The Correlation of Evolutionary Rate with Pathway Position in Plant Terpenoid Biosynthesis

Heather Ramsay,*† Loren H. Rieseberg,† and Kermit Ritland*

*Faculty of Forestry, Department of Forest Sciences, University of British Columbia, Vancouver, British Columbia, Canada; and †Department of Botany, University of British Columbia, Vancouver, British Columbia, Canada

Genes are expected to face stronger selective constraint and to evolve more slowly if they encode enzymes upstream as opposed to downstream in metabolic pathways, because upstream genes are more pleiotropic, being required for a wider range of end products. However, few clear examples of this trend in evolutionary rate variation exist. We examined whether genes involved in plant terpenoid biosynthesis exhibit such a pattern, using data for 40 genes from five fully sequenced angiosperms, *Oryza*, *Vitis*, *Arabidopsis*, *Populus*, and *Ricinus*. Our results show that $d_{s}/d_{S}$ does in fact correlate with pathway position along pathways converting glucose to the terpenoid phytohormones abscisic acid, gibberellic acid (GA), and brassinosteroids. Upstream versus downstream rate variation is particularly strong in the GA pathway. In contrast, we found no or little apparent variation in $d_{s}/d_{S}$ with gene copy number. We also introduce a new measure of pathway position, the Pathway Pleiotropy Index (PPI), which counts groups of enzymes between pathway branch points. We found that this measure is superior to pathway position in explaining variation in $d_{s}/d_{S}$ along each pathway. Although at least 8 of the 40 genes showed evidence of positive selection, correlations of $d_{s}/d_{S}$ with PPI remain significant when these genes are removed. Therefore, our results are consistent with the prediction that selective constraint is progressively relaxed along metabolic pathways.

Introduction

As species diverge through evolution, their genes diverge at unequal rates, and genomes become a mosaic of conservation and divergence. These differences in divergence can be explained as differential selective forces acting on the various loci, but what sets up these differences in selective pressure among genes?

Networks of gene interactions offer an intriguing context for answering this question. Functional genomics and whole-genome sequencing have opened the way to construct accurate representations of cellular networks and to more fully understand interactions among genes. Studies on a variety of network types (metabolic, protein–protein interaction, signaling, and development) have shown that network properties such as node connectivity and centrality within the network correlate with evolutionary rate variation in a variety of species (Cork and Purugganan 2004; Jordan et al. 2004). For example, highly connected nodes (i.e., enzymes or other proteins) consistently undergo stronger selective constraint than nodes with fewer connections in both metabolic and protein–protein interaction networks (Fraser et al. 2002, 2003; Jordan et al. 2003a, 2003b; Hahn et al. 2004; Makino et al. 2006; Vitkup et al. 2006).

A less-explored aspect is the structure of metabolic pathways. Many known metabolic pathways are structured such that the initial substrates in the pathway are ultimately transformed into several end products, based on the channeling of substrate into various branches of the pathway downstream. Therefore, upstream genes should be more pleiotropic than those downstream and so should face greater selective constraint (Waxman and Peck 1998; Otto 2004). This gradient of selective constraint should then produce low evolutionary rates upstream and high rates downstream, a pattern of evolutionary rate variation that has been termed “positional rate variation” (PRV) (Rausher et al. 1999).

PRV has been rigorously demonstrated in the anthocyanin pathway of higher plants. Rausher et al. (1999) show that among six enzymes involved in anthocyanin synthesis, the three upstream enzymes evolve more slowly than the three downstream enzymes. Lu and Rausher (2003) and Rausher et al. (2008) confirm that this is due to a relaxation of selective constraint, not stronger positive selection, on downstream genes. These studies establish an occurrence of PRV that holds true in three plant lineages and that can be demonstrated over both recent and ancient evolutionary time scales.

Yet, is PRV truly a general effect in the evolution of metabolic genes? A study of four genes involved in plant isoprene synthesis suggests a similar trend of stronger selective constraint on upstream enzymes (Sharkey et al. 2005), but work on nucleotide variation in *Arabidopsis thaliana* phenylpropanoid metabolism reveals no correlation of pathway position and various measures of nucleotide divergence (Ramos-Onsins et al. 2008). Clearly, PRV requires further study to be established as a general phenomenon.

In this study, we examine whether PRV occurs among 40 genes involved in plant terpenoid synthesis, using data from five fully sequenced angiosperms. To capture meaningful changes in pleiotype along the pathway, we trace the flow of metabolites from glycolysis through to the synthesis of four terpenoid derivatives. As the most upstream portion of this system, glycolysis plays a critical role in producing small precursors that are required for nearly every primary biosynthetic process (Plaxton 1996; Heldt and Heldt 2005). One of these processes is the synthesis of terpenoid precursors such as farnesyl diphosphate, which is also used in the production of a wide range of metabolites needed for growth and plant defense (Liu et al. 2005; Cheng et al. 2007). These include lutein, a carotenoid important for the assembly and function of the photosynthetic light harvesting complex (Pogson et al. 1996), as well as the phytohormones abscissic acid (ABA), gibberellic acid (GA), and brassinosteroids (BRs) (Buchanan et al. 2000). The enzymes involved in the transformation of glucose to these four end products are therefore less...
pleiotropic downstream than they are upstream, and offer an ideal setting to study PRV. The entire pathway is depicted in figure 1 (information derived from Plaxton 1996; Clouse and Sasse 1998; Szekeres and Koncz 1998; Buchanan et al. 2000; Hedden and Phillips 2000; Naik et al. 2003; Sakamoto et al. 2004; Liu et al. 2005; Nambara and Marion-Poll 2005; Grennan 2006; Cheng et al. 2007; KEGG PATHWAY Database 2008).

To improve detection of PRV, we use a new measure termed the “Pathway Pleiotropy Index” (PPI). Rather than counting enzymes in sequence from most upstream to most downstream, the PPI numbers “groups of enzymes” relative to pathway branch points. We compare this measure with the simpler approach of correlating evolutionary rates with pathway positions. We expect that the PPI will correlate more strongly with evolutionary rates because it tracks changes in pleiotropy along the pathways; enzymes between pathway branch points are responsible for the same set of products and therefore do not differ in pleiotropy. In fact, the anthocyanin pathway may show such strong PRV precisely because virtually every enzyme acts as a pathway branch point and therefore each enzyme is successively less pleiotropic (Harbourne 1988; Winkel 2006).

In addition, if \( d_{sl}/d_s \) does not correlate with pathway position, it may be that rate variation is more strongly influenced by another factor that mitigates the role of pathway structure. For example, after gene duplication, selective constraint is relaxed for a time, ostensibly due to redundancy (Lynch and Conery 2000). Also, gene duplication may offer an escape from the constraining effects of pleiotropy (Hoekstra and Coyne 2007), resulting in a negative correlation of copy number with pathway position, rather than a positive correlation of pathway position and \( d_{sl}/d_s \). Therefore, variation in gene copy number may explain instances where PRV is not found.

Here, we examine whether plant terpenoid synthesis exhibits PRV by addressing four specific questions. First, we test whether statistically significant variation in \( d_{sl}/d_s \) exists among enzymes. Second, we test whether this variability is explained by pathway position and/or by PPI. Third, we examine whether \( d_{sl}/d_s \) varies with gene copy number, and whether copy number varies with pathway position. Finally, we use likelihood-based methods (Church et al. 2007) to explore whether differences in \( d_{sl}/d_s \) arise from variation in selective constraint versus positive selection.

Methods

Study Species

We used all angiosperm genomes that were available at the time of study: *Oryza sativa*, *Vitis vinifera*, *A. thaliana*, *Ricinus communis*, and *Populus trichocarpa*. Protein and coding sequences for each species were downloaded from the following websites on March 27, 2008:

- http://genome.jgi-psf.org/Poptr1_1/Poptr1_1.download.ftp.html
- ftp://ftp.plantbiology.msu.edu/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseudomolecules/version_5.0/all.chrs

Finding Orthologs and Counting Paralogs

Using the KEGG PATHWAY Database (2008), 47 genes involved in the conversion of glucose to lutein,
abscisic acid, gibberellic acid, and brassinosteroids were selected. *Arabidopsis* protein sequences for the enzymes were gathered from the KEGG PATHWAY Database February 27, 2008 (glycolysis enzymes were gathered on May 23, 2008). A list of genes and accession numbers for these queries are provided in supplemental table 1, Supplementary Material online. Proteins from all five genomes were Blasted with the *Arabidopsis* protein queries (BlastP; Altschul et al. 1997). A multisquence query was used if more than one sequence was given in the KEGG PATHWAY Database (2008). Blast results were parsed for 25% amino acid identity over 50% of the query sequence length (if more than one query was used, this was taken to be 50% of the length of the shortest query). Processed hits for all species were aligned using MUSCLE (Edgar 2004). If the alignment quality was poor, Blast results were reprocessed by increasing the percent identity cutoff by 5% until the resulting multiple alignment was acceptable. The initial methionine was deleted before alignment in all cases. Next, these alignments were used to construct maximum likelihood protein trees with PhyML (Guindon and Gascuel 2003). For each alignment PhyML was run with the following parameters: Jones-Taylor-Thornton protein evaluation model (Jones et al. 1992) with 4 gamma categories; estimated proportion of invariable sites and gamma distribution parameter; 100 bootstrap replicates; Neighbor-Joining starting tree; and optimized both topology and branch lengths.

The resulting trees were visually inspected for an orthologous group: a clade with 50% or greater bootstrap support and which contained the *Arabidopsis* query plus at least one representative of each of the other four taxa (for *Arabidopsis*, the chosen ortholog was always one of the query sequences). If no such clade could be found, the gene was not included in any further analysis. If collapsing nodes with poor bootstrap support reduced the protein family tree to a large polytomy, the gene was also rejected from further analysis due to ambiguous orthology. If any taxon had more than one sequence in the ortholog clade, the sequence on the shorter branch was chosen. Because multiple sequences from any one species tended to appear together in a subclade, the presence of multiple sequences per species did not obscure orthology for the genes used in subsequent analyses. Paralogs were counted as the number of sequences for each lineage in the ortholog clade. Names of orthologs, as used in the genome file for each species, are provided in supplemental table 2, Supplementary Material online. Note that glycolytic genes are known to have neutral evolution and therefore can be left unambiguously aligned regions; analyses were carried out with both the original alignments and with the ambiguous regions masked in these four genes in order to ensure that alignment ambiguity did not sway results (all alignments are available on request).

Using this alignment of five orthologs, *dS/dG* ratios along each lineage were calculated in CODEML (Yang 1997) using the tree (*Oryza*, *Vitis*, *(Arabidopsis, (Ricinus, Populus)))*, with model = 1 (free ratios, to give separate *dS/dG* values for each branch), and get SE = 1 (obtain standard errors of *dS/dG* estimates). Five replicate runs were carried out, and the run with greatest absolute value of ln L was selected. From this selected run, *dS/dG* values for *Vitis, Arabidopsis, Ricinus,* and *Populus* were estimated (*Oryza* served as an outgroup; see fig. 2; tree topology from Soltis et al. 2005). Values of *dS/dG* for which *dS* was less than 0.005 or more than 2.00 were discarded, as were values for which *S* was less than 10. Except for tests for positive selection, the four ingroup lineages were treated separately to avoid complications due to lineage-specific differences in variability and to allow us to ask whether lineages show similar trends.

**Variability of *dS/dG***

Using standard errors from the CODEML output, we tested whether *dS/dG* varies among genes of each pathway for each lineage based on the following test statistic, summed over all *i* genes in a pathway: \( \sum (\omega_i - \mu)^2 / \sigma_i \)

Where \( \omega_i \) is the *dS/dG* value for gene *i*, \( \mu \) is the standard error for \( \omega_i \) (from CODEML), and \( \mu \) is the mean \( \omega \) across all genes sampled (calculated separately for each lineage, summed over all genes). This statistic is distributed as \( \chi^2 \) with \( n - 1 \) degrees of freedom, as each of the terms is distributed as normal with zero mean and unit variance, which is the definition of the chi-square statistic (Fisher 1922). There are \( n - 1 \) degrees of freedom because the estimate of the mean is correlated with the estimate of the variance.

**Relationship of *dS/dG* to Pathway Structure and Gene Copy Number***

To determine the relationship of *dS/dG* to each of two measures of pathway position, Kendall’s \( \tau \) rank correlation coefficient was calculated using JMP 7.0. First, Kendall’s \( \tau \) was calculated using pathway position. Enzymes were sequentially numbered from the first enzyme involved in the conversion of glucose to pyruvate through to the last
enzyme involved in the production of each of the four end products. Second, Kendall’s τ was calculated using our PPI. “Groups” of enzymes that act between pathway branch points were numbered from most upstream to most downstream (see fig. 1). If an enzyme (such as CrtL-b; see KEGG PATHWAY Database 2008) was involved in more than one path downstream of a branch point, it was counted as part of the group above that branch point. (Creating a separate category for such genes made no difference to our results). For each ingroup lineage, correlations of $d_{S}/d_{S}$ with pathway position and with PPI were calculated using data from all enzymes along the four branches of plant terpenoid synthesis shown in figure 1. We pooled the data in this way to avoid multiple comparisons that would result if we used the same upstream parts of the pathway for correlations considering each pathway branch (from glucose to end product) separately. As this was done for statistical reasons, it is not intended as a claim that a pathway position of 12 or a PPI value of 4 is biologically equivalent across different pathway branches.

The robustness of each correlation was tested by removing the highest $d_{S}/d_{S}$ value and recalculating the correlation to ensure that significant results were not driven by such outliers. The relationship of $d_{S}/d_{S}$ to gene copy number was examined visually, and because no apparent relationship was seen, Kendall’s τ was not calculated for this variable. The relationship of gene copy number and pathway position was also examined visually.

Tests for Selection

To verify whether any pattern of rate variation is due to relaxation of selective constraint, and not positive selection, we tested for selection using the following comparisons of models in CODEML (Goldman and Yang 1994; Yang 1997, 1998), after Church et al. (2007). The same alignments as for $d_{S}/d_{S}$ calculation were used to test for selection, except that the outgroup Oryza was omitted. This test of selection is the only point in the analysis in which lineages were not treated separately. First, we tested for variable rates among codons by comparing likelihood values under M3 and M0. Second, we tested for a class of codons with $d_{S}/d_{S} > 1$ by comparing likelihood values under M2a versus M1a, and M7 versus M8. Likelihood values were compared between each pair of models, but the $d_{S}/d_{S}$ values for sites were ignored as these may not be reliable (Church et al. 2007, and references therein). We carried out three likelihood ratio tests (LRTs) for each of the 40 genes: one comparing M3 versus M0 to test for variable $d_{S}/d_{S}$ among sites, and one each for comparing M2a versus M1a and M8 versus M7 to test for a class of codons with $d_{S}/d_{S} > 1$.

Results

Ortholog Availability and Rate Variation among Enzymes

Of 47 genes examined, orthologs were found for 40. After discarding $d_{S}/d_{S}$ values with $d_{S}$ above 2.00 or below 0.005, all 40 genes remained for Ricinus and Populus. Arabidopsis retained 26 and Vitis 36 $d_{S}/d_{S}$ values. All four ingroup lineages showed statistically significant variation in $d_{S}/d_{S}$ among the enzymes considered. In each lineage, $d_{S}/d_{S}$ varied by an order of magnitude ($d_{S}/d_{S}$ ranged from 0.040 to 0.28 for Arabidopsis, 0.047 to 0.28 for Populus, 0.031 to 0.29 for Ricinus, and 0.048 to 0.59 for Vitis).

Variation of $d_{S}/d_{S}$ with Pathway Position

Variability in $d_{S}/d_{S}$ correlated significantly with pathway position in Arabidopsis and Populus. The correlation in Arabidopsis remains significant when the highest $d_{S}/d_{S}$ value is removed, though that for Populus does not. Kendall’s τ values are given in table 1. Scatterplots of data for Populus are given in figure 3a–d; plots of data from other lineages can be found in supplemental figures 1–3e–h, Supplementary Material online. Because the data in these plots are separated into each of the pathway branches from glucose to each end product, they show that the positive correlation found with the pooled data is not driven by high $d_{S}/d_{S}$ in the downstream genes of one pathway branch alone. In these graphs, pathway branches leading to the three phytohormones (especially GA) show the strongest trends, whereas the lutein branch exhibits little positive trend if any.

Variation of $d_{S}/d_{S}$ with the PPI

Variability in $d_{S}/d_{S}$ correlated significantly with PPI in all four ingroup lineages, and these correlations remain significant when the highest $d_{S}/d_{S}$ values are removed, except in Vitis. Kendall’s τ values are given in table 1. These rank correlation coefficients are 0.13 higher, and the associated P values are 0.07 lower than those for pathway position, averaged across the four lineages. Data for Populus are shown in figure 3e–h; plots of data from other lineages can be found in the supplemental figures 1–3e–h, Supplementary Material online.

Gene Copy Number

Copy number varied among genes in all four lineages. Vitis, Arabidopsis, and Ricinus had mostly one copy per gene, whereas most Populus genes had two copies. Copy number was 1, 2, or 3 for Vitis; 1, 2, 3, 4, or 11 for Arabidopsis; 1 or 2 for Ricinus; and 1, 2, or 3 for Populus. Singleton genes in Vitis, Arabidopsis, and Ricinus, and both

<table>
<thead>
<tr>
<th>Pathway Position</th>
<th>PPI</th>
<th>τ</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>τ</td>
<td>P</td>
</tr>
<tr>
<td>Arabidopsis</td>
<td>0.33</td>
<td>**0.021</td>
<td>0.46</td>
</tr>
<tr>
<td>Populus</td>
<td>0.22</td>
<td>*0.048</td>
<td>0.34</td>
</tr>
<tr>
<td>Ricinus</td>
<td>0.18</td>
<td>0.115</td>
<td>0.34</td>
</tr>
<tr>
<td>Vitis</td>
<td>0.18</td>
<td>0.123</td>
<td>0.28</td>
</tr>
</tbody>
</table>

*P < 0.05. **P < 0.05 both before and after removal of highest $d_{S}/d_{S}$ value in downstream genes.
singleton and two-copy genes in Populus spanned a wide range of \( dN/dS \) values, whereas higher copy gene fell in the lower range of \( dN/dS \). Nevertheless, graphs of \( dN/dS \) versus copy number did not show any strong overall trend. Graphs of copy number versus pathway position or PPI suggest a slight positive relationship. Copy number was evenly distributed along the pathway for most genes, with the few highest-copy genes appearing downstream.

Selection Analyses

LRTs for the comparison of M3 versus M0 were significant for all 40 enzymes, demonstrating that \( dN/dS \) varies among codons. LRTs for the comparison of M2a versus M1a were not statistically significant for any of the 40 genes. LRTs for the comparison of M8 and M7, however, were significant for 9 of the 40 genes. Although these genes are scattered throughout plant terpenoid synthesis, most appear downstream. A list of these genes and their pathway positions and PPI values is given in table 2.

In all the above analyses, results using alignments altered to exclude ambiguously aligned regions (see note in Methods under “Measuring evolutionary rates”) produced qualitatively similar results. The sole exception is the comparison of M8 and M7, in which one of the significant LRTs becomes nonsignificant when the alternate alignments are used.

Discussion

PRV is a pattern of evolutionary rate variation wherein genes for the downstream enzymes in a metabolic pathway evolve more quickly than the genes for upstream enzymes (Rausher et al. 1999). Our examination of plant terpenoid synthesis represents the most extensive and rigorous survey of PRV in a metabolic pathway conducted to date. We analyze sequence data for 40 enzymes along four branches of a metabolic system, identify orthologs using phylogenetic rather than reciprocal best hit methods (Poptsova and Gogarten 2007), and test for PRV using a conservative measure of correlation (Kendall’s \( \tau \)). As a consequence, our results are robust and provide the only clear example of PRV in a metabolic pathway besides that for anthocyanin synthesis (Rausher et al. 1999, 2008; Lu and Rausher 2003).

In this study, we expected that 1) longer pathways would show stronger PRV and that 2) an index that accounts for the position of an enzyme relative pathway positions and PPI values is given in table 2.

Table 2

<table>
<thead>
<tr>
<th>Pathway Segment</th>
<th>Gene</th>
<th>Pathway Position</th>
<th>PPI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glucose–pyruvate</td>
<td>5.3.1.9 (plastidial)</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Pyruvate–farnesyl-diphosphate</td>
<td>2.2.1.7</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Pyruvate–farnesyl-diphosphate</td>
<td>2.5.1.10*</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>Ent-kaur-16-ene–GA</td>
<td>5.5.1.13</td>
<td>11</td>
<td>3</td>
</tr>
<tr>
<td>Farnesyl-PP–carotenoids</td>
<td>CrtL-b</td>
<td>14</td>
<td>3</td>
</tr>
<tr>
<td>Ent-kaur-16-ene–GA</td>
<td>1.14.13.79</td>
<td>14</td>
<td>4</td>
</tr>
<tr>
<td>Carotenoids–lutein</td>
<td>LUT1</td>
<td>16</td>
<td>4</td>
</tr>
<tr>
<td>Xanthophyll cycle–ABA</td>
<td>AA03</td>
<td>20</td>
<td>5</td>
</tr>
<tr>
<td>Xanthophyll cycle–ABA</td>
<td>1.14.13.93</td>
<td>21</td>
<td>5</td>
</tr>
</tbody>
</table>

The metabolites listed under pathway segment correspond to the pathways shown in figure 1. Gene names or EC numbers are used according to the pathway diagrams used on the KEGG PATHWAY Database (2008).

* No evidence of positive selection when alternate alignment (see Methods) is used.
branch points, “PPI,” would better correlate with \( d_{S} / d_{S} \) than pathway position as such. The long pathways in our study do show strong upstream versus downstream trends in \( d_{S} / d_{S} \), and examination of the graphs reveals that such a relationship is not generally evident in the separate “pathways” (such as glycolysis) that this longer system can be subdivided into. Notably, the variation in \( d_{S} / d_{S} \) in terpenoid synthesis is equivalent to that in the anthocyanin pathway. In the four lineages studied here, \( d_{S} / d_{S} \) varied from 0.04 upstream to 0.30 downstream, an order of magnitude difference. Rausher et al. (2008) found \( d_{S} / d_{S} \) to vary from 0.027 upstream to 0.220 downstream in the anthocyanin pathway, also an order of magnitude difference. This suggests support for our initial prediction, that the anthocyanin pathway exhibits strong PRV because of the close packing of pathway branch points.

Our “PPI” offers modest but noticeable improved detection of trends, compared with the method of simply considering the position of enzymes in the pathway. First, the PPI produces values of Kendall’s \( \tau \) that are 0.13 units higher on average than those for pathway position. Second, PPI produces lower \( P \) values and identifies statistically significant correlations that pathway position does not, as in the Ricinus and Vitis lineages. Third, correlations of \( d_{S} / d_{S} \) with PPI are more robust to removal of the highest \( d_{S} / d_{S} \) point than those for pathway position.

Because glycolytic enzymes are known to be highly conserved, our significant correlations of \( d_{S} / d_{S} \) with pathway position and with PPI could be driven solely by the low \( d_{S} / d_{S} \) values of the glycolytic genes. The correlation of \( d_{S} / d_{S} \) and PPI remains significant in Arabidopsis and Ricinus when glycolytic genes are removed from the analysis; therefore, glycolysis is important to the patterns observed here, but PRV can still be observed without it. This also demonstrates again that our PPI is superior to pathway position for detecting PRV.

In future studies, the PPI may be particularly helpful for regions of the metabolic network that are cyclical or reticulate, such as fatty acid synthesis (Buchanan et al. 2000) or the later stages of BR and GA synthesis (Szekeres and Koncz 1998; Hedden and Phillips 2000), so long as these cyclic or reticulate elements can be placed between branch points in an overall metabolic flow. Further work is required to adapt the PPI to all possible pathway structures. In particular, some pathways fork into fully or partially redundant routes that rejoin downstream, such as the synthesis of terpenoid precursors (Rodríguez-Concepción and Boronat 2002). We handled this here by simply considering both routes in one PPI class, but further work to refine this metric may produce a better approach.

Positive Selection Versus Selective Constraint

We predicted (following Rausher et al. 1999) that \( d_{S} / d_{S} \) would vary with pathway position (or PPI) due to a relaxation of selective constraint along the pathway. Having detected positive selection on about a quarter of the genes examined here, we must consider whether our data are better explained by relaxed selective constraint or by increased positive selection on the downstream genes. If all genes showing positive selection are removed from the data set, correlations of \( d_{S} / d_{S} \) with PPI remain significant for all four lineages (though not for the correlations of \( d_{S} / d_{S} \) with pathway position). Kendall’s \( \tau \) values are only 0.01 lower and \( P \) values are only 0.02 higher (averaged across the four lineages) for correlations of \( d_{S} / d_{S} \) and PPI when the positively selected genes are removed. This suggests that variation in selective constraint does play a large role in producing the overall pattern of rate variation seen here.

Nevertheless, positive selection also has some part to play, and we cannot entirely rule out the possibility that it has a key role in producing the PRV seen here, for two reasons. First, the power of the LRTs used here to detect selection is quite low when few sequences are used (Anisimova et al. 2001). Second, these tests likewise have low power to detect weak positive selection (Anisimova et al. 2001), and so we cannot rule out that downstream genes show elevated \( d_{S} / d_{S} \) due to adaptive substitutions scattered across the gene such that no one codon shows \( d_{S} / d_{S} > 1 \) (Rausher et al. 2008).

We have spoken so far of selection on nonsynonymous sites. Selection for codon usage bias affects synonymous sites in both Arabidopsis and Populus (Ingvarsson 2008; Wright and Andolfatto 2008). If downstream genes were under selection for codon bias, they would have depressed synonymous substitution rates and hence inflated \( d_{S} / d_{S} \). We examined \( d_{S} \) values associated with the \( d_{S} / d_{S} \) values measured in CODEML (Yang 1997; see Methods), and found that \( d_{S} \) was either evenly distributed along the pathway, or elevated downstream (similar to Rausher et al. 1999). Therefore, we are confident that selection on synonymous sites is not driving the trends seen here.

Pleiotropy Versus Other Explanations for PRV

We initially predicted that plant terpenoid synthesis would show PRV because downstream genes are less pleiotropic than those upstream, due to the branching structure of the pathway. Although the current data do not prove that PRV is driven by variation in pleiotropy, or that variation in pleiotropy results from the branching of the pathway, the data presented here certainly fit such a scenario. We cannot, however, prove that pleiotropy is the principle explanation for these trends, and a few alternate explanations are worth commenting on. One alternate hypothesis is that PRV is driven by differences in connectivity rather than pleiotropy along the pathway. High connectivity also produces selective constraint (Vitkup et al. 2006) and many upstream enzymes used here are highly connected (KEGG PATHWAY Database 2008), but pleiotropy likely still plays a role apart from connectivity. A hypothetical pathway in which all nodes have four connections (one in, three out) should still show PRV even though all nodes have equal connectivity, because the upstream nodes are more pleiotropic. Nevertheless, further work is required to disentangle the roles of pathway position, pleiotropy, and connectivity in shaping evolutionary rate variation.

A second alternate hypothesis is that the distribution of protein families along the pathway is biased, such that certain fast-evolving protein families are concentrated...
downstream. According to the Pfam and PROSITE designations provided by the KEGG PATHWAY Database (2009), the only protein families containing multiple members in this pathway are the 2-oxoglutarate-dependent dioxygenase and the cytochrome p450 protein families, both of which are concentrated downstream. This does suggest a bias in the distribution of protein families along the plant terpenoid pathway, but it is beyond the scope of the current data set to disentangle the role of protein family membership in producing the correlations observed here. Given the different biological roles of upstream versus downstream genes in plant terpenoid synthesis, we are confident that pleiotropy is a major factor (though likely not the sole factor) in producing elevated \( d_{ob}/d_{sb} \) in the downstream genes.

**Gene Copy Number**

At the outset of this study, we hoped that examining gene copy number could explain why certain lineages show PRV and others do not. Our observation that high copy genes appear to be concentrated downstream in some cases deserves further attention in future studies. Gene copy number, however, cannot explain why PRV is observed in plant terpenoid synthesis, because high copy genes did not have high \( d_{ob}/d_{sb} \) values.

Several factors could explain why gene copy number shows no apparent relationship or a weak negative relationship with \( d_{ob}/d_{sb} \) here. Duplicate genes experience relaxed selection for only a short period after duplication (Lynch and Connelly 2000), and even then purifying selection is not wholly relaxed (Hughes and Hughes 1993). Also, gene duplication represents only one form of redundancy, “homologous redundancy.” Another form, “nonhomologous redundancy,” is the existence of parallel paths that offer an alternate route through the network. This latter form of redundancy may actually cause greater relaxation of selective constraint than gene copy number does (Wagner 2000; Kitami and Nadeau 2002).

**Future Directions**

Overall, the structure of metabolic pathways deserves greater consideration as a factor shaping gene evolution. Ideally, future studies in this area will move toward whole-network analysis. Networkwide studies of PRV will require an organism for which the total metabolic network is well understood, plus relatives with full genome sequences. Metabolic networks have been constructed for several organisms (Jeong et al. 2000; Wagner and Fell 2001), but these often consider only a core metabolism, representing metabolic networks and compared among these various pathway types, in order to draw conclusions about pathways or networks in general. Studies on developmental and signal transduction pathways (Olsen et al. 2002; Riley et al. 2003) suggest some effect of pathway position (but see Jovelin et al. 2008). The development stage in which a gene is expressed also correlates with evolutionary rate (Cutler and Ward 2005; Davis et al. 2005, but see Good and Nachman 2005), which suggests a pathway position effect. The sequencing of more and more genomes will open the way to offer increasingly sophisticated answers to questions of how network topology shapes evolutionary rate variation.

**Supplementary Material**

Supplementary tables 1 and 2, supplementary figures 1–3, and lists of *Arabidopsis* queries used to search for orthologs, lists of orthologs used for each gene, and data figures for *Ricinus, Arabidopsis,* and *Vitis* are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

**Acknowledgments**

Many thanks to Mike Barker for his advice on ortholog determination and for providing several Perl scripts, including his alignment pipeline. Many thanks also to Nolan Kane for sharing insightful comments and unpublished work on positional rate variation. We are grateful to Asher Cutter and an anonymous reviewer for a number of insightful comments that helped improve this manuscript. Sally Otto, Dilara Ally, Itay Melrose, and Allen Liu also took part in helpful discussions. NSERC funded this project through grants to H. Ramsay and K. Ritland.

**Literature Cited**


Kenneth Wolfe, Associate Editor
Accepted January 26, 2009