 1987), cyst nematode (Riggs et al. 1998), Sclerotinia stem rot (Hartman et al. 2000), and soybean rust (Burdon and Marshall 1981a, and Hartman et al. 1992). Only one successful cross to date has been done between *Glycine max* and *Glycine tomentella* (2n=78) (Chang et al. 2013). There is also a lack of high-density molecular markers from

*Glycine max* that can be used successfully to amplify fragments in perennial *Glycines* (Chang et al. 2013). Bronski et al. (2009) addressed this issue by identifying 13 microsatellite markers that were polymorphic among the A-genome perennials (*G. argyrea*, *G. canescens*, *G. clandestine*, *G. latrobeana*, *G. rubiginosa*, and *G. syndetika*). Other groups have used high-throughput Illumina sequencing of reduced representation libraries of genomic DNA of *Glycine max* parental lines to identify single-nucleotide polymorphism (SNP) and analyze corresponding mapping populations (Hyten et al. 2010 and Wu et al. 2010), but this has not been accomplished in the perennial species. A perennial *Glycine* project was developed as part of a larger NSF study to develop genomic resources in the perennial glycine. Prior to the start of this project, no sequence-based resources were available for the perennial *Glycine.* Table 1.2 summarizes some of the statistics on genomic libraries of selected number of species for this study.

## 1.3 The Flavonoid Biosynthetic Pathway

Over the past 150 years, some of the major discoveries in science have been developed around the physiological roles of compounds like anthocyanins and flavonols as pigments. These discoveries include the establishment of basic principles of genetics by Mendel (1866), where he observed seed coat color and flower color in garden pea as two of the seven traits in his study. Flavonoids were also instrumental in the discovery of transposable elements by Barbara McClintock (1950) while studying the mechanism of the mosaic color patterns of maize seeds and the RNAi phenomenon, which was first observed during a sequence of studies trying to alter flower color in petunias (Winkel-Shirley 2001).

Flavonoids, a class of polyphenols, are a diverse family of aromatic molecules in plants. Based on the oxidation state and substitution pattern of their C-ring, they can be classified into eight subgroups: flavanones, dihydroflavonols, flavones, flavonols, flavan-3,4-diols, flavan-3-ols, anthocyanidins, and proanthocyanidins. Legumes also synthesize isoflavonoids along with these subgroups (Hegnauer and Gpayer-Barkmeijer 1993). In addition to the flavonoids and anthocyanins providing the red, blue, and purple pigments for flowers, seeds and fruits in plants, they are well known for having key roles in a plant's defense, UV protection, seed dispersion, to attract pollinators, and as feeding deterrents. They are also being associated with some anticancer benefits (Kandaswami 2005), specifically in soy products (Barrett 2006), as well as being linked to the beneficial effects of red wine on reducing heart disease (Frankel 1993). Genomic, genetic and biochemical methods are adding to our understanding of how this pathway is controlled (Winkel-Shirley 2001).

Flavonoids are synthesized via the phenylpropanoid pathway. Six major subgroups of the flavonoid pathway that are found in plants are: chalcones, flavones, flavonols, flavandiols, anthocyanin and condensed tannins. Specialized forms of flavonoids, such as the isoflavonoids are represented in some plants. Last group is aurones that are not ubiquitous, although they are widespread. The first step is catalyzed by chalcone synthase (CHS), which uses 4-coumaroyl CoA as a substrate, yielding chalcone. Chalcone is isomerized to flavanone by chalcone isomerase, from which, the pathway diverges into several side branches, each initiating a different class of flavonoids. Dihydroflavonol reductase (DFR) catalyzes the dihydroflavonols to flavan-3,

4-diols (leucoanthocyanins), resulting in anthocyanidins by anthocyanidin synthase (Figure 1.3).

Flavonoid pigments including PAs and anthocyanins can be synthesized in the seed coat of Glycine max (Kovinich et al. 2012). Bernard R, and Weiss M (1973) identified six seed coat colors that were linked to spontaneous mutations in genetic loci by classical genetics (I, R, T, Wp, W1, and O). Alleles of I locus coding tandem genomic repeats of CHS genes are associated with the flavonoid-based seed coat colors (Bernard R, and Weiss M 1973; Todd JJ, and Vodkin LO 1996; and Tuteja et al. 2004). Kovinich et al. (2011) collected specific distinct anthocyanins, PAs, isoflavones, and phenylpropanoids in the seed coat and studied the differential expression of 20 flavonoid/phenylpropanoid isogenes. Their findings suggested that R locus encodes a regulatory gene, based on their results for black (iRT) or brown (irT) isolines. In contrast, different alleles of the pleiotropic T locus coding flavonoid 3′-hydroxylase (F3′H1) have been linked to black (iRT) and imperfect black (iRt) seed coat colors (Buzzell et al. 1987; Todd JJ, and Vodkin LO 1993; Woodworth CM 1921; Toda et al. 2002; and Zabala G, and Vodkin L 2003). Furthermore, alleles of the W1 locus, coding flavonoid-3′,5′-hydroxylase (F3′5′H), are linked to Imperfect black (W1) and buff (w1) seed coat colors in an iRt background (Bernard R, and Weiss M 1973; and Zabala G, and Vodkin L 2007) while pigmented soybeans carrying the Wp allele are black (iRTWp), whereas wp gives a lighter grayish (iRTwp) grain color (Zabala G, and Vodkin L 2005). Likewise, purple (Wp) and pink (wp) flower color have been associated with high or low levels of the flavonone 3-hydroxylase gene (F3H1), respectively (Zabala G, and Vodkin L 2005).

Also, in an irT background brown (O) and red-brown (o) grain color are controlled by the O locus (Bernard R, and Weiss M 1973).

### 1.3.1 Chalcone Synthase

Plants are an excellent model system to study the evolution of the flavonoid pathway and its different mechanisms. Except for chalcone synthase (CHS) and chalcone isomerase (CHI), which are found only in plants (Jez et al. 2000), most of the enzymes in the pathway are not limited to plants and can be found in all organisms (Dixon and Steele 1999) allowing for the studies in plants to have far reaching impacts. CHS is the starting step of flavonoid pathway and yielding naringenin chalcone as the major product of this step. The CHS gene family has been through repeated duplication and specialization events over the course of evolution in plants (Durbin et al. 2000). Work by Rausher et al. (1999) has suggested that upstream genes of the pathway, such as CHS have evolved more slowly than downstream genes (Rausher et al. 1999).

CHS has been reported numerously as key regulatory enzyme of flavonoid biosynthesis (Hahlbrock and Scheel 1989) and the study of its gene expression could be used as a molecular marker of flavonoid production (Laplaze et al. 1999). cDNA clones of the complementary CHS mRNA (Kreuzaler et al. 1983, and Ryder et al. 1984), have been used to show self-up-regulation which leads to escalation of mRNA and enzyme level, hence an increase in flavonoid pigmentation and isoflavonoid phytoalexins (Chappell and Hahlbrock 1984, Cramer et al. 1985, and Lawton and Lamb 1987).

Ryder et al. (1987) has demonstrated that there are at least six different CHS genes within the haploid *Phaseolus vulgaris* genome and some of them are tightly

clustered (Ryder et al. 1987). It has been shown in some studies in different species of plants such as petunias, (Koes et al. 1989) Ipomoea species (Durbin et al. 1995), and legumes (Ryder et al. 1987, Wingender et al. 1989, An et al. 1993, Junghans et al. 1993, and Howles et al. 1995) that CHS is a member of a multigene family. Other studies have reported multiple paralogs for CHS in Sinapis (Durbin et al. 1995).

One of the tools to study the functional effects of chalcone synthase is the introduction of sense and antisense genes, relating to anthocyanins, or flower colors. For instance, Van Der Krol et al. (1988) introduced the antisense CHS gene to petunia and tobacco and reported a suppression in the formation of flower pigments. It has also been shown that the introduction of the sense CHS gene (Napoli et al. 1990) or sense DFR gene (van der Krol et al. 1990) results in a reduced flower color pigmentation in transgenic petunia plants. A similar change of flower color from pink to white was observed in transgenic chrysanthemum, after the introduction of the sense CHS gene (Courtney-Gutterson et al. 1994) Elomaa et al. (1993) introduced the antisense CHS gene in gerbera which resulted in transgenic plants with a change in their flower colors from red to pink. The same procedure produced transgenic lisianthus plants with white flowers instead of purple (Deroles et al. 1995). In each of these cases, it is clear that the pigmentation or intensity of pigmentation is related to the expression of chalcone synthase suggesting (and shown) that this gene functions early in the pathway.

Aida et al (2000) phenotypically altered sense-introduced and antisense-introduced plants of torenia. . In their study they report that all of their transgenic torenia showed lighter color than that of wild-type plants, and had reduced mRNA levels of CHS and DFR as well as reduced anthocyanin contents of the corolla. Their results suggest

that gene suppression and subsequent flower pigments reduction, triggers modifications in flower color, although it does not clarify the cause of these alterations from a pigment composition aspect. It also puts forward a common sense/antisense pattern in CHS/DFR gene-introduced plants (Aida et al. 2000).

In Petunia hybrid, CHS is produced in the flower corolla, tube and anthers. Van Der Krol et al. (1988) has shown that reduction in levels of mRNA for the enzyme and the enzyme itself results in altered flower pigmentation. The observation of various pigmentation patterns in different transgenic plants, indicates that DNA sequences impact the activity of the anti-sense gene in both a quantitative and a qualitative way. They also showed through backcrossing experiments how these phenotypes are stably inherited displaying how the manipulation of secondary metabolites in plants is conceivable by synthesizing anti-sense RNA (Van Der Krol et al. 1988).

CHS has a multigene family in petunia (Koes et al., 1987) wherein only one gene is expressed to high levels in petal tissue (Koes et al., 1989a). Anthocyanin production rate can be limited by CHS in maize (Coe and Neuffer, 1977; Dooner, 1983), even though it is not limiting in Anfirrhinum (Sommer et al., 1988). Van der Krol et al. (1990b) have shown that endogenous CHS and DFR gene activities can be inhibited by introducing an intact CHS and DFR genes, respectively (Napoli et al. 1990).

Napoli et al. (1990) over expressed CHS in pigmented petunia petals by introducing its chimeric gene that created an unexpected block in anthocyanin biosynthesis. This resulted in forty two percent of plants producing either white flowers or patterned flowers with white or pale non-clonal parts on a wild-type pigmented background. Such phenotypes were not observed in any of the control transgenic plants.

Using a progeny test they also demonstrated how the novel color phenotype co-segregates with the introduced CHS gene. Wild-type phenotypes were observed in offspring without this introduced gene.

Analysis on the RNA isolates from white flowers revealed that the level of the mRNA produced by this gene was reduced 50-fold from wild-type even though the developmental timing of mRNA expression of the endogenous CHS gene was not altered. The coordinate increase of the steady-state levels of the mRNAs produced by both the endogenous and the introduced CHS genes is associated with the reversion of plants with white flowers to phenotypically parental violet flowers meaning that the expression of both genes was coordinately suppressed in the altered white flowers. Thus for suppressing endogenous CHS transcript levels the expression of the introduced CHS gene is not adequate on its own. The unpredictable and reversible nature of this occurrence advocates the possibility of being associated with methylation (Napoli et al. 1990).

### 1.3.2 Chalcone Isomerase

The distributions of CHIs are highly family specific and they are usually classified into two types. In non-legumes, they only isomerize 6′-hydroxychalcone to 5-hydroxyflavanone (CHI type I) whereas in legumes have activities toward both 6′-deoxychalcone and 6′-hydroxychalcone, yielding 5-deoxyflavanone and 5-hydroxyflavanone, respectively (CHI type II). In *Lotus japonicus* the identity between the type I and type II CHI's is about 50%, with more than 70% identity between the amino acid sequences of the same type of CHI (Shimada et al. 2003). In their study, Shimada et al. They have designed degenerate oligonucleotide primers based on sequences of CHI 1

and 2 using cDNA synthesized from mRNA of whole-plant organs and then based on the result they deducted the amino acid sequence. From the model legume, *Lotus japonicus*, three CHI cDNAs (cCHI1–3) and the corresponding genes as well as a putative CHI gene (CHI4) have been identified. Contrary to previous studies that screened only CHI type II in legumes, Shimada et al. (2003) found evidence of a hybrid of the type I and type II CHI that exist in Lotus. The only functional CHI gene (TT5) in Arabidopsis is essential for the biosynthesis of anthocyanin and other flavonoids (Winkel-Shirley et al. 1995). Presumably, in Arabidopsis, several enzymes including CHS, CHI, F3H, and DFR form a macromolecular complex on endomembranes while interacting with each other (Winkel-Shirley 2001, Burbulis and Winkel-Shirley 1999, and Saslowsky and Winkel-Shirley 2001). The TT5 mutant fully complements the CHI cDNA in maize (Dong et al. 2001). Nevertheless, enzymes such as chalcone synthase and CHI have been suggested to be parts of the putative enzyme complex in leguminous plant cells (Dixon et al. 1996).

French bean was the first plant that CHI was successfully isolated from by Mehdy and Lamb (1987). Later other studies cloned both types of CHI form different plant species such as: *Vitis vinifera* (Sparvoli et al. 1994), maize (Grotewold and Peterson 1994), *Lotus japonicus* (Shimada et al. 2003), and soybean (Ralston et al. 2005). Based on expression profiles, it has been revealed that in soybean type II CHIs are expressed in root, are regulated during plant-microbe interactions with other isoflavonoid pathway enzymes, and nodulation signals extremely induces them (Ralston et al. 2005; Marinova et al. 2007). In contrast type I CHIs are regulated by flavonoid pathway enzymes, hence associated with flavonoid biosynthesis, while type II is associated with isoflavonoid biosynthesis (Kim et al. 2007). Both type I and II CHIs have been isolated from some of

the legumes including: *Lotus japonicus*, *Medicago truncatula*, and *Glycine max* (Shimada et al. 2003; Ralston et al. 2005).

### 1.3.3 Aureusidin Synthase

Aureusidin synthase catalyzes the synthesis of aurones, another class of flavonoids. Together with anthocyanins they are considered to provide a nectar guide for pollination in plants, thus having an evolutionary importance regarding the plant–pollinator interaction (Lunau et al. 1996). Aurones cause a bright yellow color with characteristic fluorescence on some ornamental flowers, such as the snapdragon (*Antirrhinum majus* (Scrophulariaceae)) (Schwarz-Sommer et al. 2003, and Asen et al. 1972). Other forms of aurones, such as bracteatin, sulfuretin, and maritimetin, also have been identified in the Angiosperm genera *Antirrhinum, Dahlia, Oxalis, Linaria, Limonium, Coreopsis,* and *Bidens* (Harborne and Baxter 1999). Remarkably, distribution of aurone 6-O-glucosides develops to the Bryophyte such as *Marchantia* and *Conocephalum* (Harborne and Baxter 1999), proposing that aurone 6-O-glucosides-biosynthetic machinery predates floral evolution (Ono et al. 2006). Remarkable little wok has been done in the legumes to identify and characterize the aureusidin synthase genes. Although, yellow flowers have been implicated to be a result of expression from the aurone biosynthetic pathway (Ono et al. 2006).

### 1.3.4 Flavanone 3-Hydroxylase

One of the core enzymes acting at the branching point of anthocyanins and flavonols is flavanone 3-hydroxylase (F3H) (Shen et al. 2006). The function of F3H was first described from extracts of *Matthiola incama* and was further described through parsley cell cultures (Forkmann et al. 1980, Britsch et al. 1981, and Heller and Forkmann 1993). The F3H gene was first cloned from *Petunia hybrida* (Britsch et al. 1992). Subsequently, more F3H genes have been cloned and characterized from *Hordeum vulgare* (Meldgaard 1992), *Malus* (Davies 1993), *Medicago sativa* (Charrier et al. 1995), *Zea mays* (Deboo et al. 1995), *Arabidopsis thaliana* (Pelletier and Winkel-Shirley 1996) and *Perilla frutescenes* (Gong et al. 1997). Amino acid sequence of *Medicago* anthocyanidin synthase is 40% identical to other plants flavonol synthase (FLS) and flavanone 3-hydroxylase (F3H) (Prescott and John 1996). Pang et al. (2007) deducted the ANS amino acid sequence of *Medicago* and aligned them with amino acid sequences of FLS and F3H from several plant species, using ClustalW (Thompson et al. 1994), and build phylogenetic trees using a Neighbor Joining method (Saitou and Nei 1987). They observed three distinctive clusters of FLS, F3H, and ANS.

Zabala and Vodkin (2005) studied differential expression between the purple and pink flower isolines in soybean using cDNA microarrays. In combination with RNA gel blotting experiments they were able to confirm that there is a stronger correlation between F3H and a young purple flower bud (WpWp) rather than pink flower (wpwp), making it a strong candidate for the flower color gene Wp. In 2007, they furthered their study on this gene, discovering that a pink F3H1 in soybean is encoded by a pink flower

locus. They were also able to show the strong expression of F3H1 in the seed coats while it was not as much expressed in cotyledons.

### 1.3.5 Isoflavone synthase

Isoflavone synthase (IFS) is a unique compound that is only synthesized in legumes, such as *Glycine max* (soybean), *Phaseolus vulgaris* (green beans), *Pisum sativum* (peas), and *Medicago sativa* (alfalfa), and a small number of non-legume plants. 3-deoxyanthocyanins (or phlobaphenes in the polymerized form) can be found in a few species including sorghum (*Sorghum bicolor*), maize (*Zea mays*), and gloxinia (*Sinningia cardinalis*) (Winkel-Shirley 2001). Taking these facts and how other closely related compounds to flavonoids are synthesized under consideration, one may conclude that this pathway has changed multiple times and might even have been lost from some plant lineages (Winkel-Shirley 2001).

In the phenylpropanoid pathway one branch produces isoflavone synthesis while the other branches result in lignin and anthocyanin pigments (Dixon and Pavia 1995). IFS is the first step of the branch of the pathway that through several steps synthesizes Isoflavones (Woosuk et al., 2000). IFS in soybean was identified to have two gene copies, IFS1 and IFS2 (Akashi et al., 1999; Steele et al., 1999; Jung et al., 2000; Yu et al., 2000). Both of these copies are involved in isoflavone synthesis and compete with other enzymes like F3H over substrates like naringenin and liquiritigenin. Kim et al. (2005) isolated and sequenced IFS1 and IFS2 from form eighteen different soybean varieties though amino acid level variation among them was not discovered. Cheng et al. (2013) discovered one start codon mutation of IFS1 and four frameshift mutations of IFS2 that

could result in no function of these enzymes. However, Cheng et al. (2008) have shown that silencing of IFS genes had no significant effect on the isoflavone content of the accessions in their study. Based on these results and the paper published by Jung et al (2000) that showed both IFS1 and IFS2 enzyme convert naringenin to genistein and liquiritigenin to daidzein, one can conclude that IFS1 and IFS2 compensate for each other (Cheng et al., 2013).

Rhizobium nodulation genes have been shown to be induced by IFS. They also function as allelopathic agents (Dixon 1999), and act as antimicrobial phytoalexins as part of the plant defense system against microorganisms and herbivores' attack (Dixon 1999). Examples of anti-fungal isoflavonoids phytoalexins include medicarpin from alfalfa (Medicago sativa; Higgins 1972), pisatin from garden pea (Pisa sativum; Cruickshank and Perrin 1960), and maackiain from chickpea (Cicer arietinum; Daniel et al. 1990). The major isoflavones in soybean are daidzein, genistein, and glycitein (Graham, 1991).

### 1.3.6 Anthocyanidin 3-O-Glucosyltransferase

In addition to factors such as co-pigmentation, vacuolar pH, and cell shape, buildup of various structures of anthocyanins in the petal vacuoles results in different flower colors including, orange, red purple, blue, and blue-black (Morita, Y et al., 2005; Brouillard and Dangles, 1994; Davies and Schwinn, 1997; Honda and Saito, 2002; Mol et al., 1998). One of the main reactions involved in producing flavonoids with different varieties of structures and colors is glycosylation, which frequently occurs in the final biosynthetic steps (Gachon et al. 2005). Anthocyanidin 3-o-glucosyltransferase (UF3GT)

is one of the best-studied enzymes responsible for glycosylation, it accelerates the formation of first stable anthocyanin (Sui X. et al.; 2011). Its role is essential not only in modifying flower color, but also for hydrophobic flavonoids resulting in a rise of the solubility and stability (Hondo et al. 1992; Yoshida et al. 2000). Out of seventy seven identified families of glucosyltransferases (http://afmb.cnrs-mrs.fr/CAZY/), family one incorporates over 100 members in Arabidopsis (Arabidopsis GI 2000) and approximately 150 members in Medicago truncatula (Modolo et al. 2007). This family contains a forty four amino acid consensus sequence called the plant secondary product glycosyltransferase signature sequence (PSPG box) in the carboxy-terminal domain. Studies have detected PSPG box in the open-reading frames of animal, plant, yeast and bacterial genomes. However, general sequence similarity of UF3GTs, particularly in the N-terminal domain where it relates to binding-acceptor regions, is quite low (Sui X. et al.; 2011). Offen et al. (2006) explains how UF3GTs synthesizing a large number of products could be the result of this feature, which is a possible cause for the combination of huge variety of acceptors on UF3GTs.

Anthocyanidin 3-O-glucosyltransferases have been isolated from flowers of many ornamental plants, including *Gentiana triflora* (Tanaka et al. 1996), Petunia hybrida (Yamazaki et al. 2002) and Iris hollandica (Yoshihara et al. 2005). They have also been isolated from *Zea mays* (Goto et al. 1982), *Antirrhinum majus (*Martin et al. 1991), *Vitis vinifera* (Ford et al. 1998), *Hordeum vulgare* (Wise et al. 1990), *Perilla frutescens* (Gong et al. 1997), and *Fragaria ananassa* (Almeida et al. 2007). The flowers of the butterfly pea (Clitoria ternatea) are a great example of the biosynthesis of a group of anthocyanins named ternatins that are accumulated in their petals starting with the transfer of glucose

to specific anthocyanins like delphinidin. Anthocyanidin 3-o-glucosyltransferase catalyzes this reaction in butterfly pea (Hiromoto, T. et al., 2013). Its putative amino acid sequence is forty five percent identical to cyaniding 3-o-glycosyltransferase from *Vitis vinifera* (red grape) involved in the formation of anthocyanins (Offen et al., 2006).

### 1.3.7 Anthocyanidin Reductase

Anthocyanidin reductase is an NADPH- and/or NADH-dependent enzyme (Xie et al. 2003, 2004). It has been shown in several studies that ANR transfers two hydrides from two molecules of NADPH or NADH to anthocyanidins, such as cyaniding, delphinidin, and pelargonidin, and creates three types of isomeric flavan-3-ols (Fujita et al. 2005; Gargouri et al. 2009a; Pfeiffer et al. 2006; Punyasiri et al. 2004; Xie et al. 2003). In 2003 Xie et al. discovered this reductase in both *Medicago truncatula* and *Arabidopsis thaliana*, and by using an in vitro experiment with a recombinant protein indicated that in its presence, anthocyanidin reductase catalyzes 2R,3R-cis-flavan-3-ols and 2S,3R-trans-flavan-3-ols. Other groups have validated that this reductase transforms ANR pathway to flavan-3-ols and proanthocyanidins (PAs, also called condensed tannins), responsible for brown pigmentation of seeds (Kovinich et al. 2012; Fujita et al. 2005; Peng et al. 2012; Zhu et al. 2013).

Because of the economic significance of plant seeds and fruits, investigators generally have focused on roles of ANR in plant PA biosynthesis in these two organs (Zhu et al. 2014). Some of the plants that have been studied for this purpose include *Arabidopsis thaliana* (Albert et al. 1997; Devic et al. 1999; Nesi et al. 2001; Xie et al. 2003), *Medicago truncatula* (Pang et al. 2007), grape (Bogs et al. 2005; Fujita et al. 2005;

Gargouri et al. 2009b), apple (Pfeiffer et al. 2006; Takos et al. 2006), soybean (Kovinich et al. 2012), persimmon (Ikegami et al. 2007), and others (Peng et al. 2012).

In 2010, Yang et al. mapped a sequence from soybean, with 86 percent sequence identity to an ANR gene from *Medicago truncatula*, to the O locus coding region. Surprisingly, the functional relationship between an ANR encoding gene and the O locus was not further explored. Earlier studies in Arabidopsis had revealed that by knocking out mutations in ANR gene, cyanic pigments will buildup temporarily during early seed development (Albert et al. 1997). Kovinich et al. (2012) suggested that a similar phenomenon in red-brown soybean seeds suggest a defect in ANR gene. Likewise, flavonoid-3-O-glucosyltransferase gene (UGT78K1) was isolated from the black soybean seed coat (Kovinich et al. 2010), and by studying the developmental stages of the seed coat of black soybeans, anthocyanin biosynthesis was found to be up-regulated with a second flavonoid-3-O-glucosyltransferase gene (UGT78K2) together with anthocyanidin 3′-O-methyltransferase gene (OMT5) (Kovinich et al. 2011). The differences between brown and red-brown soybean seed coats were studied in 2012 by Kovinich et al. and they identified two ANR genes (ANR1 and ANR2) from soybean seed coat tissue. They concluded that manipulation of ANR1 can lead to red-brown soybean color, based on how ANR1 transcript profiles, seed coat PA content.

### 1.3.8 Anthocyanin Synthase

In late stages of flavonoid biosynthesis pathway, leucoanthocyanidins are converted into 3-OH-anthocyanidin by anthocyanin synthase (ANS, also known as leucoanthocyanidin dioxygenase, LDOX) (Wilmouth et al. 2002). Genes and cDNAs

encoding ANS have been isolated from a number of plant species including *Arabidopsis thaliana* (Abrahams et al. 2003; Xie et al. 2003; Devic et al. 1999; Pelletier et al. 1997; and Wilmouth et al. 2002), *Medicago truncatula* (Xie et al. 2003; and Pang et al. 2007 ), *Vitis vinifera* (Bogs et al. 2005), *Perilla frutescens* (Saito et al. 1999), *Spinacia oleracea*, *Phytolacca americana,* and other plants (Shimada et al. 2005; Wellmann et al. 2006; and Shih et al. 2008). Mutants of LDOX in Arabidopsis cause a lack of anthocyanin accumulation in hypocotyls and PA deposition in seeds. This results in a transparent testa phenotype (Abrahams et al. 2003). Pang et al.'s 2007 study in Medicago showed decreased levels of both anthocyanins in leaves and PAs in seeds as a result of an antisense down-regulation.

Alternatively, an up-regulation of an ANS gene in *Arabidopsis thaliana* and *Vitis vinifera* causes a solid increase in the accumulation of PA and anthocyanin (Abrahams et al. 2003; Gollop et al. 2001; Lepiniec et al. 2006; Pelletier et al. 1997; Solfanelli et al. 2006). ANS in Arabidopsis is encoded by the gene TT18 (TDS4) and its mutation results in decreased levels of PA in the seed coat and consequently yellow seeds in comparison to the brown seed in wild-type *Arabidopsis thaliana* (Abrahams et al. 2003; Lepiniec et al. 2006). Studies to date have indicated a lack of the pattern of proanthocyanidins that corresponds to the pigmentation phenotype observed in yellow-seeded rapeseed cultivars; PAs synthesized in embryos suggest seed coat as the independent target in the study of yellow-seeded rapeseed (Marles and Gruber 2004; Nesi et al. 2009; Yan 2007). Mingli et al. (2010) suggested that proanthocyanidin biosynthesis in black-seeded rapeseeds could be activated by ANS gene expression.

### 1.3.9 Dihydroflavonol 4-Reductase

Dihydroflavonol-4-reductase (DFR, EC 1.1.1.219) is one of the rate-limited enzymes of the flavonoid biosynthetic pathway. The stereospecific reduction of three dihydroflavonols (dihydrokaempferol, dihydroquercetin and dihydromyricetin) to leucoanthocyanidins (flavan-3,4-diols) is catalyzed by this enzyme (Martens et al. 2003; Kristiansen and Rohde 1991; and Peters and Constabel 2002). It is also known that in the formation of PAs, the precursors of the anthocyanin branch, leucoanthocyanidins, are essential (Xie et al. 2004). Flavonol synthase (FLS) can catalyze dihydroflavonol of DFR to flavonols. Likewise, it can result in the production of leucoanthocyanidins that can later be converted to proanthocyanidin by leucoanthocyanidin reductase (LAR) (Davies et al. 2003; Martens et al. 2010; and Yoshida et al. 2010).

Numerous DFR cDNAs have been isolated from different species, such as *Arabidopsis thaliana*, maize (Zea mays), Vitis vinifera, barley (Hordeum vulgare), trembling aspen (Populus tremuloides), Medicago truncatula, and Petunia hybrida (Xie et al. 2004; Beld et al. 1989; Sparvoli et al. 1994; Liew et al. 1998; Fisher et al. 2003; Lo Piero et al. 2006; Helariutta et al. 1993; Tanaka et al. 1995; Bernhardt et al. 1998; Inagaki et al. 1999; and Himi and Noda 2004). A single copy DFR is present in A. thaliana, barley, tomato (Lycopersicon esculentum), grape (Vitis vinifera), snapdragon (Antirrhinum majus) and rice (Oryza sativa), while multi-copy DFRs exist in Petunia hybrida (Line V30), Ipomoea purpurea, Ipomoea nil, and Medicago truncatula (Peters and Constabel 2002; Xie et al. 2004; Beld et al. 1989; Devic et al. 1999; Inagaki et al. 1999; and Ostergaard et al. 2001). In mutants of barley and Arabidopsis, loss of anthocyanins and PAs has occurred after deactivating DFR gene (Olsen et al. 1993;

Shirley et al. 1995). DFR activities are also absent in tt3 (transparent testa) mutants of Arabidopsis. These mutants not only lacked brown tannins of proanthocyanidin in their seat coats but also, unlike wild Arabidopsis seedlings with strong red pigmentation, exhibited zero anthocyanin pigments within the cotyledon or hypocotyl (Shirley et al. 1995; Hsieh et al. 1998). It was shown that the pigmentation within the cotyledon and seed coat in tt3 mutants could be restored by the expression of the maize A1 gene that encodes DFR (Dong et al. 2001). Aida et al. (2000) were able to modify flower colors in ornamental plants by regulating the expression levels of DFR genes, and other studies have introduced a DFR gene into a forage legume Lotus corniculatus, to alter PA levels (Bavage et al., 1997; Robbins et al., 1998). Similarly, Norimoto et al. (2005) reported on the structure and function of five DFR genes in Lotus japonicus that formed a cluster within a 38 kb region. Min Xu et al. (2010) studied flowers of four soybean lines and the presence of flavonols and anthocyanins in them. They discovered that the multi-colored flower color in petals is regulated by the w4-m allele, and is a result of encoding of DFR from DFR2 that its CACTA-type transposable element has been removed.

### 1.3.10 Flavonol Synthase

Flavonols are the most abundant among the eight major subgroups of flavonoids in plants, being linked to many physiological functions such as: regulation of auxin transport, modulation of flower color, protection against ultraviolet radiation, prevention against microorganism and pest invasion, and signaling interactions with insects and microbes (Bohm et al. 1998; Harborne and Williams 2000; Winkel-Shirley 2001).

Glycosylated and methylated forms of flavonols can be found in vacuoles and in the cell walls of plants (Grotewold 2005; Yazaki 2005).

Dihydroflavonols dihydrokaempferol, dihydroquercetin or dihydromyricetin can convert to three common structures of flavonols; kaempferol, quercetin, and myricetin. These three classes vary by a single hydroxyl group on the flavonoid B ring. The hydroxylation can occur at the 3′-position or at the 3′- and 5′-positions, depending on the enzyme involved; flavonoid 3′-hydroxylase (F3′H) or flavonoid 3′,5′-hydroxylase (F3′5′H). (Toh et al. 2012; Hagmann et al. 1983). In flavonol synthesis, flavonol synthase (FLS) forms two precursor compounds of rutin, quercetin, and kaempferol and competes with F3′H and F3′5′H for substrates (Kaltenbach et al. 1999). FLS also competes with dihydroflavonol 4-reductase (DFR) for the common substrate dihydroflavonols, leading to production of flavonols and anthocyanidins/catechins, respectively (Mahajan et al. 2011) FLS belongs to the 2-oxoglutarate-dependent dioxygenase (2-ODD) superfamily and for full activity needs ascorbic acid (Holton et al. 1993; Prescott and John 1996). FLS is believed to have hydroxylation and desaturation activities and is generally classified as a bifunctional dioxygenase since it converts both flavanones and dihydroflavonols to the related flavonols (Lukačin et al. 2003; Prescott et al. 2002).

In 2004, Turnbull et al. characterized the first captured FLS activity in irradiated parsley cells, and a FLS cDNA was later cloned and identified in Petunia hybrida, Arabidopsis thaliana, Eustoma grandiflorum, Solanum tuberosum, Malus domestica, Matthiola incana, Citrus unshiu (Holton et al 1993; Wisman et al. 1998; Van Eldik et al. 1997; and Li et al. 2012). Mahajan et al. (2011) performed post-transcriptional gene silencing of FLS encoding mRNA in order to produce less-seeded or seedless fruits in

tobacco. They observed reduction in FLS encoding gene expression, and quercetin and anthocyanidins content in FLS silenced tobacco lines, while flavan-3-ols had increased in content. Furthermore, delayed flowering, disadvantaged pollen germination, lack of functional pollen tube, and lower number of seeds in each pod were observed in these silenced tobacco lines. Several studies have suggested that in plants a multicopy gene encodes FLS (Pelletier et al. 1999; Preuß et al. 2009; Ferreyra et al. 2010; and kim et al. 2012). Takahashi et al. (2007) found that a single-base deletion in the FLS gene of Glycine max changes the purple flower color to magenta, therefore FLS has a major role in determining flower color in *Glycine max* (soybean). Additionally, the buildup of flavonols in soybean has also been associated with the induction of FLS by ultraviolet-B irradiation (Kim et al. 2008).

### 1.3.11 Flavone Synthase

Flavanones that are the products of CHI can be synthesized to flavones by flavone synthase (FNS). Flavones are plants antioxidants in addition to being involved in signaling in root nodulation of legumes (Peters et al., 1986). Despite being colorless, once the flavones form a complex with anthocyanin pigments, they can alter flower colors (Goto and Kondo, 1991). Flavone-deficient roots with reduced nodulation have been observed in Medicago truncatula as a result of FNS silencing (Zhang et al., 2007, 2009).

Two distinct FNS enzymes have been described in dicots, FNSI and FNSII. FNSI is mainly restricted to Apiaceae (Martens and Mithöfer, 2005) and only recently has been described in monocotyledonous plants as well (Kim et al., 2008). It has high sequence

identity to F3H, and they both use same flavanone substrate (du et al. 2009). Gebhardt et al., in their 2007 study in parsley suggested that FNSI has evolved from a functional diversification that happened after a gene duplication of F3H. FNSII is detected in other dicot families such as *Leguminosae, Asteraceae, Plantaginaceae,* and *Lamiaceae* (Martens and Mithöfer, 2005). Most of the characterized FNSII's convert flavanone to flavone directly such as erbera hybrids (CYP93B2) (Martens and Forkmann, 1999), *Antirrhinum majus* (CYP93B3), *Torenia hybrida* (CYP93B4) (Akashi et al., 1999), and *Perilla frutescens* (CYP93B6) (Kitada et al., 2001). In an in vitro study of FNSII in *Glycyrrhiza echinata* (licorice) for CYP93B1, and in Medicago truncatula for CYP93B10, and CYP93B11, flavones were only produced by acid treatment, while FNSII formed 2-hydroxyflavanones, since its genes encode enzymes with flavanone 2-hydroxylase activities (Akashi et al., 1998; Zhang et al., 2007).

Schüler et al. (2004) reported that after coronalon treatment, flavone 7,4′-dihydroxyflavone (DHF) accumulates in soybean cell cultures. Fliegmann et al. (2010) only detected FNSII in oxylipin-induced soybean cell cultures. The induction of FNSII for osmotic stress has been described previously by Kochs and Grisebach, 1987 and Kochs et al., 1987. Fliegmanna et al. (2010) identified Glyma12g07190.1 as CYP93B16 coding gene with a duplicated gene Glyma12g07200.1 that is not expressed.

### 1.3.12 Leucoanthocyanidin Reductase

Leucoanthocyanidin reductase (LAR) is one of the three enzymes, along with ANS and ANR, involved in biosynthesis of flavan-3-ols, the building blocks of proanthocyanidins (PAs). LAR can convert flavan-3,4-diols such as leucoanthocyanidin

to 2,3-trans-flavan-3-ol (catechin) (Tanner et al., 2003). Genes encoding LAR have been isolated from numerous plant species such as *Desmodium uncinatum*, a legume (Tanner et al., 2003), *Vitis vinifera* (Bogs et al., 2005), *Lotus corniculatus* (Paolocci et al., 2007) and *Medicago truncatula* (Pang et al., 2007). However, an intact LAR orthologue has not been discovered in genomic sequence of *Arabidopsis thaliana*, neither has catechin been measured in its seed extracts (Lepiniec et al., 2006; Tanner et al., 2003; Abrahams et al., 2003).

In *Arabidopsis thaliana* transparent testa (tt) is a result of mutations in the BANYLUS gene and associated with loss of condensed tannins in the seed coat while red anthocyanin is accumulated. The amino acid sequence of BANYLUS is very similar to DFR sequence, suggesting that BANYLUS encodes LAR (Devic et al., 1999). Xie et al. (2003) were able to identify a cDNA (AY184243) that is expressed in young seeds of *Medicago truncatula* and has fifty nine percent amino acid sequence identity to *Arabidopsis thaliana* BANYLUS (AF092912). They also reported that in transgenic tobacco with *Medicago* BANYLUS expressed, petals will change color from pink to white. However, the *Arabidopsis* BANYLUS protein is only twenty five percent identical to *Desmodium* LAR and when *Arabidopsis* BANYLUS is expressed in transgenic tobacco although epicatechin is accumulated, measurement of condensed tannin show no results. This suggests that other genes are essential for synthesis of PAs (Tanner et al., 2003).

## 1.4 Genetic Diversity Relating to Flower and Seed Coat Color

The major secondary metabolic system of the phenylpropanoid pathway comprises a complex system of sub-pathways that intervenes with the transformation of phenylalanine to pigments, phytoalexins, cell-wall components, and signaling molecules (Dooner and Robbins 1991, and Hahlbrock and Scheel 1989). Particularly, the sub-pathway leading to the synthesis of flavonoids has been thoroughly characterized, to some extent because of non-lethal mutations of flavonoid genes which results into easily detectable phenotypes such as diverse flower or seed coat color (Winkel-Shirley et al. 1995). A more detailed understanding of the enzymology of the flavonoid pathway, structural and regulatory loci, and some physiological roles of its end products has been made possible through intense genetic and biochemical studies on the pathway. There have been numerous efforts on cloning the genes corresponding to different loci in the pathway as well as revealing control of their expression during development and in response to stress (Winkel-Shirley et al. 1995). Of course, much of this has been done in model systems and the entire pathway and relationship to genetic loci has not been outlined completely in many legumes.

Buildup of flavonoids (including anthocyanins), carotenoids and betalains is what causes the coloration of plant organs (Mol et al. 1998). However, in higher plants anthocyanins are the major flower pigments and have been studied extensively (Elomaa and Holton 1994, Koes et al. 1994, Holton and Cornish 1995, and Mol et al. 1996). In addition to anthocyanins, co-pigmentation, vacuolar PH, and cell shape determine the shade of flower color (Mol et al. 1998). Although many genes that encode the enzymes of the anthocyanin pathway have been isolated (Holton and Cornish 1995), new mutations

can result in new flower or seed coat colors (Mol et al. 1998). It is nearly impossible to observe the entire possible flower colors in one species; for instance, species such as rose and chrysanthemum lack flavonoid 3′5′-hydroxylase (F3′5′H) activity, hence do not synthesize the purple delphinidin derivatives (Elomaa and Holton 1994). Likewise, petunia lacks orange pelargonidin-type anthocyanins since the dihydroflavonol 4-reductase (DFR) enzyme in petunia does not accept the dihydrokaempferol precursor as a substrate (Meyer et al. 1987). Makoi et al. (2010) have measured levels of flavonoids and anthocyanins in 45 randomly selected cowpea genotypes with differing seed coat colors and found significant differences in the levels of flavonoids and anthocyanins in seeds of the different cowpea genotypes.

Recently, mutants that alter flower and seed coat color have been isolated with the purpose of explaining flavonoid biosynthetic pathway through a molecular genetic approach. The experimental model species used in early studies of this kind were maize, snapdragon (*Antirrhinum majus*), and petunia. Scientists were able to isolate many of the flavonoid's structural and regulatory genes (Winkel-Shirley 2001, Holton and Cornish 1995, and Mol et al. 1998). *Arabidopsis* on the other hand has been chosen more recently for analysis of the regulation of the flavonoid pathway, mostly due to its single gene copy for the corresponding gene families. Flavonol synthase (FLS) and all the other enzymes of central flavonoid metabolism are controlled by single-copy genes in *Arabidopsis thaliana*. While flavonoids do not play the same roles in *Arabidopsis* as other plant species, such as in defense or male fertility, defining the function of each compound in such processes has been aided with the study of *Arabidopsis* mutants (Winkel-Shirley 2001, Li et al. 1993, Landry et al. 1995, Murphy et al. 2000, and Brown et al. 2001).

The earliest studies in associating pigment with a particular gene were conducted by Gregor Mendel before he even knew what genes were. As a result of Mendel's study on pea, and his famous ratio 1 AA:2 Aa:1 aa (Mendel 1866), A is a symbol of the gene that determines the accumulation of anthocyanin pigmentation in the plant, particularly in flowers. The accumulation of anthocyanin pigments is the reason for the purple flower color in wild pea, while the observation of white flower color in cultivated pea is bHLH transcription factor homolog (Hellens et al. 2010).

Several studies have focused on the isoflavonoid pathway, resulting in isolation of some isoflavone synthase (IFS) genes. Soybean has two isoforms of IFS, which produce genistein or daidzein using liquiritigenin and naringenin respectively (Winkel-Shirley 2001). A group at DuPont Wilmington was able to indicate that by introducing IFS1 and chalcone reductase (CHI) together, providing liquiritigenin as a substrate, daidzein can be synthesized in maize (Yu et al. 2000). It has also been shown that the control of hydroxylation by F3′H causes the production of brick-red to orange pelargonidin-based pigments; and if it is controlled by F3′5′H, it will lead to synthesis of purple and blue delphinidin-based pigments (Winkel-Shirley 2001).

While the flower color of the progenitor of cultivated soybean, *Glycine soja,* is almost exclusively purple, about one third of the soybean accessions in the USDA Soybean Germplasm Collections have white flowers. The few white flowered *Glycine soja* accessions found in Crop Germplasm Resource of China have high seed weight, suggesting a recent outcross with G*lycine max* (Takahashi et al. 2010). One of the progenies of a purple-flowered *Glycine soja* accession (PI 424008A) has white flower colors, and it is believed to be the result of a recessive allele at the W1 locus similar to

white-flowered soybeans (Chen and Nelson 2004). The hydroxylation of the b-ring of flavonoids can happen either at the 3' position or at both the 3' and 5' positions. They lead to the production of cyanidin-based pigments or they produce delphinidin-based pigments respectively. The patterns of this hydroxylation play an important role in the coloration of seed coats, flower and pubescence of soybeans. Two key enzymes involved in this pathway are flavonoid 3'-hydroxylase (F3'H) and flavonoid 3'5'-hydroxylase (F3'5'H) (Forkmann 1991).

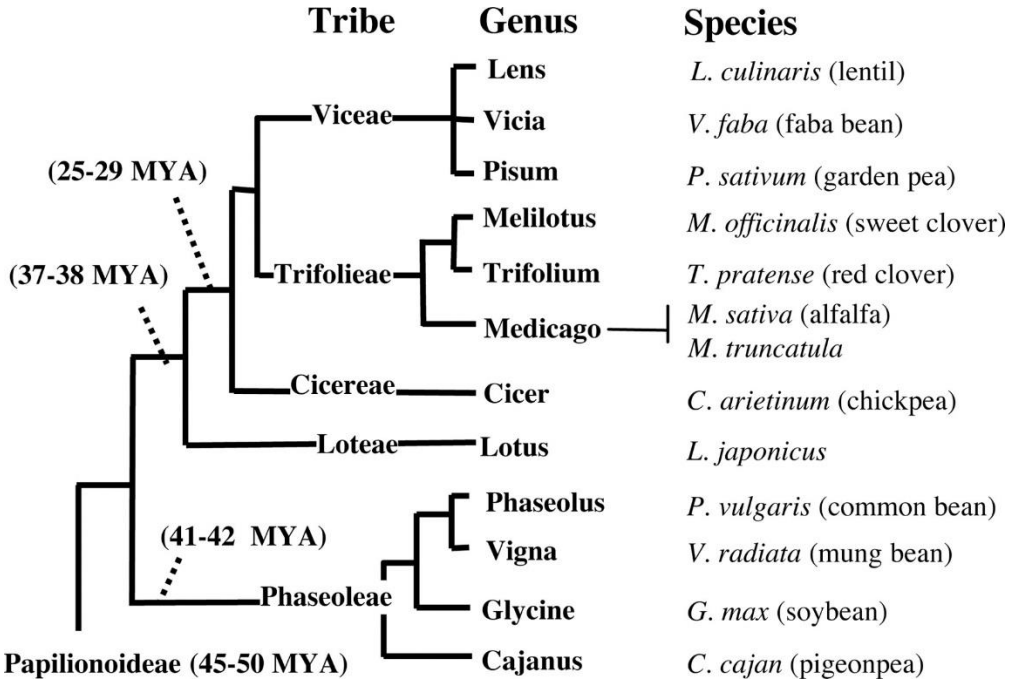



Figure 1.1: Taxonomic relationships within the two major clades of crop legumes (Choi et al. 2004)

Table 1.1: Current state of legume genetics

| Species | Genome Mbp | chrom. Num. | Ave gene Size | GC content | Predicted genes | EST based contigs |
|---|---|---|---|---|---|---|
| Glycine max | 1120 | 20 | 3.6 kb | ~34% | 66153 | |
| Medicago truncatula | 470 | 8 | ~2.5 kb | ~35% | 47529 | |
| Lotus japonicus | 471 | 6 | ~2.7 kb | ~ 45% | 47486 | |
| Cicer arietinum | 738 | 8 | 3.1kb | 26.9% | 28269 | |
| Cajanus cajan | 833.07 | 11 | 2.3 kb | ~32% | 48680 | 2,246 |
| Phaseolus vulgaris | 637 | 11 | ~2.7 kb | 39.4% | 41391 | |
| Arabidopsis thaliana | 125 | 5 | ~2 kb | ~20% | 25498 | |
| Pisum sativum | 4300 | 7 | | | | 37455 |
| Chamaecrista fasciculata | 740 | 8 | | 44% | | 21781 |

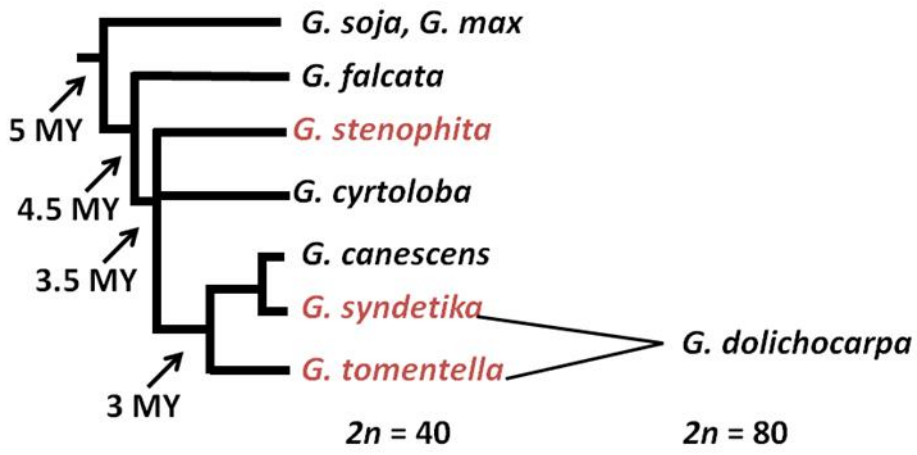



Figure 1.2: Neopolyploid progenitors. Courtesy of Sue Sherman-Broyles, Doyle lab, Cornell University.

Table 1.2: Species, genome and accessions - Glycine species selected for study and statistics on genomics libraries and sequencing (Jeff Doyle, personal communication)

| Species | Genome group | 2n | Genome size (Mbp) | Accession | #. Accession |
|---|---|---|---|---|---|
| G. soja | A | 40 | 1100 | PI468916 | |
| G. tomentella | D | 40 | 1207 | G1403 | 52 |
| G. canescens | A | 40 | 1038 | G1232 | 215 |
| G. stenophita | B | 40 | 803 | G1974 | 40 |
| G. cyrtoloba | C | 40 | 1291 | G1267 | 50 |
| G. dolichocarpa | 2(Syn. x D3)[e] | 80 | 2376 | G1134 | 37 |
| G. syndetika | A | 40 | 1340 | G1300 | 10 |
| G. falcata | | 40 | 1241 | G1155 | 53 |

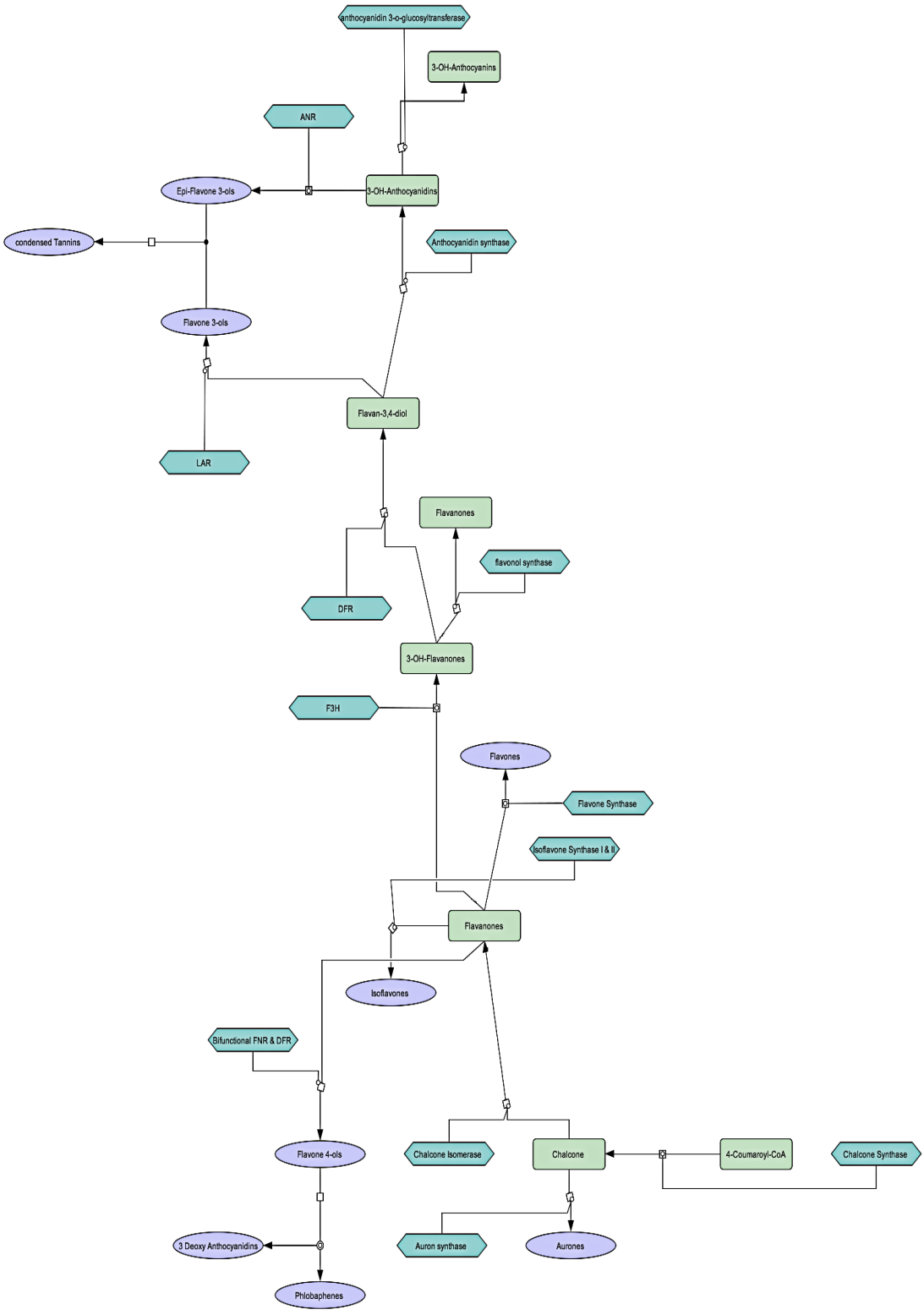



Figure 1.3. Central Flavonoid Biosynthesis Pathway

# CHAPTER 2: A PHYLOGENETIC STUDY OF THE FLAVONOID PATHWAY IN LEGUMES, AN EXAMPLE OF "TRAIT SYNTENY"

## 2.1 Introduction

Sequences that are descendants of a common precursor molecule through a chain of replication and divergence are homologous sequences (Cartmill 1994), and if their residues maintain the same position they will be homologous as well (Dewey and Pachter 2006). Sequence alignment has four main purposes; database searching, structure prediction, sequence comparison, and phylogenetic analysis, but homology is only essential when the sequence alignment is going to be used for a phylogenetic analysis (Morrison 2008). Homology is characterized by shared derived character states that are the results of the effect of some events on ancestral characters during the evolutionary history. Thus, if we align the sequences, the history of these events can be observed (Morrison 2008). For sequence comparison analyses, the objective focuses on comparing the residues that characterize conserved features of the sequence, like motifs that occur at active sites. The focus of phylogenetic analyses is a plausible hypothesis of evolutionary homology among the sequence residues (Morrison 2008). Independent sources of evidence for possible homologies are necessary in order to have a valid hypothesis in a phylogenetic context. Comparative sequence analysis can be used to obtain this evidence

through core molecular processes that result in the variations associated with the homologies (Morrison 2006).

Classical phylogenetic studies have focused on defining homology in order to define the relationships between species. To establish homology, one needs to describe characters and their states. In practice, this could be a little confusing for phenotypic characters, since there could be a large number of types of units to relate and place into characters. Although there are several units to compare, they can be used as evidence of which character they belong to. When it comes to comparing the units at the DNA level, the number of units reduces and they are more easily recognizable. At this level, the units are nucleotides A, G, C and T. However, this means that arranging them and then organizing them into columns is harder than the phenotypic level because all the units are identical. Furthermore, with fewer distinct states, the trees become heavily reliant on the accuracy of the sequence alignments. Consequently, shuffling and rearranging the units into different characters and states is the best way to demonstrate the evolutionary history of them (Morrison 2008, Patterson 1988). Statistical based alignment methods, such as Bayesian posterior probability use a probabilistic approach. By using substitutions and indels as evolutionary events, it creates distinguishable models of sequence evolution in a likelihood context, and later some of these measures are used to optimize parameters in relation to the model.

The role of polyploidy is a major force in the evolution of new plant species. Polyploidy in plants is often cyclic with periods of polyploidization followed by diploidization with the pattern repeating itself. In this study, we are investigating and characterizing the evolution of major genes that are part of the flavonoid biosynthesis

pathway. We are studying the role of duplication within soybean as well as other legumes relating to the evolution of this biosynthetic pathway. The gene families of this pathway have not been defined in many of the legume species, including those with fully sequenced genomes. Using the concept of "trait synteny", allows us to fill out the gaps of knowledge in these pathways among legume species. We have identified putative orthologs for genes in these pathways from six sequenced legumes; *Glycine max*, *Medicago truncatula*, *Lotus japonicus*, *Cicer arietinum*, *Cajanus cajan*, and *Phaseolus vulgaris*, and *Arabidopsis thaliana* as one outgroup. We also have used EST sequences of *Pisum sativum* and *Chamaecrista fasciculata*. We included EST sequences of garden pea due to the historical significance of pea as the foundation of Mendelian inheritance. *Chamaecrista fasciculata* belongs to the paraphyletic subfamily *Caesalpinioideae* within the mimosoid clade. Approximately 60 Million years ago, concurrent with the origin of legumes, this clade diverged from the common ancestor of soybean (*Glycine max*), *Medicago truncatula*, and *Lotus japonicus* (Singer et al. 2009). Its phylogenetic position makes it an ideal legume-specific outgroup in addition the non-leguminous *Arabidopsis thaliana*.

## 2.2 Materials and Methods

Soybean with the genome size of 1.12 Gb and 66,153 predicted gene models (Schmutz et al. 2010) was chosen for this study. Gmax_189_cds.fa was used to create the BLAST database and was downloaded from www.phytozome.net/soybean.php. *Medicago truncatula* with the genome size of 470 (Mbp) and 47,529 gene models (Bennett and Leitch 2010) was obtained from http://www.phytozome.net/medicago.php

and Mtruncatula_198_cds.fa was used to create the BLAST database. *Lotus japonicus* with the genome size of 471 (Mbp) and 47,486 gene models was obtained from ftp://ftp.kazusa.or.jp/pub/lotus/lotus_r2.5/ and the BLAST database was created using Lj2.5_cds.ffn..*Cicer arietinum* with 738-Mb draft whole genome shotgun sequence and 28,269 estimated genes (Varshney et al. 2013), was obtained from http://www.icrisat.org/gt-bt/ICGGC/GenomeManuscript.htm and Cicer_arietinum_GA_v1.0.gene.cds.fa used for analysis. *Cajanus cajan* with the genome size of 833.07 Mb and 48,680 predicted genes (Varshney et al. 2012) available from http://www.icrisat.org/gt-bt/iipg/genomedata.zip. *Arabidopsis thaliana* with the genome size of 125 Mb and 25,498 predicted genes was obtained from ondex.rothamsted.ac.uk/QTLNetMinerArabidopsis/html/release.html and Athaliana_167_cds.fa was used for analysis. *Phaseolus vulgaris* with the genome size of 637 Mbp and 41,391 gene models (Arumuganathan and Earle 1991), was obtained from http://genome.jgi.doe.gov/pages/dynamicOrganismDownload.jsf?organism=Pvulgaris and Pvulgaris_218_cds.fa used for analysis. *Chamaecrista fasciculata* with the genome size of ~740 (Mbp) and 21,781 EST based contigs was obtained from http://serc.carleton.edu/exploring_genomics/chamaecrista/index.html and chamae_genomic_seq.fa was used to create the BLAST database. For *Pisum sativum* (genome size of 4,300 Mb and 37,455 EST based contigs) the data Pisum sativum unigene v2 was obtained from http://www.coolseasonfoodlegume.org/sativum_unigene_v2.

All of the published sequences of enzymes of concern from species of interest, if available, were downloaded from NCBI or the appropriate databases. For each of our

genes of interest, we tried to initially identify an annotated *Glycine max* ortholog, and then search for orthologs from related legumes. If there was no identifiable gene by searching annotations in those species, we looked in closely related legumes such as *Medicago sativa* or *Glycine soja*. In one case, we had to expand our search to more distant species like *Vitis vinifera* before we could find our gene of interest. In order to obtain all the putative orthologs from each of our species of interest, we used queried our annotated gene to run tBLASTx against each of the species' CDS sequences as our database. After parsing the blast results with an e-value cutoff of e-20 or less, we were able to build a new query set based on these results and the initial sequence. Table 2.1 lists the number of candidate genes at this step. Thus, we ran a second round of tBLASTx, using the new query for each of the CDS databases, resulting in the final sequence files for all combinations of enzymes and species (Table 2.2). This iterative approach was used to ensure that we did not inadvertently miss any gene copies if our starting point was too genetically diverse. However, this approach can result in false positives particularly for enzymes that have not been fully annotated and have high similarity with other enzymes in this pathway. To overcome this issue, we manually verified annotations of these candidates if they were available, and in the cases that no annotation was found, we only included sequences in the analysis that we were able to blast to an annotated gene in one of the nine species with confidence. This step was performed manually. Table 2.3 lists the final number of gene candidates used to build the trees, and Table 2.4 lists their accession numbers.

Subsequently, we performed multiple sequence alignment (MSA) of these sequences using the iterative refinement method incorporating both the WSP and

consistency scores of MAFFT 7-055 (Katoh et al. 2002), which includes two novel techniques: 1. Homologous regions are rapidly identified by the fast Fourier transform (FFT) 2. Has a simplified scoring system that reduces CPU time and increases the accuracy of alignments even for sequences having large insertions or extensions as well as distantly related sequences of similar length. All the pairwise alignments were computed by iterative refinement method incorporating local pairwise alignment (mafft-linsi –localpair) under opening gap of 4 and extension gap of 0.05 and 1000 cycles of iterative refinement. The alignments were manually edited in Genedoc (Nicholas et al 1997). As the focus of this study is aimed at investigating the evolutionary history of these gene families rather than the exact time at which evolutionary events occurred, it was decided that keeping the gaps would help to retain the information on the recurrent indels. Hence, if any indel is observed for more than one species that gap is allowed.

The EMBOSS package (Rice et al. 2000) was used to modify file formats to nexus format for subsequent tree building using Mrbayes (Huelsenbeck and Ronquist 2001). Phylogenetic analysis and the creation of trees was done using an MPI version of Mrbayes 3.2.2. MrBayes uses Bayes's theorem to estimate the posterior probability of a phylogenetic tree and the Markov chain Monte Carlo (MCMC) method to approximate the posterior probability of a phylogenetic tree.

The evolutionary model was set to GTR with gamma-distributed rate variation across sites and proportion of invariable sites. Our original analysis began with 10000 samples from the posterior probability distribution, with the sampling frequency of every 100th generation, and continued sampling until the standard deviation of split frequencies was below 0.01. For the final sets of analysis, sampling frequency was set to 500, and a

diagnose frequency of 1000, as well as a print frequency of 1000 was used. . The default number of chains was used, and the temperature for it was set to 0.005. The number of generations varied based on the dataset. The stopping point was decided after the result was examined to make sure that the value of standard deviation of split frequency reached a value closer to zero and the Effective Sample Size (ESS) value is above 100. Finally, the log likelihood values were also examined to ensure that an adequate number of generations were produced. When it no longer increases and starts to randomly fluctuate, the run may have reached stationarity. Relburnin was set to yes, meaning that when calculating the convergence diagnostic a proportion of the sampled values was discarded as burnin. BurninFrac determines this fraction and it was set to 0.25, meaning that 25 percent of the samples were discarded. This burnin value was applied to both sump and sumt. After running sump, the program with output a table with summaries of the samples of the substitution model parameters, as well as the mean, mode and 95% credibility interval of each model parameter. This is where the parameters such as ESS were checked. Part of sumt command output includes summary statistics for the taxon bipartitions, a tree with clade credibility (posterior probability) values, and a phylogram (if branch lengths have been saved). When there is overwhelming support for a single tree all partitions will have a posterior probability of 1.0. The trees were also printed to a file that we read and visualized using FigTree (http://tree.bio.ed.ac.uk/software/figtree/). Mrbayes also generates another file that contains the list of taxon bipartitions, their posterior probability, and the branch lengths associated with them.

Examining the stationarity of LnL values, and discarding burn-in has been a common convergence test of choice. However, lack of an actual method to test

convergence can potentially lead to inadequate results. Evaluating the standard deviation of split frequencies for independent runs and their posterior properties, is another strategy to ensure that the convergence was reached. The problem arises when each of the two independent runs have very different topologies and each topology has high probability itself. As it happened for our datasets, it is going to take millions of generations to move between regions of tree space, hence there is a good possibility that they will independently find the high probability topology first, which will result in false signs of convergence based on their standard deviation of split frequency and their posterior probability similarities. There is no good way of determining the necessary number of priori generations. A common approach to this problem among researchers is to run the analysis for a large number of generations. If the runs fail to converge, it will result in affectedly inflated posterior probabilities. To overcome this issue we used two different programs to explore the convergence of runs, AWTY (Are We There Yet) (Nylander et al. 2008), and Tracer v1.6 (Rambaut et al. 2014).

Tracer is a program widely used to explore BEAST (Drummond et al. 2012) output, and has also been used extensively to evaluate the output of Mrbayes. Tracer visualizes different molecular evolutionary parameters in addition to calculating various statistics such as the mean and median. Tracer enables us to look at LnL plots over time. TRACER helps us determine if the parameters of our models have converged. However, convergence of parameters does not always state the convergence of topologies. If our data has a complex likelihood surface, this can mislead to stopping the program before the tree space has been adequately explored. AWTY is the only general diagnostic tool that rather than assessing convergence of molecular evolutionary parameters, evaluates

the convergence of topologies on the same space. A number of properties of primary interest in a Bayesian phylogenetic analysis can be visualized by using this online application. These properties include convergence rates of posterior split probabilities and branch lengths. Tracer gives a rough idea of how the analyses are behaving. It also gives a different, and more rigid ESS and burn-in value. If the ESS values shown in Tracer are below 100, the analysis needs additional runtime. AWTY on the other hand, enables us to compare posterior probabilities for independent runs, while Tracer looks at the effective sample size of parameters of interest. We used the slide, compare, and cumulative commands of AWTY. Slide commands allows us to test the subsamples of our chains and how proportional their posterior probabilities are to trees being sampled.

Initially, we performed other tests of topology, including Kishino-Hasegawa (KH), Shimodaira-Hasegawa (SH), expected likelihood weights (ELW), and obtained site-log-likelihoods for the trees, using TREE-PUZZLE (Schmidt et al. 2002). Afterwards, we used CONSEL (Shimodaira and Hasegawa 2001) to perform the approximately unbiased (AU) test (Shimodaira 2002). CONSEL uses site-log-likelihood values from TREE-PUZZLE to assess the trees. All of the tests within these programs are based on non-parametric bootstrap method, RELL (Kishino et al. 1990, Hasegawa and Kishino 1994). This means that re-estimation of parameters or branch lengths were not performed. All of the test results agreed on one topology, but the test values proved to be biased towards our data. Since all of the mentioned tests have been designed for maximum likelihood methods rather than the MCMC methods, we decided to exclude the results of these tests form this study and evaluate our trees based on Tracer and AWTY results only.

To reconstruct the species trees, we used the software *BEAST, a part of the BEAST package (Bayesian Inference of Species Trees from Multi-locus Data; Heled and Drummond, 2010). It is a method implemented in coalescent-based Bayesian species tree inference, an extension of BEAST v. 1.8.1 and simultaneously co-estimates species trees and gene trees. The BEAUti utility included in the software package was used to properly format and input files using the output trees of Mrbayes. GTR substitution model, gamma as site heterogeneity model, speciation Yule process was set as the tree prior, lognormal relaxed clock were used. We performed a run with 10,000,000 generations, sampling every 10000 generations. BEAUti outputs an XML file that we used to run BEAST. We used the program Tracer v1.6 (Rambaut and Drummond, 2007) to check for convergence. The results were obtained and summarized in TreeAnnotator v1.8.1 (Drummond and Rambaut, 2007; http://beast.bio.edu.ac.uk) and visualized in FigTree 1.4.0 (Rambaut, 2008). TreeAnnotator will take a set of trees and find the best supported tree as well as calculating the posterior clade probabilities. The burnin value of TreeAnnotator was calculated based on the TRACER burnin for *BEAST trees.

## 2.3 Results And Discussion

Here we discuss the putative gene candidates of each trait and their paralogues and orthologous relationships, as well as the underlying evolutionary events that causes different topologies in their respected inferred species trees, and improve our description of these clade orders. Following the detailed evaluation of all trees, we discuss the divergence test performed.

### 2.3.1 Aureusidin Synthase-Phylogenetic Tree

Aureusidin synthase is homologous to polyphenol oxidase and is missing the sequence for the conserved N-terminal region of polyphenol oxidase. Aureusidin synthase is responsible for flower coloration in some plants but has never been described in legumes. Using our method, we were able to identify candidate genes encoding for seven of the species of this study. To place a gene in this family and not the overall polyphenol oxidase family, it must have been missing the N-terminal region. We were not able to observe any significant sequence similarity for aureusidin synthase encoding genes in either *Arabidopsis thaliana* or *Cicer arietinum*. Figure 2.2 is the constructed gene tree for this gene family. In the absence of *Arabidopsis thaliana* sequences, we rooted the tree by *Chamaecrista fasciculata.*

The members of this gene family in *Glycine max* are overrepresented more than in the rest of the species, mostly by four to five fold. This suggests a post divergent gene expansion in *Glycine max.* Tandem duplication is observed on chromosome 7, 13 and to a lesser extent on chromosome 15 of *Glycine max.* Some of the tandem duplications have other genes inserted between them along the chromosome. The clades also support a duplication event and gene expansion sometime before the divergence of *Chamaecrista fasciculate* from other legumes. This is supported by the small two gene clade near the Medicago genes. Encoding genes of this enzyme in *Medicago truncatula* all appear on one clade, suggesting potential gene loss. *Pisum sativum* only has one gene for this family which we speculate is due to missing information as a result of using EST data.

The species tree was constructed based on the gene tree, using *BEAST (figure 2.3). This tree has a few discrepancies with the phylogenetic tree of legumes constructed by

Cannon et al in 2010 (figure 1). *Phaseolus vulgaris* is grouped with *Pisum sativum* and *Lotus japonicus* while we expected to see it grouped with *Medicago truncatula.* This could be the result of gene loss in *Medicago truncatula*.

### 2.3.2 Chalcone Isomerase-Phylogenetic Tree

Chalcone isomerase is derived from fatty-acid binding proteins. Ngaki et al. (2012) showed that the sequence identity of the characteristic chalcone isomerase domain can vary between 10 and 63 percent. Their study verifies that three of eight *Arabidopsis thaliana* chalcone isomerase encoding genes additionally encode amino-terminal chloroplast-transit sequences. This creates a challenge in determining encoding genes of chalcone isomerase in all of these species. It is possible in each of our species that some genes encoding chalcone isomerase may also encode the amino-terminal chloroplast-transit sequence. Some of the gene candidates for chalcone isomerase in some species have not been annotated. In these situations, if the sequence had high similarity to both motifs in a closely related species, we decided to pick it as a gene candidate.

The gene tree generated by Mrbayes is shown in figure 2.4 and is rooted by *Arabidopsis thaliana.* A massive gene expansion within a single species is supported by our clades. Duplication in this gene family happened prior and after speciation and was followed with more gene expansion and retention. Some of this clade expansion may be explained by tandem duplication. Tandem duplication examples include: Medtr1g115850.1 and Medtr1g115840.1, Medtr1g115870.1, Medtr1g115880.1 and Medtr1G115890.1, chr5.CM0180.670, chr5.CM0180.660 and chr5.CM0180.680, Ca_18470 and Ca_07383, C.cajan_4212582 and C.cajan_42126506. The number of

genes are overrepresented in some species like *Cajanus cajan* when compared to *Lotus japonicus* or *Glycine max.*

The species tree was obtained by *BEAST using this gene tree (figure 2.5). The phylogenetic tree of legume from figure one has major discrepancies with our species tree. This suggests that the time of divergent and the rate of evolution for this gene family is very different from other genes. In addition, when we compare the species tree to our gene tree, it is difficult to determine how *BEAST developed the species tree as clades are grouped together in an odd manner. This gene family is an excellent example of why single nuclear gene trees are not well suited for constructing species trees.

### 2.3.3 Chalcone Synthase-Phylogenetic Tree

Figure 2.6 represents the gene family of chalcone synthase and their orthologous relationships. The tree was rooted by *Arabidopsis thaliana.* The interspersed distribution of members of chalcone synthase family over different clades suggests that gene expansion happened before the divergence of these species. This pattern supports gene expansion prior to the divergence of the Papilionoideae. Surprisingly, *Medicago truncatula* and *Pisum sativum* fall as an independent clades more closely related to Chamaecrista than other Hologalegina. It appears that duplication in these two species has happened after speciation. Copy-number variations can be observed particularly in *Phaseolus vulgaris, Lotus japonicus* and *Glycine max,* result of segmental duplications/low copy repeats. Tandem duplication is also observed between chr4.CM44.260 and chr4.CM44.110 in Lotus, and Phvul.011G039700.1 and Phvul.011G026000.1 in Phaseolus.

The gene tree was used to generate species tree by *BEAST. We compared our species tree with the one shown in figure 2.1. These two topologies are not compatible. *Medicago truncatula* branch verifies what we suspected on the gene tree, and strongly points to the early divergence of this gene family in *Medicago truncatula.* However, appearance of *Chamaecrista fasciculata* after *Medicago truncatula* is most likely due to use of ESTS hence missing information on gene relationships. *Cajanus cajan, Phaseolus vulgaris,* and *Glycine max* form a similar clade as the legume phylogeny tree.

### 2.3.4 Dihydroflavonol 4-Reductase –Phylogenetic Tree

The orthologous relationships between the members of this gene family are shown in figure 2.8. The tree was rooted by *Arabidopsis thaliana.* In general, our clades show evidence for massive gene expansion prior to speciation. Duplication events happened following the speciation, and we observed tandem duplications in several species. Good examples includes: Medtr3g005170.1 and Medtr3g005210.1, Medtr8g062440.1 and Medtr8g062110.1, Medtr7g074820.1 and Medtr7g074850.1, Phvul.L005800.1 and Phvul.L005700.1, Phvul.011G022200.1 and Phvul.011G022100.1, c.cajan_32523 and c.cajan_32524, chr5.CM77.140 and chr5.CM77.150.

Some clades also support either gene retention or expansion happened post-divergence. Medicago *truncatula* has formed a polytomy. *Pisum sativum* is only observed in early parts of the tree. This could be the result of using ESTs and lack of full information of this gene family in *Pisum sativum.* However, these genes might have diverged a lot earlier in *Pisum sativum* than the rest of Papilionoideae species and this has

resulted in them forming early clusters on the tree, and other events have happened after the initial divergent.

This tree was used to infer a species tree using *BEAST (figure 2.9). The phylogenetic tree of legumes presented in figure 2.1 was used to compare the suggested phylogenetic relationships between these species in regard to this enzyme and the general consensus. *Lotus japonicus* and *Cajanus cajan* formed and peculiar clade with each other, while *Pisum sativum* is the closest species to *Chamaecrista fasciculata.* The later occurrence can be explained by this gene family evolving faster and diverging earlier in *Pisum sativum* than others.

### 2.3.5 Anthocyanidin 3-O-Glucosyltransferase-Phylogenetic Tree

3-o-glucosyltransferase family comprises many members, many of them have not been annotated further than family level. It is very hard to confidently verify that all the sequences used to build the tree belong to this particular gene family, and not other subfamilies. Nonetheless, the results presented here are the best possible outcome based on our current knowledge of 3-o-glucosyltransferase and its subfamilies. Figure 2.10 shows the orthologous relationship between the members of anthocyanidin 3-o-glucosyltransferase gene family and was rooted by *Arabidopsis thaliana.* The occurrence of *Arabidopsis thaliana* in multiple clades with multiple gene copies, suggests the presence of different members of this gene family due to mutation, in the common ancestor. Furthermore, given the size of this gene family, it is likely that we are representing more than just 3-o-glucosyltransferase as suggested above. Many of the

genes belonging to *Pisum sativum* and *Chamaecrista fasciculata* formed long polytomies, suggesting a very rapid speciation for these two species.

For the most part, our clades strongly support gene expansion within single species. But in some cases we can clearly see evidence for duplication events prior to species divergence followed by further expansion, as you would expect for a large family. An excellent example of this is between *Medicago truncatula*, *Cicer arietinum, Cajanus cajan,* and *Phaseolus vulgaris.* There is also evidence of tandem duplications in several species. A few of the *Cicer arietinum* and *Glycine max* paralogs show more similarity to each other in comparison to their orthologous genes, signifying their slower rate of evolution. The number of genes for each species is not what we expected. *Glycine max* has a smaller family comparing to *Medicago truncatula*. I am unsure if this is due to false positives, or gene retention in *Glycine max*, or expansion in *Medicago truncatula* and other species that are showing significantly higher number of genes.

Although it is likely the gene tree is of a gene family, I produced a species tree (Figure 2.11) from the gene tree using *BEAST. The phylogeny of legumes presented in figure 2.1 is used as the frame of reference. These two trees do not represent similar topologies, which is not surprising. We could speculate that positioning of *Medicago truncatula*, *Pisum sativum*, and *Chamaecrista fasciculata* is the result of their rapid speciation and long branch attraction. Long branch attraction happens when we have rapidly evolving lineages leading to phylogenetic algorithms interpreting the relationship between them as synapomorphy instead of homoplasy (convergent evolution).

### 2.3.6 Anthocyanidin Reductase-Phylogenetic Tree

The gene tree constructed by Mrbayes for members of anthocyanidin synthase gene family is illustrated in figure 2.12. The tree was rooted by *Arabidopsis thaliana*, and the presence of its genes on different clades suggests multiple copies of ancestral gene. *Chamaecrista fasciculata* three genes form one clade, which could suggest Long Branch attraction due to their rate of evolution, conversely, we suspect that in this gene family this observation is a result of missing information due to the use of ESTs. Our clades show evidence of duplication in legumes, with tandem duplication observed in: chr1.CM05.700 and chr1.CM05.730, Phvul.006G111600.1 and Phvul.006G111500.1, c.cajan_10881 and c.cajan_10882, and Glyma08g06640.1 and Glyma08g06639.1. *Pisum sativum* only has one gene, which strongly points to some missing evidence because we have only had access to EST data.

Species tree inferred from this gene tree, generated by *BEAST, (Figure 2.13) is incongruent with species tree presented in figure 2.1. This tree is the phylogeny of legumes published by Cannon S. et al in 2010. *Cicer arietinum* in on the clade with *Glycine max* and *Phaseolus vulgaris*, where we expected to see *Medicago truncatula*. *Pisum sativum* is the closest species to the outgroup species, and we suspect that it is misplaced since the information on it was incomplete. Overall, this topology suggests a different relationship between these species.

### 2.3.7 Flavanone 3-Hydroxylase-Phylogenetic Analysis

Figure 2.14 shows the orthologous relationship between the members of flavanone 3-hydroxylase gene family. This tree was generated using Mrbayes program, and was

rooted by *Arabidopsis thaliana.* The occurrence of two different clades for *Arabidopsis thaliana* with multiple gene copies, suggests the presence of different members of this gene family, due to mutation, in the common ancestor. On the other hand, we do not observe multiple clades for *Chamaecrista fasciculata*, which could suggest gene loss events. However, this absence can be due to the use of ESTs to build the gene family datasets for *Chamaecrista fasciculata*, which can lead to missing data for the gene relationships.

We can clearly see evidence of duplications events prior to the speciation, followed by gene expansion in some species. Our clades also, support gene expansion within each species for *Glycine max* and *Medicago truncatula*, as well as a separate duplication after the speciation of *Glycine max.* Few of the *Medicago truncatula* and *Glycine max* paralogs show more similarity to each other rather than their orthologues in other species, which can suggest a lower rate of evolution for them. Tandem duplication occurred between Glyma02g05470.1 and Glyma02g05450.1, Medtr5g032900.1 and Medtr5g032880.1, and Medtr8g075890.1 and Medtr8g075830.1. The absence of *Pisum sativum* genes in the Papilionoideae clade could be due to the use of EST data, or result of an evolutionary event such as gene lost, but no certain conclusion can be made.

Figure 2.15 is the species tree inferred from this gene tree, generated by *BEAST. The species tree is in incongruence with the known species tree of legumes presented in figure 2.1. *Pisum sativum* appears closer to *Chamaecrista fasciculata*, which could be explained by branch length attraction. If *Chamaecrista fasciculata* and *Pisum sativum* have slower evolution rates in comparison to other species, they can appear close to each other regardless of their true relationship. *Medicago truncatula* appears where we expect

to see *Cajanus cajan* have switched places on their clades and vice versa. *Lotus japonicus* position can be explained by the absence of *Pisum sativum* on this clade.

### 2.3.8 Leucoanthocyanidin Reductase-Phylogenetic Tree

Figure 2.16 represents the gene tree of leucoanthocyanidin reductase obtained from Mrbayes. The tree was rooted with one of the gene copies of *Arabidopsis thaliana.* The clades on this tree strongly suggest that speciation occurred prior to the duplication events. *Glycine max, Phaseolus vulgaris,* and *Medicago truncatula* went under duplication event post speciation. The whole genome duplication of *Glycine max* cannot be inferred from this tree, pointing to a possible gene lost in this family.

The species tree was generated with *BEAST using the output of Mrbayes (figure 2.17). Topology of this tree does not agree with the tree represented in figure 2.1. *Pisum sativum* appears to be the closest to ancestor. *Chamaecrista fasciculata* is positioned next to *Cicer arietinum.* The missing information in the EST data of *Pisum sativum* and *Chamaecrista fasciculata* in addition to possible gene loss in this family can help us understand the reason behind the differences between these two topologies.

### 2.3.9 Anthocyanidin Synthase-Phylogenetic Tree

Figure 2.18 is the output tree of Mrbayes for anthocyanidin synthase and was rooted by *Arabidopsis thaliana. Chamaecrista fasciculata* appears on two separate clusters, suggesting that the common ancestor of legumes had more than one copy of this gene. The clades show evidence of a duplication event prior to the divergent of species

followed by expansion within *Glycine max, Pisum sativum, Medicago truncatula,* and *Chamaecrista fasciculata.* Tandem duplication can be observed between Medtr8g074070.1 and Medtr8g074050.1, and Phvul.009G003500.1 and Phvul.009G003400.1. Absence of *Pisum sativum* from most of the clades points to either a major gene loss or missing information due to the use of ESTs. Another possible explanation is they are evolving much faster than their orthologs in other species, resulting in Long Branch Attraction. Furthermore, a couple of *Glycine max* and *Medicago truncatula* paralogs show more similarity to each other in comparison to their orthologous genes, signifying their slower rate of evolution.

The species tree generated by *Beast using the outcome of Mrbayes is presented in figure 2.19. The topology of this tree is compatible with the topology of the tree in figure 2.1.

### 2.3.10 Isoflavone Synthase-Phylogenetic Tree

Gene tree of isoflavone synthase (figure 2.20) was rooted using *Arabidopsis thaliana*. The outgroup cluster points to presence of only one copy of this gene in the ancestral genome. Our clades strongly support a massive gene expansion prior to the divergent of species which was followed by duplication events in Papilionoideae. Based on previous studies on duplication events in legumes, we suspect gene deletion has happened in *Glycine max* and *Medicago truncatula.*

*BEAST was used to generate the species tree (figure 2.21). *Phaseolus vulgaris* and *Medicago truncatula* appear closer to the root that what we expected to see based on

figure 2.1 and *Lotus japonicus* appears on the sister clade of *Pisum sativum* where we expected to see *Medicago truncatula.*

### 2.3.11 Flavonol Synthase-Phylogenetic Tree

Figure 2.22 shows the orthologous relationship between the members of flavonol synthase gene family. This tree was generated using Mrbayes program, and was rooted by *Arabidopsis thaliana.* Appearance of outgroup genes on one clade is interpreted as one copy of this gene in the ancestral genome. We can clearly see evidence of duplication events prior to the speciation, followed by gene expansion in some species. Medtr5g059130.1 and Medtr5g059140.1 point to tandem duplication.

Figure 2.23 is the species tree inferred from this gene tree, generated by *BEAST. The species tree is incongruence with the known species tree of legumes presented in figure 2.1. *Pisum sativum* appears closer to *Chamaecrista fasciculata*. *Phaseolus vulgaris* is closer to *Lotus japonicus* rather than *Glycine max*. Both of these topological differences can be explained by Long Branch Attraction, signifying higher rate of evolution for them comparing to other species; thus they can appear close to each other regardless of their true relationship. Another explanation for the placement of *Lotus japonicus* and *Phaseolus vulgaris* is deleterious mutation inferred from the absence of *Lotus japonicus* gene copies from second cluster of the gene tree. Use of EST data, hence missing information for the gene relationships could also be a likely cause of this topology.

### 2.3.12 Flavone Synthase-Phylogenetic Tree

The orthologous relationships between members of this gene family is illustrated in figure 2.24 and was rooted by *Arabidopsis thaliana.* Flavone synthase has not been annotated in most of the species, hence our search for candidate genes resulted in a batch of homologous sequences to Cytochrome 9450. The tree presented here was built using sequences with strong homology to Flavone synthase, however the lack of annotation might have resulted in false positives.

Our clades strongly support gene expansion within a single species, but in some cases we can clearly see the evidence for duplication events prior to species divergence followed by further expansion. Multiple tandem duplication was detected including: chr6.CM37.530 and chr6.CM37.560, Glyma12g07200.1 and Glyma12g07190.1, Medtr5g072980.1 and Medtr5g072930.1, Medtr8g063260.1 and Medtr8g063280.1, Medtr6g008630.1 and Medtr6g008650.1, Medtr7g028020.1 and Medtr7g027960.1, Medtr7g012860.1 and Medtr7g012330.1, Phvul.007G257400.1 and Phvul.007G257300.1, Phvul.004G159300.1 and Phvul.004G159500.1, Phvul.006G054500.1 and Phvul.006G054600.1, Phvul.009G172000.1 and Phvul.009G172100.1, and Phvul.004G021300.1 and Phvul.004G021400.1. It is hard to rule out any deleterious events after duplication events or confirm segmental duplications.

Figure 2.25 is *BEAST tree for flavone Synthase, and except for the *Pisum sativum* clade has a similar topology to the tree presented in figure 2.1. This disagreement in topology is caused by the combination of missing information due to use of ESTs, and

the difference in evolutionary rate of *Pisum sativum* and other species resulting in branch length attraction.

### 2.3.13 Convergence Tests

To diagnose the convergence in our Bayesian analyses, we first looked at the substitution model parameters using TRACER as our starting guide. In general, we looked at calculated ESS values for the combined parameter files of Mrbayes, to ensure good mixing and convergence. We also examined the LnL plots. All the ESS values were above 200, indicating convergence. Here we present the Estimates and Trace plot for LnL values. The Estimates plot is a histogram of log-likelihood values. However, Trace which is a plot of the log-likelihood through time, is generally more informative that LnL histograms.

As our primary test of convergence, we used, slide, cumulative, and compare commands in AWTY. Slide command basically acts as a sliding window of posterior probabilities, if the posterior probabilities vary wildly and show a trend in their plot, it is strong evidence that the runs have not reached convergence. Cumulative plots are used to inspect the stationarity of two independent MCMC runs. It plots the split frequencies for a number of selected splits for each individual run. Similar to the output of slide command, if a trend is observed, especially towards the end of the run, the run has not reached stationarity. Finally, Compare command is used as a second test of convergence. Compare plot is a bivariate plot of split frequencies for two independent MCMC runs. The dots indicate posterior probability values for individual nodes obtained from two separate analyses. If a tight relationship to the diagonal is observed the runs have

converged. Correspondingly, the disagreement of the results of two Bayesian analysis about support for particular nodes, is related how much these points stray from the diagonal.

TRACER was used one more time for species trees constructed by *BEAST. Adequate effective sample sizes for each parameter (ESS >200), and unimodal distribution were used to determine whether the MCMC parameter samples were drawn from a stationary. The burnin value was calculated and used in TreeAnnotator.

Figures 2.26, 2.31, 2.36, 2.41, 2.46, 2.51, 2.56, 2.61, 2.66, 2.71, 2.76, and 2.81 are the Estimates and Trace plot of LnL for aureusidin synthase, chalcone isomerase, chalcone synthase, dihydroflavonol 4-reductase, anthocyanidin 3-o-glucosyltransferase, anthocyanidin reductase, flavanone 3-Hydroxylase, leucoanthocyanidin reductase, anthocyanidin synthase, isoflavone synthase, flavonol synthase, and flavone synthase, respectively. For the most part, Estimates plots are a perfect bell shape, and does not suggest that we need to run this chain longer. For leucoanthocyanidin reductase (figure 2.61), Trace plot, often referred to as hairy caterpillar, shows the sampled values against the step in the MCMC chain. Here little to no drastic fluctuation is observed, which points to sufficient number of runs.

Figures 2.27, 2.47, 2.62, and 2.72 show slide plots of aureusidin synthase, anthocyanidin 3-o-glucosyltransferase, leucoanthocyanidin reductase, and isoflavone synthase, respectively. These plots have no to little trend which corresponds to good convergence in these runs. Figures 2.32, 2.37, 2.42, 2.52, 2.57, 2.67, 2.77, and 2.82 are slide plots for chalcone isomerase, chalcone synthase, dihydroflavonol 4-reductase, anthocyanidin reductase, flavanone 3-Hydroxylase, anthocyanidin synthase, flavonol

synthase, and flavone synthase, respectively. Although they show some trend at the start of the run, as the run progresses, they become more constant, proving that these runs have reached convergence. We adjusted our burnin values for these trees, based on these plots.

Figures 2.28, 2.33, 2.48, 2.63, and 2.73 are cumulative plots of aureusidin synthase, chalcone isomerase, anthocyanidin 3-o-glucosyltransferase, leucoanthocyanidin reductase, and isoflavone synthase, respectively. The results do not vary too much, indicating that runs have converged. Figures 2.38, 2.43, 2.53, 2.58, 2.68, 2.78, and 2.83 are cumulative plots of chalcone synthase, dihydroflavonol 4-reductase, anthocyanidin reductase, flavanone 3-Hydroxylase, anthocyanidin synthase, flavonol synthase, and flavone synthase, respectively. Similar to their slide plots, they show some trends at the beginning of runs. Eventually they stop deviating from diagonal. Overall, both slide and cumulative point to convergence of the runs, given an adjusted burnin.

Figures 2.29, 2.34, 2.39, 2.44, 2.49, 2.54, 2.59, 2.64, 2.69, 2.74, 2.79, and 2.84 are compare plots of aureusidin synthase, chalcone isomerase, chalcone synthase, dihydroflavonol 4-reductase, anthocyanidin 3-o-glucosyltransferase, anthocyanidin reductase, flavanone 3-Hydroxylase, leucoanthocyanidin reductase, anthocyanidin synthase, isoflavone synthase, flavonol synthase, and flavone synthase, respectively. All of these plots show a relative tight fit of data to the diagonal, indicating the similarity of topologies and support values between two samples of each plot. Even in the case of some scattered dots, they are poorly supported. Figures 2.29, and 2.44, show some nodes in one sample that are absent from the other, but when the posteriors are small, we can still conclude topological concordance.

All of *BEAST trees had ESS values higher than 200 for all the parameters. A unimodal distribution of the likelihood values was reached as well (Figures 2.30, 2.35, 2.40, 2.45, 2.50, 2.55, 2.60, 2.65, 2.70, 2.75, 2.80, and 2.85).

## 2.4 Conclusion

When studying closely related species, the use of DNA sequences in contrast with protein sequences can be useful with detection of synonymous changes. We used a relatively stringent set of flavone synthase related genes for our similarity search and tree building criteria. Our intention was to identify putative orthologous genes based on homology that can be used to understand the underlying evolutionary events of this pathway. We might have introduced some false positive genes using BLAST approach, but we did not want to take the chance of missing more distantly related homologs. Further steps, as described in the method section, were taken to eliminate these false positive in full measure. Nevertheless, DNA sequence is redundant, and its short alphabet results in less significance in similarity studies in comparison to protein sequences. An alternative approach would be to align at the protein level and back align the sequences to DNA from protein level alignment.

The number of species that we have used allowed us to resolve gene family topologies and to detect basal branch points. The tree topologies of our study provides more information about the evolutionary relationship between these gene families and their orthologs in other species which led us to infer the state of these gene families in the last common ancestor of these species. For some of our trees we concluded that the genes have been diversified prior to some more recent duplication divergent events in legumes.

However, some of the enzymes show evidence of evolving after major divergent events in legumes, like in the case of LAR. Based on the literature and our results, we postulate the function that gave flavonoids advantage to evolve early during the evolution of the first land plants is their role as chemical messengers and internal physiological regulators, a good example is their roles associated with the growth hormone IAA. This gave them selective value during the early stages of evolution of the pathway. In forming variety of subgroups within one group of secondary compounds, flavonoids diversity in function may have been crucial. Their rapid and parallel evolution of their subgroups within different populations could be explained by severe environmental stresses. They can act as chemical signals to other organisms. A good example of this is the evolution of mycorrhizal and symbiotic nitrogen-fixing relationships. Finally, they are essential in chemical defense against microorganisms and herbivores and the 3-hydroxy anthocyanidins are the key to angiosperm pollinator attraction and seed dispersal.

Table 2.1: Initial number of accessions used for each Species-Enzyme combination

| | AS | CHI | CHS | DFR | 3GT | ANR | F3H | LAR | ANS | IFS | FLS | FNS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Glycine max | 18 | 13 | 55 | 19 | 51 | 13 | 158 | 21 | 224 | 250 | 262 | 10 |
| Phaseolus vulgaris | 3 | 12 | 38 | 13 | 33 | 10 | 83 | 27 | 298 | 155 | 129 | 267 |
| Lotus japonicus | 2 | 7 | 72 | 23 | 16 | 17 | 18 | 16 | 24 | 148 | 110 | 250 |
| Pisum sativum | 1 | 19 | 15 | 20 | 25 | 28 | 15 | 30 | 25 | 32 | 21 | 51 |
| Medicago truncatula | 4 | 11 | 46 | 50 | 60 | 40 | 110 | 27 | 107 | 150 | 120 | 265 |
| Cicer arietinum | 1 | 8 | 21 | 13 | 32 | 23 | 88 | 11 | 89 | 104 | 106 | 184 |
| Cajanus cajan | 5 | 17 | 27 | 26 | 27 | 33 | 103 | 15 | 136 | 147 | 153 | 275 |
| Chamaecrista fasciculata | 3 | 2 | 247 | 13 | 19 | 6 | 24 | 2 | 10 | 64 | 125 | 85 |
| Arabidopsis thaliana | 0 | 12 | 10 | 27 | 57 | 30 | 75 | 22 | 92 | 165 | 13 | 257 |

Table 2.2: Number of accessions used for each Species-Enzyme combination after the second blast run

| | AS | CHI | CHS | DFR | 3GT | ANR | *F3H* | *LAR* | *ANS* | *IFS* | *FLS* | *FNS* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Glycine max | 20 | 16 | 73 | 26 | 72 | 13 | 156 | 20 | 199 | 240 | 234 | 14 |
| Phaseolus vulgaris | 9 | 12 | 44 | 19 | 41 | 14 | 84 | 26 | 106 | 151 | 126 | 259 |
| Lotus japonicus | 4 | 7 | 40 | 11 | 27 | 11 | 31 | 14 | 37 | 69 | 42 | 106 |
| Pisum sativum | 1 | 19 | 24 | 34 | 28 | 25 | 20 | 29 | 26 | 30 | 22 | 53 |
| Medicago truncatula | 16 | 11 | 51 | 40 | 64 | 32 | 101 | 25 | 101 | 140 | 113 | 246 |
| Cicer arietinum | 2 | 8 | 27 | 13 | 34 | 22 | 77 | 11 | 77 | 97 | 91 | 169 |
| Cajanus cajan | 15 | 18 | 32 | 25 | 34 | 30 | 98 | 15 | 124 | 135 | 141 | 252 |
| Chamaecrista fasciculata | 11 | 4 | 17 | 12 | 11 | 11 | 23 | 2 | 16 | 21 | 18 | 21 |
| Arabidopsis thaliana | 0 | 12 | 10 | 26 | 69 | 30 | 73 | 24 | 87 | 168 | 13 | 253 |

Table 2.3: Final number of accessions used for each Species-Enzyme combination to build the tree

| | AS | CHI | CHS | DFR | 3GT | ANR | *F3H* | *LAR* | *ANS* | *IFS* | *FLS* | *FNS* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Glycine max | 20 | 10 | 18 | 10 | 30 | 6 | 16 | 2 | 7 | 5 | 5 | 10 |
| Phaseolus vulgaris | 3 | 11 | 15 | 13 | 28 | 5 | 3 | 2 | 7 | 3 | 5 | 61 |
| Lotus japonicus | 2 | 6 | 13 | 7 | 10 | 4 | 3 | 1 | 3 | 2 | 2 | 15 |
| Pisum sativum | 1 | 12 | 13 | 20 | 28 | 1 | 4 | 1 | 4 | 4 | 2 | 18 |
| Medicago truncatula | 4 | 11 | 22 | 28 | 59 | 3 | 9 | 2 | 8 | 2 | 4 | 42 |
| Cicer arietinum | 0 | 7 | 8 | 9 | 28 | 2 | 6 | 1 | 6 | 2 | 5 | 24 |
| Cajanus cajan | 5 | 17 | 12 | 10 | 23 | 5 | 4 | 1 | 6 | 2 | 6 | 41 |
| Chamaecrista fasciculata | 6 | 4 | 5 | 8 | 9 | 3 | 3 | 1 | 8 | 3 | 5 | 9 |
| Arabidopsis thaliana | 0 | 8 | 4 | 4 | 43 | 3 | 7 | 2 | 4 | 3 | 6 | 1 |

Table 2.4: Final accession numbers used for each Species-Enzyme combination to build the tree

| | GM | PV | LJ | PS | MT | CA | CC | CF | AT |
|---|---|---|---|---|---|---|---|---|---|
| AS | Glyma04g14361.1 | Phvul.008G073200.1 | chr6.CM0336.430 | p.sativum_wa1_33776 | Medtr2g013010.1 | | C.cajan_18419 | F0SHL7R02GMOXJ | |
| | Glyma13g25181.1 | Phvul.006G188500.1 | chr6.CM0336.440 | | Medtr2g013030.1 | | C.cajan_17009 | F0SHL7R02G211F | |
| | Glyma13g25260.1 | Phvul.002G225600.1 | | | Medtr4g036400.1 | | C.cajan_17010 | F0SHL7R02GAPOJ | |
| | Glyma13g31595.1 | | | | Medtr4g036500.1 | | C.cajan_33845 | F0SHL7R02FYVE8 | |
| | Glyma13g31590.1 | | | | | | C.cajan_33846 | F0SHL7R02FVM88 | |
| | Glyma13g25150.1 | | | | | | | F0SHL7R02FOWYD | |
| | Glyma13g31571.1 | | | | | | | | |
| | Glyma01g33692.1 | | | | | | | | |
| | Glyma15g07705.1 | | | | | | | | |
| | Glyma15g07700.1 | | | | | | | | |
| | Glyma15g07710.1 | | | | | | | | |
| | Glyma07g31270.1 | | | | | | | | |
| | Glyma07g31301.1 | | | | | | | | |
| | Glyma07g31254.1 | | | | | | | | |
| | Glyma07g31310.1 | | | | | | | | |
| | Glyma07g31280.1 | | | | | | | | |
| | Glyma07g31262.1 | | | | | | | | |
| | Glyma18g45900.1 | | | | | | | | |
| | Glyma18g45895.1 | | | | | | | | |
| CHI | Glyma06g14820.1 | Phvul.003G216600.1 | chr5.CM0034.270 | Pisum_sativum_Contig352 | Medtr5g022010.1 | Ca_18470 | C.cajan_23638212 | F0SHL7R02JH84S | AT2G26310.1 |
| | Glyma04g40030.1 | Phvul.003G262100.1 | chr4.CM0119.180 | Pisum_sativum_Contig3843 | Medtr8g075950.1 | Ca_07383 | C.cajan_13724 | F0SHL7R02JE8GE | AT1G53520.1 |
| | Glyma13g33730.1 | Phvul.009G143100.1 | chr5.CM0180.660 | Pisum_sativum_Contig6152 | Medtr1g115850.1 | Ca_09149 | C.cajan_22147 | F0SHL7R02HC5VE | AT3G51230.1 |
| | Glyma19g33940.1 | Phvul.005G064600.1 | chr5.CM0180.670 | Pisum_sativum_Contig7003 | Medtr1g115830.1 | Ca_18651 | C.cajan_33731259 | F0SHL7R02GPV3O | AT3G63170.1 |
| | Glyma15g39050.1 | Phvul.005G064500.1 | chr5.CM0180.680 | Pisum_sativum_Contig7848 | Medtr1g015700.1 | Ca_18652 | C.cajan_09652 | | AT3G55120.1 |
| | Glyma03g31100.2 | Phvul.007G008600.1 | chr5.CM0180.690 | Pisum_sativum_Contig975 | Medtr1g115860.1 | Ca_18653 | C.cajan_09653 | | AT5G05270.1 |
| | Glyma20g38580.1 | Phvul.007G008500.1 | | Pisum_sativum_Contig976 | Medtr1g115820.1 | Ca_09974 | C.cajan_09655 | | AT5G66230.1 |
| | Glyma20g38560.1 | Phvul.001G152000.1 | | Pisum_sativum_Contig2414 | Medtr1g115840.1 | | C.cajan_09656 | | AT5G66220.1 |
| | Glyma20g38570.1 | Phvul.001G037700.1 | | p.sativum_contig02422 | Medtr1g115890.1 | | C.cajan_09657 | | |
| | Glyma10g43850.1 | Phvul.002G276500.1 | | p.sativum_contig24152 | Medtr1g115870.1 | | C.cajan_4032622 | | |
| | | Phvul.002G108800.1 | | p.sativum_contig25625 | Medtr1g115880.1 | | C.cajan_48225 | | |
| | | | | p.sativum_contig28704 | | | C.cajan_29159 | | |
| | | | | | | | C.cajan_2916031 | | |
| | | | | | | | C.cajan_2916107 | | |
| | | | | | | | C.cajan_2916248 | | |
| | | | | | | | C.cajan_4212582 | | |
| | | | | | | | C.cajan_42126506 | | |

| CHS | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Glyma02g14450.1 | Phvul.009G131000.1 | chr2.CM18.730 | p.sativum_04194 | Medtr3g086260.1 | Ca_08294 | C.cajan_38726 | F0SHL7R02H3MWQ | AT4G00040.1 |
| Glyma12g02670.1 | Phvul.011G039700.1 | chr2.CM18.750 | p.sativum_04195 | Medtr3g083910.1 | Ca_22507 | C.cajan_21074 | F0SHL7R02HZ2NQ | AT4G34850.1 |
| Glyma11g10380.1 | Phvul.011G026000.1 | chr2.CM18.760 | p.sativum_04197 | Medtr3g083920.1 | Ca_11408 | C.cajan_19240 | F0SHL7R02HU5X2 | AT1G02050.1 |
| Glyma11g01350.1 | Phvul.008G141200.1 | chr4.CM44.260 | p.sativum_21028 | Medtr2g058470.1 | Ca_08546 | C.cajan_20347 | F0SHL7R02FIQCG | AT5G13930.1 |
| Glyma06g12470.1 | Phvul.004G168400.1 | chr4.CM46.110 | p.sativum_22326 | Medtr7g016720.1 | Ca_11119 | C.cajan_47742 | F0SHL7R02FZULH | |
| Glyma08g11620.1 | Phvul.001G067800.1 | chr6.CM57.460 | p.sativum_23038 | Medtr7g016700.1 | Ca_11820 | C.cajan_08958 | | |
| Glyma08g11520.1 | Phvul.001G083000.1 | chr1.CM284.240 | p.sativum_23066 | Medtr7g016800.1 | Ca_11821 | C.cajan_45286 | | |
| Glyma08g11635.1 | Phvul.002G184300.1 | chr1.CM284.250 | Pisum_sativum_1434 | Medtr7g016820.1 | Ca_11822 | C.cajan_24072 | | |
| Glyma08g11530.1 | Phvul.002G038700.1 | chr3.CM590.770 | Pisum_sativum_1435 | Medtr7g113410.1 | | C.cajan_23441 | | |
| Glyma08g11630.2 | Phvul.002G039300.1 | chr3.CM590.840 | Pisum_sativum_1436 | Medtr7g084300.1 | | C.cajan_23442 | | |
| Glyma08g11610.1 | Phvul.002G038900.1 | chr1.CM591.490 | Pisum_sativum_1437 | Medtr7g016780.1 | | C.cajan_03204 | | |
| Glyma13g09640.2 | Phvul.002G038600.1 | chr1.CM591.500 | Pisum_sativum_3145 | Medtr5g007770.1 | | C.cajan_43674 | | |
| | Phvul.002G039000.1 | chr1.CM591.520 | Pisum_sativum_983 | Medtr5g007760.1 | | | | |
| Glyma05g28610.1 | Phvul.002G039200.1 | | | Medtr5g007720.1 | | | | |
| Glyma01g43880.1 | Phvul.002G039100.1 | | | Medtr5g007730.1 | | | | |
| Glyma01g13900.1 | | | | Medtr5g007740.1 | | | | |
| Glyma01g22880.1 | | | | Medtr8g085200.1 | | | | |
| Glyma19g27930.1 | | | | Medtr1g097910.1 | | | | |
| Glyma09g08780.1 | | | | Medtr1g097900.1 | | | | |
| | | | | Medtr1g098140.1 | | | | |
| | | | | Medtr1g098150.1 | | | | |
| | | | | Medtr4g078730.1 | | | | |

D

F

R

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Glyma14g07940.1 | Phvul.L005800.1 | chr6.CM66.350 | Pisum_sativum_1543 | Medtr3g101500.1 | Ca_07958 | C.cajan_18359 | F0SHL7R02GO06F | AT2G45400.1 |
| Glyma02g18380.1 | Phvul.L005700.1 | chr5.CM77.110 | Pisum_sativum_1544 | Medtr3g043600.1 | Ca_09533 | C.cajan_39978 | F0SHL7R02GEU4Q | AT4G35420.1 |
| Glyma12g02250.1 | Phvul.005G034700.1 | chr5.CM77.120 | Pisum_sativum_3012 | Medtr3g031650.1 | Ca_09534 | C.cajan_21020 | F0SHL7R02GHEU6 | AT5G58490.1 |
| Glyma12g02230.3 | Phvul.005G034800.1 | chr5.CM77.140 | Pisum_sativum_3423 | Medtr3g005170.1 | Ca_09535 | C.cajan_19849 | F0SHL7R02JC316 | AT5G42800.1 |
| Glyma04g04020.1 | Phvul.005G034900.1 | chr5.CM77.150 | Pisum_sativum_3756 | Medtr3g005210.1 | Ca_10786 | C.cajan_28332 | F0SHL7R02FRYV5 | |
| Glyma17g37060.1 | Phvul.005G034600.1 | chr5.CM77.210 | Pisum_sativum_6418 | Medtr2g013230.1 | Ca_10787 | C.cajan_28333 | F0SHL7R02I29HE | |
| Glyma01g20030.4 | Phvul.005G034500.1 | chr1.CM55.100 | Pisum_sativum_6433 | Medtr2g013250.1 | Ca_01535 | C.cajan_24381 | F0SHL7R02G065Z | |
| Glyma09g40590.1 | Phvul.011G022200.1 | | Pisum_sativum_7077 | Medtr2g028620.1 | Ca_11190 | C.cajan_32523 | F0SHL7R02F7PZL | |
| Glyma09g40580.1 | Phvul.011G022100.1 | | p.sativum_04688 | Medtr2g101330.1 | Ca_19534 | C.cajan_32524 | | |
| Glyma18g45250.1 | Phvul.011G022400.1 | | p.sativum_1195 | Medtr2g013260.1 | | C.cajan_05379 | | |
| | Phvul.007G246200.1 | | p.sativum_12111 | Medtr7g074870.1 | | | | |
| | Phvul.001G012700.1 | | p.sativum_12112 | Medtr7g074820.1 | | | | |
| | Phvul.001G012800.1 | | p.sativum_12559 | Medtr7g074730.1 | | | | |
| | | | p.sativum_12560 | Medtr7g074880.1 | | | | |
| | | | p.sativum_12923 | Medtr7g074710.1 | | | | |
| | | | p.sativum_22011 | Medtr7g074850.1 | | | | |
| | | | p.sativum_23137 | Medtr5g072620.1 | | | | |
| | | | p.sativum_23418 | Medtr5g056600.1 | | | | |
| | | | p.sativum_23443 | Medtr5g029990.1 | | | | |
| | | | p.sativum_24968 | Medtr8g062440.1 | | | | |
| | | | | Medtr8g062110.1 | | | | |
| | | | | Medtr4g009040.1 | | | | |
| | | | | Medtr4g077130.1 | | | | |
| | | | | Medtr4g009690.1 | | | | |
| | | | | Medtr4g023730.1 | | | | |
| | | | | Medtr4g077100.1 | | | | |
| | | | | Medtr3g101500.1 | | | | |
| | | | | Medtr3g043600.1 | | | | |
| | | | | Medtr3g031650.1 | | | | |
| | | | | Medtr3g005170.1 | | | | |
| | | | | Medtr3g005210.1 | | | | |
| | | | | Medtr2g013230.1 | | | | |
| | | | | Medtr2g013250.1 | | | | |
| | | | | Medtr2g028620.1 | | | | |
| | | | | Medtr2g101330.1 | | | | |
| | | | | Medtr2g013260.1 | | | | |
| | | | | Medtr7g074870.1 | | | | |
| | | | | Medtr7g074820.1 | | | | |
| | | | | Medtr7g074730.1 | | | | |
| | | | | Medtr7g074880.1 | | | | |
| | | | | Medtr7g074710.1 | | | | |
| | | | | Medtr7g074850.1 | | | | |
| | | | | Medtr5g072620.1 | | | | |
| | | | | Medtr5g056600.1 | | | | |
| | | | | Medtr5g029990.1 | | | | |
| | | | | Medtr8g062440.1 | | | | |
| | | | | Medtr8g062110.1 | | | | |
| | | | | Medtr4g009040.1 | | | | |
| | | | | Medtr4g077130.1 | | | | |
| | | | | Medtr4g009690.1 | | | | |
| | | | | Medtr4g023730.1 | | | | |
| | | | | Medtr4g077100.1 | | | | |

| 3G | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| T | | | | | | | | | |
| Glyma14g37770 | Phvul.010G071900 | chr6.CM66.350 | Pisum_sativum_1543 | Medtr3g101500.1 | Ca_07958 | C.cajan_18359 | F0SHL7R02GO06F | AT2G29730 |
| Glyma14g37741 | Phvul.009G218800 | chr5.CM77.110 | Pisum_sativum_1544 | Medtr3g043600.1 | Ca_09533 | C.cajan_39978 | F0SHL7R02GEU4Q | AT2G22590 |
| Glyma14g37170 | Phvul.005G093600 | chr5.CM77.120 | Pisum_sativum_3012 | Medtr3g031650.1 | Ca_09534 | C.cajan_21020 | F0SHL7R02GHEU6 | AT2G22930 |
| Glyma02g47990 | Phvul.005G093300 | chr5.CM77.140 | Pisum_sativum_3423 | Medtr3g005170.1 | Ca_09535 | C.cajan_19849 | F0SHL7R02JC316 | AT2G30140 |
| Glyma02g39700 | Phvul.005G093500 | chr5.CM77.150 | Pisum_sativum_3756 | Medtr3g005210.1 | Ca_10786 | C.cajan_28332 | F0SHL7R02FRYY5 | AT2G30150 |
| Glyma02g39090 | Phvul.005G149900 | chr5.CM77.210 | Pisum_sativum_6418 | Medtr2g013230.1 | Ca_10787 | C.cajan_28333 | F0SHL7R02I29HE | AT2G29750 |
| Glyma02g39680 | Phvul.011G138500 | chr1.CM55.100 | Pisum_sativum_6433 | Medtr2g013250.1 | Ca_01535 | C.cajan_24381 | F0SHL7R02G065Z | AT2G29740 |
| Glyma12g28270 | Phvul.011G138600 | | Pisum_sativum_7077 | Medtr2g028620.1 | Ca_11190 | C.cajan_32523 | F0SHL7R02F7PZL | AT2G18570 |
| Glyma12g34030 | Phvul.011G136400 | | p.sativum_04688 | Medtr2g101330.1 | Ca_19534 | C.cajan_32524 | | AT2G15480 |
| Glyma12g34045 | Phvul.011G136700 | | p.sativum_1195 | Medtr2g013260.1 | | C.cajan_05379 | | AT2G29710 |
| Glyma12g14050 | Phvul.008G290200 | | p.sativum_12111 | Medtr7g074870.1 | | | | AT2G18560 |
| Glyma12g15870 | Phvul.008G262000 | | p.sativum_12112 | Medtr7g074820.1 | | | | AT2G15490 |
| Glyma11g05680 | Phvul.008G291000 | | p.sativum_12559 | Medtr7g074730.1 | | | | AT4G09500 |
| Glyma06g36520 | Phvul.008G090600 | | p.sativum_12560 | Medtr7g074880.1 | | | | AT4G34135 |
| Glyma06g36530 | Phvul.008G091000 | | p.sativum_12923 | Medtr7g074710.1 | | | | AT4G27560 |
| Glyma06g22820 | Phvul.008G290300 | | p.sativum_22011 | Medtr7g074850.1 | | | | AT4G34138 |
| Glyma06g43880 | Phvul.008G090700 | | p.sativum_23137 | Medtr5g072620.1 | | | | AT4G34131 |
| Glyma08g19290 | Phvul.008G262100 | | p.sativum_23418 | Medtr5g056600.1 | | | | AT4G27570 |
| Glyma08g44550 | Phvul.008G090500 | | p.sativum_23443 | Medtr5g029990.1 | | | | AT1G50580 |
| Glyma13g36500 | Phvul.008G090800 | | p.sativum_24968 | Medtr8g062440.1 | | | | AT1G07240 |
| Glyma13g36490 | Phvul.007G020600 | | | Medtr8g062110.1 | | | | AT1G30530 |
| Glyma13g36520 | Phvul.007G020800 | | | Medtr4g009040.1 | | | | AT1G07250 |
| Glyma13g32770 | Phvul.007G020700 | | | Medtr4g077130.1 | | | | AT1G64910 |
| Glyma05g31500 | Phvul.006G208500 | | | Medtr4g009690.1 | | | | AT1G64920 |
| Glyma15g03670 | Phvul.006G208300 | | | Medtr4g023730.1 | | | | AT1G73880 |
| Glyma03g03850 | Phvul.002G016400 | | | Medtr4g077100.1 | | | | AT1G07260 |
| Glyma03g03870 | Phvul.002G016500 | | | AC229724_1047.1 | | | | AT3G50740 |
| Glyma03g03830 | Phvul.002G096500 | | | AC235668_2.1 | | | | AT3G21750 |
| Glyma09g09910 | | | | | | | | AT3G29630 |
| Glyma16g03715 | | | | | | | | AT3G21760 |
| | | | | | | | | AT3G21800 |
| | | | | | | | | AT3G21790 |
| | | | | | | | | AT3G21780 |
| | | | | | | | | AT5G49690 |
| | | | | | | | | AT5G53990 |
| | | | | | | | | AT5G12890 |
| | | | | | | | | AT5G17040 |
| | | | | | | | | AT5G54010 |
| | | | | | | | | AT5G17050 |
| | | | | | | | | AT5G17030 |
| | | | | | | | | AT5G54060 |
| | | | | | | | | AT5G66690 |
| | | | | | | | | AT5G26310 |

| ANR | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Glyma08g06640.1 | Phvul.004G012500.1 | chr2.CM81.100 | Pisum_sativum_7873 | Medtr2g013260.1 | Ca_00291 | C.cajan_10881 | F0SHL7R02IL9ZU | AT2G23910.1 |
| Glyma08g06630.1 | Phvul.006G111600.1 | chr1.CM05.700 | | Medtr7g117190.1 | Ca_05336 | C.cajan_10882 | F0SHL7R02ISK76 | AT1G61720.1 |
| Glyma19g44360.1 | Phvul.006G111500.1 | chr1.CM05.730 | | Medtr4g006940.1 | | C.cajan_10883 | F0SHL7R02IUYB1 | AT5G14700.1 |
| Glyma19g00980.1 | Phvul.002G218600.1 | chr4.CM16.680 | | | | C.cajan_33121 | | |
| Glyma19g44370.1 | Phvul.002G218700.1 | | | | | C.cajan_28125 | | |
| Glyma03g41740.1 | | | | | | | | |
| F3H | | | | | | | | |
| Glyma14g06400.1 | Phvul.003G261900.1 | chr1.CM39.520 | p.sativum_07211 | Medtr7g081710.1 | Ca_14829 | C.cajan_11586 | F0SHL7R02I26TS | AT1G49390.1 |
| Glyma02g15360.1 | Phvul.009G053100.1 | chr4.CM19.240 | p.sativum_07212 | Medtr7g102820.1 | Ca_03975 | C.cajan_28266 | F0SHL7R02CGTEO3 | AT1G78550.1 |
| Glyma02g05470.1 | Phvul.008G047500.1 | chr6.CM02.250 | p.sativum_07213 | Medtr5g032900.1 | Ca_13408 | C.cajan_08850 | F0SHL7R02H0FXT | AT3G51240.1 |
| Glyma02g05450.1 | | | p.sativum_07214 | Medtr5g032880.1 | Ca_09972 | C.cajan_43594 | | AT5G20550.1 |
| Glyma11g35430.1 | | | | Medtr5g022020.1 | Ca_20981 | | | AT5G24530.1 |
| Glyma06g14190.1 | | | | Medtr8g075890.1 | Ca_01969 | | | AT5G54000.1 |
| Glyma06g11590.1 | | | | Medtr8g074070.1 | | | | AT5G20400.1 |
| Glyma06g13370.1 | | | | Medtr8g075830.1 | | | | |
| Glyma01g03120.1 | | | | Medtr4g115990.1 | | | | |
| Glyma01g37120.1 | | | | | | | | |
| Glyma03g42250.2 | | | | | | | | |
| Glyma20g01200.1 | | | | | | | | |
| Glyma07g29650.1 | | | | | | | | |
| Glyma16g01990.1 | | | | | | | | |
| Glyma16g23880.1 | | | | | | | | |
| Glyma18g03020.1 | | | | | | | | |
| LAR | | | | | | | | |
| Glyma20g32681.1 | Phvul.007G102100.1 | chr1.CM05.700 | p.sativum_10508 | Medtr7g117190.1 | Ca_08438 | C.cajan_30532 | F0SHL7R02GJ3LG | AT1G19540.1 |
| Glyma10g34720.1 | Phvul.006G111500.1 | | | Medtr1g093720.1 | | | | AT5G19440.1 |
| ANS | | | | | | | | |
| Glyma14g06400.1 | Phvul.010G000300.1 | chr2.CM02.140 | p.sativum_22948 | Medtr3g117420.1 | Ca_08884 | C.cajan_18992 | F0SHL7R02I5XQC | AT2G38240.1 |
| Glyma02g42470.1 | Phvul.009G003500.1 | chr1.CM13.530 | p.sativum_25793 | Medtr3g070860.1 | Ca_13408 | C.cajan_13147 | F0SHL7R02MB3T | AT4G22880.1 |
| Glyma11g03010.1 | Phvul.009G003400.1 | chr6.CM02.250 | p.sativum_26504 | Medtr5g011250.1 | Ca_05192 | C.cajan_29237 | F0SHL7R02B6E6 | AT3G55970.1 |
| Glyma03g07680.3 | Phvul.008G223500.1 | | p.sativum_29710 | Medtr5g082750.1 | Ca_17705 | C.cajan_43296 | F0SHL7R02FUXAZ | AT5G05600.1 |
| Glyma07g18280.1 | Phvul.008G098500.1 | | | Medtr8g074070.1 | Ca_02395 | C.cajan_02682 | F0SHL7R02H0JCY | |
| Glyma18g03020.1 | Phvul.001G244000.1 | | | Medtr8g074050.1 | Ca_22198 | C.cajan_05536 | F0SHL7R02C9MLG | |
| Glyma01g42350.1 | Phvul.002G152700.1 | | | Medtr4g015790.1 | | | F0SHL7R02GNFGE | |
| | | | | AC235756_31.1 | | | F0SHL7R02H374E | |
| IFS | | | | | | | | |
| Glyma13g24200.1 | Phvul.003G051800.1 | chr3.CM27.620 | Pisum_sativum_179 | Medtr6g042540.1 | Ca_14219 | C.cajan_13947 | F0SHL7R02G4K81 | AT4G15360.1 |
| Glyma07g34560.1 | Phvul.003G074000.1 | chr1.CM18.690 | Pisum_sativum_4827 | Medtr6g042610.1 | Ca_06358 | C.cajan_04386 | F0SHL7R02HAMVX | AT1G50520.1 |
| Glyma07g31390.2 | Phvul.003G051700.1 | | p.sativum_19217 | | | | F0SHL7R02HAOUD | AT1G50560.1 |
| Glyma07g32330.1 | | | p.sativum_21770 | | | | | |
| Glyma07g34540.1 | | | | | | | | |
| FLS | | | | | | | | |
| Glyma11g35430.1 | Phvul.003G011800.1 | chr1.CM39.520 | Pisum_sativum_6596 | Medtr3g072810.1 | Ca_14829 | C.cajan_11586 | F0SHL7R02HZ87G | AT5G63600.1 |
| Glyma06g11590.1 | Phvul.009G053100.1 | chr4.CM87.690 | p.sativum_22405 | Medtr3g072820.1 | Ca_03975 | C.cajan_28266 | F0SHL7R02HZF63 | AT5G63590.1 |
| Glyma06g13370.1 | Phvul.008G166000.1 | | | Medtr5g059130.1 | Ca_13408 | C.cajan_43231 | F0SHL7R02GLIYH | AT5G63580.1 |
| Glyma13g02740.1 | Phvul.004G052400.1 | | | Medtr5g059140.1 | Ca_23835 | C.cajan_16130 | F0SHL7R02IOXSP | AT5G43935.1 |
| Glyma05g12770.1 | Phvul.002G149600.1 | | | | Ca_24703 | C.cajan_32369 | F0SHL7R02HGWGF | AT5G63595.1 |
| | | | | | | C.cajan_01360 | | AT5G08640.1 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| FNS | Glyma12g07200.1 | Phvul.003G009200.1 | chr6.CM37.530 | Pisum_sativum_3027 | AC225458_44.1 | Ca_26807 | C.cajan_21106 | F0SHL7R02HSZ34 | AT1G28430.1 |
| | Glyma12g07190.1 | Phvul.009G061600.1 | chr6.CM37.560 | Pisum_sativum_3168 | Medtr3g077460.1 | Ca_28260 | C.cajan_21139 | F0SHL7R02JJVHS | |
| | Glyma11g15330.2 | Phvul.009G172200.1 | chr6.CM37.580 | Pisum_sativum_3371 | Medtr3g020780.1 | Ca_00181 | C.cajan_21420 | F0SHL7R02F0W6Z | |
| | Glyma19g32880.1 | Phvul.009G061500.1 | chr6.CM37.910 | Pisum_sativum_8093 | Medtr3g076560.1 | Ca_10571 | C.cajan_24440 | F0SHL7R02I122T | |
| | Glyma19g32650.1 | Phvul.009G061400.1 | chr3.CM0127.620 | p.sativum_07271 | Medtr3g076530.1 | Ca_17249 | C.cajan_24441 | F0SHL7R02HAOUD | |
| | Glyma03g29950.1 | Phvul.009G078400.1 | chr3.CM0243.350 | p.sativum_07272 | Medtr3g058060.1 | Ca_22302 | C.cajan_41297 | F0SHL7R02IG3KX | |
| | Glyma03g29790.1 | Phvul.009G172000.1 | chr4.CM0288.420 | p.sativum_08045 | Medtr3g076550.1 | Ca_22375 | C.cajan_32626 | F0SHL7R02F4FSM | |
| | Glyma03g29780.1 | Phvul.009G172100.1 | chr6.CM0314.520 | p.sativum_08046 | Medtr2g010380.1 | Ca_08416 | C.cajan_32628 | F0SHL7R02GZ942 | |
| | Glyma10g12060.1 | Phvul.005G002400.1 | chr5.CM0345.740 | p.sativum_10908 | Medtr2g010280.1 | Ca_10754 | C.cajan_33039 | F0SHL7R02G04C4 | |
| | Glyma10g12100.1 | Phvul.011G213700.1 | chr4.CM0387.970 | p.sativum_10909 | Medtr2g010250.1 | Ca_14218 | C.cajan_40817 | | |
| | | Phvul.011G016700.1 | chr1.CM0442.470 | p.sativum_11756 | Medtr2g010300.1 | Ca_14219 | C.cajan_08534 | | |
| | | Phvul.011G159600.1 | chr5.CM0698.200 | p.sativum_13736 | Medtr2g010290.1 | Ca_07435 | C.cajan_08538 | | |
| | | Phvul.011G068100.1 | chr2.CM0803.400 | p.sativum_18694 | Medtr2g010330.1 | Ca_07437 | C.cajan_08544 | | |
| | | Phvul.011G173200.1 | chr3.CM1089.100 | p.sativum_19217 | Medtr2g010320.1 | Ca_05762 | C.cajan_09526 | | |
| | | Phvul.008G062700.1 | chr4.CM2115.200 | p.sativum_19384 | Medtr7g012860.1 | Ca_06339 | C.cajan_09528 | | |
| | | Phvul.008G277500.1 | | p.sativum_20223 | Medtr7g028020.1 | Ca_13864 | C.cajan_39176 | | |
| | | Phvul.008G155000.1 | | p.sativum_20242 | Medtr7g027960.1 | Ca_16721 | C.cajan_42811 | | |
| | | Phvul.008G248600.1 | | p.sativum_21770 | Medtr7g012330.1 | Ca_03361 | C.cajan_04386 | | |
| | | Phvul.004G021300.1 | | | Medtr6g008630.1 | Ca_11711 | C.cajan_32105 | | |
| | | Phvul.004G118900.1 | | | Medtr6g042540.1 | Ca_11712 | C.cajan_30959 | | |
| | | Phvul.004G119200.1 | | | Medtr6g008530.1 | Ca_14529 | C.cajan_18623 | | |
| | | Phvul.004G021900.1 | | | Medtr6g042610.1 | Ca_02088 | C.cajan_28898 | | |
| | | Phvul.004G159800.1 | | | Medtr6g008650.1 | Ca_02089 | C.cajan_11253 | | |
| | | Phvul.004G159300.1 | | | Medtr5g072980.1 | Ca_02090 | C.cajan_11260 | | |
| | | Phvul.004G159900.1 | | | Medtr5g018480.1 | | C.cajan_11261 | | |
| | | Phvul.004G021700.1 | | | Medtr5g034900.1 | | C.cajan_11263 | | |
| | | Phvul.004G118800.1 | | | Medtr5g072930.1 | | C.cajan_12595 | | |
| | | Phvul.004G021600.1 | | | Medtr5g094570.1 | | C.cajan_12596 | | |
| | | Phvul.004G021100.1 | | | Medtr5g045770.1 | | C.cajan_43848 | | |
| | | Phvul.004G021800.1 | | | Medtr5g095260.1 | | C.cajan_13938 | | |
| | | Phvul.004G159600.1 | | | Medtr5g023680.1 | | C.cajan_13947 | | |
| | | Phvul.004G021200.1 | | | Medtr5g073020.1 | | C.cajan_14174 | | |
| | | Phvul.004G159500.1 | | | Medtr5g026520.1 | | C.cajan_14419 | | |
| | | Phvul.004G085000.1 | | | Medtr5g070710.1 | | C.cajan_29243 | | |
| | | Phvul.004G119300.1 | | | Medtr8g063260.1 | | C.cajan_29246 | | |
| | | Phvul.004G118700.1 | | | Medtr8g063280.1 | | C.cajan_24634 | | |
| | | Phvul.004G022000.1 | | | Medtr1g039590.1 | | C.cajan_47317 | | |
| | | Phvul.004G021400.1 | | | Medtr1g023720.1 | | C.cajan_22785 | | |
| | | Phvul.004G021500.1 | | | Medtr1g023730.1 | | C.cajan_31856 | | |
| | | Phvul.007G257500.1 | | | Medtr1g023700.1 | | C.cajan_40735 | | |
| | | Phvul.007G257400.1 | | | Medtr4g026320.1 | | C.cajan_32305 | | |
| | | Phvul.007G257300.1 | | | Medtr4g131830.1 | | | | |
| | | Phvul.007G104800.1 | | | | | | | |
| | | Phvul.001G008900.1 | | | | | | | |
| | | Phvul.001G139400.1 | | | | | | | |
| | | Phvul.001G037100.1 | | | | | | | |
| | | Phvul.001G139500.1 | | | | | | | |
| | | Phvul.006G054500.1 | | | | | | | |
| | | Phvul.006G039800.1 | | | | | | | |
| | | Phvul.006G138100.1 | | | | | | | |
| | | Phvul.006G209600.1 | | | | | | | |
| | | Phvul.006G054600.1 | | | | | | | |
| | | Phvul.002G025000.1 | | | | | | | |
| | | Phvul.002G125200.1 | | | | | | | |
| | | Phvul.002G097000.1 | | | | | | | |
| | | Phvul.002G025200.1 | | | | | | | |
| | | Phvul.002G226700.1 | | | | | | | |
| | | Phvul.002G161900.1 | | | | | | | |
| | | Phvul.002G022900.1 | | | | | | | |
| | | Phvul.002G173900.1 | | | | | | | |
| | | Phvul.002G097100.1 | | | | | | | |

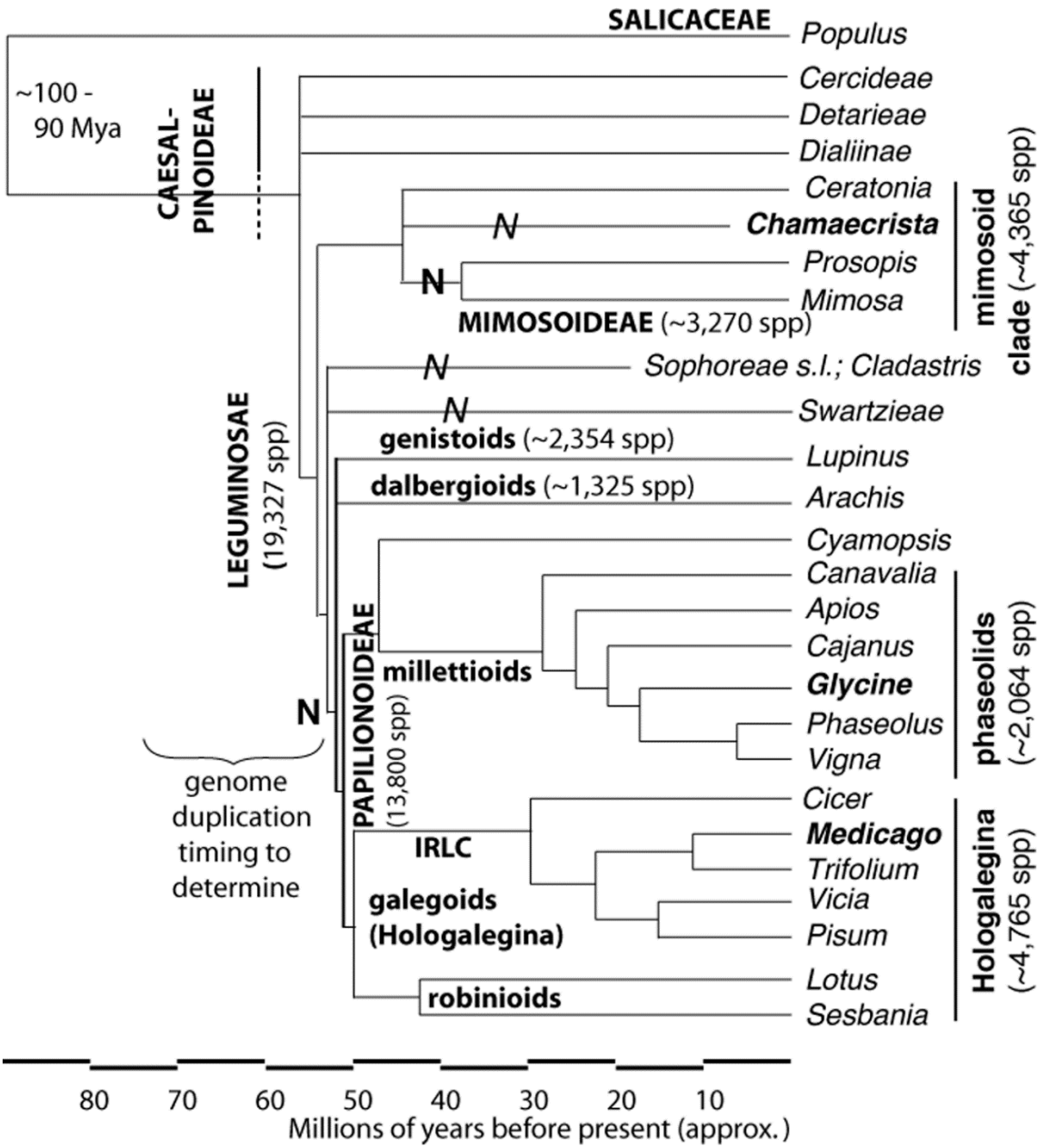



Figure 2.1: Skeleton phylogeny of the legumes (Cannon S. et al 2010)

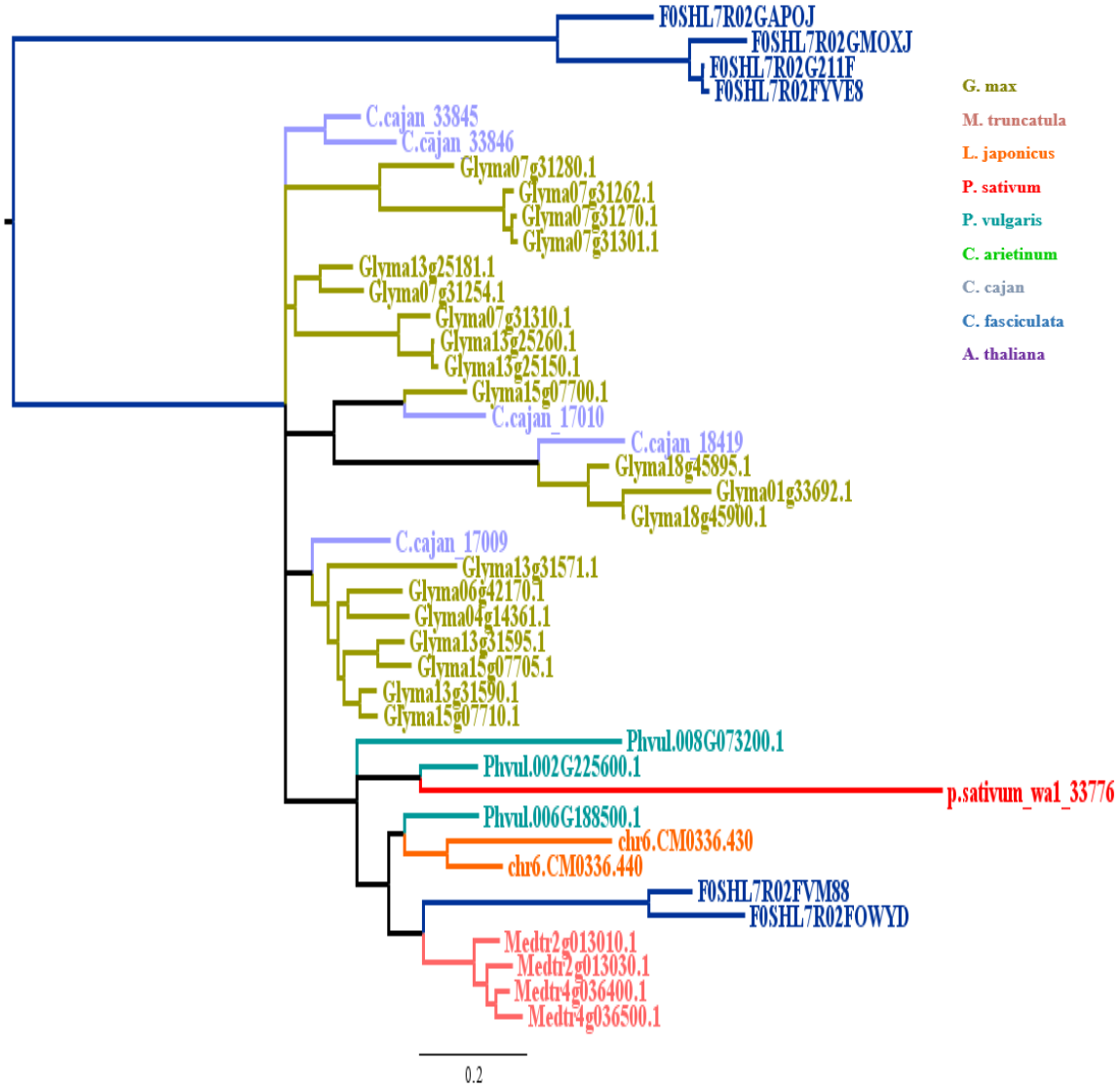



Figure 2.2: Aureusidin synthase gene tree result of Mrbayes

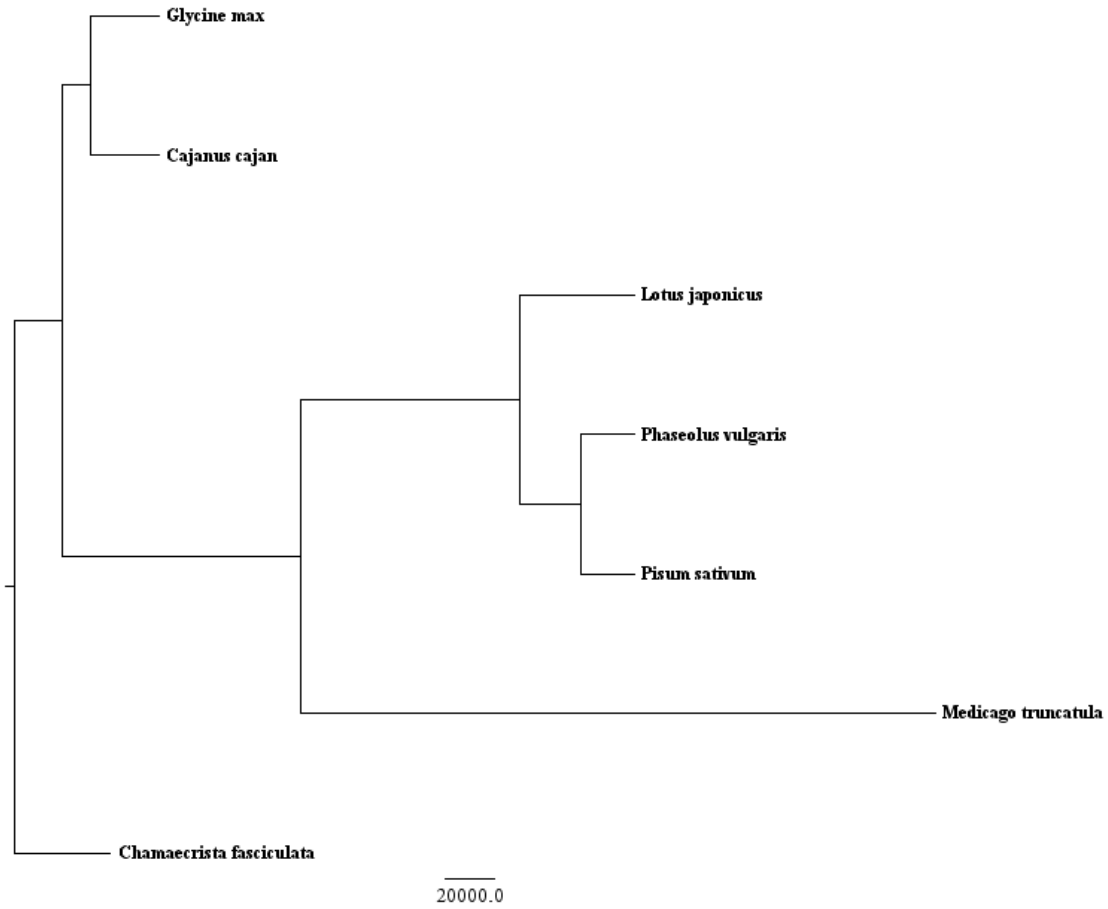



Figure 2.3: Species tree of aureusidin synthase from Beast


Figure 2.4: Chalcone isomerase gene tree result of Mrbayes

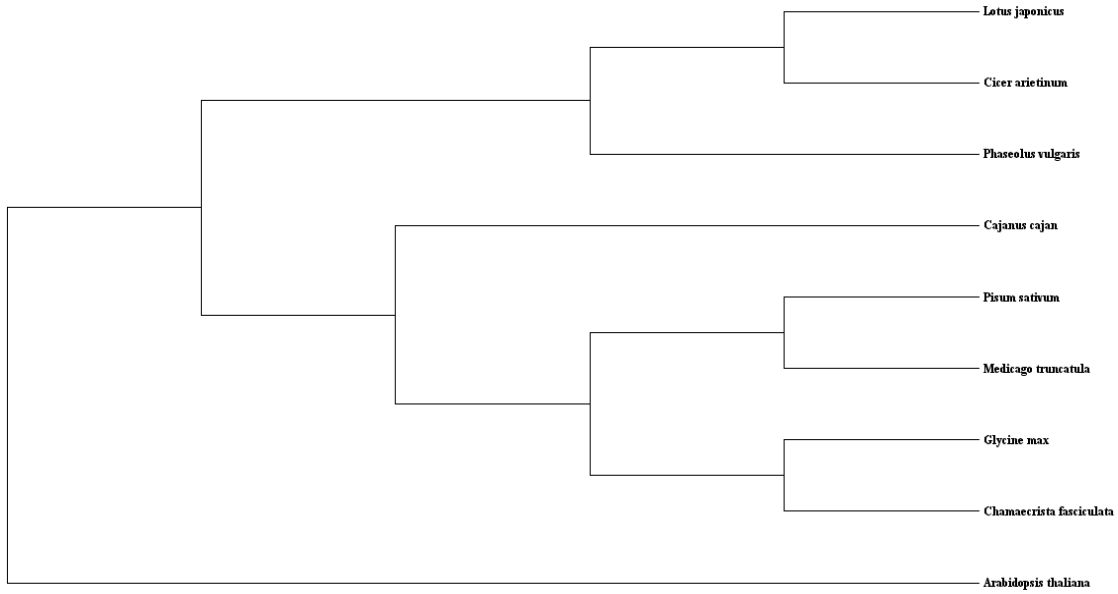



Figure 2.5: Species tree of chalcone isomerase from Beast


Figure 2.6: Chalcone synthase gene tree result of Mrbayes

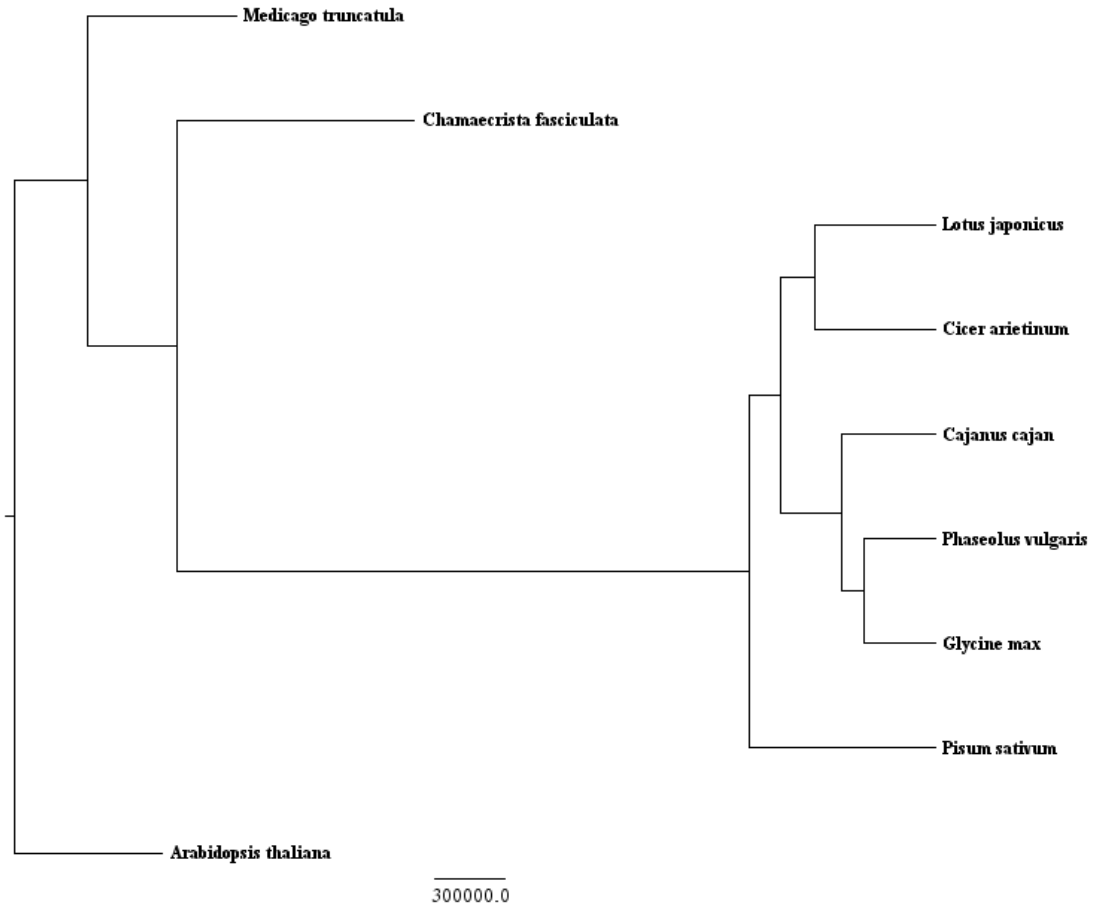



Figure 2.7: Species tree of chalcone synthase from Beast

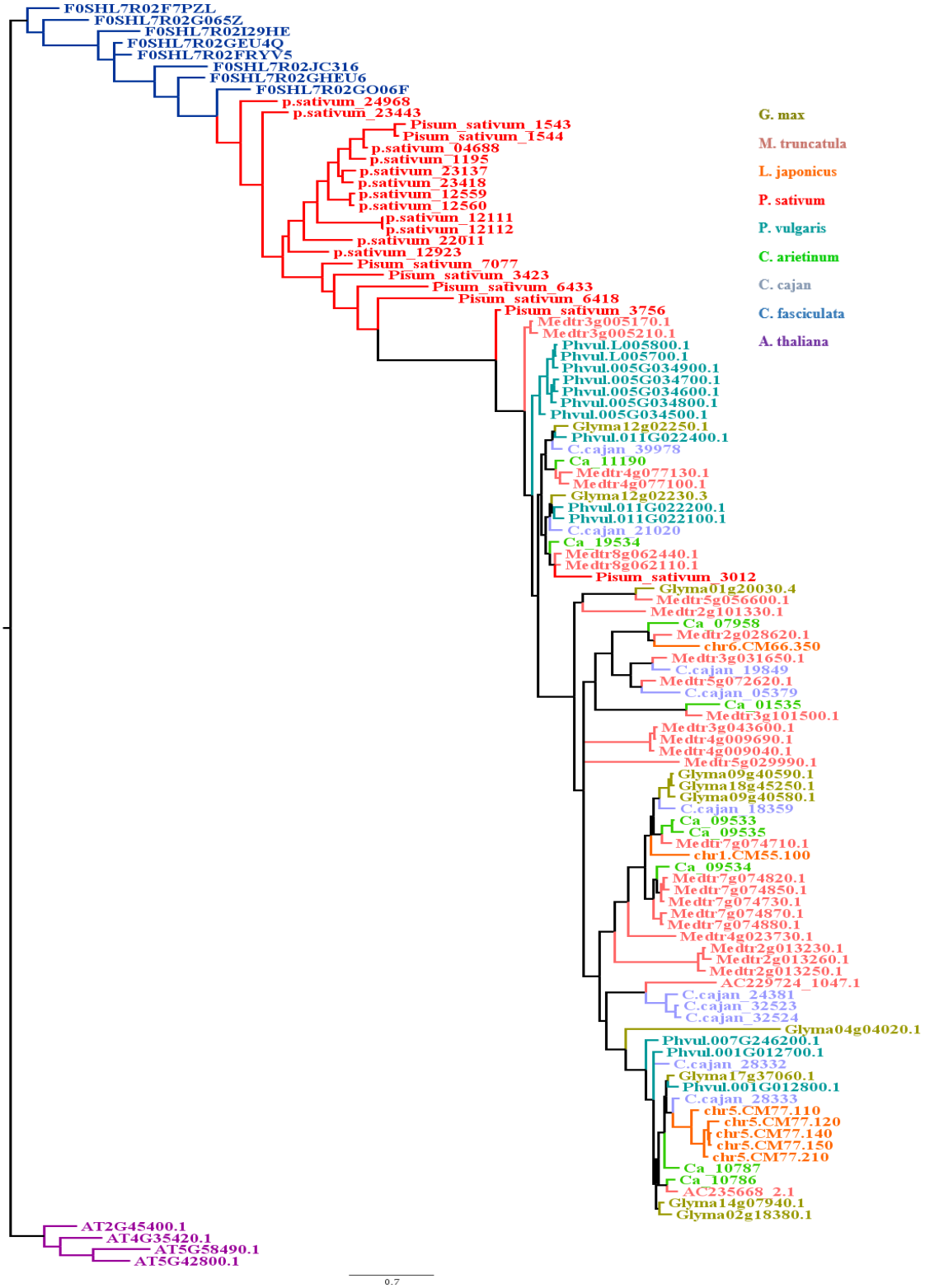



Figure 2.8: Dihydroflavonol 4-Reductase gene tree, Mrbayes output

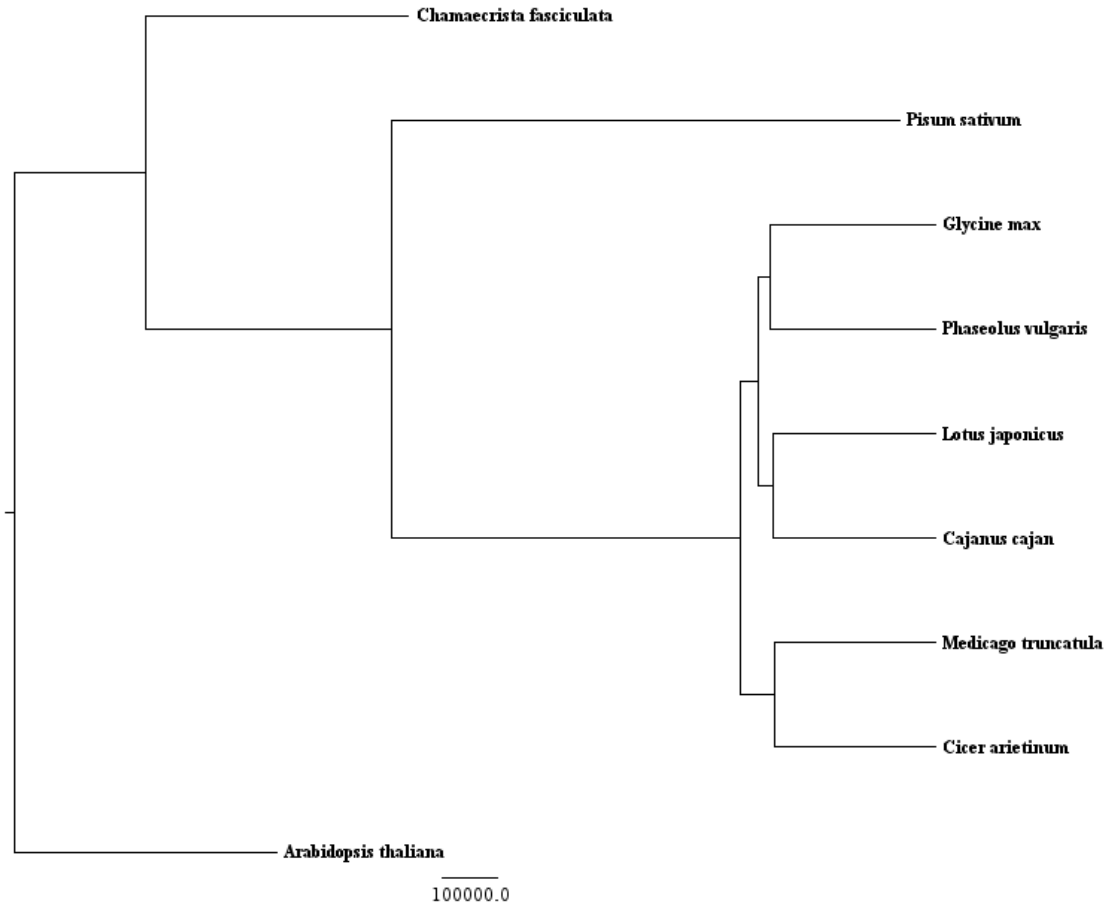



Figure 2.9: Species tree of Dihydroflavonol 4-Reductase from Beast

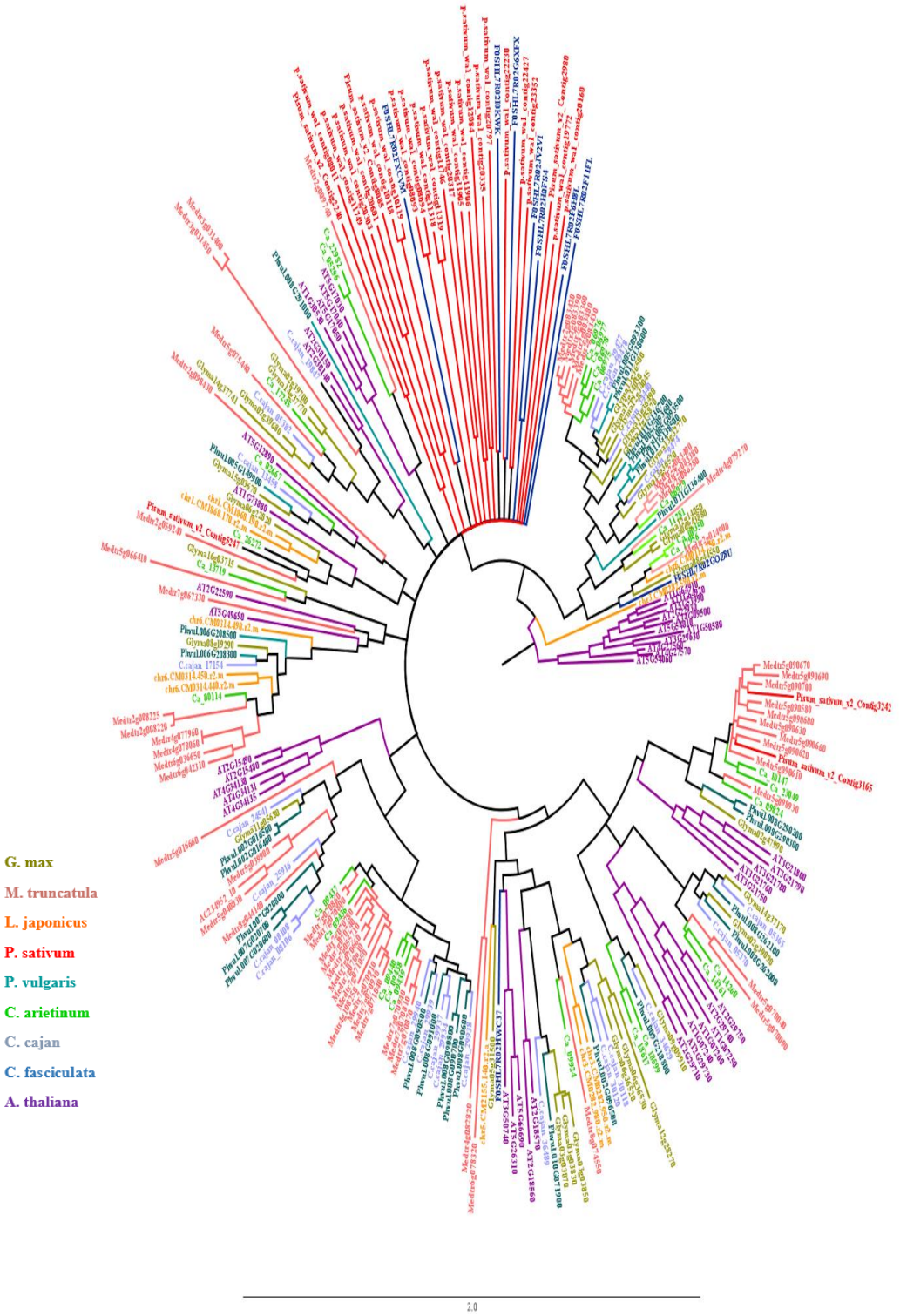



Figure 2.10: Anthocyanidin 3-O-Glucosyltransferase gene tree, Mrbayes output

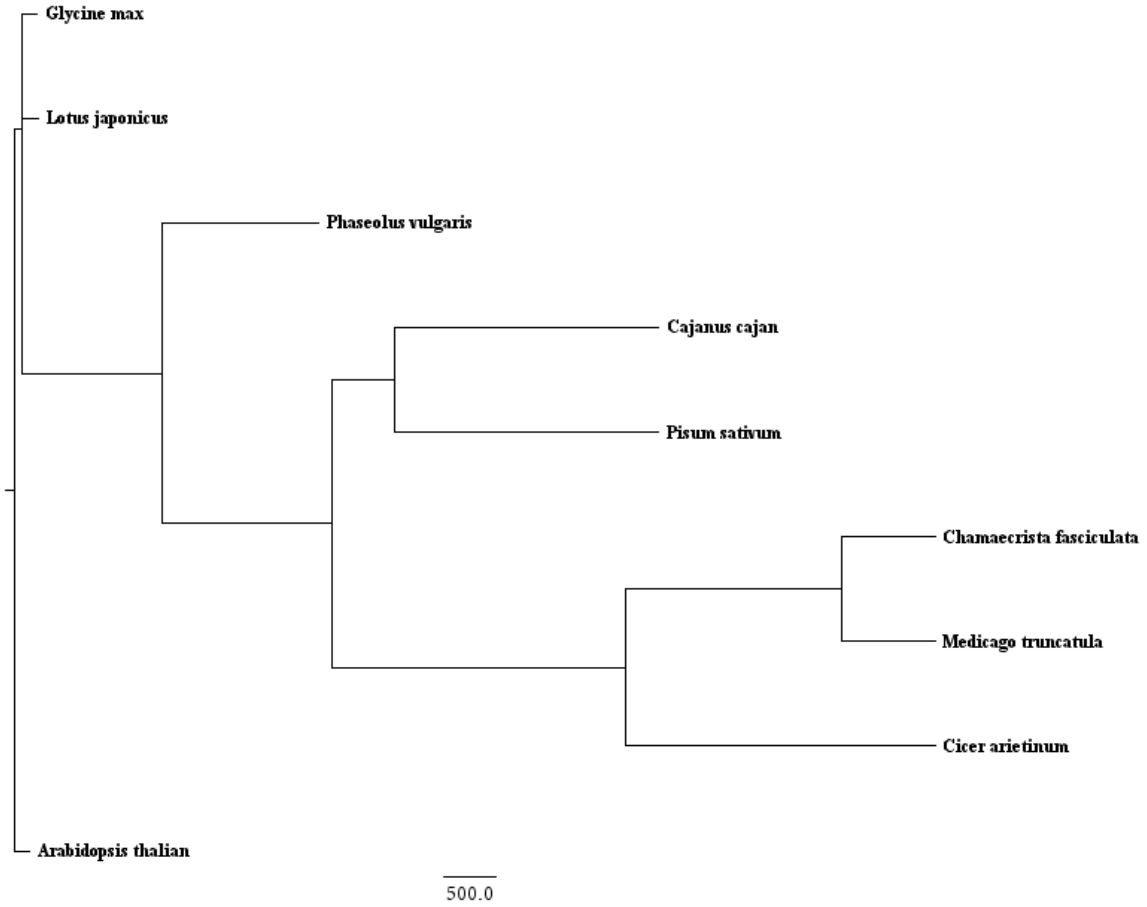



Figure 2.11: Species tree of anthocyanidin 3-O-Glucosyltransferase from Beast

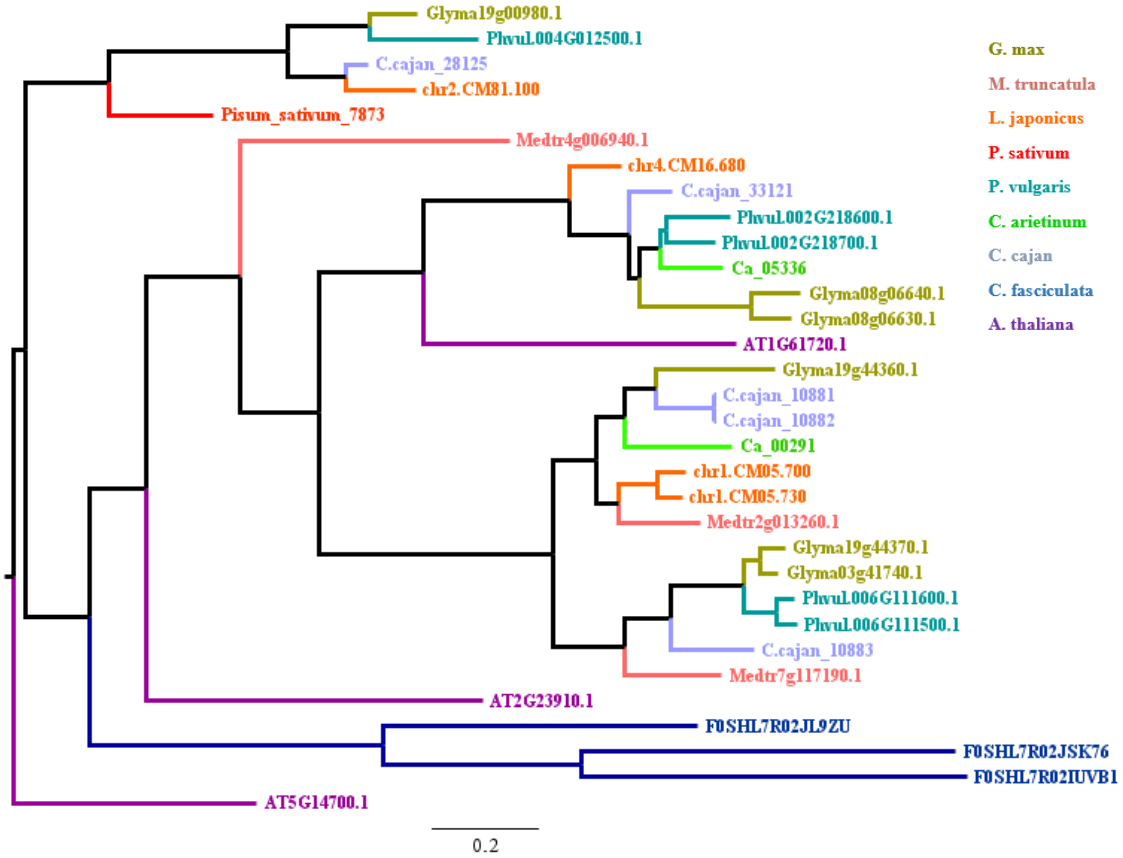



Figure 2.12: Anthocyanidin reductase gene tree from Mrbayes

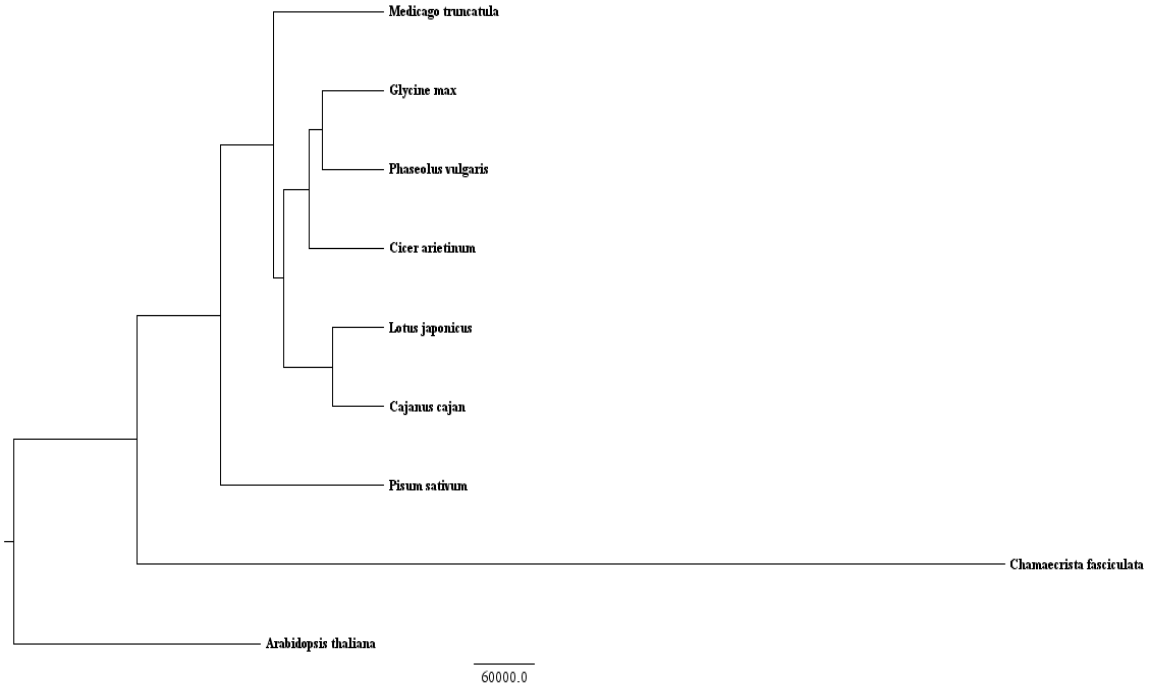



Figure 2.13: Species tree of anthocyanidin reductase from Beast

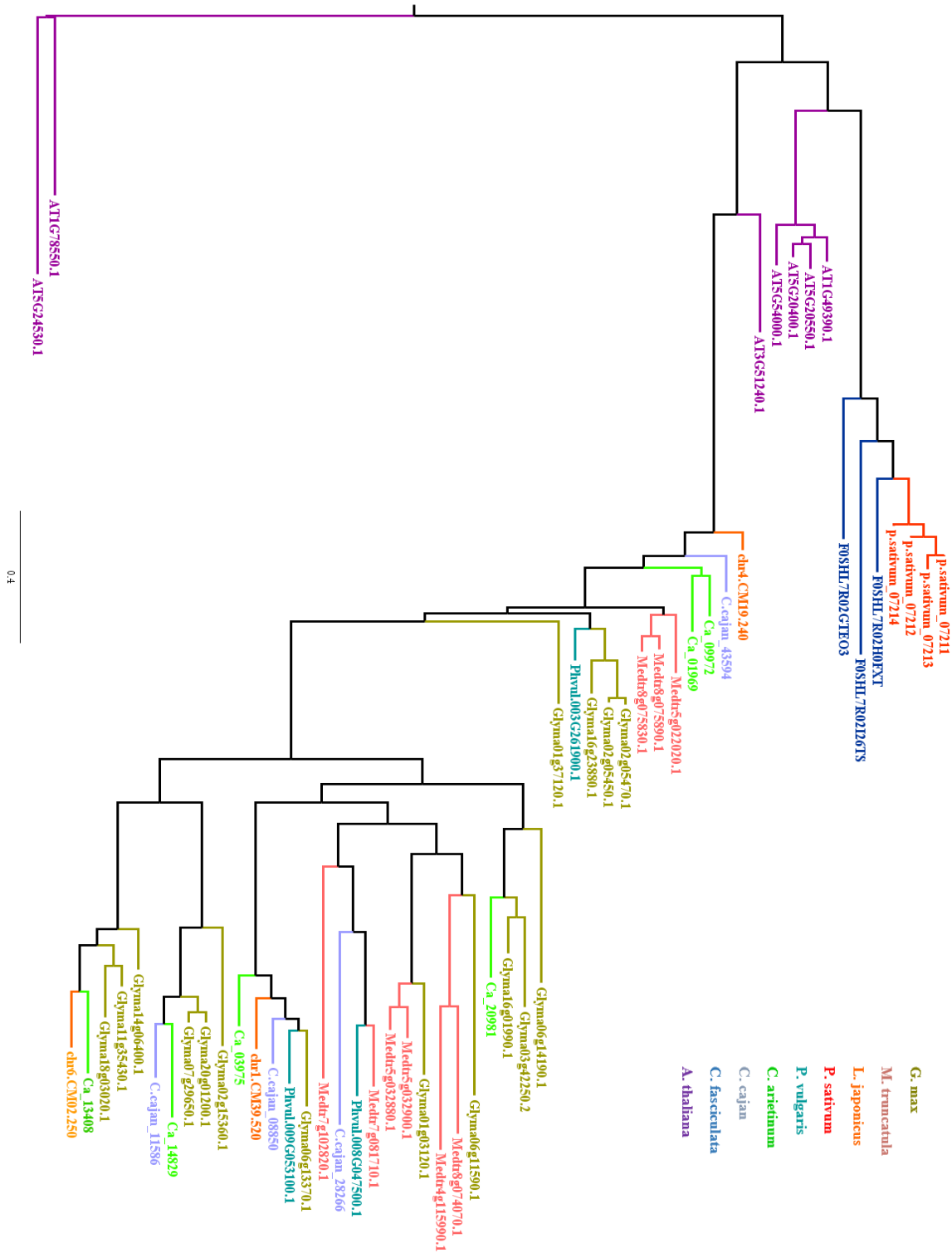



Figure 2.14: Flavanone 3-Hydroxylase gene tree, Mrbayes output

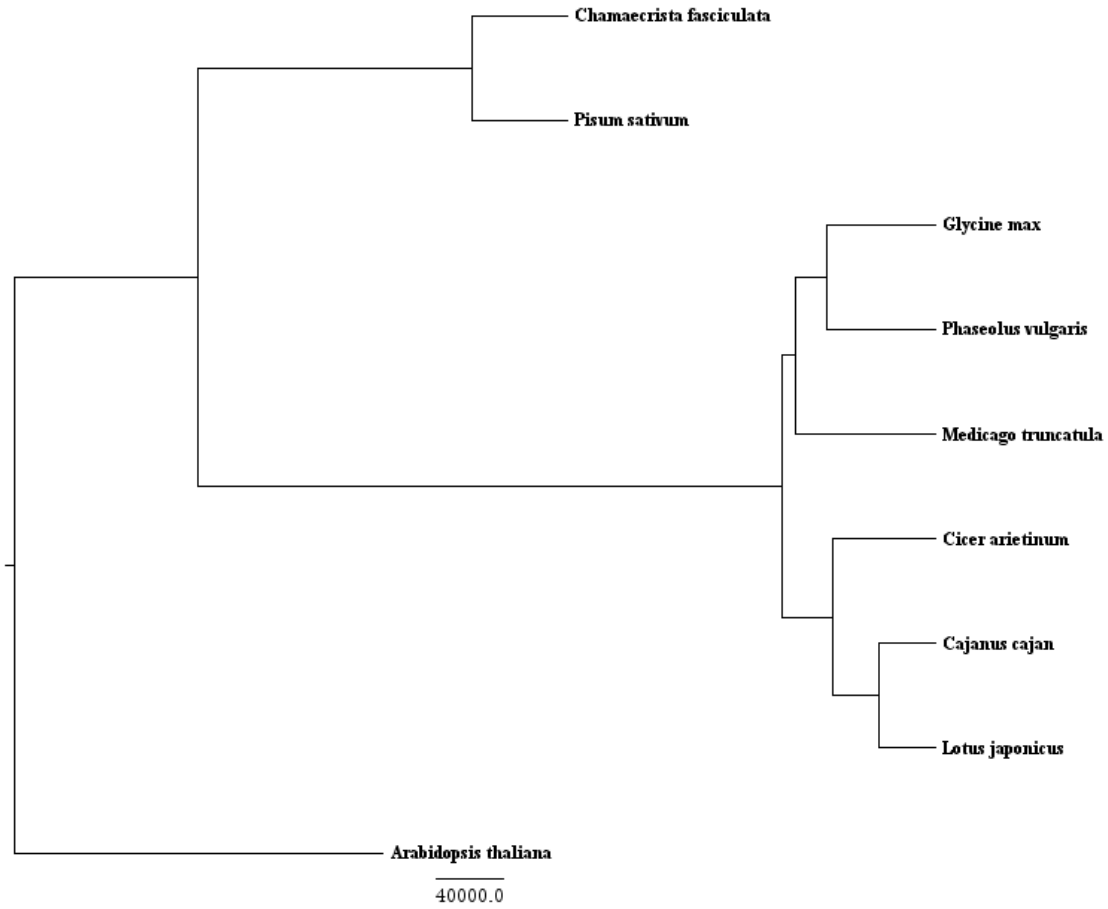



Figure 2.15: Species tree of Flavanone 3-Hydroxylase from Beast

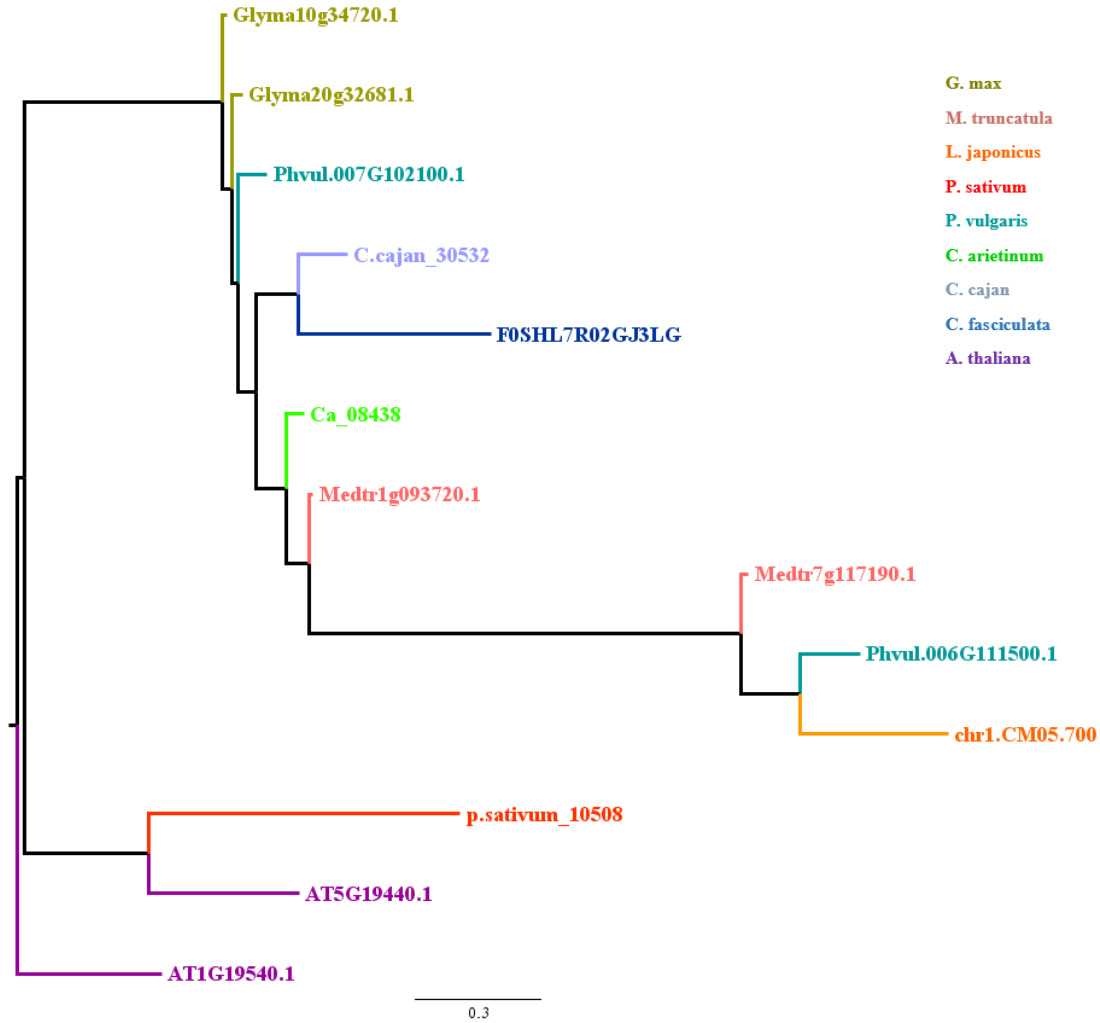



Figure 2.16: Leucoanthocyanidin reductase gene tree, Mrbayes output

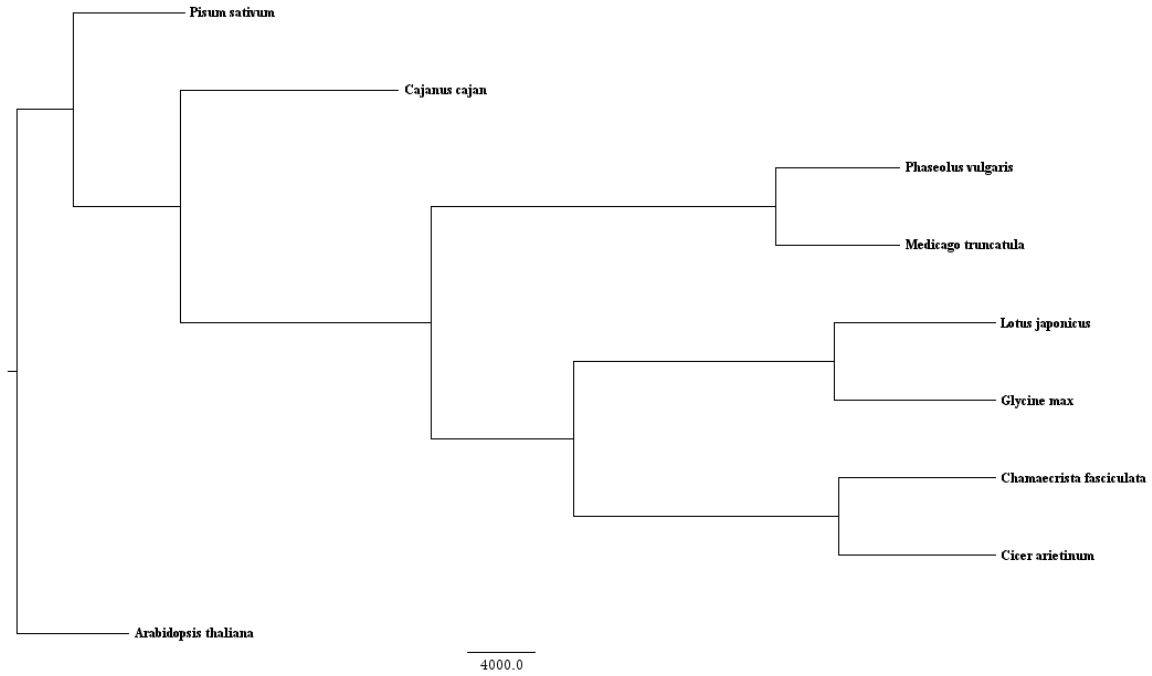



Figure 2.17: Species tree of leucoanthocyanidin reductase from Beast

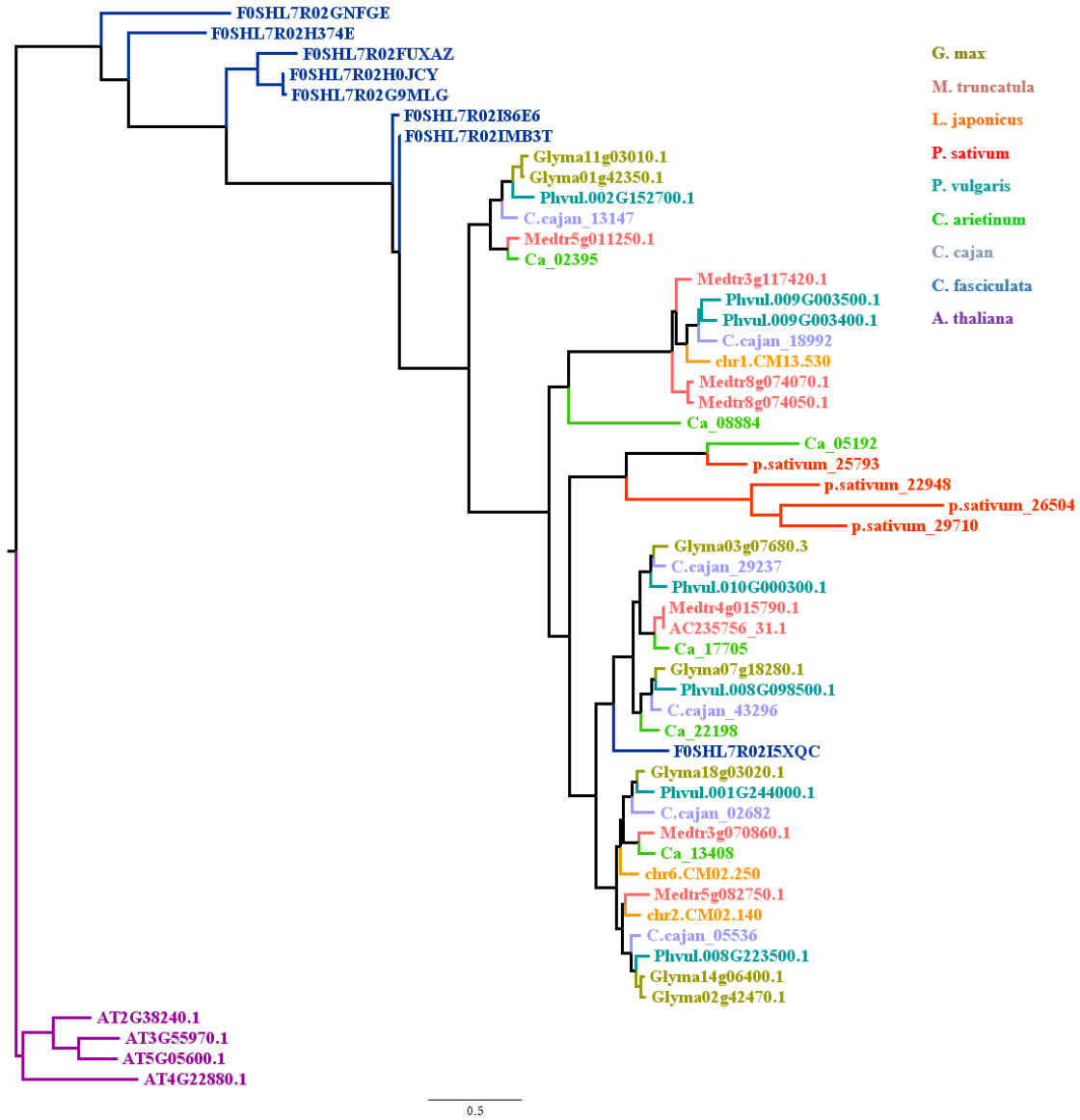



Figure 2.18: Anthocyanidin synthase gene tree, Mrbayes output

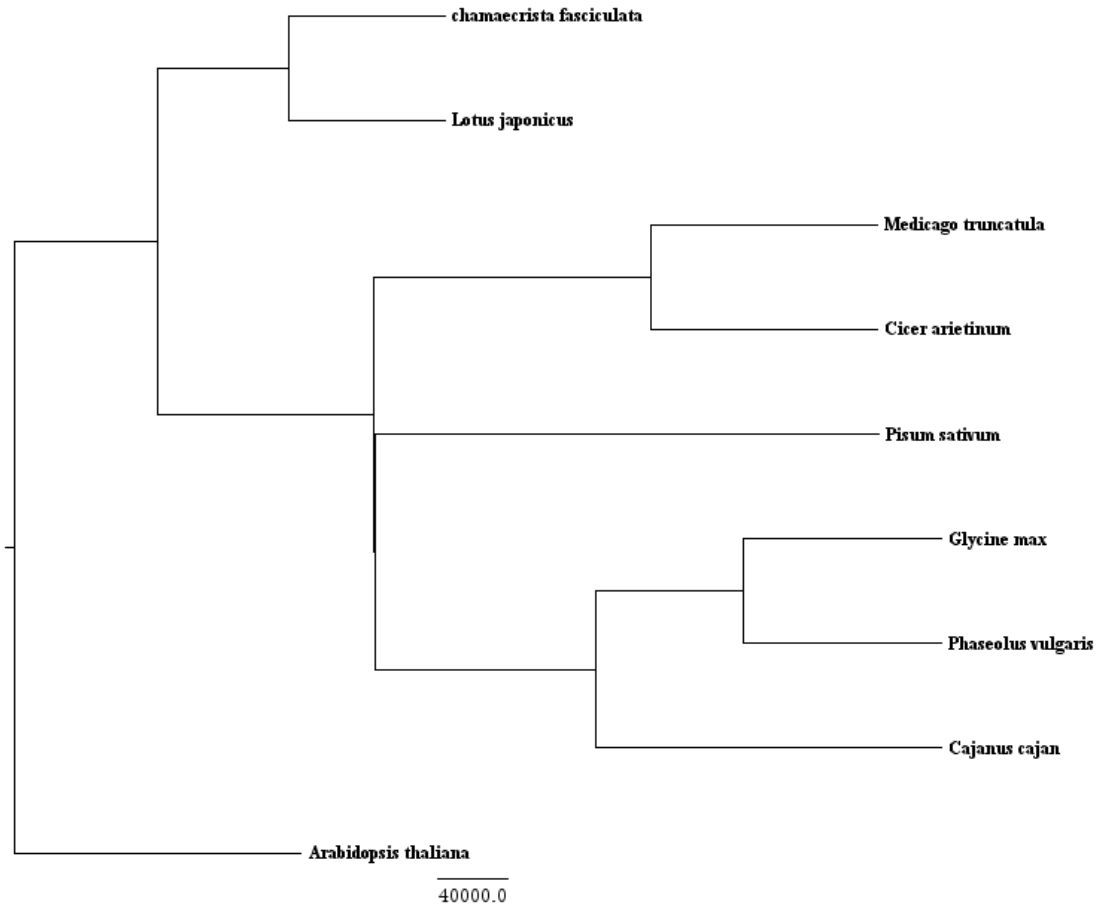



Figure 2.19: Species tree of Anthocyanidin synthase from Beast

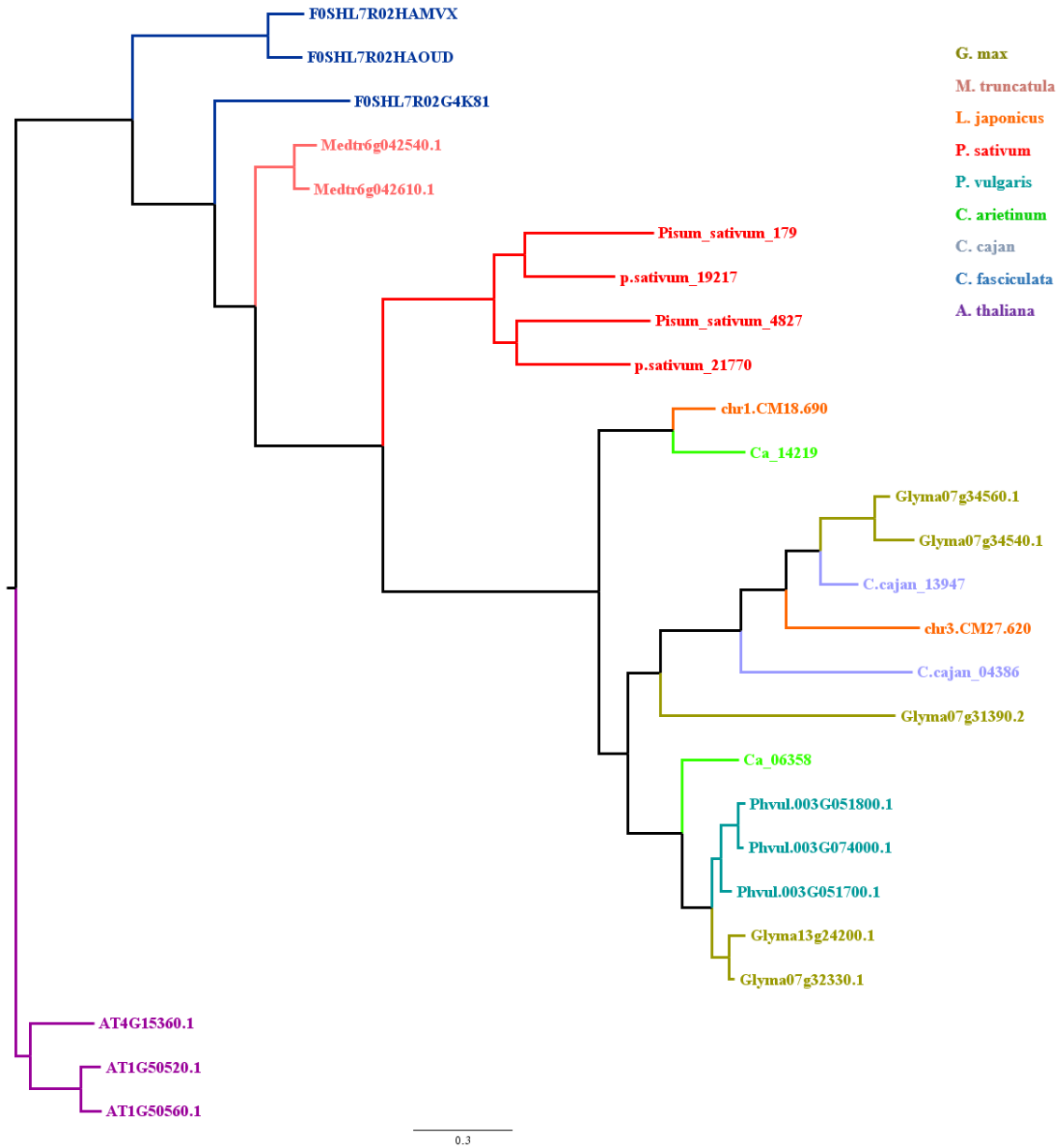



Figure 2.20: Isoflavone synthase gene tree, Mrbayes output

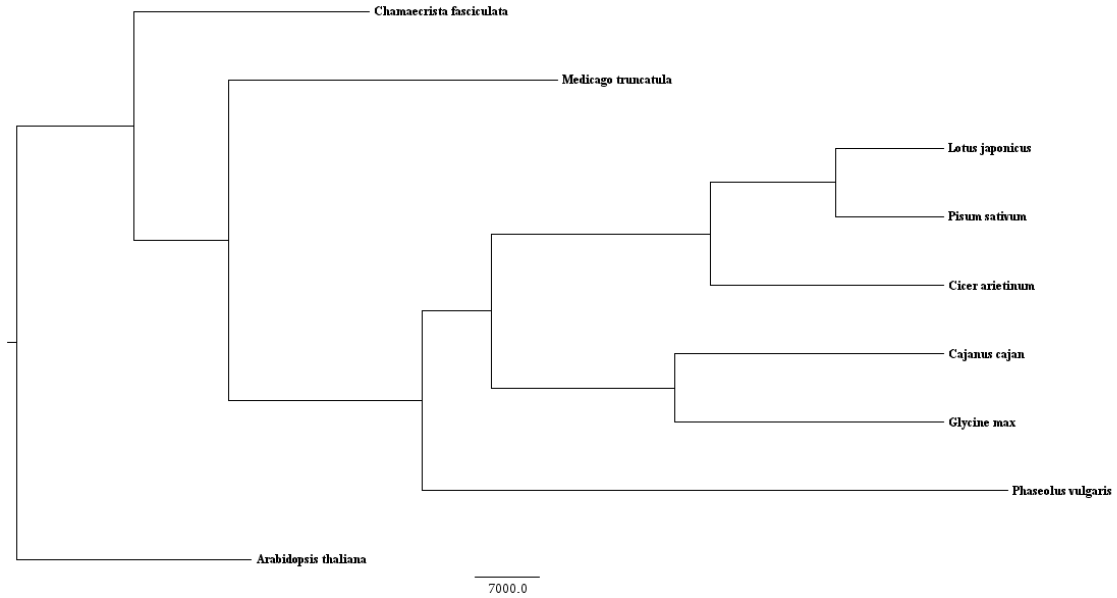



Figure 2.21: Species tree of isoflavone synthase from Beast

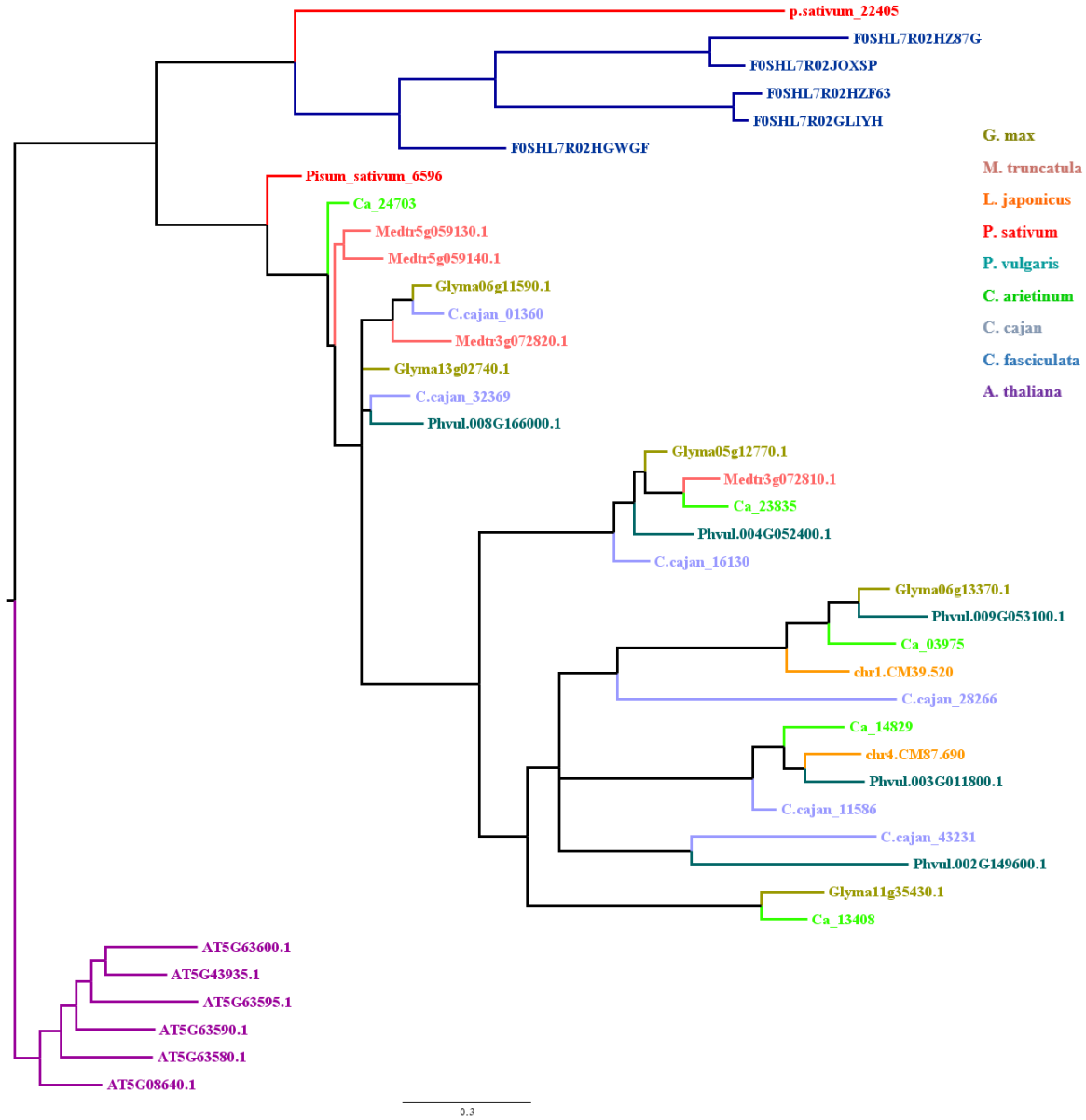



Figure 2.22: Flavonol synthase gene tree, Mrbayes output

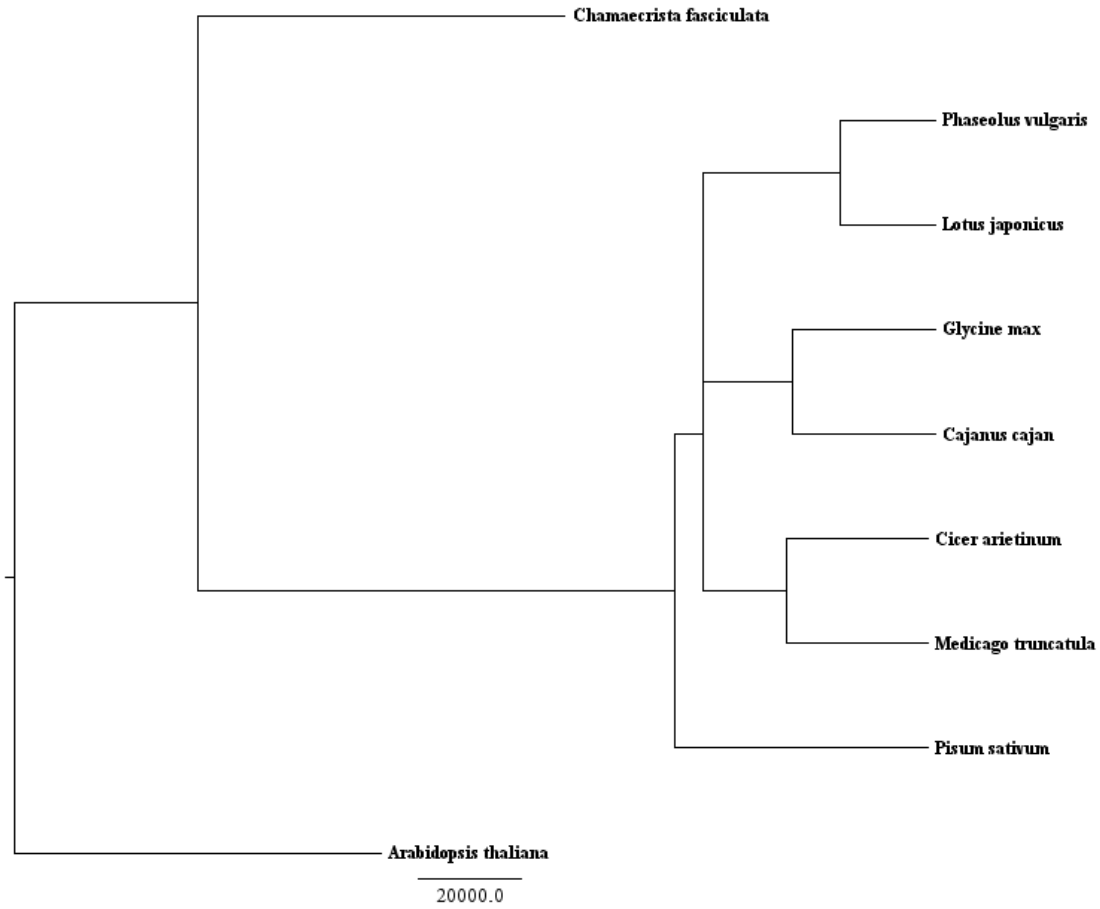



Figure 2.23: Species tree of flavonol synthase from Beast

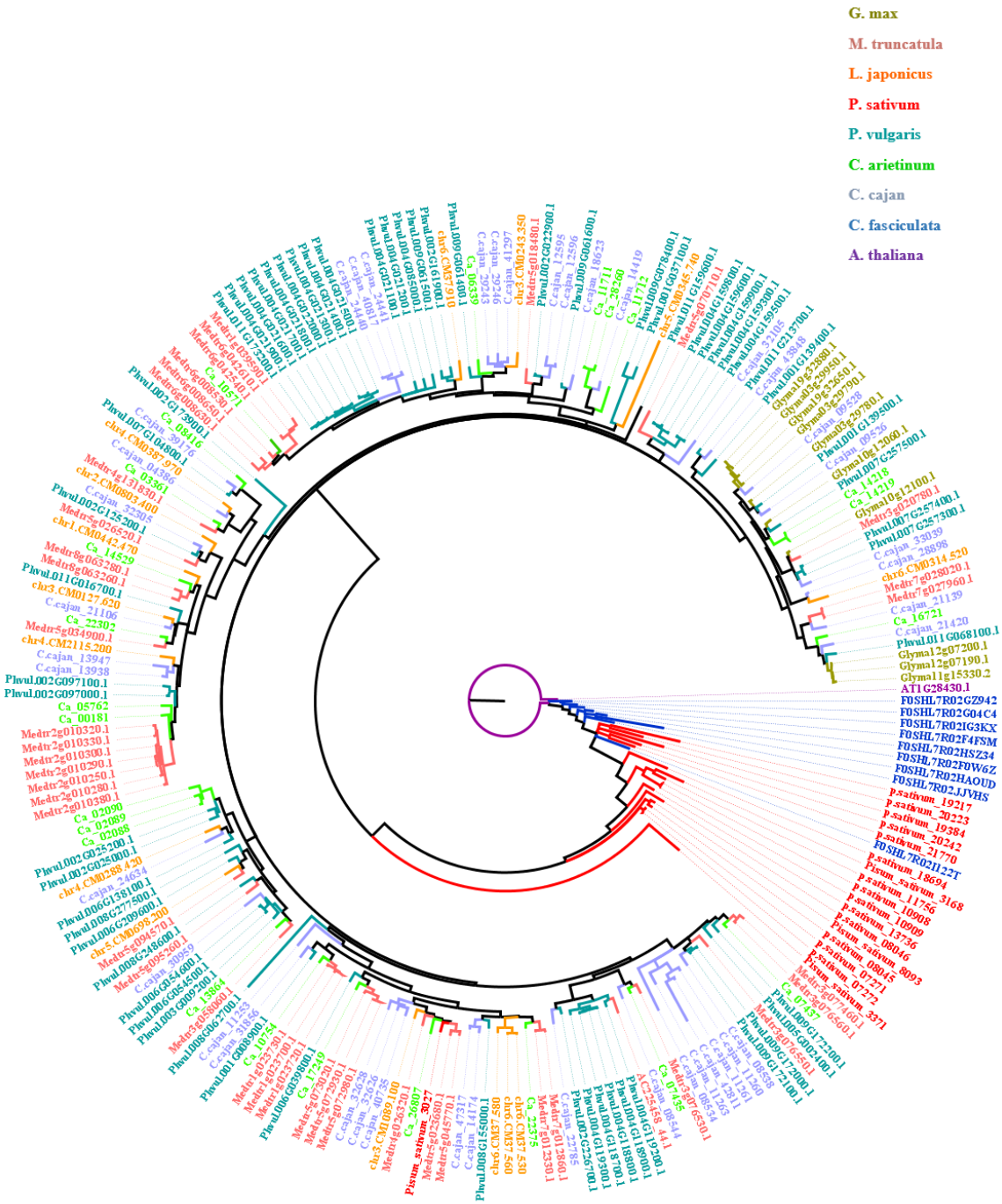



Figure 2.24: Flavone synthase gene tree, Mrbayes output

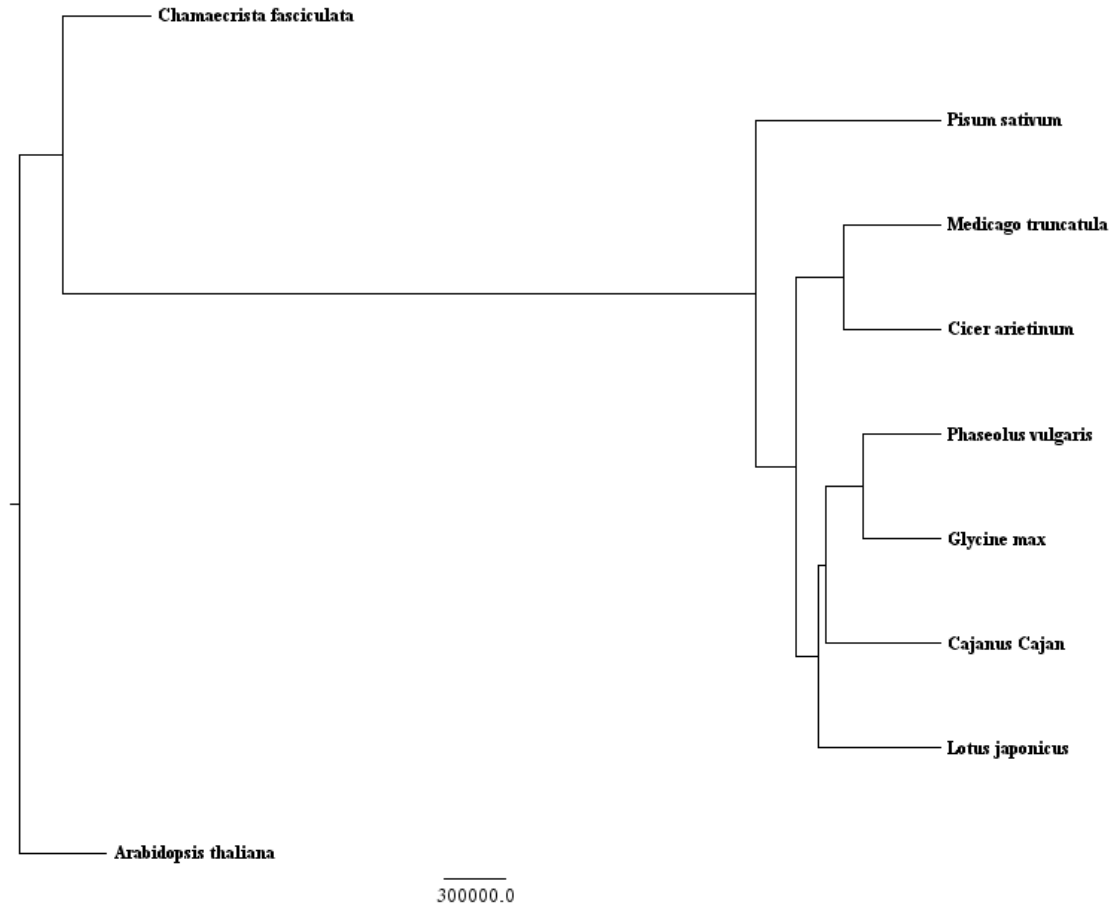



Figure 2.25: Species tree of flavone synthase from Beast

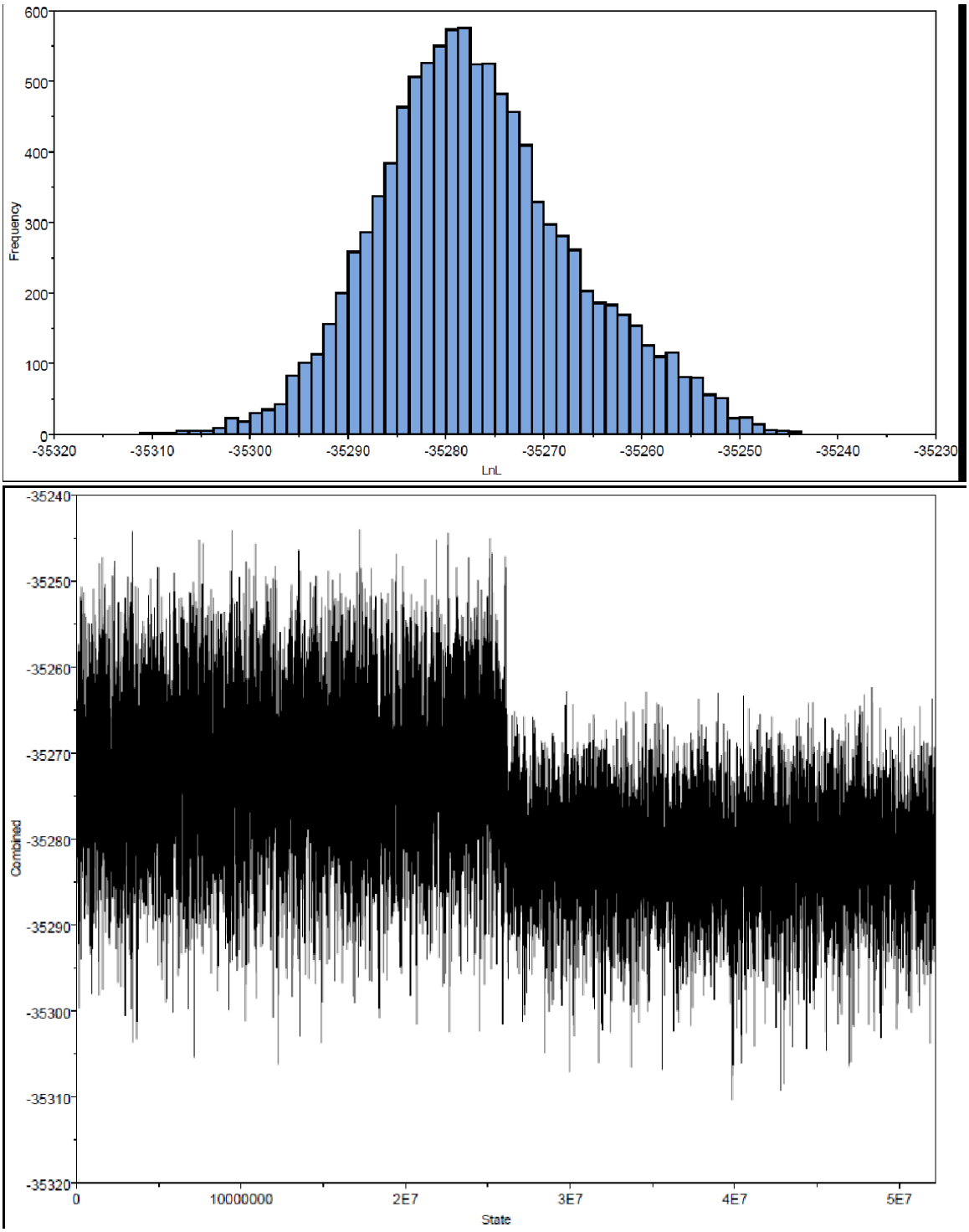



Figure 2.26: TRACER lnl estimates and trace plots of gene tree respectively in aureusidin synthase

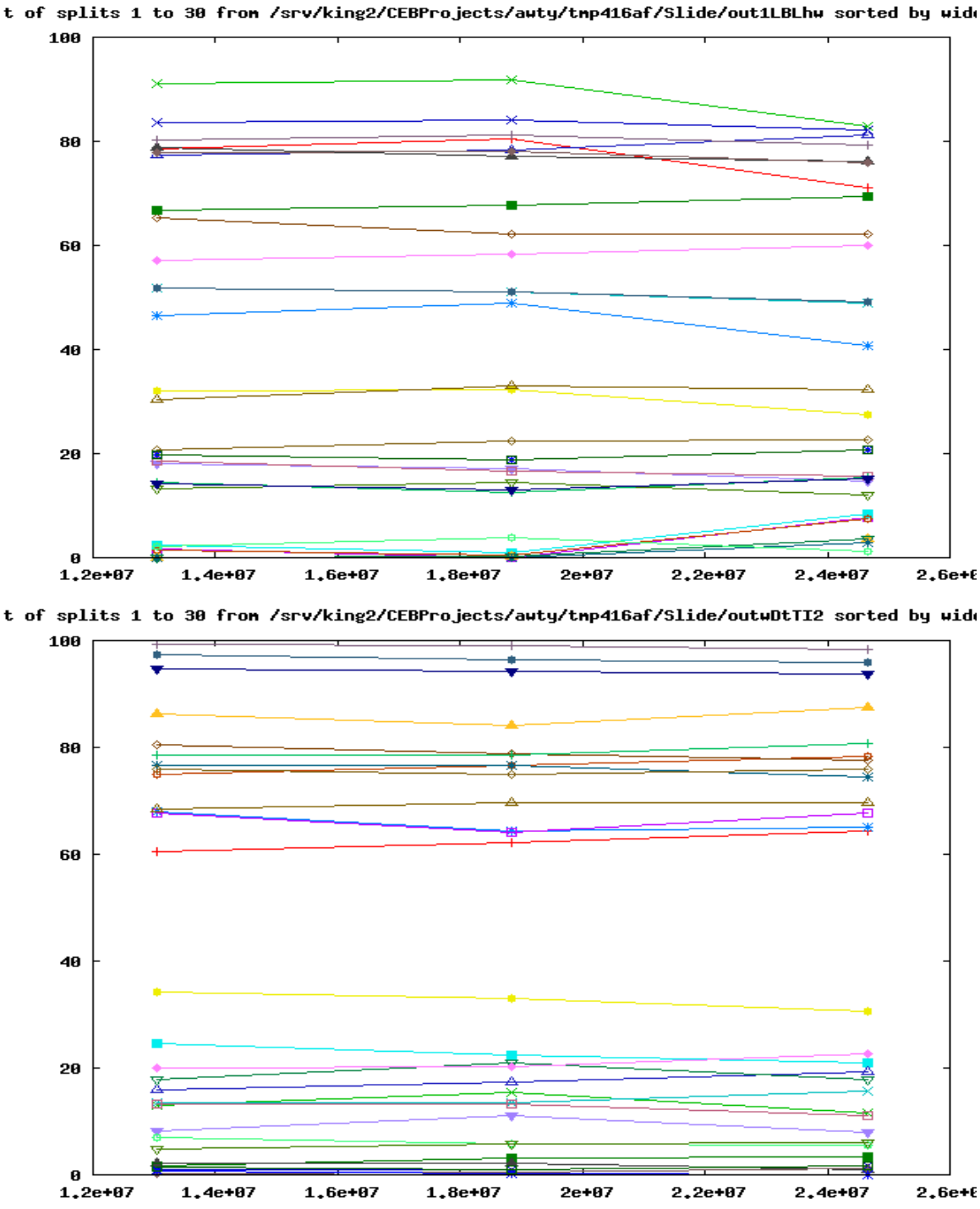



Figure 2.27: AWTY slide plot for first and second run of aureusidin synthase respectively

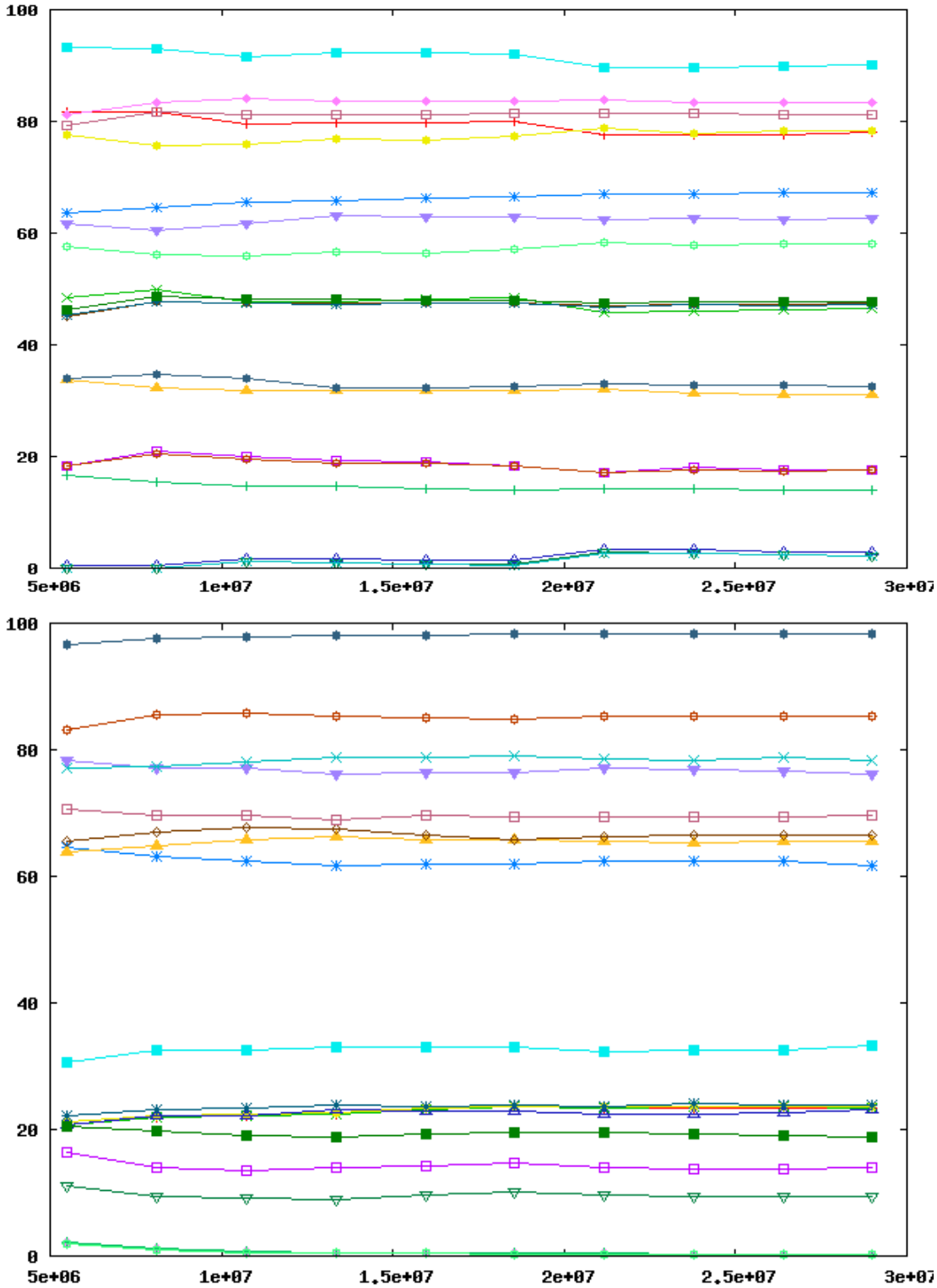



Figure 2.28: AWTY cumulative plot for first and second run of aureusidin synthase respectively

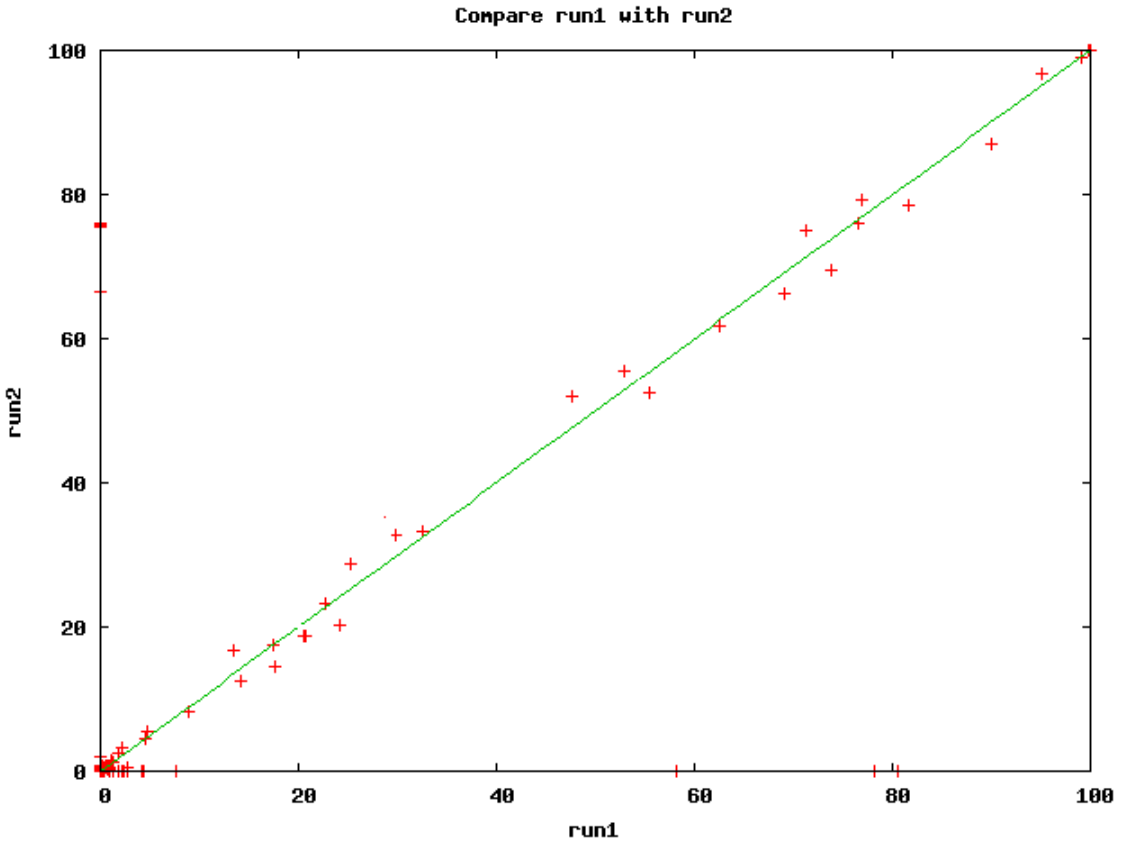



Figure 2.29: AWTY compare plot for first and second run of aureusidin synthase

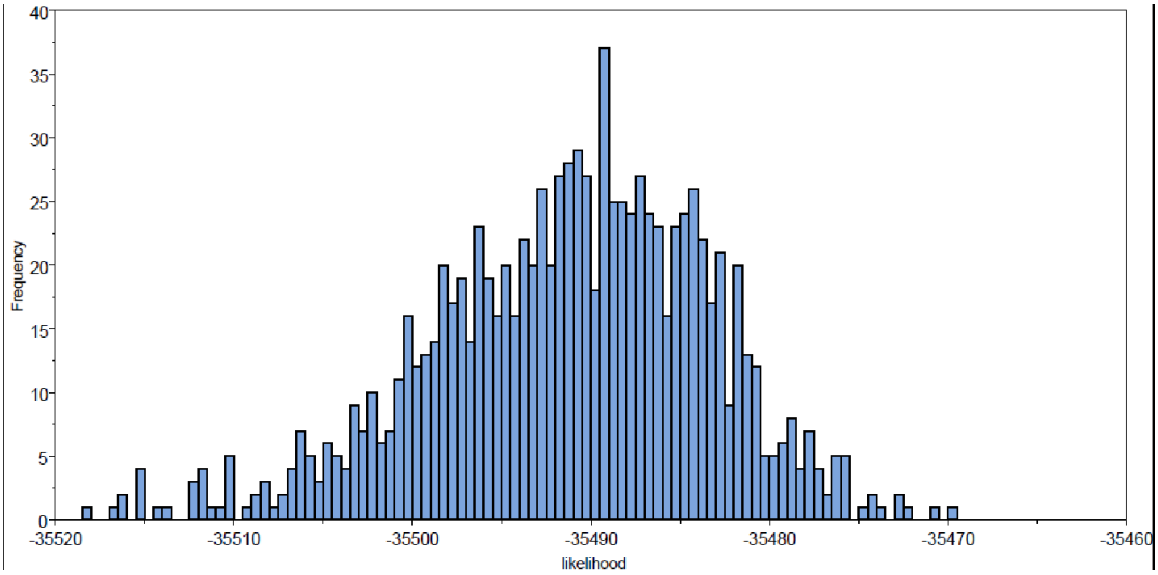



Figure 2.30: TRACER lnl estimates plot for *BEAST tree of aureusidin synthase

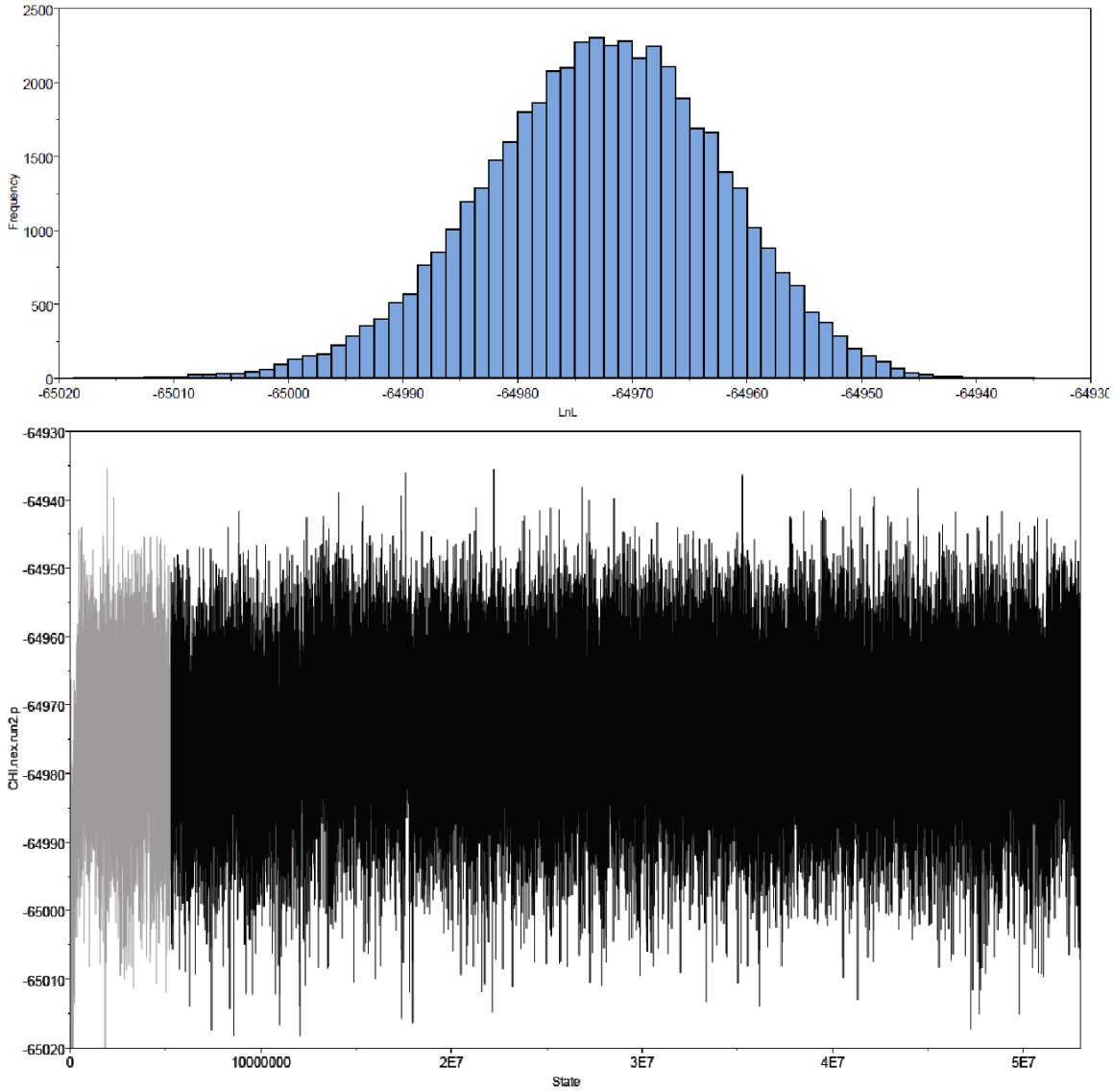



Figure 2.31: TRACER lnl estimates and trace plots for gene tree respectively in chalcone isomerase

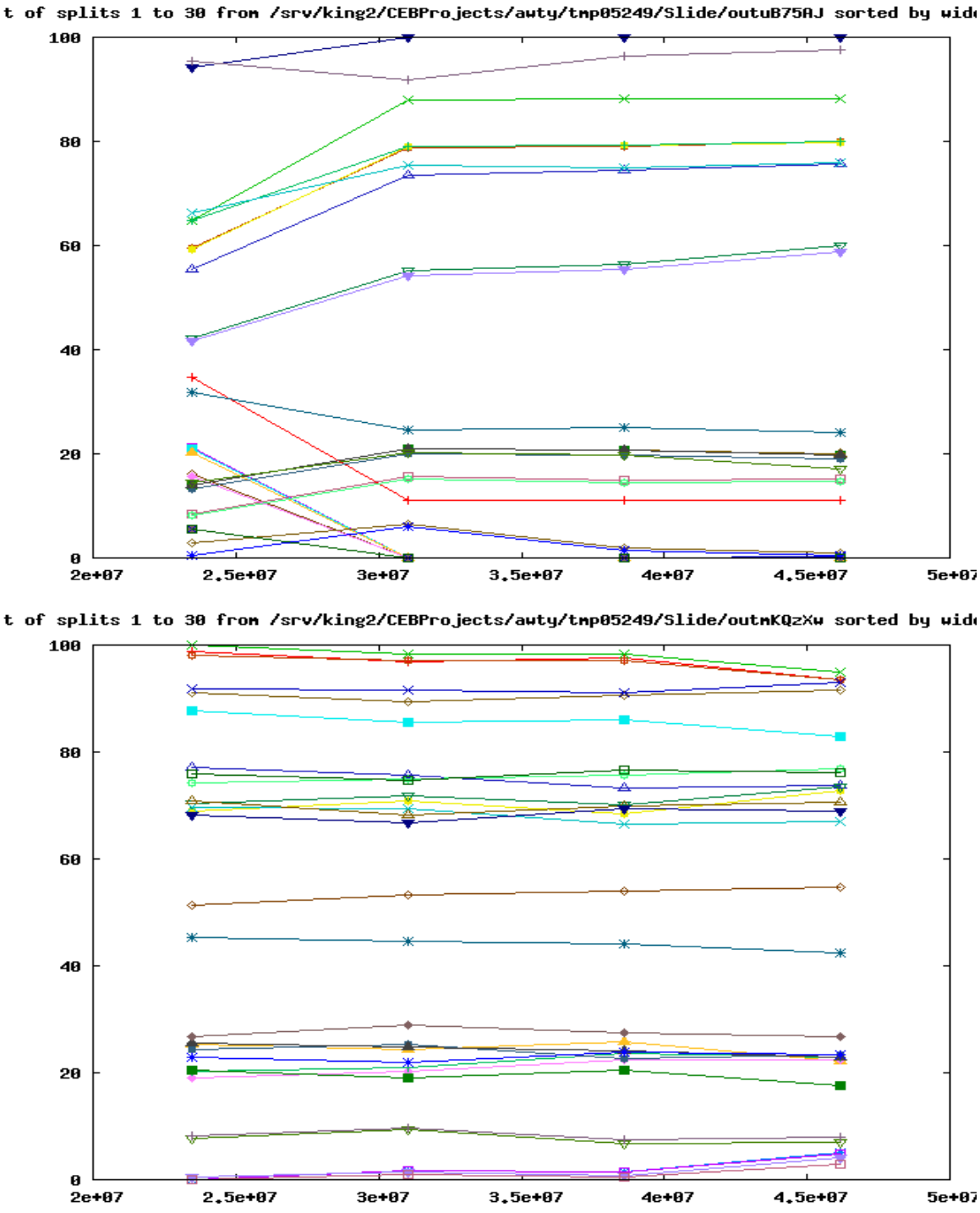



Figure 2.32: AWTY slide plot for first and second run of chalcone isomerase respectively

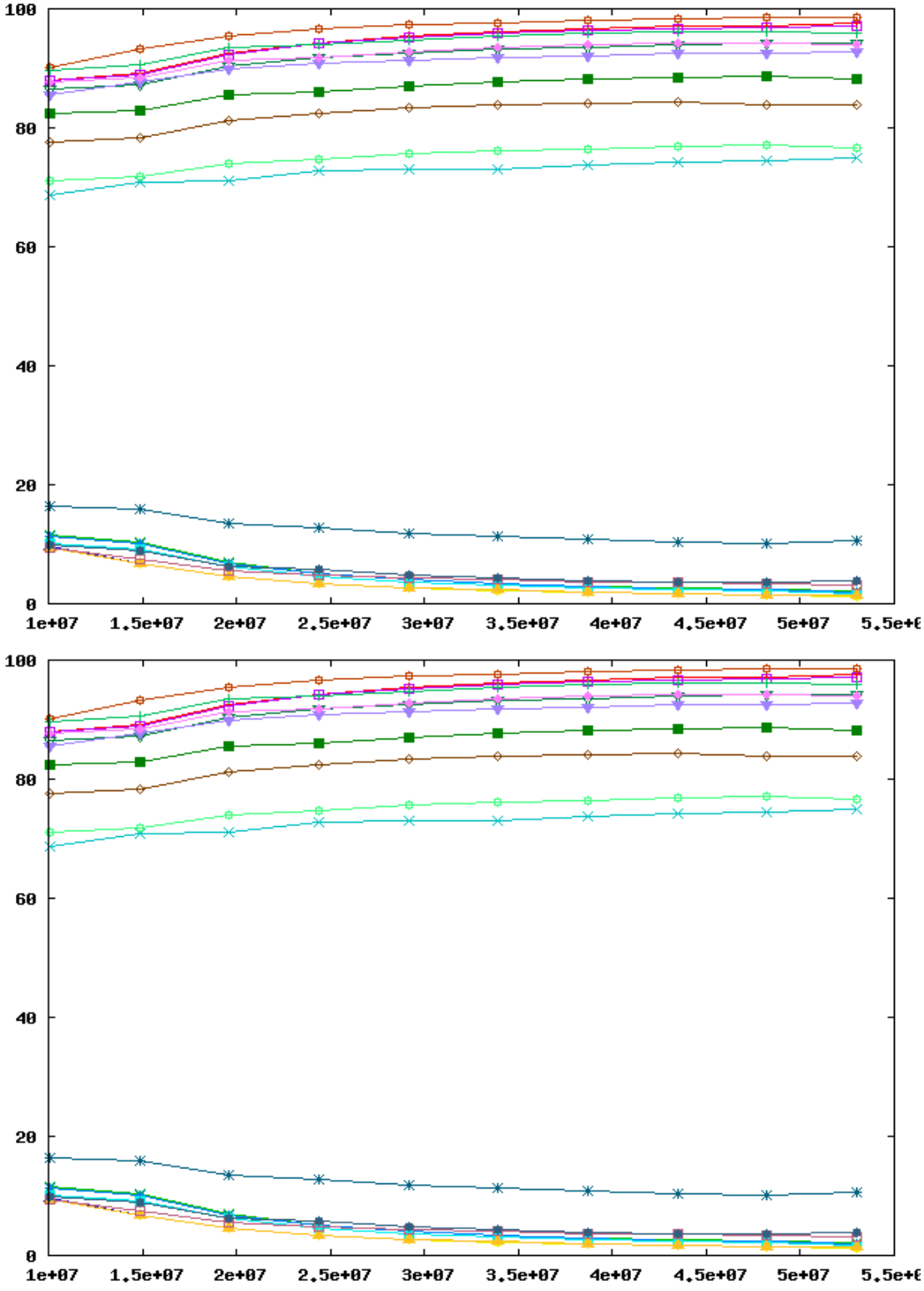



Figure 2.33: AWTY cumulative plot for first and second run of chalcone isomerase respectively

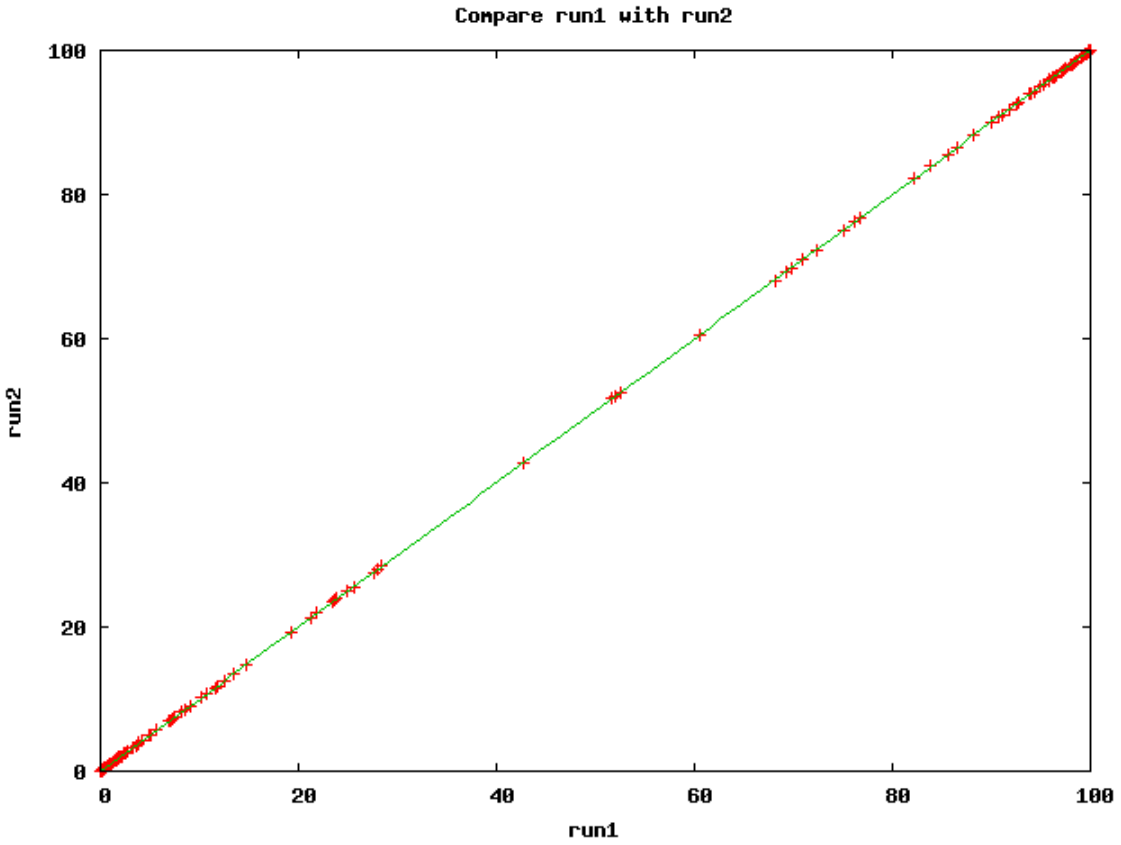



Figure 2.34: AWTY compare plot for first and second run of chalcone isomerase

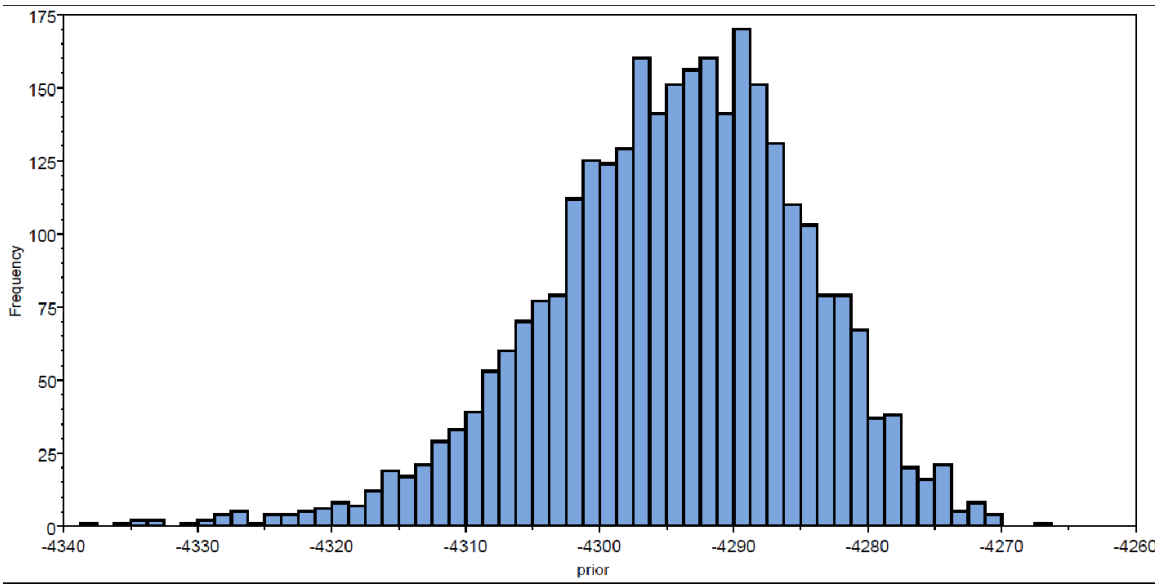



Figure 2.35: TRACER lnl estimates plot for *BEAST tree of chalcone isomerase

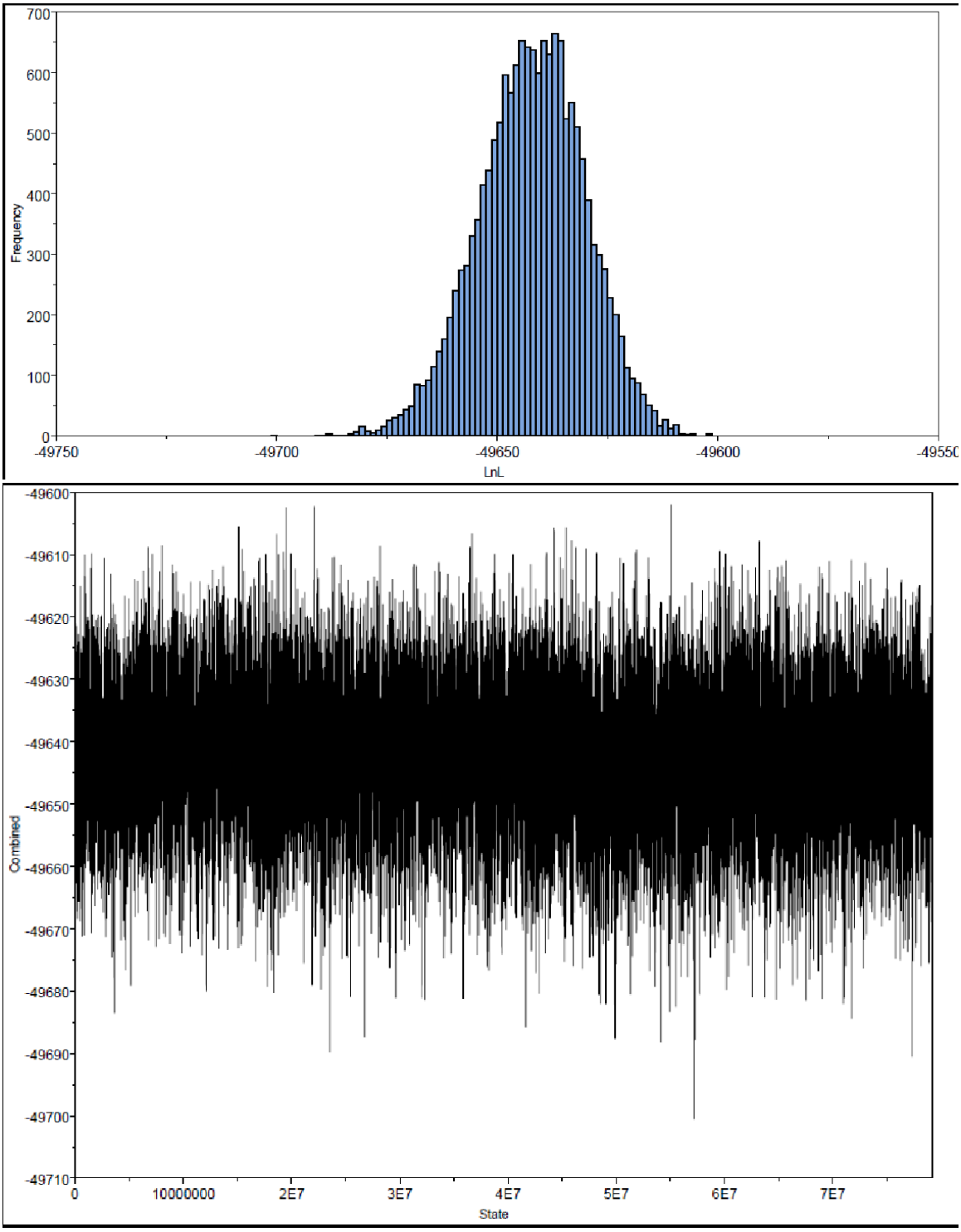



Figure 2.36: TRACER lnl estimates and trace plots for gene tree in chalcone synthase

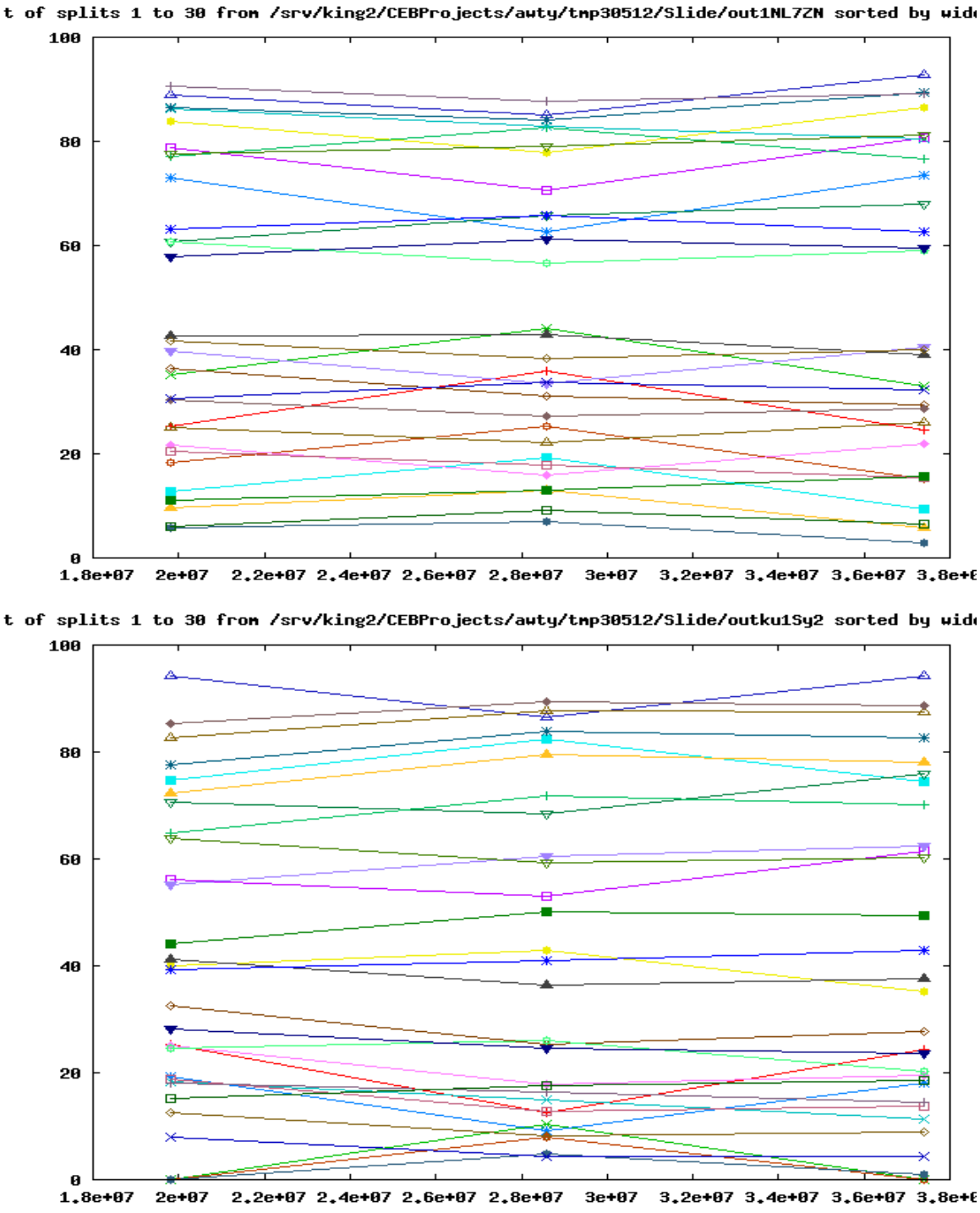



Figure 2.37: AWTY slide plot for first and second run of chalcone synthase respectively

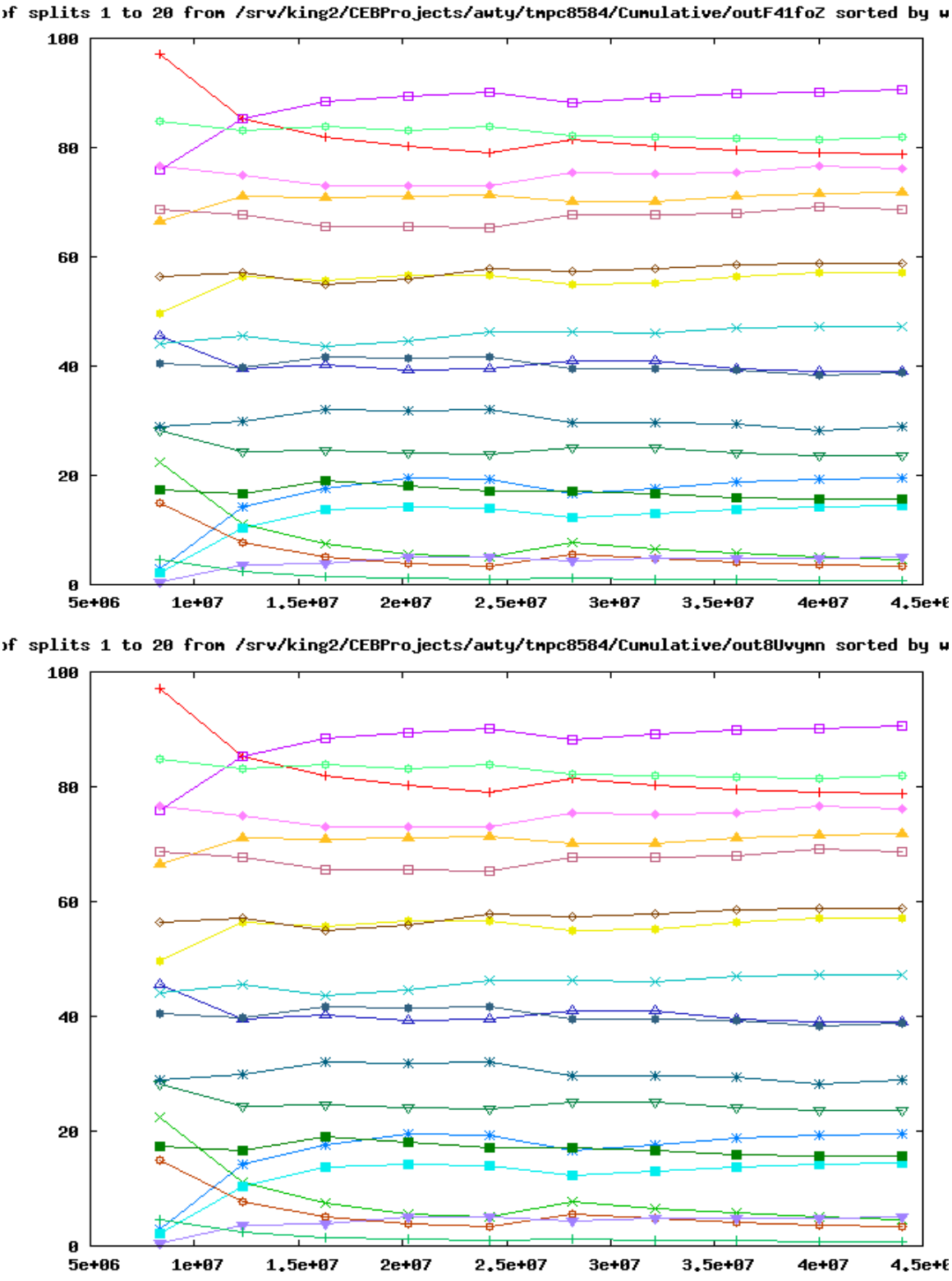



Figure 2.38: AWTY cumulative plot for first and second run of chalcone synthase respectively

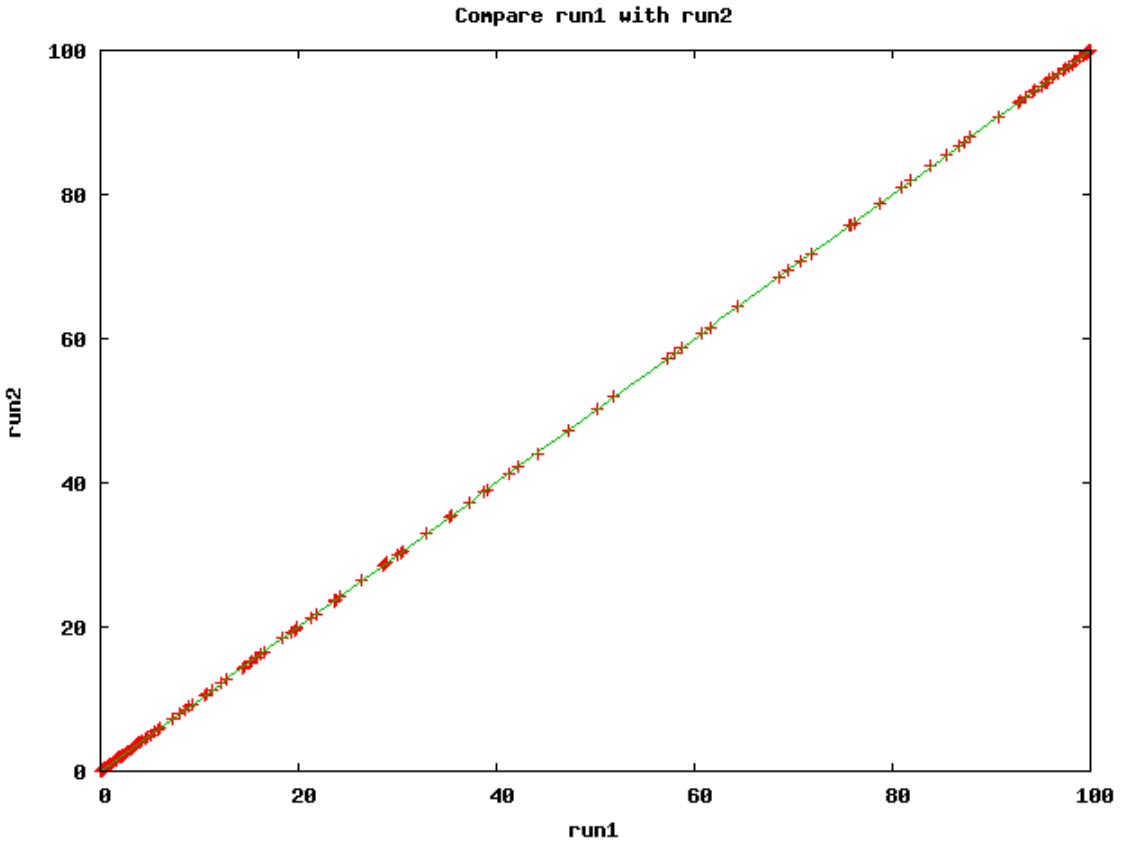



Figure 2.39: AWTY compare plot for first and second run of chalcone synthase

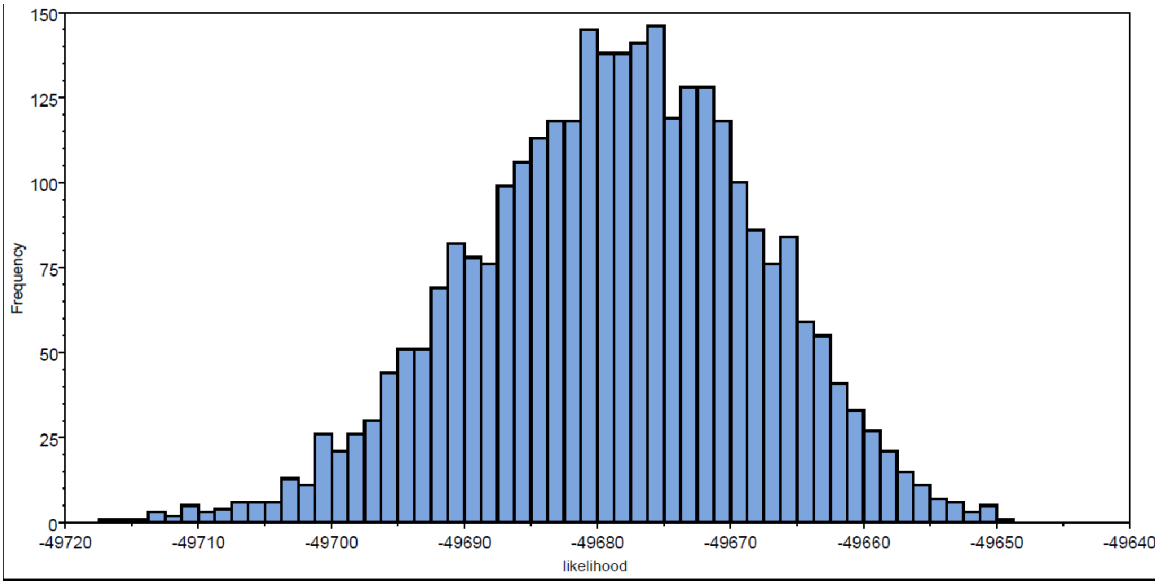



Figure 2.40: TRACER lnl estimates plot of *BEAST tree in chalcone synthase

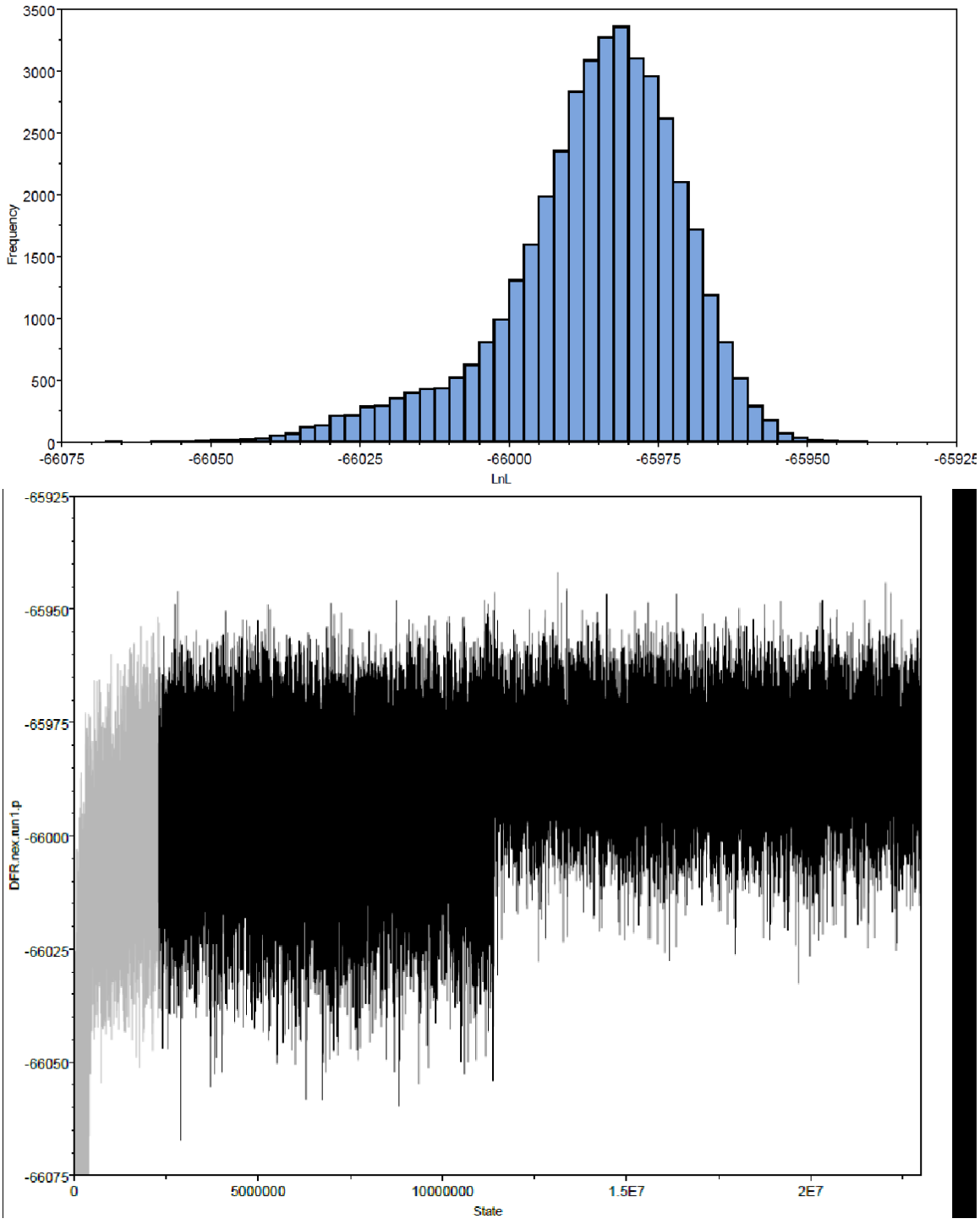



Figure 2.41: TRACER lnl estimates and trace plot of gene tree for Dihydroflavonol 4-Reductase

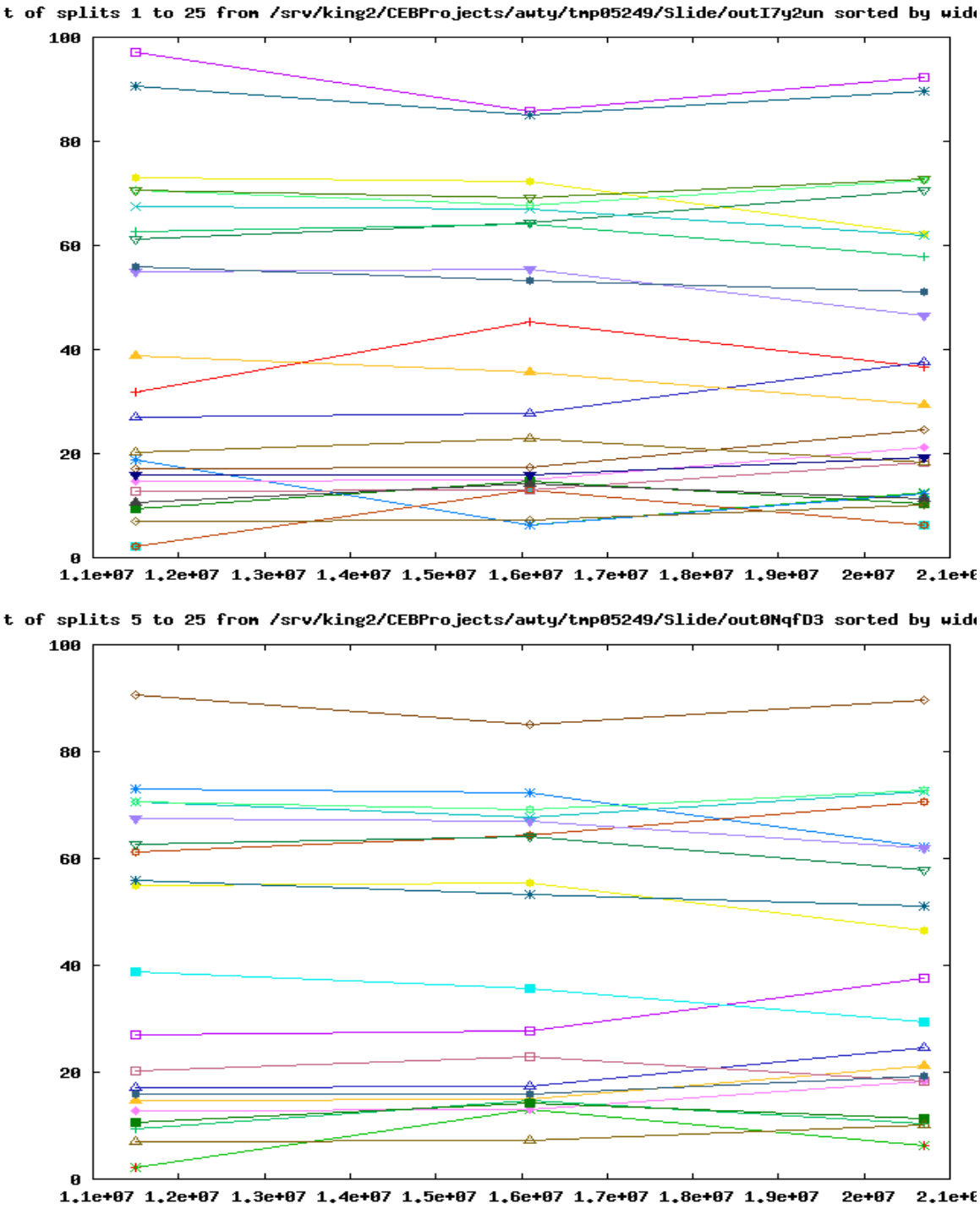



Figure 2.42: AWTY slide plot for first and second run of Dihydroflavonol 4-Reductase respectively

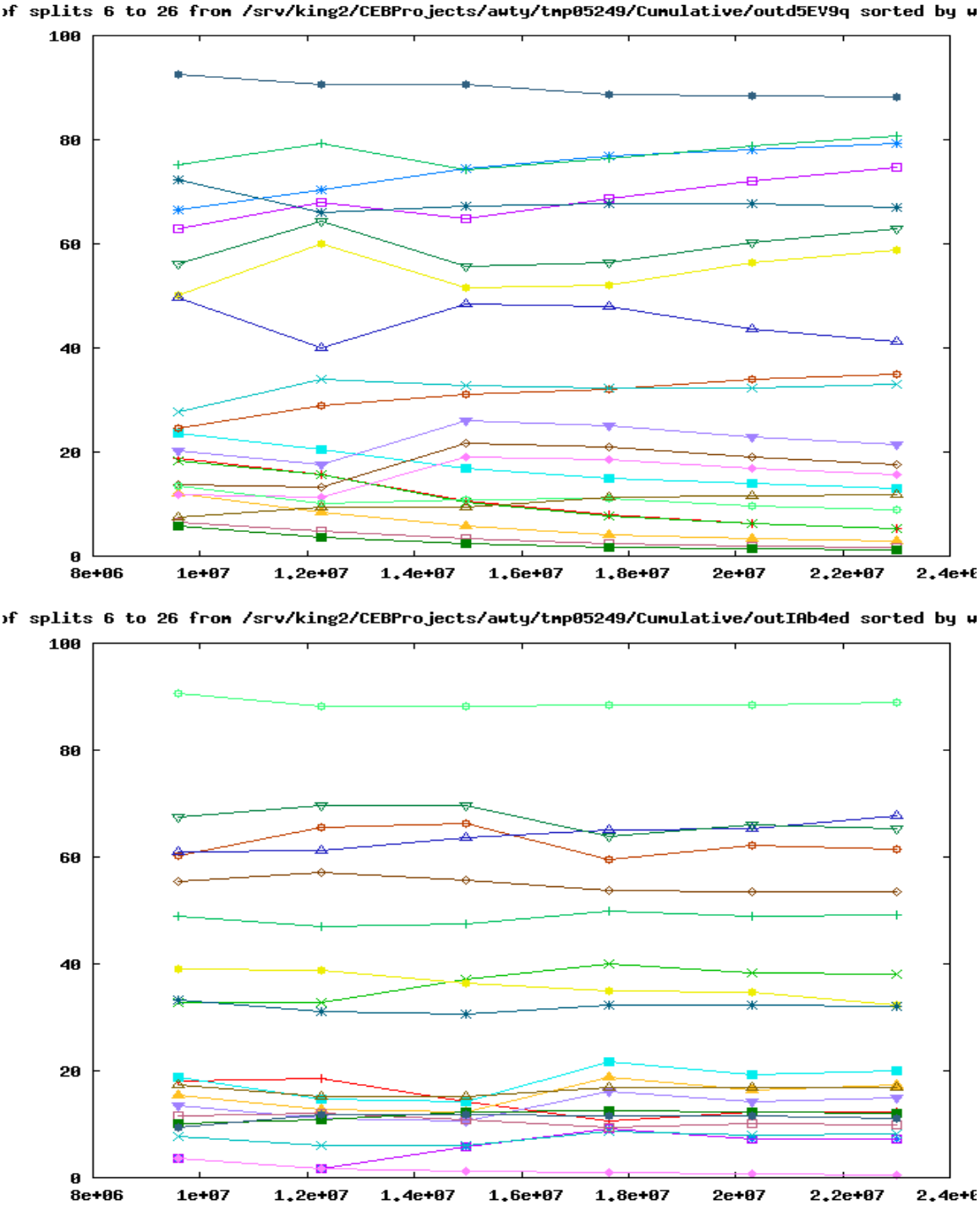



Figure 2.43: AWTY cumulative plot for first and second run of Dihydroflavonol 4-Reductase respectively

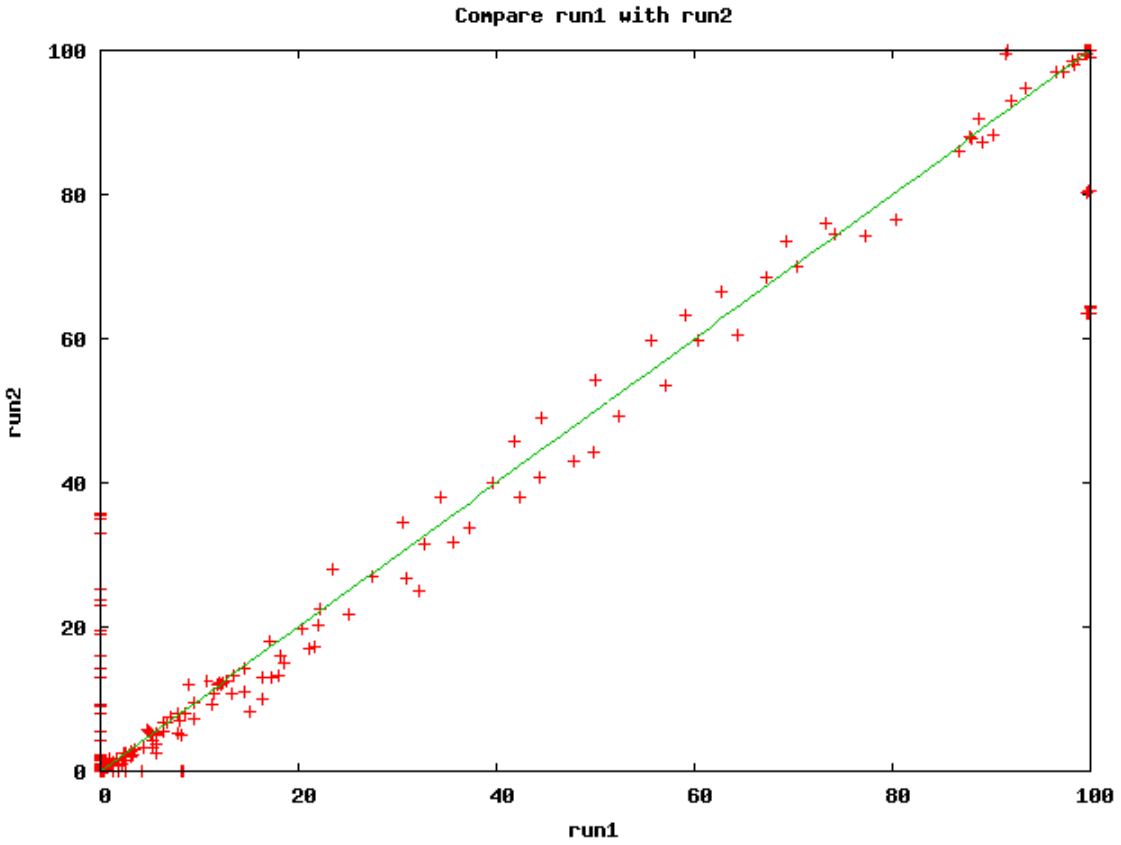



Figure 2.44: AWTY compare plot for first and second run of Dihydroflavonol 4-Reductase

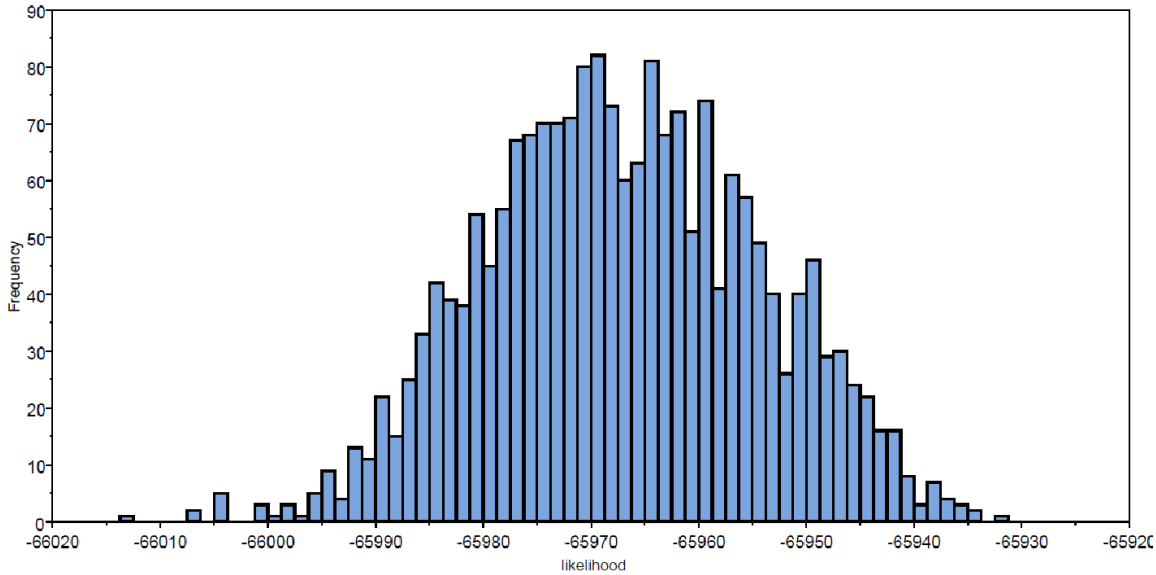



Figure 2.45: TRACER lnl estimates plot for *BEAST tree in Dihydroflavonol 4-Reductase

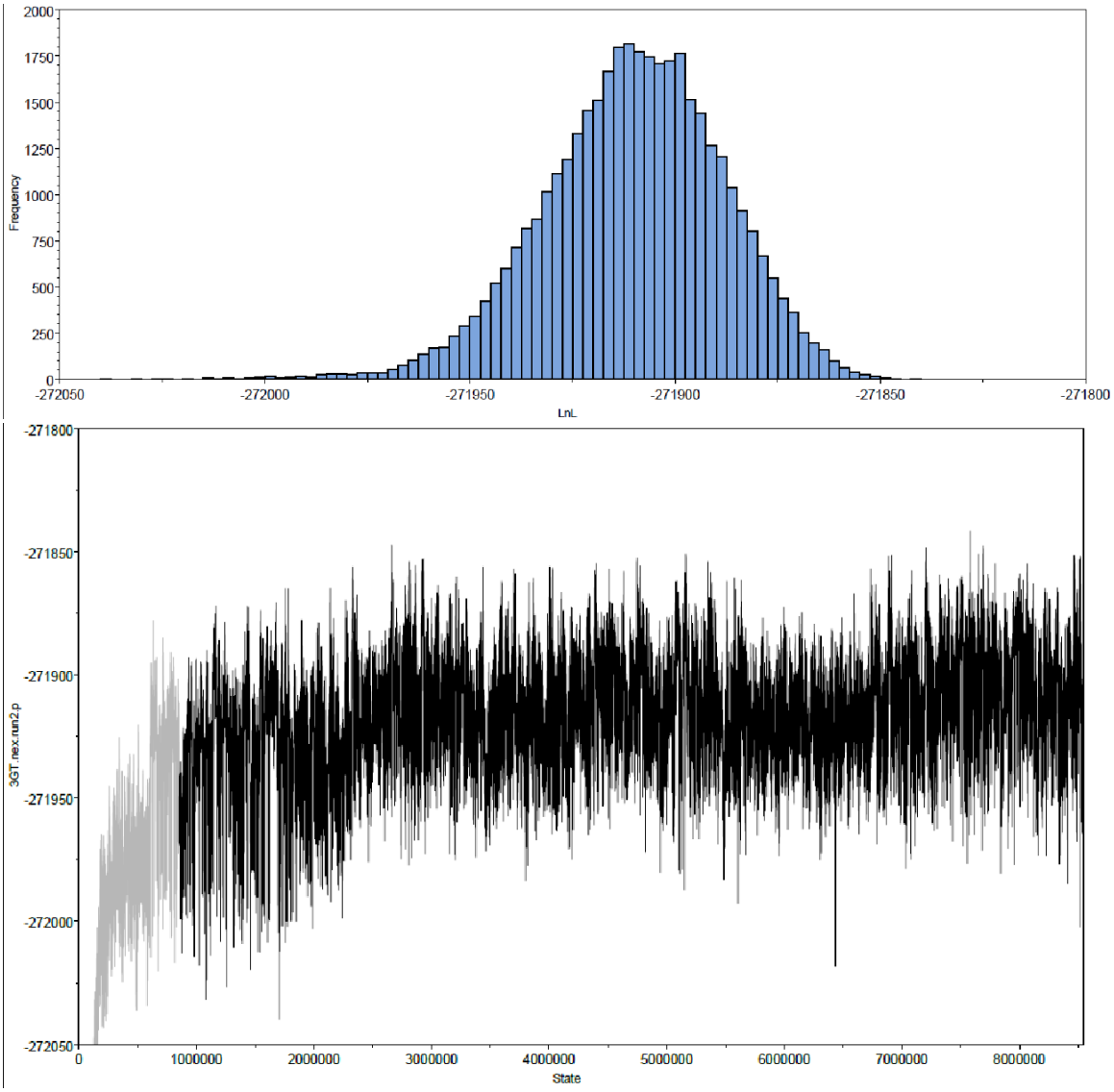



Figure 2.46: TRACER lnl estimates and trace plots for gene tree in anthocyanidin 3-O-Glucosyltransferase

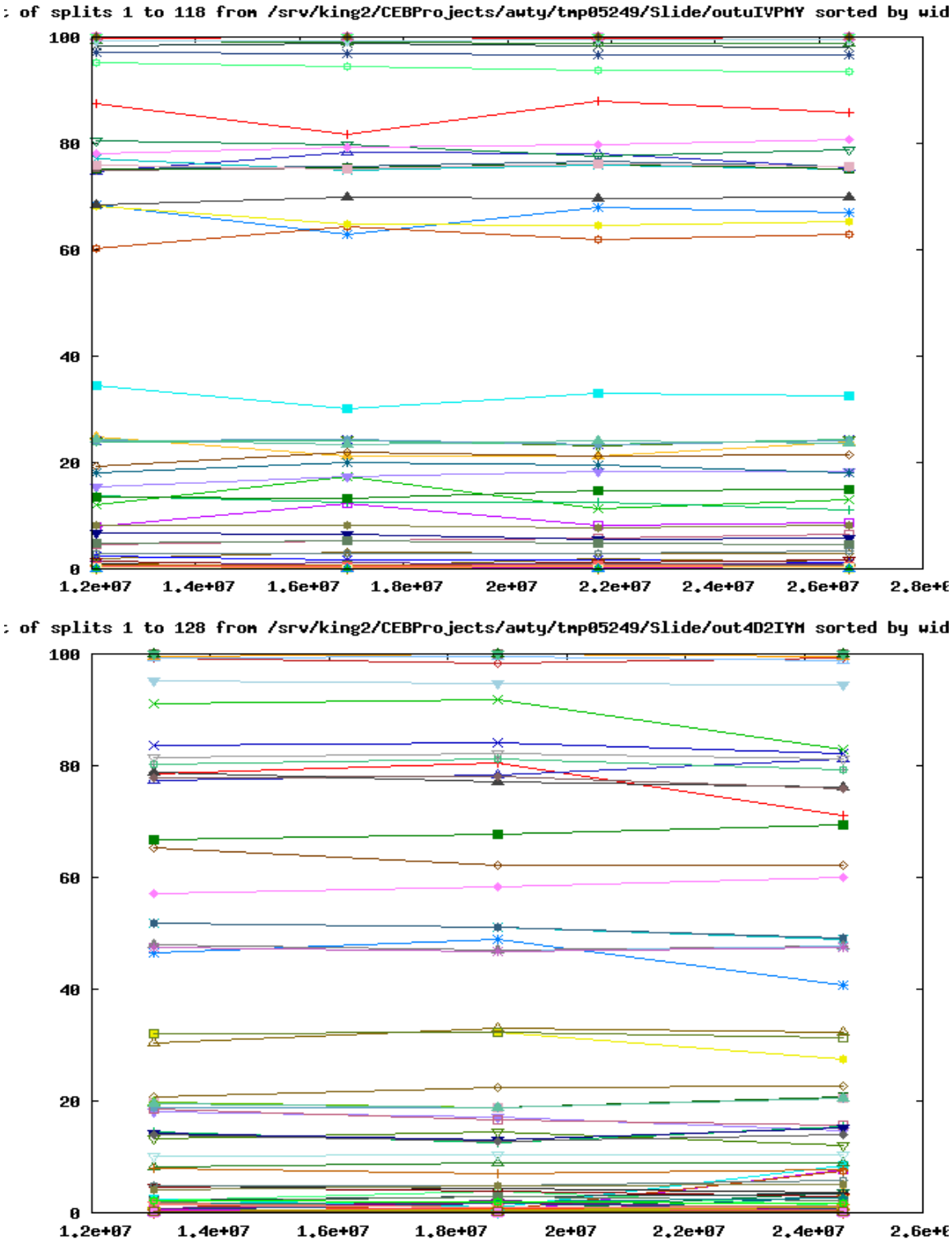



Figure 2.47: AWTY slide plot for first and second run of anthocyanidin 3-O-Glucosyltransferase respectively

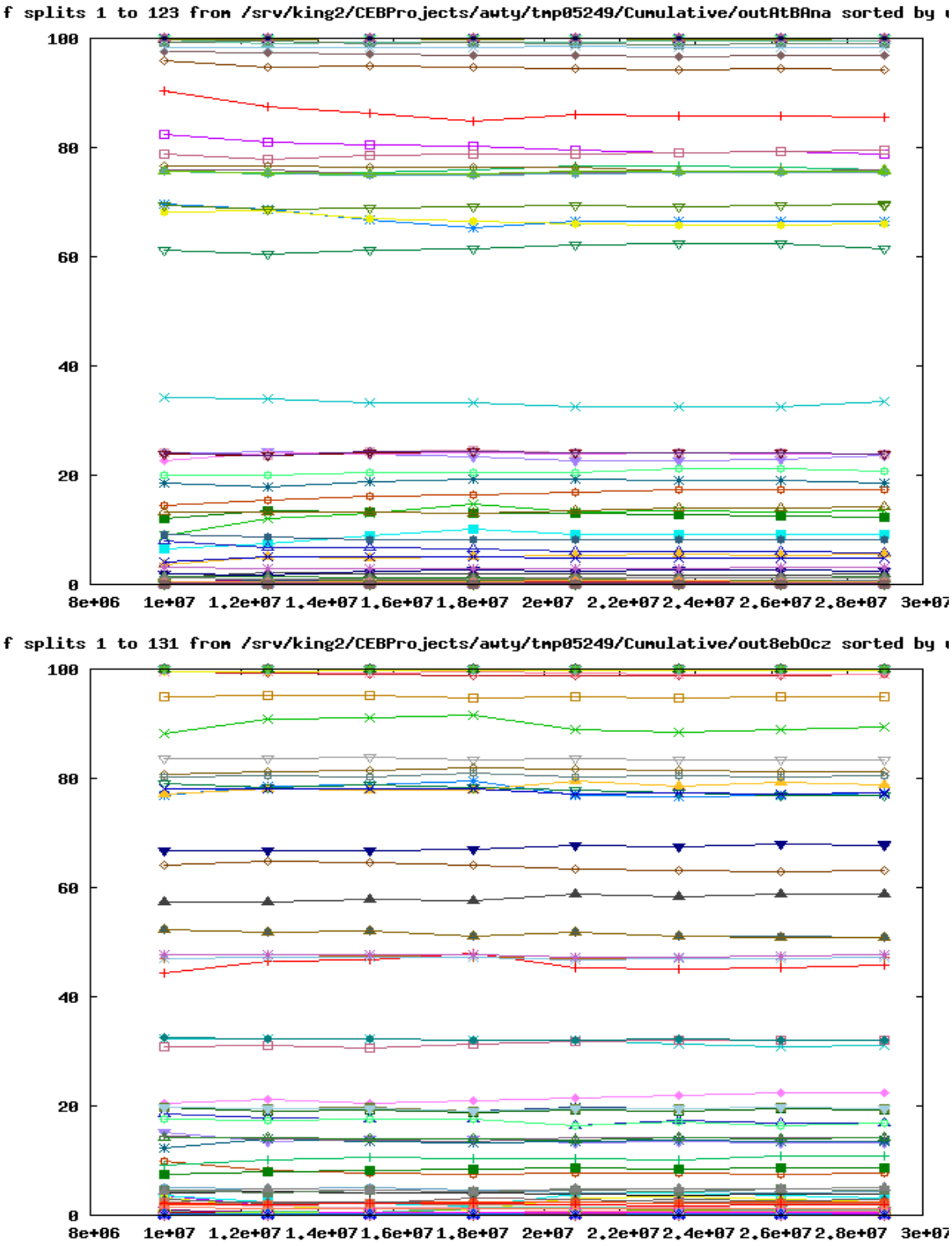



Figure 2.48: AWTY cumulative plot for first and second run of anthocyanidin 3-O-Glucosyltransferase respectively

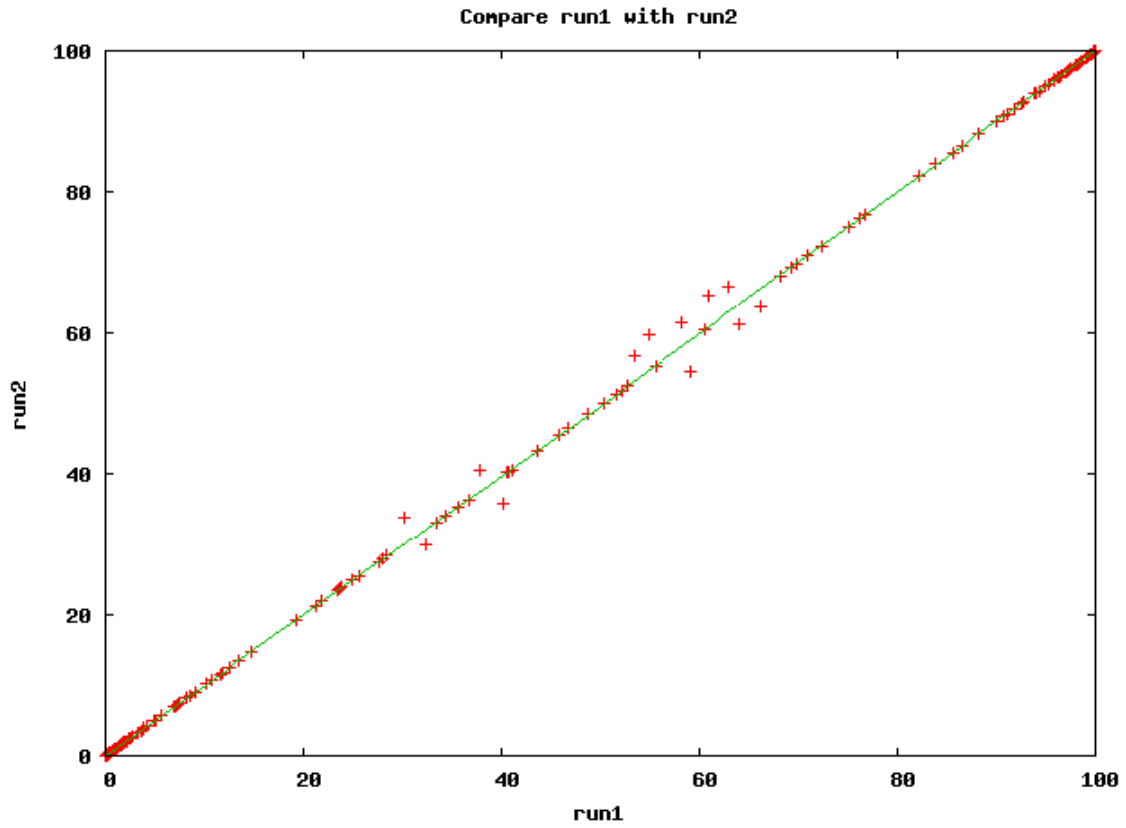



Figure 2.49: AWTY compare plot for first and second run of anthocyanidin 3-O-Glucosyltransferase

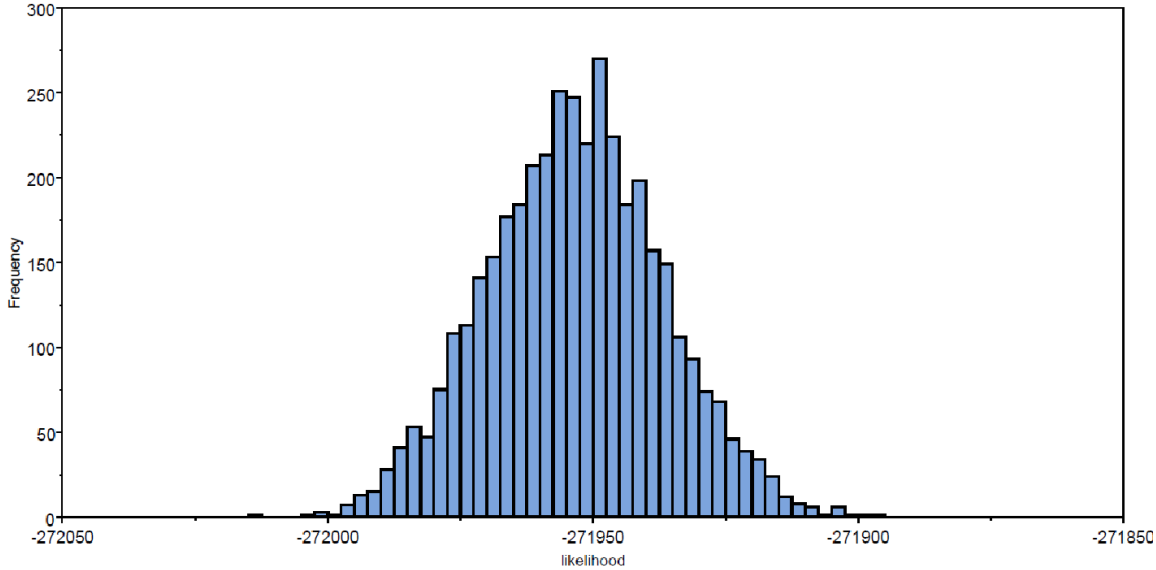



Figure 2.50: TRACER lnl estimates plot of *BEAST tree in anthocyanidin 3-O-Glucosyltransferase

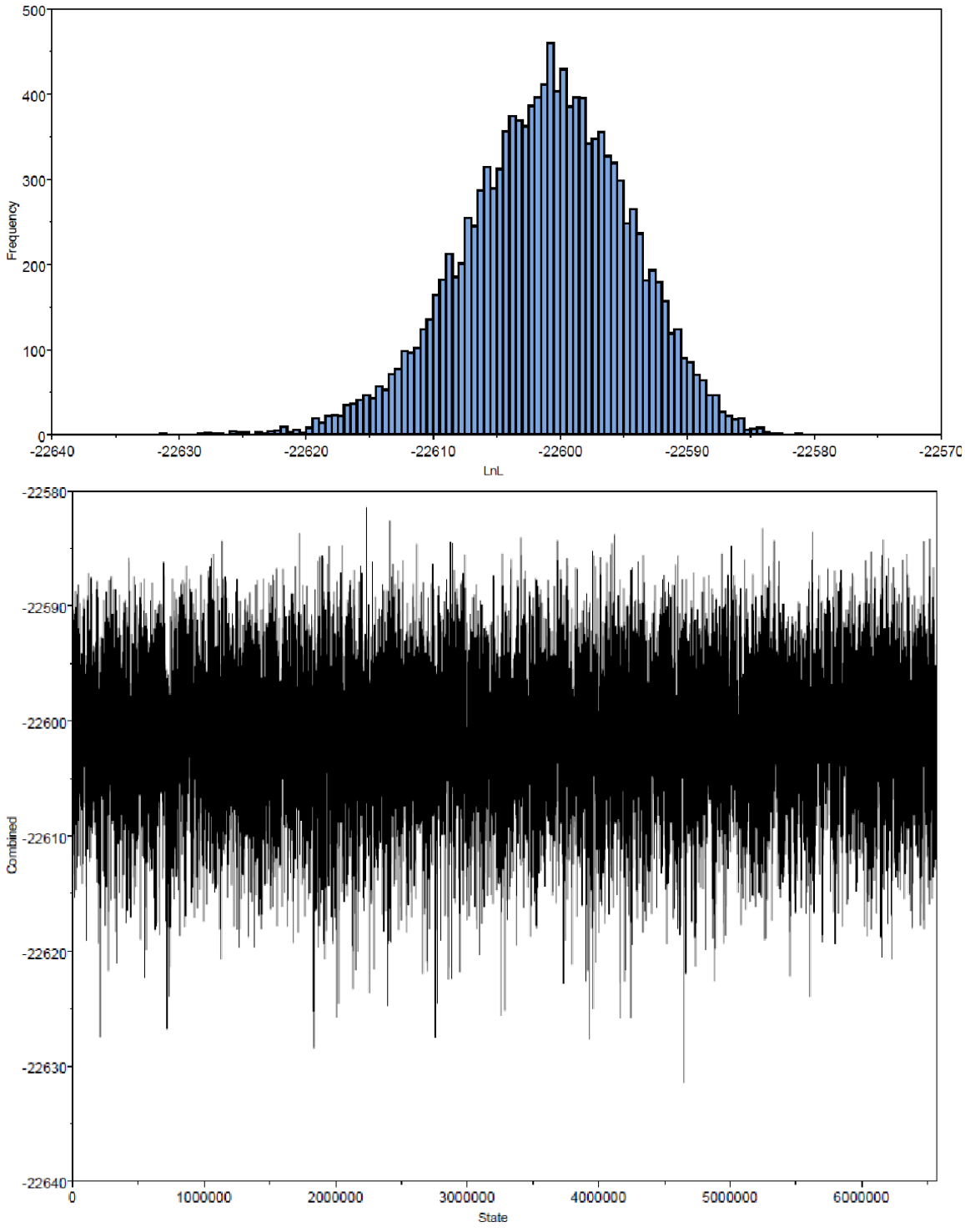



Figure 2.51: TRACER lnl estimates and trace plots for gene tree in anthocyanidin reductase

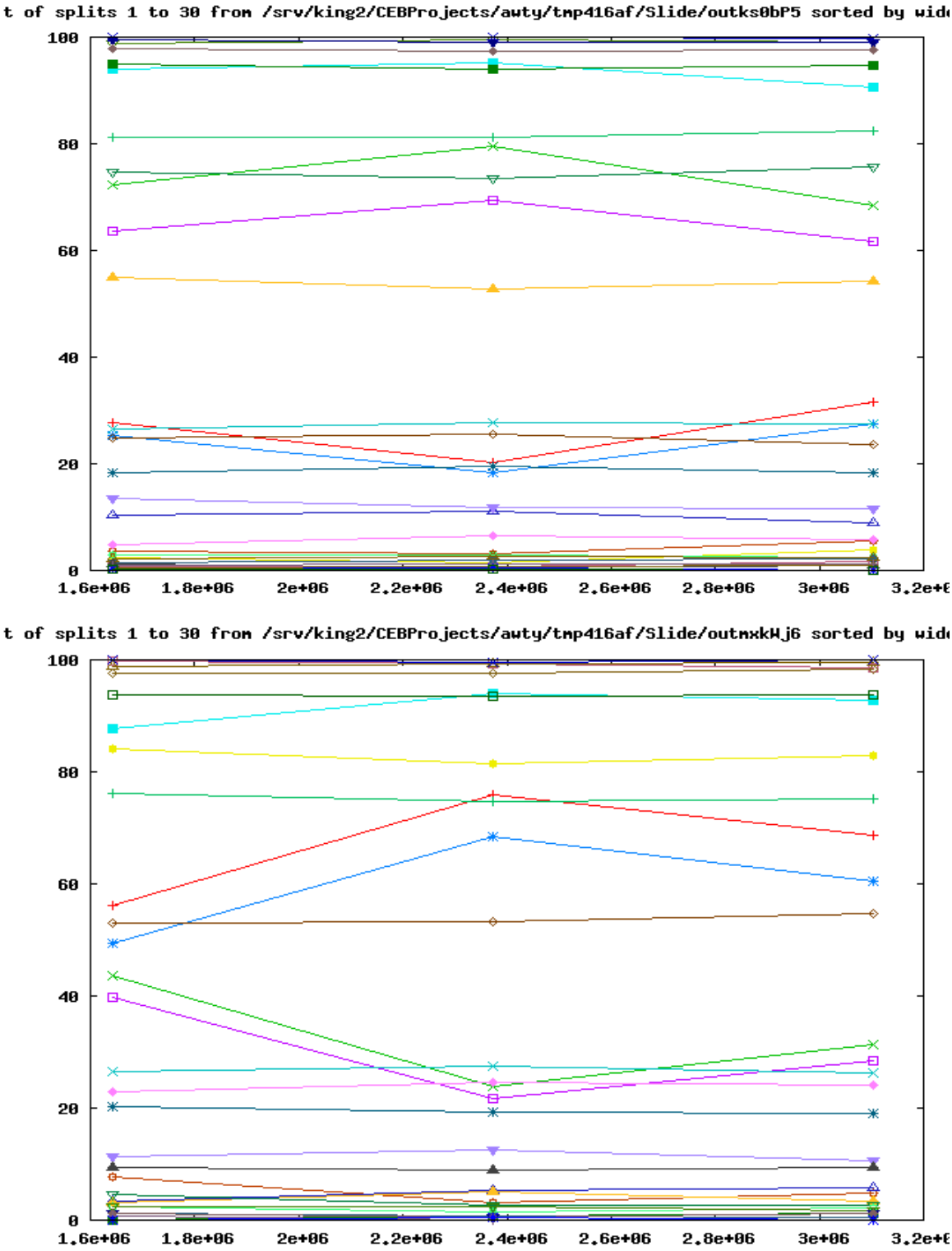



Figure 2.52: AWTY slide plot for first and second run of anthocyanidin reductase respectively

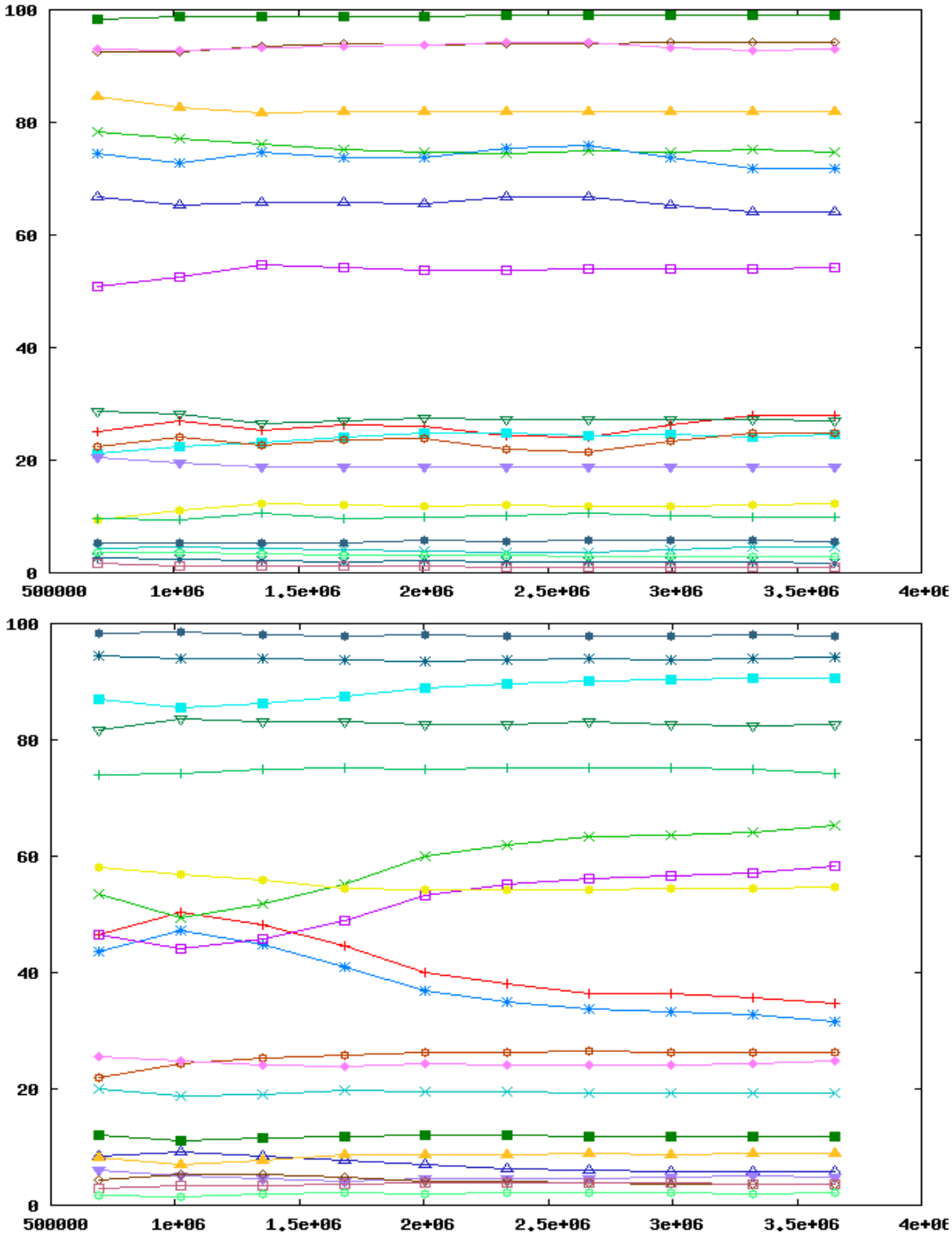



Figure 2.53: AWTY cumulative plot for first and second run of anthocyanidin reductase respectively

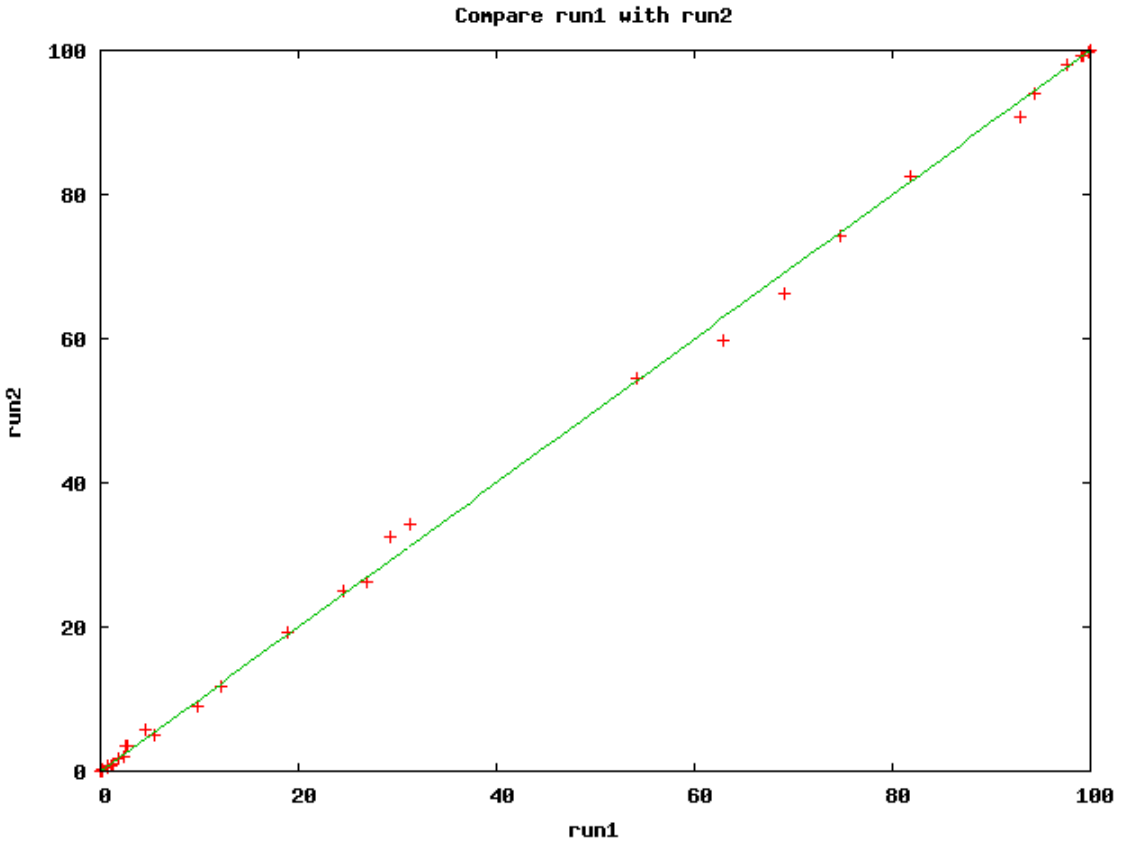



Figure 2.54: AWTY compare plot for first and second run of anthocyanidin reductase

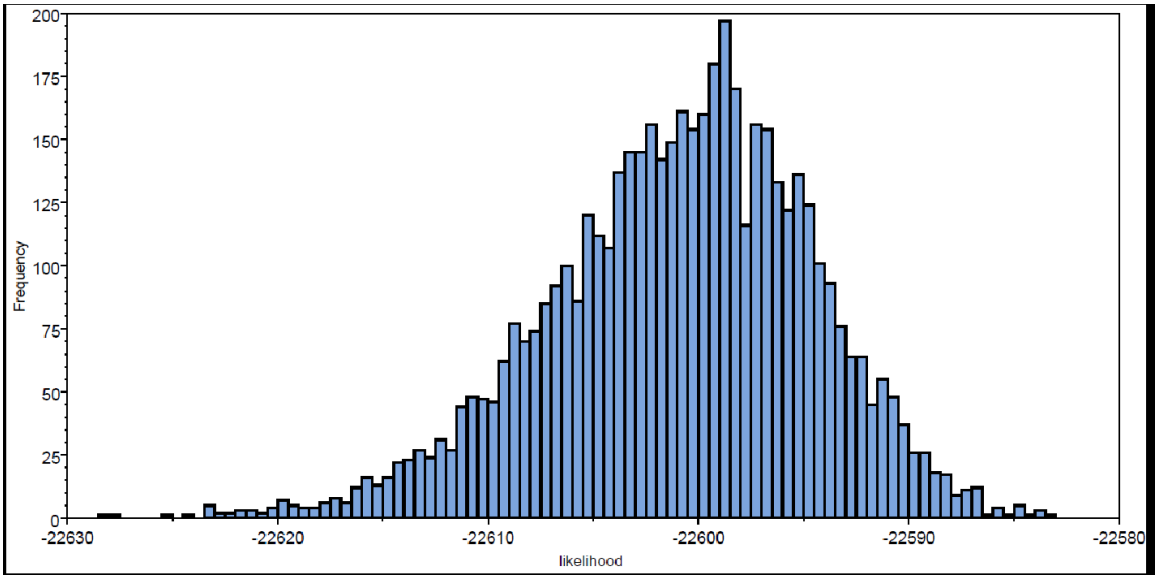



Figure 2.55: TRACER lnl estimates plot of *BEAST tree in anthocyanidin reductase

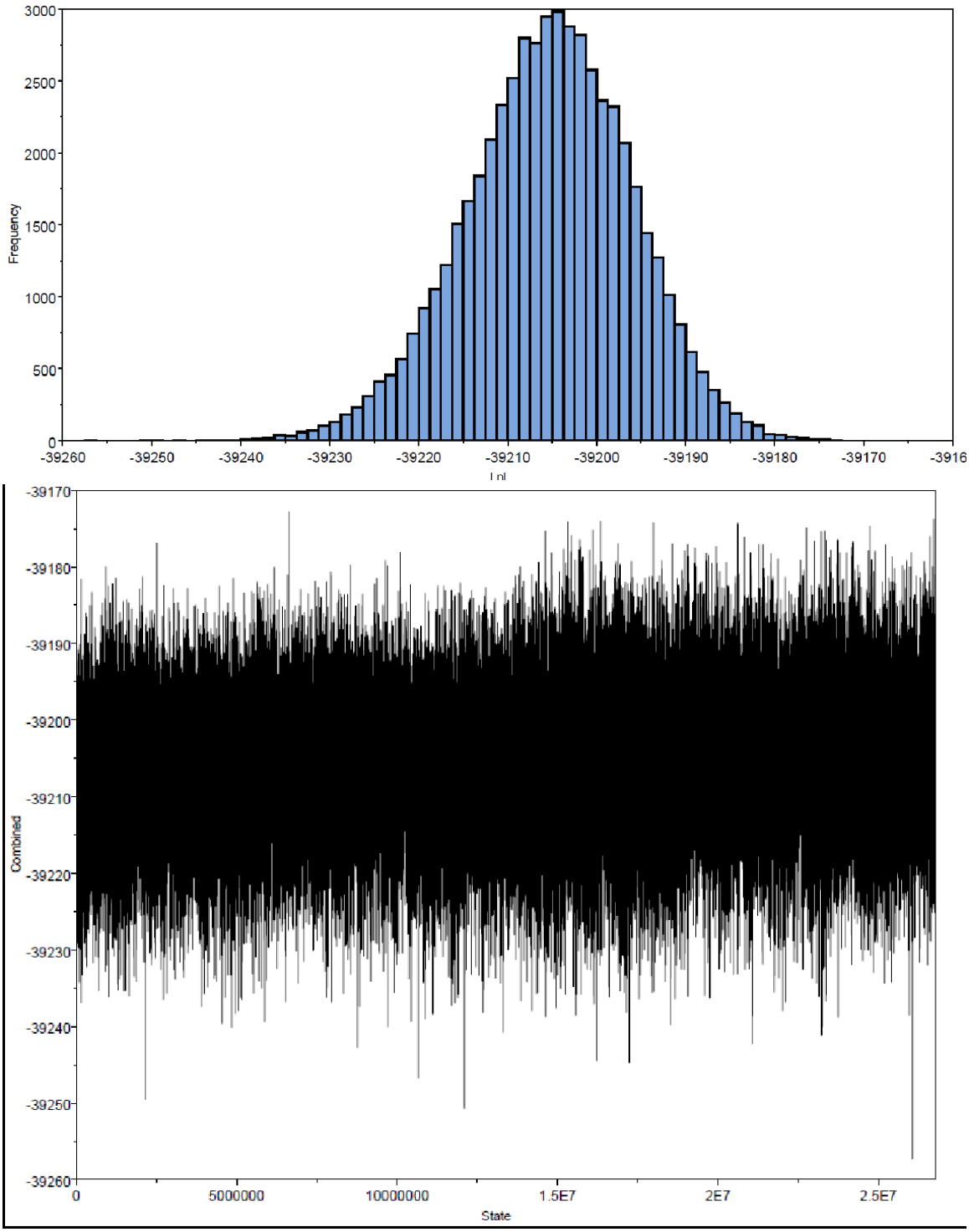



Figure 2.56: TRACER lnl estimates and trace plots of gene tree of Flavanone 3-Hydroxylase respectively

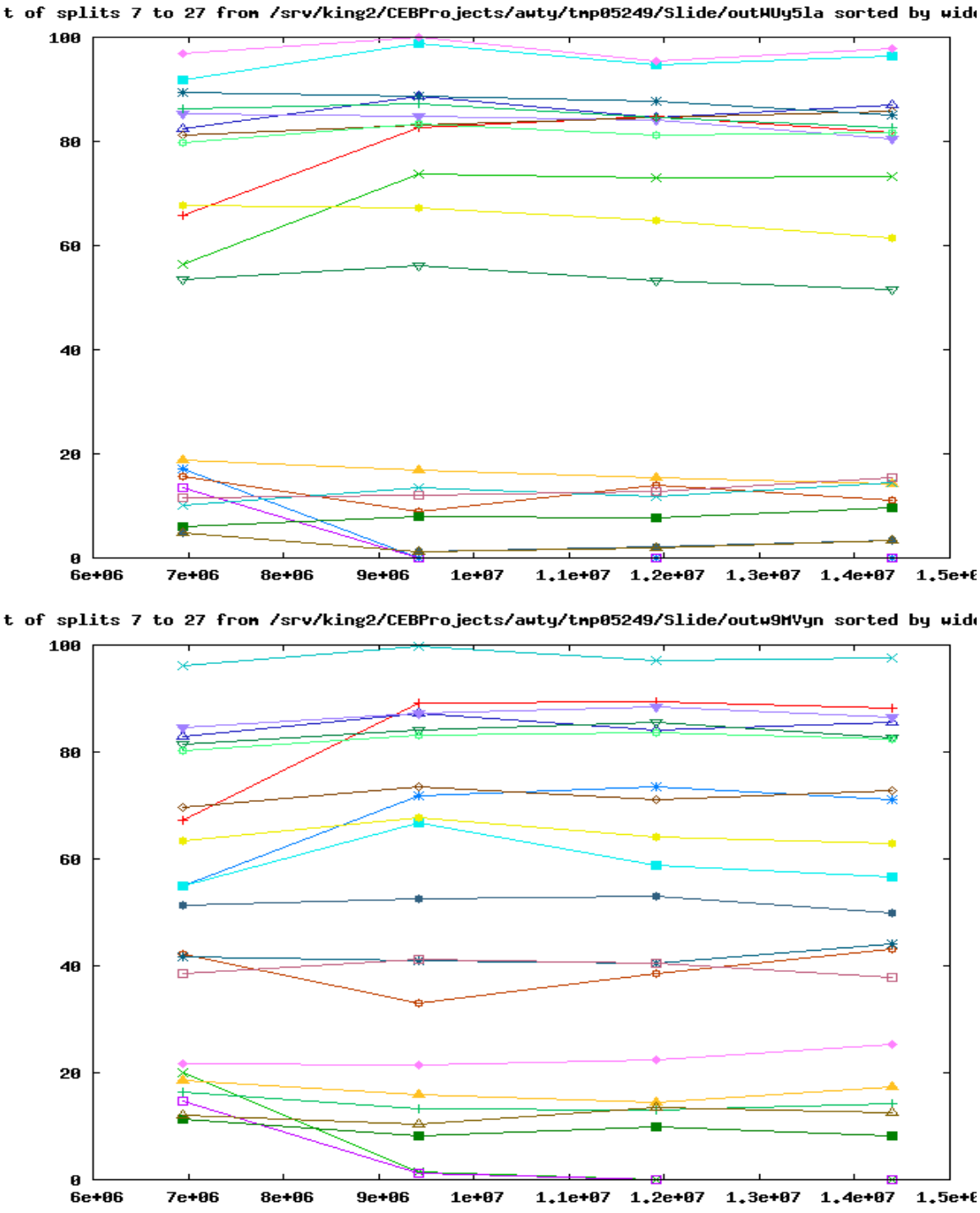



Figure 2.57: AWTY slide plot for first and second run of Flavanone 3-Hydroxylase respectively

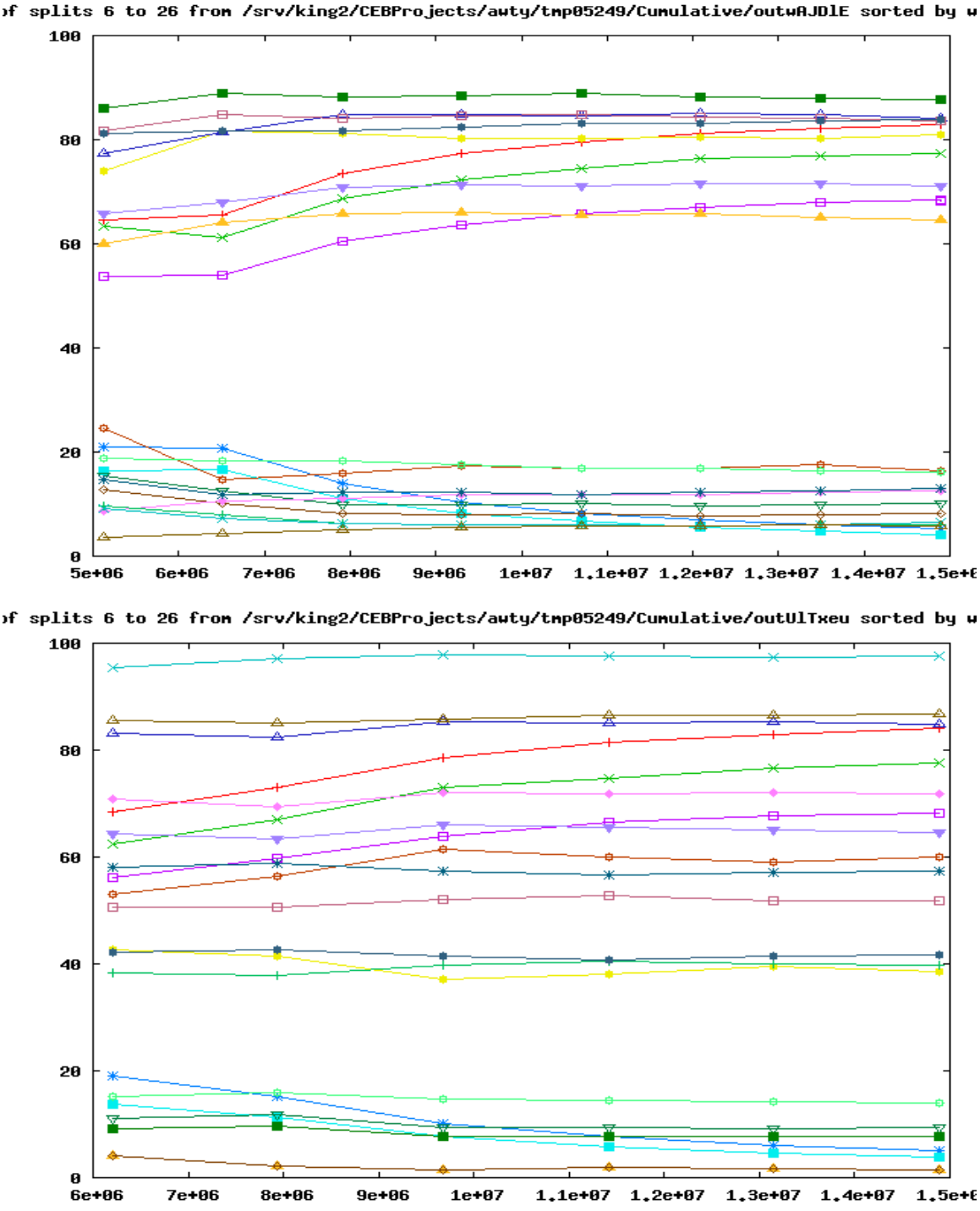



Figure 2.58: AWTY cumulative plot for first and second run of Flavanone 3-Hydroxylase respectively

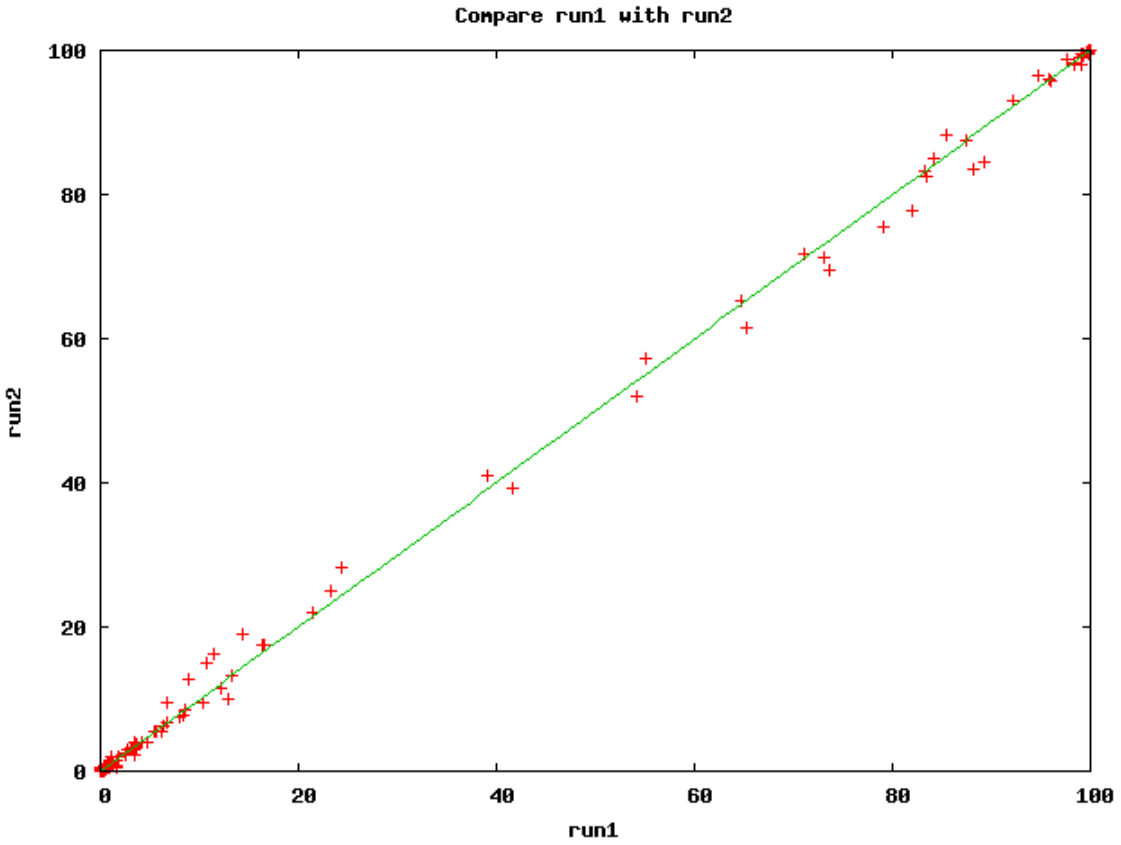



Figure 2.59: AWTY compare plot for first and second run of Flavanone 3-Hydroxylase

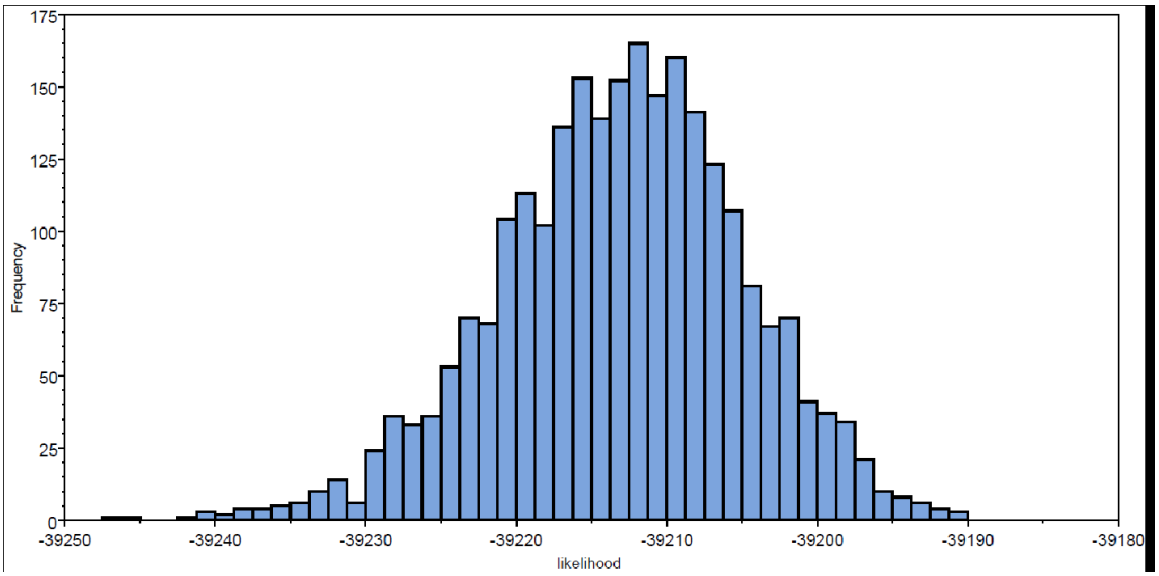



Figure 2.60: TRACER lnl estimates plot of *BEAST tree for Flavanone 3-Hydroxylase

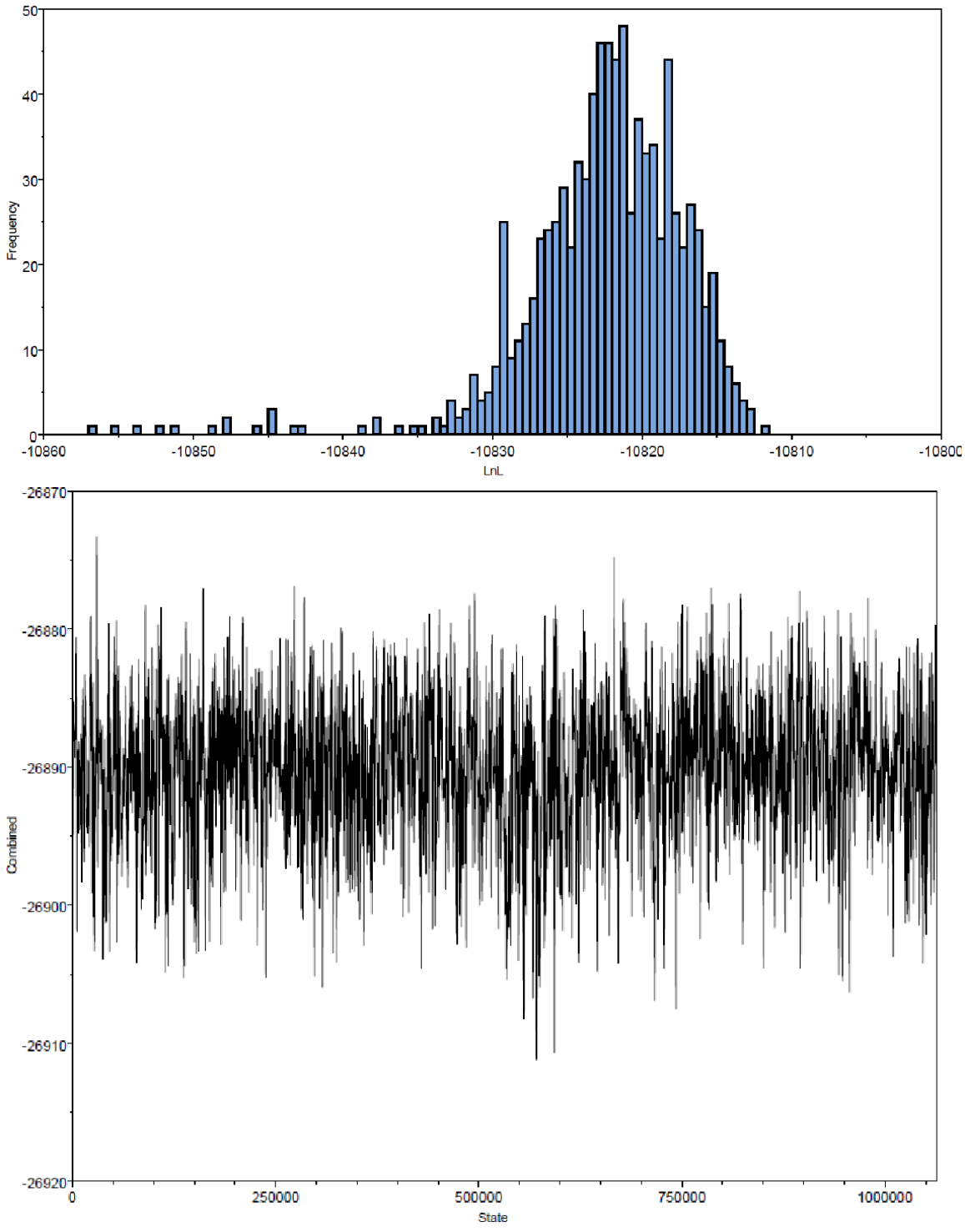



Figure 2.61: TRACER lnl estimates and trace plots of gene tree for leucoanthocyanidin reductase

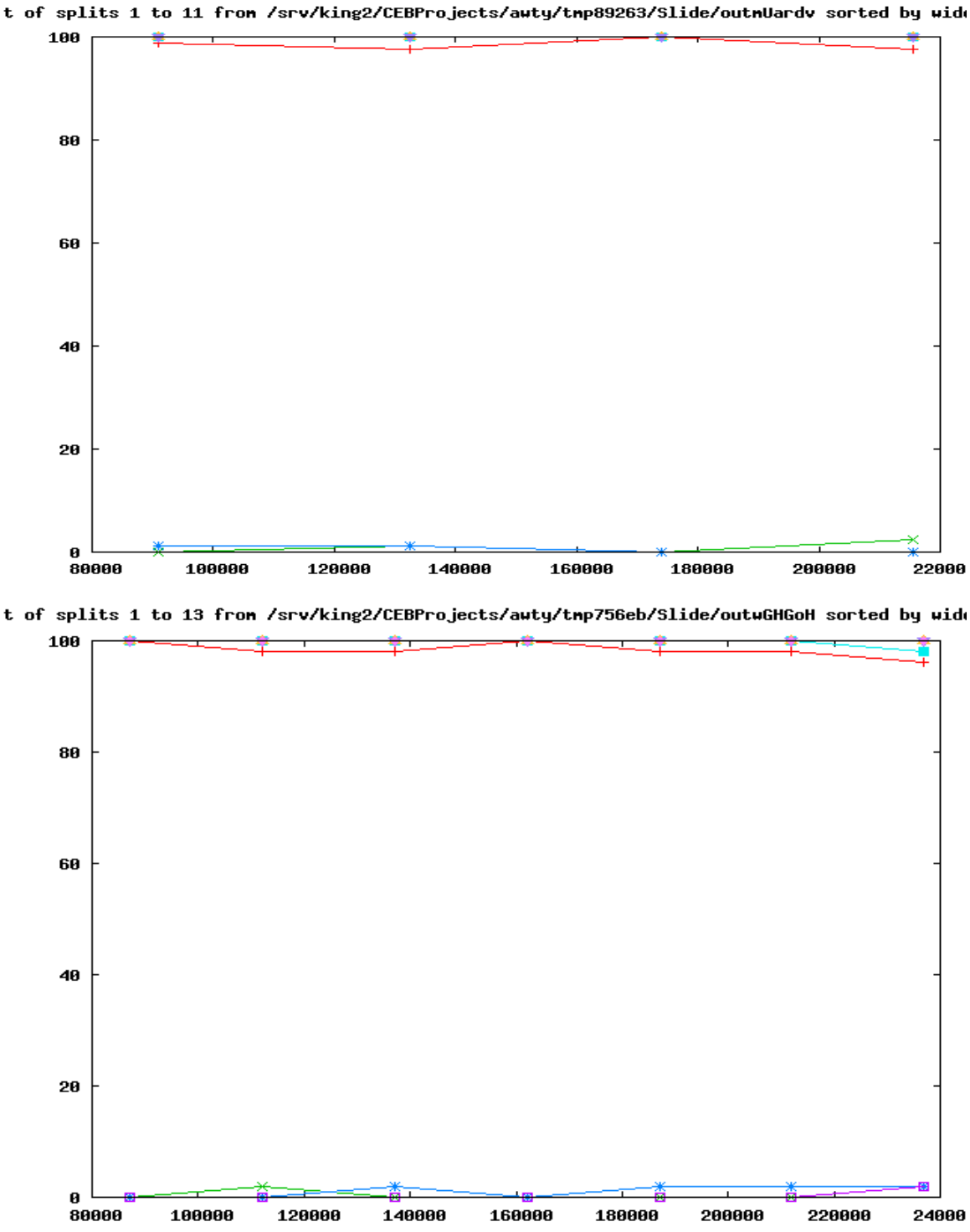



Figure 2.62: AWTY slide plot for first and second run of leucoanthocyanidin reductase respectively

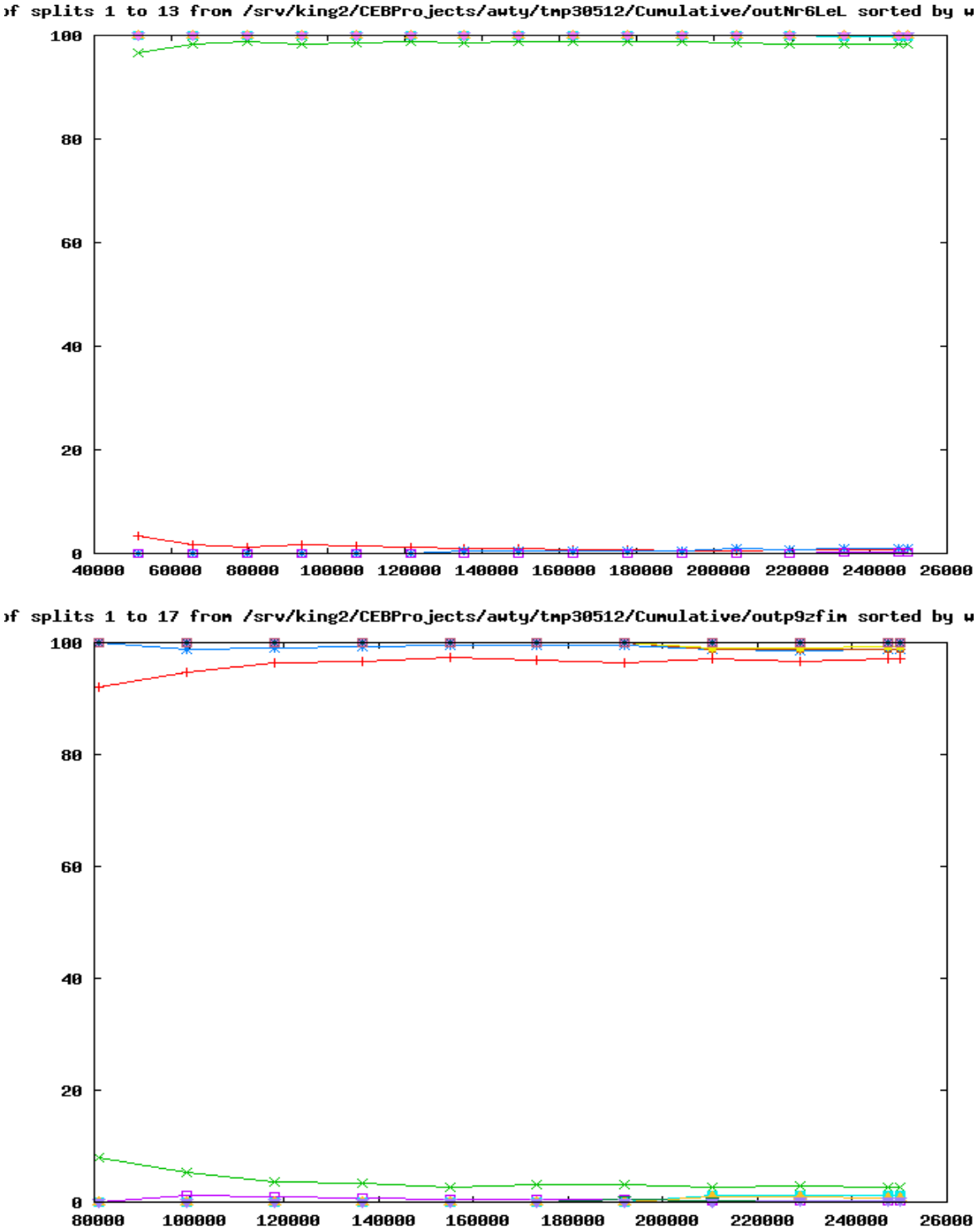



Figure 2.63: AWTY cumulative plot for first and second run of leucoanthocyanidin reductase respectively

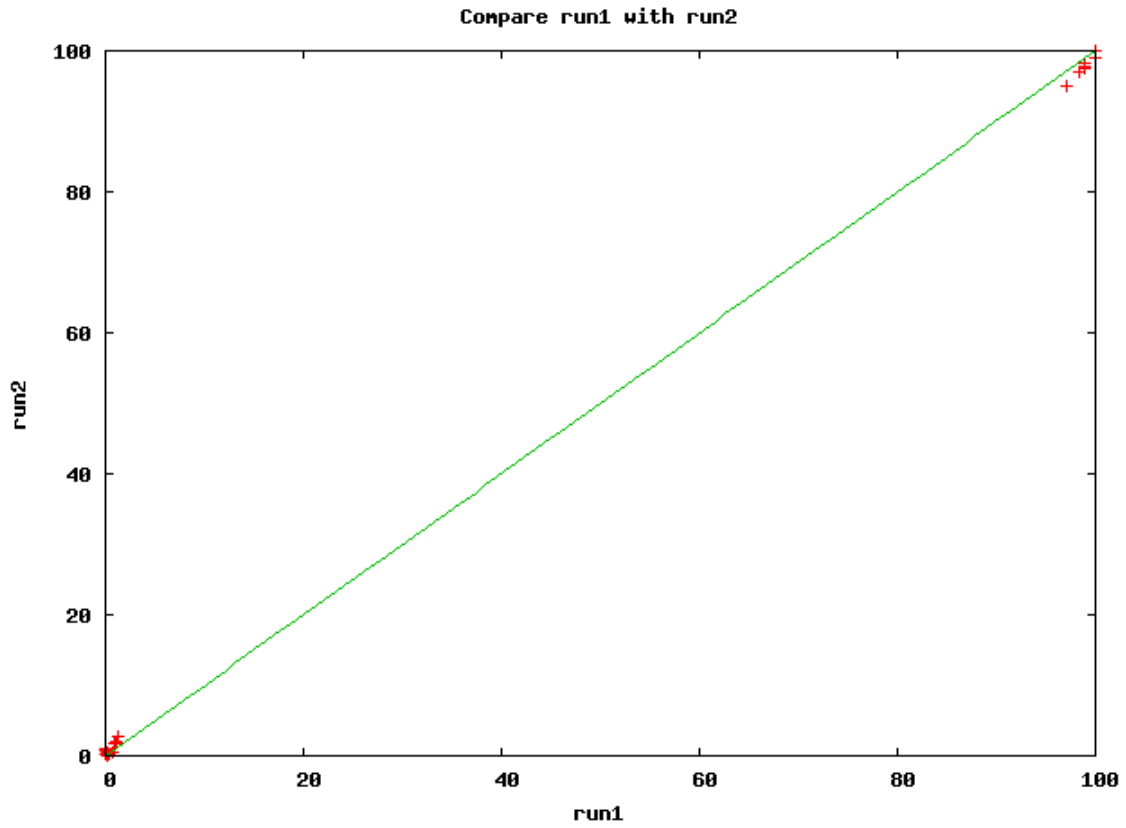



Figure 2.64: AWTY compare plot for first and second run of leucoanthocyanidin reductase

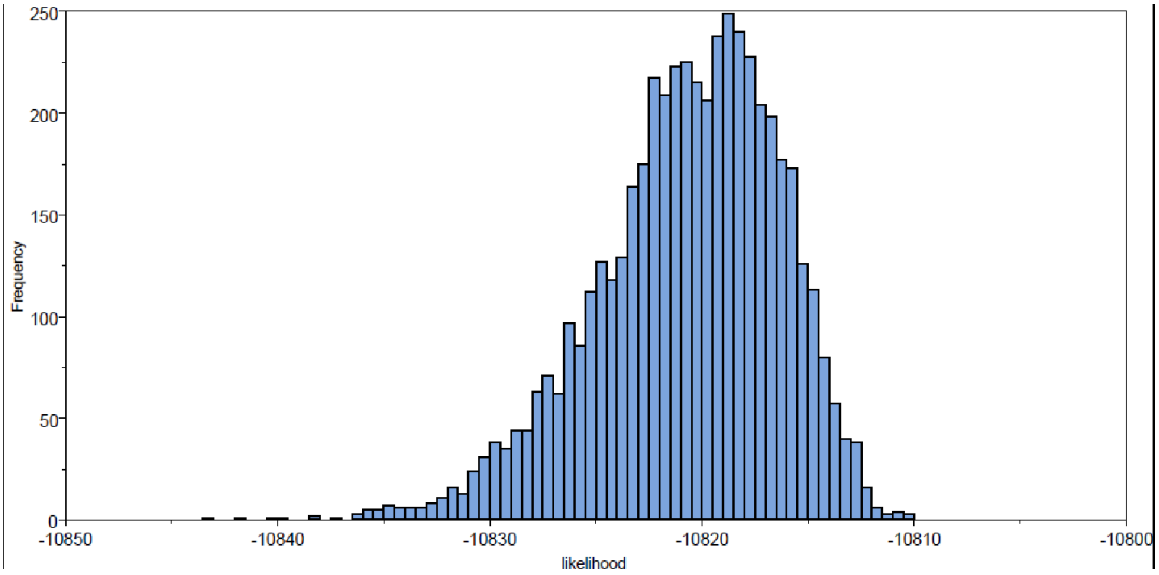



Figure 2.65: TRACER lnl estimates plot of *BEAST tree in leucoanthocyanidin reductase

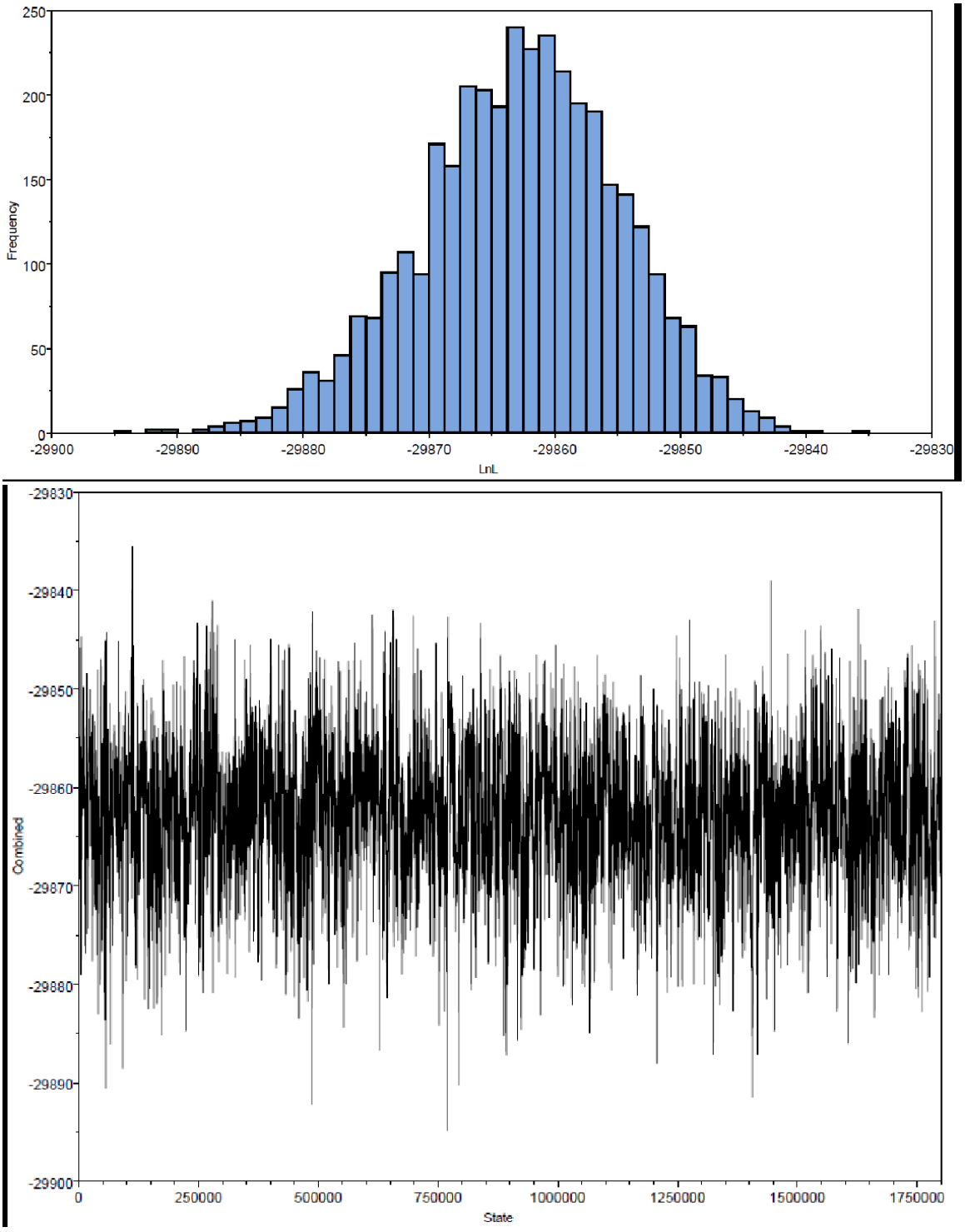



Figure 2.66: TRACER lnl estimates and trace plots of gene tree of anthocyanidin synthase

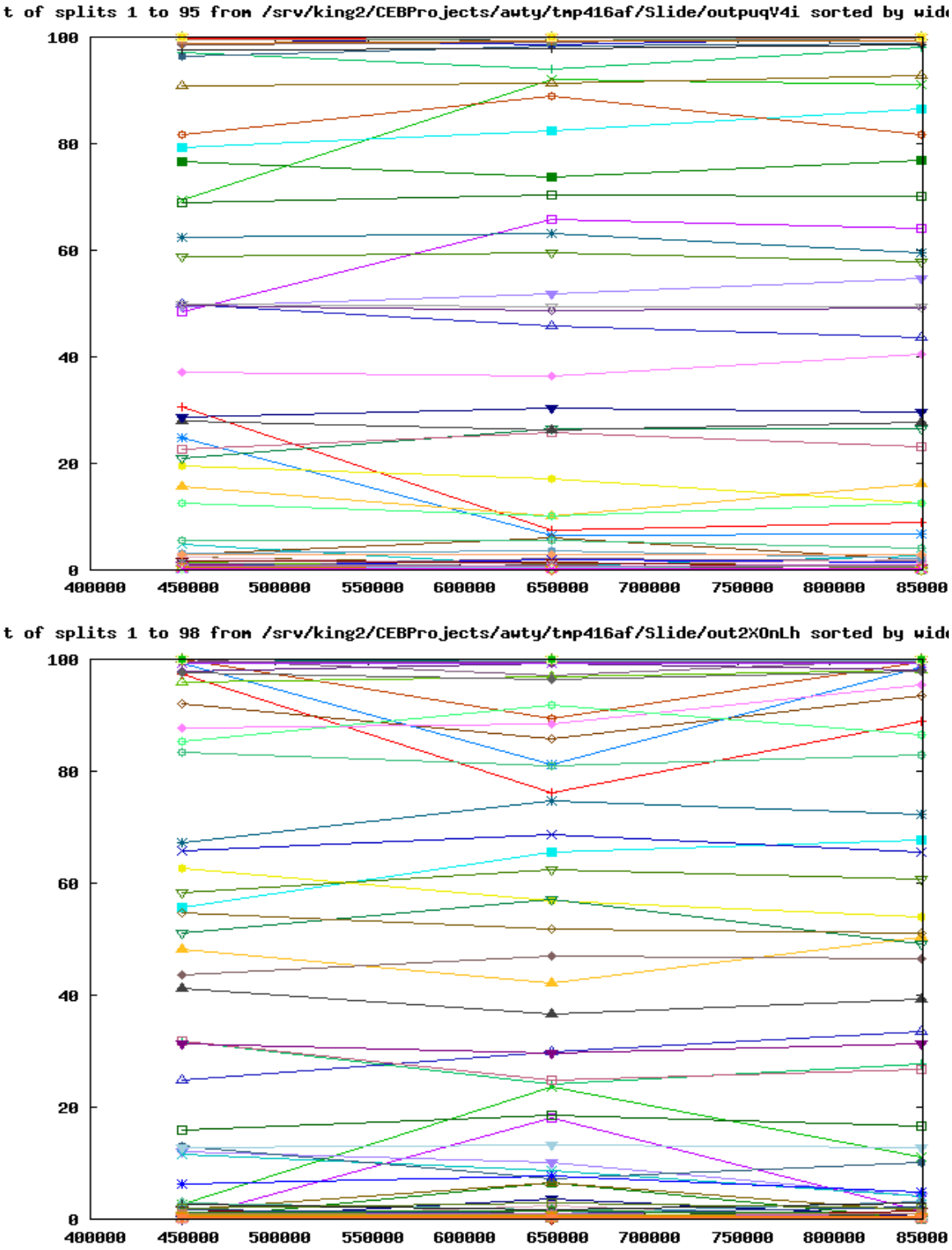



Figure 2.67: AWTY slide plot for first and second run of anthocyanidin synthase respectively

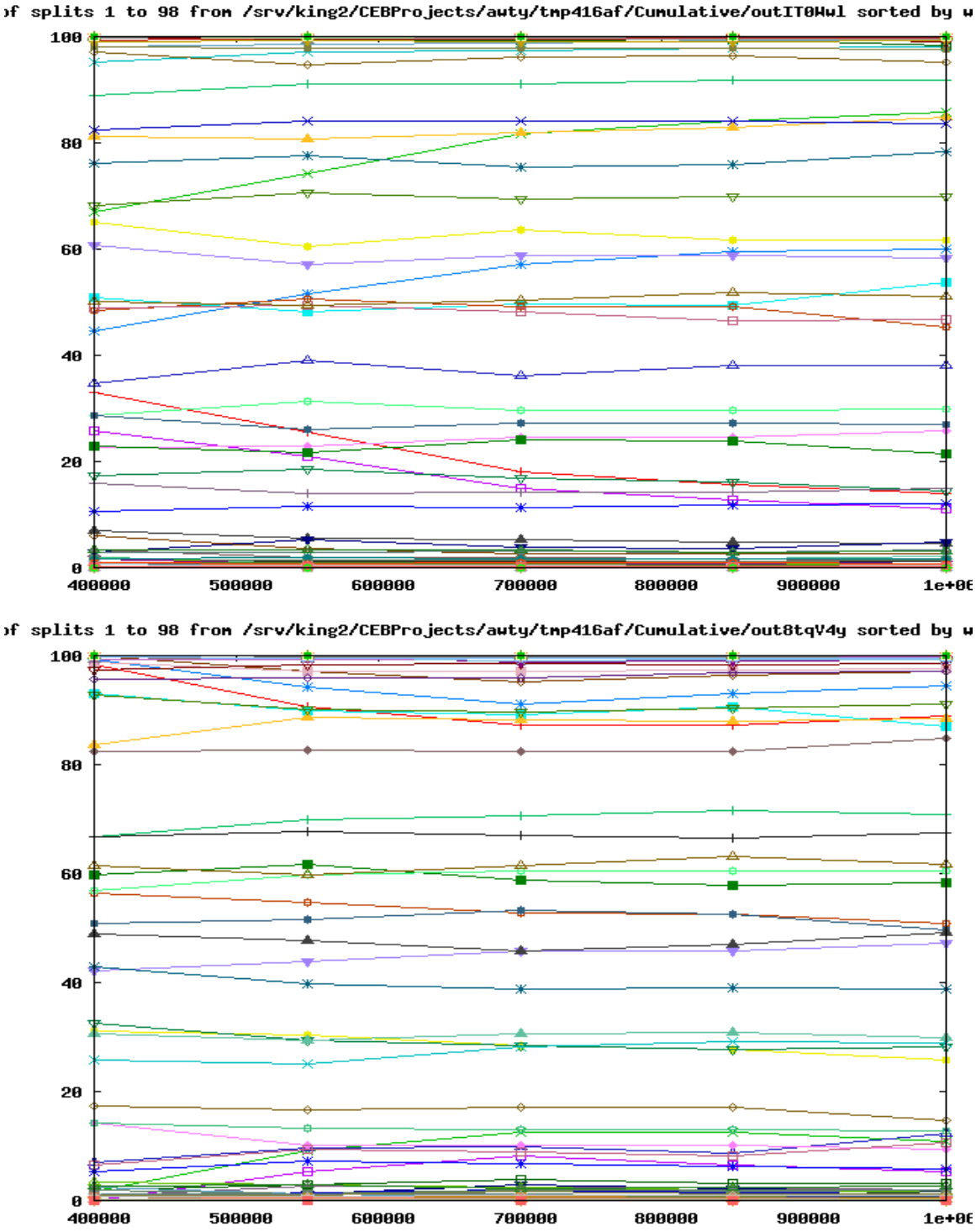



Figure 2.66: AWTY cumulative plot for first and second run of anthocyanidin synthase respectively

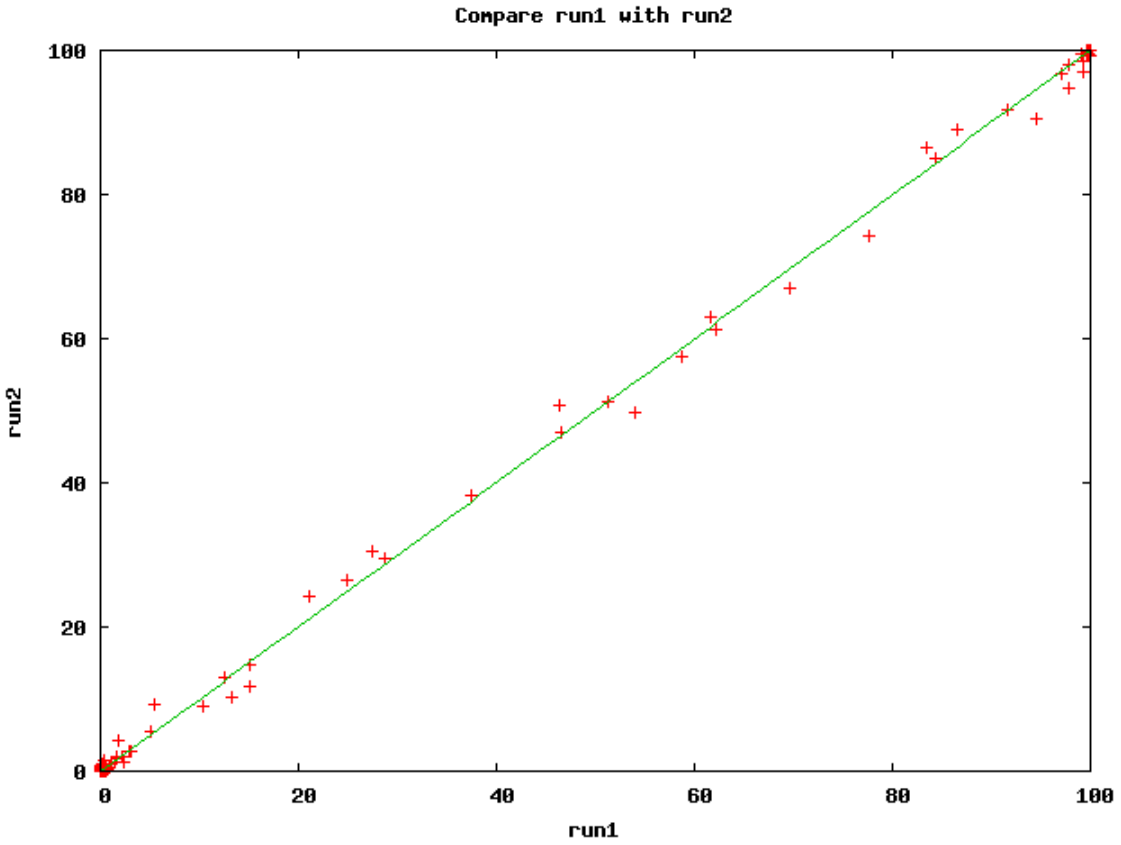



Figure 2.69: AWTY compare plot for first and second run of anthocyanidin synthase

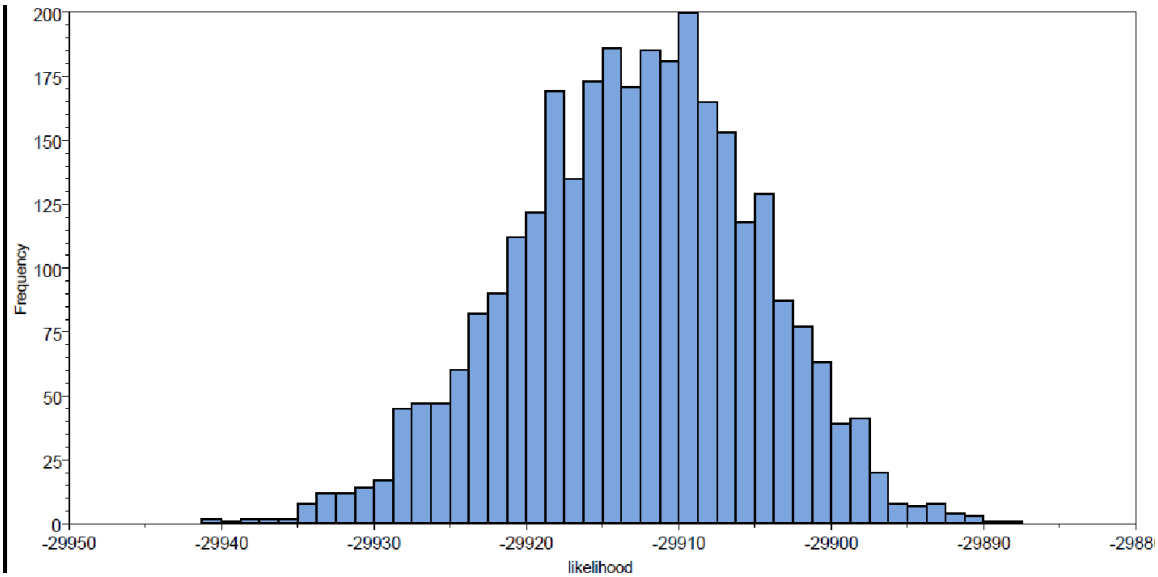



Figure 2.70: TRACER lnl estimates plot for *BEAST tree of anthocyanidin synthase

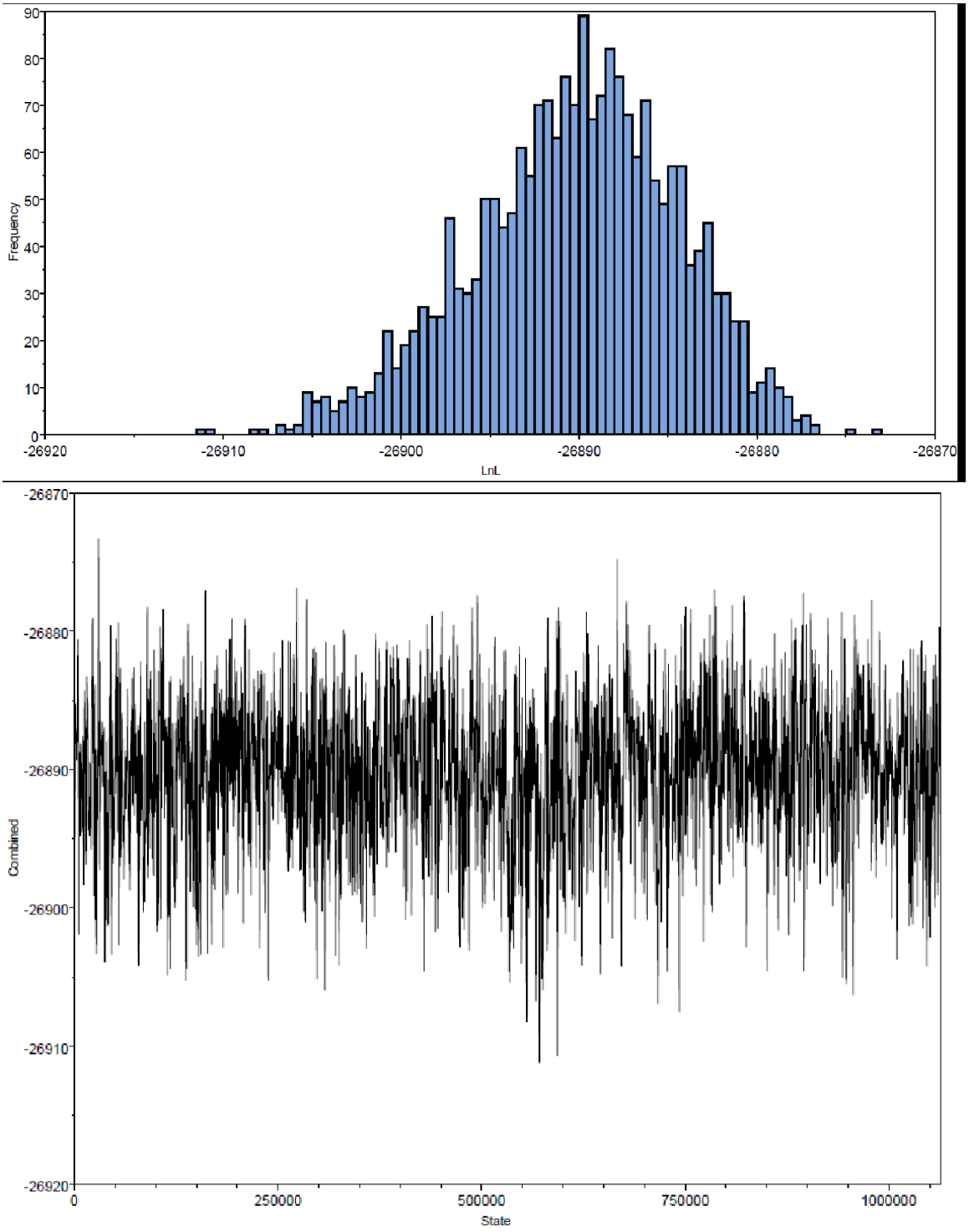



Figure 2.71: TRACER lnl estimates and trace plots of gene tree of isoflavone synthase

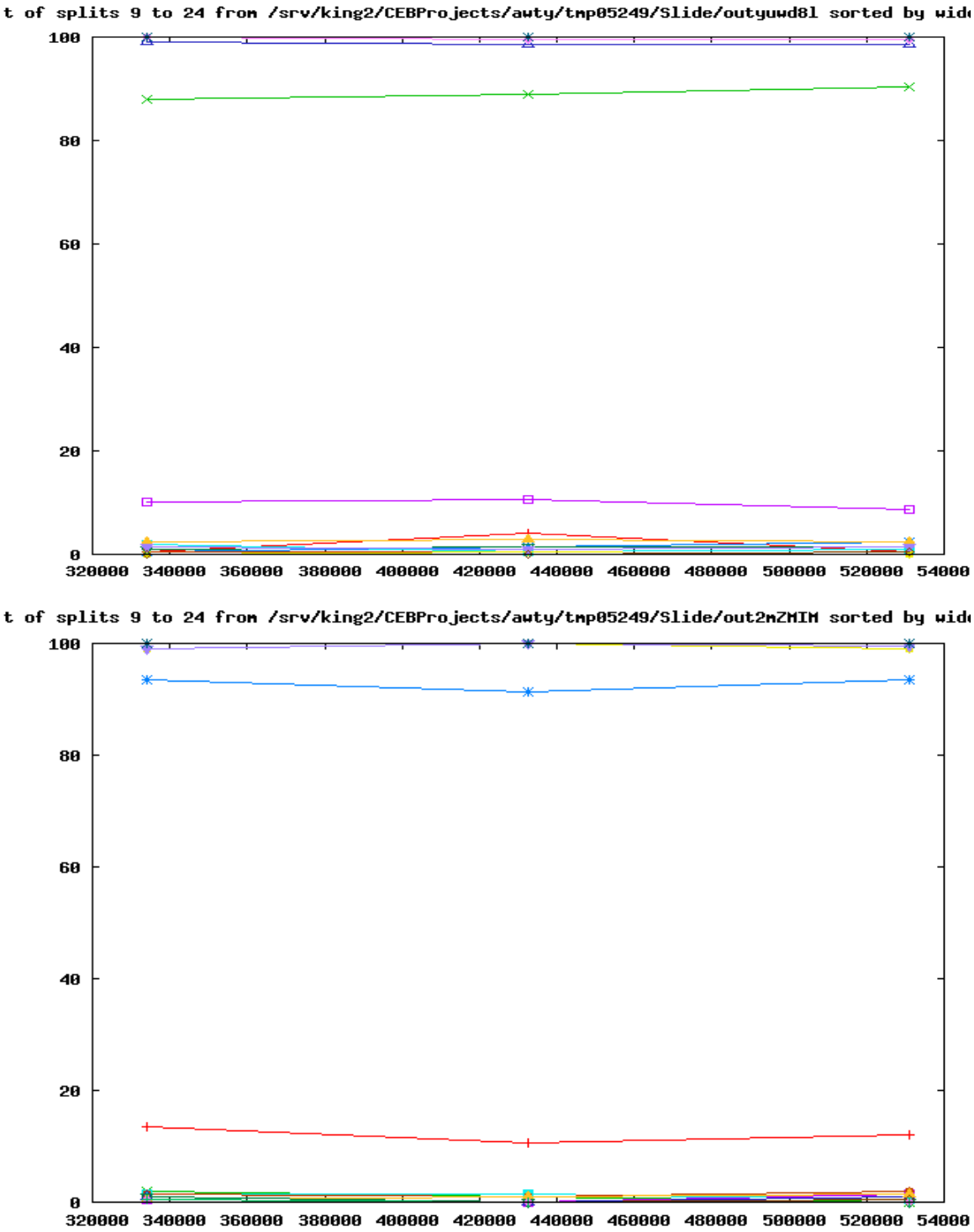



Figure 2.72: AWTY slide plot for first and second run of isoflavone synthase respectively

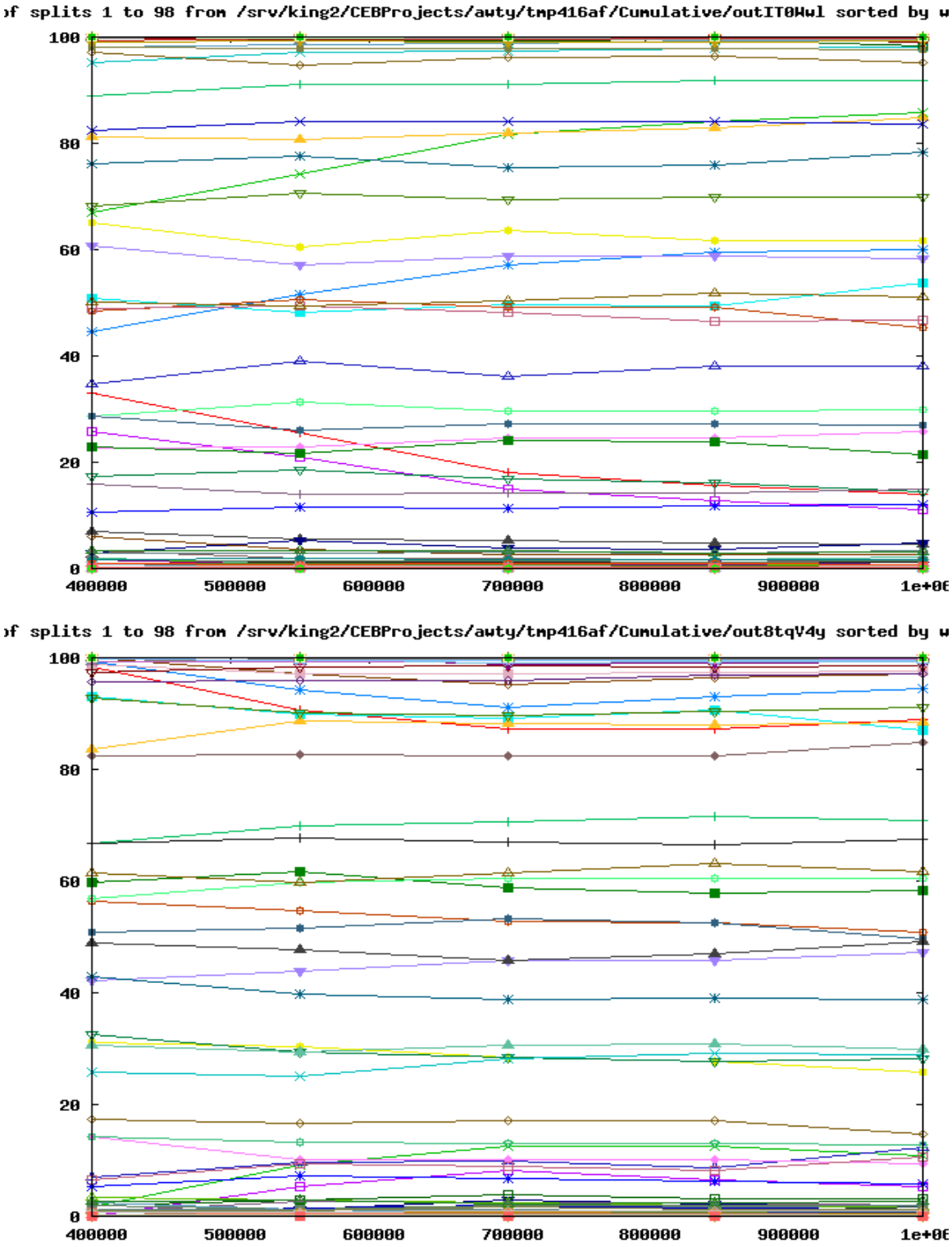



Figure 2.73: AWTY cumulative plot for first and second run of isoflavone synthase respectively

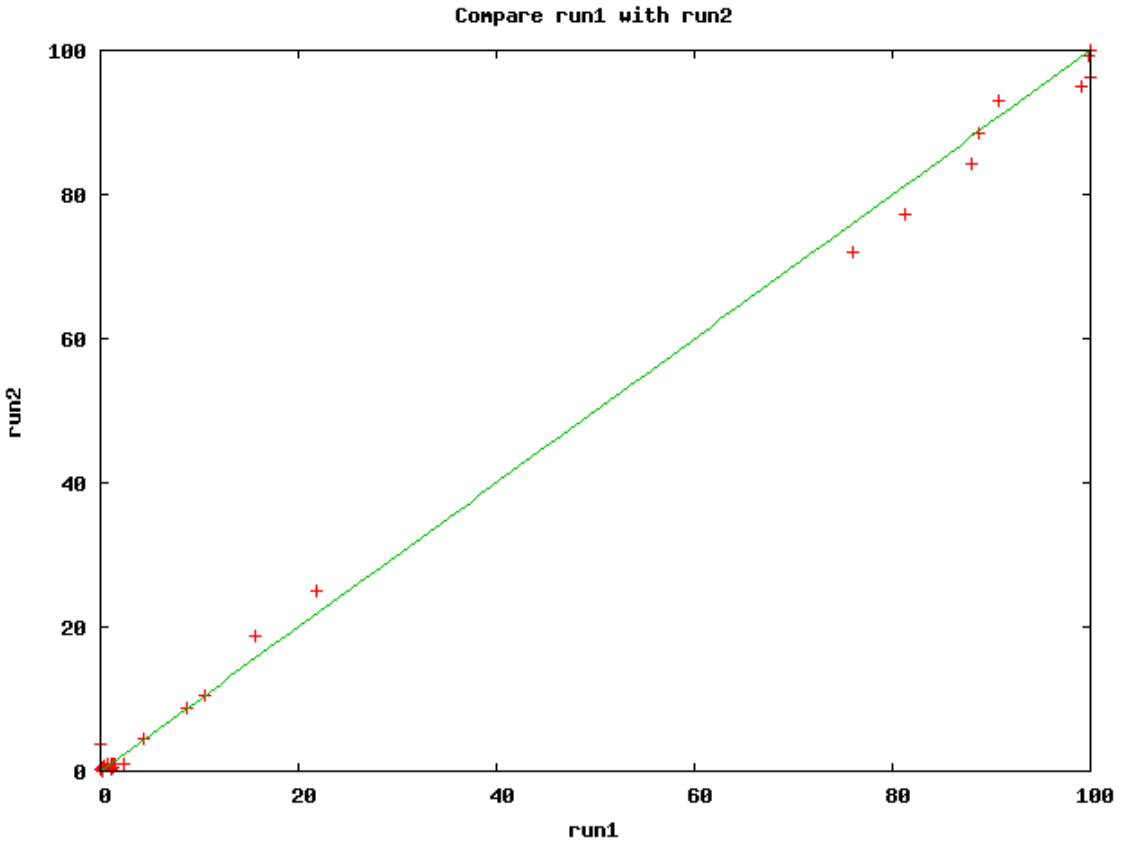



Figure 2.74: AWTY compare plot for first and second run of isoflavone synthase

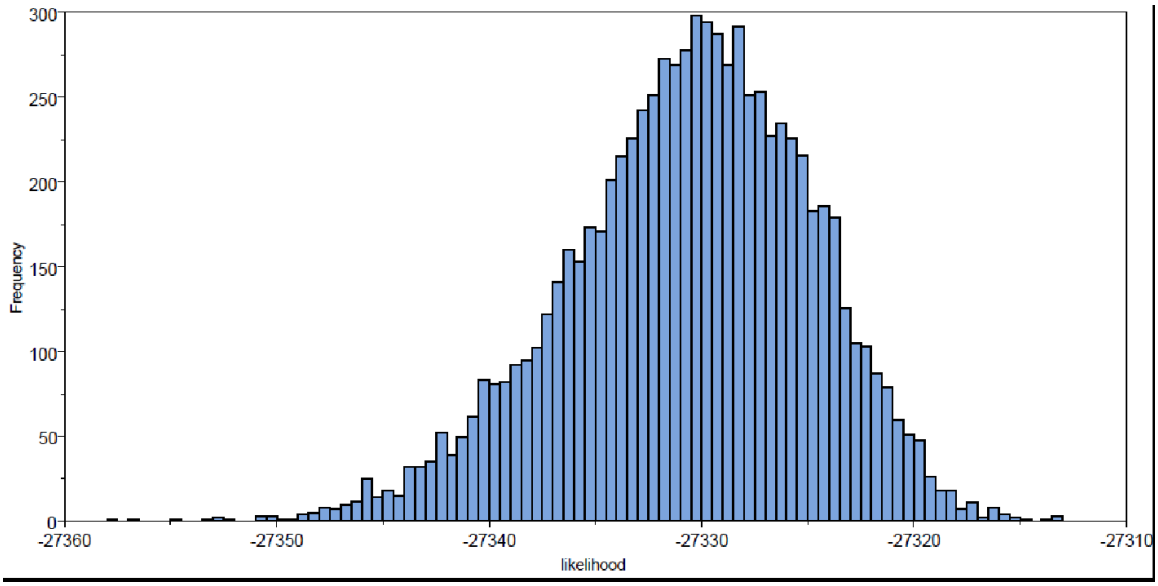



Figure 2.75: TRACER lnl estimates plot for *BEAST tree of isoflavone synthase

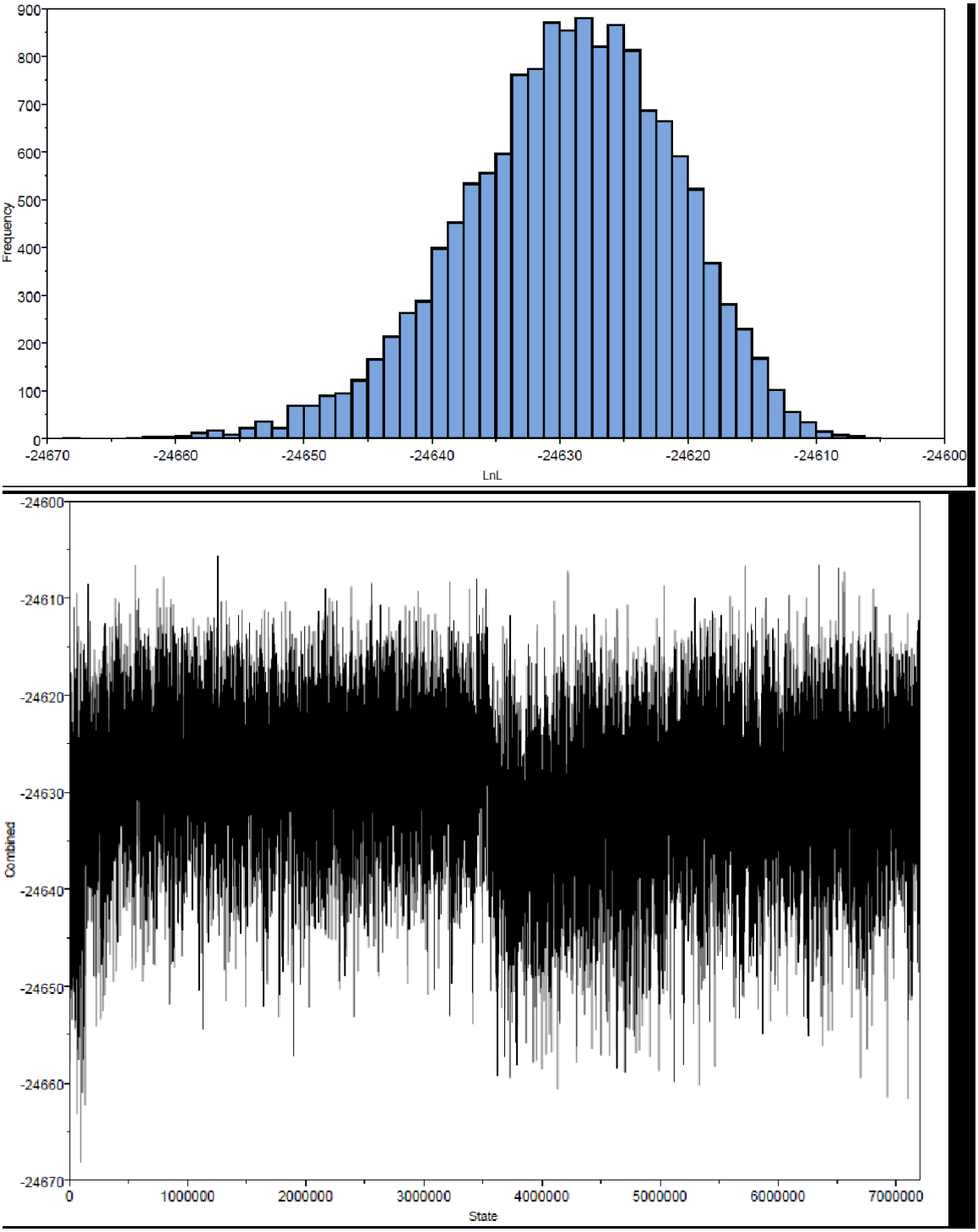



Figure 2.76: TRACER lnl estimates and trace plots for gene tree of flavonol synthase

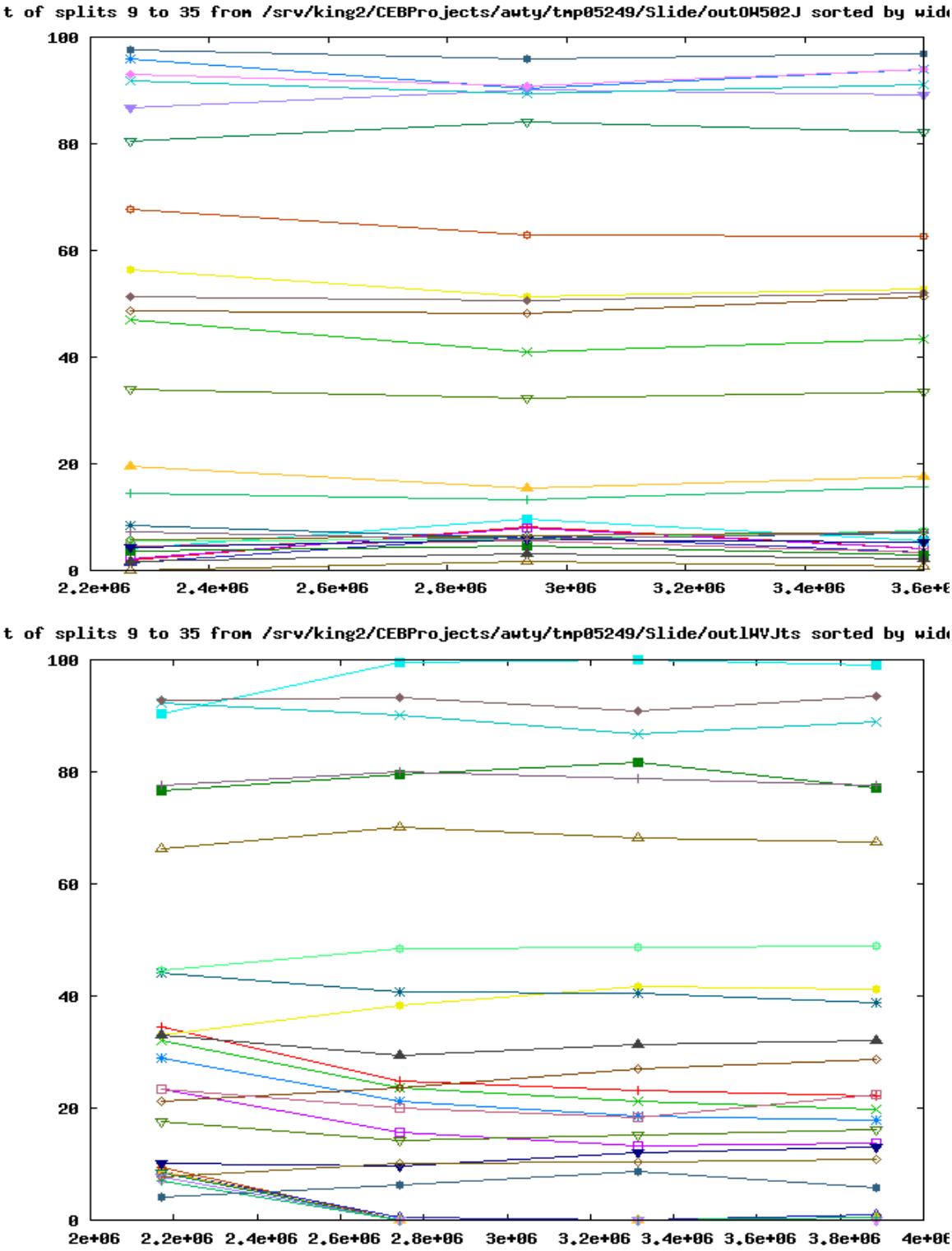



Figure 2.77: AWTY slide plot for first and second run of flavonol synthase respectively

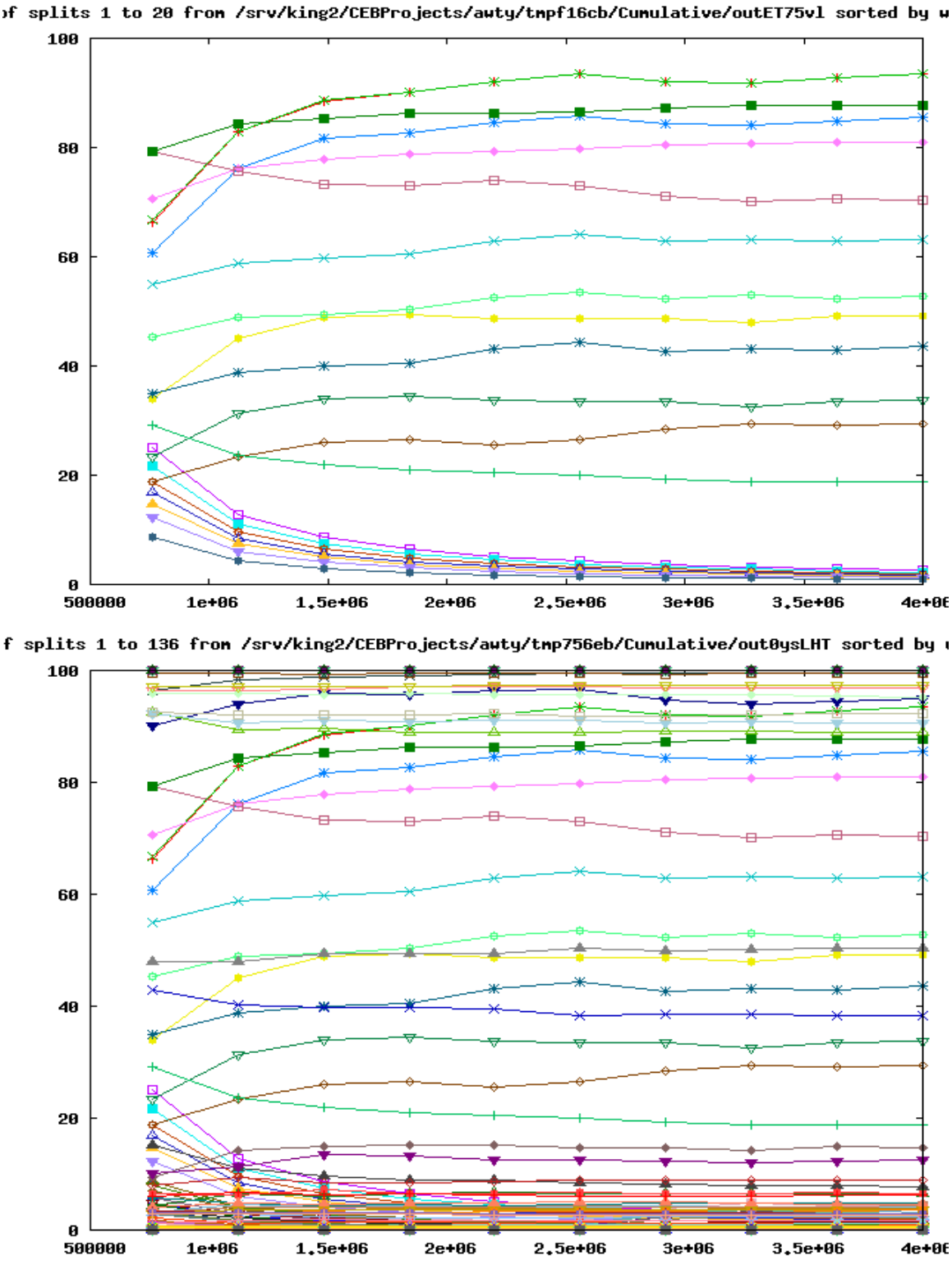



Figure 2.78: AWTY cumulative plot for first and second run of flavonol synthase respectively

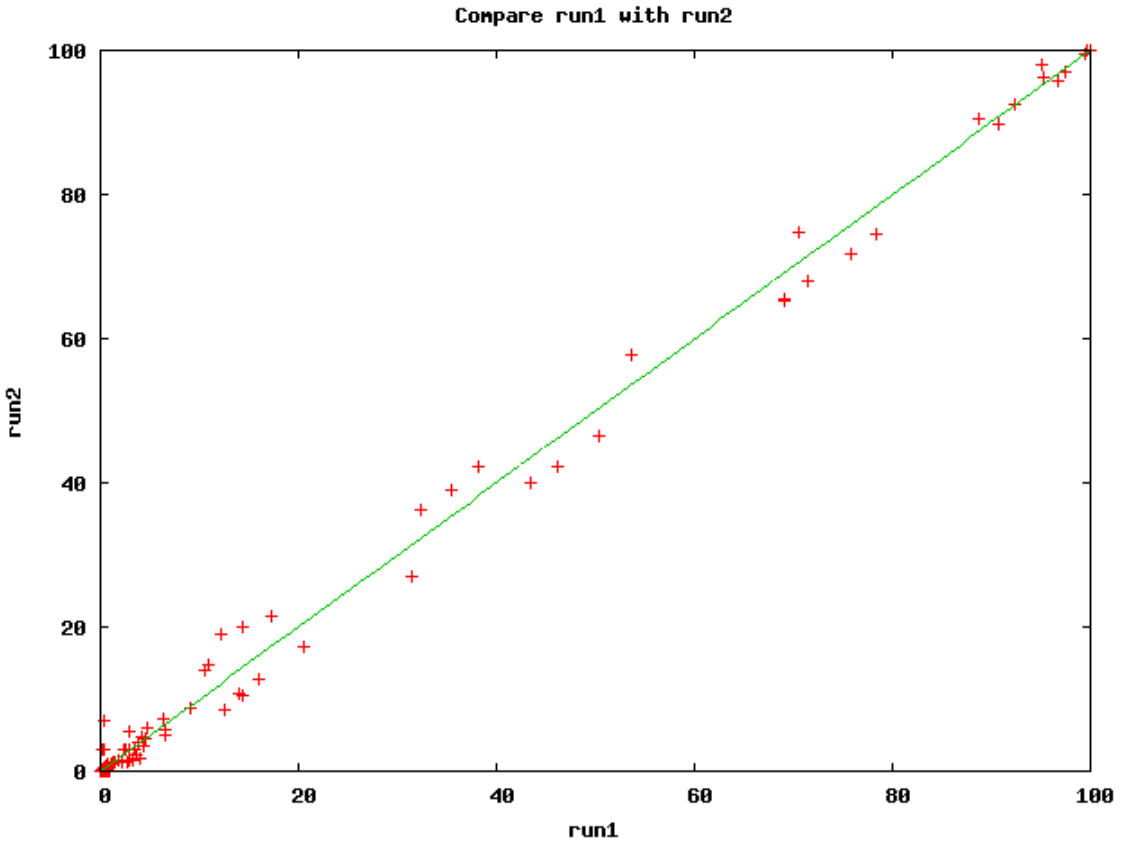



Figure 2.79: AWTY compare plot for first and second run of flavonol synthase

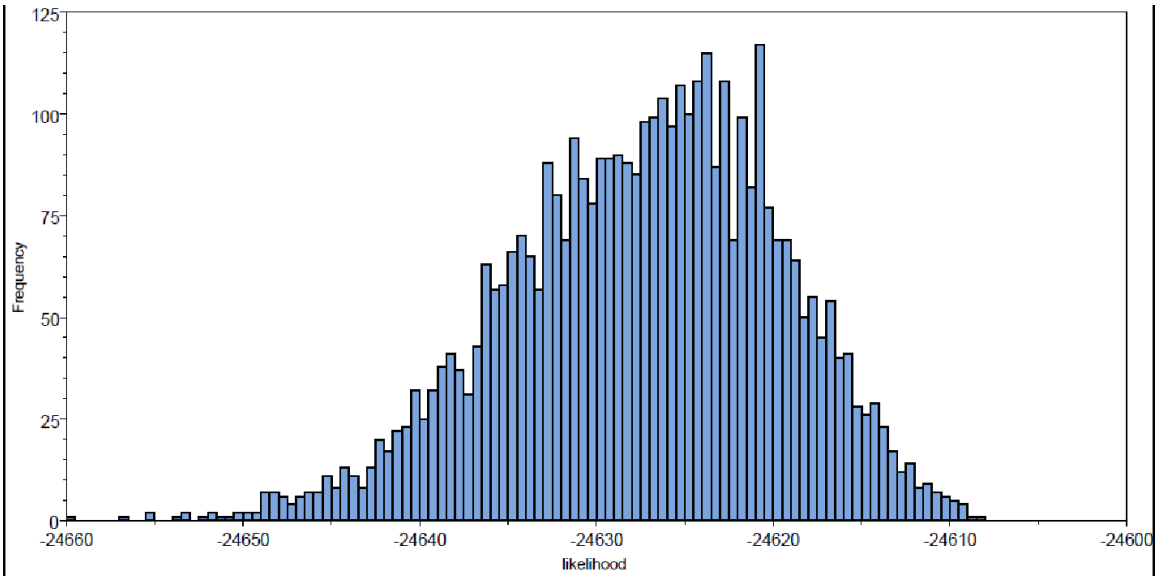



Figure 2.80: TRACER lnl estimates plot for *BEAST tree of flavonol synthase

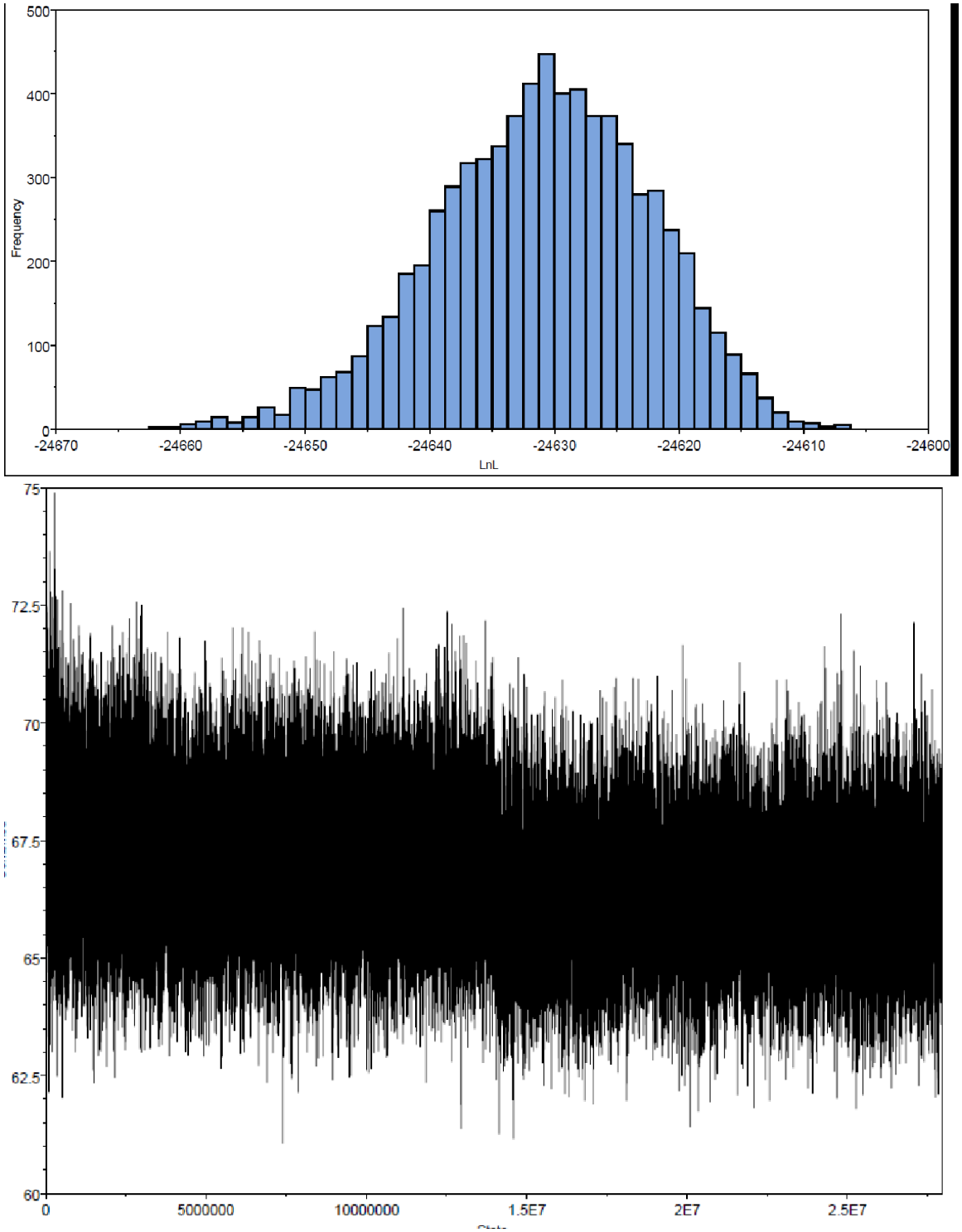



Figure 2.81: TRACER lnl estimates and trace plots for gene tress of flavone synthase

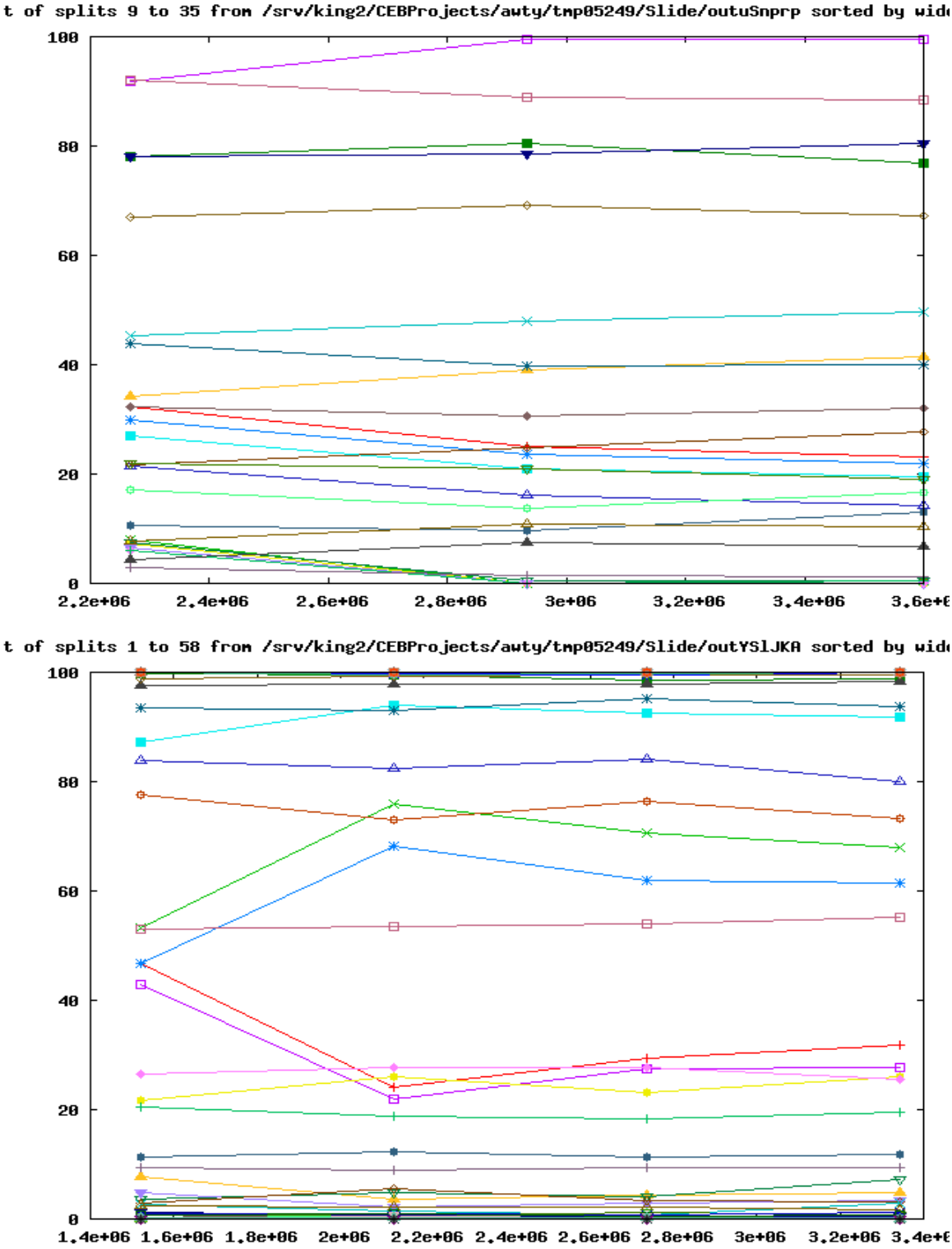



Figure 2.82: AWTY slide plot for first and second run of flavone synthase respectively

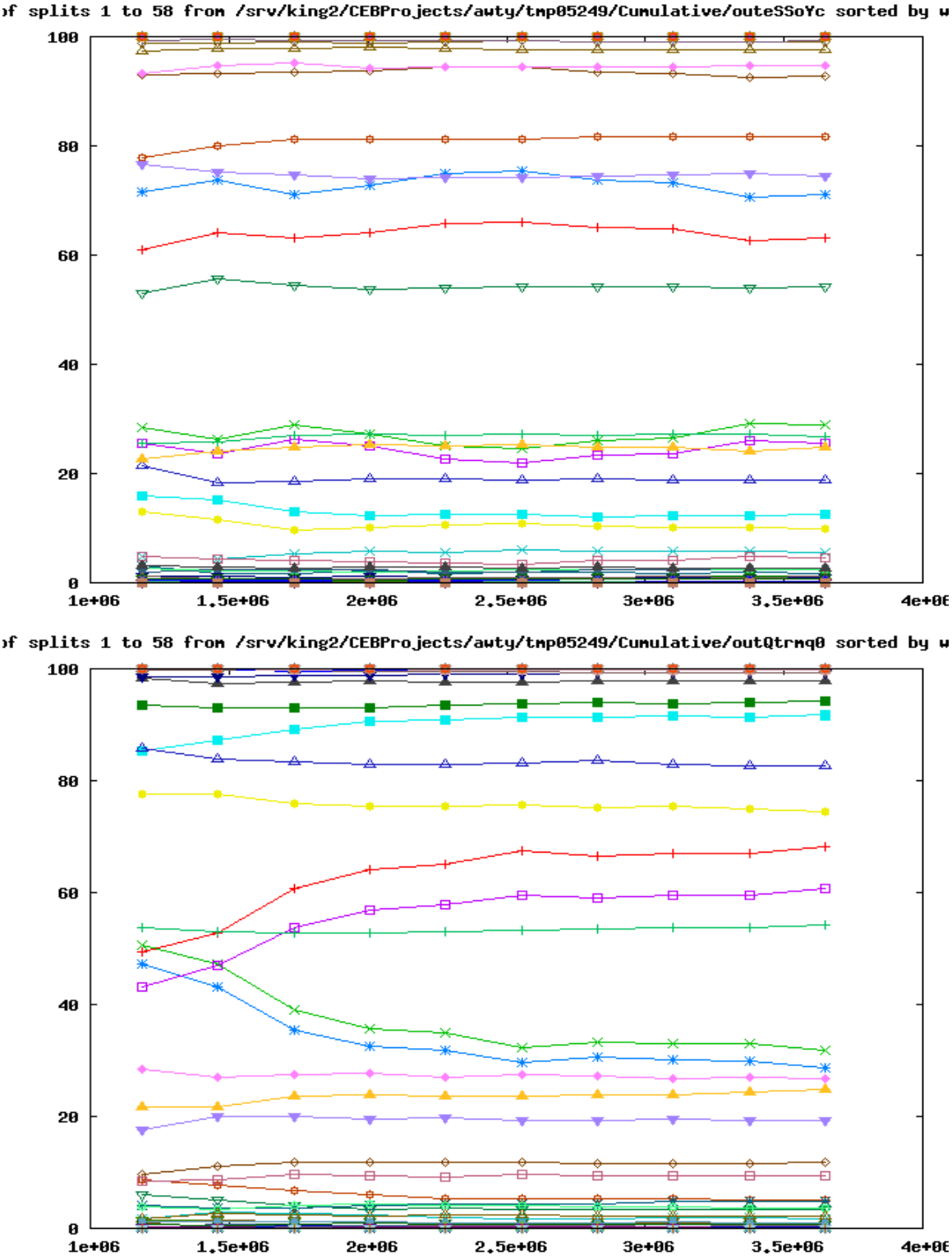



Figure 2.83: AWTY cumulative plot for first and second run of flavone synthase respectively

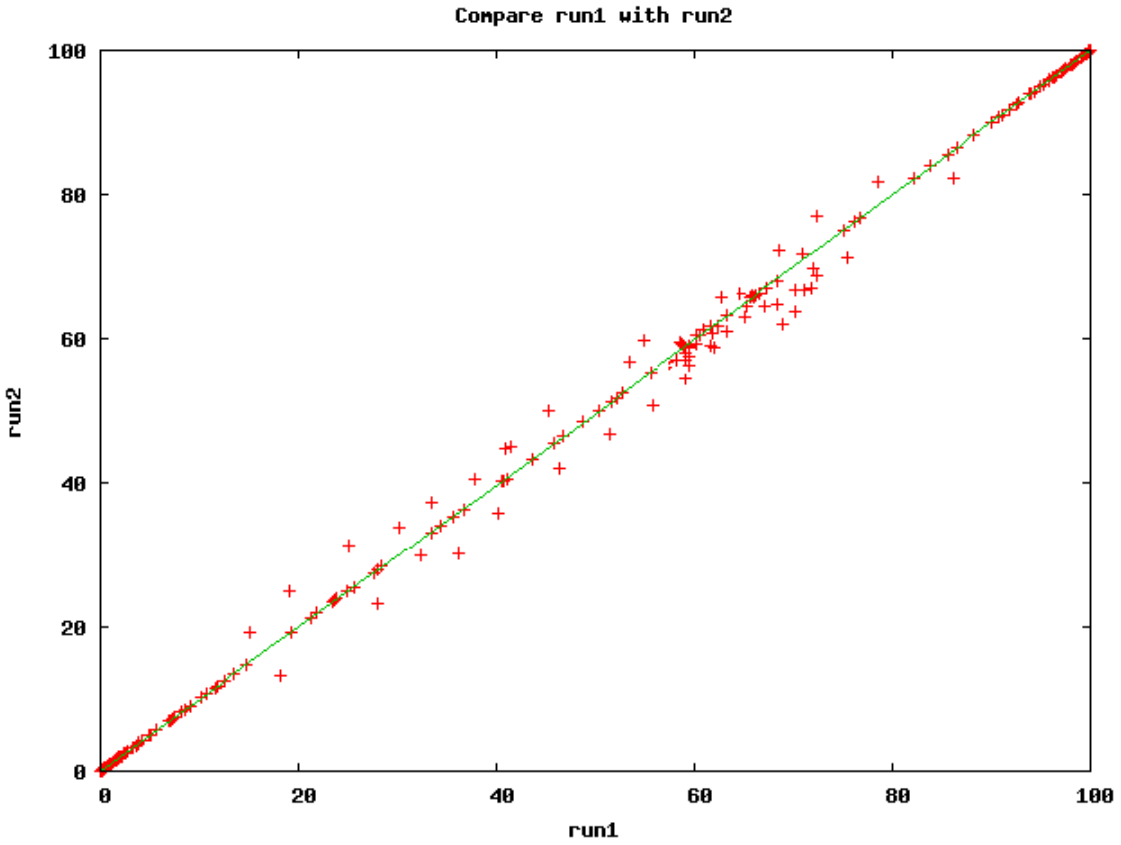



Figure 2.84: AWTY compare plot for first and second run of flavone synthase

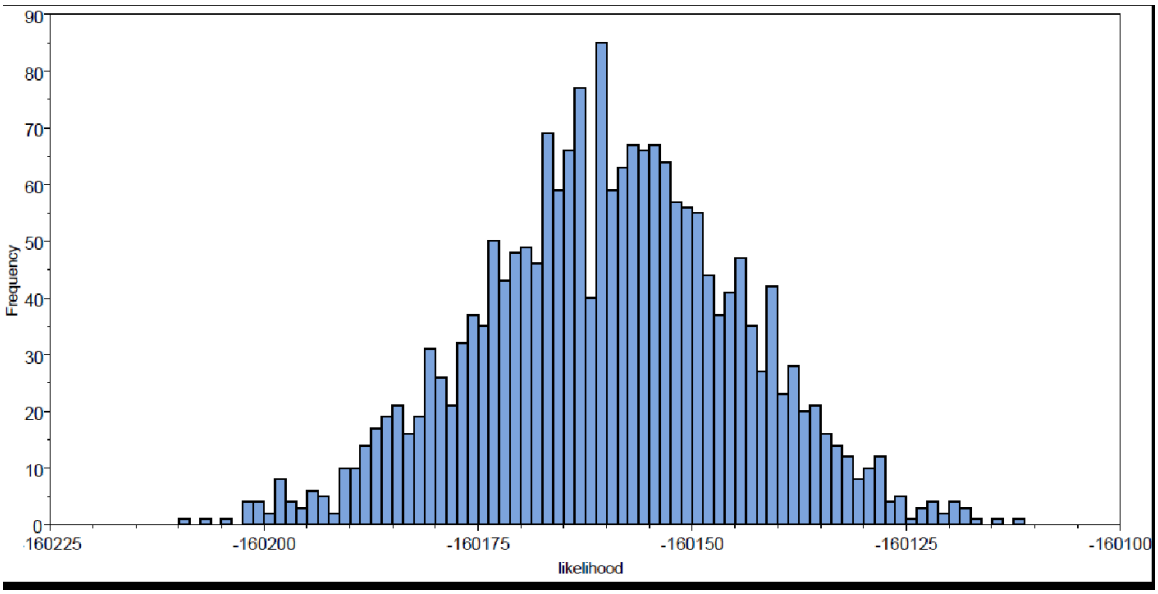



Figure 2.85: TRACER lnl estimates plot for *BEAST tree of flavone synthase

# CHAPTER 3: ION TORRENT BASED LONG AMPLICON RESEQUENCING OF FLAVONOID GENES TO SURVEY THE GENETIC DIVERSITY ACROSS SOYBEAN VARIETIES.

## 3.1 Introduction

Several studies have been done on the flavonoid biosynthetic pathway to determine the connections between phenotype and genetics. However this connection has not been fully described in soybeans. These compounds function as the major red, blue, and purple pigments in plants and are of particular interest to this study. Anthocyanins, one of the products of the flavonoid pathway, are believed to be the main controlling factors for pigmentation in soybean, causing a variety of flower and seed coat colors in different *Glycine* species (Holton and Cornish 1995). Our aim is to provide a comparative genetic analysis across soybean varieties chosen to have a wide range of flower and seed coat colors by sequencing targeted genes from the flavonoid pathway. Ideally, we hope to identify variation in the coding regions of these genes that may indicate mutations leading to variation. Many of these pigments can be used as markers to study different biological and evolutionary processes within the legumes. Based on our literature review, isoflavone synthase (IFS) and Flavanone 3-Hydroxylase (F3H) have been correlated with different flower or seed coat color in several species (Elomaa and Holton 1994, Winkel-Shirley 2001, and Yu et al. 2000). Isoflavone synthase

catalyzes the first committed step of isoflavone biosynthesis, a branch of the phenylpropanoid pathway. Flavanone 3-hydroxylase (F3H) catalyzes an early step in flavonoid metabolism leading to the formation of dihydroflavonols from flavanones. In our study, we have chosen to take an Ion Torrent based amplicon resequencing approach to survey the genetic diversity of different *Glycine* species with various combinations of flower color and seed coat color built upon our predicted gene families of IFS and F3H from chapter 2.

Ion Torrent was introduced to the scientific community in early 2010 and is based on ion detection and allows sequencing in real time. Our aim is to provide a comparative analysis across the different varieties of soybean, to identify any mutations that may reside in the IFS and F3H genes. Then begin the process of correlating those mutations, if we see them, with flower colors and seed coat colors. In addition, having the sequences of parts of the flavonoid pathway will allow us to do additional targeted sequencing of legume varieties with a wide range of flower and seed coat colors.

## 3.2 Materials and Methods

### 3.2.1 Primer Design

For isoflavone synthase and flavanone 3-hydroxylase from soybean, specific primers were designed from the genomic sequence flanking the coding region. In order to achieve this, we took all the gene candidates for each enzyme and extended them 400bp upstream and 200bp downstream, using PlantGDB. We designed our primers for the start

and end of these extended candidate regions using primer3Plus (Untergasser A. et al. 2012). Table 3.1 is the list of primers.

### 3.2.2 DNA Preparation

Six to eight seeds of each variety from Table 3.8 were planted in regular garden soil and grown in environmental growth chambers under standard greenhouse conditions with 12 hours of day and at 25°C (days) and 14°C (nights). The first to fourth trifoliate were harvested and flash frozen. A modified CTAB extraction protocol was used to extract the DNA (Doyle JJ. And Doyle JL. 1987). Modifications included changing the incubation and spinning times. For all the tissues, we added a second chloroform extraction step. We also decided to incubate the samples in 60 degrees longer than 60 minutes for a few of the samples and for some of the tissues faster and longer spinning was necessary to form a pellet. After the extraction, the DNA was quantified using nanodrop.

### 3.2.3 Long-PCR and Pooling the Products

Gradient PCR was used to test the primers described above and find the optimum conditions using Williams 82 as the standard. Williams 82 is also the genotype that primers were designed from. Some of our products were long amplicons, hence we used Taq DNA polymerase specifically designed for this. SequalPrep$^{TM}$ Long PCR Kit with dNTPs (Cat. No: A 10498) was used for the amplicons longer than 10000 bp. The PCR conditions are shown in Table 3.2. For Amplicons with length between 5000 and 10000

bp, Velocity DNA Polymerase kit (Cat. No: BIO-21099) was used. The PCR conditions for this kit are shown in Table 3.3. For the rest of our primers, with length less than 5000 bp, we used Taq DNA Polymerase kit (Cat. No: M0273X) was used. Table 3.4 shows the PCR conditions for this kit. All the annealing temperatures can be found in Table 3.1. The standard cycling conditions were used for the last two kits (Table 3.5). The cycling conditions of the first kit is shown in Table 3.6.

PCR products were tested either by running an agarose gel, or measuring the DNA using Qubit fluorometer. All of the PCR products for each variety were pooled together with equal Pico molar concentrations. For samples that were measured by Qubit, we simply converted the unit from ug/ml to pmol. To calculate Pico molar amount of samples that were ran on gel can be calculated by dividing the amount of DNA in ug is multiplied by two and divided by (DNA size in bp * 660). This number is in uM and needs to be converted to pmol. Samples were pooled, derived by lowest concentration. Table 3.7 shows the final concentrations.

### 3.2.4 Library and Template Preparation

The Ion shear™ plus reagents kit (Cat. No. 4471269) was used to shear DNA enzymatically to around 200 bp for use in the library prep workflow. Ion plus fragment library kit was used to produce high quality DNA libraries, and purify the fragmented DNA. The fragments were ligated to the barcodes using Ion Xpress™ barcode adapters 1-16 Kit. Table 3.8 lists all the samples with their corresponding barcodes, as well as their flower and seed coat color. Next, to complete the linkage between adapters and DNA inserts, a nick-repair was performed. Libraries were quantified using qPCR. We

used KAPA library quantification kit for next-generation sequencing (Kit code: KK4944) following their standard protocol. The conditions and results of qPCR library quantification are represented in Table 3.9, Figure 3.1, and Figure 3.2. After qPCR, we determined the dilution factor for the unamplified library, then calculated the number of template preparation reactions that can be performed with the unamplified library. The estimated number of template preparation reactions was sufficient, and no amplification was necessary. These quantitative steps are done using the KAPA Library Quantification Ion Torrent TDS kit (Kapa Biosystems, Inc., Wilmington, MA). Using the library dilution acquired form qPCR, each individual barcoded library was diluted and pooled together in equal volumes.

Emulsion PCR was performed using Ion OneTouch™ 2 Instrument. Multiplexed barcoded libraries were amplified by emulsion PCR on Ion Sphere particles (ISPs) at a 1:1 ratio. Ion PGM™ Template OT2 400 Kit was used for the diluted library, and enrichment was performed using the Ion OneTouch™ ES. The quality of the un-enriched and enriched template particles was assessed by Qubit 2.0 Fluorometer. The optimal template signal ratio was determined to be between 10% and 40%. Ion 314™ Chip was loaded with the enriched library. Sequencing was performed on a PGM sequencer (Ion Torrent) using the Ion PGM 400 sequencing kit according to the manufacturer's instructions.

### 3.2.5 Sequence Analysis

Torrent Suite software version 4.2.1 (Ion Torrent) was used to parse barcoded reads, to align reads to the reference genome, and to generate run metrics, including chip loading efficiency and total read counts and quality. We also used Mira 4.0.2 (Chevreux B. et al. 1994) with default setting for the Denovo and mapping assembly. We also used bwa 0.7.13 (Li H. and Durbin R. 2009), and Bowtie 0.12.7 (Langmead B. et al. 2009) to align these reads to the reference genome. Bwa uses BWT-based reference genome indexing. It runs fast and gives accurate results. It also gives gapped alignment for single-end reads, and outputs the result in SAM format. Bowtie aligns the reads one character at a time to the genome. If the alignment is successful, Bowtie examines all the positions to which the read might map. If it fails to perfectly align the character to a location, the algorithm backtracks to an earlier character, makes a substitution and resumes the search. Bowtie then builds up on the alignments by aligning two characters and so on. We then used samtools 1.3 ( Li H. et al. 2009) to convert the sorted outputs of both bwa and bowtie into BAM format. At the end, we used IGV 2.3.69 (Robinson J. et al. 2011) to visualize the BAM files.

## 3.3 Results And Discussion

### 3.3.1 Quality Assessment And Alignments To Reference Genome

Figures 3.3-3.6, Table 3.10, and Table 3.11 show the result of our sequencing and the quality of run. Unfortunately 69% of our readable data was polyclonal, meaning that they were carrying more than one template. The remaining reads were too short and did not

have enough depth to map to our reference genome. Bowtie is known as one of the best programs to align short reads to the reference genome, but was not able to effectively map our sequences to *Glycine max*. In addition to the poor quality of the sequences, our failure to adequately map them to the reference genome can also be the result of the reads mapping to multiple locations in the genome, while the program only reports on alignments that mapped to a unique location in the genome.

### 3.3.2 Denovo assemblies

Denovo assembly resulted in 1511 contigs. Most of our sequences would not assemble into long contigs. The output of assembly with the quality numbers are in the supplementary data. All of the sequences that did not assemble into a contig are also presented in the supplementary data. We used different methods of assembly and the programs mentioned in this chapter gave the best result. The highest percentage of the data that we were able to assemble was around 18 percent. The poor quality of sequences can be caused by multiple factors, rooting in both experimental design and the library preparation. We will be discussing some of the possible scenarios in the following paragraph.

All of the primer sets were designed based on our preliminary data from chapter 2. At that point of the study, we had a lot of false positives as potential gene candidates. Also, multiple number of our longer primers did not amplify, mostly due to the large difference between their forward and reverse $T_m$. This gap has led to the loss of overlap between our regions, causing difficulty in assembly. We also did not get enough depth or good quality

sequences. Almost 70% of our sequenced data was polyclonal. Furthermore, Mira is very conservative in merging regions, which often results in short contigs.

## 3.4 Conclusion

Although we were able to sequence some of our targets, the end result was not sufficient enough for further analysis or a good quality assembly. As discussed before, this was partly due to the fact that most of our sequencing data was polyclonal. Consequently, we were not able to introduce potential genes/SNPs that are involved in pigmentation of plants, which was the goal of this study. For the purpose of this study, re-sequencing and further analysis was not possible due to the lack of more funding. Nevertheless, in future studies we can use our PCR products and re-sequence these regions. This should be done after a re-evaluation of gene candidates for IFS and F3H. Secondly, it is very crucial to make sure we have good overlap between different primers. The primers that fail to amplify, need to be re-designed. Library prep and sequencing protocols should be followed more thoroughly to avoid low quality results.

Table 3.1: Primers

| Name | | Sequence | nmoles | Tm | Length |
|---|---|---|---|---|---|
| 1 | A01 | CCG TTG TCA ACT TCC AGG GA | 12.03 | 52 | <5000 bp |
| 1 | A02 | RAC AAG GTA TCA TCA ATC CAG ACT | 11.95 | 52 | <5000 bp |
| 1 | A03 | AAC TTG AAA AGC CTC AGT GG | 11.99 | 50 | <5000 bp |
| 1 | A04 | CGT GCA TGA GAC GGG TAG AA | 12.06 | 50 | <5000 bp |
| 1 | A05 | TGG GCT CCT TTC AAA ACA GA | 11.95 | 52 | <5000 bp |
| 1 | A06 | CCA CAA GAC TTA AAT TAT GTG AGC | 11.98 | 52 | <5000 bp |
| 1 | A07 | AAT TAT ACT CGC ATG CAA CAA A | 12.05 | 50 | <5000 bp |
| 1 | A08 | AGG TTG TTT CCA AAT TGA GAG T | 11.99 | 50 | <5000 bp |
| 1 | A11 | TGC AGA TCA TAC GCT TAT GGC T | 12.06 | 52 | <5000 bp |
| 1 | A12 | ACC CGT GTT AAA ACT TTG GAC A | 12.03 | 52 | <5000 bp |
| 1 | B01 | ACC TAA GTG ACA TCA ACG ATT ACA | 11.98 | 51 | <5000 bp |
| 1 | B02 | TGA GTT GAA GGA TTT CAG TGT GG | 12.04 | 51 | <5000 bp |
| 1 | B03 | TGT TGA TTC GTG GAA GCA GGT | 11.94 | 54 | <5000 bp |
| 1 | B04 | AGT TAC GCA GGA AGA GCA CT | 11.98 | 54 | <5000 bp |
| 1 | B05 | GCG GAC GAC AAC GAT CTC AA | 12.07 | 50 | <5000 bp |
| 1 | B06 | AAG CTG TAT ATC ATT TGG TGG T | 11.95 | 50 | <5000 bp |
| 1 | B09 | TTG TGA TGG TTT TTA AGT GCA A | 11.95 | 49 | <5000 bp |
| 1 | B10 | TCA GAC AAA AGA TCA GTC TTG A | 12.02 | 49 | <5000 bp |
| 1 | B11 | TGC ATC TGC ATC TAC ATC TAC CT | 11.92 | 52 | <5000 bp |
| 1 | B12 | ACC TCT TGC TTC TGT TGC TT | 11.95 | 52 | <5000 bp |
| 1 | C01 | ACC ACC CAA CCT TAT GAT AGG TG | 12.01 | 52 | <5000 bp |
| 1 | C02 | ACA CTT TGA AGC TCA TTT TGC T | 12.08 | 52 | <5000 bp |
| 1 | C05 | ACA TAA GTC ATT TTG GGT TGT TT | 12.06 | 50 | <5000 bp |
| 1 | C06 | CTT CCA CAT GTG GCT CTC CC | 11.99 | 50 | <5000 bp |
| 1 | C07 | TGT TGA TAG ATT GTG TTC ATT GCG | 12.01 | 52 | <5000 bp |
| 1 | C08 | TCT CAA TCC ACA GGC TAT GT | 12.09 | 52 | <5000 bp |
| 1 | C09 | ACC AAA CAA ACT AAC GTG ACA CT | 12.01 | 54 | <5000 bp |
| 1 | C10 | TGC TAT TTT TGA GCT ACT TAG GCG | 12.07 | 54 | <5000 bp |
| 1 | C11 | TCT ACT ATA CAA TGT CTG CCA AA | 12.06 | 49 | <5000 bp |
| 1 | C12 | AGG TTG ACA AAT ATT TTA GAG TGC | 11.97 | 49 | <5000 bp |
| 1 | D05 | TTC CTT AGA CAA ATG TCA TTG TTT | 12.03 | 50 | <5000 bp |
| 1 | D06 | AGC TGT AGT CTA CCG TTT TCT | 12.01 | 50 | <5000 bp |

| | | | | | |
|---|---|---|---|---|---|
| 1 | D07 | TCC TGC ACA GTT CAA TTT TAA CA | 12.05 | 51 | <5000 bp |
| 1 | D08 | CAC CCA ACT TAA GCA TTC AGT CA | 12 | 51 | <5000 bp |
| 1 | D09 | TCA ACC CTT TCC ATG CTT CA | 11.93 | 52 | <5000 bp |
| 1 | D10 | TGG CGA TAC CAT CAG TGC C | 12.05 | 52 | <5000 bp |
| 1 | D11 | GAG CAT CTC GAA ATC CGG C | 12.04 | 54 | <5000 bp |
| 1 | D12 | TGG ACA AGG GGT AGG TGT AGT | 12.06 | 54 | <5000 bp |
| 1 | E01 | AAA TGT TTT TAA GGG ACG AAA GAA | 12.04 | 49 | <5000 bp |
| 1 | E02 | CTG TTA ACA GGC AAC CAC AA | 11.96 | 49 | <5000 bp |
| 1 | E03 | CAT GAG AGG GGT TTC GAC GA | 11.97 | 51 | <5000 bp |
| 1 | E04 | TCC AAC TCA TCC AAT TCC CC | 11.94 | 51 | <5000 bp |
| 1 | E05 | CCC CGT CCT TGG TTT AGA TTT G | 12 | 53 | <5000 bp |
| 1 | E06 | CGA GTA ATT GCA TTT TAG GGC CA | 11.97 | 53 | <5000 bp |
| 1 | E07 | TGT CAG TGT TCC AAA AGC TGA | 11.97 | 50 | >10000 bp |
| 1 | E08 | CAC CAA AGG AGT GTC TTT TCT | 11.95 | 50 | >10000 bp |
| 1 | E09 | TGT CAG TGT TCC AAA AGC TGA | 12.02 | 50 | >10000 bp |
| 1 | E10 | TGT TCT ATG TTT CTA GTG TTG TGT | 12.06 | 50 | >10000 bp |
| 1 | F05 | TGG TGT ACA ATA AAT GTT GAC CCG | 11.99 | 50 | <5000 bp |
| 1 | F06 | TTG CGG TCC AAC ATA TCT ATA A | 11.95 | 50 | <5000 bp |
| 1 | F09 | AAA CCG CTT ATG AAT TTT CAT GA | 12.05 | 49 | <5000 bp |
| 1 | F10 | TTT CAT GAA ACC TCC GTT AGT | 11.95 | 49 | <5000 bp |
| 1 | F11 | TGG TGT ACT TGA CAT CCG GG | 11.96 | 54 | <5000 bp |
| 1 | F12 | TTC TTT CCA CGC GGA GGA AA | 12.02 | 54 | <5000 bp |
| 1 | G05 | GGG TGG TTG ACA CTC CTT GA | 12.04 | 52 | >10000 bp |
| 1 | G06 | GCT ACG AGG CAC TAC CAG AC | 12.09 | 52 | >10000 bp |
| 1 | G07 | GGG CAG ATG GCT GGA TTG AA | 11.98 | 54 | <5000 bp |
| 1 | G08 | GTC GGA TGG AAG CTA GTA GAT GA | 11.92 | 54 | <5000 bp |
| 1 | G09 | AGG TCA ACA ACC CAT TCC CA | 11.96 | 51 | <5000 bp |
| 1 | G10 | GCA ATA TTT TAA CGA CCC TAG ACC | 11.99 | 51 | <5000 bp |
| 1 | H05 | ACT TAT TAC CGA CTG AGT TGA TT | 12.06 | 51 | <5000 bp |
| 1 | H06 | TGC TTA AGT TAG CAA AGA ATT CC | 11.97 | 51 | <5000 bp |
| 1 | H07 | ATG CTA ACC AAA CCA GCA CT | 11.98 | 49 | <5000 bp |
| 1 | H08 | GTG TAC CAT ATG TTG TTA GAC AAA | 11.99 | 49 | <5000 bp |
| 1 | H11 | TTT TGC AAC CAC AAT AAG CG | 12.02 | 50 | <5000 bp |
| 1 | H12 | AGC TTT TGG ACA CTT GCA TAC | 12.01 | 50 | <5000 bp |
| 4 | A05 | GGG AAC CCC ATT CTG TAC CT | 12.01 | 49 | <5000 bp |

| | | | | | |
|---|---|---|---|---|---|
| 4 | A06 | TGA TTT TTC TCT TAT CAG ACC TGT | 12.04 | 49 | <5000 bp |
| 4 | A09 | GCA TGA ACC TTA TTA GAA CTC TCA | 12.06 | 49 | <5000 bp |
| 4 | A10 | TGA GTT CTC ATC TCA TTT TTG TTG | 11.96 | 49 | <5000 bp |
| 4 | A11 | AGG TTC AGG GCT TCA AGA CG | 12.05 | 54 | <5000 bp |
| 4 | A12 | GCT ATT GGG CTT CGT TTG TGT | 11.99 | 54 | <5000 bp |
| 4 | B01 | TTA ATT TCT GAT GCC AAT ACG TTT | 12 | 49 | <5000 bp |
| 4 | B02 | AGC GAG ATT TGA TAT TGC TCT CT | 11.99 | 49 | <5000 bp |
| 4 | B05 | ACA GCA AAA ATT TGA AGA GAC A | 12.03 | 49 | <5000 bp |
| 4 | B06 | ACC CAA CAA TCT AAC TAG AGT CCA | 12.08 | 49 | <5000 bp |
| 4 | B11 | TGG TCA AAC CTC CTA TCT TCA | 11.98 | 51 | <5000 bp |
| 4 | B12 | TGA ACC TTA GGT TAG TTA CCA GT | 12.02 | 51 | <5000 bp |
| 3 | A03 | ACA CTC TTA ACA GAC GAA GCG T | 12.05 | 53 | <5000 bp |
| 3 | A04 | TGT TCA ACT GAA GAA CCG TGG | 12.05 | 53 | <5000 bp |
| 3 | A09 | TGT TTG GGG ATG GAG ATC ATT GA | 11.98 | 55 | <5000 bp |
| 3 | A10 | TTC CAA TGG AGT GAC CCG TA | 12.01 | 55 | <5000 bp |
| 3 | A11 | ATA TTG GCC GGG TAG GTC CT | 11.93 | 51 | <5000 bp |
| 3 | A12 | AGA GTT TCA CCT CTG GTA ATG A | 11.93 | 51 | <5000 bp |
| 3 | B01 | TCA TTA AAT GAT GGA GTG ACG A | 11.97 | 49 | <5000 bp |
| 3 | B02 | AGC CAC TTC TAA AGA AAC TTA TCT | 11.94 | 49 | <5000 bp |
| 3 | B03 | TGT TTT GCT TAG AGA TGT GGA GA | 12.06 | 51 | <5000 bp |
| 3 | B04 | CCC AAC TTT TCA TTT TGT TGA CAG | 12.05 | 51 | <5000 bp |
| 3 | B07 | GTG TCT GTC TAA CAT ATG TTG ACA | 11.96 | 50 | <5000 bp |
| 3 | C01 | GCT CAA TTG CAA ATA AAA GAA ACG | 12 | 50 | <5000 bp |
| 3 | C02 | AGG AGG GGG TTA AAA GTT TGT | 12.02 | 50 | <5000 bp |
| 3 | C03 | TCC CTA TCC TTG GTT ATT CTT CCA | 11.94 | 53 | <5000 bp |
| 3 | C04 | CCC AGA GTC GAA AAG CTT CCT | 11.98 | 53 | <5000 bp |
| 3 | C07 | GTT CAT CCG TTC ATG CAT CCC | 11.99 | 53 | <5000 bp |
| 3 | C08 | TTT GCA ACC AGT ACG ATC AAA AGT | 11.95 | 53 | <5000 bp |
| 3 | C11 | GCT GTT ACG AGT CAA AAT ACC TGA | 12.01 | 52 | <5000 bp |
| 3 | C12 | GGA GTA TAA ACT AGT ATG TGG TGC | 11.95 | 52 | <5000 bp |
| 3 | D01 | AGT GTT TGT TTT TCT AAC TCT TGT T | 12.05 | 49 | <5000 bp |
| 3 | D02 | ACC TAC CAT GCA TGA GAC GT | 11.99 | 49 | <5000 bp |
| 3 | D07 | TCT GAT TTT CAT TGA TAA CTC CGT | 12.01 | 50 | <5000 bp |
| 3 | D08 | GCC TAT CAG GTT TTC ATG CAG C | 11.97 | 50 | <5000 bp |
| 3 | D09 | ACA TAT TTA AAG GCA AAT GTG AGA | 11.97 | 48 | <5000 bp |

| | | | | | |
|---|---|---|---|---|---|
| 3 | D10 | CCG GGG ATG GGC AGA TAA AA | 11.96 | 48 | <5000 bp |
| 3 | D11 | TGT TTG GCA GTC ATA TGT AAC GT | 11.94 | 51 | <5000 bp |
| 3 | D12 | GCT CAA TGC CAT GAA AGC GT | 11.94 | 51 | <5000 bp |
| 3 | E01 | TGG ATG CAA GTT AAA CGT GAC G | 11.95 | 54 | <5000 bp |
| 3 | E02 | AGA AGA ATG GAA TGG AGC TTG A | 11.94 | 54 | <5000 bp |
| 3 | E05 | TGT CAT GTG AGT GCA GCT ACA | 12.02 | 49 | 5000<L<10000 bp |
| 3 | E06 | CAT CTA TGC ATG CAT TAT ATG ACA | 11.96 | 49 | 5000<L<10000 bp |
| 3 | E07 | ACA TGA TTG AAA ATG AGA ACA ACC | 11.98 | 50 | <5000 bp |
| 3 | E08 | ACC ATT CCT TTT GAG GGC GA | 12.03 | 50 | <5000 bp |
| 3 | E09 | CCT TGG GTT GGA CTC AAG TGA | 12.05 | 48 | <5000 bp |
| 3 | E10 | ACT ATT TTG GAA AAA GAT TAT CTT CT | 12.06 | 48 | <5000 bp |
| 3 | E11 | TCG CAA GTA CTT GTT ATC GTA A | 11.94 | 49 | <5000 bp |
| 3 | E12 | AAA AAC CCC GTA GAA AAC AGA | 11.96 | 49 | <5000 bp |
| 3 | F03 | ACT ATG AAG CAT TGA CCA CGA | 12.05 | 52 | 5000<L<10000 bp |
| 3 | F04 | ACA ATG AGT TAA TGG GTC AAG CT | 12.02 | 52 | 5000<L<10000 bp |
| 3 | F09 | TTT TGG TGT TGG GTG TGA AT | 12.09 | 50 | <5000 bp |
| 3 | F10 | GAA GTA ATT GTG TGG TGA AAA GGC | 11.96 | 50 | <5000 bp |
| 3 | F11 | GTA CGA CCT GAT AGC GGC AA | 12 | 54 | <5000 bp |
| 3 | F12 | AAA GGG AAG GTT GCG TGT GT | 12.04 | 54 | <5000 bp |
| 3 | G01 | TTA GGC CAG TTA AAA GGA AAC GG | 12.03 | 53 | <5000 bp |
| 3 | G02 | TGG ATT TCC CAG ATT TGC CTC T | 11.99 | 53 | <5000 bp |
| 3 | G03 | TCC CAT GCA AAG TAG CTG CA | 12.01 | 48 | 5000<L<10000 bp |
| 3 | G04 | AAG CTA GTA TCT TAT CCT TTA CCA | 11.99 | 48 | 5000<L<10000 bp |
| 3 | G05 | AGT GAT AAA ATC CGT GAA AGT ACA | 12.06 | 48 | 5000<L<10000 bp |
| 3 | G06 | AAG CTA GTA TCT TAT CCT TTA CCA | 11.99 | 48 | 5000<L<10000 bp |
| 3 | G07 | AAC AAA GAG TTG GGG TGG GT | 11.99 | 53 | <5000 bp |
| 3 | G08 | TGC TTT CTT TAC TGC TCT TGG G | 11.95 | 53 | <5000 bp |
| 3 | G09 | ACA TTC TTT GTC AAA GAT GGA CA | 12.07 | 50 | 5000<L<10000 bp |
| 3 | G10 | GAT CGC AAC TTG AGC TCT TGA | 12.05 | 50 | 5000<L<10000 bp |
| 3 | G11 | ACT ACA TTC TTT GTC AAA GAT GGA | 11.94 | 50 | 5000<L<10000 bp |
| 3 | G12 | ACA ATG TCA GTG CTC CTT TGG A | 11.92 | 50 | 5000<L<10000 bp |
| 3 | H01 | TCA GAC AAA TAT TAA CCA CCA ACA | 11.93 | 50 | 5000<L<10000 bp |
| 3 | H02 | ACA TAA AAA GTT TCG TGT GGG T | 11.95 | 50 | 5000<L<10000 bp |
| 3 | H03 | TGG TCA TGT AAG ATG AAA ACC A | 12 | 46 | <5000 bp |

| | | | | | |
|---|---|---|---|---|---|
| 3 | H04 | TGT GAA ATT TGT CTA CAA AGT | 11.99 | 46 | <5000 bp |
| 3 | H05 | TGA GTT TTT GTT GGT ATA GAA TGG | 12.03 | 49 | <5000 bp |
| 3 | H06 | TCT TGC ACG GAT GGG TTC AG | 11.94 | 49 | <5000 bp |
| 3 | H07 | ACT GTT TGA ATT TGA AGA CTA GGA | 11.94 | 49 | 5000<L<10000 bp |
| 3 | H08 | AGC TTT AAG ACA AAT ACT AGT AGC T | 11.96 | 49 | 5000<L<10000 bp |
| 3 | H11 | ACC TAA ATG GGT CTG TTT TGA | 11.97 | 49 | <5000 bp |
| 3 | H12 | CAA CTT CTC CCC TTC CCA TGT | 12.06 | 49 | <5000 bp |
| 2 | A01 | TGA TGT ACA AGC ACT ACT ACA GCT | 12.05 | 53 | <5000 bp |
| 2 | A02 | AGT ATC ACC TGG CAA GTG GT | 12 | 53 | <5000 bp |
| 2 | A03 | GGT GAT TTC AGA AGT GAT AAA AGA | 11.97 | 49 | <5000 bp |
| 2 | A04 | ACT TGT AAC TCT TTA TCC ATT AAG T | 11.96 | 49 | <5000 bp |
| 2 | A05 | CTT TAT AAC TAC TTT GGT CCT CCA | 12.07 | 50 | <5000 bp |
| 2 | A06 | ACA GCG CAT TAC TGA CTC TT | 12 | 50 | <5000 bp |
| 2 | A07 | TTC AAA ATT TGT AAT GAC TTT GGC | 11.98 | 49 | <5000 bp |
| 2 | A08 | AGG GAT CTA TTA AGT TCC AAG GA | 12.01 | 49 | <5000 bp |
| 2 | A09 | AGG AGA CCT ATC CTC TTT GCC T | 11.94 | 51 | <5000 bp |
| 2 | A10 | TGT GCA ACA TAA CCC TTA AAA GT | 11.93 | 51 | <5000 bp |
| 2 | B01 | AGT CGC TGA TTG TTG ATT CCA | 12.05 | 52 | <5000 bp |
| 2 | B02 | TCG AAC CAA GTT ATC TTT CTT CGT | 12.02 | 52 | <5000 bp |
| 2 | B03 | GGG ACG TGA AAG ATC CTA ACC T | 12.04 | 50 | <5000 bp |
| 2 | B04 | ATA TAG CCT AGC TCA CAG GAA | 12.02 | 50 | <5000 bp |
| 2 | B05 | TGT TAA GCA AAT CAG CCG GC | 12.07 | 54 | <5000 bp |
| 2 | B06 | CCT CGT TGG TTA ACA CAA GCG | 11.98 | 54 | <5000 bp |
| 2 | B11 | ACG TAA GAA ATG GGT CCA ATG T | 11.99 | 52 | <5000 bp |
| 2 | B12 | GGT GGT TGG TTT CAC GTT AGA C | 11.93 | 52 | <5000 bp |
| 2 | C03 | ACT CCA TGA CTG AAA CAA ACT | 11.97 | 50 | <5000 bp |
| 2 | C04 | AAA ATT CAA GAC TTT TGG TTC TGA | 11.99 | 50 | <5000 bp |
| 2 | C05 | AGT GGA GGT TAG AGA AGC ATG A | 12.04 | 52 | <5000 bp |
| 2 | C06 | TCT TCT TCA GAT CAA CAC ATG CT | 12.07 | 52 | <5000 bp |
| 2 | C07 | CAG AAA ATT TTA AGC CAA AGT GGG | 12.01 | 48 | <5000 bp |
| 2 | C08 | TAA GAA ACT TGT GAA TTT TAT CCT CT | 12.05 | 48 | <5000 bp |
| 2 | C11 | TCG AAT GCA ATT TGG GAA GCA | 12.05 | 53 | <5000 bp |
| 2 | C12 | CGA GTT AAT ACC TAC ACA CAC GTG | 11.94 | 53 | <5000 bp |
| 2 | D03 | AGA TTG TTG TTC GTT GGA GCA | 11.95 | 52 | <5000 bp |

| | | | | | |
|---|---|---|---|---|---|
| 2 | D04 | AAG ATG GGT GTC AGT TGG GC | 11.99 | 52 | <5000 bp |
| 2 | D05 | AGT GAT AGG AAT TGA ATG AAT GGA | 12.04 | 50 | <5000 bp |
| 2 | D06 | TCA GCT CGA AGT CAT GCC TT | 12.03 | 50 | <5000 bp |
| 2 | D07 | TGC CTA ACA AAA TAC ATT ATT GCA | 11.93 | 49 | <5000 bp |
| 2 | D08 | GAG CAC CAT ATA AAT GAC GAA AA | 12.07 | 49 | <5000 bp |
| 2 | E01 | CTA TAG TAT TGT GAT GCA CTT TCC | 11.99 | 48 | >10000 bp |
| 2 | E02 | ATT ATG AGT GGT AGA CTC GT | 11.93 | 48 | >10000 bp |
| 2 | E03 | TCT TAT TAG TGG ATA GTG TGC CA | 11.97 | 48 | <5000 bp |
| 2 | E04 | ATT ATG AGT GGT AGA CTC GT | 12.02 | 48 | <5000 bp |
| 2 | E05 | ACA TGG TCT ATC AAA ATC TTG GGC | 11.96 | 53 | <5000 bp |
| 2 | E06 | TGA GAC GGG TAC AAA ATC CAG T | 11.96 | 53 | <5000 bp |
| 2 | E09 | CGG CTA TTC ACA ATC ACA ATG GT | 12.06 | 50 | <5000 bp |
| 2 | E10 | TGT TGT TTT TGC CAT TAC ATT GG | 11.96 | 50 | <5000 bp |
| 2 | F01 | ACC AAC ACC GGA GTG AGA AT | 12 | 49 | <5000 bp |
| 2 | F02 | TTG ACA TTT GAA CGG AGT ATT TT | 12.03 | 49 | <5000 bp |
| 2 | F05 | CCA AGT CAA CGG ATT TGT GA | 11.99 | 51 | <5000 bp |
| 2 | F06 | AGC GAA GGA ATT GTG AGG GT | 11.99 | 51 | <5000 bp |
| 2 | F07 | ACC AAC ACC AAG TAA CCG TAA GT | 11.96 | 50 | <5000 bp |
| 2 | F08 | TGT ATG TTG ATT GGA TTG AGT TTT | 11.91 | 50 | <5000 bp |
| 2 | F09 | TCT CAA TGA AAT GCA TGT GGA CA | 12 | 49 | <5000 bp |
| 2 | F10 | TTT ATC TTT GAA AAT GTG ACG TCT | 11.97 | 49 | <5000 bp |
| 2 | F11 | TGT CAT CGG CTT TCA CTT AGA | 12.03 | 50 | <5000 bp |
| 2 | F12 | GAC GAG GGG AGA TAG CAT TCT | 11.97 | 50 | <5000 bp |
| 2 | G01 | GCG ATG GCT TTC CAA TCA AA | 11.96 | 52 | <5000 bp |
| 2 | G02 | ACA GCA AGA AGG ACT TTG CAG | 12 | 52 | <5000 bp |
| 2 | G03 | TGA AAA GTA GAT GTT AAC ACG AGC | 11.97 | 49 | <5000 bp |
| 2 | G04 | ACT GAA TGA ATA GGA AGT TGG T | 11.99 | 49 | <5000 bp |
| 2 | G05 | TGA GGC TAA ATA TTC CAT CAT CGA | 11.96 | 48 | <5000 bp |
| 2 | G06 | TAG TTT GAA ACT GAT TAA TGA AGG A | 12.07 | 48 | <5000 bp |
| 2 | G07 | ACA GTC TAA AAC ACA AGC TAG T | 12.02 | 49 | <5000 bp |
| 2 | G08 | CGT TGG GCA GCA TTT TTA CT | 11.97 | 49 | <5000 bp |
| 2 | G11 | ATC CCT TTG AAT AAA TTC TAT AAC TCA | 11.93 | 48 | <5000 bp |
| 2 | G12 | TCT CAA TTG GAT TCC TAA ACT AGT | 12.07 | 48 | <5000 bp |
| 2 | H01 | TCA AGA CCA TTG TCA CCG AA | 11.93 | 52 | <5000 bp |

| | | | | | |
|---|---|---|---|---|---|
| 2 | H02 | TGC ACA AGG CAT GGT ATC AA | 12.06 | 52 | <5000 bp |
| 2 | H03 | TGA CCA GCG TCA CTA TCC TT | 11.95 | 54 | <5000 bp |
| 2 | H04 | GTT GGC AAT TCA CGC ATG GT | 11.99 | 54 | <5000 bp |
| 2 | H05 | ATC GCA CGG TTA ACA TCA GT | 12.03 | 52 | <5000 bp |
| 2 | H06 | TGT CGG TCT ATT TTG GCT TGG A | 11.96 | 52 | <5000 bp |
| 2 | H07 | TGC AGT TCA TGT TAG GGG AGG | 12.02 | 53 | <5000 bp |
| 2 | H08 | TCC TCT TGT CAA ACA GCA TCT C | 11.96 | 53 | <5000 bp |
| 2 | H11 | GGA CCG GAG GGA ATA ATC TGT | 11.93 | 56 | <5000 bp |
| 2 | H12 | TGT ACA CCT TGT GCA TGT AGC T | 11.91 | 56 | <5000 bp |

Table 3.2: PCR Conditions for SequalPrep™ Long PCR Kit

| Component | Volume |
|---|---|
| SequalPrep™ 10X reaction buffer | 1 ul |
| DMSO | 0.2 ul |
| SequalPrep™ 10X enhancer A | 1 ul |
| SequalPrep™ Long Polymerase, 5 U/ul | 0.18 ul |
| Primer mix (10 uM each) | 1 ul |
| Template DNA (50 ng/ul) | 1 ul |
| DNase-free water | 5.62 ul |

Table 3.3: PCR Conditions for Velocity DNA Polymerase kit

| Component | Volume |
|---|---|
| SequalPrep™ 10X reaction buffer | 1 ul |
| DMSO | 0.2 ul |
| SequalPrep™ 10X enhancer A | 1 ul |
| SequalPrep™ Long Polymerase, 5 U/ul | 0.18 ul |
| Primer mix (10 uM each) | 1 ul |
| Template DNA (50 ng/ul) | 1 ul |
| DNase-free water | 5.62 ul |

Table 3.4: PCR Conditions for Taq DNA Polymerase kit

| Component | Volume |
|---|---|
| 10X KCL buffer (+Mg) | 1 ul |
| 100 mM dNTP mix | 0.2 ul |
| Taq DNA Polymerase | 0.2 ul |
| Primer mix (10 uM each) | 0.4 ul |
| Template DNA (50 ng/ul) | 1 ul |
| DNase-free water | 7.2 ul |

Table 5: Standard Cycling Conditions

| Step | Temp | Time | Repeat |
|---|---|---|---|
| Initial denaturation | 98ºC | 2 min. | 1 |
| Denaturation | 98ºC | 30 sec. | 25-35 |
| Annealing | $T_m$ | 30 sec. | |
| Extension | 72 ºC | 15-30 sec/kb | |
| Final extention | 72 ºC | 4-10 min. | 1 |

Table 3.6: Long-Range PCR protocol

| Step | Temp | Time | Repeat |
|---|---|---|---|
| Initial denaturation | 94ºC | 2 min. | 1 |
| Denaturation | 94ºC | 30 sec. | 10 |
| Annealing | $T_m$ | 30 sec. | |
| Extension | 68 ºC | 15-30 sec/kb | |
| Denaturation | 94ºC | 30 sec. | 20-30 |
| Annealing | $T_m$ | 30 sec. | |
| Extension | 68 ºC | 15-30 sec/kb | |
| Final extention | 72 ºC | 5 min. | 1 |

Table 3.7: Final Concentration of Each sample after pooling

| Sample ID | Pooled concentration |
|---|---|

| | |
|---|---|
| PI 548463 | 114.8 ng/ul |
| PI 224271 | 75.3 ng/ul |
| PI 72232 | 49.5 ng/ul |
| PI 593647 | 93.4 ng/ul |
| FC 30692 | 51.5 ng/ul |
| PI 64698 | 57.9 ng/ul |
| PI 315701 | 106.0 ng/ul |
| PI 70587 | 178.8 ng/ul |
| PI 30594 | 75.3 ng/ul |
| PI 54855 | 50.9 ng/ul |
| PI 398471 | 95.6 ng/ul |
| PI 54853 | 68.6 ng/ul |
| PI 84807 | 145.5 ng/ul |
| PI 86142 | 71.6 ng/ul |
| PI 230972 | 103.0 ng/ul |
| PI 86144 | 105.8 ng/ul |

Table 3.8: Sample List and Corresponding barcodes

| Soybean ID | Flower Color | Seed Coat Color | Barcode |
|---|---|---|---|
| PI 548463 | Light purple | Black | Ion Xpress™ Barcode 1, 1 tube, 20 µL |
| PI 224271 | Blue | Yellow | Ion Xpress™ Barcode 2, 1 tube, 20 µL |
| PI 72232 | Dark purple | Yellow | Ion Xpress™ Barcode 3, 1 tube, 20 µL |
| PI 593647 | White | Yellow | Ion Xpress™ Barcode 4, 1 tube, 20 µL |
| FC 30692 | White | Brown | Ion Xpress™ Barcode 5, 1 tube, 20 µL |
| PI 64698 | White | Black | Ion Xpress™ Barcode 6, 1 tube, 20 µL |
| PI 315701 | Dilute purple | Black | Ion Xpress™ Barcode 7, 1 tube, 20 µL |
| PI 70587 | Purple | Reddish brown | Ion Xpress™ Barcode 8, 1 tube, 20 µL |
| PI 30594 | White | Green | Ion Xpress™ Barcode 9, 1 tube, 20 µL |
| PI 54855 | White | Reddish brown | Ion Xpress™ Barcode 10, 1 tube, 20 µL |
| PI 398471 | Purple | Reddish buff | Ion Xpress™ Barcode 11, 1 tube, 20 µL |
| PI 54853 | Purple | Gray | Ion Xpress™ Barcode 12, 1 tube, 20 µL |
| PI 84807 | Purple | Grayish green | Ion Xpress™ Barcode 13, 1 tube, 20 µL |
| PI 86142 | Purple | Buff | Ion Xpress™ Barcode 14, 1 tube, 20 µL |
| PI 230972 | White | Reddish buff | Ion Xpress™ Barcode 15, 1 tube, 20 µL |
| PI 86144 | Dark purple | Buff | Ion Xpress™ Barcode 16, 1 tube, 20 µL |

Table 3.9: qPCR protocol

| | | | |
|---|---|---|---|
| Cycle 1: | ( 1X) | | |
| | Step 1: | 95.0ºC | for 05:00 |
| Cycle 2: | ( 35X) | | |
| | Step 1: | 95.0ºC | for 00:30 |
| | Step 2: | 60.0ºC | for 00:45 |
| | Data collection enabled. | | |
| Cycle 3: | ( 80X) | | |
| | Step 1: | 55.0ºC | for 00:10 |
| | Increase setpoint temperature after cycle 2 by 0.5ºC | | |
| | Melt curve data collection and analysis enabled. | | |

Table 3. 10: Well Information and Library ISP Details

| **Addressable Wells** | **1,262,519** | |
|---|---|---|
| With ISPs | 889,399 | 70.4% |
| Live | 884,188 | 99.4% |
| Test Fragment | 4,298 | 00.5% |
| Library | 879,890 | 99.5% |
| **Library ISPs** | **879,890** | |
| Filtered: Polyclonal | 551,199 | 62.6% |
| Filtered: Low Quality | 97,990 | 11.1% |
| Filtered: Primer Dimer | 377 | 00.0% |
| **Final Library ISPs** | **230,324** | **26.2%** |

*Table 3.11: Run summary*

| Barcode Name | Sample | Bases | $\geq Q20$ | Reads | Mean Read Length |
|---|---|---|---|---|---|
| IonXpress_007 | 7 | 1,036,599 | 552,193 | 24,727 | 41 bp |
| IonXpress_001 | 1 | 7,934,456 | 5,883,322 | 80,587 | 98 bp |
| IonXpress_002 | 2 | 1,082,815 | 808,108 | 9,691 | 111 bp |
| IonXpress_003 | 3 | 1,550,077 | 1,183,387 | 12,742 | 121 bp |
| IonXpress_004 | 4 | 5,323,840 | 4,020,540 | 43,718 | 121 bp |
| IonXpress_005 | 5 | 1,453,340 | 1,102,365 | 12,278 | 118 bp |
| IonXpress_006 | 6 | 755,743 | 575,785 | 6,165 | 122 bp |
| IonXpress_008 | 8 | 1,328,666 | 1,002,606 | 11,519 | 115 bp |
| IonXpress_009 | 9 | 144,787 | 110,161 | 1,250 | 115 bp |
| IonXpress_010 | 10 | 897,937 | 683,774 | 7,737 | 116 bp |
| IonXpress_012 | 12 | 852,023 | 653,637 | 7,108 | 119 bp |
| IonXpress_013 | 13 | 183,408 | 139,462 | 1,517 | 120 bp |
| IonXpress_014 | 14 | 740,867 | 564,798 | 6,296 | 117 bp |
| IonXpress_016 | 16 | 426,477 | 321,433 | 3,616 | 117 bp |

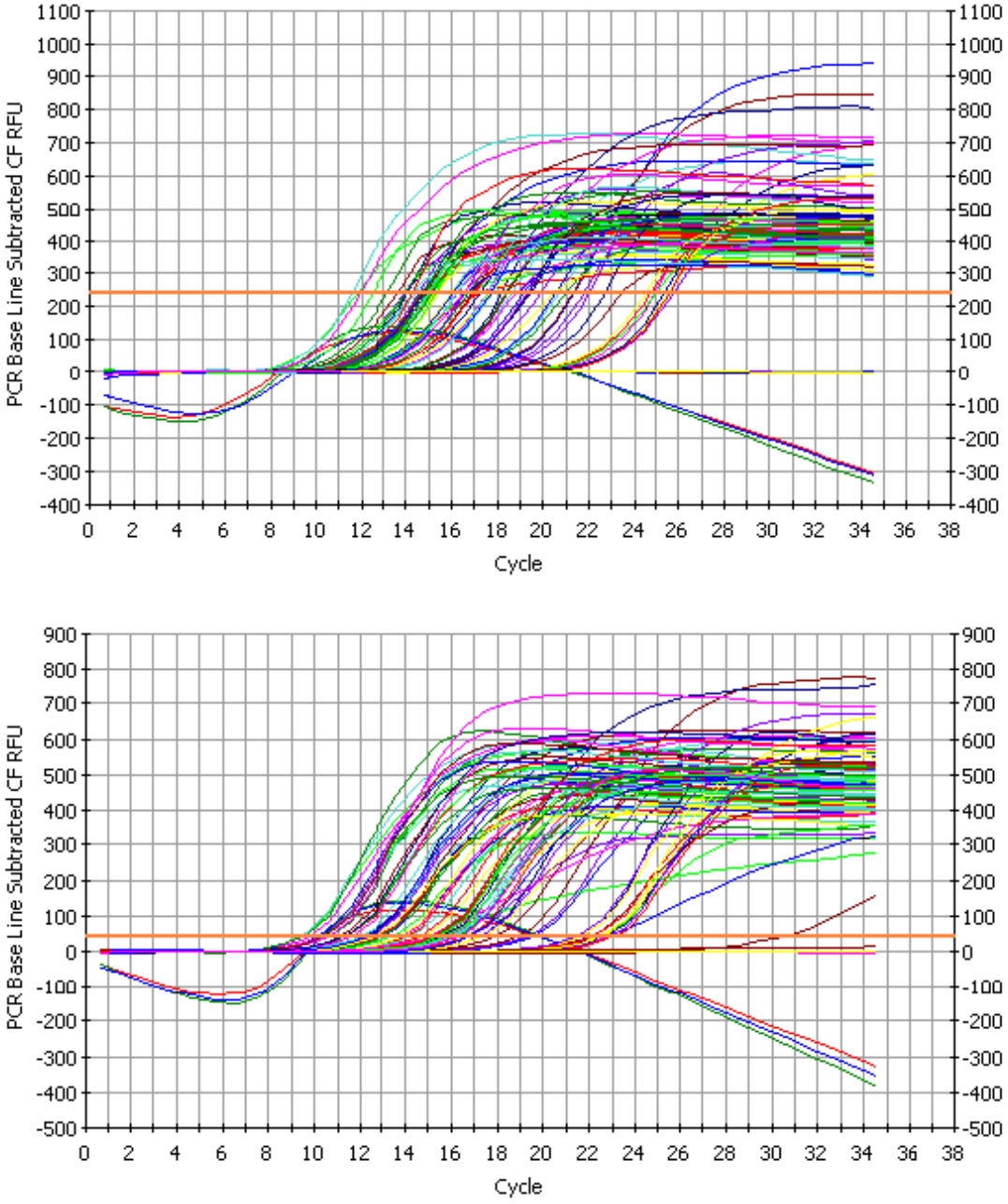



Figure 3.1: PCR Amp/Cycle Graph for SYBR-490 –first and second plate

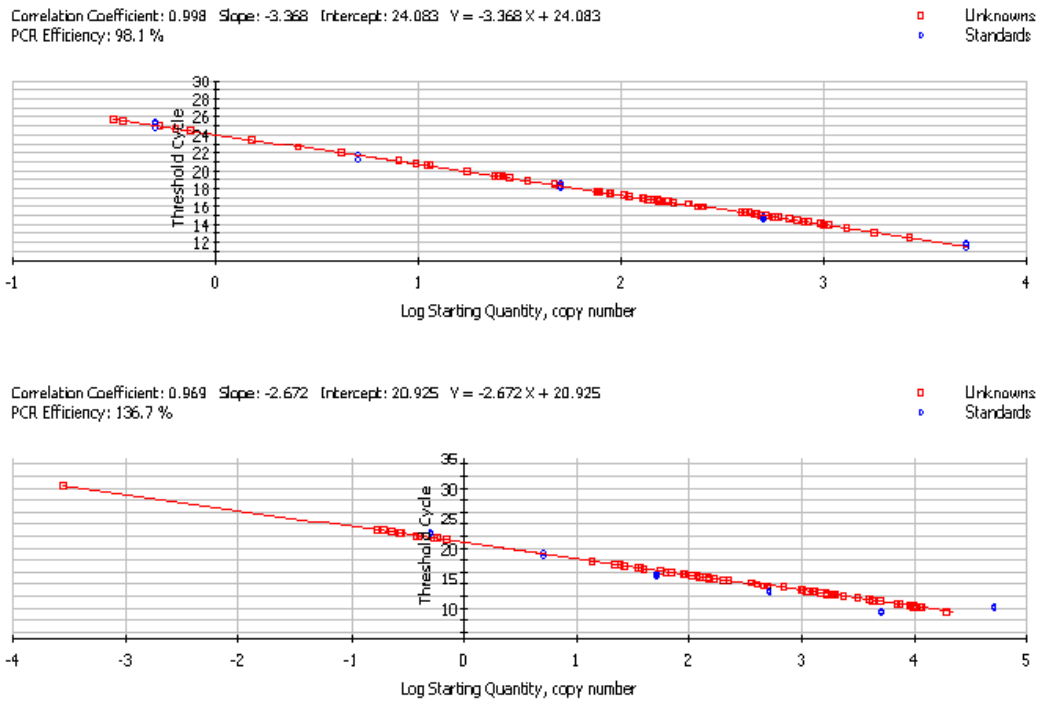



Figure 3.2: Standard Curve Graph for SYBR-490 –First and Second Plate

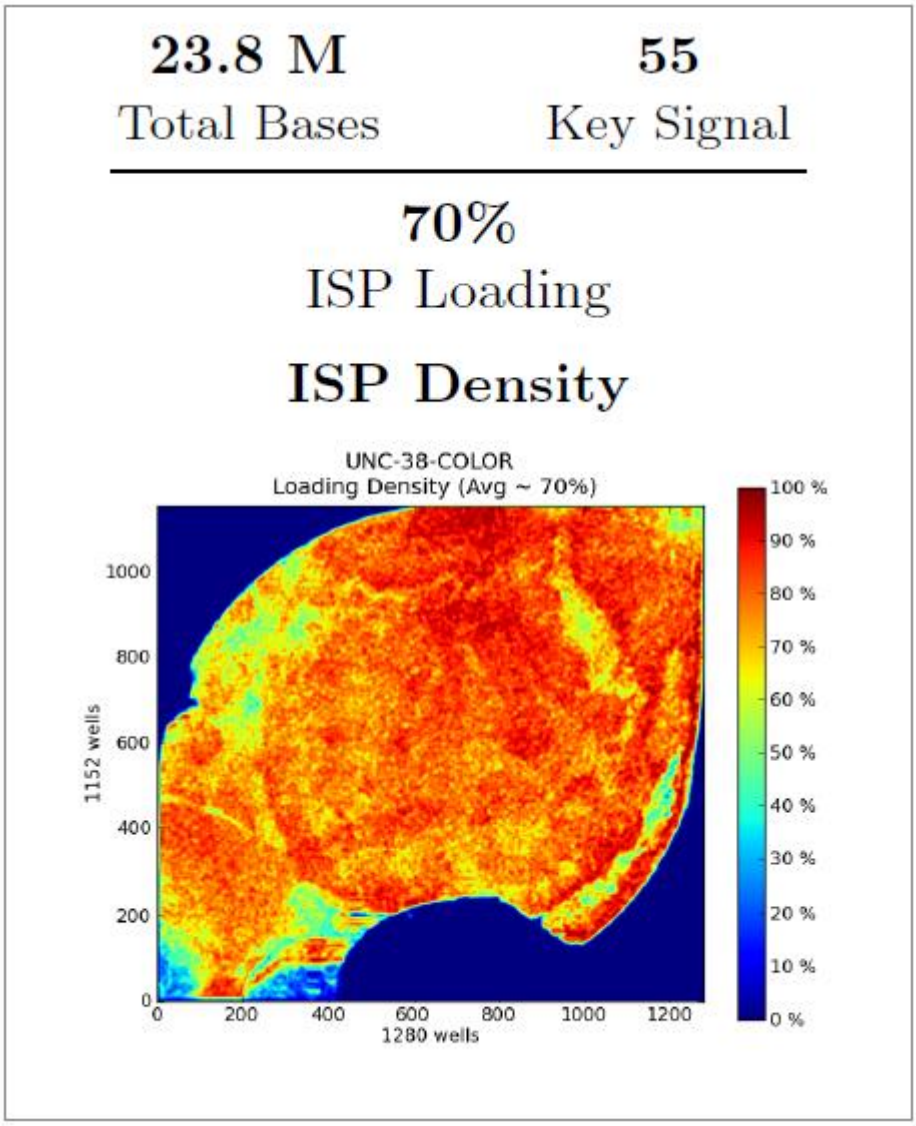



Figure 3.3 : Ion torrent ISP Density

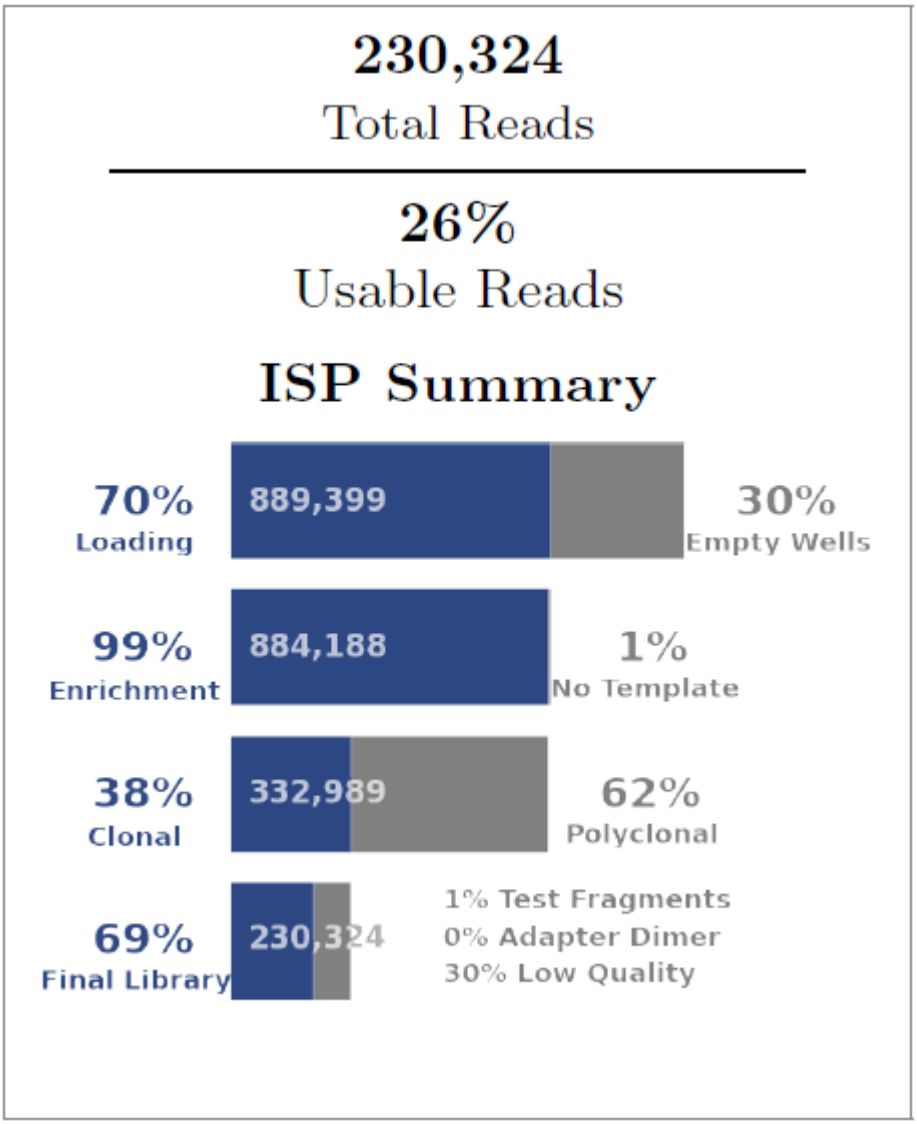



Figure 3.4: The ISP summary: Total number of filtered and trimmed reads independent of length reported in the output BAM, and FASTQ files; Percentage of chip wells that contain a live ISP; Predicted number of Live ISPs that have a key signal identical to the library key signal; Percentage of clonal ISPs; Percentage of polyclonal ISPs (ISPs carrying clones from two or more templates).

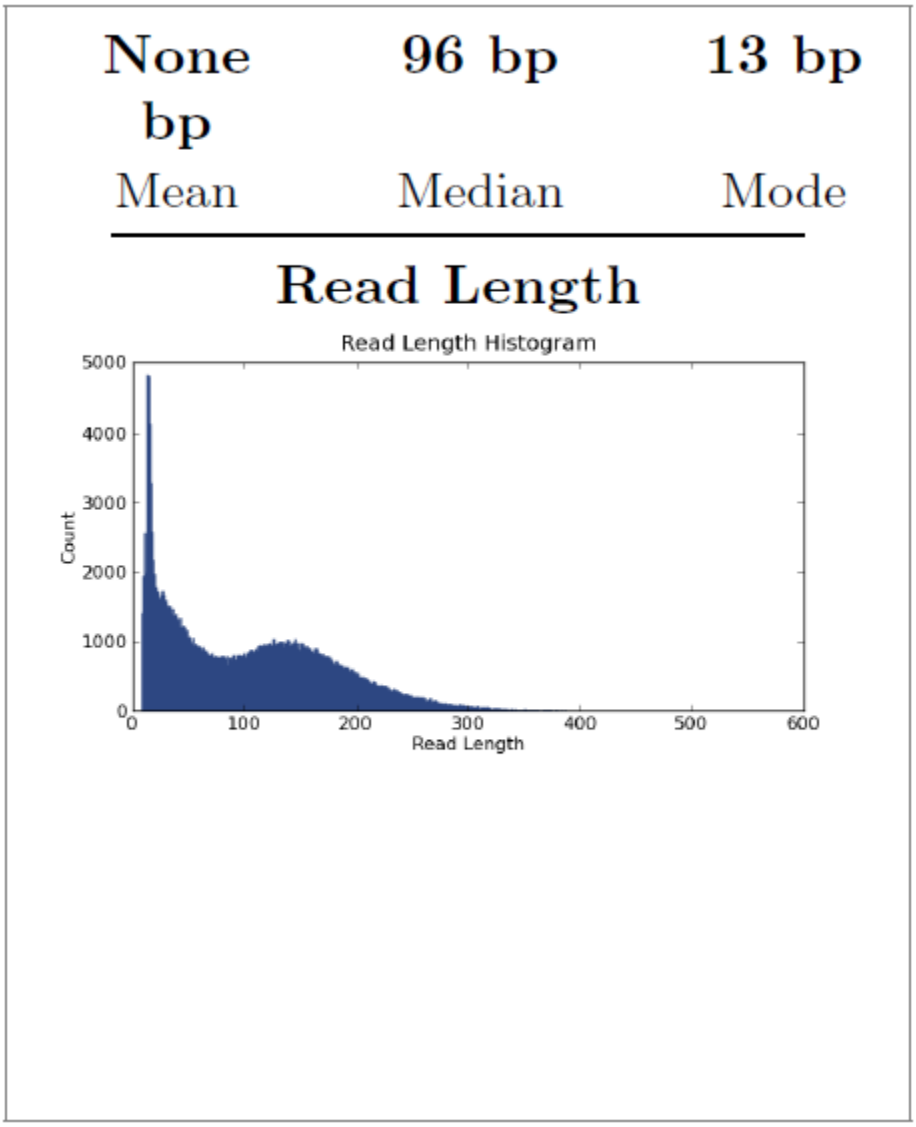



Figure 3.5: Average length, in base pairs, of all filtered and trimmed library reads reported in the output BAM, and FASTQ files.

Fi

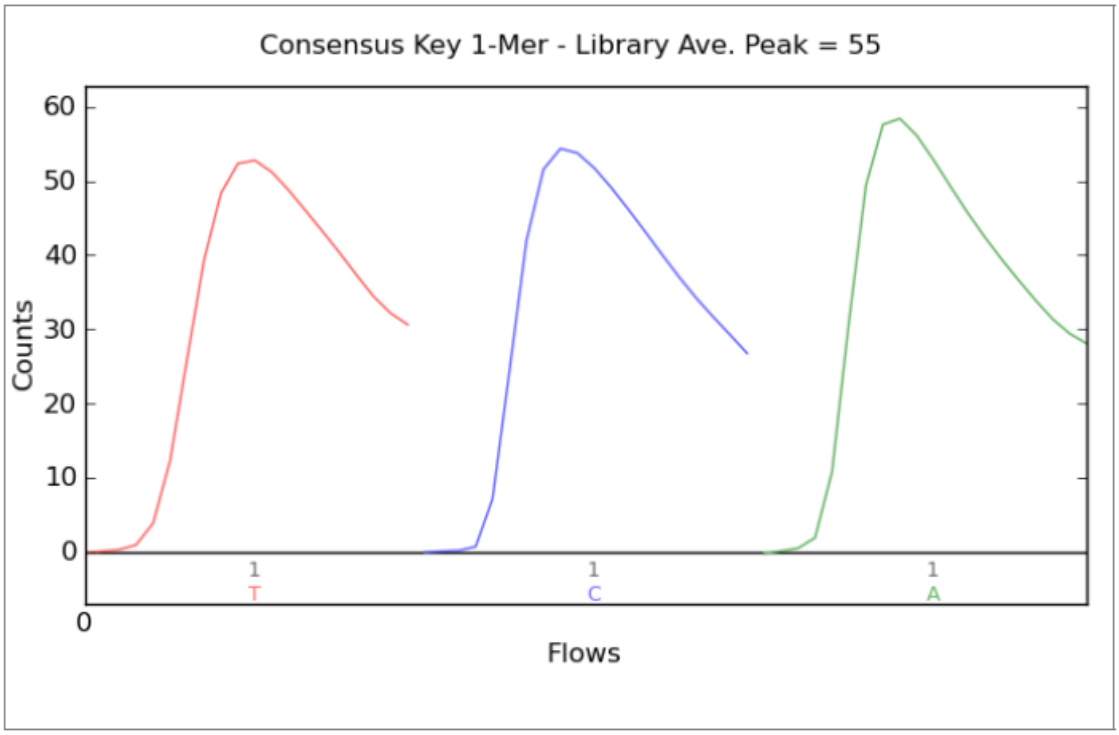



Figure 3.6: The Consensus Key 1-Mer graph shows the strength of the signal from the first three one-mer bases of the Test Fragment key.

# CHAPTER 4: ANNOTATION AND COMPARATIVE GENOMIC ANALYSIS OF ISOFLAVONE SYNTHASE REGIONS ACROSS THE PERENNIAL GLYCINES

## 4.1 Introduction

With the advent of next-generation sequencing and consistent assembly protocols, the comparative analysis of genomes is possible and has presented an opportunity to interpret these assemblies to understand biological processes at a comparative level (Stein L. 2001). The interactions and function of cellular components, alterations over evolutionary time, and their spatial location can be characterized in four different levels of genome annotation. If only genes are identified, and their predicted or known function is assigned to them, we are looking at a one dimensional genome annotation. This level of annotation is widely known as the "genome annotation". However, other levels of details can also be annotated (Reed J. 2006). When the annotation specifies interactions such as protein-protein interactions, regulatory interactions, or metabolite transformations that could be used to reconstruct a network, the annotation is called two dimensional. Moreover, three dimensional annotation can give us information on the inter-cellular arrangement of chromosomes and other cellular components, leading us to more data about the functionality of them. Finally, if one is seeking information on changes that happen to the genome sequences during adaptive evolution, they should utilize four dimensional genome annotation (Reed J. 2006). The other three levels of annotation, the

network reconstruction, ultra-structural reconstruction, and genome plasticity and new network states, are time consuming, very experimentally intensive, and outside of the scope of this study.

One-dimensional annotation starts with the identification of genes. It is also the most visible part of this process. Gene finding processes can vary from species to species. For example, gene finding in small prokaryotic genomes, mainly involves identifying long open reading frames (ORFs). However, some ambiguity might arise as well. The process becomes more complex as the genome size increases; signal to noise ratio being the main problem (Stein L. 2001). In addition to the size of the genome, the proportion of the genome that is actually coding dictates the complexity of the gene finding process. For instance, only eighty five percent of Haemophilus influenzae genome (Fleischmann et al. 1995), seventy percent of yeast's genome (Goffeau et al. 1996), twenty five percent of worm's genome (The C. elegans Sequencing Consortium, 1998), and less than two percent in humans (International Human Genome Sequencing Consortium, 2001) are coding genes. In eukaryotic genomes, annotation is further complicated with the presence of introns and exons as well as alternative splicing sites. Moreover, the size of introns and exons varies between genomes. For example, in the human genome, we can face impediments in locating the start and stop positions of a gene, or the splicing patterns of an exon. These can be caused by, among other things, exons that are smaller than introns, ambiguity between intergenic regions before exons, and adjacent genes (Stein L. 2001).

In light of these challenges, a number of gene prediction software algorithms have been developed specifically for eukaryotic genomes, including GENSCAN (Burge C. and Karlin S. 1997), Genie (Reese MG. et al. 2000), GeneMark.hmm (Besemer J. and

Borodovsky M. 1999), Grail (Uberbacher E. and Mural R. 1991), HEXON (Solovyev V. et al. 1994), MZEF (Zhang M. 1997), Fgenesh (Solovyev V. et al. 1995), GeneFinder (P. Green, U. Washington), and HMMGene (Krogh A. 1997). In general, these algorithms look for splice sites or scan the genome for signatures such as GC rich regions, since they are associated with transcriptional start sites and transcribed regions. Some exon predictors such as HEXON and MZEF, base their prediction on finding only one feature and will stop the process as soon as one feature is detected. However, most of these algorithms will generate a gene model by using the output of multiple signatures; computed by: neural networks (Grail), a rule based system (GeneFinder), or with a hidden markov model (HMM) (GenScan, Genie, HMMGene, GeneMark.hmm and Fgenes). The HMM has the ability to model how the individual probabilities of a sequence of features are combined into a probability estimate for the whole gene (Stein L. 2001).

Given the economic importance of soybean, it should be no surprise that the majority of the research attention has focused on this particular species from the Glycine genus. There is, however, an untapped resource in the related perennial Glycine. This genus has two subgenus, subgenus *soja* contains cultivated soybean, *Glycine max*, and subgenus *Glycine*, includes the perennial Glycines. The members of this subgenus are diploid genomes and are paleoallopolyploids. Although the polyploidy event in this genus has happened over 15 million years ago, some members of the genus, like *Glycine dolichocarpa*, part of the tomentella complex, have a more recent polyploidy event with many of them displaying recurrent origins. Despite these similarities, there are many notable variations among certain closely related polyploids (Doyle et al. 2004a). These

variations have been observed and documented for characteristics such as the differences of morphology (Newell and Hymowitz 1978, and Costanza and Hymowitz 1987), chromosome numbers (Newell and Hymowitz 1978), pathogen resistance (Burdon and Marshall 1981b), nuclear ribosomal genes (Doyle and Beachy 1985, and Doyle et al. 1989), seed protein (Mies and Hymowitz 1973), isozymes (Broue et al. 1977, Grant et al. 1984, and Doyle and Brown 1985), and flavonoid chemistry (Vaughan and Hymowitz 1984). Given this diversity and the phylogenetic position close to soybean, these taxa have been chosen as a germplasm source to improve the current cultivated soybean. One of the approaches to better understand these differences is identifying the homeologs in seven perennial species, by choosing homeologous regions of soybean. Chromosome 7 and 13 in *Glycine max* contain isoflavone synthase genes and were chosen target regions of the data used in this study. *Glycine dolichocarpa* is the only species that should have four distinct loci of isoflavone synthase as it is a recent polyploid.

## 4.2 Materials and Methods

### 4.2.1 Genome Sequencing Strategy, and Assembly

Isoflavone synthase containing regions were identified from *Glycine max* genome. BAC libraries were designed for each of these species and then end-sequenced. Those sequences were aligned to the soybean genome to place the BACs into a framework ordering. Orthologous BACs for the targeted isoflavone synthase regions were subsequently identified and chosen for sequencing. The sequencing was done at the Hudsonalpha Institute. Regions were assembled using the standard assembly procedures

at Hudsonalpha as part of the larger collaborative SoyMap II research project. As part of the SoyMap II project, RNA-seq data derived from each of the seven perennial *Glycines* have been generated by the Perry Cregan's group at USDA-ARS. The RNA-seq data was generated as part of a linkage disequilibrium analysis, but serves as an excellent resource for annotation. The detail of data generated is shown in Table 4.1.

### 4.2.2 Genome, Functional, and GO Annotation

Genome annotation of these targeted regions was performed using the MAKER pipeline (Holt and Yandell 2011) rather than working with a single gene predictor. MAKER has been used in re-annotation of *Zea mays* (C Lawrence, MaizeGDB), *Oryza sativa* (R Buell, MSU), and *Arabidopsis thaliana* genomes (E Huala, TAIR). Both *ab initio* predictions and sequence (RNA-seq) based validations of gene models are possible within the framework of MAKER. The pipeline begins by utilizing RepeatRunner (Smith et al. 2007) to identify transposable elements and viral proteins for masking. Following this step, *ab initio* gene predictors are run to produce preliminary gene models. This prediction is based on models that describe patterns of intron/exon structure and consensus start signals. Because this pattern can vary in different organisms, training of our prediction methods was necessary for improved gene prediction. MAKER supports four different gene prediction programs. For this study, we chose to run both SNAP (Korf 2004) and Augustus (Stanke et al. 2006) as our gene predictor software. We did not use FGENESH (Salamov and Solovyev 1998) because it was not free, or GeneMark (Borodovsky and McIninch 1993) based on the problems that have been reported with it, especially in dealing with fragmented genomes, and long introns.

Before the start of the pipeline, we trained Augustus, using *Glycine max* gene models and autoAug.pl script. The advantage of using Augustus is that it takes into account the intron containing genes. For the first MAKER run, the SNAP input was *Arabidopsis thaliana* hmm files. These files are provided in the SNAP package. The est2genome option was used in this run as a resolution to the absence of a trained gene predictor. MAKER allows for two forms of EST evidence; one for the species that is being annotated, and one that serves as an alternative from a closely related species. We combined all the RNA-seq data for all seven Glycine species and used it as the first EST evidence (Table 4.1). We also used *Glycine max* RNA file for our alternative EST evidence. This file was obtained from NCBI's genome ftp page. *Glycine max* protein file was also obtained from NCBI's genome ftp page, and used as protein homology evidence (protein2genome option). MAKER runs both TBLASTX and BLASTX to align mRNA and protein evidence sequences to the contigs. BLAST is able to extend through simple masked regions, and lower case sequences, while avoiding any alignment seeding in these regions. Exonerate (Slater and Birney 2005) is then used to polish BLAST hits. The est2genome and protein2genome options enable MAKER to find exon-intron positions by realigning the blast hits around splice sites in order. We also provided the known plant repeat library for the RepeatRunner, as well as the transposable elements file, which came with MAKER.

MAKER can take *ab initio* predictions, EST alignments, and protein alignments to generate models to predict where splice sites and protein coding regions are located. It then passes these models to the gene prediction programs. Using a modified sensitivity/specificity distance metric, MAKER chooses the best possible gene model that

is supported by the multiple streams of evidence. Using evidence from EST alignments, it can also revise gene models. Finally, MAKER calculates quality control statistics, such as length of the 5' and 3' UTR, number of exons in the mRNA, and length of the protein sequence produced by the mRNA. The MAKER annotation workflow is illustrated in Figure 4. 1.

We ran MAKER for a total of three runs for each species. Before MAKER was rerun, we merged all the gff3 files into one file, and used that to train and retrain SNAP allowing each run to incrementally increase in confidence. The scripts used in this step are maker2zff, fathom (Stein 2007), forge, and hmmassembler which are all included in SNAP. MAKER allows re-annotations by providing an option to introduce the gff file from the previous MAKER run. We need to convert gff files to zff format since that is the accepted format for fathom. Fathom filters the input gene models. Fathom's output files are provided to forge. Forge catches the genomic sequence that is surrounding each model locus immediately. Hmmassembler uses these segments to produce the hmm file. We ran MAKER for a third and final time, by providing the training models and gff files from the second run. The steps to train the models are similar to previous run.

The annotations were visualized and edited using Jbrowse (Skinner M. et al. 2009). Manual selection of final annotations was performed based on the evidence and blasting one more time. Fathom (part of the snap gene predictor) and Unix commands were used to extract useful statistical information from the results such as average GC content, number of sequences, genes, single-exon, multi-exon, and mean of the exon and intron lengths. Shell scripts were used to automate the pipeline. SyMAP v4.2 (Suderlond et al. 2006) was used to detect and display syntenic relationships between the isoflavone

synthase regions. Interproscan (Quevillon E. et al. 2001) with default options was used to identify GO annotations. BLAST2GO (Conesa A. et al. 2005) with default setting was used to functionally annotate the predicted protein models. To evaluate if there was any overrepresented conserved genes in any of the GO categories, we used the Fisher's exact test.

### 4.2.3 Testing for Selection

To look for potential signatures of positive selection, from the MAKER output, predicted genes annotated in each species were aligned using MEGA6 (Kumar et al. 2001). The default setting was used unless otherwise stated. MEGA was chosen over methods like MAFFT to speed up the analysis since its analyses are performed on a genomic scale. These alignments were also used to generate a phylogenetic tree with MEGA. Evolutionary distances between conserved genes were estimated with the pairwise-deletion option of MEGA. A Z-test of selection comparing dN and dS was performed as detailed in Nei & Kumar (2000) within MEGA. Disparity in substitution among nucleotide sites (Disparity Index), and tests of substitution homogeneity (Kumar and Gadagkar 2001) of the aligned sequence was performed within MEGA.

## 4.3 Results and Discussion

### 4.3.1 Annotations

Two soybean regions (chromosome 7 and chromosome 13) anchored by the homeologous genes isoflavone synthase were chosen for targeted resequencing. These two chromosomes fall into well characterized homologous blocks identified from the soybean reference genome. We are studying each homeologous region (four regions in *Glycine dolichocarpa*) anchored by isoflavone synthase genes across each *Glycine species*. Basic information about each genome and its predicted genes are shown in tables 4.2-4.15. Final protein coding gene count for all species is 543 (out of 845 annotated). Although initially, both SNAP and Augustus had predicted new and novel genes, but since a manual control was performed, all of these final annotations have transcript or protein similarity support. All the final annotations and their corresponding orthologs from soybean are shown in Table 4.16. All the gff3 files containing annotations, are included in supplementary data.

### 4.3.2 Synteny and Functional annotation

Genome synteny alignments allows us to analyze relative gene-order conservation between sequenced regions of the perennials and soybean, and was performed using SyMAP 4.2 (Soderlund, C. et al. 2006). Shared synteny implies that the conserved regions have originated from an identical ancestor, and possibly have retained similar functions in their genes. The syntenic relationships between homeologous regions of each species,

as well as between species are shown in Figures 4.2-4.15. The synteny covers almost the entire sequenced region, suggesting accurate assembly and annotation. Figures 4.4, 4.8, 4.9, 4.10, and 4.13 show syntenic relationships between species while the rest show homeologous synteny within a single genome. Synteny between homeologs within a genome is generally stronger than the synteny between *Glycines.* This is a bit surprising given that the paleopolyploidy event leading to the homeologous relationship occurred several million years ago while speciation was significantly more recent. Circos-style displays of the syntenic relationships identified by SyMAP are shown in Figures 4.16-4.22. Shared synteny implies that the conserved regions have originated from an identical ancestor, and possibly have retained similar functions in their genes.

Duplicated genes within a genome have a faster rate of loss of duplicated genes, hence less conservation, than orthologous regions. As mentioned in the previous paragraph, observing better internal synteny was unexpected. A better quality and a clear internal synteny than the synteny between species could be missing the whole picture on all the chromosomes. In addition, some of these internal synteny regions are small with fewer genes than the synteny regions between species. Furthermore, this could be better explained by a possible tandem duplication that happened before the speciation.

InterProScan (Jones, P. et al. 2014) and BLAST2GO (Conesa A. et al. 2005) were used to perform GO and functional annotation. Selected results are shown in Figures 4.23-4.25. For all of our results the top blast hit is *Glycine max*, as expected, and *Glycine soja* was the second best hit. The top two GO term categories are cellular and metabolic processes. Biological regulation is the third largest GO term category. The results also indicate that these targeted genes are involved in numerous biological processes

including metabolic processes, cell regulation processes, salt stress response, and response to stimulus. Perhaps, these rules can contribute to our understanding of underlying differences of *Glycines* and *Glycine max.*

### 4.3.3 MEGA

MEGA was used to align all the *Glycine* gene predictions together, and to build a phylogenetic tree. These files are provided in the supplementary data. Maximum likelihood estimate of Transition/Transversion bias was calculated and the estimated Transition/Transversion bias (R) is 0.38. Substitution pattern and rates were estimated under the Kimura (1980) 2-parameter model. For estimating ML values, a tree topology was automatically computed. The maximum Log likelihood for this computation was -3293832.002. The analysis involved 338 nucleotide sequences. Codon positions that were included are 1st+2nd+3rd+Noncoding. All positions with less than 25% site coverage were eliminated. That is, fewer than 75% alignment gaps, missing data, and ambiguous bases were allowed at any position. There were a total of 14,799 positions in the final dataset.

To calculate maximum likelihood, substitution matrix, substitution pattern and rates were estimated under the Tamura-Nei (1993) model. The result is shown in Table 4.20. Rates of different transitional substitutions are shown in bold and those of transversional substitutions are shown in italics. Relative values of instantaneous r should be considered when evaluating them. For simplicity, sum of r values is made equal to 100, the nucleotide frequencies are A = 33.21%, T/U = 33.25%, C = 16.79%, and G = 16.75%. For estimating ML values, a tree topology was automatically computed. The maximum

Log likelihood for this computation was -252146.163. The analysis involved 338 nucleotide sequences. Codon positions that were included are 1st+2nd+3rd+Noncoding. All positions containing gaps and missing data were eliminated. There were a total of 567 positions in the final dataset.

Maximum composite likelihood estimate of the pattern of nucleotide substitution was calculated as well. The results are shown in Table 4.18. Rates of different transitional substitutions are shown in bold and those of transversional substitutions are shown in italics. For simplicity, the sum of r values is made equal to 100. The nucleotide frequencies are 33.21% (A), 33.25% (T/U), 16.79% (C), and 16.75% (G). The transition/transversion rate ratios are k1 = 13.987 (purines) and k2 = 7.331 (pyrimidines). The overall transition/transversion bias is R = 4.749, where R = [A*G*k1 + T*C*k2]/[(A+G)*(T+C)]. The analysis involved 338 nucleotide sequences. Codon positions included were 1st+2nd+3rd+Noncoding. All positions containing gaps and missing data were eliminated. There were a total of 567 positions in the final dataset.

Pairwise distances were calculated and the excel sheet is provided as supplementary data. The overall mean distance was 8.32727. The codon-based test of neutrality of analysis between sequences (Z-test) was performed as well. This test is a test of positive selection. The output file is provided in supplementary data. The probability of rejecting the null hypothesis of strict neutrality (dN=dS) is shown below diagonal. Values of P less than 0.05 is considered significant at the 5% level and rejects the null hypothesis. The test statistic (dN - dS) is shown above the diagonal. dS and dN are the numbers of synonymous and nonsynonymous substitutions per site, respectively. The variance of the difference was computed using the analytical method. Analyses were conducted using the

Nei-Gojobori method. The analysis involved 338 nucleotide sequences. All ambiguous positions were removed for each sequence pair. There were a total of 97028 positions in the final dataset. The presence of n/c in the results denotes cases in which it was not possible to estimate evolutionary distances. This is important in understanding the dynamics of molecular sequence evolution. While nonsynonymous mutations may be under strong selective pressure, silence (synonymous) mutations are mostly invisible to natural selection. When the rate of nonsynonymous codon changes (dN) exceeds the rate of synonymous codon changes (dS), positive selection can be inferred.

Results from Tajima's Neutrality Test (Tajima F. 1989) is shown in Table 4.19. This test distinguishes between sequencing evolving randomly and neutrally, and the ones that are under pressure. The analysis involved 338 nucleotide sequences. Codon positions included were 1st+2nd+3rd+Noncoding. All ambiguous positions were removed for each sequence pair. There were a total of 449593 positions in the final dataset. The negative Tajima's D implies that either there has been a recent population expansion after a bottleneck and a selective sweep has happened, or low frequency of rare alleles was present (purifying selection). The equality of evolutionary rate between sequences A (SeedComposition_Gm07_SOJ_1_-_ordered_1_end_21458_bp) and B (SeedComposition_Gm07_SOJ_2_-_unordered_131440_bp), with sequence C (SeedComposition_Gm07_SOJ_3_-_unordered_85451_bp) as an outgroup in Tajima's relative rate test (Tajima 1993). The test statistic was 0.52 (P = 0.46936 with 1 degree[s] of freedom). P-value less than 0.05 is often used to reject the null hypothesis of equal rates between lineages, hence the molecular clock is not rejected and is independent of

the substitution variation. The Codon positions included were 1st+2nd+3rd+Noncoding. All positions containing gaps and missing data were eliminated.

## 4.4 Conclusion

We can use comparative genomics to transform our knowledge between close relatives of cultivated legumes, and take advantage of their resourceful and various germplasm and phenotypes. Studying the origin of different features and the evolution of them among different species of legumes will arm us with the necessary knowledge to pool interesting characteristics from them to improve our crops. In this study, we focused on regions anchored by the gene isoflavone synthase and the orthologous sequences of these from perennial *Glycines*. Based on the synteny results, we concluded that our annotations were accurate. However, we were not able to find a lot of novel genes in the perennials that are not already identified in *Glycine max;* first due to our validation approach, and second because of limited data that we had and the possibility of lost sequences. Although it is hard to conclude time of divergence based on two chromosomes, clear patterns of duplication was observed within these regions. These perennial *Glycines* show high synteny, which was expected based on their short divergent time and how closely related they are. The GO term analysis, combined with further collaborative studies on other regions on these genomes, will help us understand how these species have adapted to different climates and situations.

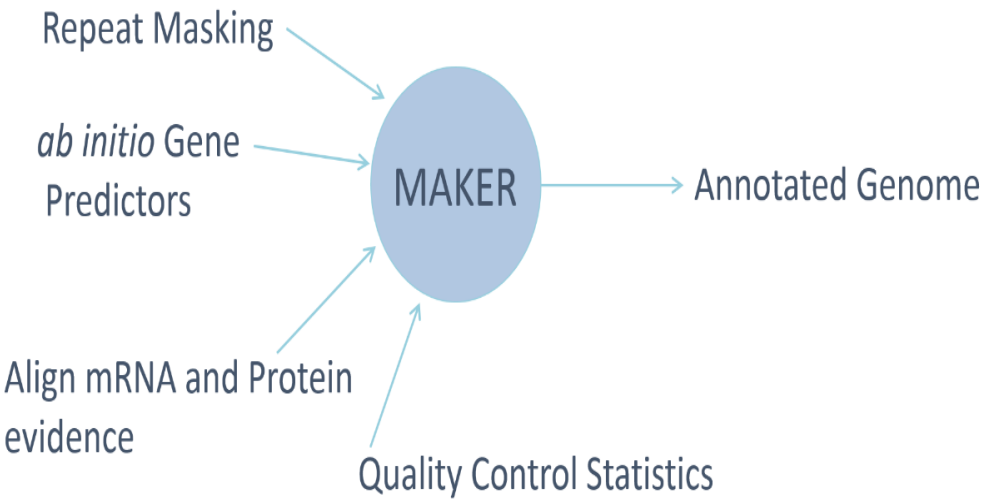



Figure 4. 4: MAKER workflow

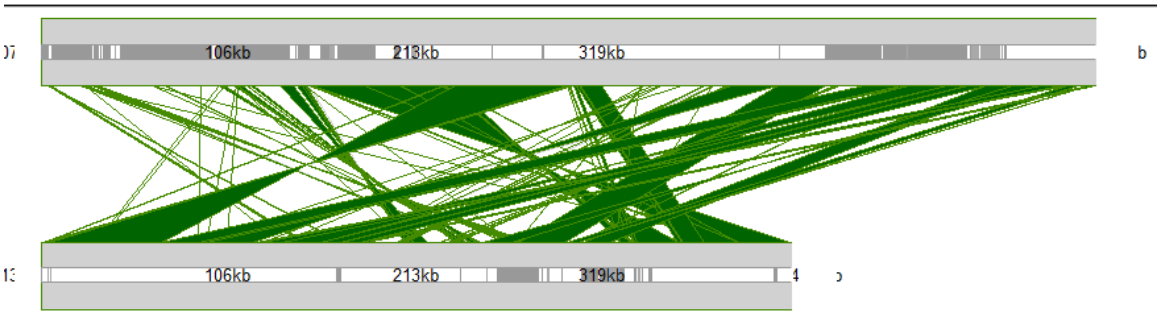



Figure 4. 2: Synteny regions between GCAN07 and GCAN13

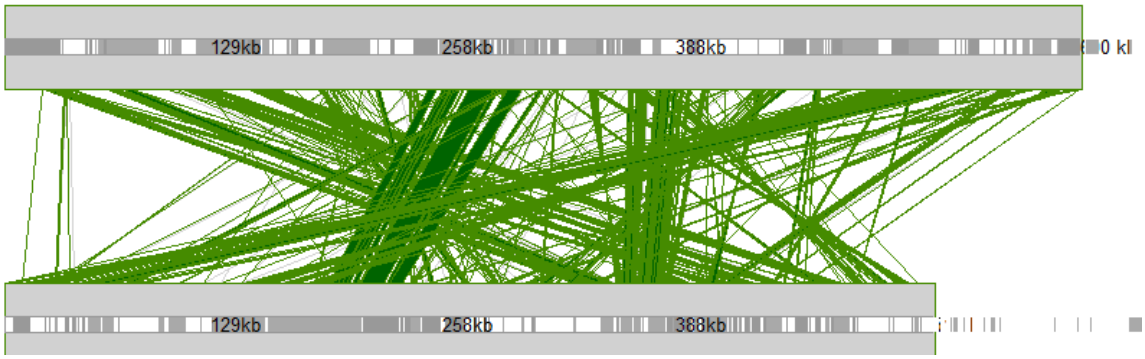



Figure 4. 3: Synteny regions between GCAN07 and GCYR07

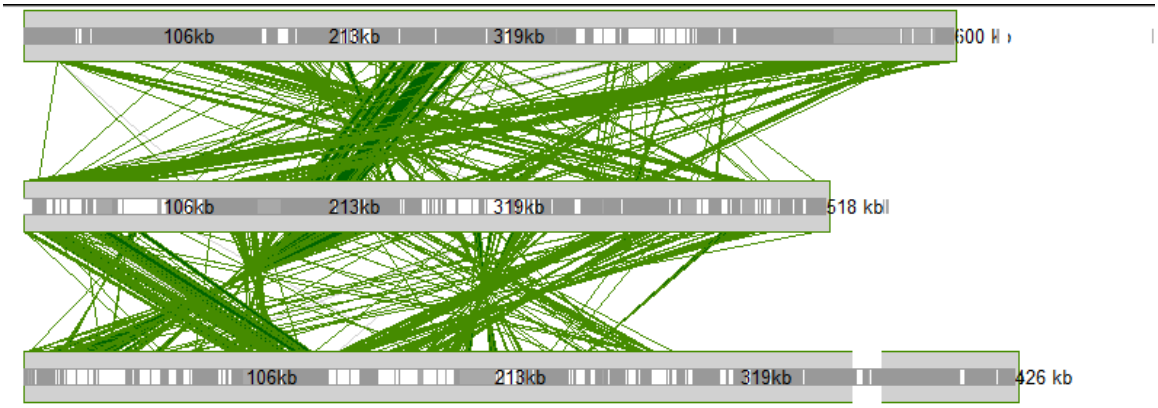



Figure 4. 4: Synteny regions between GCAN07, GCYR07, and GCAN13

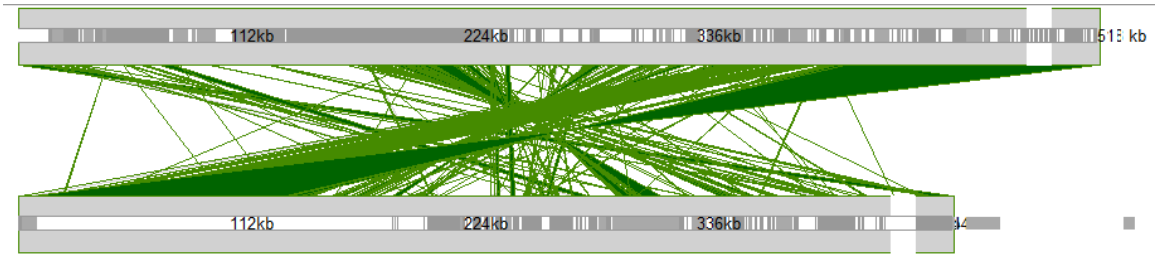



Figure 4. 5: Synteny regions between GCYR07 and GCYR13

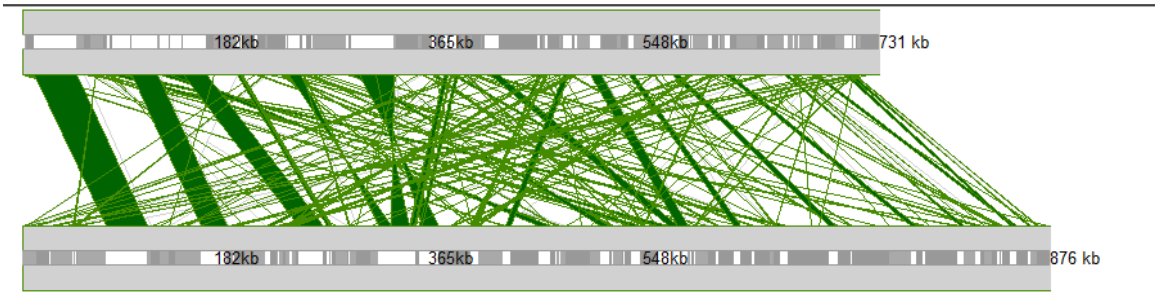



Figure 4. 6: Synteny regions between GDOL07 and GDOL13

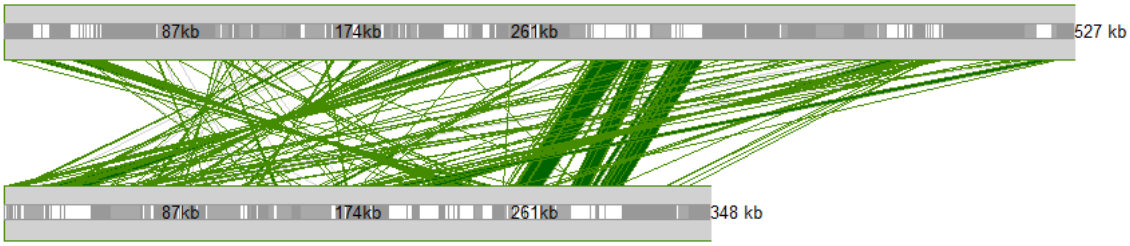



Figure 4. 7: Synteny regions between GFAL07 and GFAL13

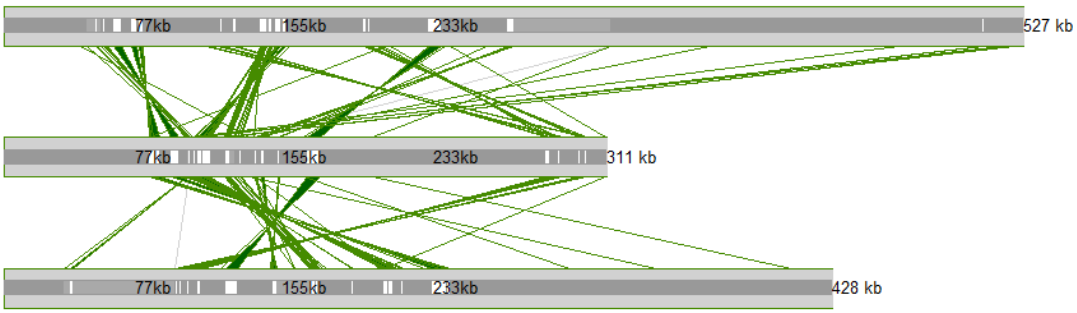



Figure 4. 8: Synteny regions between GFAL07, GSTEN07, and GTOM07

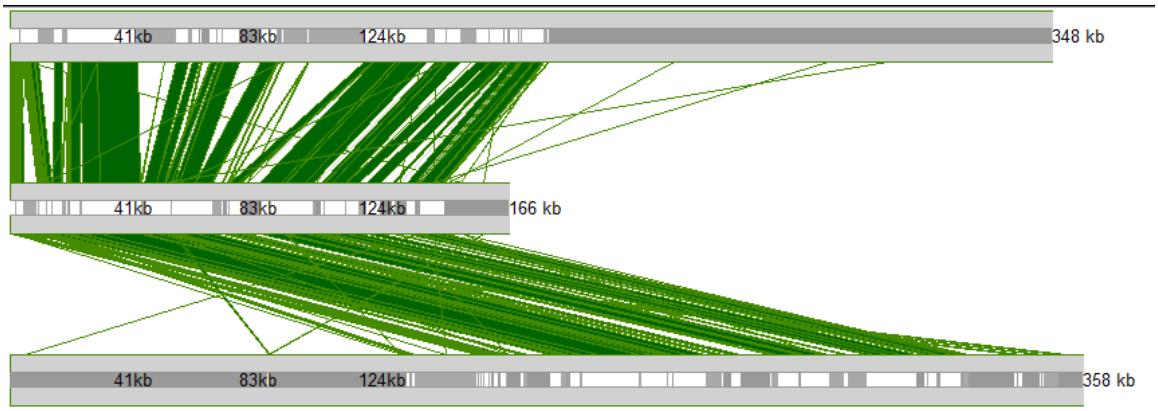



Figure 4. 9: Synteny regions between GFAL13, GSTEN13, and GTOM13

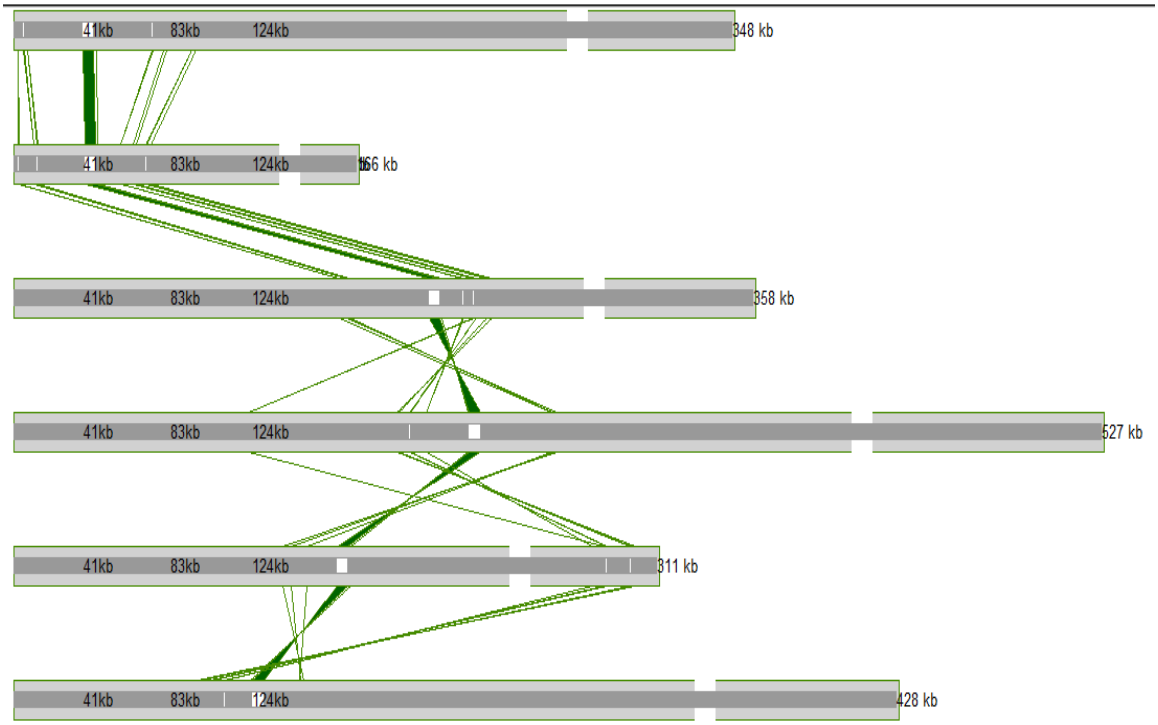



Figure 4. 10: Synteny regions between GFAL13, GSTEN13, GTOM13, GFAL07, GSTEN07, and GTOM07


Figure 4. 11: Synteny regions between GSTEN07 and GSTEN13

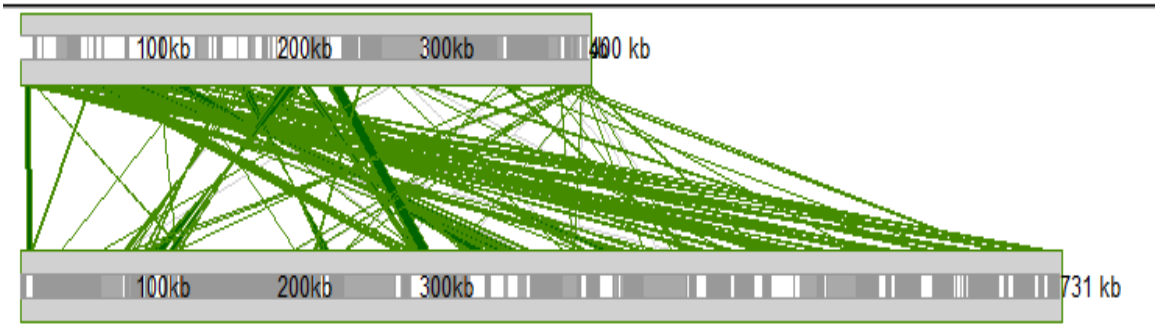



Figure 4. 12: Synteny regions between GSYN07 and GDOL07

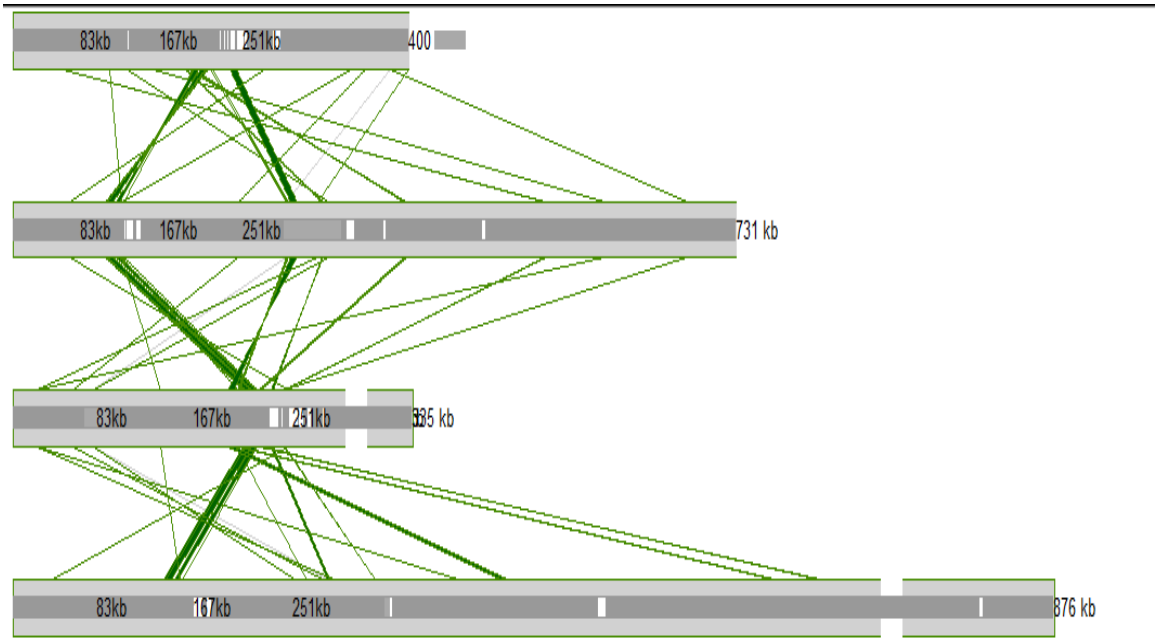


Figure 4. 13: Synteny regions between GSYN07, GDOL07, GSYN13, and GDOL13

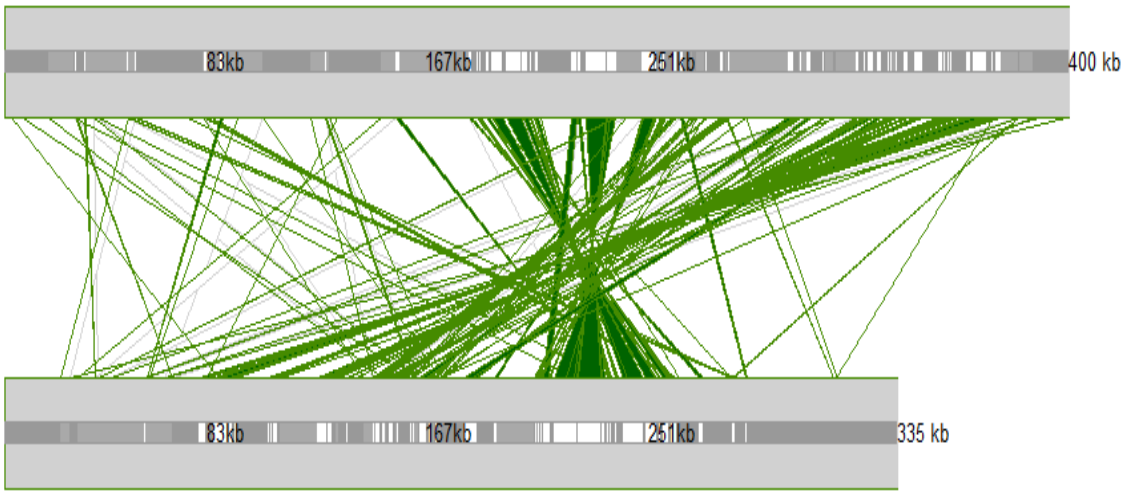



Figure 4. 14: Synteny regions between GSYN07 and GSYN13

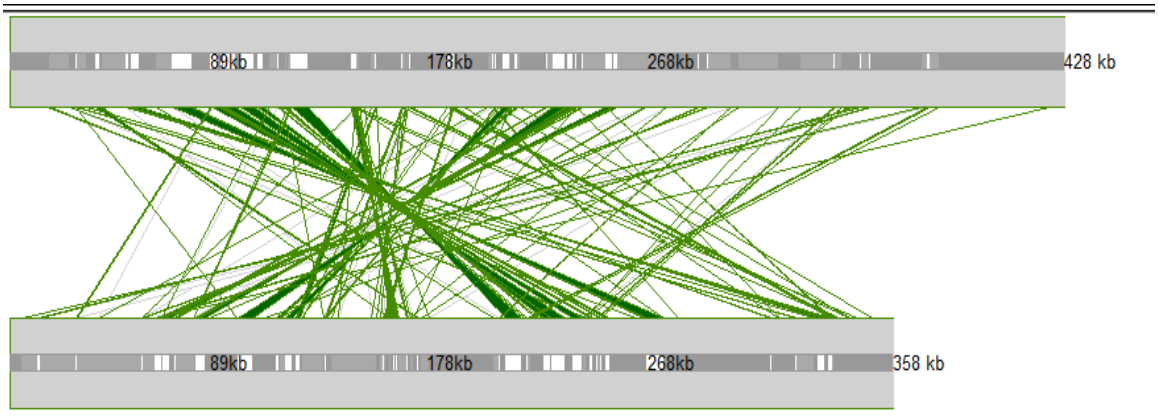



Figure 4. 15: Synteny regions between GTOM07 and GTOM13

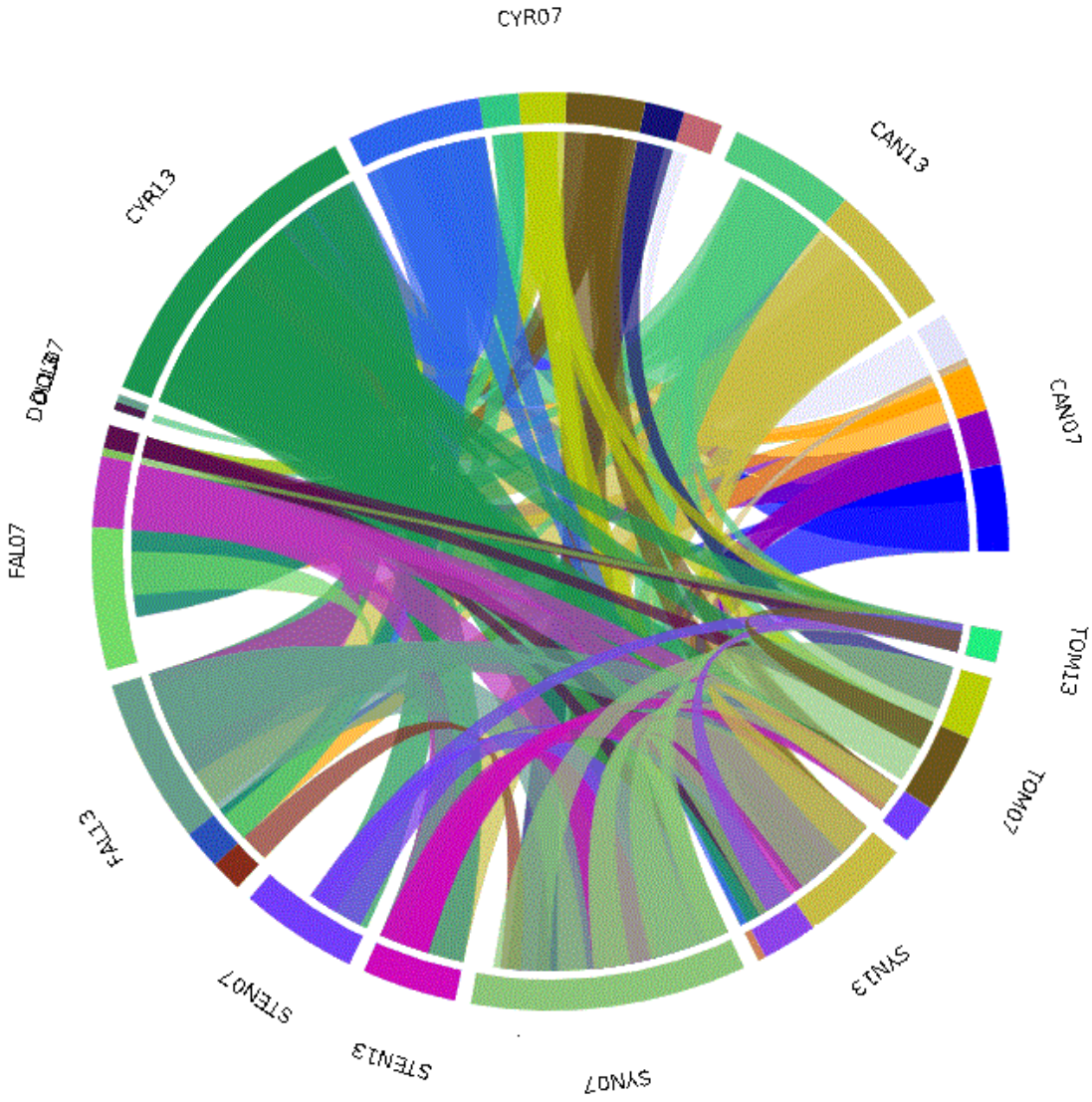



Figure 4. 16: Circular view of synteny blocks between all of the species

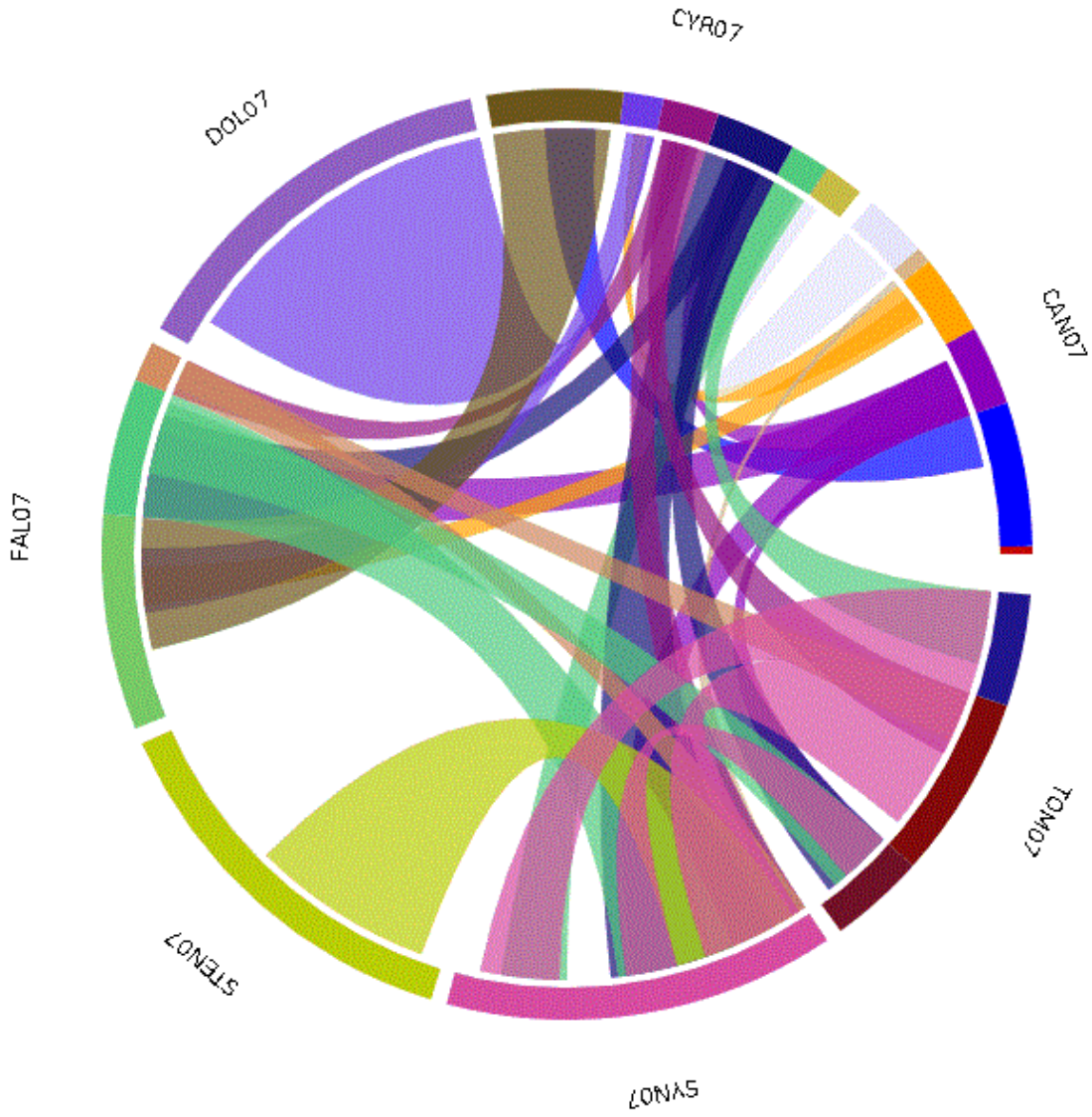



Figure 4. 17: Circular view of synteny blocks between all the regions that correspond to soybean's chromosome 07

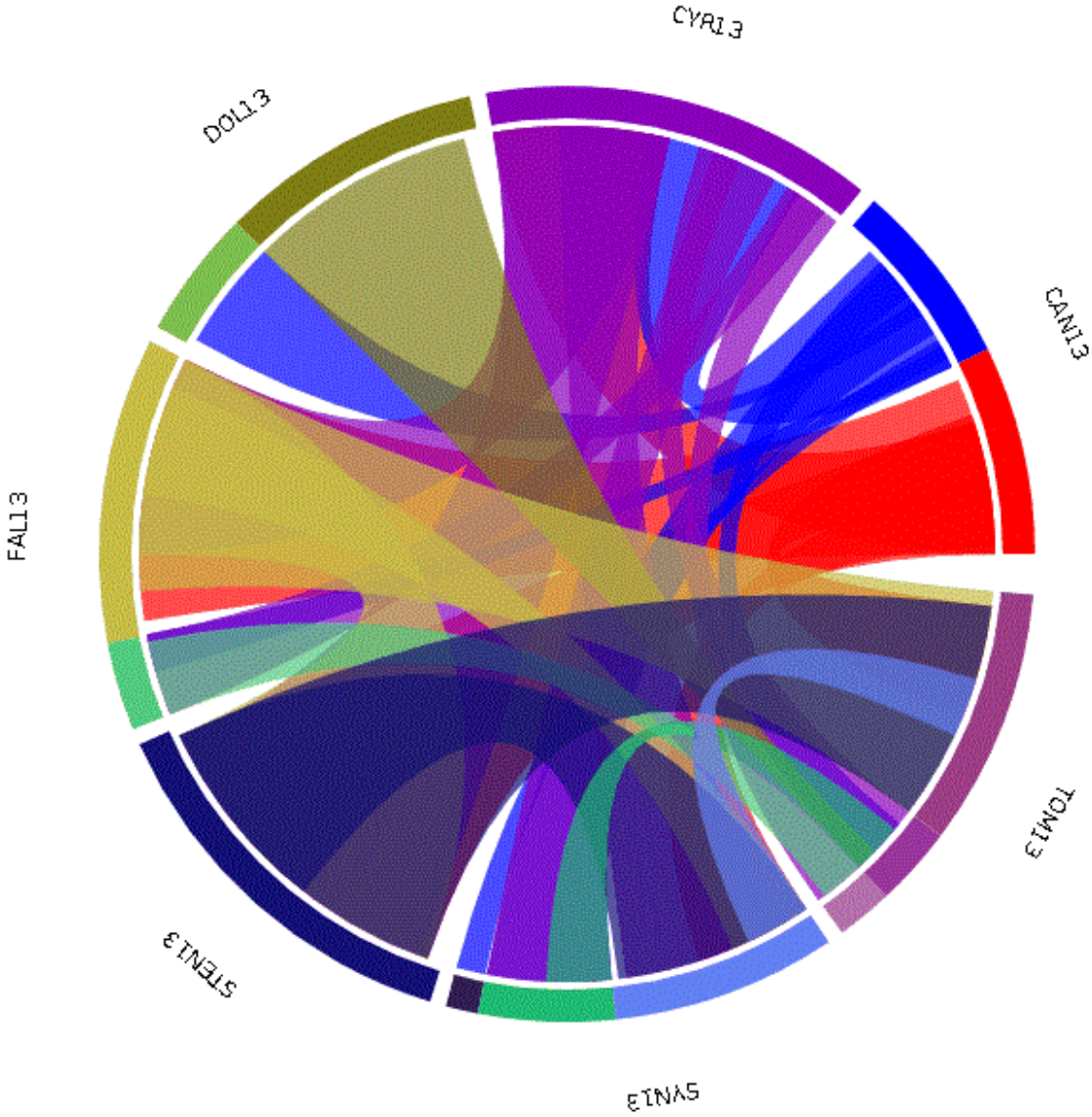



Figure 4. 18: Circular view of synteny blocks between all the regions that correspond to soybean's chromosome 13

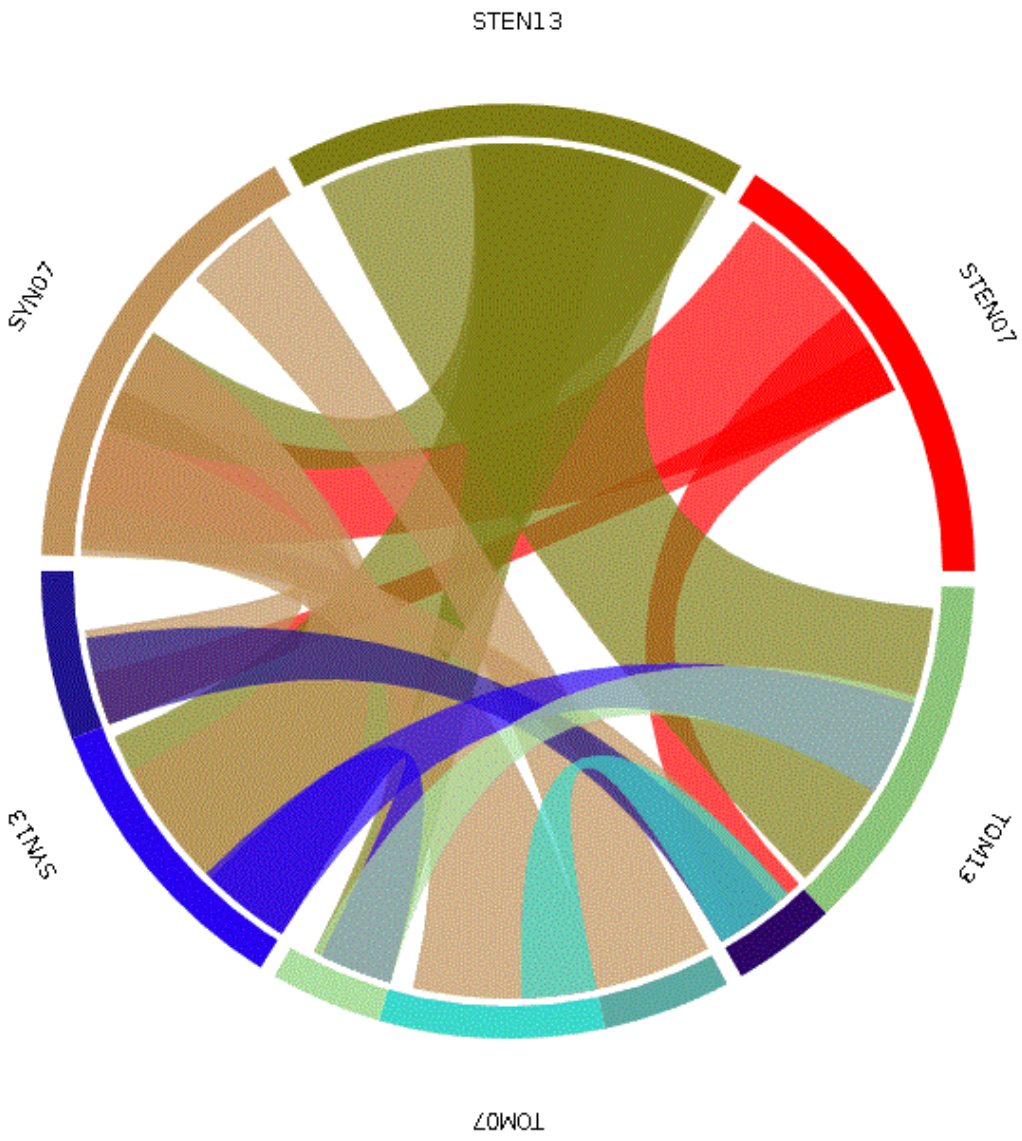



Figure 4. 19: Circular view of synteny blocks between GSYN, GTOM, and GSTEN

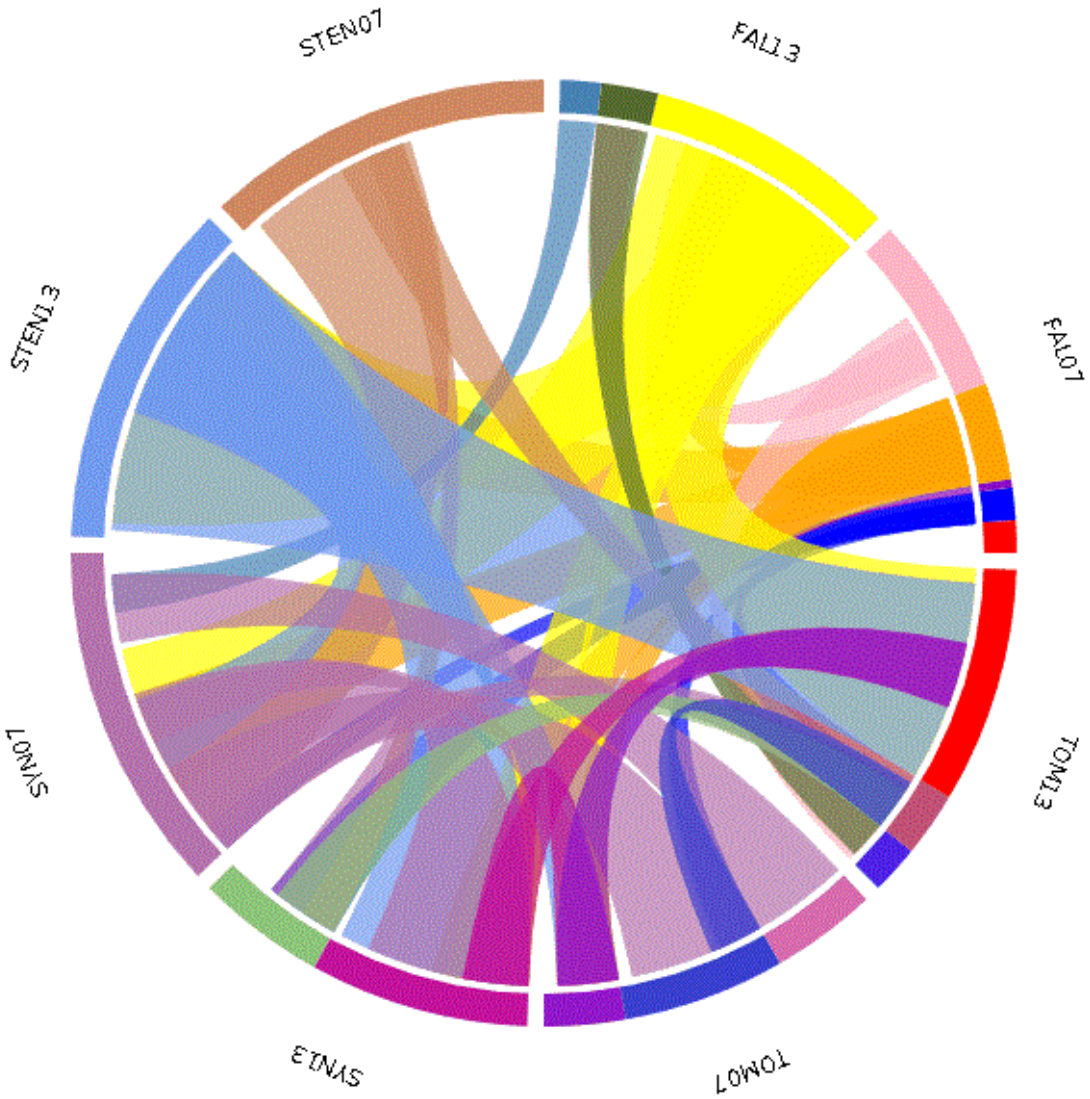



Figure 4. 20: Circular view of synteny blocks between GSYN, GTOM, GFAL, and GSTEN

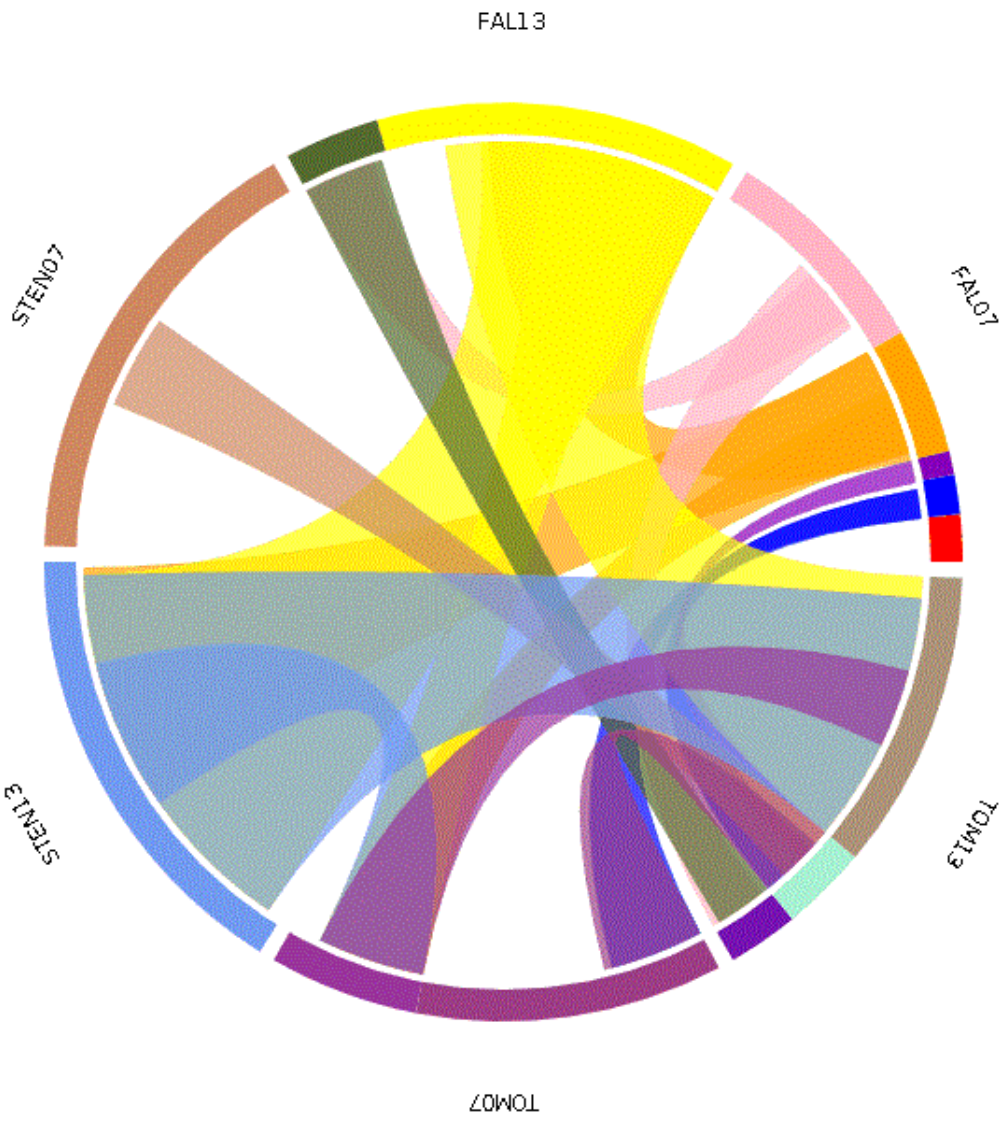



Figure 4. 21: Circular view of synteny blocks between GTOM, GFAL, and GSTEN

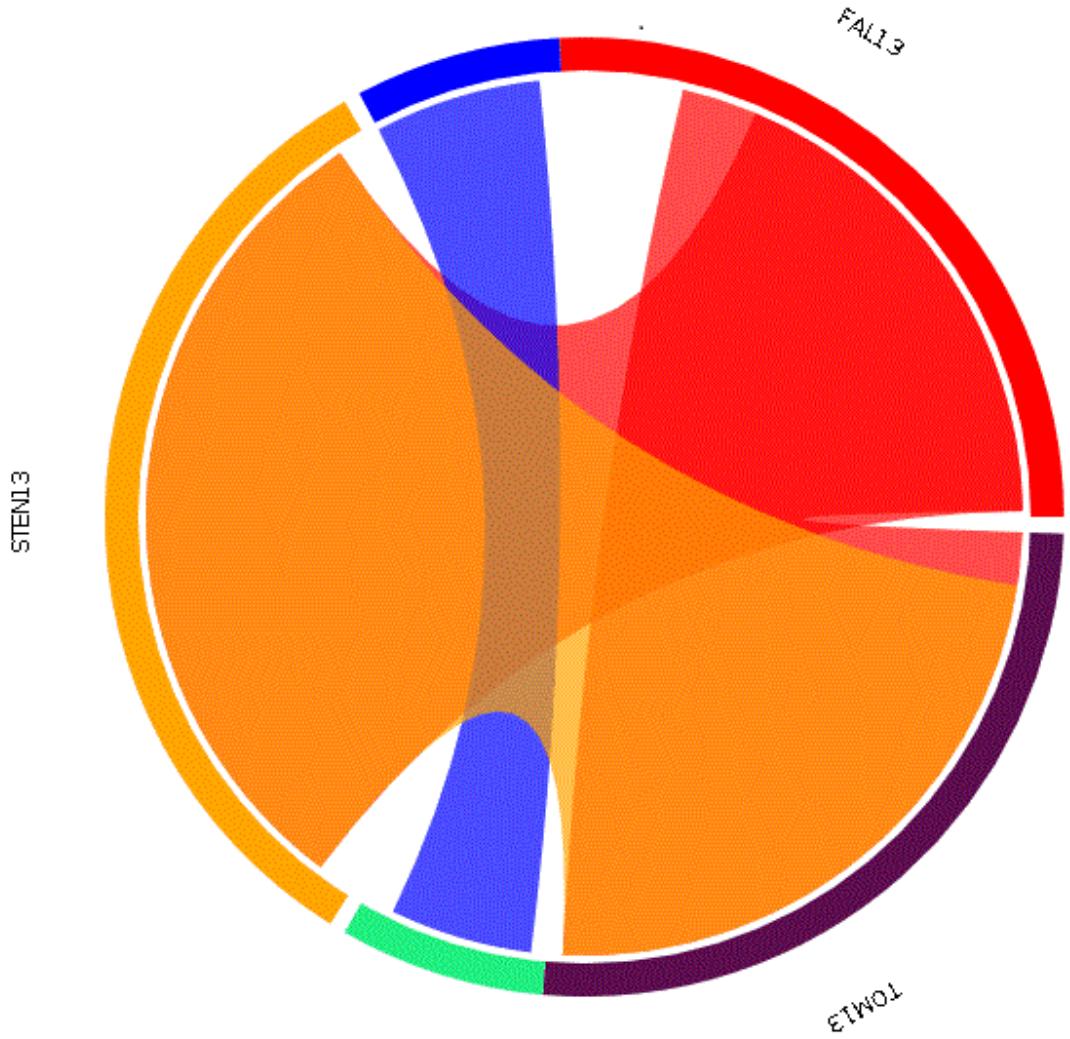
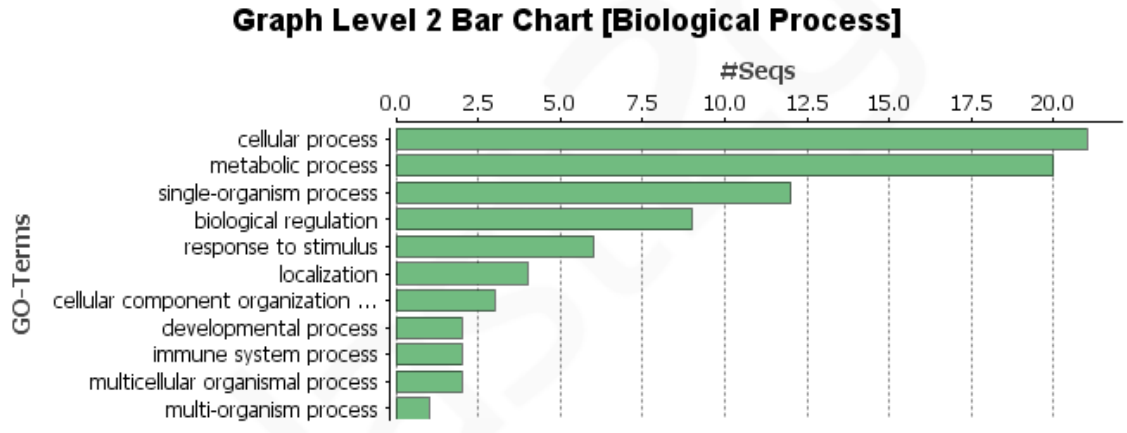
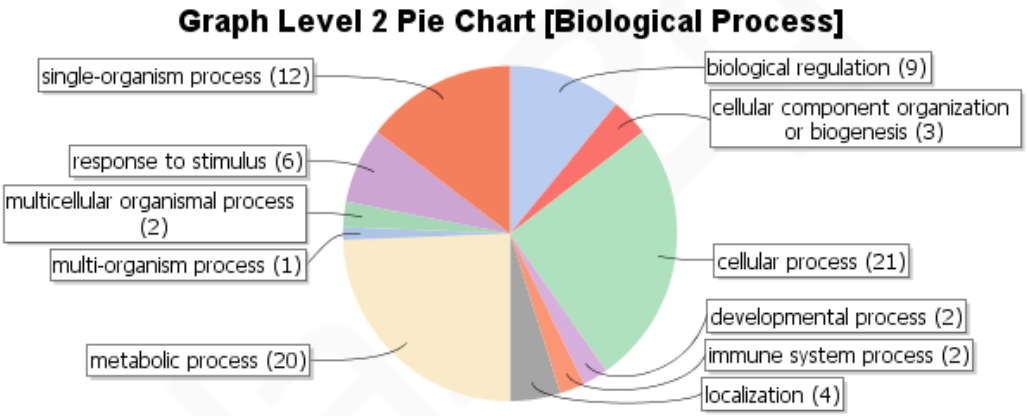
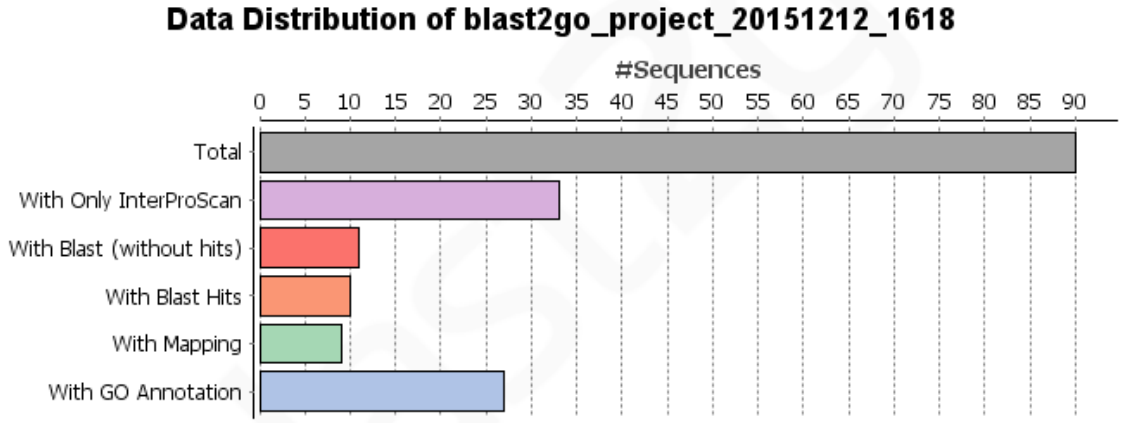
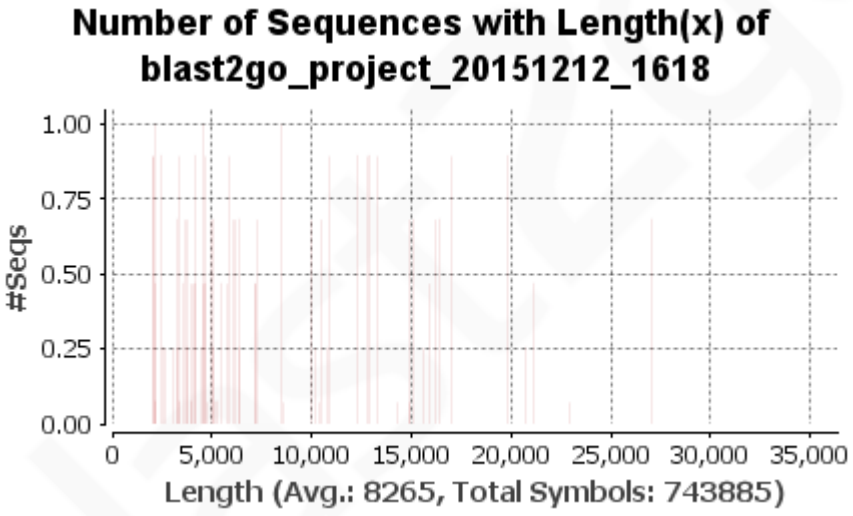



Figure 4. 22: Circular view of synteny blocks between GFAL13, GTOM13, and GSTEN13




Graph Level 2 Pie Chart [Biological Process]
single-organism process (12)
biological regulation (9)
cellular component organization or biogenesis (3)
response to stimulus (6)
multicellular organismal process (2)
multi-organism process (1)
cellular process (21)
developmental process (2)
immune system process (2)
metabolic process (20)
localization (4)


Data Distribution of blast2go_project_20151212_1618
#Sequences
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90
Total
With Only InterProScan
With Blast (without hits)
With Blast Hits
With Mapping
With GO Annotation


Number of Sequences with Length(x) of blast2go_project_20151212_1618
#Seqs
1.00
0.75
0.50
0.25
0.00
0 5,000 10,000 15,000 20,000 25,000 30,000 35,000
Length (Avg.: 8265, Total Symbols: 743885)

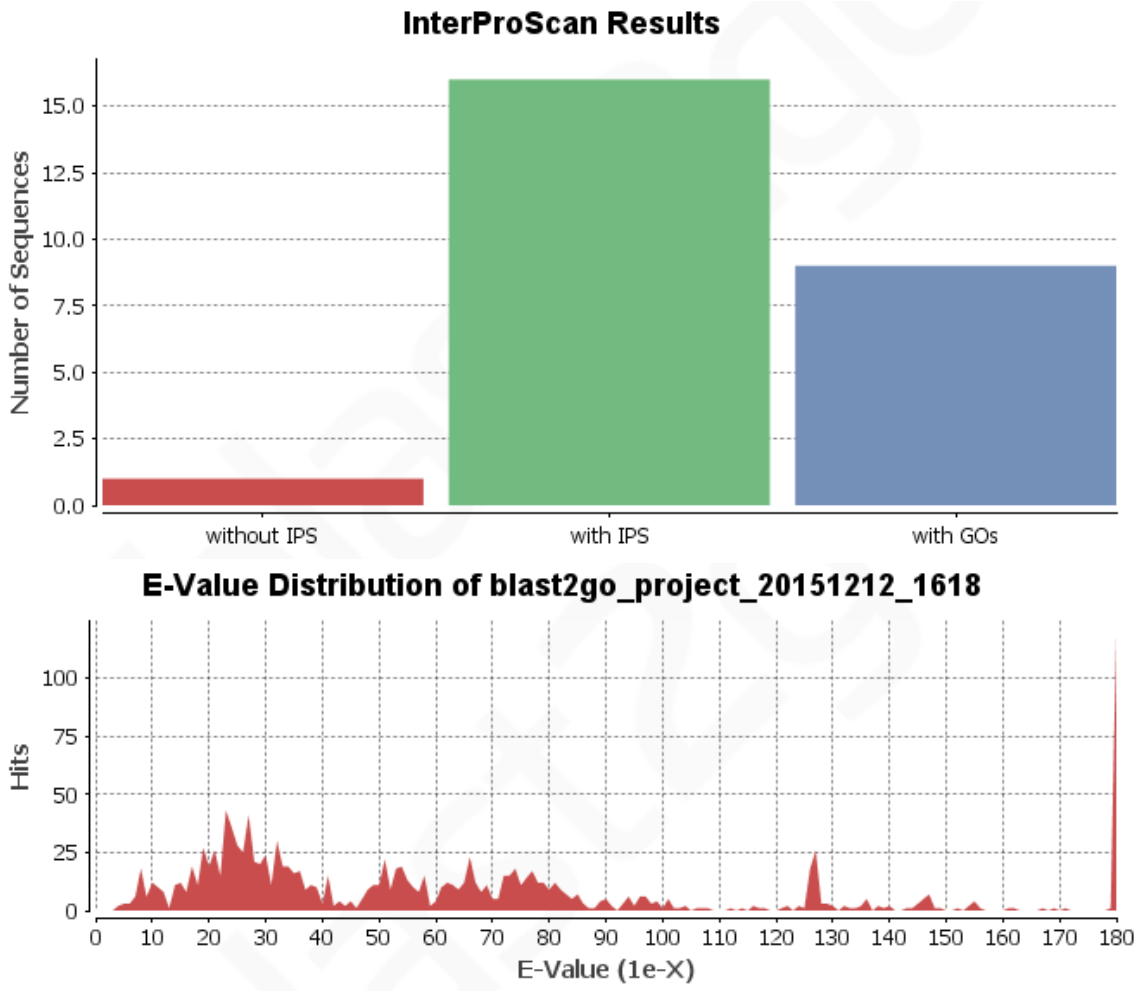
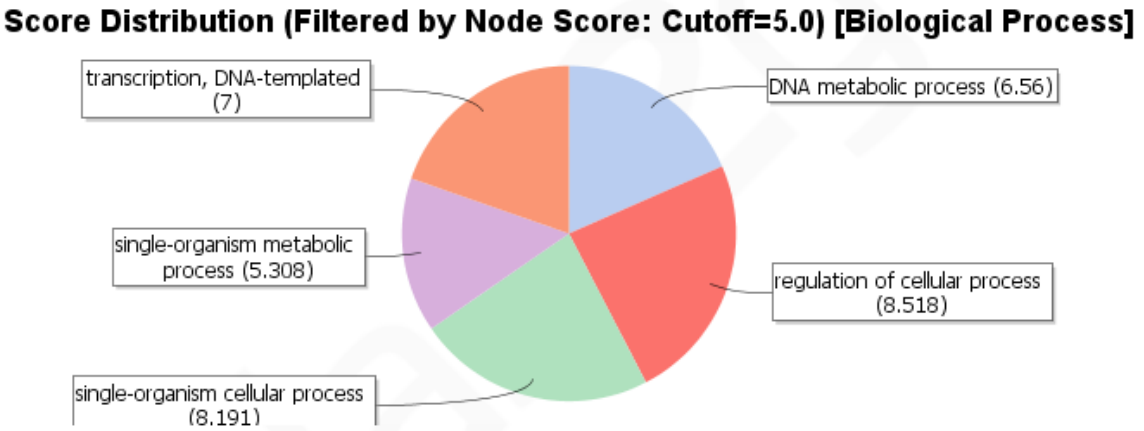
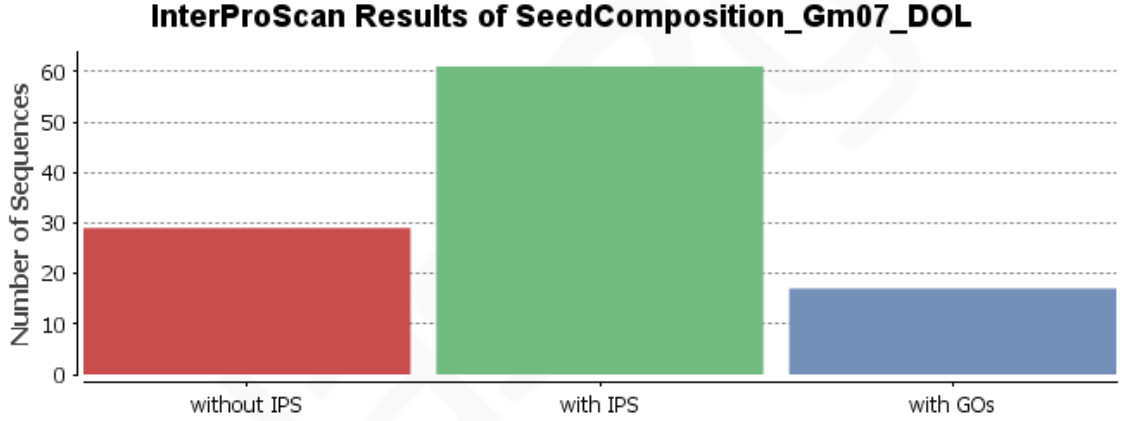
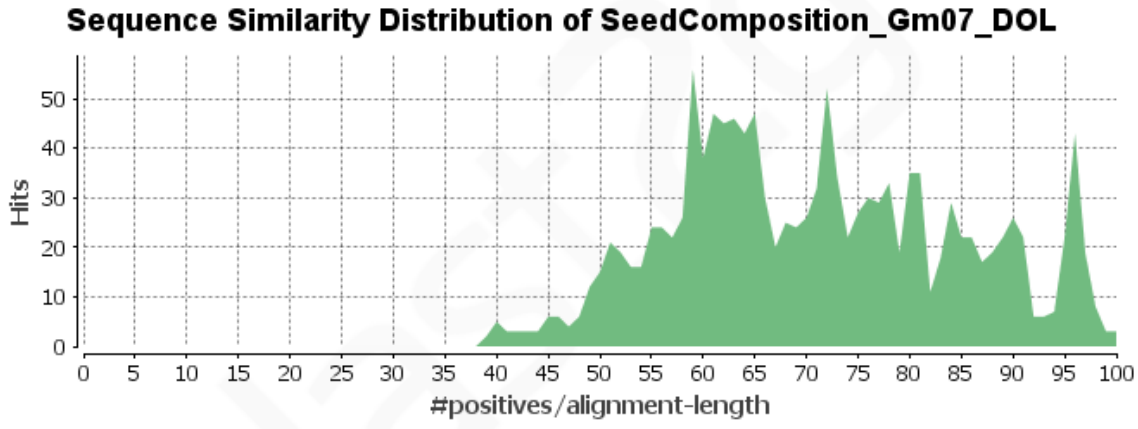



Figure 4. 23: GCAN13 InterProScan and BLAST2GO overview


Score Distribution (Filtered by Node Score: Cutoff=5.0) [Biological Process]
transcription, DNA-templated (7)
DNA metabolic process (6.56)
single-organism metabolic process (5.308)
regulation of cellular process (8.518)
single-organism cellular process (8.191)


InterProScan Results of SeedComposition_Gm07_DOL
Number of Sequences
0
10
20
30
40
50
60
without IPS
with IPS
with GOs


Sequence Similarity Distribution of SeedComposition_Gm07_DOL
Hits
0
10
20
30
40
50
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100
#positives/alignment-length

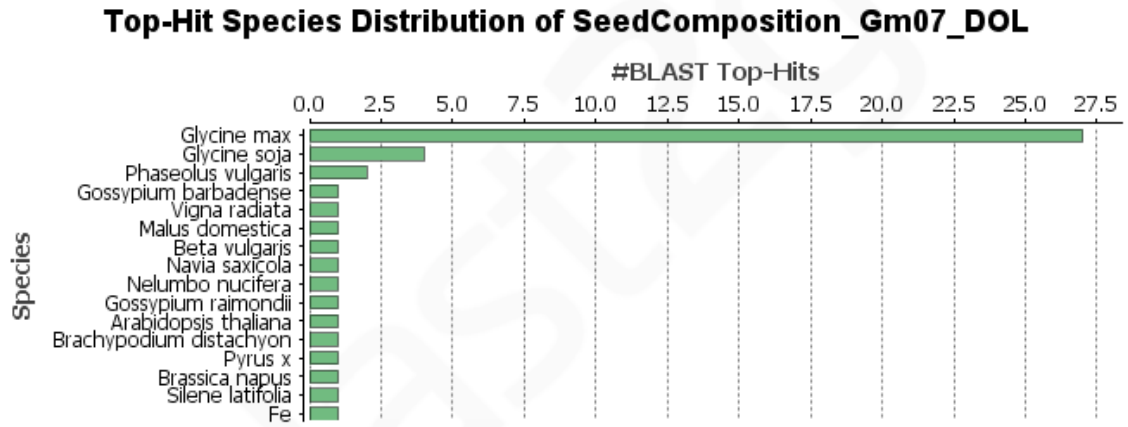
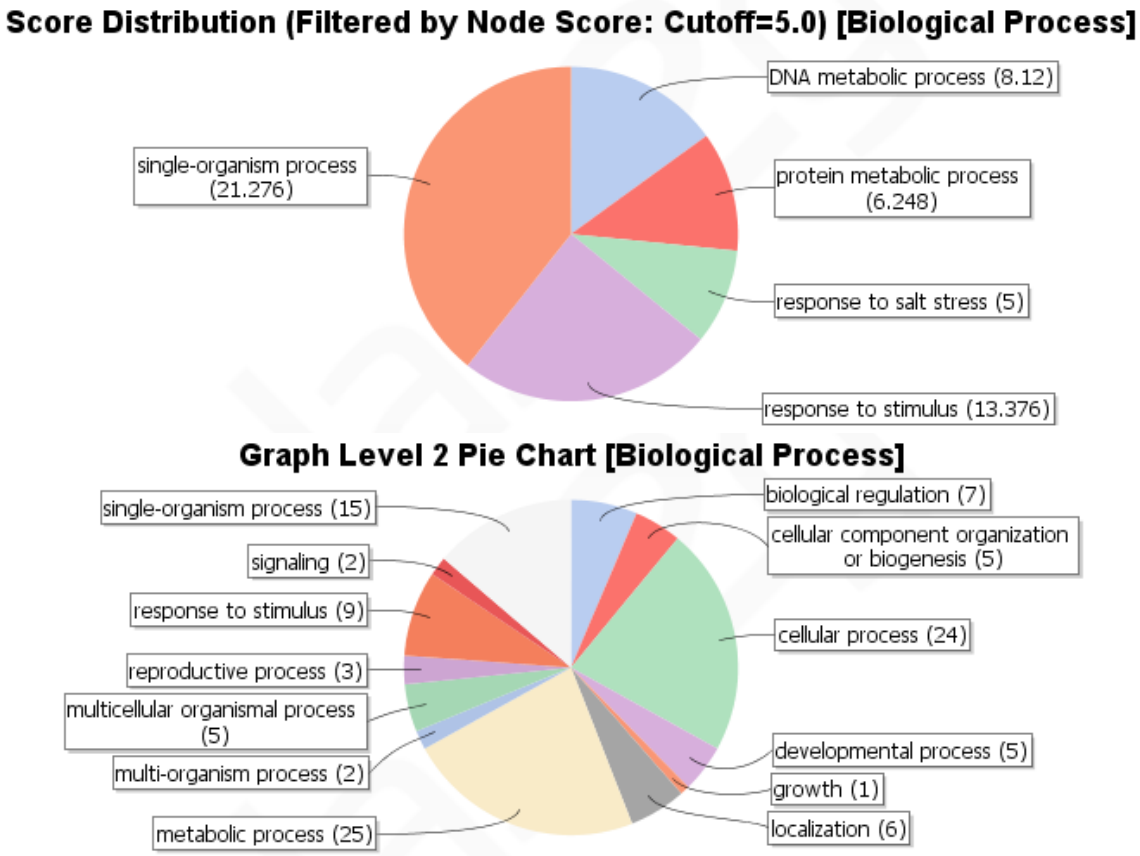



Figure 4. 4.24: GDOL07 InterProScan and BLAST2GO overview



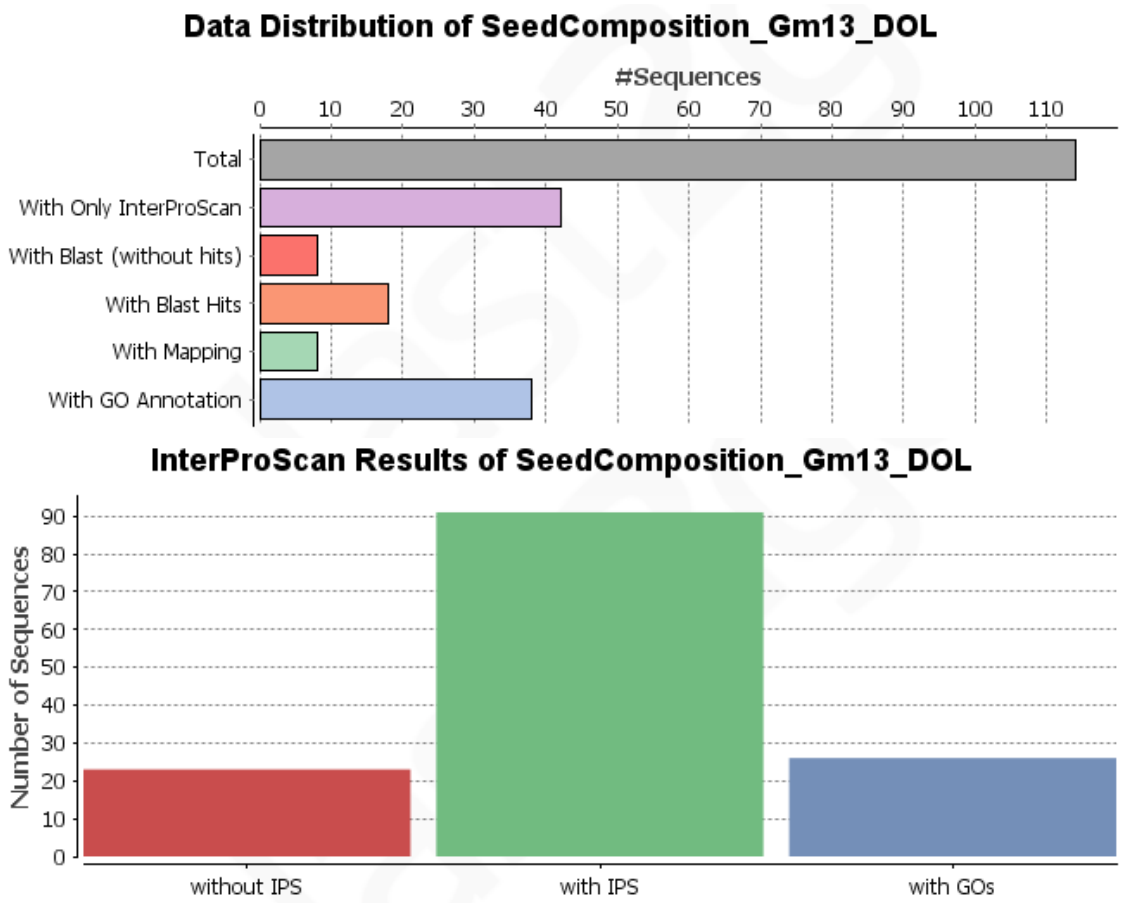



Figure 4.25: GDOL13 InterProScan and BLAST2GO overview

Table 4.1: Summary of the RNA-seq data

| Species, accessions and tissue sampled | number of reads | bases/read | total bases | number of contigs>400bp |
|---|---|---|---|---|
| Glycine canescens (G1232): Leaf_pod_root | 72891886 | 104 | 7580756144 | 22225 |
| Glycine cyrtoloba (G1267): Leaf_root | 46156858 | 104 | 4800313232 | 26631 |
| Glycine falcata(G1155): Leaf_root | 32690160 | 104 | 3399776640 | 21898 |
| Glycine dolichocarpa(G1134): Leaf_root_pod | 71064452 | 104 | 7390703008 | 18971 |
| Glycine tomentella(G1403): Leaf_pod_root | 62088568 | 104 | 6457211072 | 24658 |
| Glycine stenophita(G1974): Root_leaf | 36752214 | 104 | 3822230256 | 21295 |
| Glycine syndetika(G1300): Leaf_root_pod | 63979402 | 104 | 6653857808 | 27057 |
| **TOTAL** | 385623540 | | 40104848160 | 162735 |

Table 4.2: Annotation Summary of GCAN07

| Feature | Value |
|---|---|
| annotated genes | 86 (plus=45 minus=41) |
| protein coding genes | 48 (plus=20 minus=28) |
| avg GC fraction | 0.344288 (min=0.319062 max=0.369757) |
| single-exon | 15 (0.174419) |
| multi-exon | 71 (0.825581) |
| mean exon | 213.172668 (min=2 max=2641) |
| mean exon for non-overlapping genes | 302.943390 (min=6 max=2641) |
| mean intron | 507.081573 (min=9 max=5877) |
| mean intron for non-overlapping genes | 394.585571 (min=37 max=5877) |

Table 4.3: Annotation Summary of GCAN13

| Feature | Value |
|---|---|
| annotated genes | 56 (plus=27 minus=29) |
| protein coding genes | 30 (plus=16 minus=14) |
| avg GC fraction | 0.335514 (min=0.330695 max=0.340743) |
| single-exon | 10 (0.178571) |
| multi-exon | 46 (0.821429) |
| mean exon | 257.364716 (min=6 max=2641) |
| mean exon for non-overlapping genes | 323.845367 (min=6 max=2641) |
| mean intron | 593.326660 (min=15 max=5561) |
| mean intron for non-overlapping genes | 635.701477 (min=38 max=5561) |

Table 4.4: Annotation Summary of GCYR07

| Feature | Value |
|---|---|
| annotated genes | 60 (plus=29 minus=31) |
| protein coding genes | 41 genes (plus=20 minus=21) |
| avg GC fraction | 0.325943 (min=0.292994 max=0.351592) |
| single-exon | 8 (0.133333) |
| multi-exon | 52 (0.866667) |
| mean exon | 227.639572 (min=6 max=2638) |
| mean exon for non-overlapping genes | 260.505676 (min=6 max=2638) |
| mean intron | 418.434967 (min=13 max=3656) |
| mean intron for non-overlapping genes | 441.814819 (min=30 max=3656) |

Table 4.5: Annotation Summary of GCYR13

| Feature | Value |
|---|---|
| annotated genes | 42 (plus=26 minus=16) |
| protein coding genes | 34 (plus=20 minus=14) |

| | |
|---|---|
| avg GC fraction | 0.336405 (min=0.336405 max=0.336405) |
| single-exon | 5 (0.119048) |
| multi-exon | 37 (0.880952) |
| mean exon | 240.975006 (min=1 max=1590) |
| mean exon for non-overlapping genes | 263.587311 (min=14 max=1590) |
| mean intron | 477.818176 (min=60 max=7197) |
| mean intron for non-overlapping genes | 444.232269 (min=60 max=7197) |

Table 4.6: Annotation Summary of GFAL07

| Feature | Value |
|---|---|
| annotated genes | 62 (plus=37 minus=25) |
| protein coding genes | 45 (plus=26 minus=19) |
| avg GC fraction | 0.327172 (min=0.306552 max=0.352650) |
| single-exon | 16 (0.258065) |
| multi-exon | 46 (0.741935) |
| mean exon | 250.166672 (min=8 max=2626) |
| mean exon for non-overlapping genes | 273.134064 (min=9 max=2626) |
| mean intron | 469.471161 (min=6 max=9084) |
| mean intron for non-overlapping genes | 551.096985 (min=32 max=9084) |

Table 4.7: Annotation Summary of GFAL13

| Feature | Value |
|---|---|
| annotated genes | 41 (plus=28 minus=13) |
| protein coding genes | 25 (plus=16 minus=9) |
| avg GC fraction | 0.338124 (min=0.322387 max=0.367211) |
| single-exon | 7 (0.170732) |
| multi-exon | 34 (0.829268) |
| mean exon | 234.726059 (min=6 max=2641) |

| | |
|---|---|
| mean exon for non-overlapping genes | 259.741272 (min=6 max=2641) |
| mean intron | 623.552246 (min=29 max=6576) |
| mean intron for non-overlapping genes | 319.533905 (min=65 max=2031) |

Table 4.8: Annotation Summary of GDOL07

| Feature | Value |
|---|---|
| annotated genes | 119 (plus=61 minus=58) |
| protein coding genes | 79 (plus=38 minus=41) |
| avg GC fraction | 0.354772 (min=0.291783 max=0.443197) |
| single-exon | 25 (0.210084) |
| multi-exon | 94 (0.789916) |
| mean exon | 226.782608 (min=6 max=2632) |
| mean exon for non-overlapping genes | 282.116730 (min=10 max=2632) |
| mean intron | 371.942322 (min=8 max=6877) |
| mean intron for non-overlapping genes | 446.196625 (min=33 max=6877) |

Table 4.9: Annotation Summary of GDOL13

| Feature | Value |
|---|---|
| Annotated genes | 153 (plus=77 minus=76) |
| protein coding genes | 91 (plus=48 minus=43) |
| avg GC fraction | 0.365501 (min=0.303270 max=0.465380) |
| single-exon | 32 (0.209150) |
| multi-exon | 121 (0.790850) |
| mean exon | 247.906662 (min=6 max=5343) |
| mean exon for non-overlapping genes | 319.262238 (min=15 max=5343) |
| mean intron | 428.130280 (min=5 max=5952) |
| mean intron for non-overlapping genes | 470.875000 (min=32 max=5952) |

Table 4.10: Annotation Summary of GSTEN07

| Feature | Value |
|---|---|
| annotated genes | 41 (plus=18 minus=23) |
| protein coding genes | 30 (plus=14 minus=16) |
| avg GC fraction | 0.349633 (min=0.318335 max=0.383700) |
| single-exon | 9 (0.219512) |
| multi-exon | 32 (0.780488) |
| mean exon | 282.952087 (min=11 max=2610) |
| mean exon for non-overlapping genes | 286.620972 (min=15 max=2610) |
| mean intron | 394.880951 (min=13 max=4918) |
| mean intron for non-overlapping genes | 443.861694 (min=47 max=4918) |

Table 4.11: Annotation Summary of GSTEN13

| Feature | Value |
|---|---|
| annotated genes | 19 (plus=14 minus=5) |
| protein coding genes | 11 (plus=8 minus=3) |
| avg GC fraction | 0.354298 (min=0.331250 max=0.386881) |
| single-exon | 7 (0.368421) |
| multi-exon | 12 (0.631579) |
| mean exon | 258.147888 (min=19 max=1617) |
| mean exon for non-overlapping genes | 267.937500 (min=19 max=1617) |
| mean intron | 977.593506 (min=74 max=9160) |
| mean intron for non-overlapping genes | 352.000000 (min=74 max=2190) |

Table 4.12: Annotation Summary of GSYN07

| Feature | Value |
|---|---|
| annotated genes | 37 (plus=15 minus=22) |

| | |
|---|---|
| protein coding genes | 26 (plus=11 minus=15) |
| avg GC fraction | 0.320163 (min=0.320163 max=0.320163) |
| single-exon | 9 (0.243243) |
| multi-exon | 28 (0.756757) |
| mean exon | 264.866211 (min=3 max=2632) |
| mean exon for non-overlapping genes | 317.076935 (min=14 max=2632) |
| mean intron | 418.409515 (min=5 max=5392) |
| mean intron for non-overlapping genes | 466.115387 (min=30 max=5392) |

Table 4.13: Annotation Summary of GSYN13

| Feature | Value |
|---|---|
| annotated genes | 43 (plus=25 minus=18) |
| protein coding genes | 25 (plus=17 minus=8) |
| avg GC fraction | 0.329525 (min=0.316548 max=0.372066) |
| single-exon | 7 (0.162791) |
| multi-exon | 36 (0.837209) |
| mean exon | 230.913971 (min=6 max=1599) |
| mean exon for non-overlapping genes | 248.272720 (min=6 max=1599) |
| mean intron | 553.415283 (min=19 max=8567) |
| mean intron for non-overlapping genes | 494.148651 (min=73 max=2903) |

Table 4.14: Annotation Summary of GTOM07

| Feature | Value |
|---|---|
| Annotated genes | 48 (plus=19 minus=29) |
| protein coding genes | 35 genes (plus=14 minus=21) |
| avg GC fraction | 0.329603 (min=0.306725 max=0.349695) |
| single-exon | 10 (0.208333) |
| multi-exon | 38 (0.791667) |

| | |
|---|---|
| mean exon | 271.776550 (min=6 max=2632) |
| mean exon for non-overlapping genes | 310.695648 (min=6 max=2632) |
| mean intron | 415.045807 (min=23 max=5433) |
| Mean intron for non-overlapping genes | 465.937500 (min=33 max=5433) |

Table 4.15: Annotation Summary of GTOM13

| Feature | Value |
|---|---|
| annotated genes | 38 (plus=22 minus=16) |
| Protein coding genes | 23 genes (plus=12 minus=11) |
| avg GC fraction | 0.328253 (min=0.318133 max=0.339369) |
| single-exon | 5 (0.131579) |
| multi-exon | 33 (0.868421) |
| mean exon | 239.734985 (min=7 max=1590) |
| Mean exon for non-overlapping genes | 310.695648 (min=6 max=2632) |
| mean intron | 714.714294 (min=5 max=6765) |
| Mean intron for non-overlapping genes | 465.937500 (min=33 max=5433) |

Table 4.16: Corresponding Orthologs of Annotated Genes in Soybean

| Annotation | Soybean orthologs | Observed in: | conserved |
|---|---|---|---|
| Isoflavone synthase | Glyma13g24200.1, Glyma07g32330.1 | GCAN. GCYR, GFAL, GDOL, GSTEN. GSYN, GTOM | Yes |
| Structural maintenance of chromosomes (SMC) family protein | Glyma13g18470.1 | GCAN. GCYR, GFAL, GDOL,GSYN, GTOM | Yes |
| autophagy 9 (APG9) | Glyma13g18300.2 | GDOL, GSYN | No |
| Exostosin family protein | Glyma13g18940.2 | GCAN. GCYR, GFAL, GDOL, GSTEN. GSYN, GTOM | Yes |
| Acyl-CoA N-acyltransferase with RING/FYVE/PHD-type zinc finger domain | Glyma13g19440.1 | GCAN. GCYR, GFAL, GDOL, GSTEN. GSYN, GTOM | Yes |
| AGC (cAMP-dependent, cGMP-dependent and protein kinase C) kinase family protein | Glyma13g18670.5 Glyma13g01190.6 | GCAN. GCYR, GFAL, GDOL, GSTEN. GSYN, GTOM | yes |
| CLIP-associated protein | Glyma13g19230.1 | GDOL | No |
| regulatory particle AAA-ATPase 2A | Glyma13g19280.1 | GCAN. GCYR, GFAL, GDOL, GSTEN. GSYN, GTOM | Yes |
| Ypt/Rab-GAP domain of gyp1p superfamily protein | Glyma13g18700.4 | GDOL, GSYN | No |
| arginine/serine-rich splicing factor 35 | Glyma07g08527.1 | GCAN. GCYR, GFAL, GDOL, GSTEN. GSYN, GTOM | Yes |
| Glutathione S-transferase family protein | Glyma13g19840.2 | GCAN. GCYR, GFAL, GDOL, GSTEN. GSYN, GTOM | Yes |
| cytochrome P450, family 93, subfamily D, polypeptide 1 | Glyma07g32330.1 Glyma13g24200.1 | GCAN. GCYR, GFAL, GDOL, GSTEN. GSYN, GTOM | Yes |
| downstream target of AGL15-4 | Glyma07g01870.1 | GCAN. GCYR, GFAL, GDOL, GSTEN. GSYN, GTOM | Yes |
| Isochorismatase family protein | Glyma07g02130.1 | GCAN. GCYR, GFAL, GDOL, GSTEN. GSYN, GTOM | Yes |
| Thioesterase superfamily protein | Glyma07g02160.2 | GDOL | No |
| Cytochrome P450 superfamily protein | Glyma07g04470.1 Glyma13g00990.1 | GCAN. GCYR, GFAL, GDOL, GSTEN. GSYN, GTOM | Yes |
| Galactosyltransferase family protein | Glyma13g01060.1 | GCAN. GCYR, GFAL, GDOL, GTOM | Yes |
| SKU5 similar 4 | Glyma07g39160.1 | GCAN. GCYR, GFAL, GDOL, GSTEN. GSYN, GTOM | Yes |
| flavonol synthase 1 | Glyma13g02740.1 | GCAN. GCYR, GFAL, GDOL, GSTEN. GSYN, GTOM | Yes |
| Cellulose synthase family protein | Glyma13g27250.3 | GCAN. GCYR, GFAL, GDOL, GSTEN. | Yes |

| | | | |
|---|---|---|---|
| | | GSYN, GTOM | |
| Catechol oxidase (Precursor) | Glyma07g31254.1 | GDOL | No |
| Polyphenol oxidase | Glyma07g31262.1<br>Glyma13g25181.1<br>Glyma13g31590.1<br>Glyma13g31595.1 | GCAN. GCYR, GFAL, GDOL, GSYN, GTOM | In some species |
| Tyrosinase family protein | Glyma07g31270.1<br>Glyma07g31280.1<br>Glyma07g31310.1<br>Glyma13g25150.1 | GCAN. GCYR, GFAL, GDOL, GSTEN, GSYN, GTOM | Yes |
| Probable oxidoreductase containing common central domain of tyrosinase | Glyma07g31301.1 | GCAN, GCYR, GDOL, GSYN | In some species |
| Oxidoreductase | Glyma13g25260.1 | GCAN. GCYR, GFAL, GDOL, GSTEN. GSYN, GTOM | Yes |
| Glycosyl hydrolase superfamily protein | Glyma07g03420.2 | GCAN. GCYR, GFAL, GDOL,GSYN, GTOM | Yes |
| Peroxidase superfamily protein | Glyma13g00790.1 | GCAN. GCYR, GFAL, GDOL, GTOM | Yes |
| Lysophospholipase 2 | Glyma13g00450.1 | GCAN. GCYR, GFAL, GDOL, GSTEN. GSYN, GTOM | Yes |
| Serine/threonine-protein kinase | Glyma07g00771.1 | GCAN. GCYR, GFAL, GDOL,GSYN, GTOM | Yes |
| RNA-binding (RRM/RBD/RNP motifs) family protein | Glyma07g00560.3 | GCAN. GCYR, GFAL, GDOL, GTOM | Yes |
| fucosyltransferase 1 | Glyma07g02540.1 | GCAN. GCYR, GFAL, GDOL,GSYN, GTOM | Yes |
| calmodulin-binding family protein | Glyma07g01031.1 | GDOL | No |
| autoinhibited Ca(2+)-ATPase 9 | Glyma07g00630.1 | GCAN. GCYR, GFAL, GDOL, GTOM | Yes |
| Lactoylglutathione lyase / glyoxalase I family protein | Glyma07g03560.1 | GDOL | No |
| fatty acid desaturase 5 | Glyma07g03370.2 | GCAN. GCYR, GFAL, GDOL, GSTEN. GSYN, GTOM | Yes |
| Glycosyl hydrolases family 32 protein | Glyma07g01090.1 | GDOL | No |
| Heat shock protein 70 (Hsp 70) family protein | Glyma07g00820.1 | GCAN. GCYR, GFAL, GDOL, GSTEN. GSYN, GTOM | Yes |
| Disease resistance protein (TIR-NBS-LRR class) family | Glyma07g00991.3 | GCAN. GCYR, GFAL, GDOL, GSTEN. GSYN, GTOM | Yes |
| structural molecules | Glyma07g00830.7 | GDOL | No |
| Insulinase (Peptidase family M16) protein | Glyma07g01720.1 | GDOL | No |
| SNARE-like superfamily protein | Glyma07g01000.2 | GCAN. GCYR, GFAL, GDOL | In some species |

| | | | |
|---|---|---|---|
| hexokinase 2 | Glyma07g01791.1 | GCAN. GCYR, GFAL, GDOL | In some species |
| sulfate transporter 3;4 | Glyma07g00840.1 | GCAN, GDOL | In some species |
| Nuclear transport factor 2 (NTF2) family protein | Glyma07g07041.1 | GCAN. GCYR, GFAL, GDOL | In some species |
| lipoxygenase 1 | Glyma07g00860.2 | GCAN. GCYR, GFAL, GDOL | In some species |
| casein kinase 1-like protein 2 | Glyma07g00970.1 | GCAN, GDOL | In some species |
| Late embryogenesis abundant (LEA) hydroxyproline-rich glycoprotein family | Glyma07g01200.1 | GCAN. GCYR, GFAL, GDOL | In some species |
| beta galactosidase 1 | Glyma07g01250.1 | GCYR,, GDOL | In some species |
| Root hair defective 3 GTP-binding protein (RHD3) | Glyma07g01230.1 | GFAL, GDOL | In some species |
| Disease resistance protein (CC-NBS-LRR class) family | Glyma07g06898.1 | GCAN. GCYR, GFAL, GDOL | In some species |

Table 4.17: Maximum Likelihood Estimate of Substitution Matrix

| | A | T/U | C | G |
|---|---|---|---|---|
| A | - | 11.49 | 6.17 | **4.91** |
| T/U | 11.33 | - | **5.49** | 6.29 |
| C | 11.33 | **10.23** | - | 6.29 |
| G | **8.83** | 11.49 | 6.17 | - |

Table 4.18: Maximum Composite Likelihood Estimate of the Pattern of Nucleotide Substitution

| | A | T | C | G |
|---|---|---|---|---|
| A | - | 4.34 | 2.31 | **19.5** |
| T | 4.28 | - | **6.61** | 2.4 |
| C | 4.28 | **12.44** | - | 2.4 |
| G | **34.81** | 4.34 | 2.31 | - |

Table 4.19: Results from Tajima's Neutrality Test. Abbreviations: m = number of sequences, n = total number of sites, S = Number of segregating sites, ps = S/n, Θ = ps/a1, π = nucleotide diversity, and D is the Tajima test statistic.

| M | S | Ps | Θ | Π | D |
|---|---|---|---|---|---|
| 338 | 343229 | 0.763422 | 0.119307 | 0.008319 | -2.907913 |

# REFERENCES

Bentley, D. R. Whole-genome re-sequencing. *Curr. Opin. Genet. Dev.* **16,** 545–52 (2006).

Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269,** 496–512 (1995).

Mendel, G. Versuche über Pflanzen-Hybriden. *Verh. Naturforsch. Ver.* **4,** 3–47 (1866).

Singer, S. R. *et al.* Venturing beyond beans and peas: what can we learn from Chamaecrista? *Plant Physiol.* **151,** 1041–7 (2009).

Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18,** 821–9 (2008).

Kreuzaler, F., Ragg, H., Fautz, E., Kuhn, D. & Hahlbrock, K. UV-induction of chalcone synthase mRNA in cell suspension cultures of Petroselinum hortense. *Proc Natl Acad Sci USA* **80,** 2591–2593 (1983).

Halward, T., Stalker, T., LaRue, E. & Kochert, G. Use of single-primer DNA amplifications in genetic studies of peanut (Arachis hypogaea L.). *Plant Mol. Biol.* **18,** 315–25 (1992).

Weller, J. L. *et al.* Update on the genetic control of flowering in garden pea. *J. Exp. Bot.* **60,** 2493–9 (2009).

Tucker, S. C. Update on Floral Development Floral Development in Legumes 1. **131,** 911–926 (2003).

Zhu, H., Choi, H.-K., Cook, D. R. & Shoemaker, R. C. Update on Comparative Genomics of Legumes Bridging Model and Crop Legumes through Comparative Genomics. *Plant Physiol.* **137,** (2005).

Schmidt, H. A., Strimmer, K., Vingron, M. & von Haeseler, A. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* **18,** 502–4 (2002).

Lawton, M. & Lamb, C. Transcriptional activation of plant defense genes by fungal elicitor, wounding and infection. *Mol Cell Biol* **7,** 335–341 (1987).

Chappell, J. & Hahlbrock, K. Transcription of plant defense genes in response to UV light or fungal elicitor. *Nature* **311,** 76–78 (1984).

Rambaut, A., Suchard, M., Xie, D. & Drummond, A. Tracer v1.6. (2014).

Reed, J. L., Famili, I., Thiele, I. & Palsson, B. O. Towards multidimensional genome annotation. *Nat. Rev. Genet.* **7,** 130–41 (2006).

Zabala, G. & Vodkin, L. O. The wp Mutation of Glycine max Carries a Gene-Fragment-Rich Transposon of the CACTA Superfamily. **17,** 2619–2632 (2005).

Barrett, J. R. The science of soy: what do we really know? *Environ. Health Perspect.* **114,** A352–8 (2006).

Doyle, J. & Luckow, M. The rest of the iceberg. Legume diversity and evolution in a phylogenetic context. *Plant Physiol* **131,** 900–910 (2003).

Hecht, V. *et al.* The pea GIGAS gene is a FLOWERING LOCUS T homolog necessary for graft-transmissible specification of flowering but not for responsiveness to photoperiod. *Plant Cell* **23,** 147–61 (2011).

McClintock, B. The origin and behavior of mutable loci in maize. *Proc Natl Acad Sci USA* **36,** 344–355 (1950).

Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4,** 406–425 (1987).

Young, N. D. *et al.* The Medicago genome provides insight into the evolution of rhizobial symbioses. *Nature* **480,** 520–4 (2011).

Harborne, J. & Baxter, H. *The Handbook of Natural Flavonoids 2 (Wiley, London). Handb. Nat. Flavonoids 2 (Wiley, London),* (1999).

Koes, R., Quattrocchio, F. & Mol, J. The flavonoid biosynthetic pathway in plants: function and evolution. *BioEssays* **16,** 123–132 (1994).

Moreau, C. *et al.* The B gene of pea encodes a defective flavonoid 3',5'-hydroxylase, and confers pink flower color. *Plant Physiol.* **159,** 759–68 (2012).

Kandaswami, C. *et al.* The antitumor activities of flavonoids. *In Vivo* **19,** 895–909

Soderlund, C., Nelson, W., Shoemaker, A. & Paterson, A. SyMAP: A system for discovering and viewing syntenic regions of FPC maps. *Genome Res.* **16,** 1159–68 (2006).

VandenBosch, K. & Stacey, G. Summaries of legume genomic projects from around the globe. Community resources for crops and models. *Plant Physiol.* **131,** 840–865 (2003).

Othman, S. A., Singh, B. B. & Mukhtar, F. B. Studies on the inheritance pattern of joints, pod and flower pigmentation in cowpea [ Vigna unguiculata ( L ) walp .]. **5,** 2371–2376 (2006).

Martin, C. Structure, function, and regulation of the chalcone synthase. *Int Rev Cytol* **147,** 233–283 (1993).

An, C. *et al.* Structure and organization of the genes encoding chalcone synthase in Pisum sativum. *Plant Mol. Biol.* **21,** 789–803 (1993).

Jez, J., Bowman, M., Dixon, R. & Noel, J. Structure and mechanism of the evolutionarily unique plant enzyme chalcone isomerase. *Nat Struct Biol* **7,** 786–791 (2000).

Junghans, H., Dalkin, K. and & Dixon, R. Stress response in alfalfa (Medicago sativa L.) 15. Characterization and expression patterns of members of a subset of the chalcone synthase multigene family. *Plant Mol. Biol.* **22,** 239–253 (1993).

Yang, Z. & Bielawski, J. Statistical methods for detecting molecular adaptation. *Trends Ecol Evol* **15,** 496–503 (2000).

Tajima, F. Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. *Genetics* **123,** 585–595 (1989).

Hartman, G., Wang, T. & Hymowitz, T. Sources of resistance to soybean rust in perennial Glycine species. *Plant Dis* **76,** 396–399 (1992).

Wu, X. *et al.* SNP discovery by high-throughput sequencing in soybean. *BMC Genomics* **11,** (2010).

Harris, T. D. *et al.* Single-molecule DNA sequencing of a viral genome. *Science* **320,** 106–9 (2008).

Sheahan, J. Sinapate esters provide greater UV-B attenuation than flavonoids in Arabidopsis thaliana (Brassicaceae). *Am. J. Bot.* **83,** 679–686 (1996).

Tajima, F. Simple Methods for Testing the Molecular Evolutionary Clock Hypothesis. *Genetics* **135,** 599–607 (1993).

Nei, M. & Gojobori, T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3,** 418–26 (1986).

Wasson, A. P., Pellerone, F. I. & Mathesius, U. Silencing the flavonoid pathway in Medicago truncatula inhibits root nodule formation and prevents auxin transport regulation by rhizobia. *Plant Cell* **18,** 1617–29 (2006).

Kevei, Z. *et al.* Significant microsynteny with new evolutionary highlights is detected between Arabidopsis and legume model plants despite the lack of macrosynteny. *Mol. Genet. Genomics* **274,** 644–657 (2005).

Treutter, D. Significance of flavonoids in plant resistance and enhancement of their biosynthesis. *Plant Biol. (Stuttg).* **7,** 581–91 (2005).

Mol, J., Jenkins, G., Schäfer, E., Weiss, D. & Walbot, V. Signal perception, transduction, and gene expression involved in anthocyanin biosynthesis. *CRC. Crit. Rev. Plant Sci.* **15,** 525–557 (1996).

Makoi, J., Belane, A., Chimphango, S. & Dakora, F. Seed flavonoids and anthocyanins as markers of enhanced plant defence in nodulated cowpea (Vigna unguiculata L. Walp.). *F. Crop. Res.* **118,** 21–27 (2010).

Newell, C. & Hymowitz, T. seed coat variation in Glycine wild subgenus Glycine (Leguminosae) by SEM. *Brittonia* **30,** 76–88 (1978).

Bennett, M. & Leitch, I. *Royal Botanical Gardens angiosperm DNA C-values database*. **2010,** (2010).

Scheffler, K., Martin, D. & Seoighe, C. Robust inference of positive selection from recombining coding sequences,. *Bioinformatics* **22,** 2493–2499 (2006).

Doyle, J. & Beachy, R. Ribosomal gene variation in soybean (Glycine max) and its realtives. *Theor. APPl. Genet.* **70,** 369–376 (1985).

Smit, A., Hubley, R. & Green, P. RepeatMasker Open-3.0. **-2012,** (1996).

Murphy, A., Peer, W. & Taiz, L. Regulation of auxin transport by aminopeptidases and endogenous flavonoids. *Planta* **211,** 315–324 (2000).

Lim, S. & Hymowitz, T. Reactions of perennial wild species of genus Glycine to Septoria glycines. *Plant Dis* **71,** 891–893 (1987).

Cramer, C., Ryder, T., Bell, J. & Lamb, C. Rapid switching of plant gene expression by fungal elicitor. *Science (80-. ).* **227,** 1240–1243 (1985).

Yu, O. *et al.* Production of the isoflavones genistein and daidzein in non-legume dicot and monocot tissues. *Plant Physiol* **124,** 781–793 (2000).

Untergasser, A. *et al.* Primer3--new capabilities and interfaces. *Nucleic Acids Res.* **40,** e115 (2012).

Kimura, M. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature* **267,** 275–276 (1977).

Riggs, R., Wang, S., Singh, R. & Hymowitz, T. Possible transfer of resistance to Heterodera glycines from Glycine tomentella to soybean. *J Nematol* **30,** 547–552 (1998).

Cannon, S. B. *et al.* Polyploidy did not predate the evolution of nodulation in all legumes. *PLoS One* **5,** e11630 (2010).

Montgomery, K. T. *et al.* PolyPhred analysis software for mutation detection from fluorescence-based sequence data. *Curr. Protoc. Hum. Genet.* **Chapter 7,** Unit 7.16 (2008).

Asen, S., Norris, K. & Stewart, R. Phytochemistry. *Phytochemistry* **11,** 2739–2741 (1972).

Hahlbrock, K. & Scheel, D. Physiology and molecular biology of phenylpropanoid metabolism. *Annu Rev Plant Physiol Plant Mol Biol* **40,** 347–369 (1989).

Rausher, M., Miller, R. & Tiffin, P. Patterns of evolutionary rate variation among genes of the anthocyanin biosynthetic pathway. *Mol Biol Evol* **16,** 266–274 (1999).

Ralston, L., Subramanian, S., Matsuno, M. & Yu, O. Partial reconstruction of flavonoid and isoflavonoid biosynthesis in yeast using soybean type I and type II chalcone isomerases. *Plant Physiol.* **137,** 1375–88 (2005).

Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13,** 555–556 (1997).

Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24,** 1586–91 (2007).

Ryder, T. *et al.* Organization and differential activation of a gene family encoding the plant defense enzyme chalcone synthase in Phaseolus vulgaris. *Mol. Gen. Genet. MGG* **210,** 219–233 (1987).

Doyle, M. J. & Brown, A. H. D. Numerical analysis of isozyme variation in Glycine tomentella. *Biochem. Syst. Ecol.* **13,** 413–419 (1985).

Howles, P., Arioli, T. and & Weinman, J. Nucleotide sequence of additional members of the gene family encoding chalcone synthase in Trifolium subterraneum. *Plant Physiol.* **107,** 1035–1036 (1995).

Arumuganathan, K. & Earle, E. Nuclear DNA content of some important plant species. *Pl Mol Biol Rept* **3,** 208–218 (1991).

Zabala, G. & Vodkin, L. Novel exon combinations generated by alternative splicing of gene fragments mobilized by a CACTA transposon in Glycine max. *BMC Plant Biol.* **7,** 38 (2007).

Shimodaira, H. & Hasegawa, M. Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. *Mol Biol Evol* 16:8:1114 (1999). at <http://libra.msra.cn/Publication/2051534/multiple-comparisons-of-log-likelihoods-with-applications-to-phylogenetic-inference>

Huelsenbeck, J. P. & Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17,** 754–755 (2001).

Warwick, S. I., and L. D. B. Molecular systematics of Brassica and allied genera (subtribe Brassicinae, Brassiceae)—chloroplast genome and cytodeme congruence. *Theor. Appl. Genet.* **82,** 81–92 (1991).

Durbin, M., McCaig, B. & Clegg, M. Molecular evolution of the chalcone synthase multigene family in the morning glory genome. *Plant Mol Biol* **42,** 79–92 (2000).

Nei, M. & Kumar, S. Molecular Evolution and Phylogenetics. *Oxford Univ. Press. New York* (2000).

Britsch, L., Ruhnau-Brich, B. & Forkmann, G. Molecular cloning, sequence analysis, and in vitro expression of flavanone 3 beta-hydroxylase from Petunia hybrida. *J Biol Chem* **267,** 5380–5387 (1992).

Charrier, B., Coronado, C., Kondorosi, A. & Ratet, P. Molecular characterization and expression of alfalfa (Medicago sativa L.) flavanone-3-hydroxylase and dihydroflavonol-4-reductase encoding genes. *Plant Mol Biol* **29,** 773–786 (1995).

Elomaa, P. & Holton, T. Modification of flower colour using genetic engineering Biotechnol. *Genet. Eng. Rev.* **12,** 63–88 (1994).

Aida, R., Kishimoto, S., Tanaka, Y. & Shibata, M. Modification of flower color in torenia (Torenia fournieri Lind.) by genetic transformation. *Plant Sci.* **153,** 33–42 (2000).

Courtney-Gutterson, N. *et al.* Modification of flower color in florist's chrysanthemum: production of a white-flowering variety through molecular genetics. *Biotechnology. (N. Y).* **12,** 268–71 (1994).

Schijlen, E., Ric de Vos, C., van Tunen, A. & Bovy, A. Modification of flavonoid biosynthesis in crop plants. *Phytochemistry* **65,** 2631–2648 (2004).

Schlueter, J. *et al.* Mining the EST databases to resolve evolutionary events in major crop species. *Genome* **47,** 868–76 (2004).

Dixon, R., Lamb, C., Masoud, S., Sewalt, V. & Paiva, N. Metabolic engineering: prospects for crop improvement through the genetic manipulation of phenylpropanoid biosynthesis and defense responses—a review. *Gene* **179,** 61–71 (1996).

Tamura, K. *et al.* MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* **28,** 2731–9 (2011).

Kumar, S., Tamura, K., Jakobsen, I. & and Nei, M. MEGA2: Molecular Evolutionary Genetics Analysis software. *Bioinformatics* **17,** 1244–1245 (2001).

Kishino, H., Miyata, T. & Hasegawa, M. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *J. Mol. Evol.* (1990). at <http://link.springer.com/article/10.1007/BF02109483>

Kishino, H., Miyata, T. & Hasegawa, M. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *J. Mol. Evol.* **31,** 151–160 (1990).

Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12,** 491 (2011).

Katoh, Misawa, Kuma & Miyata. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30,** 3059–3066 (2002).

Saslowsky, D. & Winkel-Shirley, B. Localization of flavonoid enzymes in Arabidopsis roots. *Plant J* **27,** 37–48 (2001).

Goldman, N., Anderson, J. P. & Rodrigo, A. G. Likelihood-based tests of topologies in phylogenetics. *Syst. Biol.* **49,** 652–70 (2000).

Goffeau, A. *et al.* Life with 6000 Genes. *Science (80-. ).* **274,** 546–567 (1996).

Graham, P. & Vance, C. Legumes: importance and constraints to greater use. *Plant Physiol* **131,** 872–877 (2003).

Vaughan, D. & Hymowitz, T. Leaf flavonoids of Glycine subgenus Glycine in relation to systematics. *Biochem. Syst. Ecol.* **12,** 189–192 (1984).

Skinner, M. E., Uzilov, A. V., Stein, L. D., Mungall, C. J. & Holmes, I. H. JBrowse: A next-generation genome browser. *Genome Res.* **19,** 1630–1638 (2009).

Bronski, M. *et al.* Isolation and characterization of thirteen polymorphic microsatellite loci in the A-genome perennial group of the legume genus Glycine. *Molec Ecol Res* **9,** 1548–1550 (2009).

Grotewold, E. & Peterson, T. Isolation and characterization of a maize gene encoding chalcone flavonone isomerase. *Mol. Gen. Genet.* **242,** 1–8 (1994).

Napoli, C., Lemieux, C. & Jorgensen, R. Introduction of a Chimeric Chalcone Synthase Gene into Petunia Results in Reversible Co-Suppression of Homologous Genes in trans. *Plant Cell* **2,** 279–289 (1990).

Quevillon, E. *et al.* InterProScan: protein domains identifier. *Nucleic Acids Res.* **33,** W116–20 (2005).

Jones, P. Binns, D. [...], and Hunter, S. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30,** 1236–1240 (2014).

Burbulis, I. & Winkel-Shirley, B. Interactions among enzymes of theArabidopsis flavonoid biosynthetic pathway. *Proc Natl Acad Sci USA* **96,** 12929–12934 (1999).

Burdon, J. & Marshall, D. inter and intraspecific diversity in the disease response of Glycine to the lead rust fungus Phakopsora pachyrhizi. *J, Ecol.* **69,** 381–390 (1981).

Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).

Frankel, E. ., German, J. ., Kinsella, J. ., Parks, E. & Kanner, J. Inhibition of oxidation of human low-density lipoprotein by phenolic substances in red wine. *Lancet* **341,** 454–457 (1993).

Strimmer, K. & Rambaut, A. Inferring confidence sets of possibly misspecified gene trees. *Proc. Biol. Sci.* **269,** 137–42 (2002).

Talavera, G. & Castresana, J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* **56,** 564–77 (2007).

Smith, C. *et al.* Improved repeat identification and masking in Dipterans. *Gene* **389,** 1–9 (2007).

Hyten, D. *et al.* Impacts of genetic bottlenecks on soybean genome diversity. *PNAS* **103,** 16666–16671 (2006).

Markowitz, V. M. *et al.* IMG ER: a system for microbial genome annotation expert review and curation. *Bioinformatics* **25,** 2271–8 (2009).

Hellens, R. P. *et al.* Identification of Mendel's white flower character. *PLoS One* **5,** e13230 (2010).

Chang, S. *et al.* Identification of high-quality single-nucleotide polymorphisms in Glycine latifolia using a heterologous reference genome sequence. *Theor. Appl. Genet.* **126,** 1627–1638 (2013).

Chen, Y. & Nelson, R. Identification and characterization of a white-flowered wild soybean plant. *Crop Sci* **44,** 339–342 (2004).

Zhu, T., Schupp, J., Oliphant, A. & Keim, P. Hypomethylated sequences: characterization of the duplicate soybean genome. *Mol. Gen. Genet.* **244,** 638–645 (1994).

Sergei, L., Kosakovsky, P., Frost, S. & Muse, S. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* **21,** 676–679 (2005).

Mol, J., Grotewold, E. & Koes, R. How genes paint flowers and seeds. *Trends Plant Sci.* **3,** 212–217 (1998).

González-Orozco, C., Brown, A., Knerr, N., Miller, J. & Doyle, J. Hotspots of diversity of wild Australian soybean relatives and their conservation in situ - Springer. *Conserv. Genet.* (2012). at <http://link.springer.com/article/10.1007/s10592-012-0370-x/fulltext.html>

Hyten, D. *et al.* High-throughput SNP discovery through deep resequencing of a reduced representation library to anchor and orient scaffolds in the soybean whole genome sequence. *BMC Genomics* **11,** (2010).

Fedorova, M. *et al.* Genome-wide identification of nodule-specific transcripts in the model legume Medicago truncatula. *Plant Physiol.* **130,** 519–37 (2002).

Vieler, A. *et al.* Genome, Functional Gene Annotation, and Nuclear Transformation of the Heterokont Oleaginous Alga Nannochloropsis oceanica CCMP1779. *PLoS Genet.* **8,** (2012).

Schmutz, J. *et al.* Genome sequence of the palaeopolyploid soybean. *Nature* **463,** 178–83 (2010).

The C. elegans Sequencing Consortium. Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology. *Science (80-. ).* **282,** 2012–2018 (1998).

Devos, K. & Gale, M. Genome relationships: the grass model in current research. *Plant Cell* **12,** 637–646 (2000).

Shoemaker, R. *et al.* Genome duplication in soybean (Glycine subgenus Soja). *Genetics* **144,** 329–338 (1996).

Stein, L. Genome annotation: from sequence to biology. *Nat. Rev. Genet.* **2,** 493–503 (2001).

Srinivasan, S. & Gaur, P. M. Genetics and characterization of an open flower mutant in chickpea. *J. Hered.* **103,** 297–302 (2011).

Holton, T. & Cornish, E. Genetics and biochemistry of anthocyanin biosynthesis. *Plant Cell* **7,** 1071–1083 (1995).

Gettys, L. a. Genetic control of white flower color in scarlet rosemallow (Hibiscus coccineus Walter). *J. Hered.* **103,** 594–7 (2012).

Wang, Z. *et al.* Genetic control of floral zygomorphy in pea (Pisum sativum L.). *Proc. Natl. Acad. Sci. U. S. A.* **105,** 10414–9 (2008).

Sadowski, J., P. Gaubier, M. Delseny, and C. F. Q. Genetic and physical mapping in Brassica diploid species of a gene cluster defined in Arabidopsis thaliana. *Mol. Gen. Genet.* **251,** 298–306 (1996).

Du, H., Huang, Y. & Tang, Y. Genetic and metabolic engineering of isoflavonoid biosynthesis. *Appl. Microbiol. Biotechnol.* **86,** 1293–312 (2010).

Domoney, C. *et al.* Genetic and genomic analysis of legume flowers and seeds. *Curr. Opin. Plant Biol.* **9,** 133–41 (2006).

Dooner, H. & Robbins, T. Genetic and developmental control of anthocyanin biosynthesis. *Ann. Rev. Genet.* **25,** 173–199 (1991).

Yang, K. *et al.* Genetic analysis of genes controlling natural variation of seed coat and flower colors in soybean. *J. Hered.* **101,** 757–68 (2010).

Schlueter, S. D., Dong, Q. & Brendel, V. GeneSeqer@PlantGDB: Gene structure prediction in plant genomes. *Nucleic Acids Res.* **31,** 3597–600 (2003).

Deroles, S. Bradley, M. Davies, K. Schwinn, K. Manson, D. Generation of novel patterns in lisianthus flowers using an antisense chalcone synthase gene. *Int. Soc. Hortic. Sci. 420 Ornam. Plant Improv. Class. Mol.* (1995). at <http://www.actahort.org/books/420/420_5.htm>

Borodovsky, M. & McIninch, J. D. GeneMark: Parallel gene recognition for both DNA strands. *Comp. Chem.* **17,** 123–133 (1993).

Nicholas, K. B., Jr., N. H. B. & Deerfield, D. W. I. GeneDoc: Analysis and Visualization of Genetic Variation. *EMBNEW.NEWS* (1997). at <http://www.nrbsc.org/gfx/genedoc/gdfeedb.htm>

Korf, I. Gene finding in novel genomes. BMC Bioinformatics. **5,** 1471–210 (2004).

Dong, X., Braun, E. & Grotewold, E. Functional conservation of plant secondary metabolic enzymes revealed by complementation of Arabidopsis flavonoid mutants with maize genes. *Plant Physiol* **127,** 46–57 (2001).

Forkmann, G. Flavonoids as flower pigments: The formation of the natural spectrum and its extension by genetic engineering. *Plant Breed* **106,** 1–26 (1991).

Dixon, R. and & Steele, C. Flavonoids and isoflavonoids: a gold mine for metabolic engineering. *Trends Plant Sci* **4,** 394–400 (1999).

Brown, D. *et al.* Flavonoids act as negative regulators of auxin transport in vivo in Arabidopsis. Plant Physiol. **126,** 524–535 (2001).

Talukdar, D. Flavonoid-deficient mutants in grass pea (Lathyrus sativus L.): genetic control, linkage relationships, and mapping with aconitase and S-nitrosoglutathione reductase isozyme loci. *ScientificWorldJournal.* **2012,** 345983 (2012).

Van der Krol, A. R. Flavonoid Genes in Petunia: Addition of a Limited Number of Gene Copies May Lead to a Suppression of Gene Expression. *PLANT CELL ONLINE* **2,** 291–299 (1990).

Stafford, H. A. Flavonoid Evolution: An Enzymic Approach. *PLANT Physiol.* **96,** 680–685 (1991).

Winkel-Shirley, B. Flavonoid biosynthesis. A colorful model for genetics, biochemistry, cell biology, and biotechnology. *Plant Physiol.* **126,** 485–493 (2001).

Marinova, K., Kleinschmidt, K., Weissenböck, G. & Klein, M. Flavonoid biosynthesis in barley primary leaves requires the presence of the vacuole and controls the activity of vacuolar flavonoid transport. *Plant Physiol.* **144,** 432–44 (2007).

Redmond, J. W. *et al.* Flavones induce expression of nodulation genes in Rhizobium. *Nature* **323,** 632–635 (1986).

Deboo, G., Albertsen, M. & Taylor, L. Flavanone 3-hydroxylase transcripts and flavonol accumulation are temporally coordinate in maize anthers. *Plant J* **7,** 703–713 (1995).

Laplaze, L. *et al.* Flavan-containing cells delimit Frankia-infected compartments inCasuarina glauca nodules. *Plant Physiol.* **121,** 113–122 (1999).

Rambaut, A. FigTree v.1.4.0. (2008). at <http://tree.bio.ed.ac.uk/software/figtree/>

Salamov, A. & Solovyev, V. Fgenesh multiple gene prediction program: httpd. *genomic. sanger. ac. uk* (1998).

Curriculum, K. *Fathom Dynamic Data Software: Version 2*. (2006).

Weeden NF, Muehlbauer FJ, L. G. Extensive conservation of linkage relationships between pea and lentil genetic maps. *J Hered* **83,** 123–129 (1992).

Meldgaard, M. Expression of chalcone synthase, dihydroflavonol reductase and flavanone-3-hydroxylase in mutants of barley deficient in anthocyanidin and proanthocyanidin biosynthesis. *Theor Appl Genet* **83,** 695–706 (1992).

Xu, M., Brar, H. K., Grosic, S., Palmer, R. G. & Bhattacharyya, M. K. Excision of an active CACTA-like transposable element from DFR2 causes variegated flowers in soybean [Glycine max (L.) Merr.]. *Genetics* **184,** 53–63 (2010).

Doyle, J., Doyle, J., Rauscher, J. & Brown, A. Evolution of the perennial soybean polyploid complex (Glycine subgenus Glycine): a study of contrasts. *Biol. J. Linn. Soc.* **82,** 583–597 (2004).

Ngaki, M. N. *et al.* Evolution of the chalcone-isomerase fold from fatty-acid binding to stereospecific catalysis. *Nature* **485,** 530–3 (2012).

Durbin, M., Learn, G., Huttley, G. and & Clegg, M. Evolution of the chalcone synthase gene family in the genus Ipomoea. *Proc. Natl. Acad. Sci. USA* **92,** 3338–3342 (1995).

Kishino, H. & Hasegawa, M. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. *J. Mol. Evol.* **29,** 170–179 (1989).

Hartman, G., Gardner, M., Hymowitz, T. & Naidoo, G. Evaluation of perennial Glycine species for resistance to soybean fungal pathogens that cause Sclerotinia stem rot and sudden death syndrome. *Crop Sci* **40,** 545–549 (2000).

Burdon, J. & Marshall, D. Evaluation of Australian native species of Glycine for resistance to soybean rust. *Plant Dis* **65,** 44–45 (1981).

Tamura, K. & Nei, M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* **10,** 512–26 (1993).

Nielsen, R. & Yang, Z. Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA. *Mol Biol Evol* **20,** 1231–1239 (2003).

Choi, H. *et al.* Estimating genome conservation between crop and model legume species. *PNAS* **101,** 15289–15294 (2004).

Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16,** 276–7 (2000).

Ryder, T. *et al.* Elicitor rapidly induces chalcone synthase mRNA in Phaseolus vulgaris cells at the onset of the phytoalexin defense response. *Proc Natl Acad Sci USA* **81,** 5724–5728 (1984).

Pang, Y., Peel, G., Wright, E., Wang, Z. & Dixon, R. Early steps in proanthocyanidin biosynthesis in the model legume Medicago truncatula. *Plant Physiol.* **145,** 601–15 (2007).

Varshney, R. *et al.* Draft genome sequence of pigeonpea (Cajanus cajan), an orphan legume crop of resource-poor farmers. *Nat Biotech* **30,** 83–89 (2012).

Varshney, R. *et al.* Draft genome sequence of chickpea (Cicer arietinum) provides a resource for trait improvement. *Nat Biotech* **31,** 240– 246 (2013).

Kumar, S. & Gadagkar, S. Disparity Index: a simple statistic to measure and test the homogeneity of substitution patterns between molecular sequences. *Genetics* **158,** 1321–1327 (2001).

Doyle, J., Doyle, J., Rauscher, J. & Brown, A. Diploid and polyploid reticulate evolution throughout the history of the perennial soybeans (Glycine subgenus Glycine). *New Phytol.* **161,** 121–132 (2004).

Prescott, A. & John, P. Dioxygenases: molecular structure and role in plant metabolism. *Annu Rev Plant Physiol Plant Mol Biol* **47,** 245–271 (1996).

Wingender, R., Röhrig, H., Höricke, C., Wing, D. & Schell, J. Differential regulation of soybean chalcone synthase genes in plant defense, symbiosis and upon environmental stimuli. *Mol. Gen. Genet.* **218,** 315–322 (1989).

Ministerial Meeting on Population of the Non-Aligned Movement (1993: Bali). Denpasar Declaration on Population and Development. *Integration* 27–9 (1994). doi:10.1234/2013/999990.

Reinhardt, J. *et al.* De novo assembly using low-coverage short read sequence data from the rice pathogen Pseudomonas syringae pv. oryzae. *Genome Res.* **19,** 294–305 (2009).

Li, R. *et al.* De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20,** 265–72 (2010).

Pond, S. & Frost, S. Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics* **21,** 2531–2533 (2005).

Delport, W., Poon, A., Frost, S. & Kosakovsky Pond, S. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics* (2010).

Goldblatt, P. Cytology and the phylogeny of Leguminosae. In: PolhillRM, RavenPH, eds. Advances in legume systematics. *R. Bot. Gard.* 427–463 (1981).

Grant, J., Brown, A. & Grace, J. Cytological and Isozyme Diversity in Glycine tomentella Hayata (Leguminosae). *Aust. J. Bot.* **32,** 665 (1984).

Frazer, K. A., Elnitski, L., Church, D. M., Dubchak, I. & Hardison, R. C. Cross-species sequence comparisons: a review of methods and available resources. *Genome Res.* **13,** 1–12 (2003).

Britsch, L., Heller, W. & Grisebach, H. Conversion of flavanone to flavone, dihydroflavonol and flavonol with an enzyme system from cell cultures of parsley. *Z Naturforsch Sect C. Biosci* **36,** 742–750 (1981).

Foucher, F. *et al.* Conservation of Arabidopsis Flowering Genes in Model Legumes 1 [ w ]. **137,** 1420–1434 (2005).

Shimodaira, H. & Hasegawa, M. CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* **17,** 1246–1247 (2001).

Sherman-Broyles, S., Bombarely, A., Grimwood, J., Schmutz, J. & Doyle, J. Complete Plastome Sequences from Glycine syndetika, and Six Additional Perennial Wild Relatives of Soybean. *G3 (Bethesda).* g3.114.012690– (2014). doi:10.1534/g3.114.012690

Bennetzen, J. Comparative sequence analysis of plant nuclear genomes: microcolinearity and its many exceptions. *Plant Cell* **12,** 1021–1029 (2000).

Chang, S. *et al.* Comparative mapping of the wild perennial Glycine latifolia and soybean (G. max) reveals extensive chromosome rearrangements in the genus Glycine. *PLoS One* **9,** e99427 (2014).

Menancio-Hautea D, Fatokum CA, Kumar L, Danesh D, Y. N. Comparative genome analysis of mungbean (Vigna radiata (L.) Wilczek) and cowpea (V. unguiculata (L.) Walpers) using RFLP mapping data. *Theor Appl Genet* **86,** 797–810 (1993).

Mies, D. & Hymowitz, T. Comparative electrophoretic studies of trypsin inhibitors in seed of the genus Glycine. *Bot. Gaz.* **134,** 121–125 (1973).

Kovinich, N., Saleem, A., Arnason, J. T. & Miki, B. Combined analysis of transcriptome and metabolite data reveals extensive differences between black and brown nearly-isogenic soybean (Glycine max) seed coats enabling the identification of pigment isogenes. *BMC Genomics* **12,** 381 (2011).

Lunau, K., Wacht, S. & Chittka, L. Colour choices of naive bumble bees and their implications for colour perception. *J. Comp. Physiol. A* **178,** (1996).

Thompson, J., Higgins, D. & Gibson, T. ClustalW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22,** 4673–4680 (1994).

Djordjevic, M. A., Redmond, J. W., Batley, M. & Rolfe, B. G. Clovers secrete specific phenolic compounds which either stimulate or repress nod gene expression in Rhizobium trifolii. *EMBO J.* **6,** 1173–9 (1987).

Koes, R., Spelt, C. and & Van den Elzen, P. Cloning and molecular characterization of the chalcone synthase multigene family of Petunia hybrida. *Gene* **81,** 245–257 (1989).

Sparvoli, F., Martin, C., Scienza, A., Gavazzi, G. & Tonelli, C. Cloning and molecular analysis of structural genes involved in flavonoid and stilbene biosynthesis in grape (Vitis vinifera L.). *Plant Mol. Biol.* **24,** 743–55 (1994).

Gong, Z., Yamazaki, M., Sugiyama, M., Tanaka, Y. & Saito, Y. Cloning and molecular analysis of structural genes involved in anthocyanin biosynthesis and expressed in 13 forma-specific manner in Perilla frutescenes. *Plant Mol Biol* **6,** 915–927 (1997).

Shen, G. *et al.* Cloning and Characterization of a Flavanone 3-Hydroxylase Gene from Ginkgo biloba. *Biosci. Rep.* **26,** 19–29 (2006).

Wink, M. & Waterman, D. Chemotaxonomy in relation to molecular phylogeny of plants. *Annu. plant Rev* **2,** 300–341 (1999).

Mehdy, M. C. & Lamb, C. J. Chalcone isomerase cDNA cloning and mRNA induction by fungal elicitor, wounding and infection. *EMBO J.* **6,** 1527–33 (1987).

Zhu, H., HK, C., Cook, D. & Shoemaker, R. Bridging model and crop legumes through comparative genomics. *Plant Physiol.* **137,** 1189–96 (2005).

Conesa, A. *et al.* Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21,** 3674–6 (2005).

Broue, P., Marshall, D. & Muller, W. Biosystematics of Subgenus Glycine (Verdc.): Isoenzymatic Data. *Aust. J. Bot.* **25,** 555 (1977).

Heller, W. & Forkmann, G. Biosynthesis of flavonoids. In: Harborne J.B. (eds) The flavonoids, advances in research since 1986. *Chapman & Hall, London* 399–425 (1993).

Wharton, J. *et al.* Bean Briefs. 1–8 (2012). doi:10.1155/2012/829238.In

Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29,** 1969–73 (2012).

Heled, J. & Drummond, A. J. Bayesian inference of species trees from multilocus data. *Mol. Biol. Evol.* **27,** 570–80 (2010).

Nylander, J. A. A., Wilgenbusch, J. C., Warren, D. L. & Swofford, D. L. AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. *Bioinformatics* **24,** 581–3 (2008).

Sergei, L. *et al.* Automated Phylogenetic Detection of Recombination Using a Genetic Algorithm. *Mol. Biol. Evol.* **23,** 1891–1901

Slater, G. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6,** 1471–2105 (2005).

Stanke, M., Tzvetkova, A. & Morgenstern, B. AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. *Genome Biol.* **7,** S11–S18 (2006).

Landry, L., Chapple, C. & Last, R. Arabidopsis mutants lacking phenolic sunscreens exhibit enhanced ultraviolet-B injury and oxidative damage. *Plant Physiol* **109,** 1159–1166 (1995).

Li, J., Ou-Lee, T. M., Raba, R., Amundson, R. G. & Last, R. L. Arabidopsis Flavonoid Mutants Are Hypersensitive to UV-B Irradiation. *Plant Cell* **5,** 171–179 (1993).

Lewis, S. *et al.* Apollo: a sequence annotation editor. *Genome Biol.* **3,** research0082.1–0082.14 (2002).

Forkmann, G., Heller, W. & Grisebach, H. Anthocyanin biosynthesis in flowers of Matthiola incana. Flavanone 3- and flavonoid 3′-hydroxylases. *Z Naturforsch Sect C. Biosci* **35,** 691–695 (1980).

The Arabidopsis Genome, I. Analysis of the genome sequence of the flowering plant Arabidopsis Thaliana. *Nature* **408,** 796–815 (2000).

Pelletier, M. & Winkel-Shirley, B. Analysis of flavanone 3-hydroxylase in Arabidopsis seedlings. *Plant Physiol* **111,** 339–345 (1996).

Winkel-Shirley, B. *et al.* Analysis of Arabidopsis mutants deficient in flavonoid biosynthesis. *Plant J* **8,** 659–671 (1995).

Shimodaira, H. An approximately unbiased test of phylogenetic tree selection. *Syst. Biol.* **51,** 492–508 (2002).

Van Der Krol, A. *et al.* An anti-sense chalcone synthase gene in transgenic plants inhibits flower pigmentation. *Nature* **333,** 866–869 (1988).

Elomaa, P. *et al.* Agrobacterium-Mediated Transfer of Antisense Chalcone Synthase cDNA to Gerbera hybrida Inhibits Flower Pigmentation. *Bio/Technology* **11,** 508–511 (1993).

Costanza, S. & Hymowitz, T. adventitious roots in Glycine subg. Glycine (Leguminosae): Morphological and taxonomic indicators of the B genome. *Plant Syst. Evol.* **158,** 37–46 (1987).

Polhill, R. & Raven, P. Advances in legume Systematics part 1. *R. Bot. Gard. Kew* **425,** (1981).

Hasegawa, M. & Kishino, H. Accuracies of the simple methods for estimating the bootstrap probability of a maximum-likelihood tree. *Mol. Biol. Evol.* (1994). at <http://mbe.oxfordjournals.org/content/11/1/142.short>

Simpson, J. T. *et al.* ABySS: a parallel assembler for short read sequence data. *Genome Res.* **19,** 1117–23 (2009).

Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16,** 111–20 (1980).

Schmutz, J. *et al.* A reference genome for common bean and genome-wide analysis of dual domestications. *Nat. Genet.* **46,** 707–13 (2014).

Newell, C. & Hymowitz, T. A reappraisal of the subgenus Glycine. *Amer. J. Bot* **65,** 168–179 (1978).

Doyle, JJ. Doyle, J. A Rapid DNA Isolation Procedure for Small Quantities of Fresh Leaf Tissue. *Phytochem. Bull.* **19,** 11–15 (1987).

MEYER, P., HEIDMANN, I., FORKMANN, G. & SAEDLER, H. A new petunia flower colour generated by transformation of a mutant with a maize gene. **330,** 677–678 (1987).

Takahashi, R., Dubouzet, J. G., Matsumura, H., Yasuda, K. & Iwashina, T. A new allele of flower color gene W1 encoding flavonoid 3’5'-hydroxylase is responsible for light purple flowers in wild soybean Glycine soja. *BMC Plant Biol.* **10,** 155 (2010).

Muse, S. & Gaut, B. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol Biol Evol* **11,** 715–724 (1994).

Kosakovsky, P. & Frost, S. A Genetic Algorithm Approach to Detecting Lineage-specific Variation in Selection Pressure. *Mol. Biol. Evol.* **22,** 478–485 (2005).

Morrison, D. A. A framework for phylogenetic sequence alignment - Springer. *Plant Syst. Evol.* (2008). at <http://link.springer.com.librarylink.uncc.edu/article/10.1007/s00606-008-0072-5/fulltext.html#CR15>

Goldman, N. & Yang, Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol* **11,** 725–736 (1994).

Shimada, N. *et al.* A cluster of genes encodes the two types of chalcone isomerase involved in the biosynthesis of general flavonoids and legume-specific 5-deoxy (iso)flavonoids in Lotus japonicus. *Plant Physiol.* **131,** 941–51 (2003).

Davies, K. A cDNA clone for flavanone 3-hydroxlase from Malus. *Plant Physiol* **103,** 291 (1993).

Doyle, J., Doyle, J. & Brown, A. 5S nuclear ribosomal gene variation in the Glycine tomentella polyploid complex ( Leguminosae). *Syst. Boi.* **14,** 398–407 (1989).

Ono, E. *et al.* ‘Yellow flowers generated by expression of the aurone biosynthetic pathway’. *Proc. Natl. Acad. Sci. United States Am.* **103,** 11075–80 (2006).

Gordon, A. & Hannon, G. ‘Fastx-toolkit.’ FASTQ/A short-reads pre-processing tools. (2010).

Altschul, S., Gish, W., Miller, W., Myers, E. . & Lipman, D. ‘Basic local alignment search tool.’ *J. Mol. Biol.* **215,** 403–410 (1990).