Evolution of Flux Control in the Glucosinolate Pathway in *Arabidopsis thaliana*

Carrie F. Olson-Manning,* 1 Cheng-Ruei Lee, 1 Mark D. Rausher, 1 and Thomas Mitchell-Olds 1

1Department of Biology, Duke University
*Corresponding author: E-mail: cfo3@duke.edu.
Associate editor: Michael Purugganan

Abstract

Network characteristics of biochemical pathways are believed to influence the rate of evolutionary change in constituent enzymes. One characteristic that may affect rate heterogeneity is control of the amount of product produced by a biochemical pathway or flux control. In particular, theoretical analyses suggest that adaptive substitutions should be concentrated in the enzyme(s) that exert the greatest control over flux. Although a handful of studies have found a correlation between position in a pathway and evolutionary rate, these investigations have not examined the relationship between evolutionary rate and flux control. Given that genes with greater control will experience stronger selection and that the probability of fixation is proportional to the selective advantage, we ask the following: 1) do upstream enzymes have majority flux control, 2) do enzymes with majority flux control accumulate adaptive substitutions, and 3) are upstream enzymes under higher selective constraint? First, by perturbing the enzymes in the aliphatic glucosinolate pathway in *Arabidopsis thaliana* with gene insertion lines, we show that flux control is focused in the first enzyme in the pathway. Next, by analyzing several sequence signatures of selection, we also show that this enzyme is the only one in the pathway that shows convincing evidence of selection. Our results support the hypothesis that natural selection preferentially acts on enzymes with high flux control.

Key words: flux control, evolution, *Arabidopsis thaliana*, natural selection, glucosinolates.

Introduction

Several recent investigations have documented correlations between network properties of genes and selective constraint or rates of substitution, suggesting that network characteristics of enzymes influence their rates of evolution (Cork and Purugganan 2004; Flowers et al. 2007; Eanes 2011; Slotte et al. 2011; Zera 2011). Such properties include the number of connected enzymes (Pfeiffer et al. 2005), enzyme expression (Yang and Gaut 2011), or flux control (Flowers et al. 2007; Wright and Rausher 2010). Although some of these correlations likely represent differences in the neutral substitution rate due to differences in constraint (e.g., Rausher et al. 2008), others represent differences among pathway genes in rates of adaptive substitution (Flowers et al. 2007). If natural selection preferentially acts on certain genes or pathway positions, this may result in repeated or parallel evolution, and contribute to extensive heterogeneity of evolutionary rates among enzymes. Recent studies have considered sequence signatures of selection on enzymes in their network context as a way of understanding what forces influence their evolutionary fate (reviewed in Gaut et al. 2011) and several patterns have emerged.

One pattern is that the rate of evolution is often correlated with enzyme position in a metabolic pathway. Several studies have found, for example, that genes at the beginning of a pathway are under greater selective constraint, as reflected in the dN/dS ratios for downstream genes (Rausher et al. 1999; Lu and Rausher 2003; Rausher et al. 2008; Livingstone and Anderson 2009; Ramsay et al. 2009; Ma et al. 2010). Differences in pleiotropy among genes at different pathway positions might cause this pattern, as genes in the beginning of a pathway may influence a larger number of downstream products than downstream genes (Rausher et al. 1999; Ramsay et al. 2009). However, such differential pleiotropy has not been demonstrated for any pathway. Moreover, Wright and Rausher (2010) provide an alternative theoretical explanation for this pattern: in linear metabolic pathways, genes coding for upstream enzymes tend to evolve the largest control over flux, which means that slightly deleterious substitutions are more likely to occur in genes coding for downstream enzymes. Their model also predicts that in linear pathways, adaptive substitutions will tend to be concentrated in upstream enzymes because 1) their greater control over flux means that on average mutations in these genes will experience stronger selection; and 2) the probability an advantageous mutation will be fixed is proportional to its selection coefficient. The model thus provides theoretical support for the oft-expressed expectation that adaptive substitutions will be concentrated in enzymes that exert the most control over flux (Hartl et al. 1986; Eanes 1999; Watt and Dean 2000).

Despite these expectations, we are unaware of any investigations that have tested them by both estimating the magnitude of flux control for enzymes in a metabolic pathway and also assessing patterns of substitution in the genes coding for those enzymes. In an attempt to bridge this gap between theory and evidence, we have conducted an explicit test of whether these expectations are met by the glucosinolate pathway in *Arabidopsis*.
Because of the ease of quantifying total pathway outputs, glucosinolate (GLS) production in Arabidopsis thaliana provides an excellent pathway for addressing the relationship between flux control and patterns of substitution. GLSs are important defensive compounds (Hopkins et al. 2009; Fan et al. 2011) found in the Brassicales. They are derived from amino acid precursors (methionine as part of the aliphatic GLS pathway and tryptophan as part of the indolic pathway GLS in A. thaliana) and are stored in the vacuole (fig. 1) (Grubb et al. 2004). Intact GLSs have low toxicity until a leaf is damaged and they come in contact with the hydrolytic enzyme myrosinase, resulting in the production of compounds (isothiocyanates, nitriles, thiocyanates, epithionitriles, and oxazolidine-2-thiones) that slow growth and development of herbivores (Blau et al. 1978; Kliebenstein et al. 2005).

The biosynthesis of methionine- and tryptophan-derived GLSs occurs through a series of reactions, with the initial reactions occurring in the chloroplast, and subsequent reactions in the cytosol. In the chloroplast reactions, methionine undergoes side-chain elongation. These products are then transferred to the cytosol, where they undergo a series of reactions (fig. 1). Reactions occurring in the cytosol will be examined as the "core" GLS pathway (fig. 1 adapted from Sonderby et al. 2010).

Two features of this core pathway are relevant. First, several steps are catalyzed by two different enzymes. These pairs have resulted from gene duplication and are typically co-expressed. The first pair, CYP79F1 and CYP79F2, have slightly different substrate specificities, with CYP79F1 using both short- and long-chain substrates, whereas CYP79F2 tends to use only long-chain substrates. The second important feature is that the pathway leading to the synthesis of GLSs from methionine shares one enzyme, SUR1, with a parallel pathway that synthesizes GLSs derived from tryptophan (fig. 1). Although each pathway is linear, this shared enzyme could provide some cross-talk between them. In particular, if SUR1 is saturated or nearly saturated in vivo, reduction in flux down one pathway could increase flux down the other pathway. This effect is not expected, however, if SUR1 is relatively unsaturated, because there would be little competition among the precursors for access to that enzyme. Thus, a test of whether enzymes with predominant flux control in one pathway also exert control over flux down the other pathway provides information about SUR1 saturation and cross-talk.

In this analysis of A. thaliana, we asked whether the enzymes in the core GLS pathway show differential flux control resulting in heterogeneous substitution patterns. In particular, we address the following questions: 1) Do upstream enzymes exert the majority of control over flux? 2) Are adaptive substitutions concentrated in enzymes that exert strong flux control? 3) Do upstream enzymes exhibit the greatest selective constraint?

**Materials and Methods**

**Insertion Lines**

To determine whether differences in the pattern of selection on different pathway enzymes are correlated with the magnitude of flux control, we approximated flux control of each step in the pathway by perturbing the amount of each enzyme using Agrobacterium TDNA insertion lines (Sussman et al. 2000; Alonso et al. 2003). These TDNA insertions disrupt the gene of interest and in heterozygous form may substantially decrease the amount of mRNA produced, and thus the total activity of that enzyme available in the cell. A. thaliana insertion lines Wisconsin (Sussman et al. 2000) background Wassilewskija, SALK Institute Genomic Analysis Laboratory (Alonso et al. 2003) Columbia-0 (Col-0, CS60000), and Syngenta Arabidopsis Insertion Library were collected for as many genes in the aliphatic GLS pathway as were available (supplementary table S1, Supplementary Material online). Each of these lines contained an insertion causing loss-of-function (LOF) in one aliphatic GLS pathway gene, either in heterozygous or homozygous form.

**Plant Growth Conditions**

Seeds from each line were grown for one generation to determine whether they were heterozygous or homozygous for the LOF allele. For each line, approximately 20 seeds from heterozygous maternal individuals were placed on soil in a randomized complete block design in a 24-cell flat. Seeds were allowed to imbibe, and then were stratified for 3 days at 4°C to overcome dormancy. Plants were maintained under long day conditions (16 h) at 18°C for 3 weeks, when tissue was harvested. One true rosette leaf was collected over two consecutive mornings for each of the following analyses: insertion genotyping, mRNA analysis, and GLS quantification. The tissue for RNA analysis was flash frozen, and the tissue for

---

**Figure 1.** Core aliphatic and indolic GLS pathways in Arabidopsis thaliana. The enzyme that catalyzes each reaction is found to the side of the arrow. *Metabolizes short- and long-chain aliphatic GLS. ^Metabolizes long-chain substrates. The aliphatic GLS pathway is bolded and is the focus of subsequent analyses. Adapted from figure 1 in Sonderby et al. (2010) and used with permission from Elsevier.
GLS quantification was stored in 2 ml 70% methanol for GLS analysis.

Genotyping

Frozen tissue was ground with ball bearings on liquid nitrogen in a Geno/Grinder (SPEX, CentriPrep 2000). DNA was extracted with a modified CTAB protocol (Rogers and Bendich 1988) and resuspended in TE buffer. Primers were designed using the SALK SIGnAl isect primer design tool (http://signal.salk.edu/tdnaprimers2.html; last accessed 31 Aug 2012) (supplementary table S1, Supplementary Material online) and genotyped according to Alonso et al. (2003) with GoTaq (Promega) Taq. Genotypes were scored on a 0.7% agarose gel.

Quantitative Real-Time Polymerase Chain Reaction

To quantify the relative amounts of mRNA transcripts, duplicate samples of 3-week old rosette leaves were collected and flash frozen on liquid nitrogen from heterozygous (HET) or wild-type (WT) plants. Total RNA was extracted from with the SV RNA kit (Sigma). CDNA synthesis was performed with the Dynamo cDNA synthesis kit (Finnzymes). If possible, primers were designed to span the intron of the gene of interest (MWG Operon) (supplementary table S3, Supplementary Material online). For GSTF11, two different primer sets were designed that produced the same result. GLS genes were normalized to the transcript levels of the reference gene UBQ10 (Atg05320) (Czechowski et al. 2005). Duplicate quantitative reverse-transcriptase polymerase chain reaction (qRT-PCR) reactions were performed for each primer pair. qRT-PCR was performed with the Dynamo SYBR® Green qPCR kit (Finnzymes). Data were analyzed on the Mastercycler ep realplex 2 (Eppendorf). The reactions were carried out at 95°C for 2 min, and 40 cycles of 95°C for 15 s, 56°C for 15 s and 68°C for 20 s.

Analysis of GLS Concentration

Leaf tissue was first weighed and then leached in 2 ml of 70% methanol for 3 weeks at 4°C and 1 week at room temperature. Sinigrin (Sigma) was added to each sample to 1 μg/ml to each well to serve as an internal reference. Then the entire 2 ml leaching volume was added to a 96-well plate containing equilibrated DEAE Sephadex and cleaned (Mikkelsen et al. 2009). We cleaved the GLSs into desulfo-glucosinolates with 30 μg sulphatase. The desulfo-glucosinolates were run on high-pressure liquid chromatography (Kliebenstein et al. 2001) on an Agilent 1100 high-pressure liquid chromatography with 96-cell autoloader. Separation of GLSs was carried out on a C-18 column (Zorbax Eclipse XDB C18, 4.6 × 150 mm and 5 μm) and peaks were called manually based on retention time and UV absorption spectra at A_{229nm} (Windsor et al. 2005). Seven GLSs were quantified: four from the aliphatic pathway and three from the indolic pathway (supplementary table S2, Supplementary Material online). The area of each peak was calculated and normalized by the weight of the tissue collected; the area of the Sinigrin peak and molar concentrations were calculated given the predetermined calibration curve from pure desulfo-glucosinolates (Brown et al. 2003). Approximately 30 replicates were run for each WT line and 10 for each heterozygote type.

Sequence Analysis

To investigate the pattern of natural selection on GLS pathway genes, we used the polymorphism data of A. thaliana with A. lyrata as outgroup. Eighty-one A. thaliana genomes (including the Col-0 reference genome) were downloaded from the MPICao2010 subset (Cao et al. 2011) of the 1001 genome project (http://1001genomes.org/index.html; last accessed 31 Aug 2012), and coding sequences (CDS) of A. lyrata genes and their orthology information with A. thaliana were downloaded from a recently published dataset (Hu et al. 2010). Based on the annotation from TAIR10 release of The Arabidopsis Information Resource (Lamesch et al. 2011), we extracted CDS from both species (Cao et al. 2011). Alignment was performed by the codon model in PRANK (Loytynoja and Goldman 2005; Fletcher and Yang 2010), and GSTU20 (AT1G78370) was excluded from the following interspecific analyses due to the lack of orthologs in A. lyrata.

To examine the pattern of selective constraint or positive selection within each enzyme of the aliphatic GLS pathway in A. thaliana, we calculated the average pairwise difference per site (π) and the ratio of synonymous and nonsynonymous π (π_S/π_N) in DnaSP (Rozas et al. 2003). Significance of deviation of this ratio from 1 was tested by bootstrapping (1,000 replicates) using a program written in APL by one of the authors (M.D.R.).

We used DnaSAM (Eckert et al. 2010) to calculate additional population genetic statistics based on the site frequency spectrum (Tajima’s D and normalized Fay and Wu’s H). The statistics of genes in the GLS pathway were then compared with the distribution of all A. thaliana genes with known A. lyrata orthologs. Additionally, we plotted the site-frequency spectrum of the derived amino acid polymorphisms in all the genes in the GLS pathway. The derived amino acid was determined based on either the ortholog in A. lyrata, or, in the case of SOT18 and GSTU20 with no A. lyrata ortholog, the ancestral reconstruction software ANCESCON (Cai et al. 2004). To explore the geographic distribution of the acidic amino acid polymorphisms, we plotted the geographic distribution of ancestral and derived polymorphisms for CYP79F1 and SUR1 with the R package chplot (Vidmar and Pohar 2005). Finally, we searched the worldwide A. thaliana Regmap panel (Horton et al. 2012) for departures from genome-wide values of Fst in the GLS pathway.

To address constraints on divergence between species (A. thaliana and A. lyrata), we calculated the dN/dS ratio in DnaSP. Additionally, a Perl script (Holloway et al. 2007) was used to perform MK tests (McDonald and Kreitman 1991) for possible adaptive substitution. As an index of whether deviations from neutrality are due to excess adaptive substitutions or an excess of nonsynonymous variation, we calculated the statistic DoS (direction of selection), a variant of the McDonald Kreitman test. This statistic is zero under neutrality.
(Stoletzki and Eyre-Walker 2011). Positive DoS indicates an excess of nonsynonymous substitutions between species, or a deficiency of nonsynonymous polymorphisms within species. Negative DoS indicates an excess of nonsynonymous polymorphism within species, or a deficit of nonsynonymous divergence.

We used a permutation procedure to control levels of statistical significance for multiple tests. With \( N \) loci, here we focused on two loci which were potentially significant based on univariate tests. Under the null hypothesis that synonymous/nonsynonymous status is independent of whether variation is polymorphic within species or fixed between species, we computed the null distribution for Fisher’s exact test (FET) for the most extreme locus, and the second most extreme locus, in a sample of \( N \) loci. We permuted synonymous/nonsynonymous status across single nucleotide polymorphisms and calculated FET for each locus in each permutation. The most extreme and second most extreme FET values were identified among these loci and saved to a null distribution based on 50,000 permutations. Finally, the two loci (CYP79F1 and SUR1) with most extreme actual FET values were compared with these statistical null distributions. Calculations employed a Python program written by one of the authors (T.M.O.).

Analysis of Flux Control

The magnitude of flux control exerted by an enzyme is typically assessed in one of three ways: 1) by perturbing enzyme concentration or activity; 2) by calculation from elasticity coefficients; or 3) by quantifying transient metabolite concentrations (Delgado and Liao 1992). We present here a novel approach that is based on comparing the GLS concentration of individuals heterozygous for a LOF mutant in a gene in the GLS pathway with that of WT individuals. A reduction in GLS concentration in heterozygous individuals implies that reducing enzyme concentration reduces GLS production and thus that the enzyme has substantial control over flux. The magnitude of the flux control coefficient for enzyme \( i \), \( \lambda_i \) (sensitivity coefficient of Kacser and Burns 1973) was estimated from the following equation:

\[
\lambda_i = \frac{C_i(1 - (F_R)_i)}{[(F_R)_i(1 - C_i)]}
\]

where \( C_i \) is the relative concentration of the enzyme \( i \) in heterozygotes (i.e., concentration in heterozygotes divided by concentration in wild type) and \( (F_R)_i \) is the relative flux of enzyme \( i \) in heterozygotes (flux in heterozygotes divided by flux in wild type) (see Appendix for derivation).

As an approximation, we estimated \( C_i \) by the relative expression, \( R_e \), the ratio of heterozygote expression level of enzyme \( i \) divided by WT expression level. Although we did not measure enzyme concentrations directly, it is expected that concentrations in heterozygotes for a LOF allele will be approximately half that in WT homozygotes unless there is feedback regulation of transcription. In some cases, mRNA expression level may be an imperfect indicator of enzyme activity due to possible posttranslational regulation, which might reduce the correlation between mRNA and protein concentration (Vogel and Marcotte 2012). However, a recent large-scale analysis in humans showed that the correlation between mRNA and protein abundance is strong (Schwanhausser et al. 2011). Additionally, our study compares the relative expression between isogenic WT and HET lines, so differential patterns of posttranslational modification between WT and HET lines should be minimal. \( (F_R)_i \) was estimated indirectly by the relative GLS level in heterozygotes (concentration in heterozygotes divided by concentration in wild type). This approach assumes that final GLS production is proportional to flux, as would be the case if GLS production occurred over a fixed period of time. Confidence intervals for the estimates of \( \lambda_i \) were calculated by bootstrapping (1,000 replicates).

Statistical Analysis

Comparison of expression levels in heterozygotes and corresponding WT homozygotes was performed using appropriate contrast statements in an analysis of variance. Expression levels were log-transformed before analysis. Relative expression levels, \( R_e \), were calculated by first estimating the difference in log (expression) between heterozygotes and WT, \( D_i \), then calculating \( R_e = e^{D_i} \). Confidence intervals for \( R_e \) were calculated by first calculating the confidence intervals for \( D_i \) from their standard errors and then transforming by exponentiation.

Comparison of GLS production between WT and heterozygous plants was conducted using multivariate analysis of variance (MANOVA). Dependent variables were concentrations of different GLSs. When the multivariate effect of treatment (WT vs. HET) was significant, as judged by Wilk’s \( \lambda \) statistic (Timm 1975) univariate analyses were performed to determine which GLSs were affected by treatment. Two types of analysis were performed. In one analysis, we pooled all WT individuals of different lines and compared these individuals to HET individuals. This pooling was justified because preliminary analyses did not show a significant line effect for WT individuals at any of the genes (supplementary table S4, Supplementary Material online). However, we also performed a second analysis for each gene in which HET individuals were compared with only WT individuals from the same line. Because the results of the two analyses were fully concordant, we report only the results of the analyses with WT lines pooled. Analyses were performed using PROC GLM in the SAS statistical software package (SAS 9.3, Cary, NC).

Results and Discussion

Relative Gene Expression Levels in Heterozygotes and WT Individuals

To confirm that enzyme expression levels are reduced in heterozygotes and to estimate the relative expression in heterozygotes (\( C_i \)), we used quantitative real-time PCR to estimate expression levels for heterozygotes of knockdown lines with insertions in the promoter or exon for each gene, as well as for WT homozygotes. Two of the enzymes, GSTF11 and GSTU20, do not exhibit significant reduction in expression
levels in heterozygotes (fig. 2, table 1; unrelativized expression levels: supplementary tables S5 and S6, Supplementary Material online), as judged by contrast statements in ANOVA. For the remainder of the enzymes, the relative expression ratio \( R_i \) was significantly less than 1 (fig. 2 and table 1), indicating that having just one functional copy of the gene results in substantially reduced expression. For five of these enzymes, the \( R_i \) were not distinguishable from the expected value of 0.5, as judged by confidence intervals overlapping 0.5 (table 1). However, for \( SUR1 \) and \( SOT18 \), \( R_i \) was significantly less than 0.5, indicating that expression in heterozygotes was reduced by more than 50% compared with WT individuals.

**Estimation of Flux Control in the GLS Pathway**

To determine whether differences in enzyme concentrations between heterozygotes and WT individuals, as reflected in differences in expression level, affect flux, we compared GLS production in 3-week-old rosette leaves. We first compared WT lines. ANOVA indicated that the lines do not differ significantly in GLS production for any of the aliphatic or indolic GLSs (supplementary table S4, Supplementary Material online), indicating no detectable effect of genetic background on GLS production. For subsequent comparisons between HET and WT individuals, we therefore pooled WT individuals from the three lines.

We first examined flux control through the aliphatic pathway. We did not compare HET and WT GLS concentrations for enzymes GSTF11 or GSTU20 because they showed no evidence that expression levels were reduced in heterozygotes. Of the remaining enzymes, only the first enzyme in the pathway, CYP79F1, exhibited a significant reduction in GLS production in heterozygotes as judged by \( t \) test after correction with sequential Bonferroni correction (supplementary table S6, Supplementary Material online). This reduction was confined to the two short-chain aliphatic GLSs, 3MSOP and 4MSOB, and remained significant even after a Bonferroni correction. Estimated flux control coefficients were high for these two compounds (\( \lambda = 0.8394, \text{CI} = [0.4873, 1.5947] \), and \( \lambda = 0.4016, \text{CI} = [0.1573, 0.9556] \) for the two compounds, respectively). In agreement with the statistical analysis, all other enzymes exhibited low (all \( \lambda < 0.27 \)) and nonsignificant flux control coefficients for all four GLSs (table 2). It thus appears that for at least the short-chain aliphatic GLSs 3MSOP and 4MSOB, flux control is primarily vested in the first enzyme of the pathway. These results thus support the prediction (Wright and Rausher 2010) that flux control should evolve to be vested mainly in the first enzyme of a pathway.

![Fig. 2. qRT-PCR comparison of HET with WT genotypes of the different insertion lines. UBQ10 was used as the reference gene to normalize the GLS gene (abbreviations on horizontal axis) of interest in each RNA extraction. All comparison between the expression of HET (dark gray bars) and WT (light gray bars) are significantly different except GSTF11 and GSTU20. The vertical hashes indicate standard error bars. Asterisks above each pair of bars show relative expression ratio of less than one, as judged by appropriate contrasts in ANOVA.](https://academic.oup.com/mbe/article-abstract/30/1/14/1019691/1019691)
An apparent exception to this principle is the absence of flux control over the production of the long-chain aliphatic GLSs 5MSOP and 6MSOH, either in the first enzyme or any of the other enzymes. One possible explanation for this is that flux control associated with the production of these compounds is vested in CYP79F2, an enzyme that was not examined in this investigation because the same kind of LOF mutants are not available. Previous investigations suggest that activity of the CYP79F1 and CYP79F2 enzymes is redundant, but only for long-chain precursors. In a tandem deletion study of CYP79F1 and CYP79F2 (Tantikanjana et al. 2004), the authors found that both CYP79F1 and CYP79F2 produce long-chain aliphatic GLS but only CYP79F1 produced short-chain aliphatic GLS. Likewise, biochemical studies also indicate that the synthesis of long-chain aliphatic GLS by CYP79F1/CYP79F2 is redundant, because both enzymes catalyze tri-, tetra- penta- and hexahomomethionine, but only CYP79F1 can catalyze short-chain aliphatics to their corresponding aldoximes (Hansen et al. 2001; Chen et al. 2003). Because both CYP79F1 and CYP79F2 make long-chain aliphatic compounds, it is unclear whether changing CYP79F2 should also change the amount of long-chain GLSs produced. Given that both enzymes make long-chain aliphatic compounds, it is possible that the majority of flux control over these compounds is vested in CYP79F2 rather than in CYP79F1. However, because CYP79F2 cannot metabolize short-chain precursors, it is not likely to exert significant flux control over their production. If this hypothesis is true, the separation of flux control for short- vs. long-chain aliphatic GLSs between two different enzymes could allow for evolutionary flexibility in the relative concentrations of these two classes of GLSs.

Indolic GLSs are produced by a parallel pathway that is largely independent of the pathway for aliphatic GLSs (fig. 1). However, the two pathways share one enzyme, SUR1. This sharing creates the possibility that altering the concentrations of enzymes in the aliphatic pathway could affect the flux through the indolic pathway if there is strong competition among aliphatic and indolic precursors for access to SUR1. In this situation, which reflects near saturation of SUR1, control coefficients of aliphatic-pathway enzymes for the production of indolic products would be negative: a reduction in enzyme concentration would reduce flux through the aliphatic pathway, reducing competition of substrates for SUR1, and increasing indolic flux.

Given that only CYP79F1 exerts detectable control over aliphatic flux, we would a priori expect only this enzyme to exhibit this type of negative flux coefficient. In fact, neither it, nor any of the other aliphatic enzymes exhibit detectable control over indolic GLS production: indolic GLS concentrations do not differ significantly between HET and WT for any of the aliphatic enzymes (supplementary table S6, Supplementary Material online). This result indicates that despite the potential for interaction between the two pathways, there is little effect of aliphatic flux level indolic flux level, implying that SUR1 is far from saturation in vivo.

Relative Substitution Rates

We first examine the prediction that adaptive substitutions are expected to be more frequent in upstream genes, particularly in the first enzyme of the pathway. Repeated adaptive substitutions in a gene can be detected with a MacDonald–Kreitman test (McDonald and Kreitman 1991). We applied this test to each gene, and corrected significance levels for multiple comparisons. None of the genes exhibited a significant excess of nonsynonymous substitutions, which would be reflected in a DoS > 0 (table 3). In fact, DoS was positive only for GSTF11, and this was not significant. However, both CYP79F1 and SUR1 showed statistically significant negative DoS values based on a permutation test. We used FET to perform MacDonald–Kreitman tests on 10 GLS biosynthetic genes, with a permutation procedure to account for multiple tests on these loci. These results found $P = 0.0029$ for CYP79F1, and $P = 0.0001$ for SUR1, showing that both loci show significantly negative DoS.

One interpretation of these results is that balancing selection has operated at both of these loci. If this were true, we would expect to find positive Tajima’s D values. Only UGT74B1 has a slightly positive D (supplementary table S7, Supplementary Material online), which is not significant. Indeed, none of the genes show Tajima’s D values in the top 5% of the genome-wide distribution. However, two alternative interpretations are also consistent with the data. The first is that the recent population expansion (Sharbel et al. 2000; Beck et al. 2008) in A. thaliana has allowed the accumulation of deleterious alleles and thus has led to negative

### Table 2. Estimated Flux Control Coefficients Calculated Using All Three Lines for WT Individuals and 95% Bootstrap CIs After 1,000 Bootstrapping Runs.

<table>
<thead>
<tr>
<th>Gene</th>
<th>3MSOP</th>
<th>4MSOB</th>
<th>5MSOP</th>
<th>6MSOH</th>
</tr>
</thead>
<tbody>
<tr>
<td>CYP79F1</td>
<td>0.8394 (0.4873, 1.5947)</td>
<td>0.4016 (0.1573, 0.9556)</td>
<td>-0.0238 (-0.1933, 0.2934)</td>
<td>-0.1124 (-0.2486, 0.0486)</td>
</tr>
<tr>
<td>CYP83A1</td>
<td>-0.0411 (-0.2265, 0.1172)</td>
<td>-0.0808 (-0.2410, 0.0536)</td>
<td>-0.1667 (-0.4194, 0.0647)</td>
<td>-0.3315 (-0.6369, -0.1405)</td>
</tr>
<tr>
<td>SUR1</td>
<td>0.0673 (-0.0182, 0.2436)</td>
<td>0.0726 (-0.0161, 0.2639)</td>
<td>-0.0560 (-0.3199, 0.6351)</td>
<td>0.2607 (-0.0114, 1.5369)</td>
</tr>
<tr>
<td>UGT74B1</td>
<td>0.0002 (-0.1781, 0.2325)</td>
<td>0.0002 (-0.1829, 0.1870)</td>
<td>0.2042 (-0.1094, 1.3732)</td>
<td>-0.0604 (-0.4579, 0.4414)</td>
</tr>
<tr>
<td>UGT74C1</td>
<td>0.0061 (-0.1741, 0.2416)</td>
<td>0.0246 (-0.1574, 0.2832)</td>
<td>-0.3603 (-0.1065, 0.1197)</td>
<td>0.2183 (-0.2090, 1.4085)</td>
</tr>
<tr>
<td>SOT17</td>
<td>0.0352 (-0.0526, 0.1522)</td>
<td>-0.0413 (-0.1529, 0.0793)</td>
<td>0.0617 (-0.2282, 0.6739)</td>
<td>-0.1483 (-0.4511, 0.3360)</td>
</tr>
<tr>
<td>SOT18</td>
<td>-0.0148 (-0.0733, 0.0542)</td>
<td>-0.0291 (-0.0832, 0.0265)</td>
<td>-0.0600 (-0.1442, 0.0196)</td>
<td>-0.1194 (-0.2086, -0.0532)</td>
</tr>
</tbody>
</table>

Note.—Negative values should be interpreted as being equal to 0.
values of DoS at most loci in the genome and negative Tajima’s D values. The observation that 9 of the 10 genes exhibit negative DoS, a pattern that is significantly different from chance based on a binomial test (P < 0.01), is consistent with this interpretation. In this situation, the significant negative DoS for SUR1 might simply reflect the most extreme example of this stochastic process. Alternatively, the combination of negative Tajima’s D values and positive Fay and Wu’s H (supplementary table S7, Supplementary Material online) is consistent with an old population bottleneck that is regaining neutral variation (Haddrill et al. 2005). These measures are influenced by demographic factors and the genes do not deviate from the genome-wide distribution (supplementary fig. S1 and supplementary table S7, Supplementary Material online).

By contrast, at CYP79F1, the \( \pi_N/\pi_S \) ratio is substantially and significantly > 1 (bootstrap 99% confidence interval: 1.15–8.02, 1,000 replicates). Although a \( \pi_N/\pi_S \) ratio greater than one can reflect repeated adaptive substitution (Kryazhimskiy and Plotkin 2008), our failure to detect any adaptive substitutions with an M–K test strongly suggests that this locus is subject to strong balancing selection. At SUR1, however, the \( \pi_N/\pi_S \) ratio substantially less than one and is comparable with that of the other pathway genes (table 4), providing no indication that this gene has historically been subject to balancing selection. Under this interpretation, which we favor, CYP79F1 is the only gene subject to selection. The observation that this gene codes for the first enzyme in the pathway, which is also the only enzyme with demonstrable control over flux, is consistent with expectations.

A third possible interpretation is a variant of the preceding one: what we have taken for excess nonsynonymous polymorphism may be attributable to divergence among A. thaliana populations. The accessions used for our M–K analysis were collected from widespread geographic localities (Cao et al. 2011). Differences among accessions thus might reflect fixation of different alleles at different locations, i.e., between-population substitution rather than within-population variation. If so, the apparent excess of nonsynonymous polymorphisms and \( \pi_N/\pi_S \gg 1 \) at CYP79F1 could represent repeated episodes of positive selection at these two loci. In that case, the theoretical prediction that adaptive substitutions should be concentrated in the first pathway enzyme would be upheld for CYP79F1. Indeed, the site-frequency spectrum of derived nonsynonymous polymorphisms in CYP79F1 (supplementary fig. S2, Supplementary Material online) finds some high-frequency polymorphisms. The geographic distributions of these polymorphisms show that they are widespread (supplementary fig. S3, Supplementary Material online) and might represent between-population substitutions. This is consistent with the Fst values, we find for the genes in the GS pathway (Horton et al. 2012) (supplementary table S8, Supplementary Material online) where none deviate from the genome-wide distribution.

Although the last two interpretations seem more compelling to us than the inference that both CYP79F1 and SUR1

| Table 3. McDonald–Kreitman Test of Polymorphism in Arabidopsis thaliana Compared with Divergence from A. lyrata. |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| CYP79F1 | 13 | 5 | 20 | 42 | -4.46 | 0.0029 |
| CYP79F2 | 8 | 8 | 39 | 50 | -0.28 | ns |
| CYP83A1 | 2 | 14 | 6 | 26 | 0.62 | ns |
| GSTF11 | 6 | 5 | 12 | 13 | -0.30 | ns |
| SUR1 | 14 | 13 | 4 | 31 | -7.35 | 0.0001 |
| UGT74B1 | 18 | 25 | 16 | 37 | -0.67 | ns |
| UGT74C1 | 18 | 25 | 19 | 34 | -0.29 | ns |
| SOT17 | 6 | 8 | 19 | 32 | -0.26 | ns |
| SOT18 | 8 | 3 | 86 | 80 | -1.48 | ns |

Note.—NS, nonsynonymous; S, synonymous; DoS, direction of selection; ns, not significant.

| Table 4. Synonymous and Nonsynonymous within Species Variation of Arabidopsis thaliana and between Species Variation of A. thaliana Compared with A. lyrata. |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| Gene | Site Type | Polymorphism | Divergence |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| CYP79F1 | Syn | 0.0007 | 2.4054 | 0.1075 | 0.1709 |
| | Nonsyn | 0.0018 | — | — | — |
| CYP79F2 | Syn | 0.0043 | 2.1899 | 0.1327 | 0.2699 |
| | Nonsyn | 0.0010 | — | — | — |
| 83A1 | Syn | 0.0158 | 0.0095 | 0.1358 | 0.0200 |
| | Nonsyn | 0.0002 | — | — | — |
| GSTF11 | Syn | 0.0036 | 0.3609 | 0.0975 | 0.2668 |
| | Nonsyn | 0.0013 | — | — | — |
| GSTU20 | Syn | 0.0249 | 0.083 | — | NA |
| | Nonsyn | 0.0021 | — | — | — |
| SUR1 | Syn | 0.0086 | 0.2733 | 0.0950 | 0.0651 |
| | Nonsyn | 0.0024 | — | — | — |
| UGT74B1 | Syn | 0.0158 | 0.2212 | 0.1492 | 0.1215 |
| | Nonsyn | 0.0035 | — | — | — |
| UGT74C1 | Syn | 0.0047 | 0.2201 | 0.1332 | 0.1532 |
| | Nonsyn | 0.0010 | — | — | — |
| SOT17 | Syn | 0.0030 | 0.1000 | 0.1430 | 0.1701 |
| | Nonsyn | 0.0003 | — | — | — |
| SOT18 | Syn | 0.0012 | 0.9741 | 0.1298 | 0.1212 |
| | Nonsyn | 0.0011 | — | — | — |

Note.—Nonsyn, nonsynonymous; Syn, synonymous; NA, not applicable; A. lyrata has no ortholog of GSTU20.
have experienced balancing selection, we cannot rule out that interpretation. If that interpretation is true, balancing selection at SUR1 is difficult to account for by the principle that selection preferentially targets enzymes with high flux control. Instead, the elevated nonsynonymous variation would presumably reflect some other phenomenon, for example if SUR1 displays greater pleiotropy than other genes in the pathway, which has been implicated as a possible enzyme property that correlates with rate of evolution. For example, Ramsay et al. (2009) found that the rate of evolution in the plant terpenoid biosynthetic pathway was correlated with inferred levels of pleiotropy (Ramsay et al. 2009). We note that as the only enzyme involved in the production of both aliphatic and indolic GLSs, SUR1 has the potential to incur greater pleiotropy than the other pathway enzymes.

Given the evidence for positive or balancing selection in the flux-controlling enzyme, we expect that this signal will overwhelm any signature of stronger purifying selection and indeed, our results do not find stronger purifying selection in the flux controlling enzyme CYP79F1. The dN/dS ratio is much lower for genes CYP83A1 and GSTU20 than it is for CYP79F1. The site-frequency spectrum of CYP79F1 has many high-frequency–derived amino acid substitutions, which is not expected of an enzyme under strong purifying selection.

In summary, the data for evaluating the relationship between flux control and patterns of selection are equivocal but suggestive. The observation of elevated nonsynonymous polymorphism indicates that the first pathway enzyme with the greatest control over flux is subject to selection, as is expected. This may be due to several adaptive substitutions in portions of the species range, because there is no clear sequence signature of balancing selection. However, it is not clear whether this pattern reflects balancing selection or repeated adaptive substitution. With the exception of SUR1, the other pathway genes exhibit no detectable flux control and no evidence of either balancing or positive selection, as expected. The expectation of lack of selection on SUR1 is consistent with the data, but we cannot completely rule out the possibility that it is also subject to balancing selection. If so, this is not explained by the expected relationship between flux and selection, but possibly by greater pleiotropy because of its operation in both the aliphatic and indolic pathways.

Conclusion

Although there is growing evidence from signatures of selection that flux control may be unevenly distributed and focused at the beginning of pathways, we do not know how general our results may be. For one, it is conceivable that flux control could change quickly on evolutionary time scales, or even during development. The simulations of Wright and Rausher (2010) actually showed that flux control can evolve to be centered in different enzymes, although the probability is high that flux control will be vested in the first enzyme. However, studies in anthocyanin (Rausher et al. 1999; Lu and Rausher 2003) and plant terpenoid (Ramsay et al. 2009) pathways suggest that patterns of sequence rate variation have persisted over very long evolutionary time scales (since the divergence monocots and dicots). Flux control may play some part in shaping these patterns of sequence rate variation, but this possibility remains to be demonstrated.

The results of this study are consistent with theoretical predictions of flux control and sequence rate variation, in that flux control is unevenly distributed and that the beginning of a pathway often shows majority flux control. Although we do not know the range of environmental and physiological conditions to which these results may be generalized, studies of pathway rate variation in plants and animals suggest that these patterns may be common.

Supplementary Material

Supplementary figures S1–S3 and tables S1–S8 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors thank the Salk Institute Genomic Analysis Laboratory for providing the sequence-indexed Arabidopsis T-DNA insertion mutants. They also thank Jill Anderson and three anonymous for helpful discussion and comments. This work was supported by the National Institutes of Health award R01 GM086496 to T.M.O.; National Science Foundation Doctoral Dissertation Improvement grants 1011167 to C.O.M. and T.M.O., and 1110445 to C.L. and T.M.O.; and National Science Foundation grant DEB0841521 to M.D.R.

Appendix

We show here how the flux control coefficient was estimated for each enzyme. A standard result from metabolic control theory is that the flux control coefficient for enzyme i in a linear pathway with n enzymes, λi, is

$$\lambda_i = (1/E_i)/(1/E_i + 1/E_1 + \cdots + 1/E_n)$$

(A.1)

where $1/E_i = M_iK_i/V_i$. $M_i$ is the Michaelis–Menten parameter for enzyme i, $K_i$ is the equilibrium constant for the initial substrate and the product of enzyme i, and $V_i$ is the $V_{max}$ parameter for enzyme i (Kacser and Burns 1973).

Equation (A.1) can be re-expressed as

$$\lambda_i = (1/E_i)/(1/E_i + \Phi_i)$$

(A.2)

where $\Phi_i = (\Sigma_{j=1}^n E_j - 1/E_i$. Dividing equation (A.2) by $\Phi_i$ yields

$$\lambda_i = (M_iK_i/V_i\Phi_i)/[(M_iK_i/V_i\Phi_i) + 1]$$

(A.3)

and rearrangement yields

$$(M_iK_i/V_i\Phi_i) = V_i[\lambda_i/(1 - \lambda_i)].$$

(A.4)

Also from metabolic control theory, flux for a pathway in which the final product is sequestered is given by

$$F = [SK_{in}]/[(M_iK_i/V_i) + \Phi_i]$$

$$= [SK_{in}V_i/\Phi_i]/[(M_iK_i/V_i) + V_i].$$

(A.5)
where \( S \) is the concentration of the initial substrate (Kacser and Burns 1973). Substituting equation (A.4) into (A.5) yields

\[
F = \frac{F_{\text{HET}}}{(1 - \lambda_i) + 1}.
\]

(A.6)

Suppose the amount of enzyme \( i \) is reduced in heterozygotes to a fraction \( C_i \) of that in WT individuals. Then the new \( V_{\text{max}} \) of the reaction catalyzed by enzyme \( i \) is \( hV_i \) (Siegel 1975). The flux of this reaction is then

\[
F_{\text{HET}} = \frac{[SK_{\text{in}}/\Phi_i]/[(\lambda_i/(1 - \lambda_i) + 1)}

(A.7)

The relative flux for heterozygotes is then

\[
F_R = F_{\text{HET}}/F = C_i/(\lambda_i/(1 - \lambda_i) + 1)\]\n
(A.8)

Given that \( F_R \) and \( C_i \) are known, equation (A.8) can be solved for \( \lambda_i \) to yield

\[
\lambda_i = \frac{F_R(1 - C_i)}{[F_R(1 - C_i)]}.
\]

(A.9)

References


Cao J, Schneeberger K, Ossowski S, et al. (17 co-authors). 2011. Substituting equation (A.4) into (A.5) yields (A.6)

Suppose the amount of enzyme \( i \) is reduced in heterozygotes to a fraction \( C_i \) of that in WT individuals. Then the new \( V_{\text{max}} \) of the reaction catalyzed by enzyme \( i \) is \( hV_i \) (Siegel 1975). The flux of this reaction is then

\[
F_{\text{HET}} = \frac{[SK_{\text{in}}/\Phi_i]/[(\lambda_i/(1 - \lambda_i) + 1)}

(A.7)

The relative flux for heterozygotes is then

\[
F_R = F_{\text{HET}}/F = C_i/(\lambda_i/(1 - \lambda_i) + 1)\]\n
(A.8)

Given that \( F_R \) and \( C_i \) are known, equation (A.8) can be solved for \( \lambda_i \) to yield

\[
\lambda_i = \frac{F_R(1 - C_i)}{[F_R(1 - C_i)]}.
\]

(A.9)
Flux Control and Evolution of the Glucosinolate Pathway  


