The Dissertation Committee for Erika Nicole Schwarz Certifies that this is the approved version of the following dissertation:

Plastid and Mitochondrial Genome Evolution of Legumes (Fabaceae)

Committee:

Robert K. Jansen, Supervisor

David M. Hillis

Craig Randal Linder

Stanley J. Roux, Jr.

Edward C. Theriot

Plastid and Mitochondrial Genome Evolution of Legumes (Fabaceae)

by

Erika Nicole Schwarz, B.S.; M.S.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

August 2016

Plastid and Mitochondrial Genome Evolution of Legumes (Fabaceae)

Erika Nicole Schwarz, PhD
The University of Texas at Austin, 2016

Supervisor: Robert K. Jansen

Plastid genome (plastome) organization is highly conserved across seed plants with a quadripartite structure including the small single copy (SSC), the large single copy (LSC) and two copies of an inverted repeat (IR). There are several unrelated lineages that have experienced extensive structural rearrangements such as inversions and gene/intron losses and indels. Fabaceae is typically recognized as having three subfamilies: Caesalpinioideae, Mimosoideae and Papilionoideae. Publicly available plastid genomes of legumes have for the most part been limited to the subfamily Papilionoideae due to their economic importance and known structural rearrangements. In several other angiosperm lineages, correlations between accelerated rates of genomic rearrangements and nucleotide substitition rates in the plastome have been identified. Additionally, increased frequency of plastome structural changes and accelerated nucleotide substitutions have been shown to be correlated with increased evolutionary rates in the mitochondrial genome (mitogenome). To date, few legume mitochondrial genomes (7) are publicly available. My dissertation research uses Fabaceae to investigate 1) plastid genomic changes and rearrangements across all three subfamilies and 2) correlations between biological features and nucleotide substitution rates of both plastid and mitochondrial genes. Chapter two focuses on plastid structural evolution across three subfamilies of Fabaceae and shows papilionoids have smaller genomes with varying degrees of genomic rearrangements, and they have experienced multiple, independent gene/intron losses and inversions that limit the phylogenetic utility of these changes. Chapter three finds accelerated substitution rates in protein coding plastome genes among papilionoid taxa, especially those lacking one copy of the inverted repeat (IR), decreased rates in genes previously contained in the IR, and faster rates in herbaceous versus woody taxa. Chapter four focuses on substitution rates of mitochondrial genes and shows a correlation between plastid and mitochondrial substitution rates in addition to an acceleration in the papilionoid taxa, where, again, the herbaceous habit is correlated with higher rates.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

While the plastome organization in angiosperms is typically well conserved (Ruhlman and Jansen 2014), previous studies have shown extreme rearrangements in Campanulaceae (Cosner et al. 2004; Haberle et al. 2008; Knox 2014), Ericaceae (Fajardo et al. 2013; Martınez-Alberola et al. 2013), Geraniaceae (Chumley et al. 2006; Blazier et al., 2011, 2016a; Guisinger et al. 2011; Weng et al., 2014), Oleaceae (Lee et al. 2007) and Fabaceae (Palmer 1985; Milligan et al. 1989; Cai et al. 2008; Sabir et al. 2014). Correlations between nucleotide substitution rates and genomic rearrangements in the plastid have been shown in Caryophyllaceae (Sloan et al., 2012) and Geraniaceae (Guisinger et al., 2008, 2011; Weng et al., 2014; Grewe et al., 2015). In addition, rates of both structural and nucleotide change in the plastid are correlated with nucleotide substitution rates in the mitochondria (Wolfe et al., 1987; Zhu et al., 2014). In Fabaceae three subfamilies have been traditionally recognized: Caesalpinioideae, Mimosoideae, Papilionoideae. Previously, studies of plastomes in Fabaceae have been limited to the papilionoids due to their economic importance (Wojciechowski et al. 2004; LPWG 2013). Within the papilionoids a major focus has been on a monophyletic group comprising taxa lacking one copy of the inverted repeat (IR), known as the IRLC (Palmer and Thompson 1981; Wojciechowski et al., 2004). Because of the varying levels of plastid rearrangements within Fabaceae it provides an excellent group to study changes of the plastid organization across all three subfamilies and to investigate correlations between plastid and mitochondrial substitution rates.

Chapter two focuses on plastid genome organization of legumes. Previously, Sabir et al. (2014) focused on just one subset of one subfamily of the legumes, the IR-lacking clade within the papilionoids, and found size differences along with varying amounts of repetitive DNA. In order to expand upon these earlier studies, taxon sampling was increased to include species from all three subfamilies to determine the extent and cause of genome size differences within legumes. With a better representation of the legumes, I found that the basal subfamilies, Caesalpinioideae and Mimosoideae, retain ancestral gene content and order of angiosperms. However, Papilionoideae have smaller plastomes, gene losses and genome rearrangements throughout the subfamily. Additionally, I found evidence that genome rearrangements within legumes may not be as phylogenetically informative as previously thought. This is due to a 36 kb inversion that is present in two distantly related papilionoid taxa that appears to have been caused by the same mechanism, a 29 bp repeat that flanks both sides of the inversion.

Chapter three explores plastome-wide substitution rates in across four legume subgroups: caesalpinioids, mimosoids, papilionoids that contain an inverted repeat and papilionoids lacking the IR (IRLC). I found consistently accelerated rates in papilionoid legumes compared to caesalpinioid and mimosoid taxa. In addition, genes in the IR region have lower substitution rates than genes in either the large single copy (LSC) or the small single copy (SSC) regions across all legume subgroups examined. Genes formerly in the IR also have lower substitution rates than genes in the LSC and SSC but have higher rates than genes contained within the IR in other papilionoids, caesalpinioids and mimosoids. Lastly, I detected a negative correlation between genome size and nucleotide substitution rates and positive correlations between genome rearrangements and number of indels, and nucleotide substitution rates.

Chapter four focuses on nucleotide substitution rates in mitochondrial genes across legumes. Values of *dN* and *dS* of mitochondrial genes are accelerated in papilionoid legumes compared to caesalpinioids and mimosoids. In addition, several genes and certain lineages within the legumes exhibit accelerated rates. The accelerated genes also have fewer RNA editing sites than other genes suggesting mutagenic retroprocessing may play a role in legume rate variation. The branches leading to all legumes and to several individual taxa (i.e., *Prosopis glandulosa, Arachis hypogaea, Medicago trunctula, Trifolium repens*) have many genes with highly accelerated *dN* values. Values of *dS* show similar accelerations in the branch leading to all legumes and the branch leading to *A. hypogaea*. Comparisons of mitochondrial and plastid rates revealed that *dS* in plastid genes are approximately four times higher than rates of the mitochondrial genes and *dN* is about one and a half times faster in plastid genes. Comparison of rates in both mitochondrial and plastid genomes also revealed that both *dN* and *dS* are accelerated in the papilionoid lineage of legumes compared to the basal caesalpinioid and mimosoid lineages.

Chapter 2: Plastid genome sequences of legumes reveal parallel inversions and multiple losses of *rps16* in papilionoids[1]

INTRODUCTION

Fabaceae (legumes) are the third largest family of angiosperms with an estimated 20,000 species that vary greatly in habitat and growth habit (Lewis et al. 2005; LPWG 2013). Traditionally legumes are thought to comprise three subfamilies; Caesalpinioideae, Mimosoideae and Papilionoideae (Wojciechowski et al. 2004). Caesalpinioids, a paraphyletic grade from which mimosoids and papilionoids were derived (Wojciechowski et al. 2004; LPWG 2013), include approximately 2,250 species and are primarily tropical in nature ranging in size from shrubs to large trees (LPWG 2013). Mimosoids, the second largest group of legumes with approximately 3,270 species, are also shrubs to large trees (LPWG 2013). However, mimosoids have a much wider geographic distribution than caesalipinioids and play a vital ecological role in a variety of pantropical habitats (Luckow et al. 2003; LPWG 2013). Papilionoids, including about 13,800 species, are the largest and most well studied group of legumes due to their ecological and economical importance (Wojciechowski et al. 2004; LPWG 2013).

The plastid genome, or plastome, is highly conserved across seed plants with respect to size, gene order and its quadripartite structure consisting of a large single copy region (LSC), a small single copy region (SSC) and a large inverted repeat (IR).

[1] This chapter has been published in Schwarz, E.N., Ruhlman, T., Sabir, JSM., Hajrah, NH., Alharbi, NS., Al-Malki, AL., Bailey, CD and Jansen, R.K. 2015. Plastid genomes reveal parallel inversions and multiple losses of *rps16* in papilionoids. Journal of Systematics and Evolution. 53(5): 458-468.

Currently there are 525 seed plant plastomes available on NCBI with sizes ranging from approximately 62 to 218 kb (http://www.ncbi.nlm.nih.gov/genomes/GenomesGroup .cgi?taxid=2759&opt=plastid). The majority of photosynthetic seed plant plastomes range from approximately 110 to 170 kb with an average size of approximately 154 kb among angiosperms (Weng et al. 2014), while those taxa with plastomes under 107 kb in size, however, have parasitic lifestyles and have significantly reduced genomes due to rampant gene loss (Krause 2012).

Variation in plastome size or gene order within groups is relatively rare. Nevertheless, there are certain groups of seed plants that exhibit significant size variation, plastome rearrangements and gene and intron losses (Green 2011; Wicke et al. 2011; Jansen & Ruhlman 2012). Variation in plastome size is typically attributed to IR expansion, contraction or loss (Perry & Wolfe 2002; Chumley et al. 2006; Guisinger et al. 2011; Wicke et al. 2011). However, within two groups of gymnosperms, gnetophytes and cupressophytes, size variation is due to a decrease in intergenic spacer size (McCoy et al. 2008; Wu et al. 2009; Wu & Chaw 2014). The largest seed plant plastome is that of *Pelargonium x hortorum* (217,942 bp) in the Geraniaceae and its large size is due to an expansion of the IR to three times the normal size (Chumley et al. 2006). In contrast, smaller genome sizes are often due to IR contraction or loss. Reductions in IR size have been documented in Geraniaceae (Guisinger et al. 2011) and Pinaceae (Tsudzuki et al. 1992) and cases of

IR loss have been documented in Orobanchaceae (Bock & Knoop 2012), Geraniaceae (Blazier et al. 2011; Guisinger et al. 2011) and Fabaceae (Palmer and Thompson 1981; Lavin et al. 1990; Liston 1995). Within Fabaceae, most size variation has been attributed to the loss of the IR in one clade of the papilionoids, the IR Lacking Clade (IRLC) (Wojciechowski et al. 2004).

Similar to plastome size variation, plastid genomic rearrangements are also relatively rare with the exception of a few groups of seed plants. Extensive rearrangements have been documented in some gymnosperms (McCoy et al. 2008; Wu & Chaw 2014) as well as angiosperm families including Campanulaceae (Cosner et al. 2004; Haberle et al. 2008; Knox 2014), Ericaceae (Fajardo et al. 2013; Martınez-Alberola et al. 2013), Geraniaceae (Chumley et al. 2006; Guisinger et al. 2011), Oleaceae (Lee et al. 2007) and Fabaceae (Palmer 1985; Milligan et al. 1989; Cai et al. 2008; Sabir et al. 2014). In photosynthetic seed plants gene and intron losses are restricted to a small number of families (Jansen & Ruhlman 2012). The majority of gene losses across seed plants are found among gymnosperms within gnetophytes and Pinaceae, and within the angiosperm families Campanulaceae, Fabaceae, Geraniaceae, Passifloraceae and Poaceae (Jansen & Ruhlman 2012).

Due to the overall conserved nature of plastomes, events such as genomic rearrangements and gene and intron loss can be powerful phylogenetic markers. In the Asteraceae a 22 kb inversion identified Barnadesioideae as sister to the rest of the family, which is congruent with phylogenies based on both gene sequences and morphological characters (Jansen & Palmer 1987; Bremer, 1987; Kim et al. 2005). A large 50 kb inversion present in most of the papilionoid legume taxa has proven to be consistent with phylogenies from molecular sequences (Wojciechowski et al. 2004). The most extensive use of plastome inversions for phylogeny reconstruction was performed in the Campanulaceae (Cosner et al. 2004). Despite the large number of inversion events (84) there were lower levels of homoplasy than in trees generated from gene sequences, supporting previous suggestions that inversions are useful and reliable phylogenetic characters. Gene and intron losses can also be phylogenetically informative markers. For example, four gene losses, *chlB*, *chlL*, *chlN* and *trnP*-GGG, are synapomorphies for flowering plants and six other genes have been lost only once among angiosperms (Jansen et al. 2007). In Geraniaceae, a number of gene and intron losses are homoplasious but there are many others that are synapomorphies within the family (Guisinger et al. 2011). The transfer of *rpl32* to the nucleus at the base of the Ranunculaceae subfamily Thalictroideae was useful in supporting the monophyly of this subfamily (Park et al. 2015). Gene and intron losses have been studied extensively in legumes. Early studies utilized Southern hybridization and PCR to examine gene and intron losses in the family (Doyle et al. 1995; Bailey et al. 1997). One of the most extensive comparisons involved the gene *rpl22*, which was shown to be lost in all legumes (Doyle et al. 1995). Later, Gantt et al. (1991) confirmed that *rpl22* had been transferred to the nucleus where its transcript, including sequences encoding a plastid targeting peptide, is expressed.

More recent investigations based on plastome sequences identified several other gene and intron losses in legumes. The transfer of plastid *accD*, coding for Acetyl-CoA carboxylase, to the nucleus has been confirmed in two *Trifolium* species (Magee et al. 2010; Sabir et al. 2014). In another case of legume gene loss, *rps16* has been confirmed missing from the plastome of all IRLC species and in *Phaseolus vulgaris* (Guo et al. 2007; Magee et al. 2010; Sabir et al. 2014). In *Medicago truncatula* the *rps16* loss was facilitated by a gene substitution in which a nuclear encoded, mitochondrial targeted gene has acquired dual targeting and is now directed to both the mitochondrion and the plastid (Ueda et al. 2008). Within the IRLC, the introns of *rps12* and *clpP* have also been lost as determined by extensive survey of many individual taxa (Jansen et al. 2008; Sabir et al. 2014).

Previous plastid genome investigations have elucidated important events and characteristics of genome evolution but for a relatively small number of species restricted to only one of the three subfamilies of legumes, papilionoids. In this study we present 13 new legume plastome sequences, including the first caesalpinioid plastomes and additional members of mimosoids and papilionoids. Our comparisons of these new genomes with previously published legume plastomes show that both the caesalpinioids and mimosoids are highly conserved in gene order and content like most other angiosperms. In contrast, papilionoids have smaller genomes with varying degrees of genomic rearrangements and they have experienced multiple, independent gene/intron losses and inversions that limit the phylogenetic utility of these changes.

MATERIALS AND METHODS

Plant material

Sampling included 13 species representing each of the three subfamilies of Fabaceae (Table 1). *Apios americana*, *Caesalpinia coriaria* and *Pachyrhizus erosus* seeds were obtained from Sand Mountain Herbs (http://www.sandmountainherbs.com), eBay and Trade Winds Fruit (http://www.tradewindsfruit.com/), respectively. Seeds of the remaining 10 species were obtained from the USDA-ARS National Plant Germplasm System. Seed germination and plant growth was conducted in the UT-Austin greenhouse and vouchers were deposited in the UT Plant Resources Center (TEX-LL). Newly emerged leaves were collected, flash frozen with liquid nitrogen and stored at -80^{o} C for DNA isolation.

DNA isolation

Isolation of DNA was performed using the method of Doyle and Doyle (1987) with modifications. Cetyl trimethylammonium bromide (CTAB) buffer was augmented with 3% PVP and 3% betamercaptoethanol (Sigma, St. Louis MO). Organic phase separation was repeated until the aqueous fraction was clear. DNA pellets were resuspended in ~200 µL DNase-free water. Following treatment with RNase A (ThermoScientific, Lafayette, CO) samples were again subjected to phase separation with chloroform. DNA was recovered by ethanol precipitation, resuspended in DNase-free water and stored at -20 ^{o}C.

Genome sequencing, assembly and annotation

DNAs were sheared to yield ~800 base pair fragments for paired end library construction according to the NEBNext Ultra DNA Library Prep Kit for Illumina (New England BioLabs, Ipswich, MA). Library preparation and DNA sequencing were carried out at the UT-Austin Genome Sequencing and Analysis Facility on the Illumina HiSeq 2000 platform (Illumina, San Diego, CA). Reads were quality-filtered using FastxToolkit (hannonlab.cshl.edu/fastx_toolkit/). The quality-filtered reads were assembled using Velvet version 1.2.08 (Zerbino and Birney, 2008) at the Texas Advanced Computing Center. Multiple assemblies were performed with modified parameters (i.e., varying kmer, scaffolding on or off and manual input of insert size versus default estimation). Contigs from all assemblies were imported into Geneious version 6.1.3 (Biomatters Ltd., http://www.geneious.com/). A plastid gene database comprising closely related legume sequences was employed to identify plastid contigs from each assembly. Plastid contigs from multiple assemblies for each species were evaluated to resolve IR boundaries in addition to ambiguities or differences among contigs. Illumina reads were mapped to contigs using Bowtie2 (Langmead and Salzberg, 2012) to address potential misassembly issues. It should be noted that *Trifolium pratense* was assembled into 6 contigs and genes were extracted from those contigs as large amounts of repetitive DNA hindered full assembly.

Gene annotation of plastomes was performed in DOGMA (Wyman *et al.*, 2004). Verification of protein coding genes was performed in Geneious version 6.1.3 (Biomatters Ltd., http://www.geneious.com/) using the plastid gene database described above and tRNAs were verified using tRNAscan (Schattner *et al.* 2005).

Whole genome comparisons

Publicly available plastome sequences were downloaded from NCBI (http://www.ncbi.nlm.nih.gov/) (Supplemental Table 1). Whole genome alignments were performed to identify inversions using progressiveMauve version 2.3.1 (Darling *et al*., 2010) in Geneious. MultiPipMaker (Schwartz *et al*., 2000) was used to visualize variable regions among genomes and paired t-tests were performed to determine statistically different genome sizes between subfamilies.

Gene sequence alignment and phylogenetic analyses

Seventy-one genes (Supplemental Table 2) present in all 36 species (32 legumes and 4 outgroup taxa) were extracted and aligned using the translation align tool in Geneious with default MUSCLE (Edgar 2004) settings. Alignments were manually adjusted where necessary and deposited in the Dryad Digital Repository (http://dx.doi.org/10.5061/dryad.n85m5). The 71 gene alignments were concatenated into a single alignment and maximum likelihood trees were generated through RaxML Blackbox (Stamatakis *et al*. 2008) using the gamma model of rate heterogeneity and maximum likelihood search settings. The best scoring tree was imported into FigTree version 1.4.2 (http://tree.bio.ed.ac.uk/software/figtree/).

RESULTS

Genome size, gene content and organization of 13 new legume plastomes

Thirteen new Fabaceae plastome sequences were completed: five from subfamily Caesalpinioideae, one from Mimosoideae and seven from Papilionoideae (species in bold in Table 2). The new plastomes range in size from 151,866 bp to 163,042 bp with the LSC, SSC and IR ranging in size from approximately 82-92 kb, 17-23 kb and 25-26 kb, respectively.

The majority of the new plastomes share the same 77 protein-coding genes, 30 tRNAs and 4 rRNAs. However, *rps16* is either a pseudogene or absent in *Arachis hypogaea*, *Apios americana*, *Vigna unguiculata* and *Robinia pseudoacacia*. In addition, *rpl33* is absent in *V. unguiculata*.

While the caesalpinioid and mimosoid plastomes share the ancestral genome organization for angiosperms (Ruhlman and Jansen 2014), the new papilionoid plastomes all share a 50 kb inversion with endpoints near *rbcL* and *rps16*. In addition, *Lupinus albus* and *R. pseudoacacia* share an inversion that is nested within the 50 kb inversion (Figure 1a). The endpoints of the nested inversion lie in a repeat that occurs in opposite orientation within *trnS*-GGA and *trnS*-GCU (Figure 1b). This inversion is approximately 36 kb in *L. albus* and 39 kb in *R. pseudoacacia*.

The increased length of intergenic spacer regions in *R. pseudoacacia* accounts for the 3 kb difference in size.

Phylogenetic analysis

A maximum likelihood tree was constructed using 71 genes common to all 36 species (Supplemental Table 2). The alignment included 71,436 bp, yielding a phylogeny with an optimal likelihood score of ln(L) = - 470265.78 (Figure 2). Bootstrap values were 100% for all nodes.

The phylogeny is congruent with published gene based phylogenies (Wojciechowski et al. 2004; Bruneau et al. 2008; LPWG, 2013;) and shows caesalpinioids as a basal, paraphyletic grade with mimosoids and papilionoids forming monophyletic groups nested within the caesalpinoids.

Phylogenetic distribution of genome size, gene/intron losses and inversions across legumes

The 71-gene phylogeny (Figure 2) was used to examine the distribution of gene and intron losses and gene order changes across legumes. All legume plastomes included in this study have lost *rpl22*, whereas *rps16* has experienced five independent losses. The *rpl33* gene loss is restricted to a single papilionoid clade that includes two genera, *Phaseolus* and *Vigna*. The remaining gene losses (e.g., *accD*, *rpl23*, *psaI* and *ycf4*) are restricted to clades within the IRLC. Three different intron losses (e.g., clpP intron 1, clpP intron 2 and rps12 intron 1) are also only found among the IRLC taxa.

The 50 kb inversion present in the new papilionoid plastomes is in all other IR-containing papilionoids included in this study. The inversion in *L. albus* and *R. pseudoacacia* that is nested within the 50 kb inversion is found in two distinct lineages (Figure 2). The 36 kb and 39 kb inversions in *L. albus* and *R. pseudoacacia*, respectively, have the same gene content and have endpoints that lie within a 29 bp region of sequence present in both in *trnS*-GGA and *trnS*-GCU (Figure 1a,b). All legumes sampled in this study have *trnS*-GCU and *trnS*-GGA genes that contain the identical 29 bp of sequence at the 3' end with the exception of only a few nucleotide differences in *A. hypogaea*, *V. unguiculata* and *P. vulgaris* (Figure 1b). Inversion events within the IRLC were not included in this analysis; see Sabir et al (2014) for a summary of these events.

Genome size variation among IR-containing legumes

The caesalpinioids and mimosoids have plastomes that average approximately 160 kb in size, whereas papilionoid plastomes that contain an IR have an average size of approximately 153 kb (Figure 3). There is a statistically significant ($p \leq 0.01$) reduction in plastome size in IR-containing papilionoids compared to caesalpinioids and mimosoids. The significant decrease in size of papilionoid plastomes is due largely to deletions in intergenic spacer (IGS) regions. These deletions are more prevalent in the LSC compared to the SSC and IR (Figure 3). Whole genome alignment using MultiPipMaker identified nine hotspots (A-I in Figure 4) within papilionoid legume plastomes where most deletions are located.

These IGS deletions account for the significant reduction in genome size of papilionoids compared to caesalpinioids and mimosoids.

Discussion

Genomic rearrangements across legumes

The basal lineages of legumes have plastomes with the same gene order and content as the ancestral angiosperm genome (Ruhlman & Jansen 2014) with the exception of the loss of *rpl22*. It was previously shown that *rpl22* is missing in legumes (Gantt et al. 1991; Doyle et al. 1995), and that this gene was transferred to the nucleus (Gantt et al. 1991). In contrast, the papilionoids display numerous cases of gene and intron loss and inversions with the most extensive events occurring in the IRLC (Figure 2). The loss of *rpl33* is limited to a clade within the papilionoids comprising *V. unguiculata* and *P. vulgaris* suggesting a single loss of this gene among legumes. Conversely, in addition to its absence in four of the new genomes sequenced here, *rps16* is also absent in *Phaseolus vulgaris* (Guo et al. 2007) and all species examined to date within the IRLC (Cai et al. 2008; Jansen et al. 2008; Magee et al. 2010; Sabir et al. 2014). The phylogenetic distribution indicated five independent losses of *rps16* across legumes (Figure 2). This corroborates previous evidence of multiple, independent losses of *rps16* across legumes based on Southern hybridization or PCR screening (Doyle et al. 1995). The loss of the 3' intron of *clpP* and of *rps12* is common to all IRLC species (Jansen et al. 2008) while the second *clpP* intron has been lost only from *Glycyrrhiza glabra* (Sabir et al. 2014). All *Trifolium* species examined to date except *T. lupinaster* have lost *accD* (Cai et al. 2008; Sabir et al. 2014; Sveinsson and Cronk 2014). While both *Pisum sativum* and *Lathyrus sativus* have lost *rpl23*, *ycf4* is absent from only *P. sativum* and *psaI* is absent only in *L. sativus* (Magee et al. 2010). With our expanded sampling of legumes, including taxa from the two previously unsampled subfamilies, caesalpinioids and mimosoids, the remaining discussion will emphasize three areas of novelty: genome size differences, independent losses of *rps16* and the 36 kb parallel inversion.

Genome size difference

Compared to mimosoids and caesalpinioids, papilionoids that have an IR show a significant reduction in plastome size (Figure 3). Variation in plastome size in seed plants is typically attributed to IR expansion/contraction or loss, gene duplication or gene loss (Wicke et al. 2011; Jansen & Ruhlman 2012) or increased repetitive DNA content in intergenic regions (Blazier et al. 2011; Green 2011; Sabir et al. 2014). However, size reduction in IR-containing papilionoids is caused primarily by deletions within intergenic spacers in nine different regions, six of which are in the LSC (Figure 4). A similar phenomenon has also been reported in the cupressophytes (Wu & Chaw 2014) and gnetophytes (McCoy et al 2008; Wu et al 2009), which have species with reduced or missing IRs. In cupressophytes, the reduced sizes were attributed to intergenic deletions (Wu & Chaw 2014) whereas in gnetophytes the reductions were found in intronic regions as well as inter-operon, as opposed to intra-operon, spacers (Wu et al. 2009).

Downsizing of intergenic regions has been proposed to have selective advantages in plastids and parasitic bacteria as a means to streamline replication and minimize resources required for growth (Dufresne et al. 2005; McCoy et al. 2008; Wolf & Koonin 2013; Wu & Chaw 2014). Additionally, Lynch et al. (2006) suggested a negative correlation between genome size and mutation rates at silent sites, which was supported by Wu & Chaw (2014) in the cupressophytes. Calculating nucleotide substitution rates across legumes would be valuable for testing the generality of this correlation in the future.

Independent losses of *rps16*

Multiple losses of plastid encoded *rps16* have been documented across seed plants (reviewed in Jansen and Ruhlman 2012). The *rps16* intron has been lost in the plastomes of *Penthorum chinense*, *Trachelium caeruleum* and *Pelargonium x hortorum* (Chumley et al. 2006; Haberle et al. 2008; Dong et al. 2013) and *rps16* is completely missing in a wide diversity of taxa ranging from ferns to angiosperms (Gao et al., 2013; Magee et al., 2010; Roy et al., 2010; Sabir et al., 2014; Saski et al., 2005; Tangphatsornruang et al., 2010; Tsudzuki et al., 1992; Ueda et al. 2008). In species of *Arabidopsis* and other members of the Brassicaceae, *rps16* is in a state flux with fully functional forms in some species and pseudogenes in others (Roy et al. 2010). While we did not test the functionality of *rps16*, the situation in Fabaceae and Brassicaceae is quite similar, with some copies apparently being fully functional while others exist as pseudogenes or have been lost entirely from the plastome.

Among legumes with published plastid genome sequences, the loss of *rps16* has been previously reported for members of the IRLC in addition to *P. vulgaris* and *V. radiata* (Guo et al. 2007; Magee et al. 2010; Sabir et al. 2014; Saski et al. 2005; Tangphatsornruang et al. 2010). The detection of four additional losses of *rps16* across legumes (Figure 2) indicates that the gene has been lost independently at least five times.

Gene loss in plastomes is often associated with a functional gene transfer to the nucleus, substitution by a nuclear encoded mitochondrial targeted gene product or substitution by another nuclear encoded protein (Bock & Timmis 2008; Jansen & Ruhlman 2012). Ueda et al. (2008) surveyed *rps16* and found that the loss of the plastid encoded *rps16* was mediated by the substitution of the nuclear encoded mitochondrial-targeted *rps16* in *Populus alba* and the IRLC legume *Medicago truncatula*. Evidence of this substitution was also found in transcriptome data of two other members of the IRLC, *Trifolium repens* and *T. pratense,* where each species was missing the plastid copy of *rps16* and had two nuclear copies of *rps16* (Sabir et al. 2014). It would be worth exploring whether nuclear copies of *rps16* are present in all legume species regardless of the status of the plastid encoded gene*.* The presence of a plastid targeted Rps16 in species harboring an intact plastid gene would suggest an intermediate state in which two discrete Rps16 proteins would be present in plastids, similar to the situation described by Ueda et al. (2008) for *A. thaliana* and *Oryza sativa*. Such redundancy could permit the eventual pseudogenization of the plastid copy. In addition, Magee et al. (2010) found that *rps16* and nearby genes including *accD*, *psaI* and *ycf4* are located in a hypermutable region with a mutation rate that is higher than in the nucleus. Brandvain & Wade (2009) found a positive correlation between mutation rates and the number of transfers that occurred from the mitochondria to the nucleus. A hypermutable region in various legumes that spans *rps16*, *accD*, *psaI* and *ycf4* could be promoting gene losses via any of the mechanisms mentioned above.

Parallel inversion

Martin et al. (2014) originally described the 36 kb inversion embedded within the 50 kb inversion common to most papilionoid legumes in *Lupinus luteus* and we have identified the same inversion in *L. albus*, consistent with their suggestion that the inversion is present in core genistoid legumes. The inversion is likely caused by a 29 bp repeat within *trnS*-GCU and *trnS*-GGA that occurs approximately 36 kb apart in opposite orientation in the plastid genome (Figure 1). A novel finding of this study is that the same inversion is also found in *Robinia pseudoacacia,* a distantly related papilionoid legume. The same repeat presumed responsible for this inversion occurs in the *trnS*-GCU and *trnS*-GGA genes of all the legumes included here (Figure 1b). Thus, this repeat has the potential to facilitate flip flop recombination in the other species, much like the IR mediated recombination described by Palmer (1983) and others (Kim & Lee 2005; Jansen & Ruhlman 2012; Martin et al. 2014). As long as such conformational changes do not inhibit proper gene function, the plastomes may contain isomers with respect to this region much like the IR. For example, Gurdon & Maliga (2014) found two stable plastid configurations differing by a 45 kb inversion initiated by a run of T's nested in an imperfect repeat within *Medicago truncatula*. Guo et al. (2014) documented genomic isoforms involving a 36 kb inversion in the Cupressophytes flanked by an approximately 250 bp inverted repeat suggesting varying configurations may not be as rare as previously thought. Parallel inversions utilizing the same endpoints in distantly related taxa are extremely rare, however. Aside from our case within the legumes a similar occurrence of a parallel inversion was reported in *Clematis* and *Anemone* species (Hoot & Palmer 1994). However, this case of a parallel inversion was detected by Southern hybridization and whether the endpoints or the cause of the inversion was the same in both species was not determined. The fact that all studied legumes have the same 29 bp repeat with potential to initiate inversions in some but not all taxa is novel.

Rare genomic changes, especially inversions, have been proposed to be powerful phylogenetic characters that have little or no homoplasy (Raubeson & Jansen 2005). However, the presence of identical inversions in two distantly related genera within papilionoids, *Lupinus* and *Robinia,* as well as within other groups suggests that caution should be utilized when using inversions as phylogenetic markers.

Table 2.1. Sampling of new legume plastid genomes.

Taxa are ordered as they appear in the phylogeny in Figure 2. Dashes in the USDA ID # column indicate the species was obtained from other sources (see Methods). Accession # refers to GenBank accession numbers. Vouchers are deposited in TEX-LL.

Species	Subfamily	USDA ID No.	Accession #	Voucher ID
Caesalpinia coriaria	Caesalpinioideae	-	KJ468095	L004
Ceratonia siliqua	Caesalpinioideae	00-0031	KJ468096	L005
Cercis canadensis	Caesalpinioideae	91-0010	KF856619	L006
Haematoxylum brasiletto	Caesalpinioideae	89-0061D	KJ468097	L009
Tamarindus indica	Caesalpinioideae	90-0361	KJ468103	L017
Prosopis gladulosa	Mimosoideae	90-0502	KJ468101	L015
Apios americana	Papilionoideae	-	KF856618	L002
Arachis hypogaea	Papilionoideae	PI 536065	KJ468094	L003
Indigofera tinctoria	Papilionoideae	PI 300006	KJ468098	L010
Lupinus albus	Papilionoideae	W6 39803	KJ468099	L012
Pachyrhizus erosus	Papilionoideae	-	KJ468100	L014
Robinia pseudoacacia	Papilionoideae	PI 502585	KJ468102	L016
Vigna unguiculata	Papilionoideae	PI 313545	KJ468104	L021

Table 2.2. Features of the 13 new plastome sequences

New plastome sequences (in bold) shown in the order that they appear in the phylogeny (Figure 2). Abbreviation are: C – Caesalpinioideae, M – Mimosoideae, P – Papilionoideae, LSC – large single copy, SSC – small single copy, IR – inverted repeat.

Subfamily	Species	Genome size	LSC	SSC	IR	Number genes	Protein -coding genes	tRNA genes	rRNA genes	Genes with introns	GC %	Protein -coding %
C	***Cercis canadensis***	158,995	88,118	19,621	25,628	111	77	30	4	18	36.2	57.4
C	***Tamarindus indica***	159,551	87,967	22,800	24,392	111	77	30	4	18	36.2	56.9
C	***Ceratonia siliqua***	156,367	85,801	18,504	26,031	111	77	30	4	18	36.7	57.6
C	***Haematoxylum brasiletto***	157,728	87,465	18,193	26,035	111	77	30	4	18	36.7	57.3
C	***Caesalpinia coriaria***	158,045	87,589	18,160	26,148	111	77	30	4	18	36.5	57.5
M	***Prosopis glandulosa***	163,042	92,324	18,904	25,907	111	77	30	4	18	35.9	55.7
M	*Acacia ligulata*	158,724	88,577	18,299	25,924	111	77	30	4	18	36.2	50.6
M	*Leucaena trichandra*	164,692	93,690	18,890	26,056	111	77	30	4	18	35.6	47.6
P	***Arachis hypogaea***	156,395	85,951	19,868	25,288	110	76	30	4	17	36.4	55.9
P	***Lupinus albus***	154,140	82,266	20,070	25,902	111	77	30	4	18	36.5	59.8
P	***Indigofera tinctoria***	158,367	88,852	18,799	25,358	111	77	30	4	18	35.8	57.0
P	*Millettia pinnata*	152,968	83,401	19,051	25,258	111	77	30	4	18	34.8	56.7
P	***Apios americana***	152,828	83,092	18,272	25,732	110	76	30	4	17	35.6	56.6
P	***Pachyrhizus erosus***	151,947	83,605	18,912	24,715	111	77	30	4	18	35.3	59.3
P	*Glycine max*	152,218	83,175	17,895	25,574	111	77	30	4	18	35.4	60.0
P	***Vigna unguiculata***	151,866	81,587	17,427	26,426	109	75	30	4	17	35.2	59.2

Table 2.2 (continued)

P	*Phaseolus vulgaris*	150,284	79,825	17,609	26,425	109	75	30	4	17	35.4	57.5
P	***Robinia pseudoacacia***	154,835	86,172	19,005	24,829	110	76	30	4	17	35.9	56.3
P	*Lotus japonicus*	150,519	81,936	18271	25,156	111	77	30	4	18	36.0	57.0

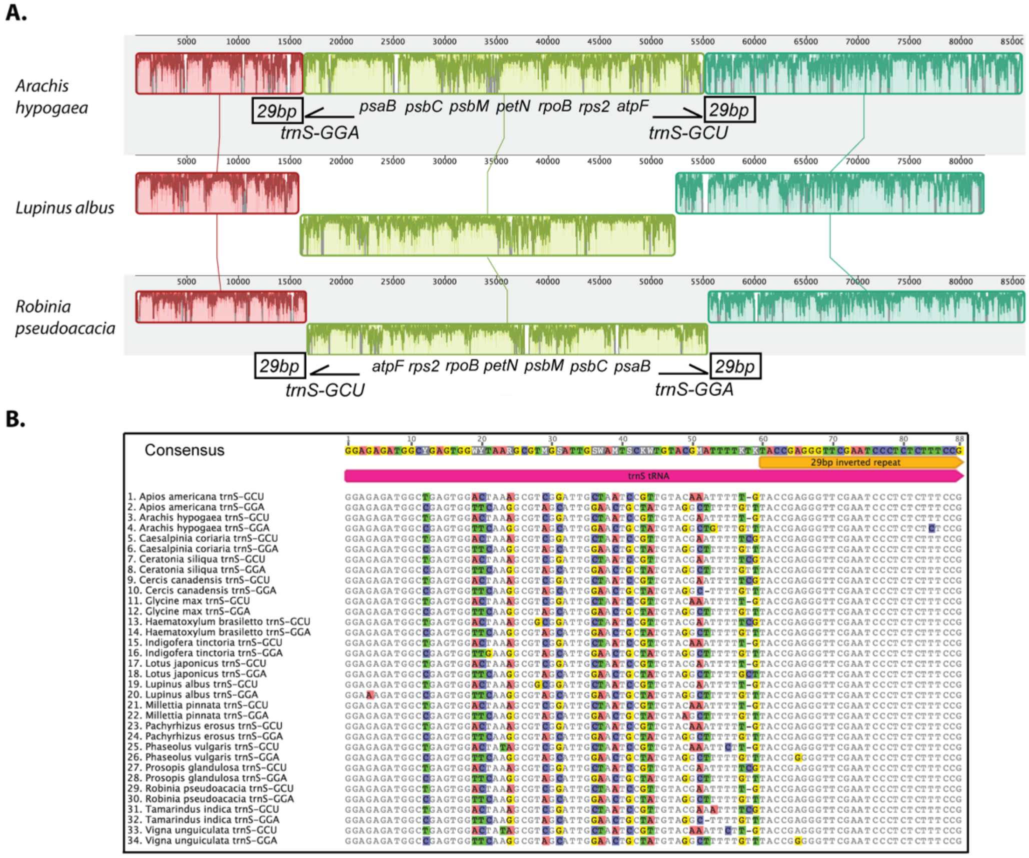

A.
Arachis hypogaea
29bp
psaB psbC psbM petN rpoB rps2 atpF
29bp
trnS-GGA
trnS-GCU
Lupinus albus
Robinia pseudoacacia
29bp
atpF rps2 rpoB petN psbM psbC psaB
29bp
trnS-GCU
trnS-GGA
B.
Consensus
29bp inverted repeat
trnS tRNA
1. Apios americana trnS-GCU
2. Apios americana trnS-GGA
3. Arachis hypogaea trnS-GCU
4. Arachis hypogaea trnS-GGA
5. Caesalpinia coriaria trnS-GCU
6. Caesalpinia coriaria trnS-GGA
7. Ceratonia siliqua trnS-GCU
8. Ceratonia siliqua trnS-GGA
9. Cercis canadensis trnS-GCU
10. Cercis canadensis trnS-GGA
11. Glycine max trnS-GCU
12. Glycine max trnS-GGA
13. Haematoxylum brasiletto trnS-GCU
14. Haematoxylum brasiletto trnS-GGA
15. Indigofera tinctoria trnS-GCU
16. Indigofera tinctoria trnS-GGA
17. Lotus japonicus trnS-GCU
18. Lotus japonicus trnS-GGA
19. Lupinus albus trnS-GCU
20. Lupinus albus trnS-GGA
21. Millettia pinnata trnS-GCU
22. Millettia pinnata trnS-GGA
23. Pachyrhizus erosus trnS-GCU
24. Pachyrhizus erosus trnS-GGA
25. Phaseolus vulgaris trnS-GCU
26. Phaseolus vulgaris trnS-GGA
27. Prosopis glandulosa trnS-GCU
28. Prosopis glandulosa trnS-GGA
29. Robinia pseudoacacia trnS-GCU
30. Robinia pseudoacacia trnS-GGA
31. Tamarindus indica trnS-GCU
32. Tamarindus indica trnS-GGA
33. Vigna unguiculata trnS-GCU
34. Vigna unguiculata trnS-GGA

Figure 2.1. Mauve alignment showing a shared inversion in two papilionoids. **A**, Diagram of the large single copy (LSC) region aligned in Geneious using progressiveMauve with *Arachis hypogaea* as the reference. Syntenic regions are indicated by colored, locally collinear blocks (LCBs). Histograms inside each block represent pairwise nucleotide sequence identity. Inversions are shown as blocks flipped across the plane. The location of *trnS*-GCU and *trnS*-GGA is annotated and a subset of protein coding genes in the inversion region is also indicated. **B**, Alignment of *trnS*-GCU and *trnS*-GGA in sampled legumes. Alignment of *trnS*-GCU and *trnS*-GGA genes (pink arrow) from 17 legume species was generated in

Geneious using MUSCLE. The 29 bp repeat is indicated by the orange arrow. Colored nucleotides are those that differ between the two *trnS* genes and across species.

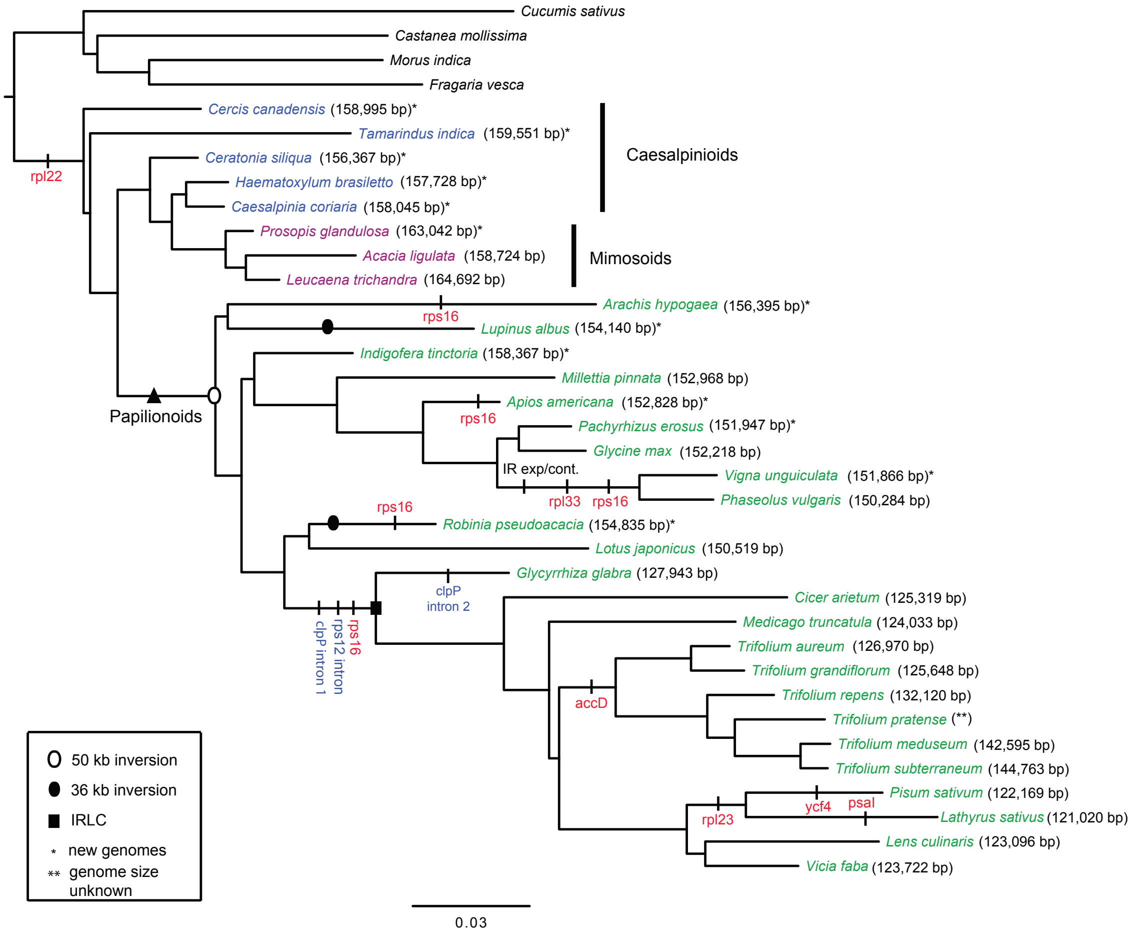

Cucumis sativus
Castanea mollissima
Morus indica
Fragaria vesca
rpl22
Cercis canadensis (158,995 bp)*
Tamarindus indica (159,551 bp)*
Ceratonia siliqua (156,367 bp)*
Haematoxylum brasiletto (157,728 bp)*
Caesalpinia coriaria (158,045 bp)*
Caesalpinioids
Prosopis glandulosa (163,042 bp)*
Acacia ligulata (158,724 bp)
Leucaena trichandra (164,692 bp)
Mimosoids
Arachis hypogaea (156,395 bp)*
rps16
Lupinus albus (154,140 bp)*
Indigofera tinctoria (158,367 bp)*
Millettia pinnata (152,968 bp)
Apios americana (152,828 bp)*
rps16
Papilionoids
Pachyrhizus erosus (151,947 bp)*
Glycine max (152,218 bp)
IR exp/cont.
Vigna unguiculata (151,866 bp)*
Phaseolus vulgaris (150,284 bp)
rpl33
rps16
rps16
Robinia pseudoacacia (154,835 bp)*
Lotus japonicus (150,519 bp)
Glycyrrhiza glabra (127,943 bp)
clpP
intron 2
clpP intron 1
rps12 intron
rps16
Cicer arietum (125,319 bp)
Medicago truncatula (124,033 bp)
Trifolium aureum (126,970 bp)
Trifolium grandiflorum (125,648 bp)
accD
Trifolium repens (132,120 bp)
Trifolium pratense (**)
Trifolium meduseum (142,595 bp)
Trifolium subterraneum (144,763 bp)
Pisum sativum (122,169 bp)
ycf4
psaI
rpl23
Lathyrus sativus (121,020 bp)
Lens culinaris (123,096 bp)
Vicia faba (123,722 bp)
○ 50 kb inversion
● 36 kb inversion
■ IRLC
* new genomes
** genome size unknown
0.03

Figure 2.2. Phylogenetic relationships among legumes with completed plastomes. The maximum likelihood tree was generated using RaXML. Bootstrap values were 100% for all branches. Species names are colored to indicate three subfamilies of Fabaceae: Caesalpinioideae (blue), Mimosoideae (purple) and Papilionoideae (green). Gene and intron losses are indicated on the branches and plastome size is included in parentheses next to each taxon in the tree. Scale bar indicates mean number of nucleotide substitutions per site along each branch. IR exp/cont.

represents inverted repeat expansion/contraction. The 50 kb inversion indicated on the phylogeny is present in all papilionoid taxa included in this study but a previous investigation showed that this inversion is not present in all members of this subfamily (Doyle et al. 1996; Wojciechowski et al. 2004).

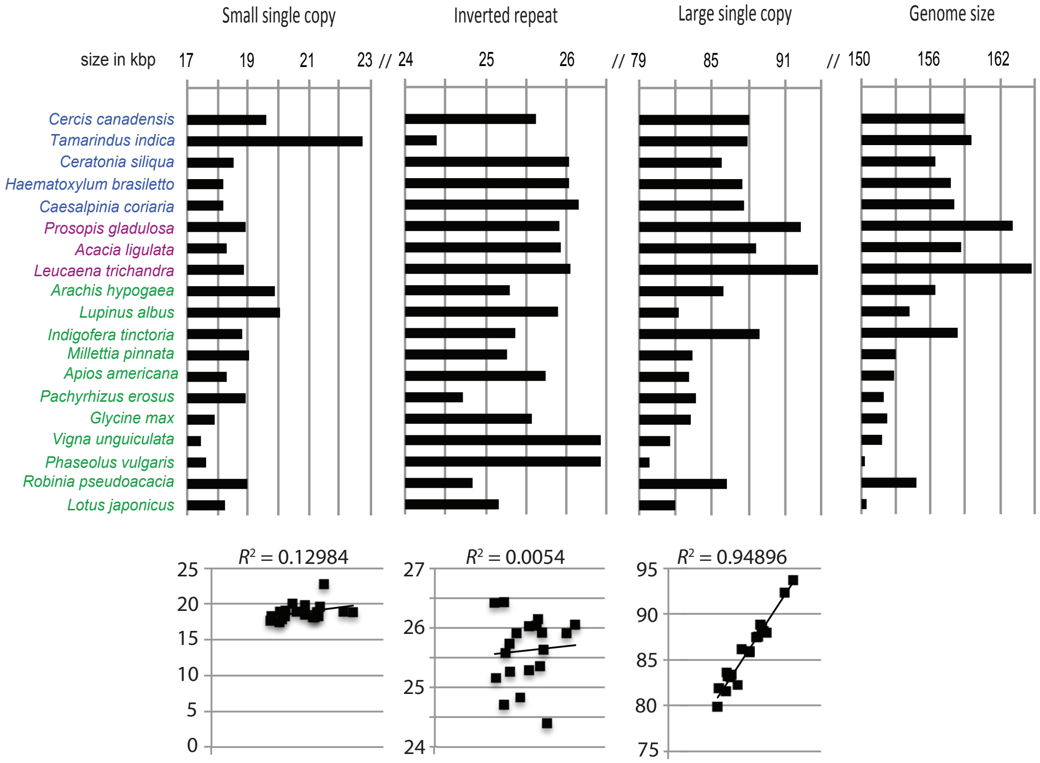


Figure 2.3. Size variation of legume plastomes. Histograms showing the sizes of distinct regions of legume plastomes and the entire genome. Species names are colored to indicate three subfamilies of Fabaceae: Caesalpinioideae (blue), Mimosoideae (purple) and Papilionoideae (green). Below each histogram is a scatter plot of the respective component size plotted against total genome size.

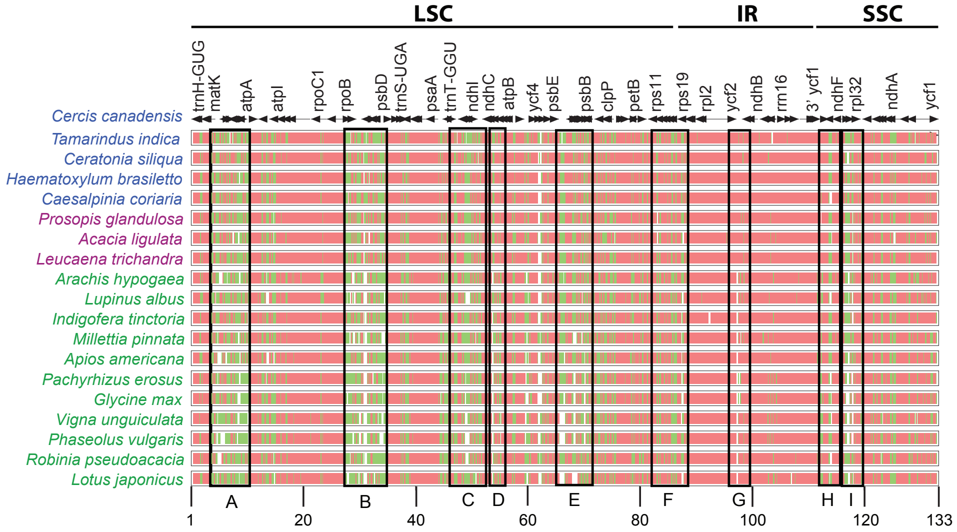


Figure 2.4. MultiPipMaker similarity plot of whole plastomes. *Cercis canadensis* was used as the reference. One copy of the inverted repeat (IR) was excluded. The regions outlined in black boxes correspond to deletion hotspots in the following intergenic regions: A – *matK*–*atpA*, B – *rpoB-psbD*, C – *trnT-ndhI*, D – *ndhC-atpB*, E – *psbE*-psbB, F – *rps11-rps19*, G – *ycf2-trnL*-CAA, H – 3' *ycf1-ndhF*, I – *ndhF-trnL*-UAG. Species names are colored to indicate three subfamilies of Fabaceae: Caesalpinioideae (blue), Mimosoideae (purple) and Papilionoideae (green). Color in similarity plot indicates nucleotide sequence percent identity: red (75-100), green (50-75) and white (<50). Numbers along the bottom indicate genome coordinates in kb. LSC = large single copy region, IR = inverted repeat and SSC = small single copy region.

Chapter 3: Plastome-wide nucleotide substitution rates reveal accelerated rates in Papilionoideae and correlations with genome features across legume subfamilies

INTRODUCTION

Angiosperm plastid genomes (plastomes) are characterized by a quadripartite structure that includes two identical copies of an inverted repeat (IR) separated by a large single copy (LSC) region and a small single copy (SSC) region (Ruhlman and Jansen, 2014). The inverted repeat is usually about 25 kilobases (kb) and houses four ribosomal genes (*rrn4.5*, *rrn5*, *rrn16*, *rrn23*), seven tRNAs and seven protein coding genes. Within angiosperms the IR has been lost in certain members of Orobanchaceae (Downie and Palmer, 1992), Geraniaceae (Blazier et al., 2011, 2016a; Guisinger et al., 2011) and Fabaceae (Koller and Delius, 1980; Palmer and Thompson, 1981; Palmer et al., 1987). Within Fabaceae (legumes), loss of the IR occurred once leading to a monophyletic group within the subfamily Papilionoideae termed the "IR lacking clade" (IRLC) (Wolfe, 1988; Lavin et al., 1990; Wojciechowski et al., 2004).

Rates of nucleotide substitution in genes located in the IR have been shown to be at least three times lower compared to single copy (SC) genes (Wolfe et al., 1987). This estimate was confirmed in a recent study that investigated synonymous substitution rates across 52 families of angiosperms, gymnosperms and ferns, which concluded that rates of genes in the IR are about four times lower than SC genes (Zhu et al., 2016). It has been suggested that lower rates in IR genes are caused by gene conversion bias whereby the mutation rate across the genome is the same but duplicated regions such as the IR are resistant to mutational change (Birky and Walsh, 1992). Species lacking one copy of the IR present an opportunity to compare rates between ancestral SC genes with genes formerly in the IR that are now single copy. Perry and Wolfe, (2002) investigated nucleotide substitution rates in SC and IR genes in four legume species with (*Glycine max, Lotus japonicus*) and without (*Medicago truncatula*, *Pisum sativum*) the IR and found that genes formerly in the IR have accelerated rates that are equivalent to rates of SC genes, in agreement with the gene conversion bias hypothesis of Birky and Walsh (1992). Plastome-wide evolutionary rate comparisons using an expanded taxon sampling in a family that includes both IR-containing and IR-lacking species have not yet been performed.

Comparison of nucleotide substitution rates between functional groups of genes also provides insight into plastome evolution. Genes encoding subunits that are important in photosynthetic processes such as ATP synthase (ATP), NAD(P)H dehydrogenase (NDH), cytochrome b6f complex (PET) and photosystems I and II (PSA and PSB) have been shown to have lower rates of nucleotide substitution than other functional groups of genes in grasses (Zhong et al. 2009; Guisinger et al., 2010) and Geraniaceae (Guisinger et al., 2008). Studies have identified a few groups of genes or individual genes that have accelerated rates. Ribosomal protein (RPL and RPS) genes are highly accelerated in Geraniaceae (Guisinger et al., 2011; Weng et al., 2012) and RNA polymerase (RPO) genes have recently been shown to be accelerated in several angiosperm lineages (Blazier et al., 2016). In *Silene* plastid genes with the most accelerated rates are *accD*, *clpP*, *ycf1* and *ycf2* (Sloan et al., 2012). *accD* encodes acetyl-CoA carboxylase, important in fatty acid biosynthesis (Kode et al., 2005) and *clpP* encodes a protein that is part of a multimeric protease (Peltier et al., 2004). Additionally, recent studies in mimosoid legumes found *clpP* to be highly divergent in certain lineages (Dugas et al., 2015; Williams et al., 2015).

In addition to rate variation between genes or functional groups of genes, rate variation in relationship to features of the plastome, such as size and genomic rearrangements (gene order changes, gene/intron loss and indels) can provide insight into forces that shape the plastome. While gene order and content is highly conserved throughout seed plants (Jansen and Ruhlman, 2012), extensive rearrangements have been found in conifers (Hirao et al., 2008; McCoy et al., 2008), Campanulaceae (Cosner et al., 2004; Haberle et al., 2008; Jansen and Ruhlman, 2012; Knox, 2014), Ericaceae (Fajardo et al., 2013; Martınez-Alberola et al., 2013), Geraniaceae (Chumley et al., 2006; Blazier et al., 2011; Guisinger et al., 2011; Weng et al., 2014; Blazier et al., 2016a), Oleaceae (Lee et al., 2007) and Fabaceae (Cai et al., 2008, Sabir et al., 2014, Schwarz et al., 2015; Sveinsson and Cronk, 2014). A positive correlation between genome rearrangement events and nucleotide substitution rates has been noted in several lineages of angiosperms (Jansen et al., 2007). This has been confirmed with more in-depth studies in the mitochondrial and plastid genomes of angiosperm families Caryophyllaceae (Sloan et al., 2012) and

Geraniaceae (Guisinger et al., 2008, 2011; Weng et al., 2014; Grewe et al., 2015) and the gymnosperm *Welwitschia mirabilis* (McCoy et al., 2008). Correlations between genome size and nucleotide substitution rates have also been investigated. In Cupressophytes, a negative correlation was found between genome size and values of *dS* (Wu and Chaw, 2014), whereas in Caryophyllaceae (Sloan et al., 2012) and Geraniaceae (Grewe et al., 2015) mitochondrial genome size was positively correlated with substitution rates. Several studies have suggested that increased rates of nucleotide substitutions and genomic rearrangements may be due to alterations in DNA repair, replication and recombination mechanisms (Guisinger et al., 2008; Weng et al., 2014; Zhang et al. 2016). Additional studies with increased sampling that focus on correlations between genome features and substitution rates are needed.

Rate heterogeneity across lineages has also been explored and rate differences between lineages are often attributed to differences in life history traits. Based on animal studies, the generation time effect hypothesis was posited, which states that nucleotide substitution rates should be negatively correlated with generation time because animals with shorter generation times undergo more germ line cell divisions (Ohta, 1993; Wu and Li, 1987). In plants, large, mostly woody plants have lower absolute growth rates leading to fewer cell divisions per unit time, whereas small, herbaceous plants have high absolute growth rates (Petit and Hampe, 1997). Studies across major monocot lineages have revealed that substitution rates of plastid *rbcL* (Doebley et al., 1990; Gaut et al., 1992; Wilson et al., 1990), nuclear *adh* (Gaut et al., 1996; MacCay et al., 1995) and mitochondrial *atpA* (Eyre-Walker and Gaut, 1997) in grasses are fast compared to palms, which have a longer generation time. Subsequent studies have found evidence supporting the generation time hypothesis in plants where woody plants have slower rates than herbaceous plants (Bousquet et al., 1992; Kay et al., 2006; Laroche et al., 1997; Smith and Donoghue, 2008). Recently, a study utilizing genes from the mitochondria (*atp1*, *matR*, *nad5*, *rps3*), the plastome (*atpB* and *rbcL*) and the nucleus (*xdh*) tested substitution rates against plant height, as a measure of life history, across multiple plant families (Bromham et al., 2015). A clear and consistent pattern emerged that plant families with shorter average height (fast generation time) have faster rates of molecular evolution. To date, however, all studies addressing generation time as it relates to substitution rates have focused on only a few loci or a handful of lineages that are either very closely or very distantly related. Family-wide sampling across multiple loci is necessary in order to elucidate any patterns reflecting effects of generation time on nucleotide substitution rates.

Legumes are an ideal group in which to investigate nucleotide substitution rates both among lineages and across genomes. The family includes three traditionally recognized subfamilies, Caesalpinioideae, Mimosoideae, Papilionoideae, that exhibit a wide variety of habitat and growth habits (LPWG, 2013). Additionally, plastome organization within legumes ranges from ancestral angiosperm gene order in Caesalpinioideae and Mimosoideae to highly rearranged gene orders in the Papilioinoideae IRLC with a gradient of change seen in the IR-containing Papilionoideae (Schwarz et al., 2015). In this study, we address the following questions: 1) Are nucleotide substitution rates higher in IRLC papilionoids? 2) Are rates of nucleotide substitutions in genes formerly in the IR in IRLC taxa accelerated compared to genes retained in the IR in IR-containing taxa? 3) Is there a correlation between rates and genome features such as size and rearrangement events? In order to address these questions, we utilize 71 genes common to 20 legume plastomes recently published by our group (Sabir et al., 2014; Schwarz et al., 2015; chapter 2) and 19 publicly available legume plastomes, representing the largest rate analysis, both in taxonomic coverage across legumes and number of genes included. Our analyses find that substitution rates in genes formerly in the IR in IRLC taxa are accelerated, but not significantly so, than IR genes in IR-containing species. Additionally, rates are accelerated significantly within IRLC papilionoids and there is a significant increase of rates in herbaceous versus woody legumes. Lastly, we show that there is a significant correlation of nucleotide substitution rates and plastome features such as size and rearrangement events (gene order changes and indels).

METHODS

Sampling

In addition to twenty plastid genomes that were recently sequenced, assembled and annotated for this analysis (Sabir et al., 2014 and Schwarz et al., 2015, chapter 2), 19 publicly available legume plastome sequences and four outgroups were downloaded from GenBank (http://www.ncbi.nlm.nih.gov/genbank/) (Table 3.1). With the exception of *Trifolium* species, sampling included only one species representing each genus when there were multiple species available.

Gene sequence alignment and phylogenetic analysis

Seventy-one protein coding genes (Table 3.2) common to all 43 species were extracted and aligned using MAFFT (Katoh and Standley, 2013) translation align in Geneious 7.1.9 (Biomatters, Ltd.). Alignments were manually edited in order to improve alignment quality and also to ensure indels were maintained in groups of three to retain the reading frame. A concatenated alignment of all 71 genes was generated and ambiguous and poorly aligned regions were removed using Gblocks (http://molevol.cmima.csic.es/castresana/Gblocks_server.html). Phylogenetic analysis was performed in RAxML Blackbox (Stamatakis et al. 2008) using the GTR model and 100 bootstrap replicates. Alignments of individual genes were also concatenated into functional groups (Table 3.3) in Geneious 7.1.9 (Biomatters, Ltd.).

Nucleotide substitution rates

Nonsynonymous (*dN*) and synonymous (*dS*) nucleotide substitution rates for each of 71 protein coding genes and eight functional groups were estimated using the codeml program in PAML 4.5 (Yang 2007). Codon frequencies were determined by the F3 x 4 model. Transition/transversion and *dN*/*dS* ratios were estimated with the initial values of 2 and 0.4, respectively. Two analyses were run: 1) runmode = 0, model = 0 in which the phylogeny generated by RAxML was used as a constraint tree and *dN*/*dS* ratios were allowed to vary among branches, and 2) runmode = -2, model = 1, for pairwise rate comparisons between each legume taxon and *Morus indica*, one of the outgroup species.

Detection of rate acceleration

Pairwise *dN* and *dS* values across the genome were compared between four major groups of legumes: caesalpinioids, mimosoids, papilionoids with the IR and IRLC papilionoids. Significant acceleration of substitution rates in genes between the four legume subgroups was tested using the nonparametric, Kruskal-Wallis test. The generated p-values were corrected using the Holm method in the p.adjust function in R version 3.2.4. Line plots were created using ggplot2 in R version 3.2.4, using the stat_summary function to plot the mean value of genes in each of the four groups. In order to test if sampling bias from a large number of papilionoid (30) taxa compared to many fewer mimosoids (4) and caesalpinioids (5) may have skewed results, an additional analysis was performed using only five IR-containing papilionoids (*Arachis hypogaea*, *Indigofera tinctoria*, *Apios americana*, *Phaseolus vulgaris*, *Robinia pseudoacacia*) and five IRLC papilionoids (*Wisteria floribunda*, *Cicer arietinum*, *Trifolium boissieri*, T. *pratense*, *Vicia faba*).

Averaged *dN* and *dS* values for each taxon were utilized in the nonparametric Wilcoxon Rank Sum Test in order to test significance of rate accelerations for two different comparisons: 1) between legumes that have both copies of the IR compared to those lacking one copy of the IR, and 2) between legumes with a herbaceous versus a woody growth habit.

Correlation between substitution rates and genome characteristics

Numbers of indels for each gene was calculated using a custom Python script and indels were summed across all genes resulting in a single indel count number for each taxon. Genomic rearrangements were calculated for each genome compared to an outgroup, *Morus indica*, using Common Interval Rearrangement Explorer (CREx) (Bernt et al., 2007). Genome size included only one copy of the IR in order to be consistent across all legumes.

Average *dN* and *dS* values were calculated for each taxon and correlation against genome size, number of indels and number of rearrangements for each genome was tested. The Pearson correlation test was performed in R version 3.2.4, using the rcorr function in the Hmisc package. The p-values were corrected using the Holm method in the p.adjust function and scatterplots were generated using the ggplot2 package in R version 3.2.4.

Results

Phylogenetic analysis

The maximum likelihood phylogeny was generated using 43 taxa, 39 of which were legumes, and 71 protein-coding genes from the plastome. The maximum likelihood score was - 449433.5397. The phylogeny has strong support for all nodes with the exception of the branch leading to *Glycyrrhiza glabra* and *Wisteria floribunda* (Figure 3.1) and is congruent with previous legume phylogenies (LPWG, 2013; Wojciechowski et al., 1994).

Rates of functional groups

Functional groups consisted of eight groups of genes in addition to eight individual genes that cannot be assigned to any of these groups (Table 3.3). Values of *dS* are higher than *dN* across all functional groups and individual genes (Figures 3.2 - 3.3, Table 3.4). Mean *dN* values range from 0.0014 (PSB) to 0.0465 (*ycf*1). Mean values of *dN* are highest in *clp*P (0.0266) and *ycf*1 (0.0465). Values of *dS* range from 0.0177 (*ycf*2) to 0.0759 (*ycf*1) with the same two genes, *clp*P and *ycf*1, having the highest mean values. Photosynthetic genes (PSA, PSB, PET) and ATP-synthase genes (ATP) have some of the lowest *dN* values while NADH-dehydrogenase genes (NDH) and *ycf*2, in addition to PSA and PSB have the lowest *dS* values.

Rate accelerations in papilionoids

Pairwise comparison across the genome of four legume subgroups

With the exception of two genes, *mat*K and *pet*L, pairwise *dN* values are consistently higher in all protein coding genes of IRLC papilionoids compared to the other three groups (Figure 3.4). Caesalpinioids and mimosoids have the lowest *dN* values across the genome with the exception of *clp*P, which is accelerated in the mimosoids. The values of *dS* show a more complex pattern of variation across the genome (Figures 3.5 - 3.6). A number of genes have a higher *dS* value in IR-containing papilionoid taxa compared to IRLC papilionoids (i.e., *psb*I, *atp*A, *psa*B, *atp*E, *psb*E, *pet*L, *rpl*20, *psb*T, *pet*B, *rps*11, *rpl*3, *rpl*16, *rps*15, *ndh*I, *ndh*E, *psa*C). Similar to the pattern seen in *dN* values, the caesalpinioids and mimosoids consistently have the lowest *dS* values compared to the papilionoid groups. There is a noticeable decrease in *dS* of genes that are contained within the IR region (Figures 3.5 – 3.6). Of the four major groups of legumes, the IRLC papilionoids have the highest *dS* rate across genes that were formerly in the IR region. With the exception of three photosynthetic genes, for both *dN* (*psb*J, *psb*M, *psb*Z) and *dS* (*psb*E, *psb*I, *psb*J) comparisons, there is a statistically significant increase in *dN* and *dS* values between the two papilionoid groups and caesalpinioid and mimosoid groups across the genome (Table 3.5). While the analysis containing just a subset of IR-containing papilionoids and IRLC papilionoids showed the exact same patterns for both *dN* and *dS* (Figures 3.7 – 3.9), the difference in rates between groups was not significant for any of the 71 genes after Holm correction (Table 3.6).

Pairwise comparison of habit and IR presence

There is a significant increase in *dN* and *dS* values (p-values of 1.01E-07 and 7.61E-04, respectively) of herbaceous legumes compared to those of legumes with a woody growth habit. Additionally, those taxa lacking one copy of the IR have significantly higher *dN* and *dS* values (1.16E-08 and 1.34E-02, respectively) compared to rates in taxa that contain both copies of the IR.

Correlations between rates and genome characteristics

There is a negative correlation of -0.371 and -0.431 between genome size and *dN* and *dS* values, respectively (Figures 3.10 - 3.11). However, only the correlation between genome size and *dS* values is significant with a p-value of 2.76E-02. There is a strong and significant positive correlation of 0.910 (p-value, 2.66E-14) and 0.727 (p-value, 1.65E-06) between *dN* and number of indels and rearrangements, respectively (Figures 3.12 – 3.13). While not as strong, *dS* also shows significant positive correlations of 0.561 (p-value, 1.47E-03) and 0.522 (p-value, 3.92E-03) with number of indels and rearrangements, respectively (Figures 3.14 – 3.15).

Lineage-specific accelerated rates

Across branches, *dN* values are generally higher and more variable within papilionoids compared to caesalpinioids and mimosoids (Figures 3.16 – 3.17). The most accelerated *dN* values among genes are in *clp*P, *ycf*1 and three ribosomal protein genes (i.e., *rps*3, *rps*8, *rps*15). Overall, the most accelerated *dN* values are on the branch leading to *Lathyrus sativus* (Figure 3.17).

Values of *dS* show a slightly different pattern across lineages with more accelerated branches in the IR-containing papilionoids (Figures 3.18 – 3.19). The two most accelerated branches of the papilionoids are those leading to *Arachis hypogaea* and *Lotus japonicus*. Within the caesalpinioids *Tamarindus indica* has the most accelerated branch. The most accelerated lineage-specific values of *dS* are similar to those of *dN* where *ycf1* and several ribosomal protein genes tend to be the fastest evolving genes.

Discussion

This represents the most comprehensive study of nucleotide substitutions rates in plastid genes across legumes both in terms of the number of taxa and genes compared. Included in the analyses were a total of 39 legume plastomes. In addition to the broad sampling across all three subfamilies, 71 protein coding genes common to all 39 plastomes were examined. Legumes are an ideal group to address questions of rate heterogeneity in relation to biological features because of the variation in both plastome organization and growth habit. Nucleotide substitution rates were consistently higher in papilionoid taxa compared to caesalpinioid and mimosoid taxa, and rates in IRLC papilionoids were generally higher than in IR-containing papilionoids. Additionally, positive correlations were uncovered between substitution rates and genome rearrangements and number of indels in protein coding genes. A negative correlation between *dS* and genome size was also revealed. The discussion will focus on rates of genes commonly housed in the IR in both IR-containing and IRLC taxa, potential explanations for why papilionoids exhibit faster rates than the other two legume subfamilies and what mechanisms may be responsible for correlations between rates and genomic characteristics such as size, rearrangements and indels.

Accelerated rates in IR genes of IRLC papilionoids

Synonymous rates (*dS*) of genes within the IR are much lower than genes in the LSC and SSC regions in all legume subgroups. This pattern is consistent with a number of studies focusing on angiosperm (Maier et al., 1995; Perry and Wolfe, 2002; Wolfe et al., 1987; Zhu et al, 2016) and gymnosperm (Wu and Chaw, 2015) plastome evolution. It has been proposed that stabilization of the two copies of the IR through copy dependent DNA repair (Perry and Wolfe, 2002; Wolfe et al., 1987) and gene conversion (Birky and Walsh, 1992) is the mechanism for reduced synonymous substitution rates within IR genes. Perry and Wolfe (2002) indicated that genes normally contained within the IR should have a mutation rate equal to the LSC and SSC genes in those species lacking an IR. While *dS* values of genes contained within the IR are lower across all legumes, the accelerated rates of the IR genes within the IRLC papilionoids are still much lower than the rest of the genome and not equal to rates of the other single copy genes. The larger taxon sampling in the present study could account for differences in general trends or patterns compared to previous studies.

Earlier studies of rates of nucleotide substitutions in legumes have emphasized *dS* and not *dN*. In this study, *dN* values across the genome were significantly higher in the papilionoids with the exception of three photosynthetic genes (*psb*J, *psb*M, *psb*Z). However, *dN* of all legumes was equally variable across the genome and did not show any pattern that was unique to genes within the IR region (Figure 3.4). This was not unexpected because *dS*, not *dN*, is representative of the underlying mutation rate given the neutral theory of molecular evolution (Kimura, 1984) so general trends in rate heterogeneity across the genome structure would be likely be more pronounced in *dS* values.

Accelerated rates in papilionoid taxa

Both *dN* and *dS* are significantly accelerated in papilionoids compared to caesalpinioids and mimosoids (Figures 3.4 – 3.6). This is consistent with other recent studies that have explored legume plastome evolution (Dugas et al. 2015; Williams et al., 2015). Nucleotide substitution rate heterogeneity between taxonomic groups has long been studied in animals (Britten, 1986; Martin and Palumbi, 1993; Mooers and Harvey, 1994; Ohta, 1993; Wu and Li, 1985) and plants (Barraclough et al., 1996; Bousquet et al., 1992; Gaut et al., 1992; Gaut et al., 1996; Smith and Donoghue, 2008). A hypothesis commonly invoked to explain rate heterogeneity between taxonomic groups is the generation time. In plants, the generation time hypothesis has been largely supported in studies comparing rates of herbaceous, short-lived plants to woody, long-lived plants (Kay et al., 2006; Laroche et al., 2008; Smith and Donoghue, 2008). Our study also supports the generation time hypothesis as papilionoid legumes are largely herbaceous while mimosoids and caesalpinioids are mostly woody. However, the validity of the generation time hypothesis is debated for two main reasons: 1) a mechanism behind generation time influencing substitution rates is unclear due to the fact that plants don't sequester their germ line cells as animals do therefore somatic mutations can be passed down, and 2) many studies addressing how rates may be influenced by generation time to-date have used either very closely related taxa, very divergent taxa or very few loci (Smith and Donoghue, 2008; Whittle and Johnston, 2003).

Recently Bromham et al. (2015) investigated, among other things, correlation between *dS* and plant height in sequences from the plastid, mitochondrion and nucleus. They found a consistently negative correlation between *dS* and plant height suggesting that taller plants, which tend to be woody, have lower rates of synonymous substitution than shorter plants. Taller plants have more cell divisions between the seed and the apical meristem, and therefore more opportunities for mutation (Bobiwash et al., 2013). A way to avoid this is by reducing the error rate per replication to reduce mutation rates, which would be reflected in values of *dS* (Bromham et al., 2015).

Studies have also shown a positive correlation between *dS* and species diversification in angiosperms (Barraclough et al., 1996; Bousquet et al., 1992; Bromham et al., 2015). In the context of legumes this correlation is also supported by the data as the papilionoids are much more species rich and diverse than either the caesalpinioids or the mimosoids.

Correlations between substitution rates and genome complexity

Size

The causes of plastome size variation within the legumes are distinct in different lineages. Expanded sizes within mimosoids are due to increased regions of tandem repeats (Dugas et al., 2015). In papilionoids containing both copies of the IR, downsizing from the ancestral state is due to deletions within intergenic hotspots, especially in the LSC (Schwarz et al., 2015). Lastly, the most drastic size reductions come from the complete loss of the IR in the IRLC (Palmer and Thompson, 1981; Lavin, 1990; Liston et al., 1995). Previous studies have shown negative (Lynch et al., 2006; Wu & Chaw, 2014) and positive (Grewe et al., 2015) correlations between nucleotide substitution rates and genome size.

Both *dN* and *dS* values of legumes as a whole reveal a negative correlation with genome size (Figures 3.4 – 3.5). However, the correlation between *dN* and genome size is not significant, whereas the correlation between *dS* and size is significant. Lynch et al. (2006) suggested that organellar genomes are shaped by mutational burden, in which case a negative correlation between genome size and mutation rate at silent sites would be present. This is the case in cupressophyte plastomes that vary in size due to intergenic downsizing (Wu and Chaw, 2014).

Alternatively, Grewe et al., (2015) have shown increased substitution rates that are correlated with increased mitochondrial genome size and decreased complexity as measured by gene and intron loss. Within the papilionoids, *Trifolium meduseum* and *T. subterraneum* have the largest genome sizes and some of the highest *dS* values. In addition, these two genomes contain the most rearrangements, repetitive regions and a large number of gene and intron losses compared to other papilionoid taxa. It may be that there are confounding processes shaping the plastome evolution in the papilionoids.

Rearrangements and indels

Within legumes both *dN* and *dS* are correlated with number of indels and rearrangements (Figures 3.8 – 3.9) but correlations of both variables with *dN* are much stronger. This is congruent with previous studies that have found positive correlations between *dN* and increased rearrangements (Guisinger et al. 2008; Weng et al. 2014). Jansen et al. (2007) also identified a positive correlation between branch lengths and gene/intron losses, indels and rearrangements. Weng et al. (2014) found a correlation in Geraniaceae between rearrangements and *dN*. In legumes a correlation between both *dN* and *dS* with respect to rearrangements and indels is present. A possible explanation for a correlation in both *dN* and *dS* and genomic rearrangements is that a single factor is influencing both substitution rates and genomic rearrangements. It has been suggested that a faulty DNA repair/recombination/replication mechanism may be responsible for both highly rearranged plastomes and increased *dN* values in Campanulaceae (Barnard-Kubow et al., 2014) and Geraniacaeae (Guisinger et al., 2008, 2011; Zhang et al., 2016).

Conclusion

Legumes provide a unique opportunity to explore plastome-wide nucleotide substitution rate heterogeneity across three subfamilies that vary in many characteristics including plastome size, number of rearrangement events, presence of an IR and growth habit. Here we find accelerated rates in papilionoid taxa compared to mimosoid and caesalpinioid taxa. Acceleration in the papilionoid lineage may be due to the fact that most species are herbaceous whereas caesalpinioids and mimosoids are largely woody. However, more work is needed to determine whether growth habit and or other features such as species richness or population size play a larger role in rate heterogeneity. Correlations were also revealed between *dN* and *dS* values and genome rearrangements, number of indels and plastome size. Accelerated rates and more genome rearrangement events could be the result of a faulty DNA repair/recombination/replication system as has been suggested in Campanulaceae and Geraniaceae. Exploring nucleotide substitution rates in mitochondrial genes may give some insight into whether this trend is present in genomes of other cellular compartments or limited to the plastome. The negative correlation between rates and plastome size could be explained by mutational burden. However, effective population size could also be a factor in this correlation and more work is needed in order to untangle the cause of this correlation.

Table 3.1. Species included in analyses, along with GenBank Accession numbers and Fabaceae subfamily to which each taxon belongs. For accession numbers of individual genes for *T. pratense*, see Table XX.

Species	Accession No.	Subfamily	Habit
Castanea mollissima	NC_014674	Outgroup	NA*
Cucumis sativus	NC_007144.1	Outgroup	NA*
Fragaria vesca	NC_015206	Outgroup	NA*
Morus indica	NC_008359.1	Outgroup	NA*
Caesalpinia coriaria	KJ468095	Caesalpinioideae	Woody
Ceratonia siliqua	KJ468096	Caesalpinioideae	Woody
Cercis canadensis	KF856619	Caesalpinioideae	Woody
Haematoxylum brasiletto	KJ468097	Caesalpinioideae	Woody
Tamarindus indica	KJ468103	Caesalpinioideae	Woody
Acacia ligulata	NC_026134.1	Mimosoideae	Woody
Inga leiocalycina	NC_028732	Mimosoideae	Woody
Leucaena trichandra	KT428297	Mimosoideae	Woody
Prosopis gladulosa	KJ468101	Mimosoideae	Woody
Apios americana	KF856618	Papilionoideae	Herbaceous
Arachis hypogaea	KJ468094	Papilionoideae	Herbaceous
Glycine max	NC_007942	Papilionoideae	Herbaceous
Indigofera tinctoria	KJ468098	Papilionoideae	Woody
Lotus japonicus	NC_002694	Papilionoideae	Herbaceous
Lupinus albus	KJ468099	Papilionoideae	Herbaceous
Millettia pinnata	NC_016708	Papilionoideae	Woody
Pachyrhizus erosus	KJ468100	Papilionoideae	Herbaceous
Phaseolus vulgaris	NC_009259.1	Papilionoideae	Herbaceous
Robinia pseudoacacia	KJ468102	Papilionoideae	Woody
Vigna unguiculata	KJ468104	Papilionoideae	Herbaceous
Astragalus nakaianus	NC_028171	Papilionoideae_IRLC	Woody
Cicer arietum	NC_011163.1	Papilionoideae_IRLC	Herbaceous
Glycyrrhiza glabra	KF201590	Papilionoideae_IRLC	Herbaceous
Lathyrus sativus	NC_014063	Papilionoideae_IRLC	Herbaceous
Lens culinaris	KF186232	Papilionoideae_IRLC	Herbaceous
Medicago truncatula	NC_003119.6	Papilionoideae_IRLC	Herbaceous
Pisum sativum	NC_014057	Papilionoideae_IRLC	Herbaceous
Trifolium aureum	KC894708	Papilionoideae_IRLC	Herbaceous
Trifolium boissieri	NC_025745	Papilionoideae_IRLC	Herbaceous
Trifolium glanduliferum	NC_024034	Papilionoideae_IRLC	Herbaceous

Table 3.1 (continued)

Trifolium grandiflorum	**KC894707**	**Papilionoideae_IRLC**	**Herbaceous**
Trifolium lupinaster	KJ788287	Papilionoideae_IRLC	Herbaceous
Trifolium meduseum	KJ476730	Papilionoideae_IRLC	Herbaceous
Trifolium pratense	Table XX	Papilionoideae_IRLC	Herbaceous
Trifolium repens	KC894706	Papilionoideae_IRLC	Herbaceous
Trifolium strictum	NC_025745	Papilionoideae_IRLC	Herbaceous
Trifolium subterraneum	NC_011828	Papilionoideae_IRLC	Herbaceous
Vicia faba	KF042344	Papilionoideae_IRLC	Herbaceous
Wisteria floribunda	NC_027677	Papilionoideae_IRLC	Woody

*NA = not applicable

Table 3.2. List of genes utilized in all analyses of this study.

Gene name		
atpA	*psaC*	*rps3*
atpB	*psaJ*	*rps4*
atpE	*psbA*	*rps7*
atpF	*psbB*	*rps8*
atpH	*psbC*	*rps11*
atpI	*psbD*	*rps12*
ccsA	*psbE*	*rps14*
cemA	*psbF*	*rps15*
clpP	*psbH*	*rps18*
matK	*psbI*	*rps19*
ndhA	*psbJ*	*ycf1*
ndhB	*psbK*	*ycf2*
ndhC	*psbL*	*ycf3*
ndhD	*psbM*	
ndhE	*psbN*	
ndhF	*psbT*	
ndhG	*psbZ*	
ndhH	*rbcL*	
ndhI	*rpl2*	
ndhJ	*rpl14*	
ndhK	*rpl16*	
petA	*rpl20*	
petB	*rpl32*	
petD	*rpl36*	
petG	*rpoA*	
petL	*rpoB*	
petN	*rpoC1*	
psaA	*rpoC2*	
psaB	*rps2*	

Table 3.3. Genes included in the functional groups utilized in substitution rate analyses

Functional group	Genes included
ATP	*atpA, atpB, atpE, atpF, atpH, atpI*
ccsA	*ccsA*
cemA	*cemA*
clpP	*clpP*
NDH	*ndhA, ndhB, ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK*
matK	*matK*
PET	*petA, petB, petD, petG, petL, petN*
PSA	*psaA, psaB, psaC, psaJ*
PSB	*psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbL, psbM, psbN, psbT, psbZ*
rbcL	*rbcL*
RPO	*rpoA, rpoB, rpoC1, rpoC2*
RPL	*rpl2, rpl14, rpl16, rpl20, rpl32, rpl36*
RPS	*rps2, rps3, rps4, rps7, rps8, rps11, rps12, rps14, rps15, rps18, rps19*
ycf1	*ycf2*
ycf2	*ycf3*
ycf3	*ycf4*

Table 3.4. Mean substitution rates of eight functional groups and eight genes that cannot be placed in any of those groups.

Mean values of functional groups		
	dN	***dS***
ATP	0.0033	0.0362
ccsA	0.0092	0.0395
cemA	0.0112	0.0295
clpP	0.0266	0.0551
matK	0.0160	0.0383
NDH	0.0081	0.0226
PET	0.0027	0.0354
PSA	0.0015	0.0294
PSB	0.0014	0.0278
rbcL	0.0032	0.0398
RPL	0.0064	0.0345
RPO	0.0070	0.0370
RPS	0.0078	0.0341
ycf1	0.0465	0.0759
ycf2	0.0113	0.0177
ycf3	0.0016	0.0299

Table 3.5. Results from Kruskal-Wallis test of dN and dS values for each gene between the four Fabaceae subgroups. All results are significant with the exception of those genes marked with a (±).

dN			***dS***		
Gene	**P_values**	**Holm**	**Gene**	**P_values**	**Holm**
atpA	1.04E-06	7.07E-05	*atpA*	8.47E-05	4.14E-03
atpB	2.51E-06	1.50E-04	*atpB*	9.07E-05	4.26E-03
atpE	7.52E-06	4.06E-04	*atpE*	1.29E-04	4.60E-03
atpF	1.55E-04	5.89E-03	*atpF*	2.71E-05	1.52E-03
atpH	1.82E-05	9.12E-04	*atpH*	2.53E-06	1.75E-04
atpI	7.68E-05	3.38E-03	*atpI*	1.76E-05	1.04E-03
ccsA	4.60E-04	1.06E-02	*ccsA*	1.33E-04	4.60E-03
cemA	3.39E-04	8.64E-03	*cemA*	1.06E-04	4.60E-03
clpP	9.48E-05	4.08E-03	*clpP*	4.92E-06	3.30E-04
matK	1.33E-04	5.33E-03	*matK*	1.19E-04	4.60E-03
ndhA	2.43E-04	6.79E-03	*ndhA*	5.06E-05	2.63E-03
ndhB	1.90E-06	1.24E-04	*ndhB*	1.83E-06	1.28E-04
ndhC	5.14E-04	1.06E-02	*ndhC*	7.92E-04	1.21E-02
ndhD	1.84E-04	6.07E-03	*ndhD*	1.15E-04	4.60E-03
ndhE	1.97E-04	6.07E-03	*ndhE*	8.45E-06	5.49E-04
ndhF	9.03E-04	1.44E-02	*ndhF*	8.13E-05	4.06E-03
ndhG	2.48E-03	2.48E-02	*ndhG*	1.27E-04	4.60E-03
ndhH	6.77E-04	1.22E-02	*ndhH*	3.91E-05	2.07E-03
ndhI	2.10E-03	2.31E-02	*ndhI*	5.19E-03	4.15E-02
ndhJ	1.76E-04	6.07E-03	*ndhJ*	1.38E-04	4.60E-03
ndhK	1.91E-04	6.07E-03	*ndhK*	1.10E-04	4.60E-03
petA	2.63E-03	2.48E-02	*petA*	2.97E-05	1.60E-03
petB	1.16E-05	6.02E-04	*petB*	1.16E-04	4.60E-03
petD	4.94E-04	1.06E-02	*petD*	1.43E-04	4.60E-03
petG	5.21E-03	3.03E-02	*petG*	6.06E-04	1.15E-02
petL	5.55E-03	3.03E-02	*petL*	9.25E-04	1.21E-02
petN	1.78E-04	6.07E-03	*petN*	2.04E-05	1.18E-03
psaA	8.99E-04	1.44E-02	*psaA*	1.05E-04	4.60E-03
psaB	7.30E-04	1.24E-02	*psaB*	3.48E-04	7.31E-03
psaC	2.56E-05	1.20E-03	*psaC*	5.73E-04	1.15E-02
psaJ	1.92E-06	1.24E-04	*psaJ*	3.85E-03	3.47E-02
psbA	1.43E-03	1.71E-02	*psbA*	1.02E-03	1.22E-02
psbB	1.85E-05	9.12E-04	*psbB*	9.99E-05	4.60E-03

Table 3.5 (continued)

psbC	1.62E-05	8.28E-04		*psbC*	1.09E-05	6.76E-04
psbD	1.74E-04	6.07E-03		*psbD*	1.33E-04	4.60E-03
psbE	4.85E-05	2.18E-03		*psbE* ±	1.64E-02	6.54E-02
psbF	3.73E-04	8.95E-03		*psbF*	1.11E-03	1.22E-02
psbH	3.66E-06	2.16E-04		*psbH*	2.21E-04	5.09E-03
psbI	2.13E-06	1.34E-04		*psbI* ±	4.96E-02	1.11E-01
psbJ ±	3.94E-01	1.00E+00		*psbJ* ±	1.12E-02	5.59E-02
psbK	2.49E-03	2.48E-02		*psbK*	3.70E-02	1.11E-01
psbL	3.93E-03	2.75E-02		*psbL*	9.22E-06	5.81E-04
psbM ±	4.16E-01	1.00E+00		*psbM*	2.31E-04	5.09E-03
psbN	3.23E-04	8.64E-03		*psbN*	6.47E-02	1.11E-01
psbT	1.11E-04	4.56E-03		*psbT*	8.21E-03	4.93E-02
psbZ ±	4.43E-01	1.00E+00		*psbZ*	2.22E-05	1.27E-03
rbcL	2.04E-04	6.07E-03		*rbcL*	6.21E-06	4.10E-04
rpl14	2.42E-06	1.48E-04		*rpl14*	1.54E-05	9.40E-04
rpl16	7.52E-07	5.34E-05		*rpl16*	1.33E-04	4.60E-03
rpl20	3.90E-06	2.26E-04		*rpl20*	2.83E-05	1.56E-03
rpl2	8.64E-06	4.58E-04		*rpl2*	4.41E-06	3.00E-04
rpl32	2.00E-05	9.61E-04		*rpl32*	7.59E-04	1.21E-02
rpl36	3.52E-05	1.62E-03		*rpl36*	1.43E-04	4.60E-03
rpoA	3.20E-04	8.64E-03		*rpoA*	6.40E-05	3.26E-03
rpoB	1.62E-04	5.89E-03		*rpoB*	1.63E-04	4.60E-03
rpoC1	9.67E-05	4.08E-03		*rpoC1*	1.18E-04	4.60E-03
rpoC2	1.34E-04	5.33E-03		*rpoC2*	1.37E-04	4.60E-03
rps11	5.53E-06	3.15E-04		*rps11*	1.45E-04	4.60E-03
rps12	1.56E-04	5.89E-03		*rps12*	8.44E-05	4.14E-03
rps14	2.33E-06	1.45E-04		*rps14*	1.04E-04	4.60E-03
rps15	1.03E-03	1.44E-02		*rps15*	8.53E-04	1.21E-02
rps18	1.10E-06	7.38E-05		*rps18*	6.41E-04	1.15E-02
rps19	6.21E-06	3.48E-04		*rps19*	2.78E-03	2.78E-02
rps2	4.72E-04	1.06E-02		*rps2*	1.44E-04	4.60E-03
rps3	6.45E-06	3.55E-04		*rps3*	5.53E-03	4.15E-02
rps4	5.05E-03	3.03E-02		*rps4*	1.20E-04	4.60E-03
rps7	1.72E-06	1.14E-04		*rps7*	8.45E-06	5.49E-04
rps8	1.10E-03	1.44E-02		*rps8*	7.04E-04	1.20E-02
ycf1	8.29E-07	5.80E-05		*ycf1*	1.01E-04	4.60E-03
ycf2	9.40E-07	6.49E-05		*ycf2*	6.18E-07	4.39E-05
ycf3	5.70E-04	1.08E-02		*ycf3*	1.66E-05	9.97E-04

Table 3.6. Kruskal-Wallis values of pairwise nucleotide substitution rates compared between four groups of legumes: mimosoids, caesalpinioids, IR-containing papilionoids and IRLC papilionoids. Holm correction was performed on all p-values.

	dN		***dS***	
Genes	**P-values**	**Holm corrected**	**P-values**	**Holm corrected**
psbA	0.0641	0.6302	0.0069	0.1719
matK	0.0066	0.2855	0.0397	0.4371
psbK	0.0075	0.2855	0.0983	0.7305
psbI	0.0080	0.2887	0.0913	0.7305
atpA	0.0022	0.1435	0.0047	0.1517
atpF	0.0033	0.1857	0.0029	0.1517
atpH	0.0040	0.2081	0.0030	0.1517
atpI	0.0041	0.2098	0.0049	0.1517
rps2	0.0165	0.3965	0.0034	0.1517
rpoC2	0.0039	0.2081	0.0029	0.1517
rpoC1	0.0029	0.1744	0.0039	0.1517
rpoB	0.0048	0.2251	0.0039	0.1517
petN	0.0123	0.3393	0.0025	0.1440
psbM	0.8414	1.0000	0.0610	0.6098
psbD	0.0337	0.4918	0.0022	0.1396
psbC	0.0201	0.4274	0.0023	0.1396
psbZ	0.3545	1.0000	0.0034	0.1517
rps14	0.0098	0.3037	0.0036	0.1517
psaB	0.0304	0.4918	0.0023	0.1396
psaA	0.0177	0.4065	0.0048	0.1517
ycf3	0.0117	0.3393	0.0015	0.1027
rps4	0.0573	0.6302	0.0019	0.1242
ndhJ	0.0069	0.2855	0.0027	0.1509
ndhK	0.0080	0.2887	0.0028	0.1517
ndhC	0.0726	0.6302	0.0107	0.2241
atpE	0.0029	0.1744	0.0033	0.1517
atpB	0.0068	0.2855	0.0022	0.1396
rbcL	0.0637	0.6302	0.0012	0.0871

Table 3.6 (continued)

cemA	**0.0061**	**0.2790**	**0.0036**	**0.1517**
petA	0.0589	0.6302	0.0026	0.1483
psbJ	0.7217	1.0000	0.1103	0.7305
psbL	0.0282	0.4918	0.0021	0.1372
psbF	0.0095	0.3037	0.0288	0.3747
psbE	0.0023	0.1489	0.9575	1.0000
petL	0.0068	0.2855	0.0233	0.3491
petG	0.0369	0.4918	0.0036	0.1517
psaJ	0.0043	0.2101	0.4638	1.0000
rps18	0.0081	0.2887	0.0210	0.3364
rpl20	0.0018	0.1267	0.0028	0.1517
clpP	0.0688	0.6302	0.0169	0.3038
psbB	0.0020	0.1360	0.0035	0.1517
psbT	0.0197	0.4274	0.0310	0.3747
psbN	0.2131	0.8523	0.2235	1.0000
psbH	0.0096	0.3037	0.0182	0.3087
petB	0.0087	0.2887	0.0046	0.1517
petD	0.0194	0.4274	0.0032	0.1517
rpoA	0.0041	0.2098	0.0138	0.2620
rps11	0.0044	0.2101	0.0033	0.1517
rpl36	0.0030	0.1744	0.0036	0.1517
rps8	0.0119	0.3393	0.0114	0.2287
rpl14	0.0028	0.1714	0.0084	0.1858
rpl16	0.0014	0.1005	0.0039	0.1517
rps3	0.0144	0.3611	0.2879	1.0000
rps19	0.0030	0.1744	0.0233	0.3491
rpl2	0.0039	0.2081	0.0017	0.1163
ycf2	0.0029	0.1744	0.0016	0.1073
ndhB	0.0020	0.1360	0.0076	0.1830
rps7	0.0070	0.2855	0.0016	0.1070
rps12	0.0024	0.1546	0.0044	0.1517
ycf1	0.0021	0.1407	0.0031	0.1517
rps15	0.0273	0.4918	0.0728	0.6550
ndhH	0.0134	0.3492	0.0021	0.1372
ndhA	0.0067	0.2855	0.0035	0.1517

Table 3.6 (continued)

ndhI	**0.0477**	**0.5726**	**0.3442**	**1.0000**
ndhG	0.0292	0.4918	0.0028	0.1517
ndhE	0.0061	0.2790	0.0022	0.1396
psaC	0.0061	0.2790	0.0052	0.1517
ndhD	0.0033	0.1857	0.0057	0.1517
ccsA	0.0223	0.4274	0.0030	0.1517
rpl32	0.0027	0.1697	0.0081	0.1854
ndhF	0.0620	0.6302	0.0029	0.1517

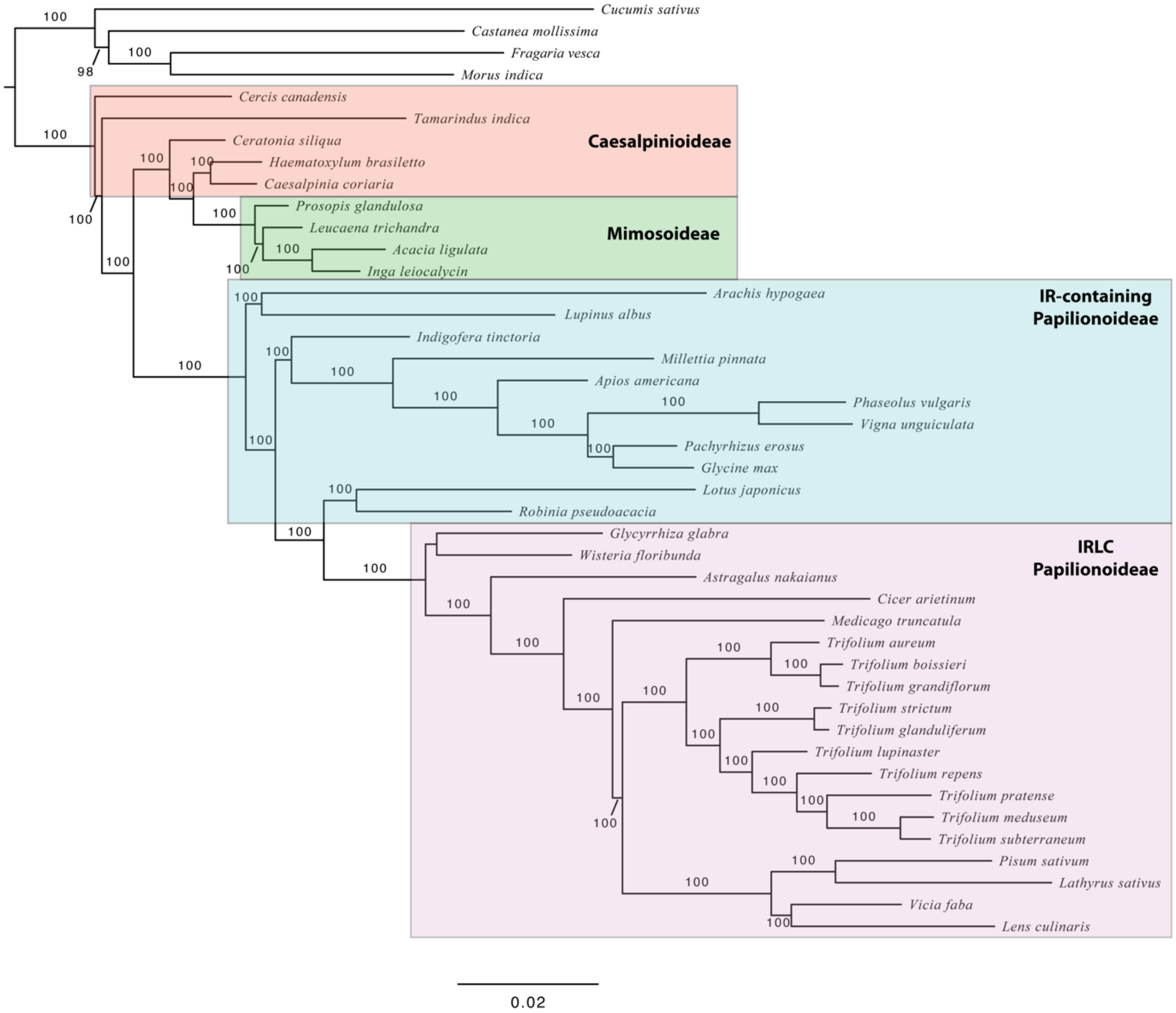


Figure 3.1. Maximum likelihood tree (-1n = -449433.539651) of Fabaceae based on 71 plastid genes.

Numbers above branches are bootstrap support values. The scale bar represents substitutions per site. The phylogeny is divided into four subgroups: Caesalpinioideae (red), Mimosoideae (green), Papilionoideae taxa containing both copies of the IR (inverted repeat) (blue) and Papilionoids lacking the IR (purple). Bootstrap values > 50 are shown at nodes.

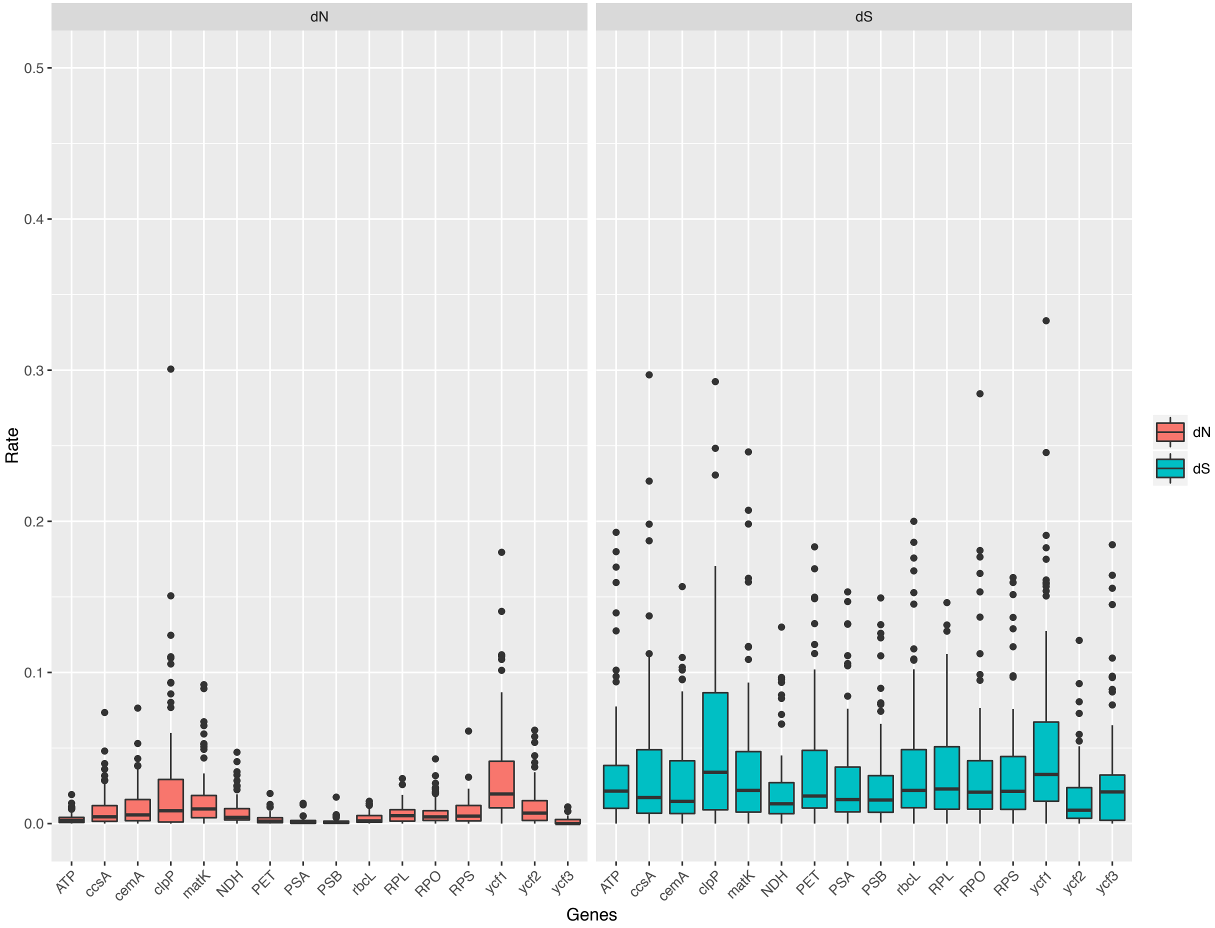


Figure 3.2. *dN* and *dS* of functional groups with the highest values removed.

Box plots of *dN* values (red) and *dS* values (green) of eight functional groups and eight individual genes. The top and bottom lines of the box represent the 75th and 25th percentiles, respectively and the middle line in each box represents the 50th percentile. The whisker lines represent the minimum to the maximum points and the points outside of the whisker lines are outliers. Two extremely high values for *ycf1* were removed in order to provide a closer view of the relationships between functional groups.

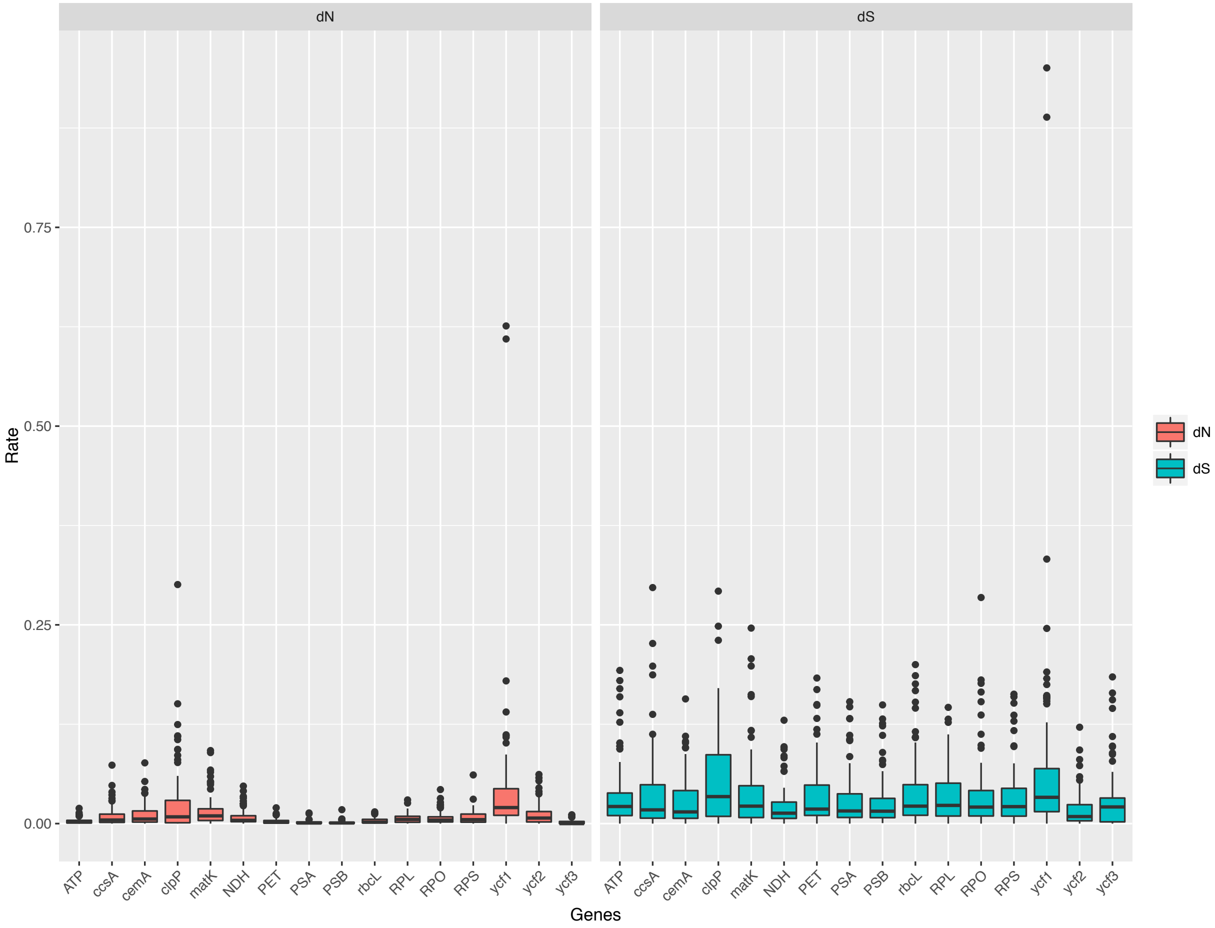


Figure 3.3. *dN* and *dS* of functional groups with all data points.

Box plots of *dN* values (red) and *dS* values (green) of eight functional groups and eight individual genes. The top and bottom lines of the box represent the 75th and 25th percentiles, respectively and the middle line in each box represents the 50th percentile. The whisker lines represent the minimum to the maximum points and the points outside of the whisker lines are outliers.

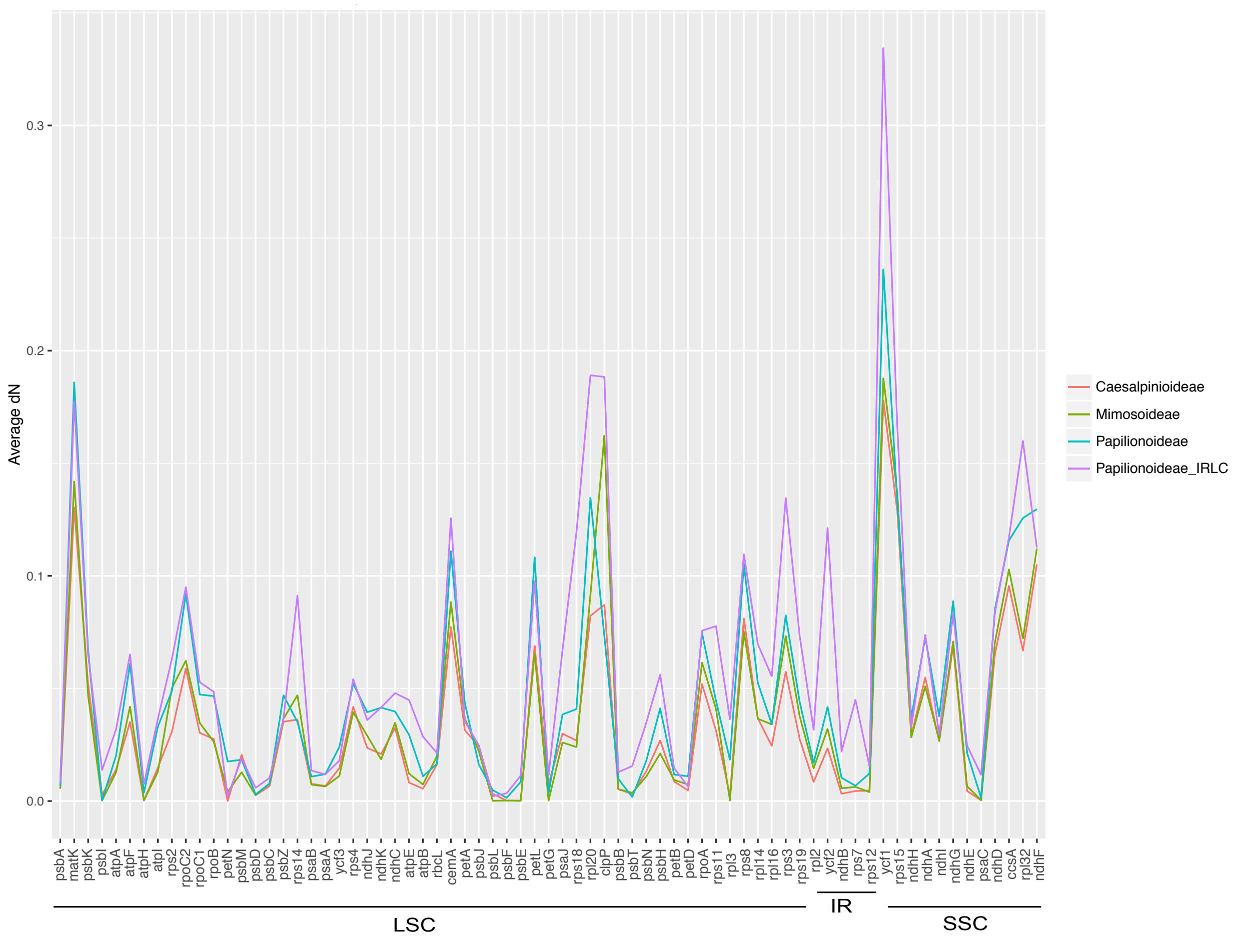


Figure 3.4. Pairwise comparison of *dN* values across the genome of four groups of Fabaceae.

Line plot representing average *dN* values for each gene within the Caesalpinioideae (red line), Mimosoideae (green line), Papilionoideae (blue line), Papilionoideae IRLC (purple). Genes are in the ancestral gene order with genes in the LSC (large single copy region), IR (inverted repeat) and SSC (small single copy region) labeled.

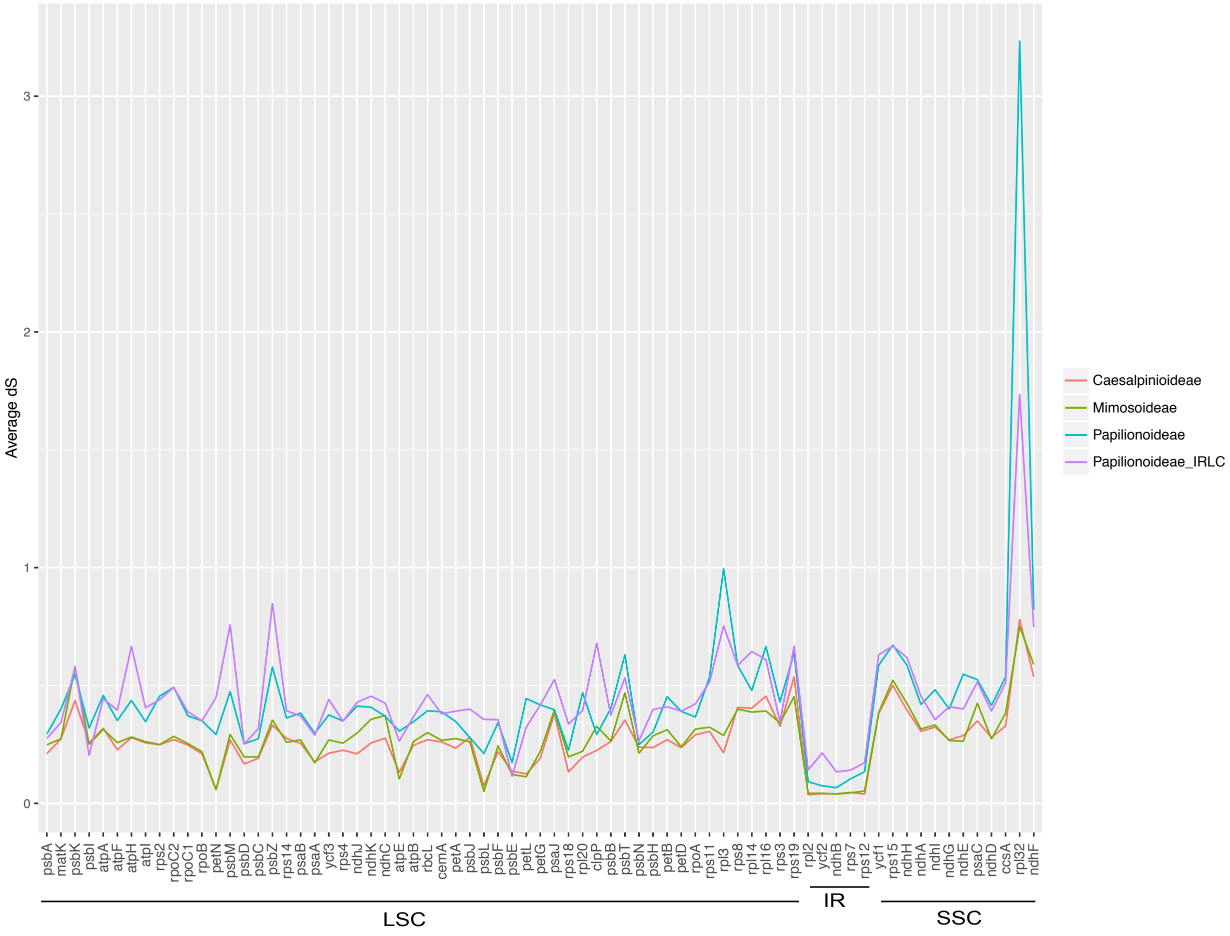


Figure 3.5. Pairwise comparison of *dS* values across the genome of four groups of Fabaceae.

Line plot representing average *dS* values for each gene within the Caesalpinioideae (red line), Mimosoideae (green line), Papilionoideae (blue line), Papilionoideae IRLC (purple). Genes are in the ancestral gene order with genes in the LSC (large single copy region), IR (inverted repeat) and SSC (small single copy region) labeled.

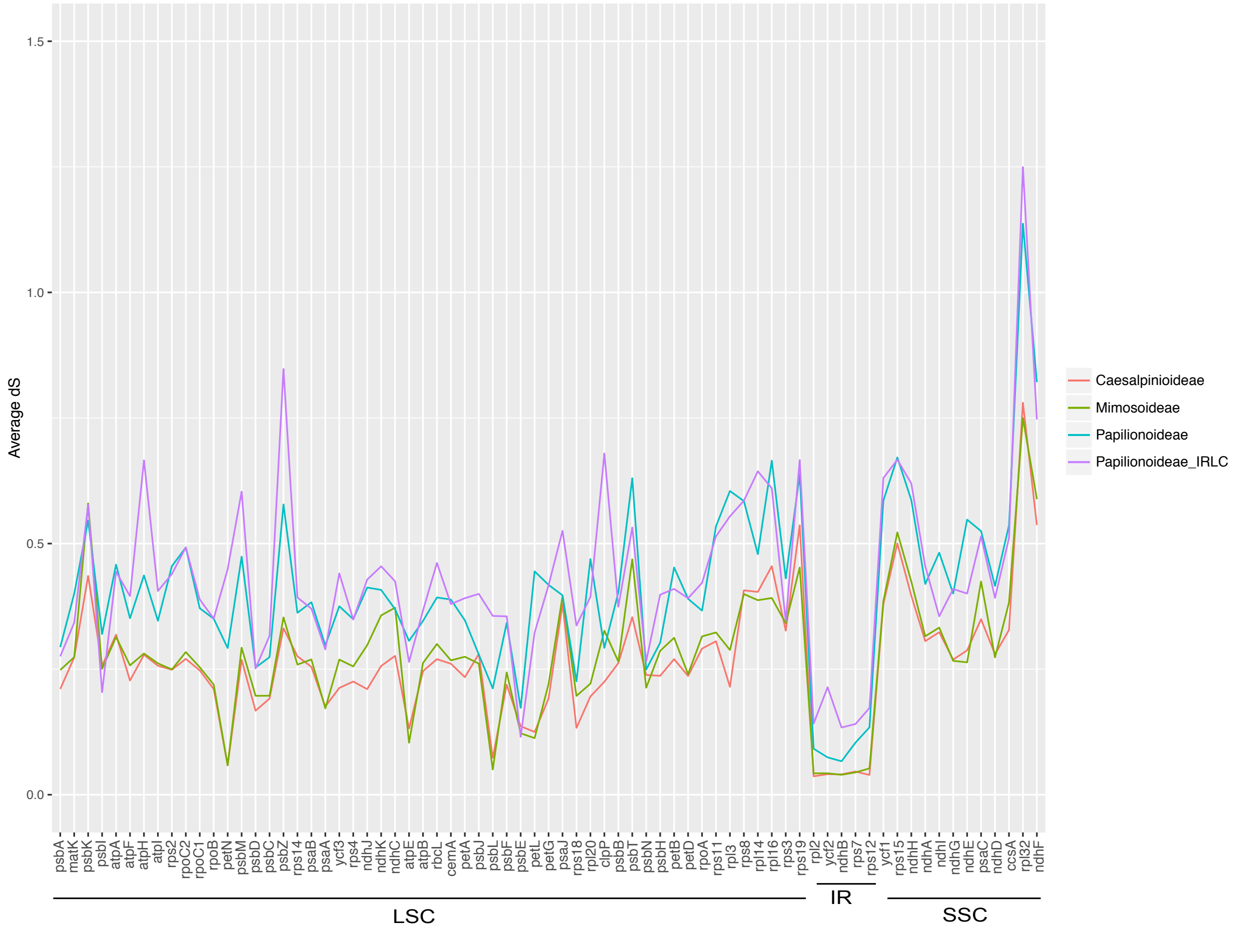


Figure 3.6. Pairwise comparison of *dS* values across the genome of four groups of Fabaceae with outliers removed.

Line plot representing average *dS* values for each gene within the Caesalpinioideae (red line), Mimosoideae (green line), Papilionoideae (blue line), Papilionoideae IRLC (purple). Genes are in the ancestral gene order with genes in the LSC (large single copy region), IR (inverted repeat) and SSC (small single copy region) labeled. Extremely high values of *rpl32* in both papilionoid lineages were cutoff at 1.5 in order to provide a better view of the relationships between the subgroups.

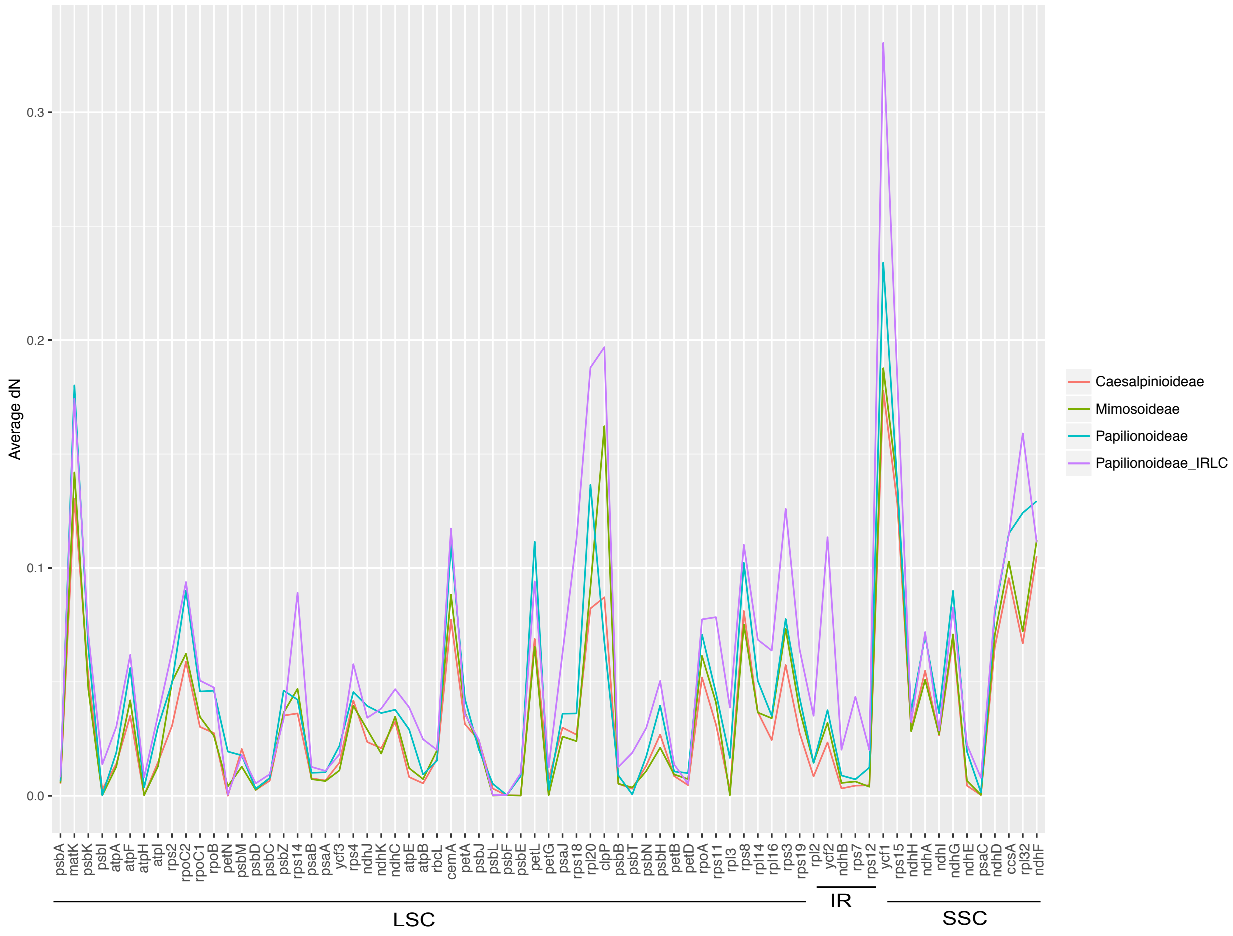


Figure 3.7. Pairwise comparison of *dN* values across the genome of four groups of Fabaceae using a subset of papilionoid taxa.

Line plot representing average *dN* values for each gene within the Caesalpinioideae (red line), Mimosoideae (green line), Papilionoideae (blue line), Papilionoideae IRLC (purple). Genes are in the ancestral gene order with genes in the LSC (large single copy region), IR (inverted repeat) and SSC (small single copy region) labeled. A subset of taxa including only five species each of papilionoid and papilionoid IRLC were utilized (see Methods).

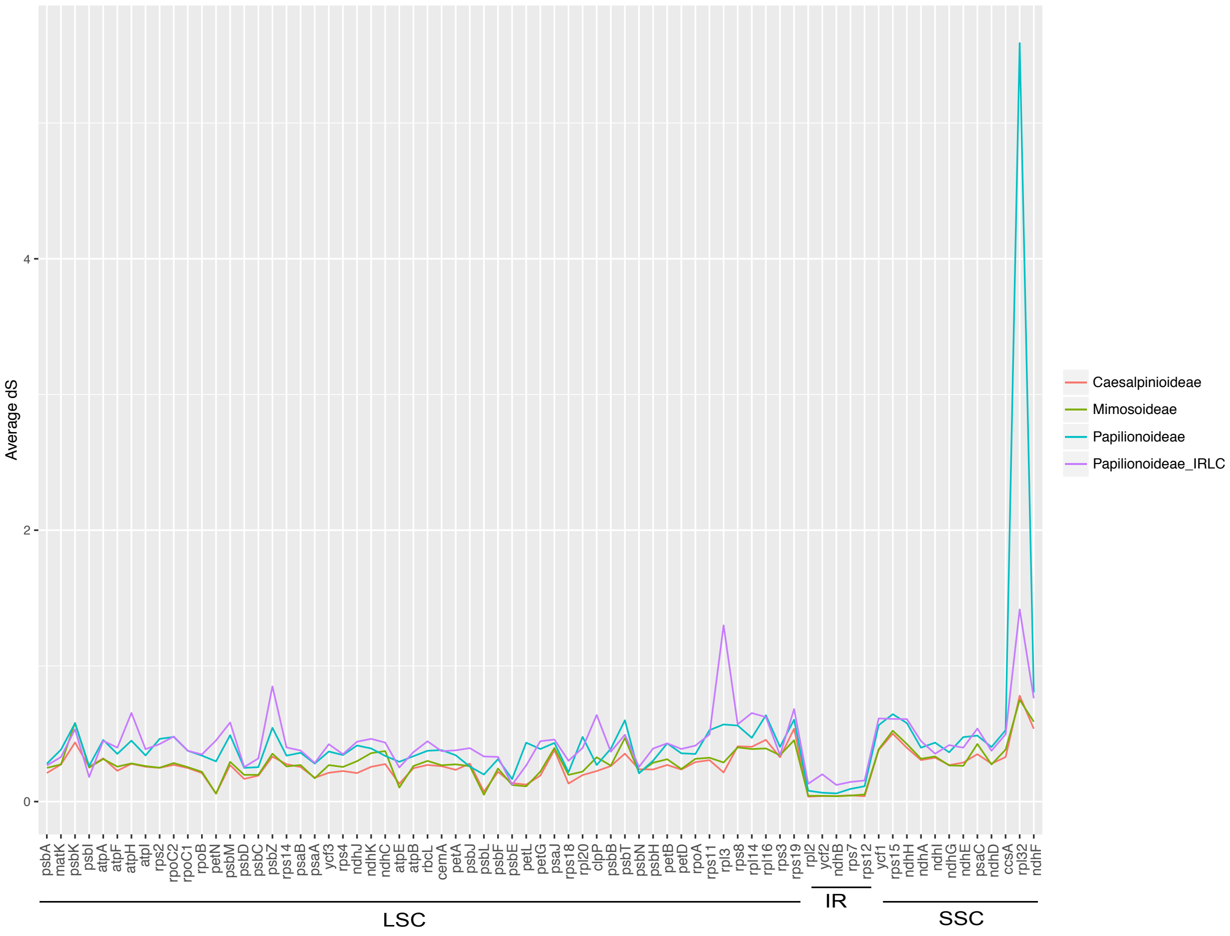


Figure 3.8. Pairwise comparison of *dS* values across the genome of four groups of Fabaceae using a subset of papilionoid taxa.

Line plot representing average *dN* values for each gene within the Caesalpinioideae (red line), Mimosoideae (green line), Papilionoideae (blue line), Papilionoideae IRLC (purple). Genes are in the ancestral gene order with genes in the LSC (large single copy region), IR (inverted repeat) and SSC (small single copy region) labeled. A subset of taxa including only five species each of papilionoid and papilionoid IRLC were utilized (see Methods).

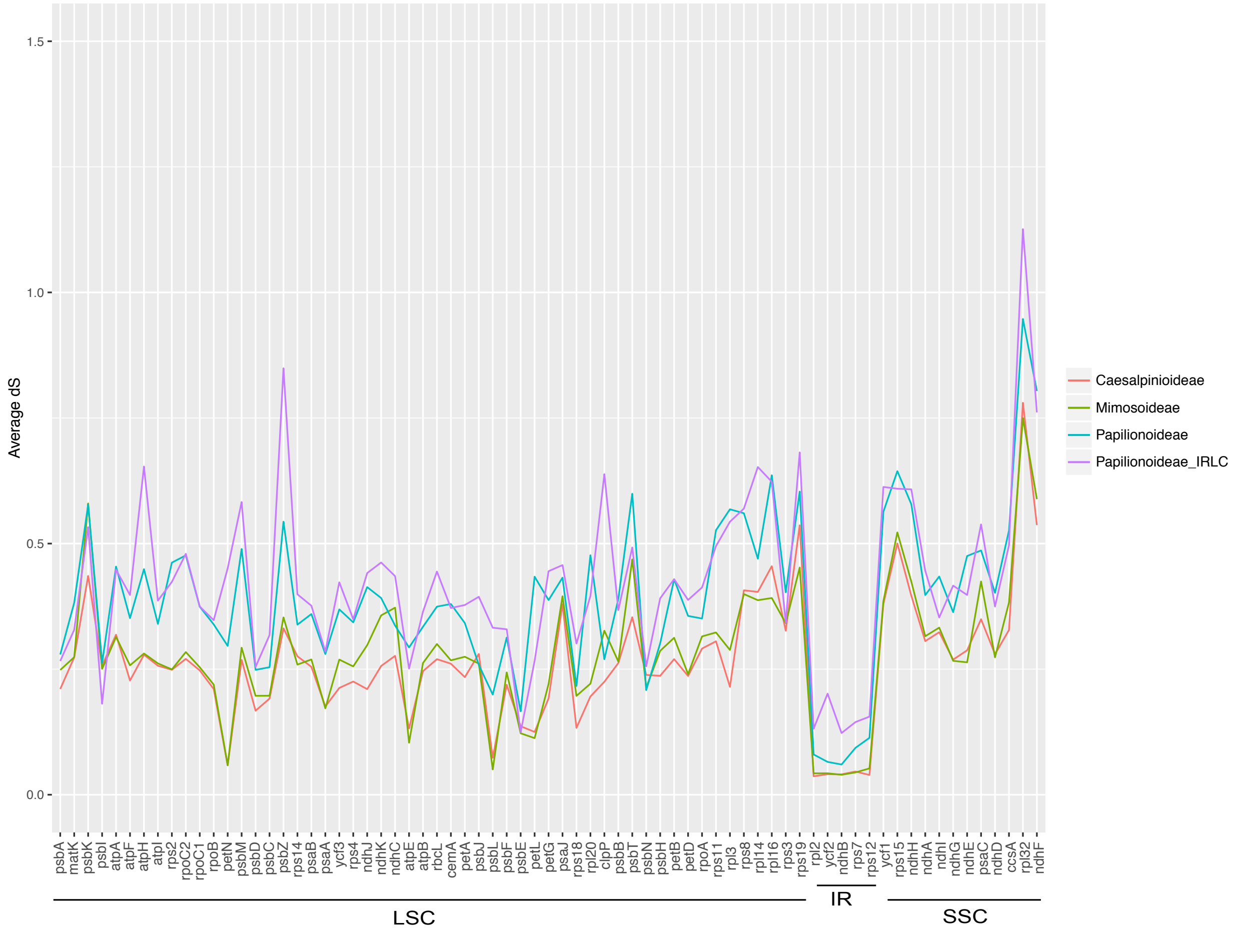


Figure 3.9. Pairwise comparison of *dS* values across the genome of four groups of Fabaceae using a subset of papilionoid taxa and with outliers removed.

Line plot representing average *dS* values for each gene within the Caesalpinioideae (red line), Mimosoideae (green line), Papilionoideae (blue line), Papilionoideae IRLC (purple). Genes are in the ancestral gene order with genes in the LSC (large single copy region), IR (inverted repeat) and SSC (small single copy region) labeled. Extremely high values of *rpl32* in both papilionoid lineages were cutoff at 1.5 in order to provide a better view of the relationships between the subgroups. A subset of taxa including only five species each of papilionoid and papilionoid IRLC were utilized (see Methods).

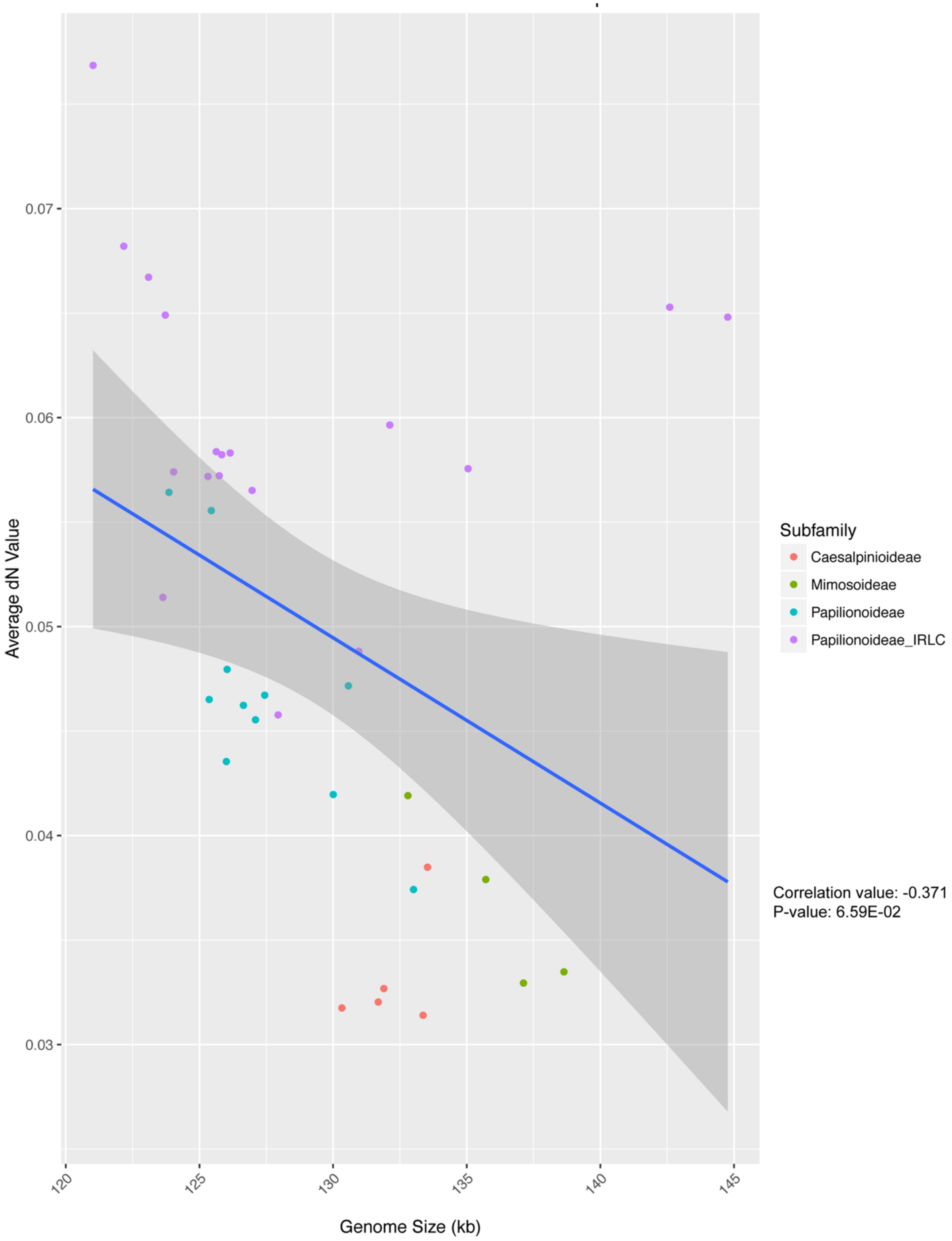

Average dN Value
0.07
0.06
0.05
0.04
0.03
120
125
130
135
140
145
Genome Size (kb)
Subfamily
● Caesalpinioideae
● Mimosoideae
● Papilionoideae
● Papilionoideae_IRLC
Correlation value: -0.371
P-value: 6.59E-02

Figure 3.10. Correlation between genome size and *dN* values.

Scatterplot with regression line (blue line) of average *dN* values for each genome compared to genome size in kilobases. The color of each point represents the subgroup to which it belongs: Caesalpiniodeae (red), Mimosoideae (green) Papilionoideae (blue), Papilionoideae IRLC (purple). The grey region surrounding the regression line represents the standard error. Correlation value = -0.371, p-value = 6.59E-02.

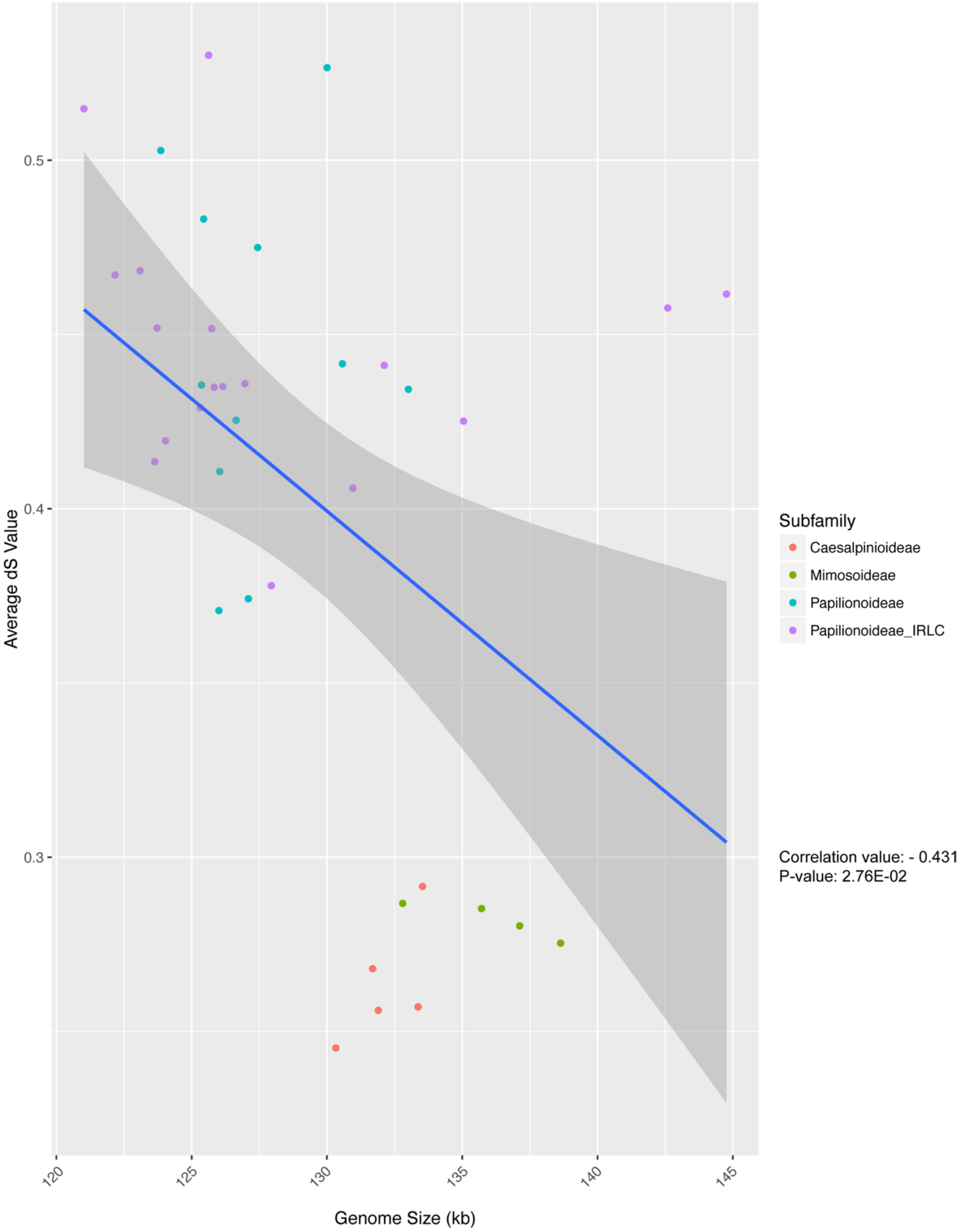

0.5
0.4
0.3
Average dS Value
120
125
130
135
140
145
Genome Size (kb)
Subfamily
● Caesalpinioideae
● Mimosoideae
● Papilionoideae
● Papilionoideae_IRLC
Correlation value: - 0.431
P-value: 2.76E-02

Figure 3.11. Correlation between genome size and *dS* values.

Scatterplot with regression line (blue line) of average *dS* values for each genome compared to genome size in kilobases. The color of each point represents the subgroup to which it belongs: Caesalpiniodeae (red), Mimosoideae (green) Papilionoideae (blue), Papilionoideae IRLC (purple). The grey region surrounding the regression line represents the standard error. Correlation value = -0.431, p-value = 2.76E-02.

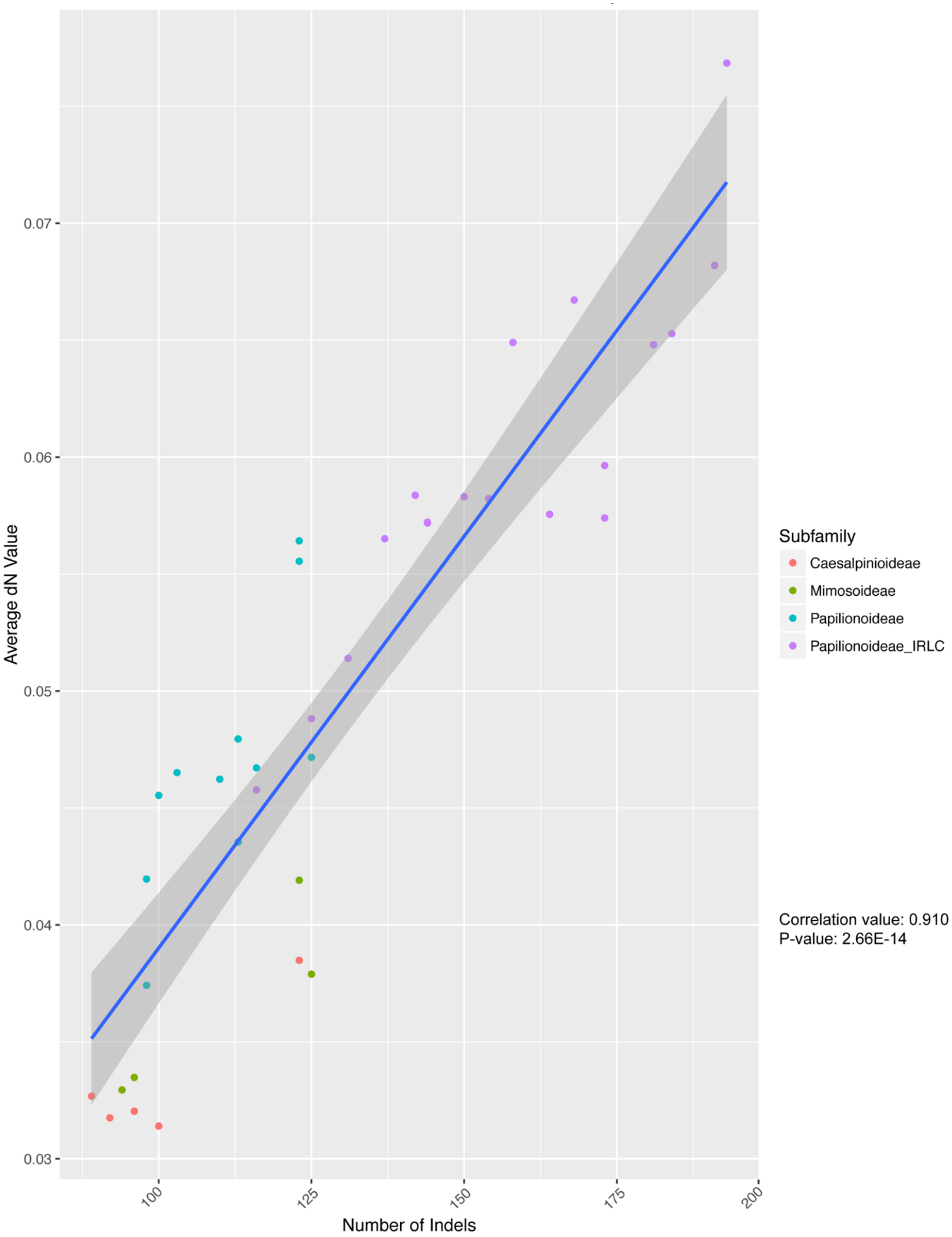
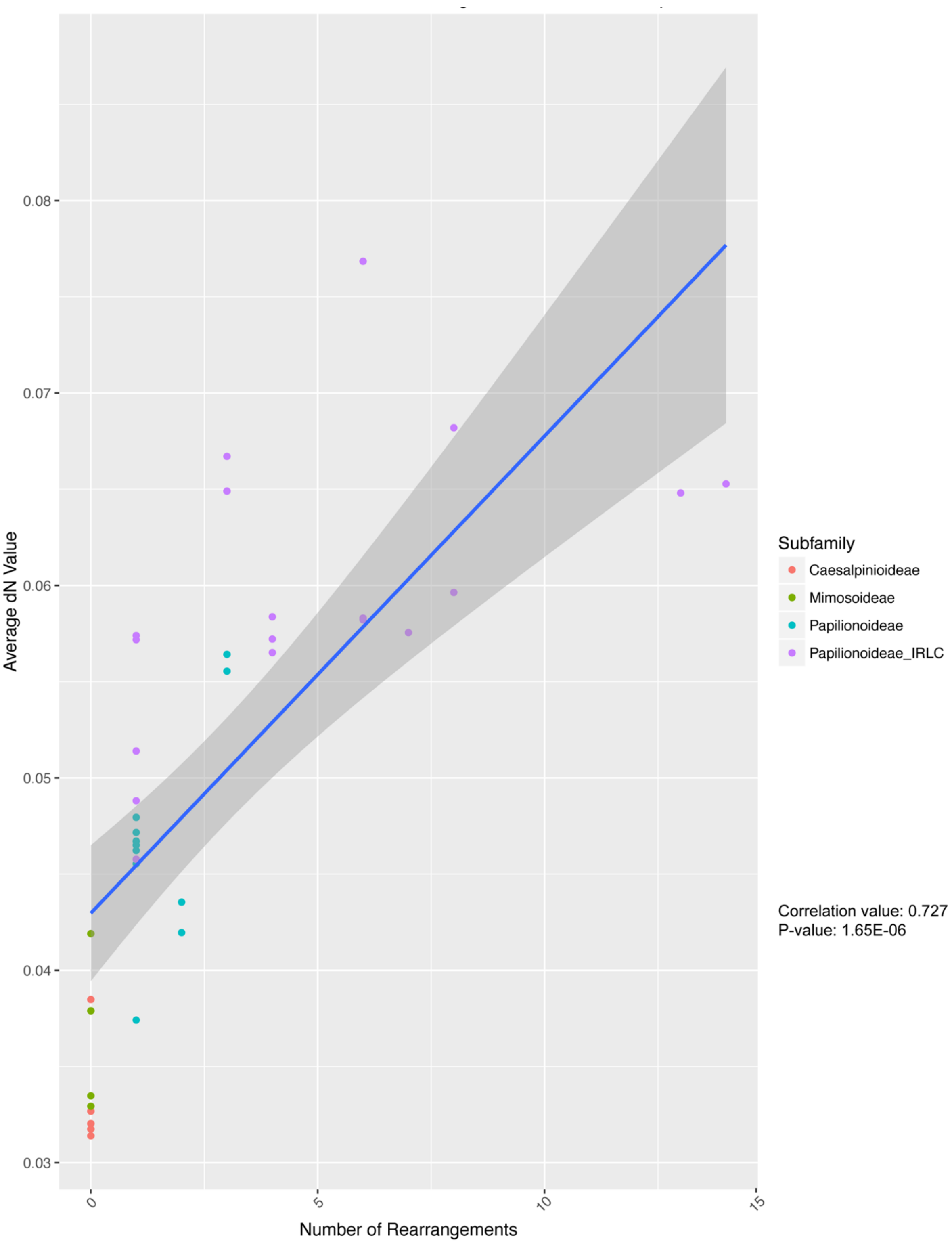

Average dN Value
0.03
0.04
0.05
0.06
0.07
100
125
150
175
200
Number of Indels
Subfamily
● Caesalpinioideae
● Mimosoideae
● Papilionoideae
● Papilionoideae_IRLC
Correlation value: 0.910
P-value: 2.66E-14

Figure 3.12 Correlation between indel count and *dN* values.

Scatterplot with regression line (blue line) of average *dN* values for each genome compared to number of indels. The color of each point represents the subgroup to which it belongs: Caesalpiniodeae (red), Mimosoideae (green) Papilionoideae (blue), Papilionoideae IRLC (purple). The grey region surrounding the regression line represents the standard error. Correlation value = 0.910, p-value = 2.66E-14.



Correlation value: 0.727
P-value: 1.65E-06

Figure 3.13. Correlation between rearrangements and *dN* values.

Scatterplot with regression line (blue line) of average *dN* values for each genome compared to number of rearrangements. The color of each point represents the subgroup to which it belongs: Caesalpiniodeae (red), Mimosoideae (green) Papilionoideae (blue), Papilionoideae IRLC (purple). The grey region surrounding the regression line represents the standard error. Correlation value = 0.727, p-value = 1.65E-06.

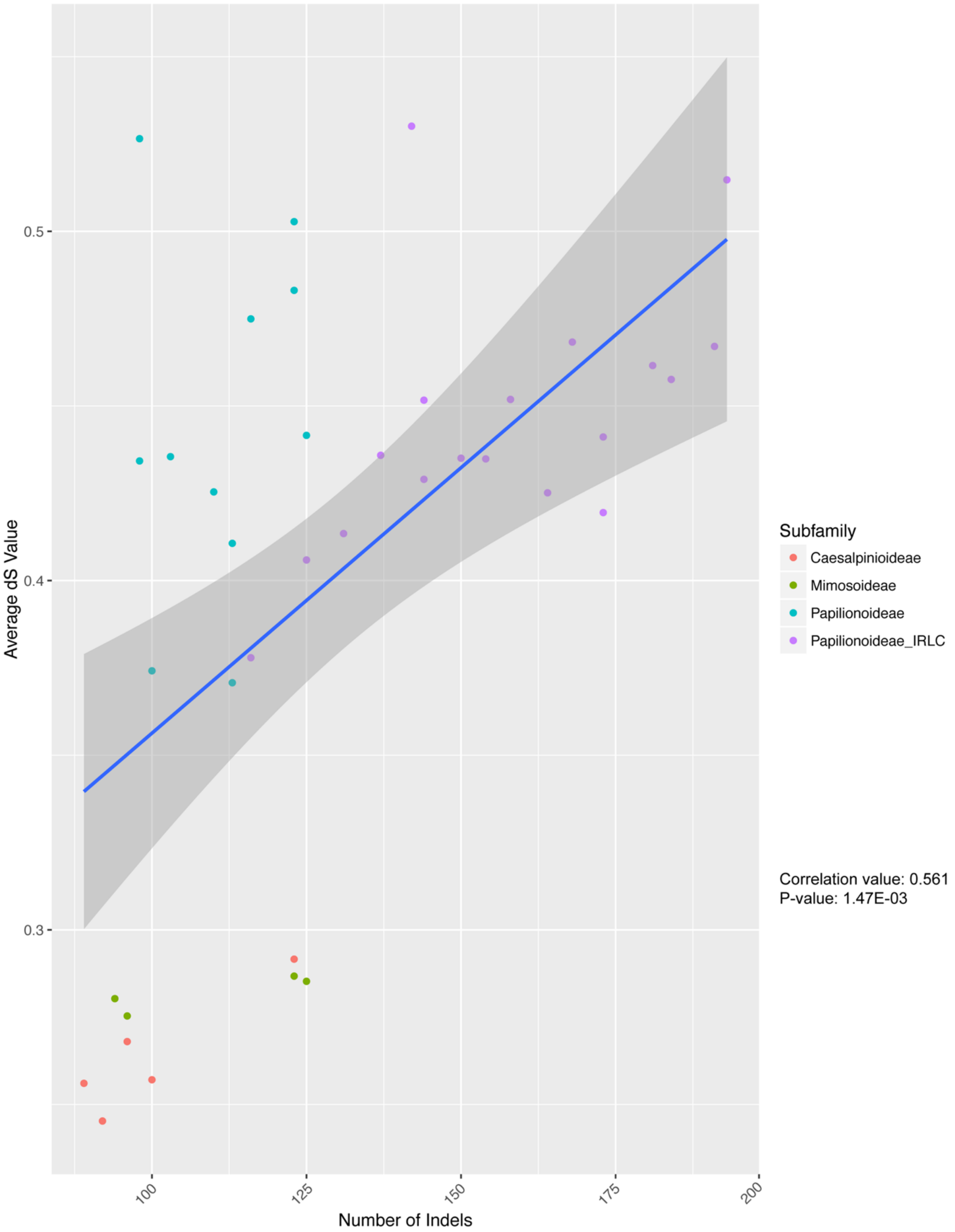

Average dS Value
0.3
0.4
0.5
100
125
150
175
200
Number of Indels
Subfamily
● Caesalpinioideae
● Mimosoideae
● Papilionoideae
● Papilionoideae_IRLC
Correlation value: 0.561
P-value: 1.47E-03

Figure 3.14. Correlation between indel count and *dS* values.

Scatterplot with regression line (blue line) of average *dS* values for each genome compared to number of indels. The color of each point represents the subgroup to which it belongs: Caesalpiniodeae (red), Mimosoideae (green) Papilionoideae (blue), Papilionoideae IRLC (purple). The grey region surrounding the regression line represents the standard error. Correlation value = 0.561, p-value = 1.47E-03.

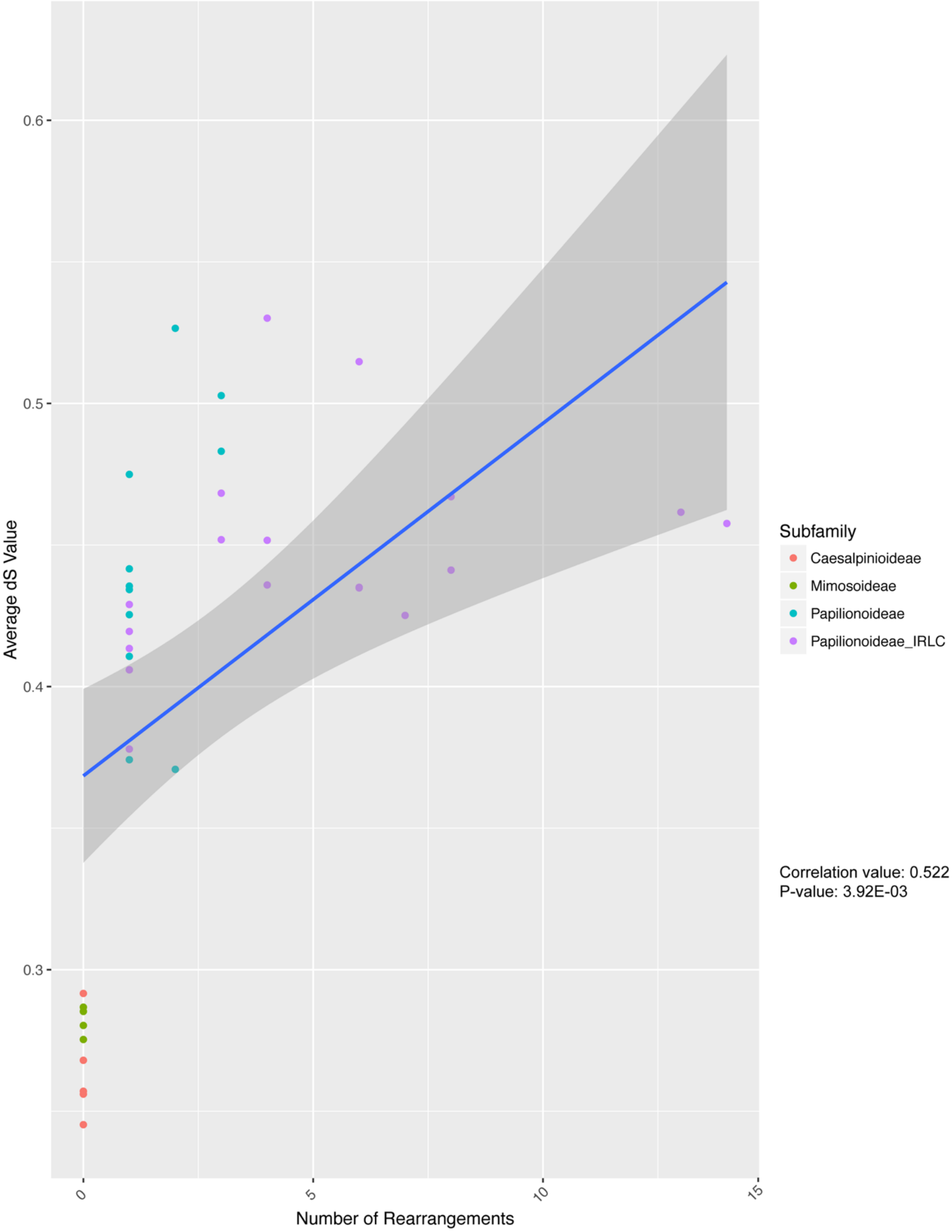

Average dS Value
0.3
0.4
0.5
0.6
0
5
10
15
Number of Rearrangements
Subfamily
● Caesalpinioideae
● Mimosoideae
● Papilionoideae
● Papilionoideae_IRLC
Correlation value: 0.522
P-value: 3.92E-03

Figure 3.15. Correlation between rearrangements and *dS* values.

Scatterplot with regression line (blue line) of average *dS* values for each genome compared to number of rearrangements. The color of each point represents the subgroup to which it belongs: Caesalpiniodeae (red), Mimosoideae (green) Papilionoideae (blue), Papilionoideae IRLC (purple). The grey region surrounding the regression line represents the standard error. Correlation value = 0.522, p-value = 3.92E-03.

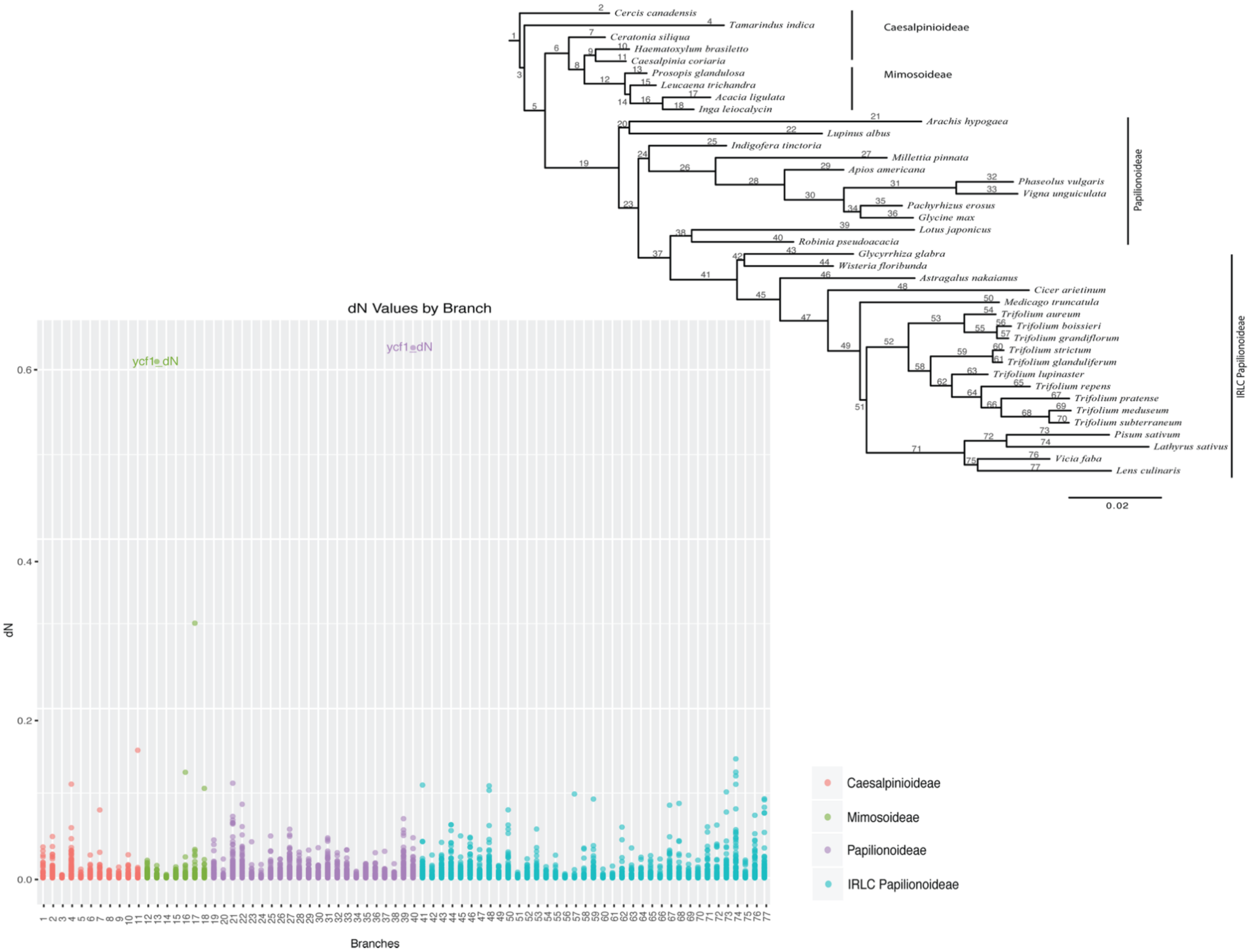

Cercis canadensis
Tamarindus indica
Ceratonia siliqua
Haematoxylum brasiletto
Caesalpinia coriaria
Prosopis glandulosa
Leucaena trichandra
Acacia ligulata
Inga leiocalycin
Arachis hypogaea
Lupinus albus
Indigofera tinctoria
Millettia pinnata
Apios americana
Phaseolus vulgaris
Vigna unguiculata
Pachyrhizus erosus
Glycine max
Lotus japonicus
Robinia pseudoacacia
Glycyrrhiza glabra
Wisteria floribunda
Astragalus nakaianus
Cicer arietinum
Medicago truncatula
Trifolium aureum
Trifolium boissieri
Trifolium grandiflorum
Trifolium strictum
Trifolium glanduliferum
Trifolium lupinaster
Trifolium repens
Trifolium pratense
Trifolium meduseum
Trifolium subterraneum
Pisum sativum
Lathyrus sativus
Vicia faba
Lens culinaris
Caesalpinioideae
Mimosoideae
Papilionoideae
IRLC Papilionoideae
0.02
dN Values by Branch
ycf1_dN
ycf1_dN
0.6
0.4
0.2
0.0
dN
Branches
Caesalpinioideae
Mimosoideae
Papilionoideae
IRLC Papilionoideae

Figure 3.16. Plot of *dN* values for each gene by branch.

Point plot representing *dN* values of all genes for each branch. Each point represents the *dN* value of one gene. The subgroups are represented by each color: Caesalpiniodeae (red), Mimosoideae (green) Papilionoideae (purple), Papilionoideae IRLC (blue). Data points with the highest values are labeled with their gene name. Branch numbers on the point plot correlate to the branch labels on the phylogeny in upper right, which is taken from Figure 1 with the outgroup taxa removed.

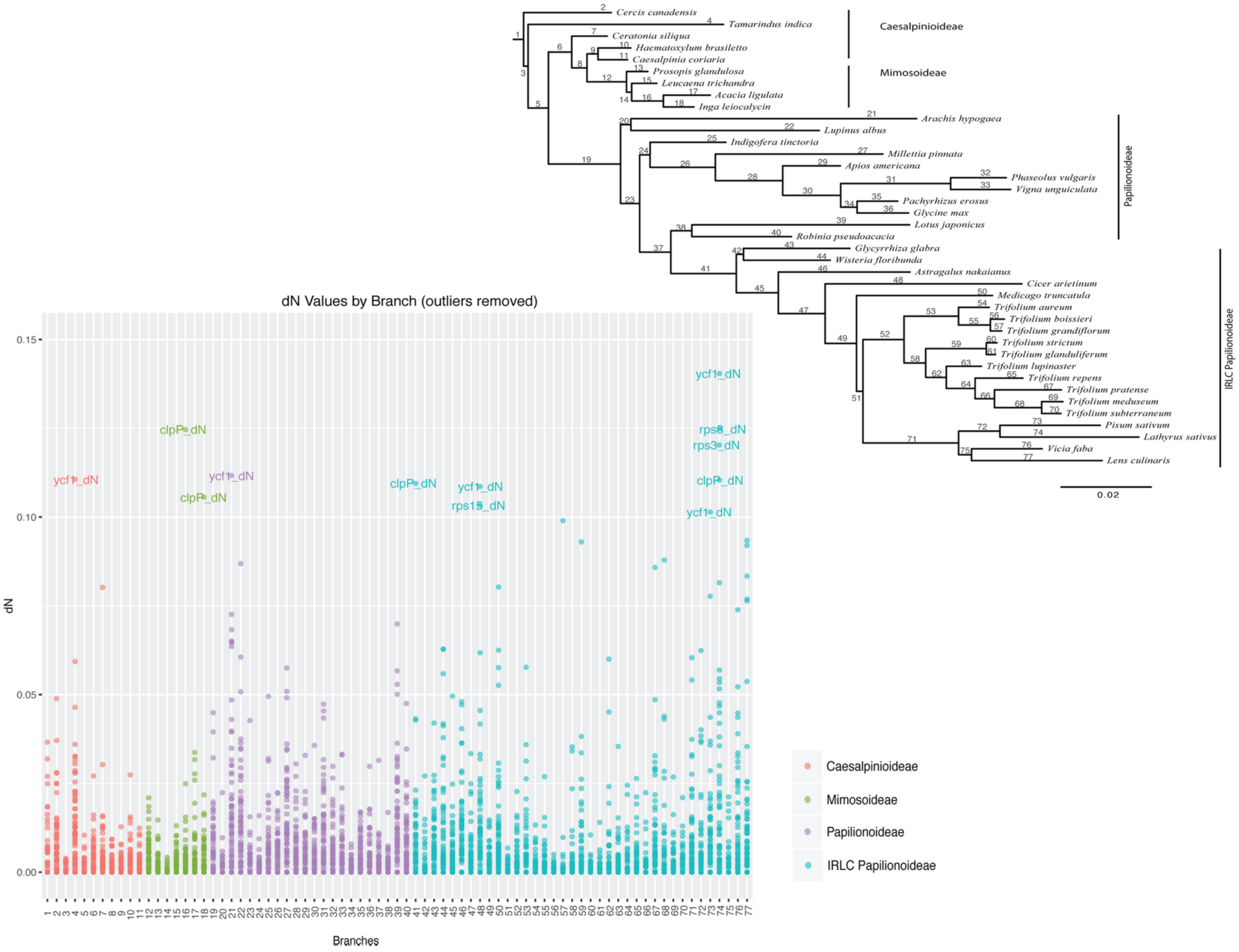

Cercis canadensis
Tamarindus indica
Ceratonia siliqua
Haematoxylum brasiletto
Caesalpinia coriaria
Prosopis glandulosa
Leucaena trichandra
Acacia ligulata
Inga leiocalycin
Arachis hypogaea
Lupinus albus
Indigofera tinctoria
Millettia pinnata
Apios americana
Phaseolus vulgaris
Vigna unguiculata
Pachyrhizus erosus
Glycine max
Lotus japonicus
Robinia pseudoacacia
Glycyrrhiza glabra
Wisteria floribunda
Astragalus nakaianus
Cicer arietinum
Medicago truncatula
Trifolium aureum
Trifolium boissieri
Trifolium grandiflorum
Trifolium strictum
Trifolium glanduliferum
Trifolium lupinaster
Trifolium repens
Trifolium pratense
Trifolium meduseum
Trifolium subterraneum
Pisum sativum
Lathyrus sativus
Vicia faba
Lens culinaris
Caesalpinioideae
Mimosoideae
Papilionoideae
IRLC Papilionoideae
0.02
dN Values by Branch (outliers removed)
dN
0.15
0.10
0.05
0.00
Branches
ycf1_dN
clpP_dN
ycf1_dN
clpP_dN
clpP_dN
ycf1_dN
rps15_dN
ycf1_dN
rps8_dN
rps3_dN
clpP_dN
ycf1_dN
Caesalpinioideae
Mimosoideae
Papilionoideae
IRLC Papilionoideae

Figure 3.17. Plot of *dN* values for each gene by branch with outliers removed.

Point plot representing *dN* values of all genes for each branch cutoff at 0.15. Each point represents the *dN* value of one gene. The subgroups are represented by each color: Caesalpiniodeae (red), Mimosoideae (green) Papilionoideae (purple), Papilionoideae IRLC (blue). Data points above 0.10 are labeled with their gene name. Branch numbers on the point plot correlate to the branch labels on the phylogeny in upper right, which is taken from Figure 1 with the outgroup taxa removed.

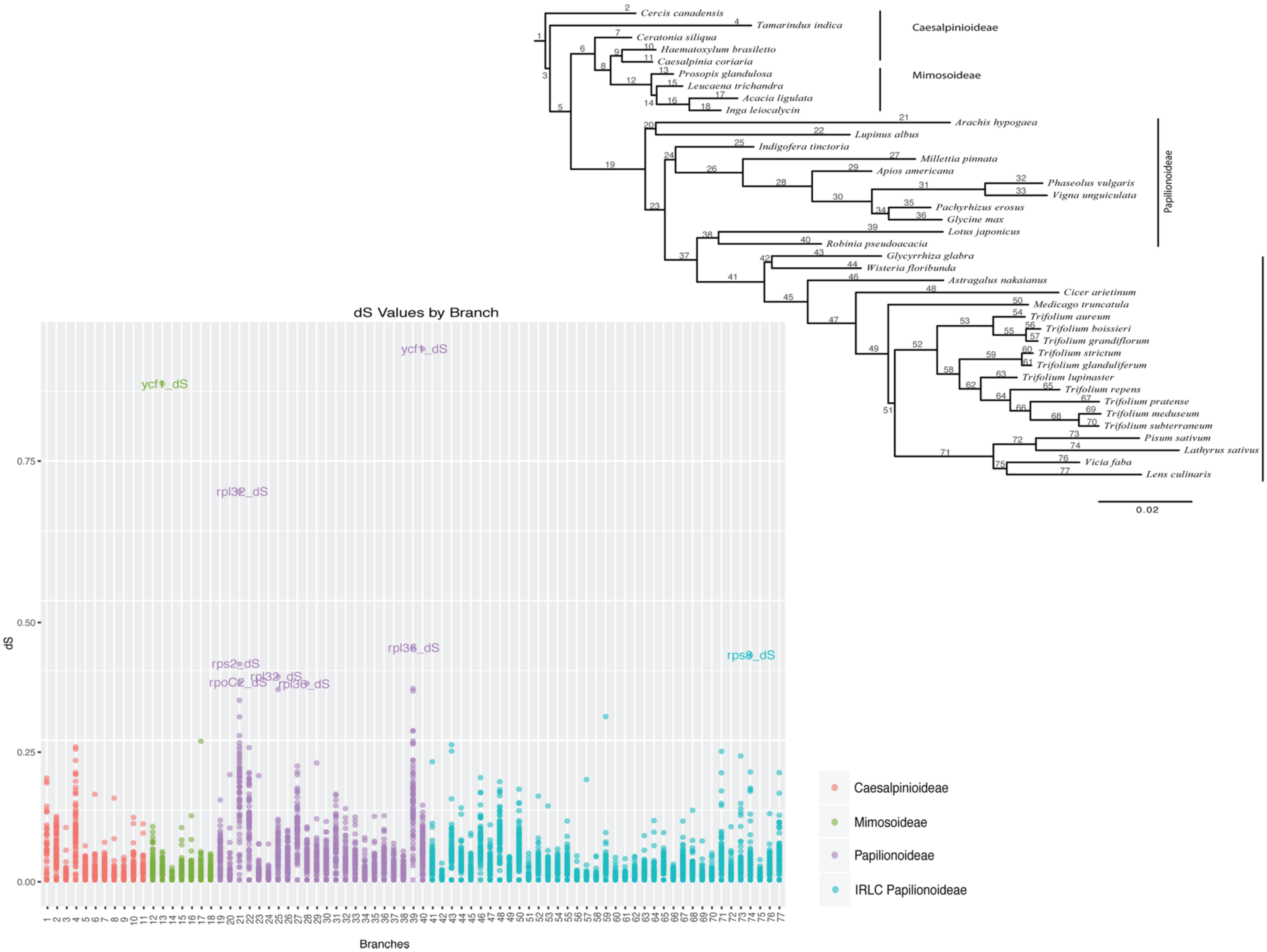

1
2 Cercis canadensis
3
4 Tamarindus indica
5
6
7 Ceratonia siliqua
8
9
10 Haematoxylum brasiletto
11 Caesalpinia coriaria
12
13 Prosopis glandulosa
14
15 Leucaena trichandra
16
17 Acacia ligulata
18 Inga leiocalycin
19
20
21 Arachis hypogaea
22 Lupinus albus
23
24
25 Indigofera tinctoria
26
27 Millettia pinnata
28
29 Apios americana
30
31
32 Phaseolus vulgaris
33 Vigna unguiculata
34
35 Pachyrhizus erosus
36 Glycine max
37
38
39 Lotus japonicus
40 Robinia pseudoacacia
41
42
43 Glycyrrhiza glabra
44 Wisteria floribunda
45
46 Astragalus nakaianus
47
48 Cicer arietinum
49
50 Medicago truncatula
51
52
53
54 Trifolium aureum
55
56 Trifolium boissieri
57 Trifolium grandiflorum
58
59
60 Trifolium strictum
61 Trifolium glanduliferum
62
63 Trifolium lupinaster
64
65 Trifolium repens
66
67 Trifolium pratense
68
69 Trifolium meduseum
70 Trifolium subterraneum
71
72
73 Pisum sativum
74 Lathyrus sativus
75
76 Vicia faba
77 Lens culinaris
Caesalpinioideae
Mimosoideae
Papilionoideae
IRLC Papilionoideae
0.02
dS Values by Branch
ycf1_dS
ycf1_dS
rpl32_dS
rps2_dS
rpl36_dS
rps8_dS
rpoC2_dS
rpl32_dS
rpl36_dS
0.75
0.50
0.25
0.00
dS
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77
Branches
● Caesalpinioideae
● Mimosoideae
● Papilionoideae
● IRLC Papilionoideae

Figure 3.18. Plot of *dS* values for each gene by branch.

Point plot representing *dS* values of all genes for each branch. Each point represents the *dS* value of one gene. The subgroups are represented by each color: Caesalpiniodeae (red), Mimosoideae (green) Papilionoideae (purple), Papilionoideae IRLC (blue). Data points above 0.30 are labeled with their gene name. Branch numbers on the point plot correlate to the branch labels on the phylogeny in upper right, which is taken from Figure 1 with the outgroup taxa removed.

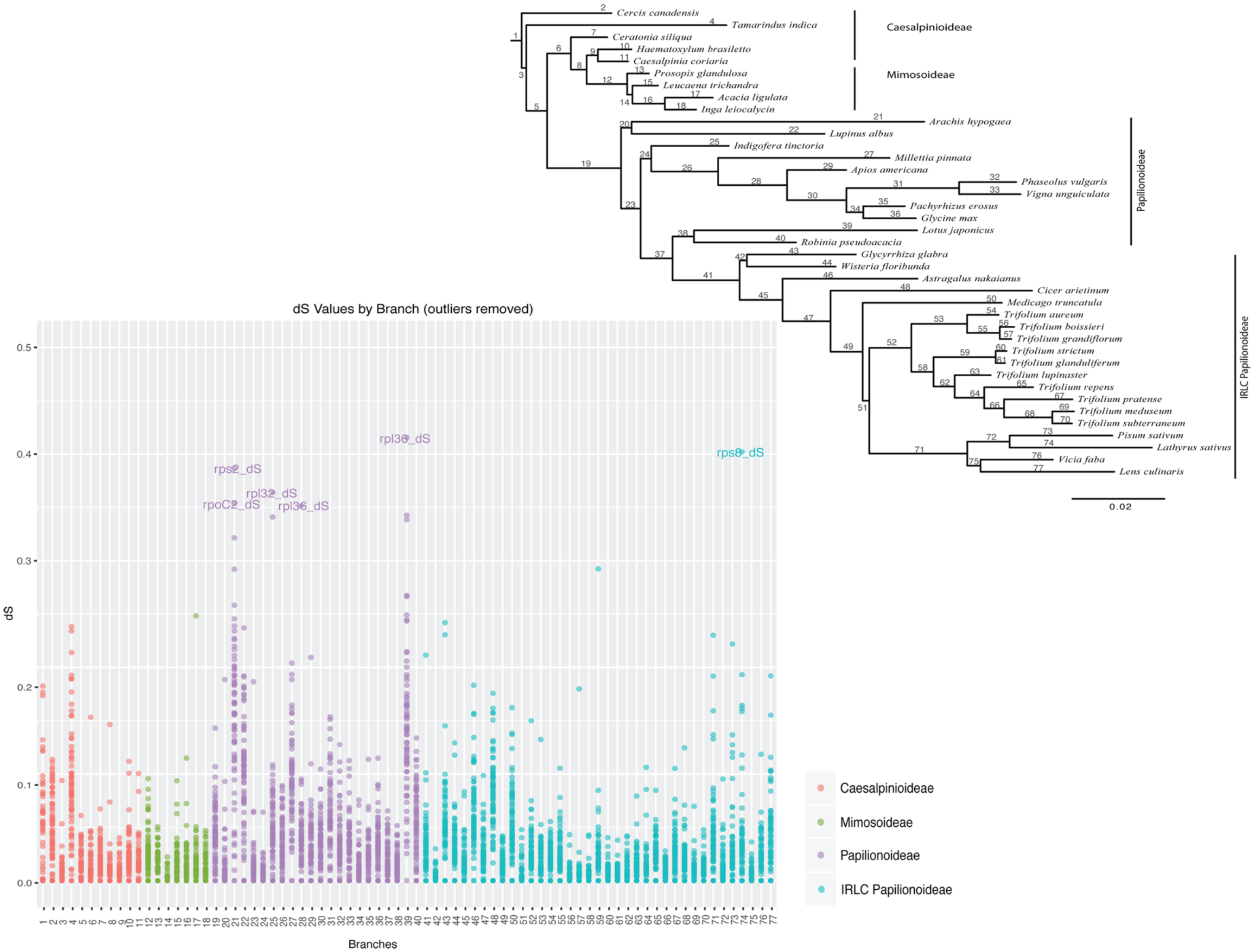

Cercis canadensis
Tamarindus indica
Ceratonia siliqua
Haematoxylum brasiletto
Caesalpinia coriaria
Prosopis glandulosa
Leucaena trichandra
Acacia ligulata
Inga leiocalycin
Arachis hypogaea
Lupinus albus
Indigofera tinctoria
Millettia pinnata
Apios americana
Phaseolus vulgaris
Vigna unguiculata
Pachyrhizus erosus
Glycine max
Lotus japonicus
Robinia pseudoacacia
Glycyrrhiza glabra
Wisteria floribunda
Astragalus nakaianus
Cicer arietinum
Medicago truncatula
Trifolium aureum
Trifolium boissieri
Trifolium grandiflorum
Trifolium strictum
Trifolium glanduliferum
Trifolium lupinaster
Trifolium repens
Trifolium pratense
Trifolium meduseum
Trifolium subterraneum
Pisum sativum
Lathyrus sativus
Vicia faba
Lens culinaris
Caesalpinioideae
Mimosoideae
Papilionoideae
IRLC Papilionoideae
0.02
dS Values by Branch (outliers removed)
dS
Branches
0.0
0.1
0.2
0.3
0.4
0.5
rps2_dS
rpl32_dS
rpoC2_dS
rpl36_dS
rpl36_dS
rps8_dS
Caesalpinioideae
Mimosoideae
Papilionoideae
IRLC Papilionoideae

Figure 3.19. Plot of *dS* values for each gene by branch with outliers removed.

Point plot representing *dS* values of all genes for each branch cutoff at 0.50. Each point represents the *dS* value of one gene. The subgroups are represented by each color: Caesalpiniodeae (red), Mimosoideae (green) Papilionoideae (purple), Papilionoideae IRLC (blue). Data points above 0.35 are labeled with their gene name. Branch numbers on the point plot correlate to the branch labels on the phylogeny in upper right, which is taken from Figure 1 with the outgroup taxa removed.

Chapter 4: Nucleotide substitution rates of legume mitogenomes reveal accelerated rates in Papilionoideae consistent with plastome-wide substitution rates

INTRODUCTION

Mitochondrial genomes (mitogenomes) of angiosperms are the largest compared to other eukaryotic lineages and most fluid in terms of genome structure of any organelle (Mower et al., 2012). Genes in the mitogenome encode products involved in electron transport, ATP synthesis, intron splicing and the translation, maturation and the translocation of proteins (Mower et al., 2012). Gene content is known to vary within angiosperms with some basal angiosperms retaining all 40 ancestral genes while others such as *Lachnocaulon* (Eriocaulaceae) have lost all 14 ribosomal proteins and the succinate dehydrogenase (sdh) genes (Adams et al., 2002). Mitochondrial gene loss is scattered phylogenetically with most variation in gene content involving ribosomal protein genes (Palmer et al., 2000; Adams and Palmer, 2003; Mower et al., 2012). The driving force behind many mitochondrial gene losses is RNA-mediated functional transfer to the nucleus whereby a copy of a mitochondrial mRNA is reverse transcribed and integrated into the nucleus where it gains function and eventually either the nuclear or mitochondrial copy is silenced or lost (Adams et al., 1999; Adams et al., 2001). The most interesting examples of functional transfer intermediates are found in *cox2* within legumes in which most species (except *Vigna*) have both nuclear and mitochondrial copies but one or the other has been inactivated (Covello and Gray, 1992; Adams et al., 1999; Palmer et al., 2000).

While mitogenomes change rapidly in organization, nucleotide substitution rates of mitochondrial genes in plants are the slowest compared to other groups such as animals and to other organelles (Wolfe et al., 1987; Palmer and Herbon, 1988). Mitochondrial genes within plants evolve 10-20 times slower than the nuclear genes, three times slower than the plastid and 40-100 times slower than mammalian mitochondrial genes (Wolfe et al., 1987). However, exceptions to this ratio are found within Geraniaceae, Plantaginaceae, Lamiaceae and Caryophyllaceae. Palmer et al. (2000) surveyed 281 angiosperms for gene loss using Southern blot analyses and showed *Pelargonium hortorum* (Geraniaceae) and *Plantago rugelii* (Plantaginaceae) to be highly divergent in their mitochondrial genes due to the lack of hybridization. Nucleotide substitution rates for both *Pelargonium* and *Plantago* have been explored further and accelerated rates of mitochondrial genes have been shown to extend to multiple genera within the Geraniaceae (Parkinson et al., 2005; Bakker et al., 2006; Weng et al., 2012) and Plantaginaceae (Cho et al., 2004; Bakker et al., 2006). Additionally, *Silene* (Caryophyllaceae) is known to have extremely high levels of mitochondrial sequence divergence; however, its case is exceptional because of an 8-fold difference in substitution rates between species even when the fastest evolving species, *Silene noctiflora*, is removed from the comparisons (Mower et al., 2007; Sloan et al., 2008). More recently Zhu et al. (2014) found extreme synonymous rate heterogeneity (up to 340-fold) within the mitogenome of *Ajuga reptans* (Lamiaceae).

While genes in the mitochondria, plastid and nucleus evolve at very different rates in plants, typically in a 1:3:16 ratio, respectively (Wolfe et al., 1987), nucleotide substitution rates are generally correlated across all three genomes (Eyre-Walker and Gaut, 1997). Studies comparing substitution rates of grasses and palms have shown elevated synonymous rates of a plastid (*rbcL*), mitochondrial (*atp1*) and nuclear gene (*Adh*) in grasses (Bousquet et al., 1992; Gaut et al., 1992; Gaut et al., 1996; Eyre-Walker and Gaut, 1997). Due to the extreme acceleration in mitochondrial genes in *Silene*, Sloan et al. (2012) sequenced four *Silene* plastomes to compare substitution rates and genomic rearrangements between the two genomes. They found the two species, *S. noctiflora* and *S. conica*, with fast evolving mitochondrial genes also had accelerated rates in a subset of plastid genes in addition to high levels of rearrangement in the plastome. Similarly, *S. latifolia* and *S. vulgaris*, which have slower rates in the mitochondrial genes, also have correspondingly low rates and unrearranged plastomes. Geraniaceae also exhibits a high number of plastid genomic rearrangements and accelerated rates of plastid genes (Chumley et al., 2006; Guisinger et al., 2008, 2011; Blazier et al., 2011; Weng et al., 2012, 2014; Blazier et al., 2016b) in addition to accelerated rates of mitochondrial genes (Parkinson et al., 2005; Weng et al., 2012). Accelerated rates in *Plantago* genes seems to be limited to the mitochondrion as rate analyses of two plastid genes, *rbcL* and *ndhF*, revealed limited variation in *Plantago* compared to other taxa (Cho et al., 2004). While the lack of rate variation in the plastid of *Plantago* may be due to the fact that only two genes were examined, in the case of *Ajuga* 78 and 27 protein coding genes were utilized from the plastid and mitochondrion, respectively, to compare substitution rates (Zhu et al., 2014) and while mitochondrial rates were increased, there was no increase in plastid rates, uncoupling the correlation between plastid and mitochondrial rates.

Within some legumes, especially papilionoids, high levels of plastid rearrangements are well documented (Cai et al., 2008, Sabir et al., 2014, Schwarz et al., 2015; Sveinsson and Cronk, 2014). Studies of rate heterogeneity in legumes have been restricted to a few taxa and largely focused on plastid genes relative to the presence/absence of the inverted repeat (IR) (Wolfe et al., 1987; Perry and Wolfe, 2002), or single, highly divergent genes such as *ycf4* (Magee et al., 2010) and *clpP* (Dugas et al., 2015; Williams et al., 2015). However, chapter three explored plastid rate heterogeneity on a broader scale across all three subfamilies and in comparison to biological features such as genome size, genome rearrangements and growth habit in order to uncover trends in legume plastid gene evolution. Given the large amount of information that is now available on legume plastid organization and substitution rates, investigating rates of evolution of mitochondrial genes may provide insights into the causes of evolutionary changes in both organellar genomes.

To date, there are currently seven mitogenomes publicly available on GenBank (https://www.ncbi.nlm.nih.gov/). In this study, we present sequence data for 19 draft legume mitogenomes across the entire family to investigate: 1) mitochondrial gene content in legumes, 2) acceleration of rates of individual mitochondrial genes or functional groups of genes, 3) lineage specific variation of rates, and 4) comparison of rates in the mitochondria and plastids. This is the most comprehensive investigation of mitochondrial nucleotide substitution rates in the legumes both in terms of taxon sampling and number of genes examined.

MATERIALS AND METHODS

Taxon sampling, contig assembly and mitochondrial gene identification

Illumina reads from genomic DNA previously generated for 19 species of legumes (Sabir et al., 2014; Schwarz et al., 2015) were assembled with 200X coverage using a range of kmer sizes (71, 73, 75, 77) and scaffolding turned off with Velvet (Zerbino and Birney, 2008). Contigs from all assemblies were imported into Geneious version 7.1.9 (Biomatters Ltd., http://www.geneious.com/). A database of mitochondrial protein-coding genes comprising closely related legume sequences (Table 4.1) was employed to identify mitochondrial genes in contigs from each assembly. Illumina reads were mapped to reference genes for each gene that could not be found in assembled contigs using Bowtie2 (Langmead and Salzberg, 2012). An additional eight mitochondrial genomes (seven legumes and one outgroup, *Populus tremula*) publicly available on GenBank (http://www.ncbi.nlm.nih.gov/genbank/) were also utilized (Table 4.1).

Plastid genes common to the same 27 species were extracted from plastomes generated for previous studies (Sabir et al., 2014; Schwarz et al., 2015) and from NCBI (http://www.ncbi.nlm.nih.gov/genbank/) (Table 4.1).

Sequence alignment and phylogenetic analyses

Twenty-six mitochondrial genes and 70 plastid genes (Table 4.2) present in all 27 species (26 legumes and one outgroup) were extracted and aligned using the translation align tool in Geneious with default MAFFT (Katoh and Standley, 2013) settings. Alignments were manually edited to improve quality by ensuring that indels were maintained in groups of three to retain the reading frame. In addition to individual gene alignments, three concatenated alignments were generated: 1) 70 plastid genes, 2) 26 mitochondrial genes and 3) all 96 plastid and mitochondrial genes. Ambiguous and poorly aligned regions were removed using Gblocks (http://molevol.cmima.csic.es/castresana/Gblocks_server.html). Maximum likelihood trees were generated using RaxML Blackbox (Stamatakis *et al*. 2008) on CIPRES (Miller et al., 2010) with the gamma model of rate heterogeneity, rapid bootstrapping and the "auto" setting, which determines when there are a sufficient number of replicates. The tree with the maximum likelihood value was imported into FigTree version 1.4.2 (http://tree.bio.ed.ac.uk/software/figtree/).

Nucleotide substitution rates

Nonsynonymous (*dN*) and synonymous (*dS*) nucleotide substitution rates for each of 96 (26 mitochondrial, 70 plastid) protein-coding genes were estimated using the codeml program in PAML 4.5 (Yang 2007). Codon frequencies were determined by the F3 x 4 model. Transition/transversion and *dN*/*dS* ratios were estimated with the initial values of 2 and 0.4, respectively. Two analyses were run: 1) runmode = 0, model = 1 in which the mitochondrial phylogeny generated by RAxML was used as a constraint tree, and 2) runmode = -2, model = 1, for pairwise rate comparisons between each legume taxon and *Populus tremula*, the outgroup species.

Detection of RNA editing sites

RNA editing sites were predicted using PREP-Mt in the Predictive RNA Editor for Plants (PREP) Suite (Mower, 2005, 2009). Individual sequences for each gene were submitted and only those predicted edit sites that had a confidence interval of ≥ 0.50 were counted. An average number of editing sites for each gene was calculated across all species. Correlation testing between *dS* and *dN* values and number of RNA editing sites was performed in R (v3.2.2) using the rcorr function in the Hmisc package using the Pearson method. Boxplots and scatterplots were generated using the ggplot package in R (v3.2.2). One outlier, *atp9*, was removed from the scatterplot of *dS* versus RNA editing sites in order to visualize the trend but this outlier was not left out of the correlation testing.

Detection of rate acceleration

Values of *dN* and *dS* for each branch were plotted using the ggplot package in R (v3.2.2). Pairwise *dN* and *dS* values for all genes across all 26 legume species in addition to the average values for each legume species were also plotted using the ggplot package in R (v3.2.2). The Wilcoxon rank sum test (wilcox.test, paired=FALSE) in R (v3.2.2) was used to test significance between woody versus herbaceous habit. P-values were corrected using the p.adjust function in R with the Holm method.

Comparison of rates of mitochondrial and plastid genes

Using *dN* and *dS* values from the pairwise PAML analysis, *dN* and *dS* values across all genes from both genomes were averaged for each species. Boxplots were generated using the ggplot package in R (v3.2.2). Significance was tested using the Wilcoxon rank sum test (wilcox.test, paired=FALSE) in R (v3.2.2). The Spearman correlation test was performed using the rcorr function in the Hmisc package in R (v3.2.2) to evaluate any relationship between *dN* and *dS* values of each genome.

RESULTS

Phylogenetic analysis

Phylogenetic analyses were performed using three different datasets: 1) plastid only, 2) mitochondrial only and 3) plastid and mitochondria combined. The maximum likelihood tree (-ln = -61835.918823) generated with 27 taxa and 26 protein coding genes from the mitogenome (Figure 4.1) was the most congruent topology with recent legume phylogenies (Wojciechowski et al., 1994; LPWG, 2013) and was therefore selected as the constraint tree. The rapid bootstrap search terminated after 360 replicates. There was strong support (> 90% bootstrap values) for all but two nodes in the tree (Figure 4.1).

Mitochondrial gene content

Of 32 mitochondrial genes known to be shared by most angiosperms (Mower et al., 2012), 26 were detected in all legumes and the outgroup (Table 4.3). Two genes, *rpl5* and *rps10*, were lost from the mitochondrial genome of the outgroup *Populus tremula*. Additionally, four genes (i.e., *cox2*, *rps1*, *rps3*, *rps14*) were putatively lost at least once within legumes (Table 3).

Rates of nucleotide substitutions of mitochondrial genes

Twenty six mitochondrial genes present in all 27 species examined, the outgroup *Populus* and 26 legumes, were utilized in rate comparisons (Tables 4.1 – 4.2). Mean values of *dN* ranged from 0.012 in *nad*5 to 0.101 in *atp8* (Figure 4.2). The genes in the ATP synthase functional group were the most variable in terms of *dN* values and also contained members with the highest *dN* values among all the mitochondrial genes (*i.e.*, *atp4* and *atp9*). The remaining functional groups have more uniform *dN* values compared to the ATP synthase genes with two exceptions:

nad3 and *rps4*. Values of *dS* showed a more pronounced trend than *dN* values with ATP synthase genes having highly variable *dS* values. *atp9* had the highest mean *dS* value (1.047) overall and *nad7* had the lowest mean value at 0.028 (Figure 4.2).

Lineage specific rates

Substitution rates for each lineage were calculated using the mitochondrial constraint tree. The branch leading to the legumes had the highest number of accelerated genes for both *dN* and *dS* (Figures 4.3 - 4.4). Values of *dN* were relatively uniform across all branches with the exception of a few lineages with accelerated genes. *atp4* was accelerated in the branches leading to the legumes, *Cercis canadensis* and *Ceratonia siliqua*. The only mimosoid included in this study, *Prosopis glandulosa,* showed accelerated *dN* values for *atp8* and *nad4L* genes. Within the papilionoids there were a few accelerated genes including *atp6*, *atp8* and *nad9. Trifolium repens* exhibited accelerated *dN* values in multiple genes that were not accelerated in the remaining legumes, including *mttB*, *ccmC* and *nad4*. Values of *dS* were also relatively homogeneous across all branches with a few exceptions. There was an acceleration of *dS* in multiple genes in the branch leading to the legumes. *dS* of *atp9* was accelerated in the branches leading to *Arachis hypogaea* and *Prosopis glandulosa*, and *atp6* was accelerated in the branch leading to *Medicago truncatula.* These values along with *atp1* in the branch leading to all legumes were higher than all other values of *dS* within legumes (Figure 4.4).

Pairwise *dN* and *dS* values were averaged across all genes in each species and plotted (Figure 4.5). In *Prosopis glandulosa* there was a noticeable increase in *dS* and, and to a much lesser extent, in *dN*. Similarly, values of both *dN* and *dS* were accelerated in *Arachis hypogaea, Medicago truncatula and Trifolium repens*. Values of *dS* were approximately 1.6 times higher in papilionoid taxa (mean = 0.131) compared with caesalpinioid and mimosoid taxa (mean = 0.080). In contrast, *dN* values only differed by approximately 1.1 times in the papilionoids (mean = 0.047) compared to caesalpinioids and mimosoids (mean = 0.041). In addition to comparing the differences between the subfamilies, there was a significant increase in both *dN* (p-value = 6.27E-03) and *dS* (p-value = 4.86E-05) values in those taxa that have a herbaceous versus woody habit.

Frequency of RNA editing in legumes

RNA editing sites were detected in all legume mitochondrial genes with the exception of five genes (*atp1, atp6, atp8, cox1, mttB*) belonging to *Trifolium repens* (Table 4.4). Overall, the lowest levels of editing were predicted in *atp1*, *atp8* and *atp9* with an average of 2-3 editing sites across the legumes (Table 4.4, Figure 4.6). The NADH dehydrogenase genes consistently had the highest level of editing with the most predicted sites (average of 39.5) in *nad4*. A significant ($p < 0.05$) negative correlation between the number of RNA editing sites and *dN* and *dS* was detected (Figure 4.7).

Mitochondrial versus plastid rates

Rates of both mitochondrial and plastid protein-coding genes were calculated and rates of each gene from each genome were averaged across legume species. Mean *dN* and *dS* values of the plastid genes were approximately 1.3 and 3.8 times higher than those of the mitochondria, respectively (Table 4.5). Overall, genes in the plastid were significantly accelerated in both of *dN* (p-value = 3.143E-05) and *dS* (p-value = 4.033E-15) compared to the mitochondrial genome (Figure 4.8, Table 4.5). Across legumes there was a trend of increasing *dN* values from the basal caesalpinioid and mimosoid lineages to the IRLC papilionoids in both genomes, although it was much more pronounced in the plastid; however, there was a decrease in rates in *Robinia pseudoacacia*, *Lotus japonicus* and *Glycyrrhiza glabra* (Figure 4.9). Values of *dS* exhibited a marked elevation in the papilionoid taxa for both mitochondrial and plastid genes (Figure 4.9).

Correlation between *dN* and *dS* values of each genome was calculated using the Spearman correlation test. The highest (0.783) and most significant (p-value = 2.24E-06) correlation was between *dS* of the mitochondrion and plastid. There was also a significant positive correlation (0.683, p-value = 1.19E-04) between *dN* for both genomes (Figure 4.10).

Discussion

This study represents the most comprehensive nucleotide substitution rate comparison of organellar genes in legumes. We utilized newly and previously generated sequence data to analyze substitution rates of 96 mitochondrial and plastid genes for 26 legume species representing all four major clades in the family. Legumes are an excellent group to study nucleotide substitution rates because it is a large family with variation in several biological features such as growth habit and species richness (Wojciechowski et al., 2004; Bruneau et al., 2008; LPWG, 2013). Additionally, plastid rate heterogeneity within legumes has been examined previously (Wolfe et al., 1987; Perry and Wolfe, 2002; Magee et al., 2010; Dugas et al., 2015; Williams et al., 2015; chapter 3) and provides an excellent framework to compare rate heterogeneity between multiple organellar genomes on a family-wide scale. We found four mitochondrial genes putatively lost within the legumes. Mitochondrial *dS* values were more than 1.5 times faster in the papilionoid lineage compared with the caesalpinioid and mimosoid lineages combined. Values of both *dN* and *dS* were also accelerated in a number of ATP synthase genes and several genes from other functional groups. When comparing overall rates of genes from the mitogenome and the plastome, *dS* was 3.8 times higher in the plastid compared to the mitochondrion, whereas nonsynonymous values were only 1.3 times higher in the plastid. The following discussion will focus on four topics: 1) mitochondrial gene losses in legumes, 2) intragenomic rate heterogeneity in the mitogenome, 3) acceleration of rates in mitochondrial genes in papilionoid lineages and 4) comparison of rates between the mitochondrial and plastid genomes.

Mitochondrial gene losses in legumes

Four protein coding genes (i.e., *cox2*, *rps1*, *rps3* and *rps14*) are putatively missing from the mitogenomes of the legumes examined (Table 4.3). Because mitogenomes were not completed for these species, it is possible that the missing genes may be due to low sequence coverage. However, there are two reasons that make it likely that these four genes have been lost. First, we were able to detect all other protein coding genes for these same species, which suggests that the depth of coverage is sufficient to detect any genes that are present. Second, previous studies have shown that these four genes have been lost in one or multiple lineages of angiosperms, including legumes, either by complete loss, substitution or functional transfer to the nucleus (Brandvain and Wade, 2009). The loss of *cox2* from angiosperm mitochondrial genomes is restricted to *Vigna* within legumes, which was originally suggested by southern hybridization studies (Nugent and Palmer, 1991; Adams et al., 2002) and later verified with sequencing of the *V. radiata* (Alverson et al., 2011) and *V. angularis* (Naito et al., 2013) mitogenomes. Our data show that the loss of *cox2* also occurs in *V. unguiculata*. The functional transfer of *cox2* to the nucleus occurred between 60 and 200 million years ago, however the mitochondrial copy still remains in most lineages with the only loss in angiosperms being reported in the *Vigna* lineage (Nugent and Palmer 1991). Covello and Gray, (1992) characterized a functional nuclear *cox2* gene in *Glycine*, a closely related legume, showing an intermediate stage in the functional transfer in which the nuclear copy is expressed but the mitochondrial copy is not expressed. Adams et al. (1999) were able to show that in some angiosperms both the nuclear and mitochondrial copies of *cox2* are functional while in other species only one of the copies is functional.

The remaining three mitochondrial genes (i.e., *rps1*, *rps3*, *rps14*) lost in legumes are ribosomal protein coding genes, which are known to be lost frequently across angiosperms (Adams and Palmer, 2003). Adams et al. (1999) surveyed 40 mitochondrial genes across 280 genera of flowering plants and found losses of ribosomal protein genes are much more common than genes belonging to other functional groups. They found 33, 7 and 27 losses of *rps1*, *rps3* and *rps14*, respectively, across angiosperms and the losses of other ribosomal genes show a similar frequency of loss. *rps1* is missing in four legume taxa examined in this study, *Lotus japonicus*, *Trifolium aureum*, *T. grandiflorum* and *T. meduseum*. The loss of *rps1* in *L. japonicus* was reported previously by Kazakoff et al. (2012) based on complete mitogenome sequencing, and it has been functionally transferred to the nucleus. Frequent loss of ribosomal protein genes is not limited to mitochondria. Plastid genomes also exhibit many ribosomal protein gene losses across angiosperms with 11 plastid ribosomal protein genes lost once or multiple times (Jansen et al. 2007). Within the legumes *rpl22* (Gantt et al. 1991; Doyle et al. 1995), *rpl33* (Guo et al., 2007; Schwarz et al., 2015) and *rps16* (Guo et al., 2007; Cai et al. 2008; Jansen et al. 2007; Magee et al. 2010; Sabir et al. 2014; Schwarz et al., 2015) have been lost from the plastid genome once (*rpl22*, *rpl33*) or multiple (*rps16*) times.

Rate variation in legume mitochondrial genes

Values of *dN* are elevated in *atp1, atp4, atp6* and *atp8*, *nad3* and rps4 compared with other genes (Figure 4.2). Also, *dS* values are elevated in *rps12, atp1, atp4* and *atp9* with the latter gene having especially high rates of change. Mitochondrial rate heterogeneity has been documented in a number of angiosperm species. The most extreme example is the 340-fold rate acceleration in Lamiaceae genus *Ajuga* (Zhu et al., 2014). While the differences in rates are not so extreme in legumes, many of the same genes that show rate heterogeneity in both *dN* and *dS* among legumes (i.e., *rps12* and *atp4, atp6, atp*8 and *atp9*) are the same ones detected in other angiosperm lineages (Adams and Palmer, 2003). Rate variation between genes can be explained by three different mechanisms: 1) localized hypermutation, 2) RNA editing or 3) mutagenic retroprocessing. In the plastid genome, Magee et al. (2010) identified a region surrounding *ycf4* that is a hotspot for point mutations hypothesized to be the result of repeated DNA breakage and repair. A mutational hotspot requires genes to be in close proximity in the genome, and while a few of the mitochondrial genes with accelerated rates in legumes (e.g., *nad3* and *rps12*) are near each other in *Vicia faba*, *V. radiata* and *Milletia pinnata*, most accelerated genes in legumes are not close to each other. RNA editing is a common cause of divergence in mitochondrial rates (Lu et al., 1998). Values of *dN* may be overestimated due to RNA editing but *dS* values are rarely affected (Mower et al., 2007). A survey for RNA editing sites within legumes revealed the lowest number of editing sites in genes that have the higher *dN* and *dS* values (Figures 4.2 - 4.6) suggesting that RNA editing is not likely responsible for the rate heterogeneity. This negative correlation is also seen in Geraniaceae (Parkinson et al., 2005), *Silene* (Sloan et al., 2010) and in a family of monocots (Cuenca et al., 2010). Lastly, a process referred to as mutagenic retroprocessing by Parkinson et al. (2005) explains rate heterogeneity among mitochondrial genes. This mechanism involves exceptionally high levels of reverse transcription in combination with homologous recombination (Parkinson et al., 2005; Bakker et al., 2006). This process was invoked in Geraniaceae and *Plantago*, which have high levels of rate heterogeneity but low levels of RNA editing. In the case of mutagenic retroprocessing a correlation between transcription levels and substitution rates may exist if those genes that are highly transcribed are also retroprocessed more frequently. It is noteworthy that Islam et al. (2013) analyzed mitochondrial gene expression in flower tissues of rye grass and found high normalized expression levels of *rps3*, *rps12*, *rpl16* and the highest levels in ATP synthase genes with 2.5 times higher expression of *atp9*. In view of information, mutagenic retroprocessing is the most likely candidate for the rate variation in mitochondrial genes of legumes. However, studies focusing on transcription levels of mitochondrial genes are needed to explore this mechanism more thoroughly.

A slight rate acceleration in *dS* of *atpH* was demonstrated in the plastid (Chapter 3). *atpH* is a homolog of *atp9,* which also shows highly accelerated rates in the mitochondrion. This correlated rate acceleration pattern of these two genes was also observed in *Ajuga* (Zhu et al., 2014).

Accelerated rates in papilionoids

Papilionoid legumes have 1.1 and 1.6 times higher *dN* and *dS* values than the caesalpinioid and mimosoid taxa combined (Figure 4.5). There are two major differences between caesalpinioids/mimosoids and papilionoids, numbers of species and growth habit. Papilionoids are largely herbaceous and are much larger in terms of numbers of species whereas both caesalpinioids and mimosoids are woody with many fewer species (Wojciechowski et al., 2004; Bruneau et al., 2008; LPWG, 2013). Multiple studies have shown correlations between substitution rates and species diversification (Barraclough et al., 1996; Bousquet et al., 1992) and growth habit (Kay et al., 2006; Laroche et al., 2008; Smith and Donoghue, 2008) in plants. Bromham et al. (2015) examined correlations between a number of factors and substitution rates in all three plant genomes over a range of flowering plants and found a consistent negative correlation between plant height and *dS* in both plastid and mitochondrial genes. However, there was no link between species diversification and *dS* of mitochondrial genes, although there was such a correlation for for plastid rates. We did not test for a correlation between species diversification and nucleotide substitutions rates of mitochondrial genes but in view of the limited sampling of legumes such a comparison would be more appropriate once more extensive taxon sampling is available. We did find that rates are significantly higher in herbaceous versus woody taxa, which supports the generation time hypothesis. This pattern was also observed between for rates of sequence evolution in legume plastomes (Chapter 3).

Rates in mitochondrial genes versus plastid genes

Plastid protein coding genes have 1.3 and 3.8 higher rates than the mitogenome for *dN* and *dS*, respectively (Figure 4.8). This is congruent with the ratio of *dS* of mitochondria and plastid of 1:3 that was previously reported in plants (Wolfe et al., 1987; Drouin et al., 2008). Several previous studies have focused on the correlation of substitution rates between the mitogenome, plastome and nuclear genome and have found levels of rate heterogeneity are often correlated between all three genomes (Gaut et al., 1996; Eyre-Walker and Gaut, 1997; Gaut 1998). We also detected a positive correlation between mitochondrial and plastid rates for *dN* and *dS* (Figure 4.10).

Patterns of *dN* and *dS* are similar across legumes for both genomes including an acceleration in the papilionoid legumes (Figure 4.9). Sloan et al. (2012)

compared rates in plastomes and mitogenomes of four *Silene* species and found mitogenome-wide increases in *dS* were not correlated as much with plastome rates as they were with plastomic rearrangements, such as indels, intron losses and inversions. While another study comparing genome-wide rates and biological features of both organellar genomes is not available, similar cases in which increases in sequence and/or structural evolution in either organellar genome has been shown in Geraniaceae (Parkinson et al., 2005; Mower et al., 2007; Guisinger et al., 2008, 2011; Blazier et al., 2011; Blazier et al., 2016a), gymnosperms (McCoy et al., 2008; Wu et al., 2009; Wu and Chaw, 2014), and legumes (Sabir et al., 2014; Schwarz et al., 2015; Chapter 3). Correlations between plastomic rearrangements and substitution rates were shown in Chapter three. While caesalpinioids and mimosoids have ancestral gene content and order (Dugas et al., 2015; Schwarz et al. 2015), papilionoids exhibit many rearrangements in the form of inversions, gene and intron losses, indels and the loss of one copy of the IR in one clade (Palmer et al., 1987; Cai et al., 2008; Sabir et al., 2014; Sveinsson and Cronk, 2014; Schwarz et al., 2015). The increased rates in both genomes in the papilionoid legumes may be explained by a common mechanism that results in increased substitution rates and genomic rearrangements. *MSH1*, *RECA* and Whirly proteins have been shown to play important roles in plant organellar genome stability (Shedge et al., 2007; Marechal et al., 2009; Rowan et al., 2010; Xu et al., 2011). *MSH1* is targeted to both mitochondria and plastids but only the mitogenome is affected in mutants (Shedge et al., 2007) and Whirly proteins are important in stabilizing the plastome (Marechal et al., 2009). A modification the dual-targeted *RECA2* gene could affect the evolution of both genomes but double knockouts reveal that the consequences in the plastome and mitogenome are different (Shedge et al., 2007). The relationship between the patterns of evolution in the mitogenome and plastome is still unclear. More comparative studies of all three plant genomes need to be completed to uncover evolutionary mechanisms driving these patterns.

CONCLUSION

This is the most comprehensive study of evolutionary rates of organellar genomes in legumes. Although whole mitogenomes were not generated, the rate analyses provide insights into patterns of evolution within the family. We identified four mitochondrial genes missing in one or more species of legumes (*cox2, rps1, rps3, rps14*). These same genes have been lost and functionally transferred to the nucleus in several disparate lineages of angiosperms (Mower et al. 2012). Values of *dS* of the plastome are 3.8 times faster than in the mitogenome, which are similar or slightly higher than well-established ratio of 3:1 between these two genomes (Wolfe et al., 1987; Drouin et al., 2008). In both genomes we see accelerated *dN* and *dS* in papilionoid legumes compared to caesalpinioids and mimosoids. This acceleration may be due to differences in growth habit, nuclear encoded genes involved in DNA replication, repair and recombination or a combination of these two forces. More organellar genome comparisons are needed to expand the knowledge of the evolutionary mechanisms driving genomic rearrangements and accelerations in substitution rates.

Table 4.1. List of outgroup and Fabaceae species utilized in this study with subfamily placement, and accession numbers. XX – XX indicates range of accessions numbers, which will be submitted to Genbank when the chapter is submitted for publication.

Species	Subfamily	Mitochondria Accession No.	Plastid Accession No.
Populus tremula	Outgroup	NC_028096	NC_027425
Caesalpinia coriaria	Caesalpinioideae	XX-XX	KJ468095
Ceratonia siliqua	Caesalpinioideae	XX-XX	KJ468096
Cercis canadensis	Caesalpinioideae	XX-XX	KF856619
Haematoxylum brasiletto	Caesalpinioideae	XX-XX	KJ468097
Tamarindus indica	Caesalpinioideae	XX-XX	KJ468103
Prosopis gladulosa	Mimosoideae	XX-XX	KJ468101
Apios americana	Papilionoideae	XX-XX	KF856618
Arachis hypogaea	Papilionoideae	XX-XX	KJ468094
Glycine max	Papilionoideae	NC_020455.1	NC_007942
Indigofera tinctoria	Papilionoideae	XX-XX	KJ468098
Lotus japonicus	Papilionoideae	NC_016743.2	NC_002694
Lupinus albus	Papilionoideae	XX-XX	KJ468099
Millettia pinnata	Papilionoideae	NC_016742.1	NC_016708
Pachyrhizus erosus	Papilionoideae	XX-XX	KJ468100
Robinia pseudoacacia	Papilionoideae	XX-XX	KJ468102
Vigna angularis	Papilionoideae	NC_021092.1	NC_021091
Vigna radiata var. radiata	Papilionoideae	NC_015121.1	NC_013843
Vigna unguiculata	Papilionoideae	XX-XX	KJ468104
Glycyrrhiza glabra	Papilionoideae_IRLC	XX-XX	KF201590
Medicago truncatula	Papilionoideae_IRLC	NC_029641.1	NC_003119.6
Trifolium aureum	Papilionoideae_IRLC	XX-XX	KC894708
Trifolium grandiflorum	Papilionoideae_IRLC	XX-XX	KC894707
Trifolium meduseum	Papilionoideae_IRLC	XX-XX	KJ476730
Trifolium pratense	Papilionoideae_IRLC	XX-XX	XX-XX
Trifolium repens	Papilionoideae_IRLC	XX-XX	KC894706
Vicia faba	Papilionoideae_IRLC	KC189947	KF042344

Table 4.2. List of genes utilized in all nucleotide substitution rate analyses.

Plastid		**Mitochondrion**
atpA	*psbJ*	*atp1*
atpB	*psbK*	*atp4*
atpE	*psbL*	*atp6*
atpF	*psbM*	*atp8*
atpH	*psbN*	*atp9*
atpI	*psbT*	*ccmB*
ccsA	*psbZ*	*ccmC*
cemA	*rbcL*	*ccmFC*
clpP	*rpl2*	*ccmFN*
matK	*rpl14*	*cob*
ndhA	*rpl16*	*cox1*
ndhB	*rpl20*	*cox3*
ndhC	*rpl36*	*matR*
ndhD	*rpoA*	*mttB*
ndhE	*rpoB*	*nad1*
ndhF	*rpoC1*	*nad2*
ndhG	*rpoC2*	*nad3*
ndhH	*rps2*	*nad4*
ndhI	*rps3*	*nad4L*
ndhJ	*rps4*	*nad5*
ndhK	*rps7*	*nad6*
petA	*rps8*	*nad7*
petB	*rps11*	*nad9*
petD	*rps12*	*rpl16*
petG	*rps14*	*rps4*
petL	*rps15*	*rps12*
petN	*rps18*	
psaA	*rps19*	
psaB	*ycf1*	
psaC	*ycf2*	
psaJ	*ycf3*	
psbA		
psbB		
psbC		
psbD		
psbE		

Table 4.2 (continued)

psbF		
psbH		
psbI		

Table 4.3. Mitochondria gene content in 26 legumes and the outgroup *Populus*. X indicates gene is present, red indicates absences.

Gene	*Populus_tremula*	*Apios americana*	*Arachis hypogaea*	*Caesalpinia coriaria*	*Ceratonia siliqua*	*Cercis canadensis*	*Glycine max*	*Glycyrrhiza glabra*	*Haematoxylum brasileto*	*Indigofera tinctoria*	*Lotus japonicus*	*Lupinus albus*	*Medicago truncatula*	*Millettia pinnata*	*Pachyrhizus erosus*	*Prosopis gladulosa*	*Robinia pseudoacacia*	*Tamarindus indica*	*Trifolium aureum*	*Trifolium grandiflorum*	*Trifolium meduseum*	*Trifolium pratense*	*Trifolium repens*	*Vicia faba*	*Vigna angularis*	*Vigna radiata*	*Vigna unguiculata*
atp1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
atp4	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
atp6	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
atp8	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
atp9	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
ccmB	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
ccmC	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
ccmFC	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
ccmFN	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
cob	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
cox1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
cox2	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	🟥	🟥	🟥
cox3	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
matR	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
mttB	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
nad1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
nad2	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
nad3	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
nad4	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
nad4L	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
nad5	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
nad6	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
nad7	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
nad9	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
rpl5	🟥	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
rpl16	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
rps1	X	X	X	X	X	X	X	X	X	X	🟥	X	X	X	X	X	X	X	🟥	🟥	🟥	X	X	X	X	X	X
rps3	X	X	X	X	X	X	X	X	🟥	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
rps4	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
rps10	🟥	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
rps12	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
rps14	X	X	X	X	X	X	X	X	X	🟥	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

Table 4.4. Predicted number of RNA editing sites for each mitochondrial gene predicted by PREP-Mt of the PREP Suite.

Gene	*Apios americana*	*Arachis hypogaea*	*Caesalpinia coriaria*	*Ceratonia siliqua*	*Cercis canadensis*	*Glycine max*	*Glycyrrhiza glabra*	*Haematoxylum brasiletto*	*Indigofera tinctoria*	*Lotus japonicus*	*Lupinus albus*	*Medicago truncatula*	*Millettia pinnata*	*Pachyrhizus erosus*	*Prosopis gladulosa*	*Robinia pseudoacacia*	*Tamarindus indica*	*Trifolium aureum*	*Trifolium grandiflorum*	*Trifolium meduseum*	*Trifolium pratense*	*Trifolium repens*	*Vicia faba*	*Vigna angularis*	*Vigna radiata*	*Vigna unguiculata*	*Average No. Sites*
atp1	2	1	6	5	5	1	1	6	2	1	1	1	2	1	6	1	5	1	1	1	0	0	1	1	1	1	2.077
atp4	12	12	12	12	12	12	12	12	11	12	12	12	12	12	12	12	12	12	12	12	2	5	12	12	12	12	11.308
atp6	17	15	15	15	15	17	15	15	17	18	0	3	15	17	15	17	18	7	7	7	7	0	8	16	16	16	12.615
atp8	3	2	5	3	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	0	3	3	3	3	2.885
atp9	2	2	5	5	5	2	2	5	2	1	2	2	2	2	5	2	4	2	2	2	2	2	2	2	2	2	2.615
ccmB	30	30	33	33	33	30	29	33	29	29	27	31	30	30	32	31	32	29	29	29	29	29	29	31	31	31	30.346
ccmC	28	27	30	30	29	28	27	30	27	27	28	28	26	28	30	28	30	28	26	28	28	1	28	29	29	29	27.192
ccmFC	19	19	19	20	19	19	19	19	19	19	19	19	19	19	20	19	20	19	19	19	9	9	19	19	19	19	18.346
ccmFN	33	33	32	32	34	33	33	32	32	33	33	33	33	33	34	33	33	32	32	32	6	32	33	33	33	33	31.731
cob	15	15	15	16	16	14	15	15	15	16	15	15	14	14	15	15	16	15	15	15	14	15	15	14	14	14	14.885
cox1	19	20	21	20	21	18	19	21	19	19	19	19	19	18	21	19	21	19	19	18	0	0	19	17	17	17	17.654
cox3	11	11	13	13	13	11	12	13	11	12	12	12	9	9	13	12	13	12	12	11	10	11	12	10	10	10	11.462
matR	13	14	13	13	12	13	12	13	13	13	11	12	13	13	13	13	12	12	12	12	12	12	12	15	15	15	12.808
mttB	25	25	26	25	26	25	25	26	23	24	24	25	25	24	25	25	26	25	25	25	25	0	25	26	26	26	24.115
nad1	20	20	21	20	19	20	20	19	21	21	20	19	20	20	20	20	20	19	19	19	19	8	19	20	20	20	19.346
nad2	28	30	31	32	32	28	28	31	28	29	29	27	28	28	29	28	32	28	28	28	28	16	28	27	27	27	28.269
nad3	13	13	13	13	13	12	13	11	12	13	13	12	13	12	13	13	12	12	12	12	12	12	12	12	12	12	12.385
nad4	41	41	41	42	40	41	41	40	40	42	41	42	41	40	41	41	43	41	41	42	41	2	40	41	41	41	39.538
nad4L	13	13	13	13	13	13	13	13	13	13	13	13	12	13	5	13	13	13	13	13	13	13	13	13	13	13	12.654
nad5	28	28	28	28	27	27	27	28	28	28	28	28	28	28	28	28	28	28	28	28	28	10	28	28	28	28	27.192
nad6	12	10	10	11	11	12	12	11	12	12	12	11	12	12	12	12	11	11	11	11	11	11	11	12	12	12	11.423
nad7	29	28	31	32	32	29	30	31	30	30	29	30	29	29	31	30	33	29	30	30	29	7	30	29	29	29	29.038
nad9	7	8	8	8	9	8	8	8	8	8	8	8	6	8	8	8	8	9	6	6	6	6	6	8	8	8	7.577
rpl16	4	5	6	7	6	4	4	6	4	4	4	4	4	4	7	4	7	4	4	4	4	1	4	5	5	5	4.615
rps4	16	17	16	16	16	17	16	15	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	17	17	17	16.154
rps12	7	7	7	7	7	7	7	7	6	7	7	6	7	7	6	7	7	7	7	7	7	7	7	7	7	7	6.885

Table 4.5. Mean *dN* and *dS* values of mitochondrial and plastid genes.

Genome	**Substitution Type**	**Mean**	**P-value***
Mitochondrion	*dN*	0.0408	3.14E-05
Plastid	*dN*	0.0531	
Mitochondrion	*dS*	0.1197	4.03E-15
Plastid	*dS*	0.4543	

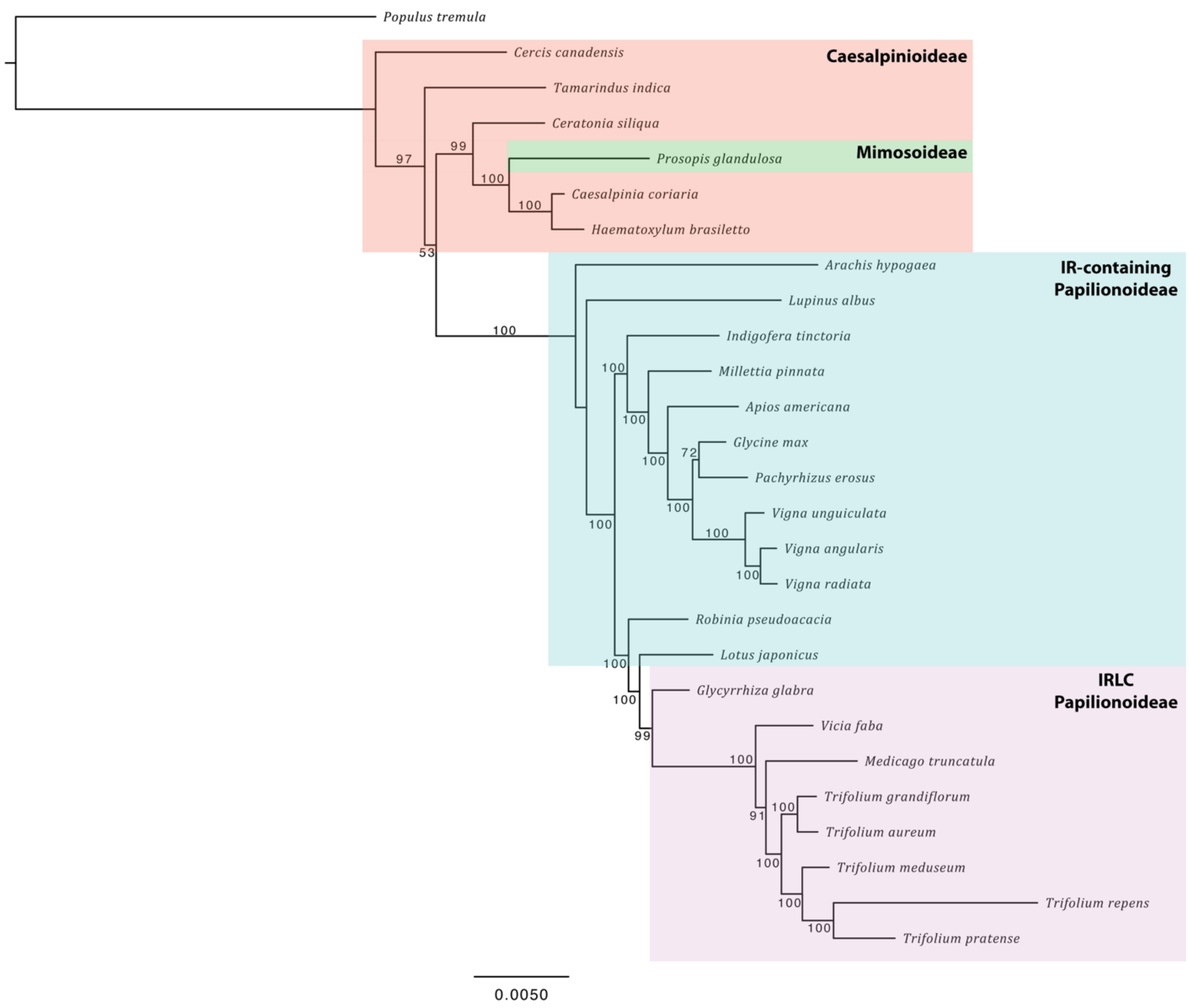


Figure 4.1. Maximum likelihood tree (-ln = -61835.919) of Fabaceae based on 26 mitochondrial genes.

Numbers at nodes are bootstrap support values. Only support values greater than 50 are shown. The scale bar represents substitutions per site. The phylogeny is divided into four subgroups: Caesalpinioideae (red), Mimosoideae (green), Papilionoideae taxa containing both copies of the IR (inverted repeat) (blue) and Papilionoids lacking the IR (purple).

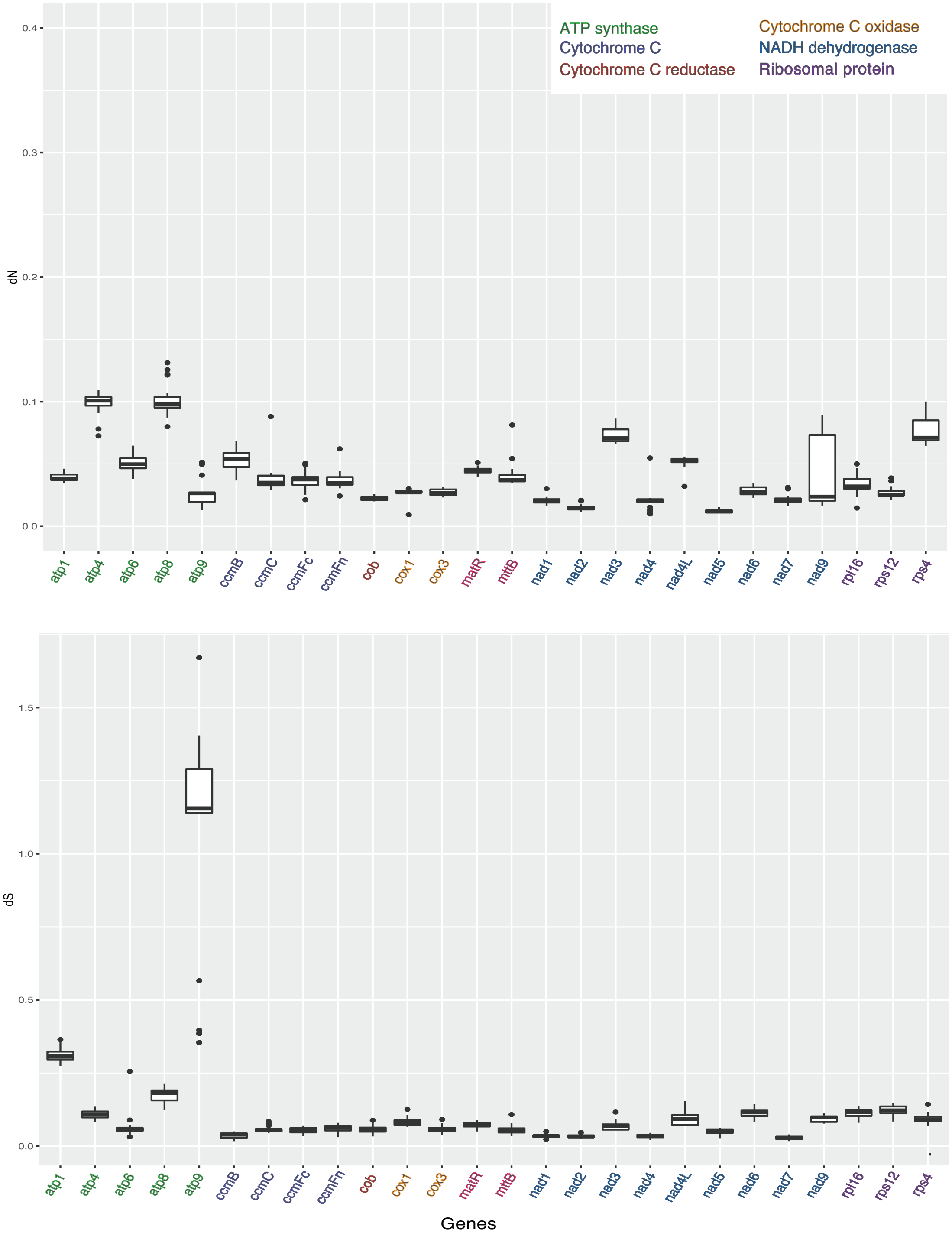

ATP synthase
Cytochrome C
Cytochrome C reductase
Cytochrome C oxidase
NADH dehydrogenase
Ribosomal protein
dN
0.0
0.1
0.2
0.3
0.4
dS
0.0
0.5
1.0
1.5
atp1
atp4
atp6
atp8
atp9
ccmB
ccmC
ccmFc
ccmFn
cob
cox1
cox3
matR
mttB
nad1
nad2
nad3
nad4
nad4L
nad5
nad6
nad7
nad9
rpl16
rps12
rps4
Genes

Figure 4.2. Box plots of *dN* and *dS* values of 26 mitochondrial genes.

The top and bottom lines of each box represent the 75th and 25th percentiles, respectively and the horizontal line in each box represents the 50th percentile. The whisker lines represent the minimum to the maximum points and the points outside of the whiskers are outliers. Gene names are colored to represent functional groups: ATP synthase genes (green), Cytochrome C (periwinkle), Cytochrome C reductase (red), Cytochrome C oxidase (orange), NADH dehydrogenase (blue) and ribosomal proteins (purple).

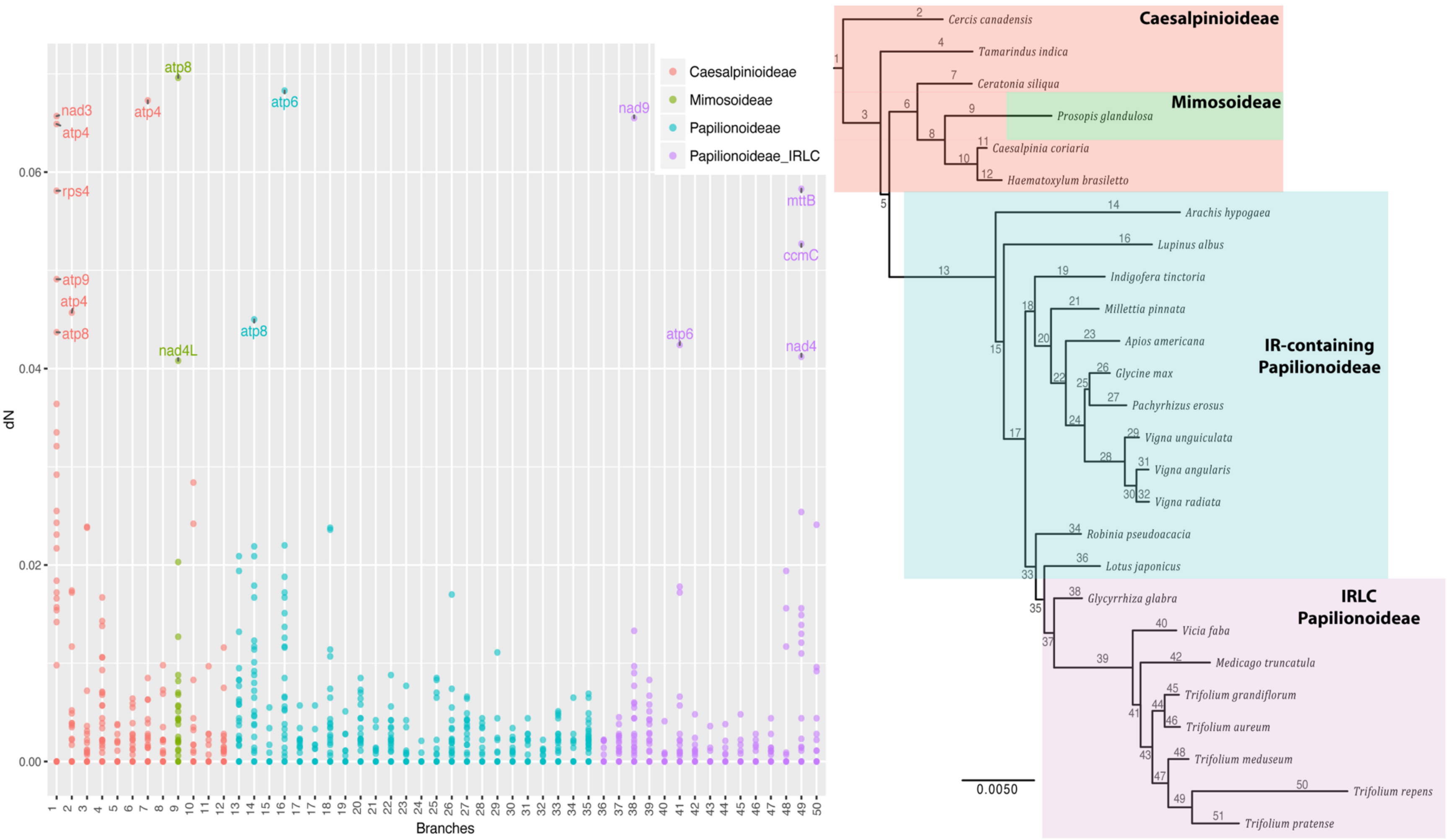

nad3
atp4
rps4
atp9
atp4
atp8
atp4
atp8
nad4L
atp8
atp6
nad9
mttB
ccmC
atp6
nad4
0.06
0.04
0.02
0.00
dN
Branches
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
Caesalpinioideae
Mimosoideae
Papilionoideae
Papilionoideae_IRLC
Caesalpinioideae
Mimosoideae
IR-containing Papilionoideae
IRLC Papilionoideae
Cercis canadensis
Tamarindus indica
Ceratonia siliqua
Prosopis glandulosa
Caesalpinia coriaria
Haematoxylum brasiletto
Arachis hypogaea
Lupinus albus
Indigofera tinctoria
Millettia pinnata
Apios americana
Glycine max
Pachyrhizus erosus
Vigna unguiculata
Vigna angularis
Vigna radiata
Robinia pseudoacacia
Lotus japonicus
Glycyrrhiza glabra
Vicia faba
Medicago truncatula
Trifolium grandiflorum
Trifolium aureum
Trifolium meduseum
Trifolium repens
Trifolium pratense
0.0050
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51

Figure 4.3. Plot of *dN* values for each gene by branch.

Point plot representing *dN* values of all genes for each branch. Each point represents the *dN* value of one gene. The Fabaceae subgroups are indicated by the following colors: Caesalpiniodeae (red), Mimosoideae (green) Papilionoideae (blue), Papilionoideae IRLC (purple). Data points with a value greater than 0.04 are labeled with their gene name. Branch numbers along the x-axis correlate to the branch labels on the phylogeny on the right, which is taken from Figure 4.1 with the outgroup taxon and bootstrap support values removed.

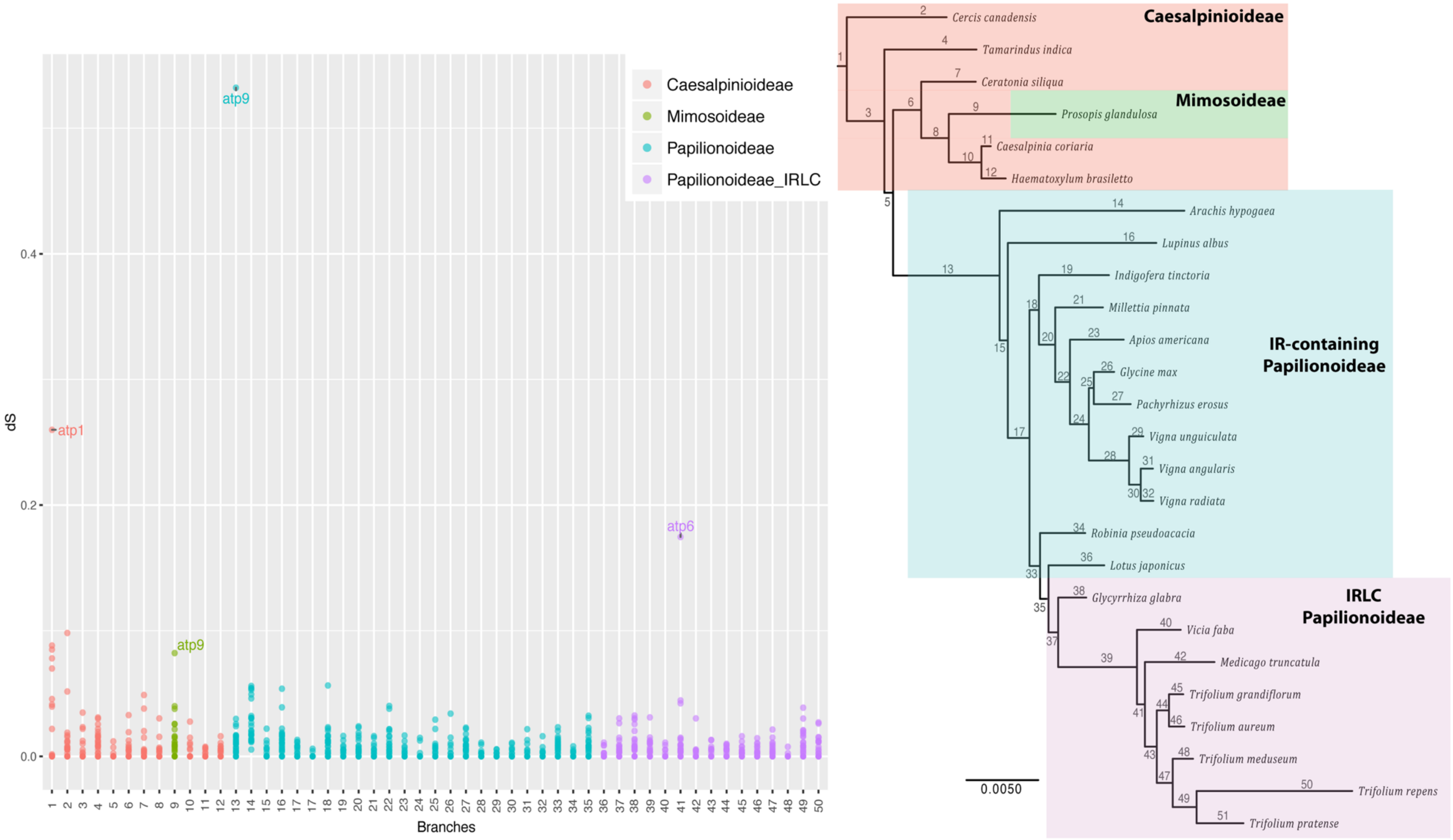

Caesalpinioideae
Mimosoideae
Papilionoideae
Papilionoideae_IRLC
dS
0.0
0.2
0.4
Branches
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
atp9
atp1
atp6
atp9
Caesalpinioideae
Mimosoideae
IR-containing Papilionoideae
IRLC Papilionoideae
Cercis canadensis
Tamarindus indica
Ceratonia siliqua
Prosopis glandulosa
Caesalpinia coriaria
Haematoxylum brasiletto
Arachis hypogaea
Lupinus albus
Indigofera tinctoria
Millettia pinnata
Apios americana
Glycine max
Pachyrhizus erosus
Vigna unguiculata
Vigna angularis
Vigna radiata
Robinia pseudoacacia
Lotus japonicus
Glycyrrhiza glabra
Vicia faba
Medicago truncatula
Trifolium grandiflorum
Trifolium aureum
Trifolium meduseum
Trifolium repens
Trifolium pratense
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51
0.0050

Figure 4.4. Plot of *dS* values for each gene by branch.

Point plot representing *dS* values of all genes for each branch. Each point represents the *dS* value of one gene. The Fabaceae subgroups are indicated by the following colors: Caesalpiniodeae (red), Mimosoideae (green) Papilionoideae (blue), Papilionoideae IRLC (purple). Data points with value greater than 0.15 are labeled with their gene name. Branch numbers along the x-axis correlate to the branch labels on the phylogeny on the right, which is taken from Figure 4.1 with the outgroup taxon and bootstrap support values removed.

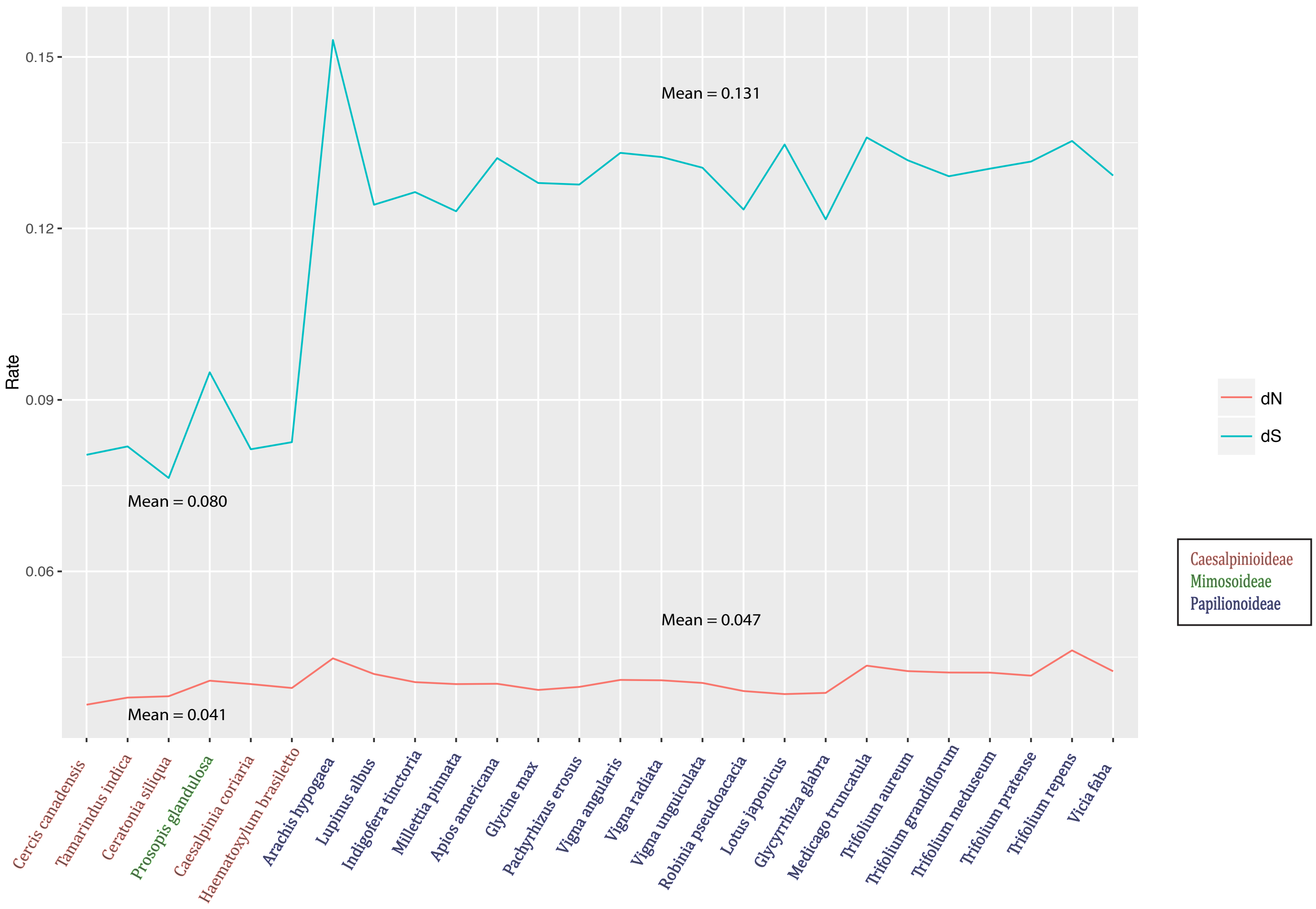

0.15
0.12
0.09
0.06
Rate
Mean = 0.131
Mean = 0.080
Mean = 0.047
Mean = 0.041
dN
dS
Caesalpinioideae
Mimosoideae
Papilionoideae
Cercis canadensis
Tamarindus indica
Ceratonia siliqua
Prosopis glandulosa
Caesalpinia coriaria
Haematoxylum brasiletto
Arachis hypogaea
Lupinus albus
Indigofera tinctoria
Millettia pinnata
Apios americana
Glycine max
Pachyrhizus erosus
Vigna angularis
Vigna radiata
Vigna unguiculata
Robinia pseudoacacia
Lotus japonicus
Glycyrrhiza glabra
Medicago truncatula
Trifolium aureum
Trifolium grandiflorum
Trifolium meduseum
Trifolium pratense
Trifolium repens
Vicia faba

Figure 4.5. Pairwise comparison of *dN* and *dS* values for each species across Fabaceae.

Line plot represents average *dN* (red line) and *dS* (green line) values for each species. Species labels are colored to indicate Fabaceae subfamilies: Caesalpinioideae (red), Mimosoideae (green), Papilionoideae (blue). Species are in the order shown in the phylogeny in Figure 4.1.

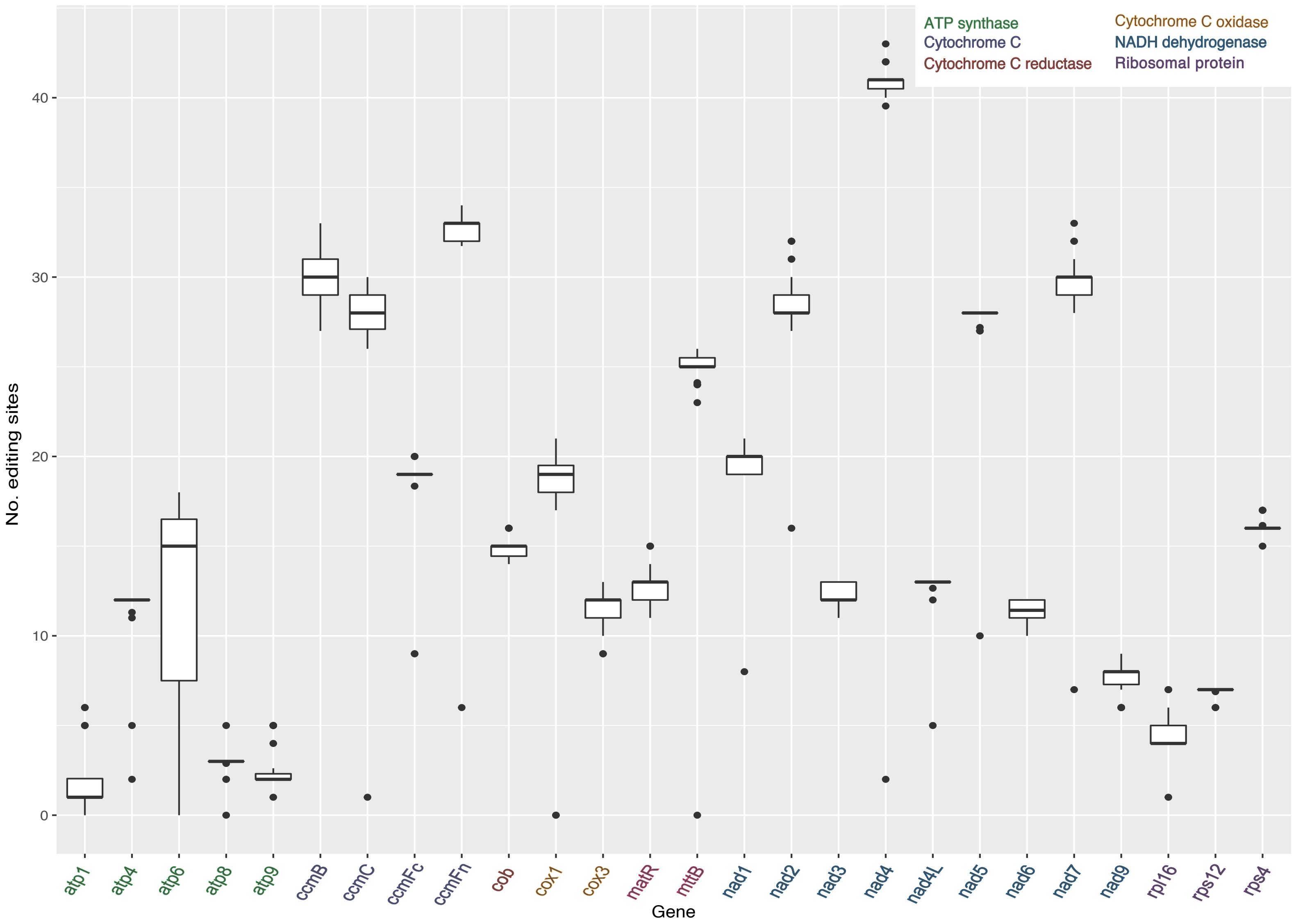
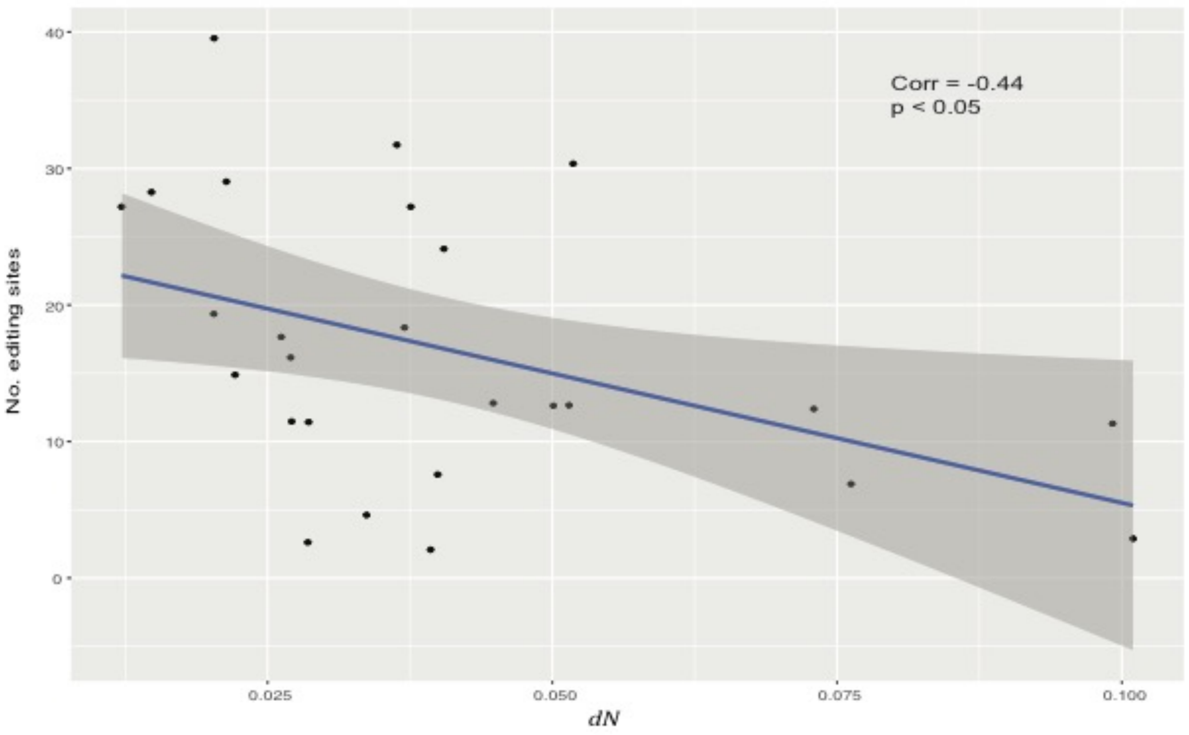

ATP synthase
Cytochrome C
Cytochrome C reductase
Cytochrome C oxidase
NADH dehydrogenase
Ribosomal protein
No. editing sites
0
10
20
30
40
Gene
atp1
atp4
atp6
atp8
atp9
ccmB
ccmC
ccmFc
ccmFn
cob
cox1
cox3
matR
mttB
nad1
nad2
nad3
nad4
nad4L
nad5
nad6
nad7
nad9
rpl16
rps12
rps4

Figure 4.6. Box plots of number of RNA editing sites of 26 mitochondrial genes.

The top and bottom lines of each box represent the 75th and 25th percentiles, respectively and the horizontal line in each box represents the 50th percentile. The whisker lines represent the minimum to the maximum points and the points outside of the whiskers are outliers. Gene names are colored to represent functional groups: ATP synthase genes (green), Cytochrome C (periwinkle), Cytochrome C reductase (red), Cytochrome C oxidase (orange), NADH dehydrogenase (blue) and ribosomal proteins (purple).


Corr = -0.44
p < 0.05
No. editing sites
40
30
20
10
0
0.025
0.050
0.075
0.100
dN

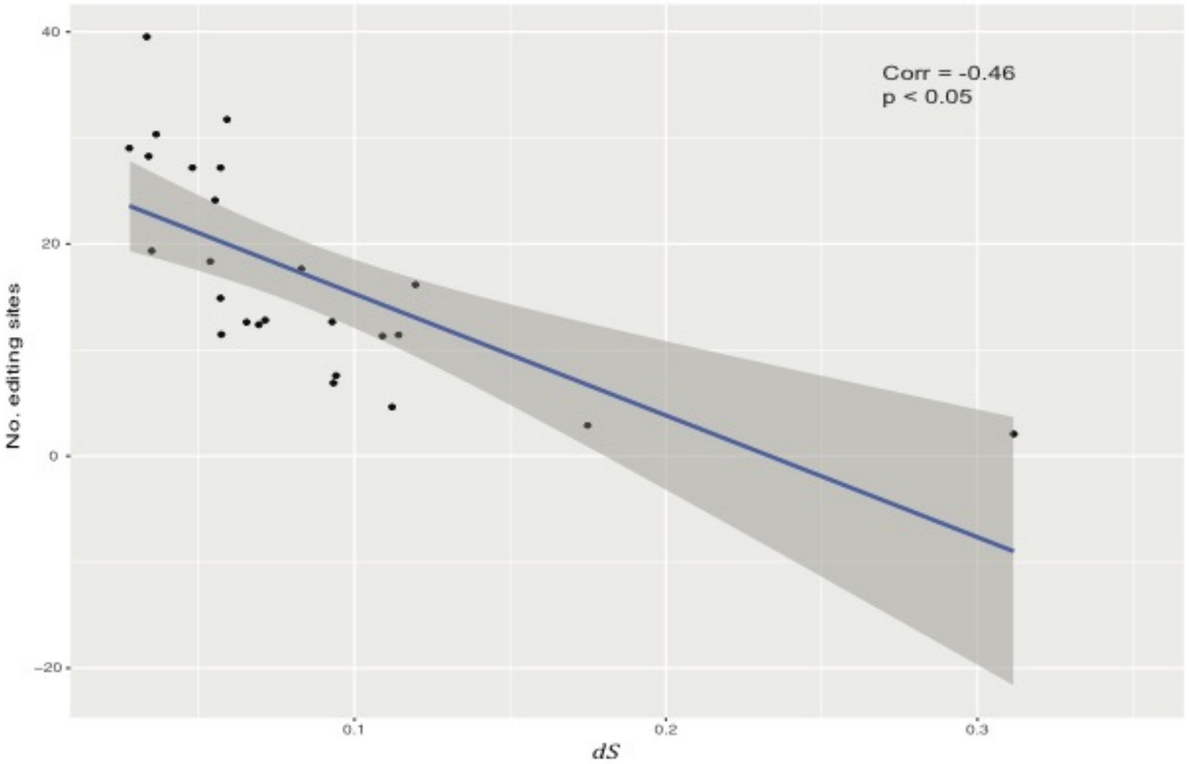

Corr = -0.46
p < 0.05
No. editing sites
40
20
0
-20
0.1
0.2
0.3
dS

Figure 4.7. Correlation scatterplots between *dN* and *dS* values of the mitchondrion versus number of predicted RNA editing sites.

Scatterplots with regression lines (blue) of *dN* (top diagram) and *dS* (bottom diagram) values from the genes of the mitochondria versus the number of predicted RNA editing sites. The grey region surrounding the regression line represents the standard error. Correlation values = - 0.440 (*dN*) and -0.460 (*dS*), p-values = 0.026 (*dN*) and 0.028 (*dS*).

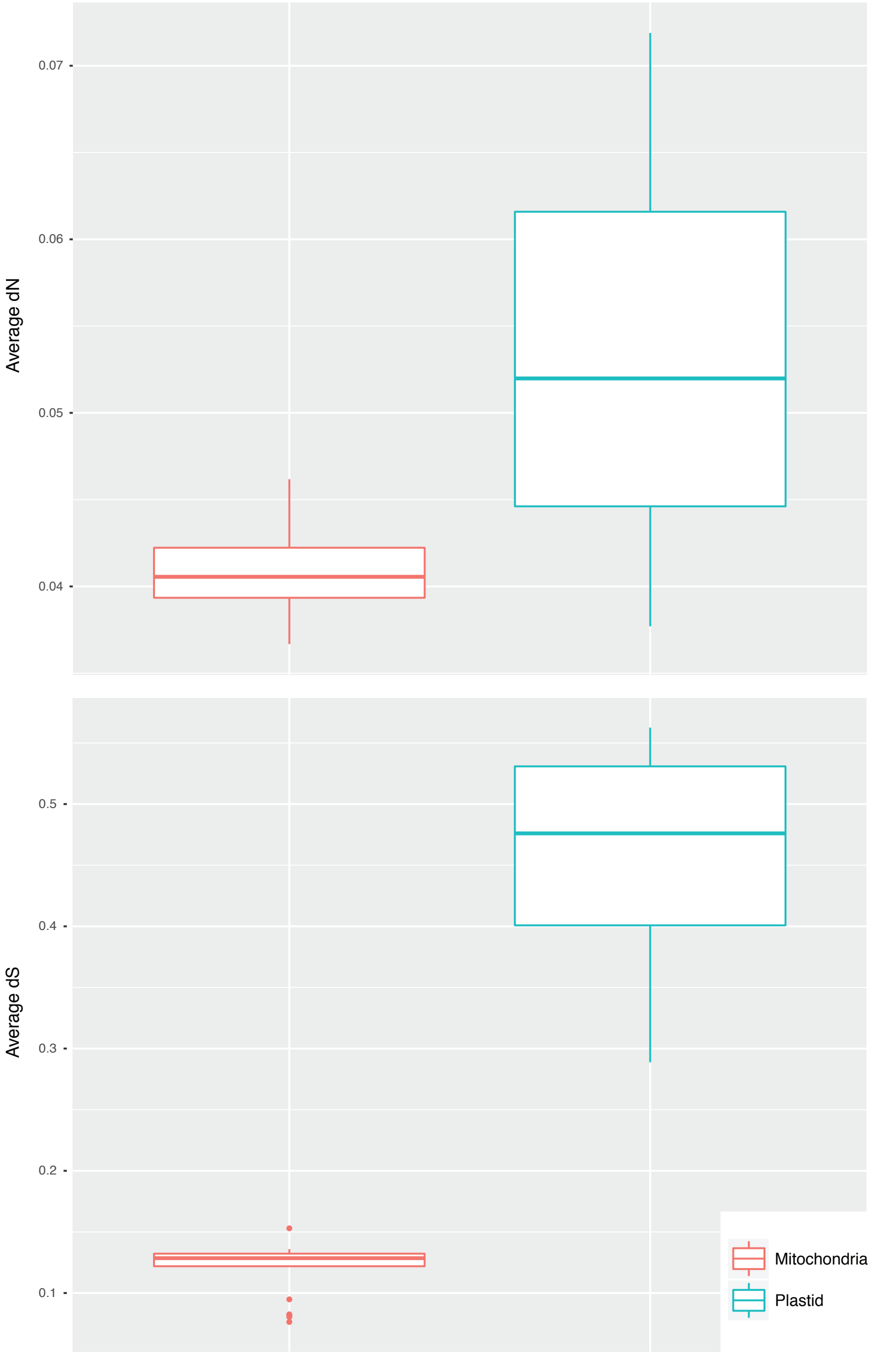

0.07
0.06
0.05
0.04
Average dN
0.5
0.4
0.3
0.2
0.1
Average dS
Mitochondria
Plastid

Figure 4.8. Box plots of overall *dN* and *dS* values of the mitochondrial and plastid genes.

Values of *dN* (top diagram) and *dS* (bottom diagram) for mitochondrial (red boxes) and plastid (green boxes) genes are shown. In each plot the top and bottom lines of the box represent the 75th and 25th percentiles, respectively and the middle line in each box represents the 50th percentile. The whisker lines represent the minimum to the maximum points and the points outside of the whisker lines are outliers.

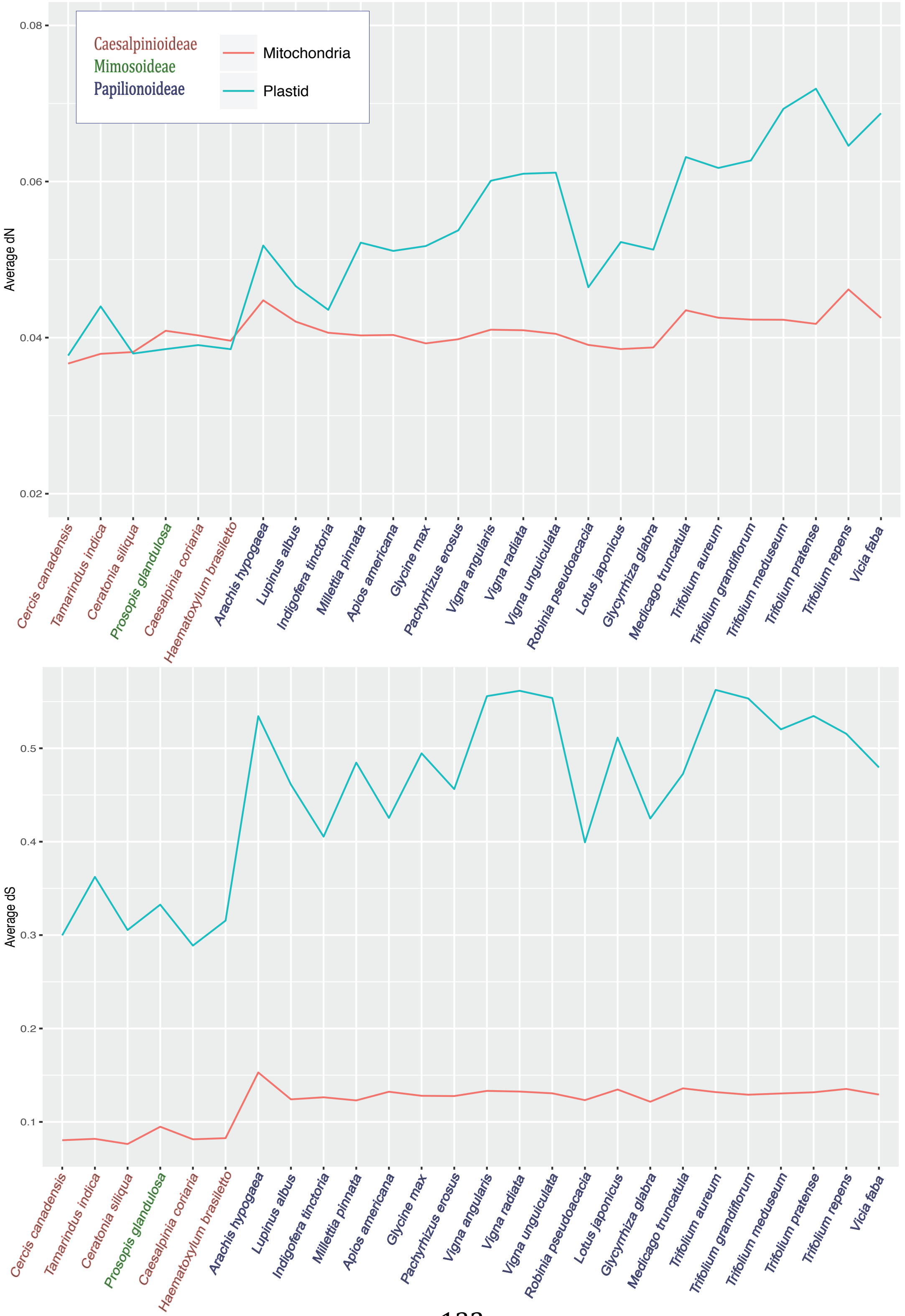
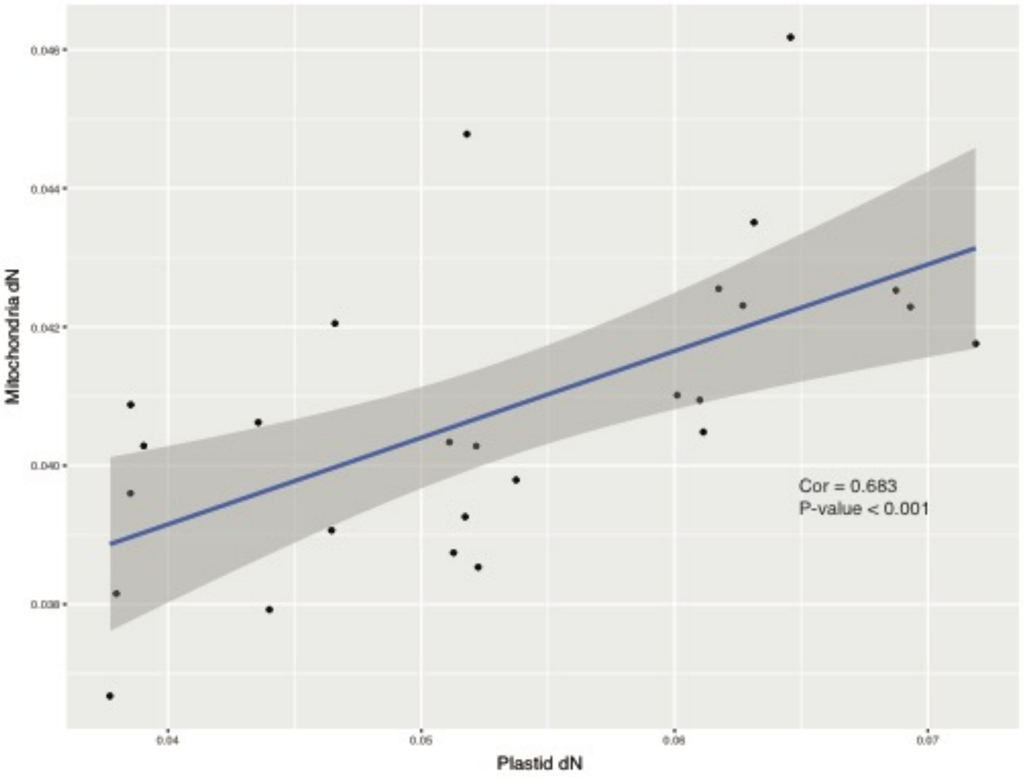

Caesalpinioideae
Mimosoideae
Papilionoideae
Mitochondria
Plastid
Average dN
0.08
0.06
0.04
0.02
Cercis canadensis
Tamarindus indica
Ceratonia siliqua
Prosopis glandulosa
Caesalpinia coriaria
Haematoxylum brasiletto
Arachis hypogaea
Lupinus albus
Indigofera tinctoria
Millettia pinnata
Apios americana
Glycine max
Pachyrhizus erosus
Vigna angularis
Vigna radiata
Vigna unguiculata
Robinia pseudoacacia
Lotus japonicus
Glycyrrhiza glabra
Medicago truncatula
Trifolium aureum
Trifolium grandiflorum
Trifolium meduseum
Trifolium pratense
Trifolium repens
Vicia faba
Average dS
0.5
0.4
0.3
0.2
0.1
Cercis canadensis
Tamarindus indica
Ceratonia siliqua
Prosopis glandulosa
Caesalpinia coriaria
Haematoxylum brasiletto
Arachis hypogaea
Lupinus albus
Indigofera tinctoria
Millettia pinnata
Apios americana
Glycine max
Pachyrhizus erosus
Vigna angularis
Vigna radiata
Vigna unguiculata
Robinia pseudoacacia
Lotus japonicus
Glycyrrhiza glabra
Medicago truncatula
Trifolium aureum
Trifolium grandiflorum
Trifolium meduseum
Trifolium pratense
Trifolium repens
Vicia faba

Figure 4.9. Pairwise comparison of *dN* and *dS* values of the mitochondria and plastid genes across Fabaceae.

Line plots represent average *dN* (top diagram) and *dS* (bottom diagram) values for each species. Mitochondrial rate values are indicated by the green and plastid rate values are indicated by the red line. Species labels are colored to indicate Fabaceae subfamilies: Caesalpinioideae (red), Mimosoideae (green), Papilionoideae (blue). Species are in the order shown in the phylogeny in Figure 4.1.


0.046
0.044
0.042
0.040
0.038
Mitochondria dN
0.04
0.05
0.06
0.07
Plastid dN
Cor = 0.683
P-value < 0.001

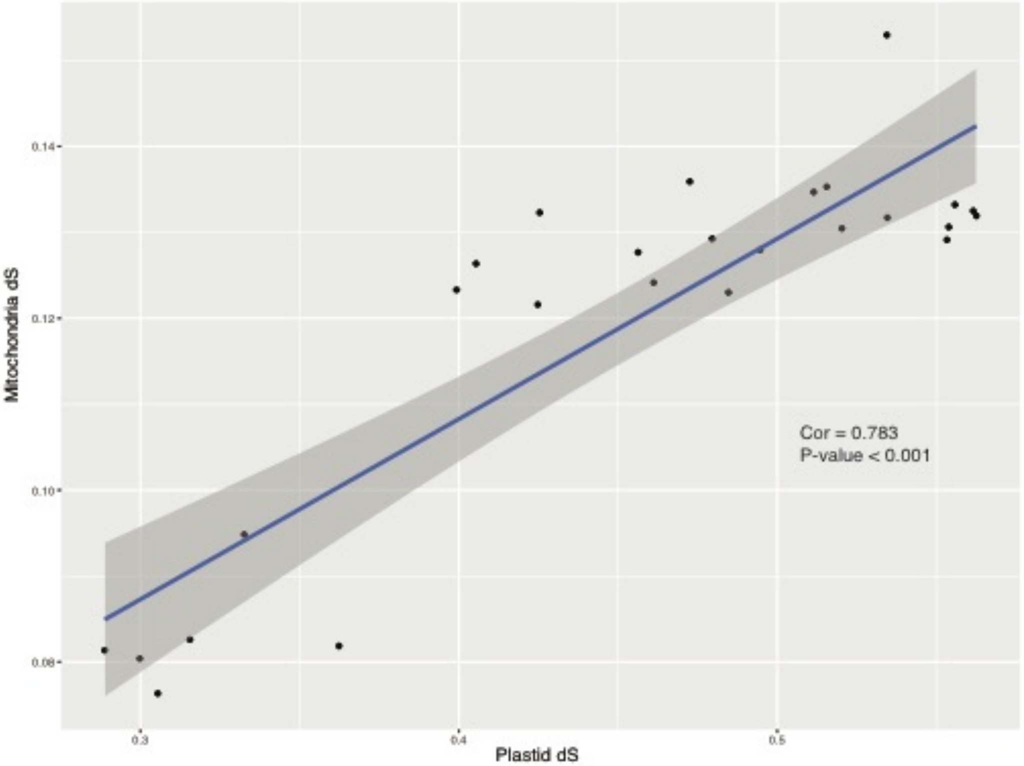

0.14
0.12
0.10
0.08
Mitochondria dS
0.3
0.4
0.5
Plastid dS
Cor = 0.783
P-value < 0.001

Figure 4.10. Correlation scatterplots between *dN* and *dS* values of the plastid versus mitochondrial genes.

Scatterplots with regression lines (blue) of *dN* (top diagram) and *dS* (bottom diagram) values from the genes of the mitochondria versus the plastid. The grey region surrounding the regression line represents the standard error. Correlation values = 0.683 (*dN*) and 0.783 (*dS*), p-values = 1.12E-04 (*dN*) and 2.24E-06 (*dS*).

References

Adams KL, Palmer JD (2003) Evolution of mitochondrial gene content: gene loss and transfer to the nucleus. Molecular Phylogenetics and Evolution 29:380-395

Adams KL, Qiu Y-L, Stoutemyer M, Palmer JD (2002) Punctuated evolution of mitochondrial gene content: high and variable rates of mitochondrial gene loss and transfer to the nucleus during angiosperm evolution. PNAS 99:9905-9912

Adams KL, Rosenblueth M, Qiu Y-L, Palmer JD (2001) Multiple losses and transfers to the nucleus of two mitochondrial succinate dehydrogenase genes during angiosperm evolution. Genetics 158: 1289-1300

Adams KL, Song K, Roessler PG, Nugent JM, Doyle JL, Doyle JL, Palmer JD (1999) Intracellular gene transfer in action: dual transcription and multiple silencings of nuclear and mitochondrial *cox2* genes in legumes. PNAS 96:13863-13868

Alverson AJ, Zhuo S, Rice DW, Sloan DB, Palmer JD (2011) The mitochondrial genome of the legume *Vigna radiata* and the analysis of recombination across short mitochondrial repeats. PLoS ONE. 6:e16404

Bailey CD, Doyle JJ, Kajita T, Nemoto T, Ohashi H. 1997. The chloroplast *rpl2* intron and ORF184 as phylogenetic markers in the legume Tribe Desmodieae Systematic Botany 22(1): 133–138.

Bakker FT, Breman F, Merckx (2006) DNA sequence evolution in fast evolving mitochondrial DNA nad1 exons in Geraniaceae and Plantaginaceae. Taxon 55(4): 887-896

Barnard-Kubow K, Sloan DB, Galloway LF (2014) Correlation between sequence divergence and polymorphism reveals similar evolutionary mechanisms acting across multiple timescales in a rapidly evolving plastid genome. BMC Evol Biol. doi: 10.1186/s12862-014-0268-y

Barraclough TG, Harvey PH, Nee S (1996) Rate of *rbc*L gene sequence evolution and species diversification in flowering plants (angiosperms). Proc R Soc Lond B 263: 589-591

Birky CW, Walsh JB (1992) Biased gene conversion, copy number, and apparent mutation rate differences within chloroplast and bacterial genomes. Genetics 130:677–683

Blazier CJ, Guisinger MM, Jansen RK (2011) Recent loss of plastid-encoded ndh genes within *Erodium* (Geraniaceae). Plt Mol Biol 76:263–272.

Blazier JC, Jansen RK, Mower JP, Govindu M, Zhang J, Weng M-L, Ruhlman TA (2016a) Variable presence of the inverted repeat and plastome stability in *Erodium*. Annals of Botany, doi:10.1093/aob/mcw065.

Blazier JC, Ruhlman TA, Weng M-L, Rehman SK, Sabir JSM, and Jansen RK (2016b) Divergence of RNA polymerase α subunits in angiosperm plastid genomes is mediated by genomic rearrangement. Scientific Reports 6:24595.

Bobiwash, K, Schultz S, and Schoen D (2013) Somatic deleterious mutation rate in a woody plant: estimation from phenotypic data. Heredity 111:338–344

Bock R, Knoop V (2012) Genomics of Chloroplasts and Mitochondria. Springer Science and Business Media

Bock R, Timmis JN (2008) Reconstructing evolution: Gene transfer from plastids to the nucleus. BioEssays 30(6): 556–566

Bousquet J, Strauss S, Doerksen A, Price R (1992) Extensive variation in evolutionary rate of *rbcL* gene sequences among seed plants. Proc Natl Acad Sci 89:7844–7848.

Brandvain Y, Wade MJ (2009) The functional transfer of genes from the mitochondria to the nucleus: the effects of selection, mutation, population size and rate of self-fertilization. Genetics 182:1129-1139

Bromham, L, Hua X, Lanfear R, Cowman PF (2015) Exploring the relationships between mutation rates, life history, genome size, environment, and species richness in flowering plants. The American Naturalist 185(4): 507-524

Bruneau A, Mercure M, Lewis GP, Herendeen PS (2008) Phylogenetic patterns and diversification in the caesalpinioid legumes. Botany 86:697–718

Cai Z, Guisinger M, Kim H-G, Ruck E, Blazier JC, McMurtry V, Kuehl JV, Boore J, Jansen RK (2008) Extensive reorganization of the plastid genome of *Trifolium subterraneum* (Fabaceae) is associated with numerous repeated sequences and novel DNA insertions. J Mol Evol 67(6): 696-704

Cho Y, Mower JP, Qiu Y-L, Palmer JD (2004) Mitochondrial substitution rates are extraordinarily elevated and variable in a genus of flowering plants. PNAS 101(51):17741-17746

Chumley TW, Palmer JD, Mower JP, Fourcade HM, Calie PJ, Boore JL, Jansen RK (2006) The complete chloroplast genome sequence of *Pelargonium × hortorum*: organization and evolution of the largest and most highly rearranged chloroplast genome of land plants. Molecular Biology and Evolution 23(11): 2175–2190.

Cosner ME, Raubeson LA, Jansen RK (2004) Chloroplast DNA rearrangements in Campanulaceae: phylogenetic utility of highly rearranged genomes. BMC Evol Biol 4:27. doi: 10.1186/1471-2148-4-27

Covello PS, Gray MW (1992) Silent mitochondrial and active nuclear genes for subunit 2 of cytochrome c oxidase (*cox2*) in soybean: evidence for RNA-mediated gene transfer. The EMBO Journal 11(11):3815-3820

Cronk Q, Ojeda I, Pennington RT (2006) Legume comparative genomics: progress in phylogenetics and phylogenomics. Current Opinion in Plant Biology 9(2): 99–103.

Cuenca A, Petersen G, Seberg O, Davis JI, Stevenson DW (2010) Are substitution rates and RNA editing correlated? BMC Evolutionary Biology 10:349

Darling AE, Mau B, Pema NT. 2010. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS ONE 5(6): e11147.

Doebley J, Durbin M, Golenberg EM, Clegg MT, Ma DP (1990) Evolutionary analysis of the large subunit of carboxylase (*rbcL*) nucleotide sequence among the grasses (Gramineae). Evolution 44(4): 1097-1108

Dong W, Xu C, Cheng T, Zhou S. 2013. Complete chloroplast genome of *Sedum sarmentosum* and chloroplast genome evolution in Saxifragales. PloS One 8(10): e77965.

Downie SR, Palmer JD (1992) Use of chloroplast DNA rearrangements in reconstructing plant phylogeny. In Soltis PS, Soltis DE, Doyle JJ (eds) Molecular Systematics of Plants, Chapman and Hall, New York, pp14–35

Doyle JJ, Doyle JL, Ballenger JA, Palmer JD (1996) The distribution and phylogenetic significance of a 50-kb chloroplast DNA inversion in the flowering plant family Leguminosae. Molecular Phylogenetics and Evolution 5: 429–438.

Doyle JJ, Doyle JL (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochemical Bulletin 19:11-15.

Doyle JJ, Doyle JL, Palmer JD (1995) Multiple independent losses of two genes and one intron from legume chloroplast genomes. Systematic Botany 20(3): 272–294.

Drescher A, Ruf S, Calsa T, Carrer H, Bock R (2000) The two largest chloroplast genome-encoded open reading frames of higher plants are essential genes. Plant J 22:97–104. doi: 10.1046/j.1365-313x.2000.00722

Dufresne A, Garczarek L, Partensky F. 2005. Accelerated evolution associated with genome reduction in a free-living prokaryote. Genome Biology 6(2): R14.

Dugas, DV, Hernandex D, Koenen E, Schwarz E, Straub S, Hughes CE, Jansen RK, Nageswara-Rao M, Staats M, Trujillo J, Hajrah NH, Alharbi NS, Al-Malki AL, Sabir JSM, Bailey CD (2015) Mimosoid legume plastome evolution: IR expansion, tandem repeat expansions, and accelerated rate of evolution in clpP. Scientific Reports 5:16958

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32(5): 1792–1797.

Eyre-Walker A, Gaut B (1997) Correlated rates of synonymous site evolution across plant genomes. Mol Biol Evol 14:455–460.

Fajardo D, Senalik D, Ames M, Zhu H, Steffan SA, Harbut R, Polashock J, Vorsa N, Gillespie E, Kron K, Zalapa JE (2013) Complete plastid genome sequence of *Vaccinium macrocarpon*: structure, gene content, and rearrangements revealed by next generation sequencing. Tree Genetics & Genomes 9: 489–498.

Gantt JS, Baldauf SL, Calie PJ, Weeden NF, Palmer JD (1991) Transfer of *rpl22* to the nucleus greatly preceded its loss from the chloroplast and involved the gain of an intron. The EMBO journal 10(10): 3073-3078.

Gao L, Wang B, Wang Z-W, Zhou Y, Su Y-J, Wang T (2013) Plastome sequences of *Lygodium japonicum* and *Marsilea crenata* reveal the genome organization transformation from basal ferns to core leptosporangiates. Genome Biology and Evolution 5(7): 1403–1407.

Gaut B, Morton B, McCaig B, Clegg M (1996) Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene *adh* parallel rate differences at the plastid gene *rbcL*. Proc Natl Acad Sci USA 93:10274–10279

Gaut B, Muse S, Clark W, Clegg M (1992) Relative rates of nucleotide substitution at the *rbcL* locus of monocotyledonous plants. J Mol Evol 35:292–303.

Gaut BS (1998) Molecular clocks and nucleotide substitution rates in higher plants. In: Hecht MK, Macintyre RJ, Clegg MT (eds) Evolutionary biology. Plenum Press, New York, pp 93–120

Green BR. 2011. Chloroplast genomes of photosynthetic eukaryotes. The Plant Journal 66(1): 34–44.

Grewe F, Gubbels EA, Mower JP (2015) The mitochondrial genome evolution of the geranium family: elevated substitution rates decrease genomic complexity. Plant and animal genome XXIII San Diego, CA, USA https://pag.confex.com/pag/xxiii/webprogram/Paper14712.html

Guisinger MM, Chumley TW, Kuehl JV, Boore JL, Jansen RK (2010) Implications of the plastid genome sequence of *Typha* (Typhaceae, Poales) for understanding genome evolution in Poaceae. J Mol Evol 70:149–166

Guisinger MM, Kuehl JV, Boore JL, Jansen RK (2008) Genome-wide analyses of Geraniaceae plastid DNA reveal unprecedented patterns of increased nucleotide substitutions. Proc Natl Acad Sci USA 105:18424–18429

Guisinger MM, Kuehl JV, Boore JL, Jansen RK (2011) Extreme reconfiguration of plastid genomes in the angiosperm family Geraniaceae: Rearrangements, repeats, and codon usage. Mol Biol Evol 28:583–600

Guo W, Grewe F, Cobo-Clark A, Fan W, Duan Z, Adams RP, Schwarzbach AE, Mower JP (2014) Predominant and substoichiometric isomers of the plastid genome coexist within *Juniperus* plants and have shifted multiple times during Cupressophyte evolution. Genome Biology and Evolution 6(3): 580–590.

Guo X, Castillo-Ramírez S, González V, Bustos P, Fernández-Vázquez JL, Santamaria RI, Arellano J, Cevallos MA, Dávila G (2007) Rapid evolutionary change of common bean (*Phaseolus vulgaris* L) plastome and the genomic diversification of legume chloroplasts. BMC Genomics 8(1): 228-244.

Gurdon C, Maliga P (2014) Two distinct plastid genome configurations and unprecedented intraspecies length variation in the accD coding region in *Medicago truncatula*. DNA Research 21: 417–427.

Haberle RC, Fourcade HM, Boore JL, Jansen RK (2008) Extensive rearrangements in the chloroplast genome of *Trachelium caeruleum* are associated with repeats and tRNA genes. Journal of Molecular Evolution 66(4): 350–361.

Hirao T, Watanabe A, Kurita M, Kondo T, Takata K (2008) Complete nucleotide sequence of the *Cryptomeria japonica* D. Don. chloroplast genome and comparative chloroplast genomics: diversified genomic structure of coniferous species. BMC Plt Biol 8:70. doi: 10.1186/1471-2229-8-70

Hoot SB, Palmer JD (1994) Structural rearrangements, including parallel inversions, within the chloroplast genome of *Anemone* and related genera. Journal of Molecular Evolution 38: 274-281.

Islam MS, Studer B, Byrne SL, Farrell JD, Panitz F, Bendixen C, Møller IM, Asp T (2013) The genome and transcriptome of perennial ryegrass mitochondria. BMC Genomics 14:202-223

Jansen RK, Cai Z, Raubeson LA, Daniell H, dePamphilis CW, Leebens-Mack J, Müller KF, Guisinger-Bellian M, Haberle RC, Hansen AK, Chumley TW, Lee S-B, Peery R, McNeal JR, Kuehl JV, Boore JL (2007) Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc Natl Acad Sci USA 104:19369–19374.

Jansen RK, Palmer JD (1987) A chloroplast DNA inversion marks an ancient evolutionary split in the sunflower family (Asteraceae). Proceedings of the National Academy of Sciences USA 84(16): 5818–5822.

Jansen RK, Ruhlman TA (2012) Plastid genomes of seed plants. In: Bock R, Knoop V (eds) Genomics of chloroplasts and mitochondria, Advances in

Photosynthesis and Respiration 35. Springer, Dordrecht Advances, pp 103–126

Jansen RK, Wojciechowski MF, Sanniyasi E, Lee S-B, Daniell H (2008) Complete plastid genome sequence of the chickpea (*Cicer arietinum*) and the phylogenetic distribution of *rps12* and *clpP* intron losses among legumes (Leguminosae). Molecular Phylogenetics and Evolution 48(3): 1204–1217.

Käss E, Wink M (1997) Phylogenetic relationships in the Papilionoideae (Family Leguminosae) based on nucleotide sequences of cpDNA (*rbcL*) and ncDNA (ITS 1 and 2). Molecular Phylogenetics and Evolution 8(1): 65–88.

Katoh K, Standley DM (2013) MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol. Biol. Evol. 30(4): 772-780

Kay K, Whittall J, Hodges S (2006) A survey of nuclear ribosomal internal transcribed spacer substitution rates across angiosperms: an approximate molecular clock with life history effects. BMC Evol Biol 6:36.

Kazakoff SH, Imelfort M, Edwards D, Koehorst J, Biswas B, Batley J, Scott PT, Gresshoff PM (2012) Capturing the biofuel wellhead and powerhouse: the chloroplast and mitochondrial genomes of leguminous feedstock tree *Pongamia pinnata*.PLoS ONE 7(12):e51687

Kim K-J, Choi K-S, Jansen RK (2005) Two chloroplast DNA inversions originated simultaneously during the early evolution of the sunflower family (Asteraceae). Molecular Biology and Evolution 22(9): 1783–1792.

Kim K-J, Lee H-L (2005) Widespread occurrence of small inversions in the chloroplast genomes of land plants. Molecules and Cells 19(1): 104–113.

Kimura, M. (1984) The neutral theory of molecular evolution. Cambridge University Press.

Knox EB (2014) The dynamic history of plastid genomes in the Campanulaceae sensu lato is unique among angiosperms. Proceedings of the National Academy of Sciences USA 111: 11097-11102.

Kode V, Mudd EA, Iamtham S, Day A (2005) The tobacco plastid *accD* gene is essential and is required for leaf development. Plt J 44:237–44.

Koller B, Delius H (1980) *Vicia faba* chloroplast DNA has only one set of ribosomal RNA genes as shown by partial denaturation mapping and R-loop analysis. Mol Gen Genetics 178:261–269.

Krause K (2012) Plastid genomes of parasitic plants: A trail of reductions and losses. In: Bullerwell CE ed. Organelle Genetics. Heidelberg: Springer Berlin. 79–103.

Langmead B, Salzberg S (2012) Fast gapped-read alignment with Bowtie 2. Nature Methods 9:357-359.

Laroche J, Li P, Maggia L, Bousquet J (1997) Molecular evolution of angiosperm mitochondrial introns and exons. Proc Natl Acad Sci USA 94:5722–7.

Lavin M, Doyle JJ, Palmer JD (1990) Evolutionary significance of the loss of the chloroplast-DNA inverted repeat in the Leguminosae subfamily Papilionoideae. Evolution 44(2): 390-402

Lee H-L, Jansen RK, Chumley TW, Kim K-J (2007) Gene relocations within chloroplast genomes of *Jasminum* and *Menodora* (Oleaceae) are due to multiple overlapping inversions. Molecular Biology and Evolution 24(5): 1161–1180.

Lewis G, Schrire B, Mackinder B, Lock M (2005) Legumes of the world. United Kingdom: Royal Botanic Gardens, Kew

Liston A (1995) Use of the polymerase chain reaction to survey for the loss of the inverted repeat in the legume chloroplast genome. In: Crisp MD, Doyle JJ eds. Advances in Legume Systematics 7, Phylogeny: 31-40. Royal Botanic Gardens, Kew.

LPWG (2013) Legume phylogeny and classification in the 21st century: Progress, prospects and lessons for other species-rich clades. Taxon 62(2): 217-248

Lu M-Z, Szmidt AE, Wang XR (1998) RNA editing in gymnosperms and its impact on the evolution of the mitochondrial cox1 gene. Plant Mol Biol 37:225-234

Luckow M, Miller JT, Murphy DJ, Livshultz T (2003) A phylogenetic analysis of the Mimosoideae (Leguminosae) based on chloroplast DNA sequence data. In: Klitgaard BB, Bruneau A (eds) Advances in Legume Systematics 10:197-220. United Kingdom: Royal Botanic Gardens, Kew.

Luo M-C, You FM, Li P, Wang J-R, Zhu T, Dandekar AM, Leslie CA, Aradhya M, McGuire PE, Dvorak J (2015) Synteny analysis in Rosids with a walnut physical map reveals slow genome evolution in long-lived woody perennials. BMC Genomics 16:707

Lynch M, Koskella B, Schaack S (2006) Mutation pressure and the evolution of organelle genomic architecture. Science 311(5768): 1727–1730

MacKay J, Liu W, Whetten R, Sederoff RR, O'Malley DM (1995) Genetic analysis of cinnamyl alcohol dehydrogenase in loblolly pine: single gene inheritance, molecular characterization and evolution. Mol Gen Genetics 247:537–45. doi: 10.1007/BF00290344

Magee AM, Aspinall S, Rice DW, Cusack BP, Sémon M, Perry AS, Stefanović S, Milbourne D, Barth S, Palmer JD, Gray JC, Kavanagh TA, Wolfe KH (2010) Localized hypermutation and associated gene losses in legume chloroplast genomes. Genome Research 20(12): 1700–1710

Maier RM, Neckermann K, Igloi GL, Kössel H (1995) Complete sequence of the maize chloroplast genome: gene content, hotspots of divergence and fine tuning of genetic information by transcript editing. J Mol Biol 251:614–28.

Marechal A, Parent JS, Veronneau-Lafortune F, Joyeaux A, Lang BF, Brisson N (2009) Whirly proteins maintain plastid genome stability in Arabidopsis. PNAS 106: 14693-14698

Martin A, Palumbi S (1993) Body size, metabolic rate, generation time, and the molecular clock. Proc Natl Acad Sci USA 90:4087–4091. doi: 10.1073/pnas.90.9.4087

Martin GE, Rousseau-Gueutin M, Cordonnier S, Lima O, Michon-Coudouel S, Naquin D, de Carvalho JF, Aïnouche M, Salmon A, Aïnouche A (2014) The first complete chloroplast genome of the Genistoid legume *Lupinus luteus*: evidence for a novel major lineage-specific rearrangement and new insights regarding plastome evolution in the legume family. Annals of Botany 113(7): 1197–1210.

Martínez-Alberola F, del Campo EM, Lázaro-Gimeno D, Mezquita-Claramonte S, Molins A, Mateu-Andrés I, Pedrola-Monfort J, Casano LM, Barreno E (2013) Balanced gene losses, duplications and intensive rearrangements led to an unusual regularly sized genome in *Arbutus unedo* chloroplasts. PLoS ONE 8(11): e79685.

McCoy SR, Kuehl JV, Boore JL, Raubeson LA (2008) The complete plastid genome sequence of *Welwitschia mirabilis*: an unusually compact plastome with accelerated divergence rates. BMC Evolutionary Biology 8(1) p130-146

Miller, M.A., Pfeiffer, W., and Schwartz, T. (2010) Creating the CIPRES Science Gateway for inference of large phylogenetic trees. Proceedings of the Gateway Computing Environments Workshop (GCE), 14 Nov. 2010, New Orleans, LA pp 1 - 8

Milligan BG, Hampton JN, Palmer JD (1989) Dispersed repeats and structural reorganization in subclover chloroplast DNA. Molecular Biology and Evolution 6(4) 355-368.

Mower JP, Sloan DB, Alverson AJ (2012) Plant mitochondrial genome diversity: the genomics revolution. In: Wendel JF, Greilhuber J, Doležel J, Leitch IJ (eds) Plant Genome Diversity, Volume 1. Springer, New York, pp 123-144.

Mower JP, Touzet P, Gummow JS, Delph LF, Palmer JD (2007) Extensive variation in synonymous substitution rates in mitochondrial genes of seed plants. BMC Evol Biol 7:135-149

Naito K, Kaga A, Tomooka N, Kawase M (2013) De novo assembly of the complete organell genome sequences of azuki bean (*Vigna angularis*) using next-generation sequencers. Breeding Science 63:176-182

Nugent JM, Palmer JD (1991) RNA-mediated transfer of the gene *coxII* from the mitochondrion to the nucleus during flowering plant evolution. Cell 66:473-481

Palmer JD (1983) Chloroplast DNA exists in two orientations. Nature 301(5895): 92–93.

Palmer JD (1985) Comparative organization of chloroplast genomes. Annual Review of Genetics 19: 325-354.

Palmer JD (1991) Plastid chromosomes: Structure and evolution. In: Bogorad L, Vasil IK (eds) Cell culture and somatic cell genetics of plants, Vol 7A, Academic Press, New York, pp 5-53

Palmer JD, Adams KL, Cho Y, Parkinson CL, Qiu Y-L, Song K (2000) Dynamic evolution of plant mitochondrial genomes: mobile genes and introns and highly variable mutation rates. PNAS 97(13): 6960-6966

Palmer JD, Herbon LA (1988) Plant mitochondrial DNA evolves rapidly in structure, but slowly in sequence. J Mol Evol 28:87-97

Palmer J, Osorio B, Aldrich J, Thompson W (1987) Chloroplast DNA evolution among legumes: Loss of a large inverted repeat occurred prior to other sequence rearrangements. Curr Genet 11:275–286

Palmer JD, Thompson WF (1981) Rearrangements in the chloroplast genomes of mung bean and pea. Proc Natl Acad Sci USA 78:5533–5537

Park S, Jansen RK, Park S (2015) Complete plastome sequence of *Thalictrum coreanum* (Ranunculaceae) and transfer of the *rpl32* gene to the nucleus in the ancestor of the subfamily Thalictroideae. BMC Plant Biology 15: 40

Parkinson CL, Mower JP, Qiu Y-L, Shirk AJ, Song K, Young ND, dePamphilis CW, Palmer JD (2005) Multiple major increases and decreases in mitochondrial substitution rates in the plant family Geraniaceae. BMC Evol Biol 5:73-85

Peltier J-B, Ripoll DR, Friso G, Rudella A, Cai Y, Ytterberg J, Giacomelli L, Pillardy J, van Wijk KJ (2004) Clp protease complexes from photosynthetic and nonphotosynthetic plastids and mitochondria of plants, their predicted three-dimensional structures, and functional implications. J Biol Chem 279:4768–4781

Perry AS, Wolfe KH (2002) Nucleotide substitution rates in legume chloroplast DNA depend on the presence of the inverted repeat. J Mol Evol 55:501–508

Petit RJ, Hampe A (2006) Some evolutionary consequences of being a tree. Ann Rev Ecol Evol Syst 37:187-214

Raubeson LA, Jansen RK (2005) Chloroplast genomes of plants. In: Henry RJ (ed) Plant Diversity and Evolution: Genotypic and Phenotypic Variation in Higher Plants. Wallingford: CAB International. 45-68

Rowan BA, Oldenburg DJ, Bendich AJ (2010) RecA maintains the integrity of chloroplast DNA molecules in Arabidopsis. J Exp Bot 61:2575-2588

Roy S, Ueda M, Kadowaki K-I, Tsutsumi N (2010) Different status of the gene for ribosomal protein S16 in the chloroplast genome during evolution of the genus *Arabidopsis* and closely related species. Genes and Genetic Systems 85(5): 319–326

Ruhlman TA, Jansen RK (2014) The Plastid Genomes of Flowering Plants. In: Maliga P (ed) Chloroplast Biotechnology: Methods and Protocols, Methods in Molecular Biology, vol 1132. Springer Science and Business Media, New York, pp 3-38

Sabir J, Schwarz EN, Ellison N, Zhang J, Baeshen NA, Mutwakil M, Jansen RK, Ruhlman TA (2014) Evolutionary and biotechnology implications of plastid genome variation in the inverted-repeat-lacking clade of legumes. Plant Biotechnology Journal 12(6): 743–754

Saski C, Lee S-B, Daniell H, Wood TC, Tomkins J, Kim H-G, Jansen RK (2005) Complete chloroplast genome sequence of *Glycine max* and comparative analyses with other legume genomes. Plant Molecular Biology 59(2): 309–322

Schattner P, Brooks AN, Lowe TM (2005) The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Research 33: 686-689

Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C, Bouck J, Gibbs R, Hardison R, Miller W (2000) Pipmaker - a web server for aligning two genomic DNA sequences. Genome Research 10(4): 577-586

Schwarz EN, Ruhlman T, Sabir JSM, Hajrah NH, Alharbi NS, Al-Malki AL, Bailey CD, Jansen RK (2015) Plastid genomes reveal parallel inversions and multiple losses of *rps16* in papilionoids. J Syst Evol 53(5): 458-468

Shedge V, Arrieta-Montiel M, Christensen AC, Mackenzie SA (2007) Plant mitochondrial recombination surveillance requires unusual *recA* and *mutS* homologs. The Plant Cell 19:1251-1264

Sloan DB, Alverson AJ, Chuckalovcak JP, Wu M, McCauley DE,Palmer JD, Taylor DR (2012) Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. PLoS Biol 10(1)

Sloan DB, Alverson AJ, Wu M, Palmer JD, Taylor DR (2012) Recent acceleration of plastid sequence and structural evolution coincides with extreme mitochondrial divergence in the angiosperm genus *Silene*. Gen Biol Evol 4:294–306

Sloan DB, Barr CM, Olson MS, Keller SR, Taylor DR (2008) Evolutionary rate variation at multiple levels of biological organization in plant mitochondrial DNA. Mol Biol Evol 25(2): 243-246

Smith SA, Donoghue MJ (2008) Rates of molecular evolution are linked to life history in flowering plants. Science 322:86–89.

Stamatakis A, Hoover P, Rougemont J (2008) A rapid bootstrap algorithm for the RAxML web servers. Systematic Biology 75(5): 758-771.

Sveinsson S, Cronk Q (2014) Evolutionary origin of highly repetitive plastid genomes within the clover genus (*Trifolium*). BMC Evolutionary Biology 14: 228

Tangphatsornruang S, Sangsrakru D, Chanprasert J, Uthaipaisanwong P, Yoocha T, Jomchai N, Tragoonrung S (2010) The chloroplast genome sequence of mungbean (*Vigna radiata*) determined by high-throughput pyrosequencing:

structural organization and phylogenetic relationships. DNA Research 17(1): 11–22

Tsudzuki J, Nakashima K, Tsudzuki T, Hiratsuka J, Shibata M, Wakasugi T, Sugiura M (1992) Chloroplast DNA of black pine retains a residual inverted repeat lacking rRNA genes: nucleotide sequences of *trnQ trnK psbA trnI* and *trnH* and the absence of *rps16*. Molecular and General Genetics 232(2): 206–214

Ueda M, Nishikawa T, Fujimoto M, Takanashi H, Arimura S, Tsutsumi N, Kadowaki K (2008) Substitution of the gene for chloroplast *rps16* was assisted by generation of a dual targeting signal. Molecular Biology and Evolution 25(8): 1566–1575

Weng M-L, Blazier CJ, Govindu M, Jansen RK (2014) Reconstruction of the ancestral plastid genome in Geraniaceae reveals a correlation between genome rearrangements, repeats and nucleotide substitution rates. Mol Biol Evol 31(3):645–659

Weng M-L, Ruhlman T, Gibby M, Jansen R (2012) Phylogeny, rate variation, and genome size evolution of *Pelargonium* (Geraniaceae). Mol Phylogen Evol 64:654–670.

Whittle C-A, Johnston M (2003) Broad-scale analysis contradicts the theory that generation time affects molecular evolutionary rates in plants. J Mol Evol 56:223–233.

Wicke S, Schneeweiss GM, dePamphilis CW, Müller KF, Quandt D (2011) The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. Plt Mol Biol 76: 273–297

Williams A, Boykin L, Howell K, Nevill PG, Small I (2015) The complete sequence of the *Acacia ligulata* chloroplast genome reveals a highly divergent *clpP1* gene. Plos One 10: e0125768

Wilson M, Gaut B, Clegg M (1990) Chloroplast DNA evolves slowly in the palm family (Arecaceae). Mol Biol Evol 7:303–14

Wojciechowski MF, Lavin M, Sanderson MJ (2004) A phylogeny of legumes (Leguminosae) based on analysis of the plastid *matK* gene resolves many well-supported subclades within the family. Amer J Bot 91(11): 1846–1862

Wolf YI, Koonin EV (2013) Genome reduction as the dominant mode of evolution. BioEssays 35(9): 829–837

Wolfe KH (1988) The site of deletion of the inverted repeat in pea chloroplast DNA contains duplicated gene fragments. Curr Genet 13:97-99

Wolfe KH, Li WH, Sharp PM (1987) Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc Natl Acad Sci USA 84:9054–9058

Wu CS, Chaw SM (2015) Evolutionary stasis in cycad plastomes and the first case of plastome GC-biased gene conversion. Gen Biol Evol 7:2000–2009

Wu CS, Chaw SM (2014) Highly rearranged and size-variable chloroplast genomes in conifers II clade (cupressophytes): evolution towards shorter intergenic spacers. Plt Biotech J 12(3): 344–353

Wu C, Li W (1985) Evidence for higher rates of nucleotide substitution in rodents than in man. Proc Natl Acad Sci USA 82:1741–5

Wu CS, Lai YT, Lin CP, Wang YN, Chaw SM (2009) Evolution of reduced and compact chloroplast genomes (cpDNAs) in gnetophytes: Selection toward a lower-cost strategy. Mol Phylogen Evol 52(1): 115–124

Wyman SK, Jansen RK, Boore JL (2004) Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20: 3252–3255.

Xu YZ, Arrieta-Montiel MP, Virdi KS, de Paula WBM, Widhalm JR, Basset GJ, Davila JI, Elthon TE, Elowsky CG, Sato SJ, Clemente TE, Mackenzie SA (2011) MutS HOMOLOG1 is a nucleoid protein that alters mitochondrial and plastid properties and plant response to high light. Plant Cell 23:3428-3441

Yokoyama S, Harry D (1993) Molecular phylogeny and evolutionary rates of alcohol dehydrogenases in vertebrates and plants. Mol Biol Evol 10:1215–26

Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Research 18(5): 821-829.

Zhang J, Ruhlman T, Sabir J, Blazier JC, Weng M-L, Park S, Jansen RK (2016) Coevolution between nuclear encoded DNA replication, recombination and repair genes and plastid genome complexity. Gen Biol Evol 8.

Zhong BJ, Yonezawa T, Zhong Y, Hasegawa M (2009) Episodic evolution and adaptation of chloroplast genomes in ancestral grasses. PLoS One 4:e5297

Zhu A, Guo W, Gupta S, Fan W, Mower JP (2015) Evolutionary dynamics of the plastid inverted repeat: the effects of expansion, contraction, and loss on substitution rates. The New Phytol 209:1747–56

Zhu A, Guo W, Jain K, Mower JP (2014) Unprecedented heterogeneity in the synonymous substitution rate within a plant genome. Mol Biol Evol 31(5): 1228-1236